Statistical Analysis on the Data Derived from Receptor Mediated Endocytosis Simulations

Md Muhtasim Billah

Executive Summary

Among multiple endocytic pathways adopted by biparticles (drugs or viruses) to enter the cells, receptor mediated enocytosis (RME) is the most common and well studied one. Previously, it was established that RME procedure is almost always accompanied by a coat-protein called clathrin and thus clathrin mediated endocytosis (CME) was interchangeably used with RME. But only three decades ago, researchers found that there might be RME pathways that don't require any coat-protein like clathrin and those were labeled as clathrin independent endocytosis (CIE). Thus, a lot of things are still unclear about this phenomenon and research in this field is evolving, both in the experimental and computational fields. But, due to its intricate nature, the experimentalists find it extremely difficult, either *in vivo* or *in vitro* to capture the characteristics of both CME and CIE and thus computational studies were undertaken to remedy this issue. While there are a few numerical methods (continuum, discrete, mesoscale) that can capture the particle binding and membrane evolution, the one adopted for this study is a Merkov Chain Monte Carlo (MCMC) type probabilistic simulation. Using this tool, several design parameters have been already studied which can provide important guidelines for analyzing virus entry to the cell as well as located drug delivery.

The entire RME process can be further divided into several stages such as particle attachment to the cell via ligand-receptor interation (A), membrane deformation (B), vesicle budding (C) and finally pinching off of the matured vesicle from the inner leaflet of the membrane (D). These four steps are shown in Figure 1(a). The difference between CME and CIE is that, for CME the steps B, C, D are governed by clathrin but for CIE it is not. What lies responsible for driving the CIE process is only hypothesized but not confirmed yet. Thus, studying this field is very important since it facilitates cell entry of multiple bioparticles in various living bodies as discovered by the researchers. The numerical models mentioned before can be modified to simulate both CME and CIE. But unfortunately, some of the existing models face issues while mimicking the RME scenario of a living body. They find it difficult to set appropriate values for two of the very important design parameters- the receptor length, L_{rec} and the reaction cutoff distance, d_c. The current simulation studies are based on these two parameters.

Reaction cutoff distance, d_c is the minimum distance required to form a bond. This means that the free tip of the ligand and receptor has to be within this range. It can be understood from the Figure 1(b) where the blue-black colors represent non-bonded ligand-receptor pairs ($d > d_c$) and the red-green colors represent the bonded ligand-receptor pairs ($d < d_c$). Since the real life procedure occurs at nanoscale, experiments often report a wide range of reaction cutoffs spanning from 0.1 to >30 nm and same is reflected by the computational models. The receptor length, L_{rec} can vary from 3 to 12 nm depending on the type of the receptor. But numerical models like coarse-grained molecular dynamics (CGMD) often model the receptors as spherical beads just like the lipid molecule of the membrane which cannot represent the actual RME process. It is because the recpetors always have their physical length and they are almost never equal to the thickness of the lipid membrane. It was found from the MCMC simulations, that these two parameters, both individually and collectively can affect the simulation dynamics significantly, both for CME and CIE. To demonstrate these effects, while new parameters can be defined that would represent the effectiveness and efficiency of the entire simulation to achieve endocytosis, statistical methods can provide a quantitative and more deterministic way. Equipped with R programming language, using statistical concepts such as completely randomized design (CRD), comparison of population means, analysis of variance (ANOVA) and, simple linear regression (SLR), the findings from the numerical simulations can be reestablished which will remain to be the purpose of this project.

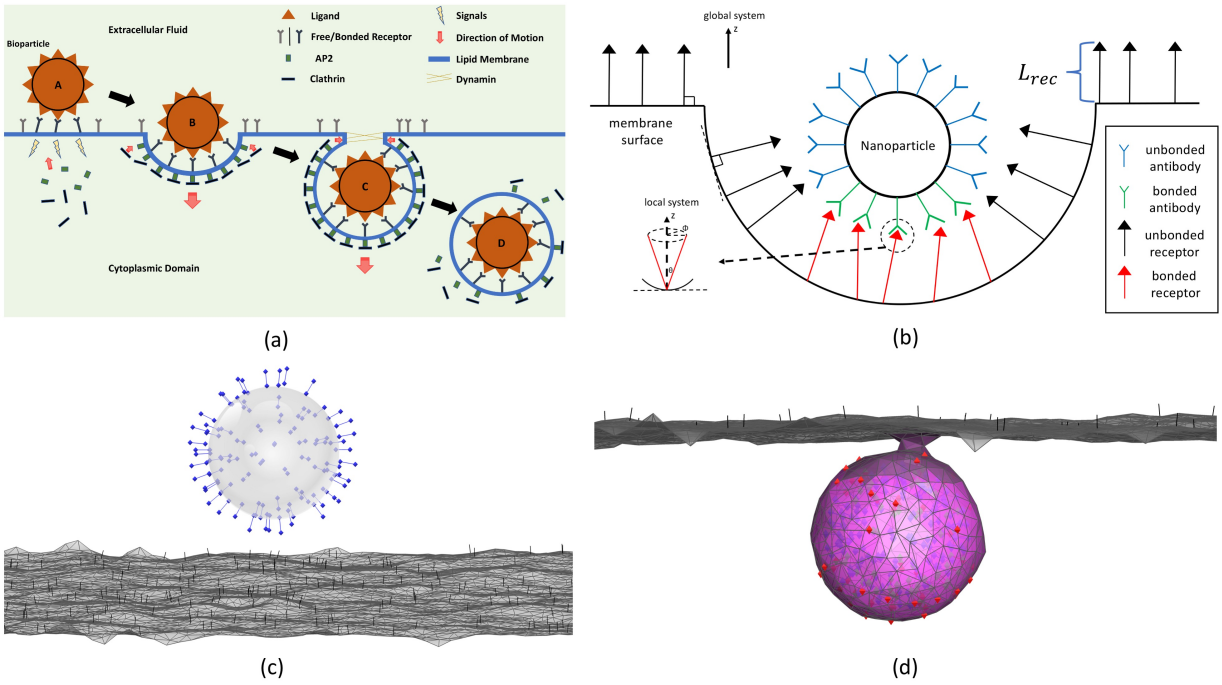

Figure 1: Receptor mediated endocytosis (RME). (a) A schematic that shows the step by step endocytosis procedure. (b) A shematic showing the particle-membrane binding model. (c) A capture from the initiation of the simulation. (d) A capture of the equilibrium profile at the end of the simulation.

1. Objectives and Expectations

The two RME pathways (CME and CIE) are strictly dependent on many biophysical and biochemical parameters. Among these parameters, two very important ones are L_{rec} and d_c. MCMC simulations have been performed on a stochastic model of this biophysical system to study the effect of clathrin as well as these two aforementioned parameters. From the simulations, an attempt was made to determine the effects of these parameters on the whole RME process by looking at the number of bonds and as well the equilibrium profiles of the cell membrane. While, those visualizations are enough to come to a conclusion, adopting a more quantitative measure to determine their significant effects can be more helpful. Thus, a more deterministic approach can be made based upon well established statistical concepts to reevaluate the conclusions made from the stochastic simulations. This will be the main objective of this project and the expectations will be met if the conclusions from the statistical approacch matches the one from the numerical simulations.

2. Data Collection

To measure the effectiveness of a parameter setup for particle internalization (essentially this means- to experience endocytosis), a new parameter $Max\ R_{Bonds}$ is defined which is calculated for each simulation. This quantity is the ratio of the maximum number of bonds formed during the entire simulation to the total number of available ligands on the particle surface and has range of $0 \leq Max\ R_{Bonds} \leq 1$. Higher the value, higher the chance for a complete endocytosis. For lower values, there might be partial or no endocytosis. This will be the response variable for all the analyses where each simulation will be an experimental unit. Two factors are considered which are the receptor length, L_{rec} (factor A) and the reaction cutoff distance, d_c (factor B). The number of replicates will be constant at 3 for any treatment which means that data from

three independent simulations will be used for each treatment. Based on these, three different statistical analysis will be performed.

2.1 CME vs CIE

The first analysis is directed towards the comparison of two population means (based on the response $Max\ R_{Bonds}$ value) of CME and CIE. Clathrin plays a significant role for most of the receptor mediated endocytosis ocurring in nature. It's importance has been proven to be unquestionable in numerous experimental and computational research. The stochastic model allows us to capture this clathrin-inclusion effect by comparing between simulations that have clathrin with the ones that do not. Under the same computational setup, CME (Treatment 1) shows full internalization of particle where CIE (Treatment 2) only shows partial wrapping. This attribute can be determined by the value of $Max\ R_{Bonds}$. Naturally, CME gives a higher value than CIE and this can also be verified statistically by comparing the population means of CME and CIE based on the response $Max\ R_{Bonds}$. The two factors were kept constant for both the treatments and thus have no effect here. The data are given in Table 1.

Nomenclature	**Treatment 1**	**Treatment 2**
Factor 1, L_{rec}	9.3	9.3
Factor 2, d_c	0.9	0.9
$Max\, R_{bonds}$	$0.759, 0.741, 0.759$	$0.049, 0.068, 0.049$
Treatment Means	$\bar{x}_{1.} = 0.753$	$\bar{x}_{2.} = 0.055$
Treatment SD	$s_1 = 0.01039$	$s_2 = 0.01097$

Table 1: Data for comparing the population mean of CME with CIE.

2.2 L_{rec} vs d_c

Since, CIE is a recent discovery comapring to the CME and many questions remains unanswered till date, along with experiments, computational studies are being carried out to determine parameter effects for this procedure. As mentioned before, two very important ones are L_{rec} and d_c. Depending on how the system is modeled, a very wide range of values for these two parameters have been used in literature which sometimes are not in line with real life scenarios. Based on their values, the simulation outcomes can vary sigificantly and that is why they are important to study. Simulations have been performed for only CIE focusing on these two parameters which will be the two factors of the experiment that are to be compared. The most reasonable values for them have been considered as the control group (Treatment 1). For Treatment 2, d_c and for Treatment 3, L_{rec} has different values from the control group. For Treatment 4, both of these values are different from control group. The response is the $Max\ R_{Bonds}$ value, just as before, and the number of replicates are 3. The data for $Max\ R_{Bonds}$ found from the simultions are provided in Table 2. Based on this dataset, a two-way ANOVA will be performed.

Nomenclature	**Treatment 1**	**Treatment 2**	**Treatment 3**	**Treatment 4**
Factor 1, L_{rec}	9.3	9.3	0.3	0.3
Factor 2, d_c	0.9	5.9	0.9	5.9
$Max\, R_{bonds}$	$0.049, 0.068, 0.049$	$0.494, 0.488, 0.531$	$0.315, 0.401, 0.500$	$0.969, 0.969, 0.981$
Treatment Means	$\bar{x}_{1.} = 0.055$	$\bar{x}_{2.} = 0.504$	$\bar{x}_{3.} = 0.405$	$\bar{x}_{4.} = 0.973$
Treatment SD	$s_1 = 0.01097$	$s_2 = 0.02329$	$s_3 = 0.09258$	$s_4 = 0.00693$

Table 2: Data for comparing two factors, L_{rec} and d_c.

2.3 Effect of d_c on CIE

The reaction cutoff distance, d_c has potential influence on the $Max\ R_{Bonds}$ and with an increase in d_c, there is an increasing trend observed for $Max\ R_{Bonds}$ which indicates to a positively linear relation among these two variables. And apparently, this relation is never affected by the value of L_{rec}. Thus, to analyze this from a more statistically intensive perspective, linear model can be applied on the dataset provided in Table 3. The receptor length, L_{rec} has been fixed on 0.3 nm and the reaction cutoff, d_c has been varied to adopt 6 different values. The number of replicates for each value is 3 which gives a total of 18 data points.

L_{rec}	d_c	$Max\ R_{Bonds}$
0.3	0.9	0.31, 0.36, 0.38
0.3	1.9	0.52, 0.51, 0.51
0.3	2.9	0.64, 0.65, 0.61
0.3	3.9	0.67, 0.66, 0.72
0.3	4.9	0.98, 0.67, 0.88
0.3	5.9	0.97, 0.97, 0.98

Table 3: Data for comparing the effect of d_c on CIE when $L_{rec} = 0.3$ nm.

3. Statistical Methods

The principles of completely randomized design (CRD) such as control, randomization, replication have been applied properly while running simulations for each study. For the first analysis, the principles for comparing two population (true) means will be used. For the second analysis, the basics of two factor analysis of variance (ANOVA) will be applicable. And for the last dataset, simple linear regression (SLR) will be utilized to investigate the relationship between the two variables under consideration. The level of significance was be fixed at $\alpha = 0.05$ for all the analysis.

4. Results and Discussions

4.1 Comparison of Population Means between CME and CIE

Since, the sample sizes are very small ($m = n = 3$) and the population standard deviations ($\sigma_1,\ \sigma_2$) are unknown, to compare the population means of $Max\ R_{Bonds}$ for CME and CIE dataset, case-III (t-based procedure) will be applicable.

Here,
Null Hypothesis, $H_0 : \mu_1 - \mu_2 = 0$.
Alternative Hypothesis, $H_a : \mu_1 - \mu_2 > 0$.
Level of significance, $\alpha = 0.05$.
Degree of freedom,

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{\left(\frac{0.01039^2}{3} + \frac{0.01097^2}{3}\right)^2}{\frac{(0.01039^2/3)^2}{3-1} + \frac{(0.01097^2/3)^2}{3-1}} = \frac{2.940874e^{-6}}{6.474253275e^{-10} + 8.0455179e^{-10}} = 2025.42 \approx 2025$$

Test statistic,

$$t = \frac{\bar{x}_{1.} - \bar{x}_{2.}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{0.753 - 0.055}{\sqrt{\frac{0.01039^2}{3} + \frac{0.01097^2}{3}}} = 80.014695$$

Here, from the table $t_{\alpha,\ \nu} = 1.645$.
Thus, it is clear that,

$$t > t_{\alpha,\ \nu}$$

Also,

$$p - value = 0 < 0.05$$

This means that there is high statistical evidence to reject the null. It means that there is a significant difference between the population means of $Max\ R_{Bonds}$ between CME and CIE. This means that, with clathrin inclusion, the value of $Max\ R_{Bonds}$ will always be significantly higher. This proves that, clathrin plays a crucial role for endocytosis.

4.2 Two Factor ANOVA for CIE

For this analysis, there are two factors. Factor A is the receptor length, L_{rec} and factor B is the reaction cutoff, d_c. Factor A has two levels ($I = 2$) which are two different lengths of the receptor- 9.3 nm and 0.3 nm. Factor B has two levels as well ($J = 2$) which are two different values of reaction cutoffs- 0.9 nm and 5.9 nm. Number of replicates for each factor combination is $K = 3$.

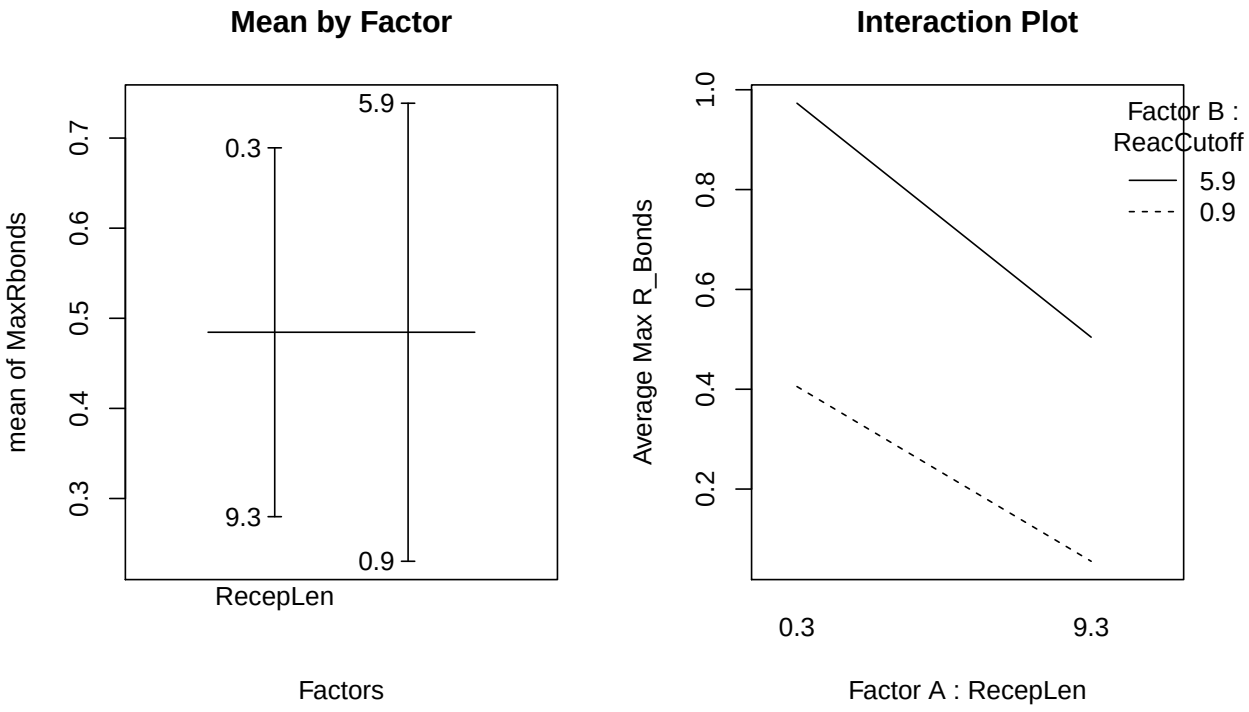

Figure 1: Plot for showing the mean of response by factors (left) and plot for checking any potential interaction among the factors (right).

The plot of the mean response with factors shows the range of response values for the two factors at their different levels and also indicates that the grand mean is close to ~0.5. From the interaction plot, it is apparent that there isn't enough evidence for interactions among these two factors since the lines are roughly parallel. So, a two way additive fixed model with replicates will be applicable for performing two-way ANOVA on this dataset.

The model,

$$X_{ijk} = \alpha_i + \beta_j + \varepsilon_{ijk}$$

Here,
α_i = The main effect of factor A at level i.
β_i = The main effect of factor B at level j.
$k = 1, 2, ...K$.

From the ANOVA output, provided in Table 4, the p-values indicate that both the factors RecepLen (L_{rec}) and ReacCutoff (d_c) have significant main effects on the response. The null and alternative hypothesis for factors A and B are as below.

For factor A, $H_0 : \alpha_1 = \alpha_2 = 0$ vs H_a : At least one $\alpha_i \neq 0$.
For factor B, $H_0 : \beta_1 = \beta_2 = 0$ vs H_a : At least one $\beta_j \neq 0$.

Source	DF	SS	MS	F	p-value
RecepLen	1	0.5027	0.5027	155.3	$5.57e-07$
ReacCutoff	1	0.7752	0.7752	239.6	$8.59e-08$
Residuals	9	0.0291	0.0032		
Total	11	1.3070			

Table 4: ANOVA table for the two-way fixed additive model.

Another way to investigate the main effects of the factors is constructing simultaneous confidence intervals (CI) using Tukey's multiple comparisons method. Based on this method, 95% confidence intervals are constructed for the mean difference in response for the factor levels. If 0 is within the interval, it can be said with 95% confidence that there is no mean difference between the factor levels. But if 0 isn't within the range, the opposite is assumed which indicates to the existence of main effects of the factors. The results are provided in Table 5 and they match with the conclusion found from the ANOVA table.

Factor	Levels	Difference	Lower Limit	Upper Limit	Main Effect?
A: L_{rec}	$\alpha_2 - \alpha_1$	-0.4093	-0.4836	-0.3350	YES
B: d_c	$\beta_2 - \beta_1$	0.5083	0.4340	0.5826	YES

Table 5: Tukey's simultaneous 95% CIs for testing the main effects.

The main effects can also be tested using Tukey's underscore (T-method). This procedure assumes fixed effects and it is only applicable for equally replicated treatments. The procedure is valid if interactions are not significant. All these assumptions hold true for the current analysis. The resulting underscore plots are provided in Figure 2.

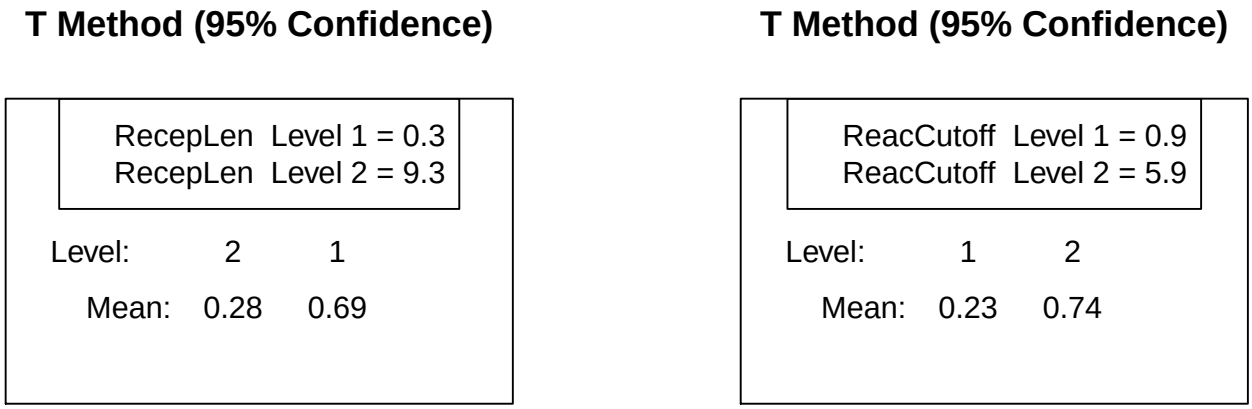

Figure 2: Tukey's underscore (T-method) for testing the main effects.

For factor A (L_{rec}), the margin of error is $w = 0.074295$. Then, (level 2 mean + w) = 0.35 which indicates that level 2 and 1 are different. This means that they have significant effects. The margin of error, w is same for factor B (d_c) and (level 1 mean + w) = 0.30 which indicates that level 1 and 2 are different. This means that they have signigicant effects as well. Thus, it is beyond any doubt that both the factors A and B have significant main effects on the response.

Since the main effects of the factors have been established, to quantify there effects, these parameters need to be estimated. The parameters estimated by the two-way additive model is provided in Table 6. Using these values, fitted reponse values for any combinations of the factor levels can be calculated.

Parameter	$\hat{\mu}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Estimate	0.4350	0.4093	−0.4093	−0.5083	0.5083

Table 6: Model parameters estimation.

ANOVA is based on two basic assumptions which are the normality of the residuals and the constancy of the variance. For checking these assumptions, both graphical measures and statistical testing can be adopted.

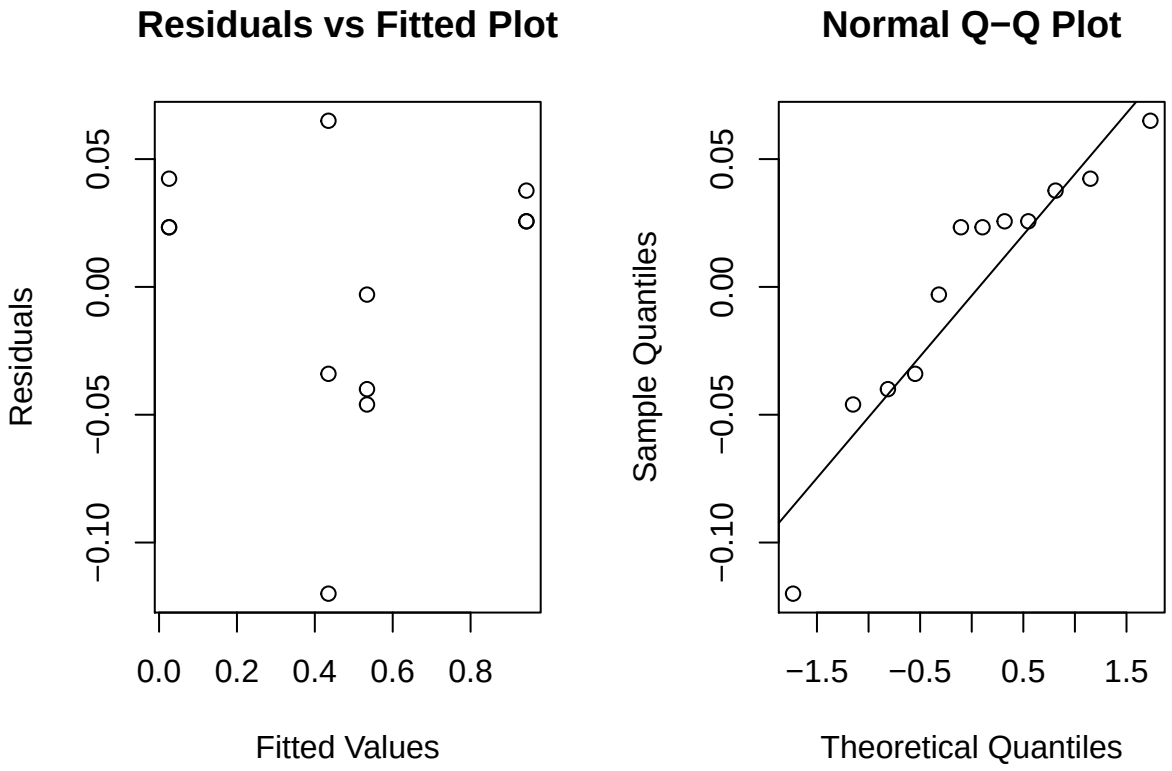

Figure 3: Graphical measures for checking assumptions.

Test Name	Assumptions	Statistic	P-value	Decision
Shaphiro-Wilk	Normality of the residuals	W = 0.82134	0.01655	Non-normal residuals
Breusch-Pagan	Homogeneity of residual variance	BP = 7.0333	0.07084	Constant variance

Table 7: Diagnostic tests for checking assumptions.

From the residuals vs fitted plot in Figure 3, it seems that for Treatment 2, the residuals are more widely spread than the other three treatments, which indicates to a possibility of the violation of the assumption of

constant variance. From the normal probability plot (Q-Q plot), it seems that most of the residuals are not very close to the line and on top of that, they move even further away at the tails. This refers to a possibility of the violation of the normality assumption. For a more deterministic approach, one statistical testing per each assumption has been done on the dataset and the results are provided in Table 7. From test statistics and p-values, it is evident that assumption of normality of the residuals are violated but the assumption of constant variance remains intact.

4.3 Regressing $Max\ R_{Bonds}$ on d_c by SLR

The dataset considered to analyze the relationship between the variables d_c and $Max\ R_{Bonds}$ is provided in Table 3 where d_c is the explanatory variable and $Max\ R_{Bonds}$ is the response variable. Just by looking at the dataset, it is noticeable that $Max\ R_{Bonds}$ increases with an increase in d_c. Now, for a more quantitative measure of how this data behaves if fit into a linear model, a simple linear regression is perfromed over this dataset which is given in Figure 4. From the figure, it is clear that the regression line has a positive slope (β_1) and a positive intercept (β_0). The line fits the data points very well becuase there is adequate linearity among the data points. The summary of the simple linear regression model is given in Table 8.

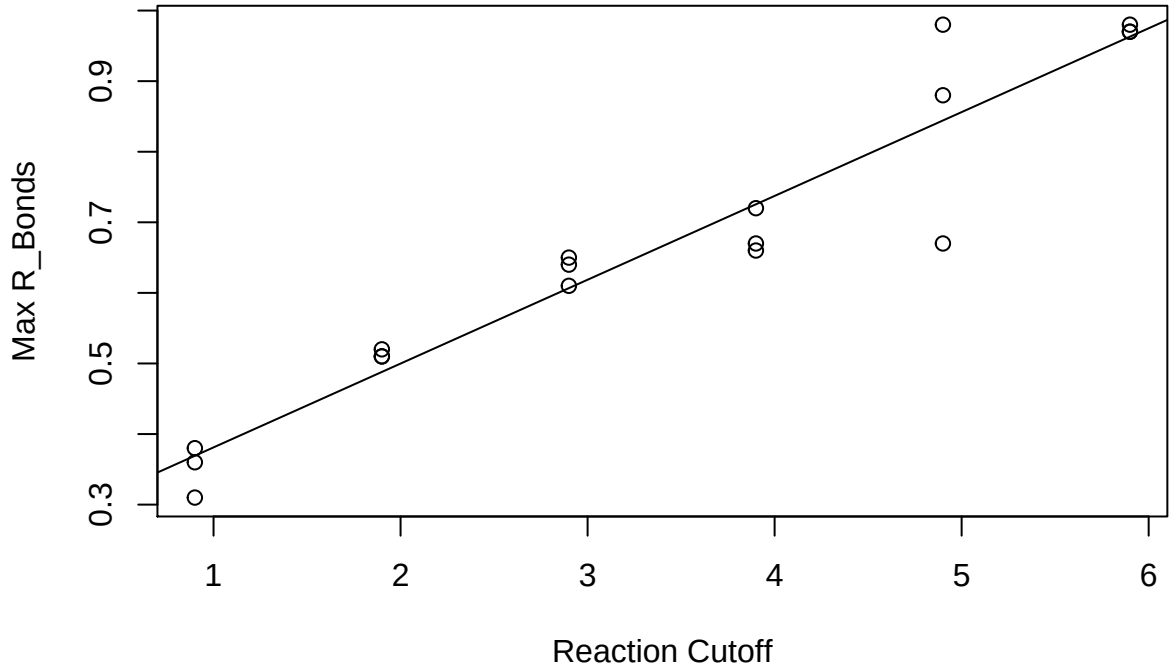

Figure 4: Regression line fitted on the dataset.

Coefficients	Estimate	Confidence Interval	Std. Error	t value	Pr(>|t|)
Intercept, β_0	0.2623	(0.1905, 0.3341)	0.0339	7.746	8.41e-07
Slope, β_1	0.1188	(0.0999, 0.1376)	0.0089	13.343	4.36e-10

Table 8: Summary of the coefficient estimates from the SLR model.

From Table 8, it is found that the values for the parameters are $\beta_0 = 0.2623$ and $\beta_1 = 0.1188$ with an standard error of 0.0339 and 0.0089 respectively. With 95% confidence, the confidence intervals for β_0 is $(0.1905, 0.3341)$ and for β_1 is $((0.0999, 0.1376))$. The value of β_0 indicates that for a reaction cutoff of 0, the maximum ratio of bonds will be only 0.2623. This doesn't have any physical meaning since for zero reaction cutoff there won't be any bond formation which will make the maximum ratio of bonds zero. The value of

the slope, $\beta_1 = 0.1188$ means that for a unit increase in reaction cutoff, the maximum ratio of bonds will go up by 0.1188. This positive slope exhibits the linearly positive relationship among the explanatory and response variables. The fitted regression model now has the form,

$$\hat{y} = 0.2623 + 0.1188x$$

From the fitted model, the maximum ratio of bonds can be predicted for any given reaction cutoff since both are continuous variables. For example, for $x = 2.0$, the fitted value of response variable would be $y = 0.4999$ and for $x = 3.2$, we will get $y = 0.6425$. The cofindence and prediction intervals for the fitted line is given in the Figure 5.

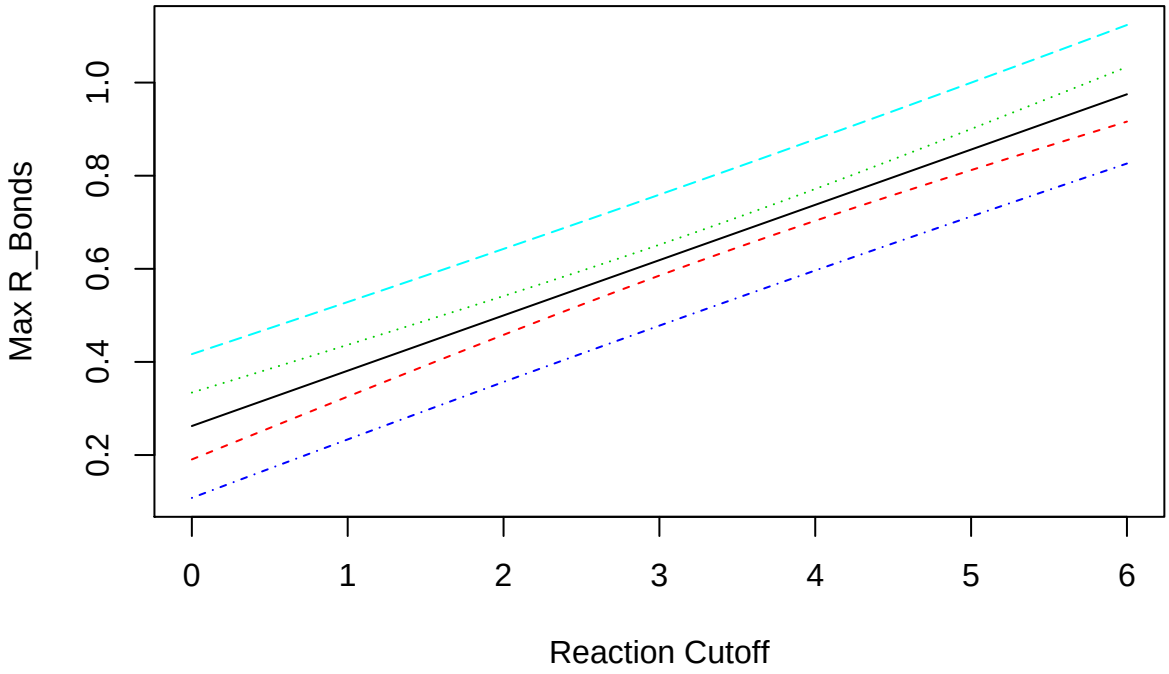

Figure 5: Confidence (red-green band) and prediction (blue-cyan band) intervals on the regression line.

Some other important attributes of the model are given in Table 9. The residual standard error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term ε. Due to the presence of this error term, it's not viable to perfectly predict the response variable from the predictor. The residual standard error is the average amount that the response will deviate from the true regression line. In this dataset, the maximum ratio of bonds can deviate from the true regression line by approximately 0.0645% on average. It's also worth noting that the residual standard error was calculated with 16 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters after taking into account these parameters (restriction). For this case, there was 18 data points and two parameters (intercept and slope).

Attributes	Values	DF
Residual standard error	0.0645	16
Multiple R-squared	0.9175	
Adjusted R-squared	0.9124	
F-statistic	178	(1, 16)
p-value	4.361e-10	

Table 9: Other attributes of the SLR model.

The R-squared (R^2) statistic provides a measure of how well the model is fitting the actual data. R^2 is a measure of the linear relationship between the predictor and response variable. It always lies between 0 and 1 and a number close to 1 indicates that the model can explain the observed variance in the response variable efficiently. For this dataset, the R^2 was found to be 0.9175. This means that roughly 91% of the variance found in the response variable (maximum ratio of bonds) were explained by the predictor variable (reaction cutoff).

Moreover, the F-statistic is a good indicator of whether there is a relationship between the predictor and the response variable. The further the F-statistic is from 1 the better it is. Generally, for large dataset, an F-statistic only a little bit larger than 1 is sufficient to reject the null hypothesis (H_0 : there is no relationship between x and y). The reverse is true as if the dataset is small, a large F-statistic is required to reject the null. For this dataset, the F-statistic is 178 on 1 and 16 DF which is very large and gives a p-value of $4.361e^{-10}$. So, the null is rejected and it indicates that there is strong relation between maximum ratio of bonds, $Max\ R_{Bonds}$ and reaction cutoff, d_c.

Regression model is based on two primary assumptions just like ANOVA which are the normality of the residuals and the constancy of variance. From the residuals vs fitted plot in Figure 6 (left) is is visible that some residuals are more widely spreadout than others which indicates towards a non constant variance. Also, it is apparent from the QQ plot (right) that the data points tend to deviate at both upper and lower tails which indicates non-normality of the residuals.

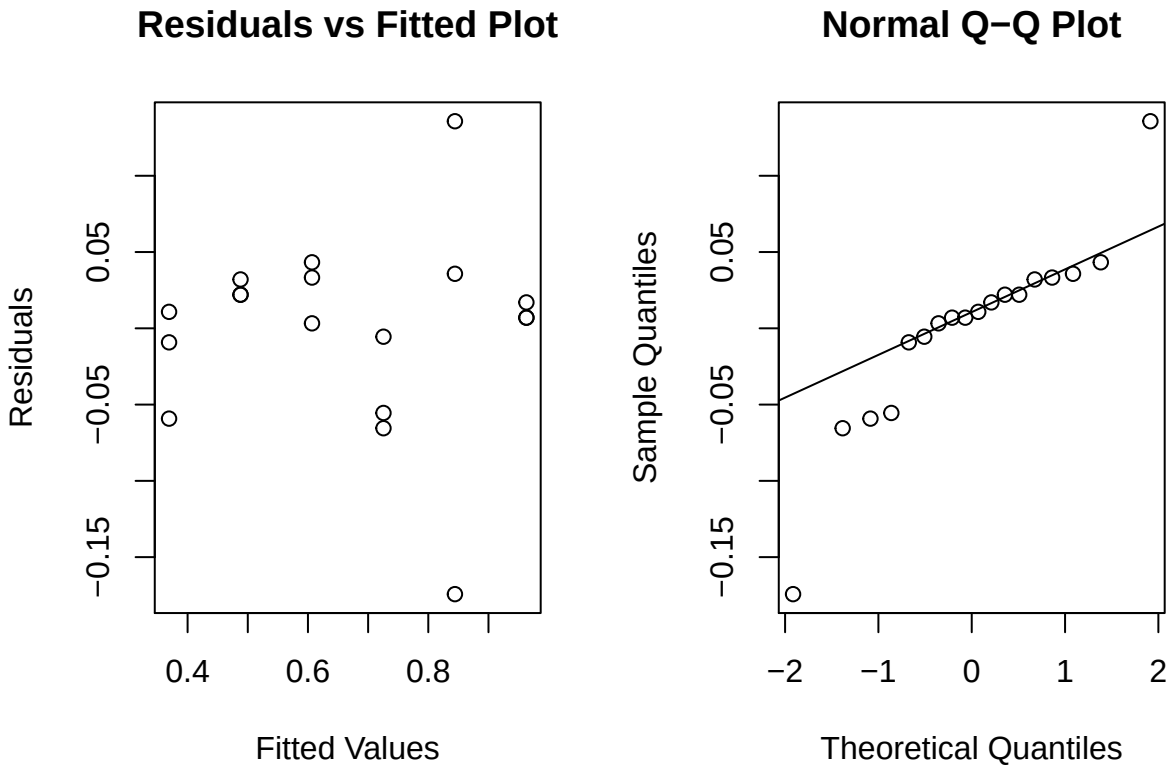

Figure 6: Dagnostic plots for checking regression model assumptions.

To get a better insight about the model assumptions, a test for each has been performed and tabulated in Table 10. The Shapiro-Wilk test confirms the possiblity of non-normality suggested by the QQ plot. And though the residuals vs fitted plot showed different spreads of the residuals, the Breusch-Pagan test indicates that the residuals have constant variance. The Durbin-Watson statistic found from the test is not very far from 2 which indicates that there is no autocorrelations issues in the model which is also verified by the p-value.

Test Name	Assumptions	Statistic	P-value	Decision
Shaphiro-Wilk	Normality of the residuals	W = 0.87284	0.01982	Non-normal residuals
Breusch-Pagan	Homogeneity of residual variance	BP = 1.4046	0.236	Constant variance
Durbin-Watson	Autocorrelation of the residuals	DW = 2.6073	0.8618	No autocorrelation

Table 10: Diagnostic tests for checking the regression assumptions.

5. Conclusions

Receptor mediated endocytosis (both CME and CIE) are crucially important in multiple biological procedures as well as for located drug delivery and thus investigated in both experimental and computational regimes. The data collected from numerical simulations, which were set up according to the principles of a completely randomized design (CRD), were divided into three different datasets to apply statistical analyses on them. The first dataset was used for comparing the CME and CIE process to prove the well-known effect of clathrin. The second dataset was used to compare two factors of CIE, the receptor length and the reaction cutoff, via two-way ANOVA analysis and it was found that both the factors have significant main effects on the response. Simple linear regression (SLR) was performed on the last dataset to analyze the relationship between the reaction cutoff and the response and it was found that they have linearly positive relationship. Model diagnostics for both ANOVA and SLR were performed and it was found that some of the assumptions were violated. While remedial measures could be taken for solving this issue, they weren't carried out due to the limited scope of the project and can remain for a task for its extension in future. The sample size or the number of replicates were very small for all the datasets due to the extreme cost of computational resources. Thus, for future works, a more parsimonious way for data collection or for designing the experiment can be adopted for achieving a better analysis. However, for further insights into RME, more data can be collected on multiple predictors and multiple linear regression can be applied.

Appendix

R code used for analyzing the data.

```
#STAT523_Project
#Md Muhtasim Billah


###getting started
#set working directory
setwd("/Users/muhtasim/Desktop/STAT523/Project")
#read in the dataset
mydata=read.csv("data.csv",header=T,colClasses=c("factor","factor","numeric"))
#manage plot window
par(mfrow=c(1,2))
#means by factor plot
plot.design(mydata, main = "Mean by Factor")
#interaction plot
interaction.plot(mydata$RecepLen,mydata$ReacCutoff,mydata$MaxRbonds,
                 xlab="Factor A : RecepLen", ylab="Average Max R_Bonds",
                 trace.label="Factor B : \n ReacCutoff", main= "Interaction Plot")
#two way additive ANOVA model
model=aov(MaxRbonds ~ RecepLen + ReacCutoff,data=mydata)
# ANOVA table for summary
summary(model)
#estimated parameter
coef(model)
#manage plot window
par(mfrow=c(1,2))
#Tukey's simultaneous difference
TukeyHSD(model,conf.level=.95)
#Tukey's underscore method
source(url("http://math.wsu.edu/math/faculty/jpascual/stat423/R/two-way-T-method.R"))
twoway.t.method(model)
#Diagnostic plots for ANOVA
par(mfrow=c(1,2))
#residuals vs fitted plot
plot(model$fitted, model$res, main="Residuals vs Fitted Plot",
     xlab = "Fitted Values", ylab = "Residuals")
#QQ plot
qqnorm(model$res)
qqline(model$res)

#Alternative options
#source(url("http://math.wsu.edu/math/faculty/jpascual/stat423/R/chap13fcns.R"))
#residplots.lm(model,std.resid=F)

#Breusch-Pagan test
library(lmtest)
bptest(model)


##For simple linear regression
#data setup
```

```r
regdata=read.csv("regdata.csv",header=T)
cutoff=regdata[,1]
max_Rbonds=regdata[,2]
#scatter plot
plot(cutoff,max_Rbonds,xlab="Reaction Cutoff",ylab="Max R_Bonds")
#SLR
model1=lm(max_Rbonds~cutoff,data=regdata)
abline(model1)
#model parameters
s=summary(model1)
#coefficients(model1)
cf=confint(model1,level=0.95)
#prediction intervals
y=max_Rbonds
x=cutoff
predict(lm(y ~ x))
#creating new values for prediction
new <- data.frame(x = seq(0.0, 6.0, 0.25))
predict(lm(y ~ x), new, se.fit = TRUE)
pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")
pred.w.clim <- predict(lm(y ~ x), new, interval = "confidence")
#prediction plots
matplot(new$x, cbind(pred.w.clim, pred.w.plim[,-1]),  type = "l",
        ylab = "Max R_Bonds", xlab = "Reaction Cutoff")
#Diagnostics plots for SLR
#Normal probability plot of Residuals
par(mfrow=c(1,2))
#plot of residual vs x and fitted values
plot(model1$fitted, model1$res, main = "Residuals vs Fitted Plot",
     xlab = "Fitted Values", ylab = "Residuals")
qqnorm(model1$res)
qqline(model1$res)
#Diagnostic tests
#test for normality
library(stats)
shapiro.test(model1$res)
#breusch pagan test
#might need to install lmtest
library(lmtest)
bptest(model1)
#durbin-watson test
dwtest(model1)
```