

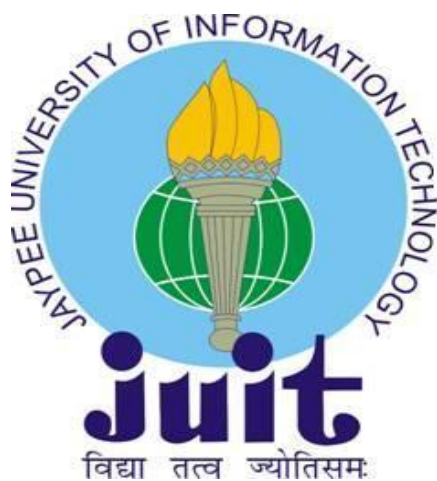
CORRELATING ACTIVITY OF DRUGS WITH LOW SOLUBILITY USING QSAR AND QSPR MODELS

By

Sumita Dutta - 131513

Under the supervision of

Dr. Chittaranjan Rout



Dec 2017

Submitted in partial fulfilment of the Degree of

Bachelor of Technology in Bioinformatics

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, SOLAN

TABLE OF CONTENTS

Serial No.	Title	Page No.
1	Certificate of Originality	3
2	Declaration	4
3	Acknowledgement	5
4	List of figures and Tables	6
5	Introduction	8
6	Review of Literature	15
7	Materials and Methods	18
8	Results and Observations	25
9	Conclusion	39
10	Publications	40
11	References	41

CERTIFICATE OF ORIGINALITY

This is to certify that the project report entitled “**Correlating Activity Of Drugs With Low Solubility Using QSAR AND QSPR Models**”, submitted by **Sumita Dutta** in partial fulfilment for degree of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology, Waknaghat has been carried out under supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of any other degree or diploma.

Signature of Supervisor

Dr.Chittaranjan Rout

(Associate Professor)

Department of Biotechnology and Bioinformatics

JUIT, Solan

Date:

Declaration

I hereby declare that the project titled “**Correlating Activity Of Drugs With Low Solubility Using QSAR AND QSPR Models**” is submitted as a project work has been carried out by me under the guidance of Dr. Chittaranjan Rout at Jaypee University of Information Technology, Solan. Any further extension, continuation or use of this project has to be undertaken with prior express written consent from the Supervisor, Jaypee University of Information Technology, Solan-173234.

I further declare that the project work or any part thereof has not been previously submitted for any degree or diploma in any university.

Signature

Name : Sumita Dutta

Date:

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. I would like to express my sincere gratitude towards my project supervisor, **Dr.Chittaranjan Rout**, Associate Professor, Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, for his support and unsurpassed guidance, and also for his unmatched patience with which he chose to nurture young student like me at every step. I am truly grateful to him.

I am highly indebted towards **Ms. Nupur S. Munjal** (PhD Scholar) for providing necessary information and relentless help regarding the project for her kind co-operation and encouragement which helped me in completion of this project.

I am very grateful to **Dr. Sudhir Sayal**, Associate Professor, Acting Head of Department, Department of Biotechnology and Bioinformatics for his invaluable suggestions, scientific attitude, continuous support, guidance, cooperation and encouragement. I owe a lot of gratitude towards former HOD **Dr. R.S. Chauhan** for his unending support. Also I am highly thankful to **DR. Vinod Kumar** , Vice- Chancellor,JUIT , **Dr. Samir DevGupta** , Director and Dean academics and **Maj Gen Rakesh Bassi (Retd.)**, Registrar,JUIT for their immense support during the project.

A special note of thanks and appreciation is towards the technical assistance provided by **Mrs. Somlata Sharma** and **Ms. Sonika Gupta** for their support during the project work.

At the end this work would not be worthy without being thankful towards my inspiring parents **Mr. Pankaj Kumar Dutta** and **Mrs. Geeta Dutta** who have always encouraged me and supported me to keep up the good work.

Thank you

List of Figures and Tables

Figure No.	Caption	Page No.
1	Solubilisation Process	11
2	Chemsketch	21
3	Open Babble	22
4	Datawarrior	22
5	Molinspiration	23
6	Dragon 7	23
7	R 3.2.2	24
8	Weka	24
9	Regression Analysis	25
10	Scatter Plot	34
11	Regression Output	36

12	Trend between experimental QSAR and QSPR	37
13	Trend between model QSAR and QSPR	38
<u>List of Tables</u>		
1	Classification of BCS	9
2	Experimental solubility values from literature	26
3	Experimental IC50 values from literature	27
4	QSAR results from Weka	31
5	QSPR results from Weka	32
6	Data from model and data warrior	33
7	Data from model and regression equation	35

Chapter 1- Introduction

1.1 Drug Solubility

Therapeutic efficacy of a drug depends on the solubility of the core compound. According to IUPAC, solubility refers to the proportion of a chosen substance in designated solvent. The dissolution of a drug compound is dependent on its structure and solution conditions. Structure of a molecule decides crystal properties, mostly physicochemical, which determine aqueous solubility[1]. Mixes with poor dissolution convey a higher risk of failure during discovery and development since inadequate dissolvability might trade off alternative property measures, veil additional unwanted properties, impact each pharmacological approach of the compound, and lastly might influence production capability of the compound. Potential difficulties arising from poor solubility:

- cut-back target specificity.
- Low bioavailability in veterinary studies

Low micromolar water solubility is often acceptable just for very potent and/or permeable compounds. The dissolution of drug is key feature that plays a terribly important role in dissolution of the drug after oral administration. In drug discovery, the amount of insoluble drug candidates has accumulated largely, with upcoming 70% of recent new drug formulations show poor water solubility. For these drug candidates, poor solubility and poor dissolution within the Gastro Intestinal fluids may be a hindering issue to the in vivo availability biologically, in the case of oral administration..

Paclitaxel is a chemotherapeutic drug that shows significant outcomes in cancer treatment. The drug is given intravenously because of lower solubility issues. Oral administration isn't most popular because the compound gets crystallizes once it comes in contact with the blood stream and thus show adverse effect, additionally, it may result fatal. Many studies are being carried out for the solubility improvement. I have taken it as the reference drug so as to validate the results of the developed model and also relate various therapeutic classes (anticancerous, antifungal, antiparasitic) on the solubility grounds. A drug is taken in account of highest solubility when the maximum strength is solubilized in 250 ml. BCS or the Bio pharmaceuticals classification system is scientific classification of a medicinal substance

based on its dissolution coefficient and organal permeability that relates to in vitro solubility and in vivo bioavailability of drug merchandise. The BCS classifies it as. (**Table 1**)

BCS class	Solubility	Permeability	Absorption pattern	Examples
I	High	High	Well absorbed	Metoprolol, Diltiazem, Propranolol
II	Low	High	Well absorbed	Phenytoin, Nifedipine, Danazol
III	High	Low	Variable	Cimetidine, Acyclovir, Captopril
IV	Low	Low	Poorly absorbed	Hydrochlorothiazide, Taxol , Furosemide

Table 1

The reliableness of QSPR and QSAR mode is usually tough to quantify because of the issues of accessing the modalities used to build the models. QSAR are data - series specific.

I have discerned the structure-activity relationships of all molecules so that a new potential drug may be tested by fitting it in this model which does not have any solubility issue. Also we can relate various drug molecules from different therapeutic classes on the grounds of Solubility.

1.2 QSAR and QSPR:

Correlation of molecular characteristics with physicochemical properties is efficient outcomes of the quantitative structure property-relationship (QSPR) models and data mining methods. Reliable models like GSE (general solubility equation)[2] and AQUAFAC were developed for solubility prediction requiring experimental data[3]. Structure-based methods were more reliable, fast and applicable for the predicting the dissolution coefficients of the new drug formulations..

QSAR model establishes a statistical relationship between the biological activity exerted by a compound and a group of parameters determined from the structures of the compounds. The central assumption of a QSAR model is that the numerical value of a given biological activity depends on the structure of those molecules [4]. If the structure for the receptor is derivable, it is attainable to analyse binding partners and to deduce 3D QSAR models from the withstood parameters [5]. The precise definition of 3D QSAR continues to be lacking.

Nowadays, an outsized amount of trial and anticipated information with respect to the 3D structure of organic molecules and bio molecules is obtainable. One among key strategies for preparing these informations is Quantitative Structure-Property Relationship (QSPR) modelling. This strategy communicates atoms by means of various numerical esteems (called descriptors), which encode the basic attributes of particles. A short time later the descriptors are used in figuring the physicochemical properties of the molecules. Subtle mathematical approaches including partial least-squares (PLS) method, multiple linear regressions (MLR), and other techniques such as artificial neural networks are used for these calculations. QSPR modelling has turned out to be exceptionally discussed in chemical, biological and pharmaceutical research. The fundamental reason is that the forecast or test assurance of physicochemical properties (e.g. dissociation constants, partition constants, dissolvability, lipophilicity, and natural movement) is exceptionally testing and QSPR gives an efective approach to appraise these qualities Moreover, the contributions for QSPR models are sub-atomic structures, unreservedly and essentially out there for a significant number of particles. Bioinformatics is targeted on massive molecules like proteins and nucleic acids. It utilizes computer science and studies the similarities and differences among these molecules, their necessary patterns and relations between their structure and biological performance . The most common application fields of bioinformatics are healthcare, genetics and biology. Chemoinformatics deals with small organic molecules (mainly drug-like compounds) and is powerfully connected with mathematics and informatics. The foremost necessary chemoinformatics topics talk over with analyzing the similarity among molecules, checking out the molecules in structural databases, finding potential drug molecules and studying the relations relating structure of molecules and their characteristic features. Chemo informatics is very beneficial for the pharmaceutical industry and drug design.

Molecular descriptors assume a basic part in science, pharmaceutical Sciences, ecological insurance arrangement, wellbeing examination and quality control, getting used to foresee

biological and physicochemical properties of particles (QSAR/QSPR) and for virtual screening of atom libraries

Exploiting these molecules which were taken from literature, motive of the project is to form a QSAR model for the solubility prediction. One can find a mathematical relationship, or quantitative structure activity relationship, between the two using the structural parameters. Statistically, if carefully validated can then be used to predict the Solubility-activity variation in synchronisation. The Dragon molecular descriptors were extracted for each derivative. Important statistical methods, such as akiake information criteria (AIC) and variance inflation factor (VIF) multicollinearity indicators (7) were used to identify independent descriptors.

1.3 Solubilisation process

The method of solubilisation involves the breaking bonds within the substance, which in turn separates the molecules of the solvent to produce house within the solvent for the substance.

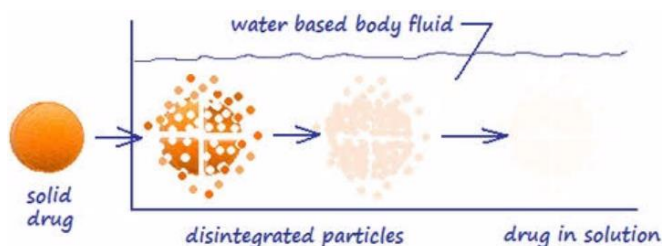


Figure 1: Solubilisation Process

1.4 Importance of enhancement of drug solubility

Oral intake is the conveniently utilised route of drug delivery. However, the most challenging aspect in designing oral dosage is due to low bioavailability.

Moreover, any drug to be absorbed should be present in the form of variety of associate degree solution in the receptor site.

The enhanced solubility is required as:

- To realise the preferred proportion of medicine in circulatory system for proper medicinal response.

- Poorly water soluble medication oftentimes need high doses and hence influencing therapeutic plasma concentrations after oral administration.
- Not good for generic medicines.
- Acidity or basicity is also a major part of the solubility criteria.
- Low solubility leads to slow drug absorption which which inturn causes GI toxicity.

1.5 Prodrugs

Prodrugs derivatives of drug molecules which are biologically reversible and undergo an some organic transformations done experimentally to release the active parent drug which helps in achieving the pharmacological results to which we desired for.. They have some undesirable properties

- Poor dissolution
- Low lipophilicity
- Chemical unstable
- Poor distribution across cell membranes
- No site specificity

I have an interest in rising the solubility of the drug supported the structural annotation and thus we tend to think about the prodrugs. Prodrugs unharness same active parent drug once undergoing associate degree accelerator and/or chemical transformation in vivo and thus provide an efficient medicine response. Objective of the project Solubility values of all the prodrugs that were taken from literature and that square measure nanomolar active in vivo square measure potential drug molecules Many approaches have been used to improve solubility like microionization, solubilisation and complexion with polymers, salt formation, and adding surfactants. However, all these strategies suffer from certain limitations[8]. I tend to overcome this problem by exploiting prodrugs.

1.6 Objective of the project

Solubility values of all the prodrugs which were taken from literature and which are nanomolar active in vivo are potential drug molecules.[9]. Many lipid formulations and other solubility improvements in different solvents except for water are being done for solubility

enhancement. Yet we need to improve the water solubility so as to reach at par with the scientific enhancement in drug formulation in the human body. Therefore, I have done an analytical study for activity and solubility values after performing substitution of certain groups which would further give a relating factor for aqueous solubility of a compound.

1.7 Computational models for predicting solubility

Jorgensen and Duffy studied a group of organic nonelectrolytes and based on which Jain and Yalkowsky projected a General solubility equation (GSE) was then employed for estimating their aqueous solubility. The sole inputs employed in the GSE were some coefficients such as Melting point (MP) in Celsius and the water octanol partition coefficient (Kow). Ran et al. predicted drug solubility by the General Solubility Equation (GSE). Based on the data set chosen by Jorgensen and Duffy, it seems that the general solubility equation is simpler and easier to use than the Monte Carlo simulation equation for estimations of the water solubility of un-ionized drugs. Melting points were obtained from The Merck Index and CHEMFINDER. Measured and calculated partition coefficients (MLOGP and CLOGP, respectively) were obtained from the CLOGP software [10]. Solubilities were calculated from both MLOGP and CLOGP by the general solubility equation.

GSE is simpler providing a more correct dissolution prediction of an equivalent group of biochemical entities. The general solubility equation calculated the molar aqueous solubility as:

$$\log S_w^{\text{solid}} = 0.5 - 0.01(\text{MP} - 25) - \log \text{Kow}$$

Where,

- **Kow**- octanol-water partition coefficient
- **MP**- Celsius melting point

1.8 Geometry optimization

Geometry Optimisation is the name for the procedure that attempts to find the configuration of minimum energy of the molecule. Quantum mechanical calculations are performed to get the optimized geometry of the molecules. The parameters on which the electronic wave function for polyatomic molecule depends – bond length, bond angle and dihedral angles of

rotation around single bonds. Four main approaches for the calculation of molecular properties are ab initio methods, semi - empirical methods, the density- functional method and the molecular mechanics method.

For studying and optimizing the structural data of prodrugs we have used the semi-empirical method **PM6** using the software **Gaussian 09**. Parameterisation method 6 is a recently developed semi-empirical method. PM6's parameters coupled with AMPAC'S wonderful capabilities when applied to any dataset can treat the system very efficiently. Comparing PM6 to existing traditional systems proves to be producing much better results.

Chapter 2- Review of Literature

The approximate solubilities of 65,500 compounds in dimethylsulfoxide, using Kohonen neural network were studied by Balakin et al. to rank into partially-soluble and insoluble categories. They achieved a 93% correctly classified dataset yet not clearly differentiable.

Mirmehrabi et al. set up an iterative model requiring the calculation of the compound activity constant and mole fraction solubility, which was later applied to for prediction of three groups of solute. Ruelle et al. delineated speculation which needs the enthalpy of fusion, that must be noted through an experiment or otherwise calculable. Also the stability constant, K_{oi} , for every hydrogen bond interaction for the solution was noted. Mobile order theory has been applied to the solubility of the pesticide diuron with good results. The outcomes relating the Gozalbesetal informational index demonstrated huge disparities with the initially detailed information, as had been resolved for the forecast of blood mind obstruction saturation. The primary wellspring of this change returns again from the altogether different sizes of the informational indexes, with the informational collection arranged by Gozalbes et al. being up to 25-times bigger than some of those utilized as a part of the prior work. The expectations exploiting demonstrative measure abusing analytic assay were similar to those announced by Gozalbesetal et al. Fundamentally, utilizing the informational indexes utilized as a part of the underlying work of alternate techniques they could get very much like coefficients of assurance (r^2) as those revealed. The effect of the database measure on the precision and pertinence of QSPR models is generally valued, and had been said in connection to Caco-2 cell porousness most importantly inside the first work of Gozalbesetal. The Gozalbesetal informational collection of the type of the Caco-2 cell penetrability covers the range - 7.6 to - 3.7; the normal mean unsigned blunder from every one of the 5 models of 0.58 speaks to 15% of the exploratory range, and mirrors the chief hopeful exactness that can be gotten from these models. In all models, the greatest mistake was identified with a proportionate atom, 2-(1-(aminomethyl) cyclohexyl) carboxylic corrosive, generally zwitterionic at unbiased watery pH. None of the models analyzed here incorporate particular descriptors that outline a zwitterion.

Gracin et al. anticipated the dissolvability of paracetamol (4-hydroxyacetanilide) in eight nonaqueous solvents with a standard deviation of 0.41 or 0.29 log units, contingent upon how

useful gatherings were allocated. The extended dissolvability parameter system has been connected to an assortment of mixes. The yield isn't forecasts of dissolvability however figured fits to the solvency parameter condition. For the dissolvability of methyl 4-hydroxybenzoate in 35 nonaqueous solvents, a standard deviation of 0.24 log units is discovered .

Hilal et al. have depicted the expectation of solubilities through the SPARC salvation demonstrate, that is presently unreservedly accessible . For 162 solute nonaqueous dissolvable blends the RMSD was 0.53 log unit (for 507 solubilities in water, RMSD/40.49).All the anticipated general techniques for expectation of solubilities, aside from that in light of SPARC, require various test amounts for their usage; the majority of the systems requiring the enthalpy of combination or the Gibbs vitality of combination that are once in a while accessible for drugs. There are distributed gathering commitment strategies which give sensible appraisals to natural exacerbates that don't display strong to-strong stage changes preceding liquefying. Be that as it may, for a given compound, amass commitment techniques give just a solitary assessed enthalpy of combination; it is unreasonable right now to anticipate enthalpies of combination for the majority of the polymorphic crystalline structures that a compound may have.

Yingqing Ran et al inferred that in view of the informational collection chose by Jorgensen and Duffy, it appears that the general dissolvability condition is less complex and less demanding to use than the Monte Carlo recreation condition for estimations of the fluid solvency of un-ionized medications.

The reconsidered general dissolvability condition (GSE) proposed by Jain and Yalkowsky is utilized to evaluate the fluid solvency of a gathering of natural nonelectrolytes contemplated by Jorgensen and Duffy. The main sources of info utilized as a part of the GSE are the Celsius liquefying point (MP) and the octanol water segment coefficient (Kow). These are for the most part known, effectively measured, or essentially computed. The GSE does not use any fitted parameters. The normal supreme mistake for the 150 mixes is 0.43 contrasted with 0.56 with Jorgensen and Duffy's computational strategy, which utilizes five fitted parameters. In this way, the overhauled GSE is more straightforward and gives a more precise estimation of fluid dissolvability of a similar arrangement of natural mixes. It is additionally more exact than the first form of the GSE.

Kolar et al. have proposed that dissolvable determination could be encouraged through setting up a general information base of solubilities, specifically one in view of mono- and bi-utility benzene subordinates in a progression of solvents, yet this appears to be excessively confined. Steroids, barbiturates, and medications in view of numerous different frameworks would be barred. It might well be that pharmaceutical scientists will utilize various distinctive strategies to anticipate dissolvability in solvents, and it is the reason for the present work to depict another general strategy.

Ketan T. Savjani et al. worked in setting that Low watery solvency is the significant issue experienced with definition improvement of new substance elements and for the bland advancement. Any medication to be ingested must be available as arrangement at the site of retention. Different strategies are utilized for the upgrade of the solvency of ineffectively dissolvable medications which incorporate physical and synthetic modifications of medication and different techniques like molecule estimate diminishment, precious stone designing, salt arrangement, strong scattering, utilization of surfactant, complexation, et cetera. Choice of dissolvability enhancing strategy relies upon sedate property, site of ingestion, and required measurement shape qualities.

Appropriate determination of dissolvability upgrade strategy is the way to guarantee the objectives of a decent definition like great oral bioavailability, lessen recurrence of dosing and better patient consistence joined with an ease of generation. Determination of technique for dissolvability upgrade relies on tranquilize attributes like solvency, concoction nature, softening point, retention site, physical nature, pharmacokinetic conduct et cetera, dose frame prerequisite like pill or case plan, quality, quick, or modified discharge et cetera, and regulative requirements like most extreme day by day dosage of any excipients and additionally sedate, endorsed excipients, diagnostic exactness etc.

Chapter 3- Materials and methods

3.1. Materials

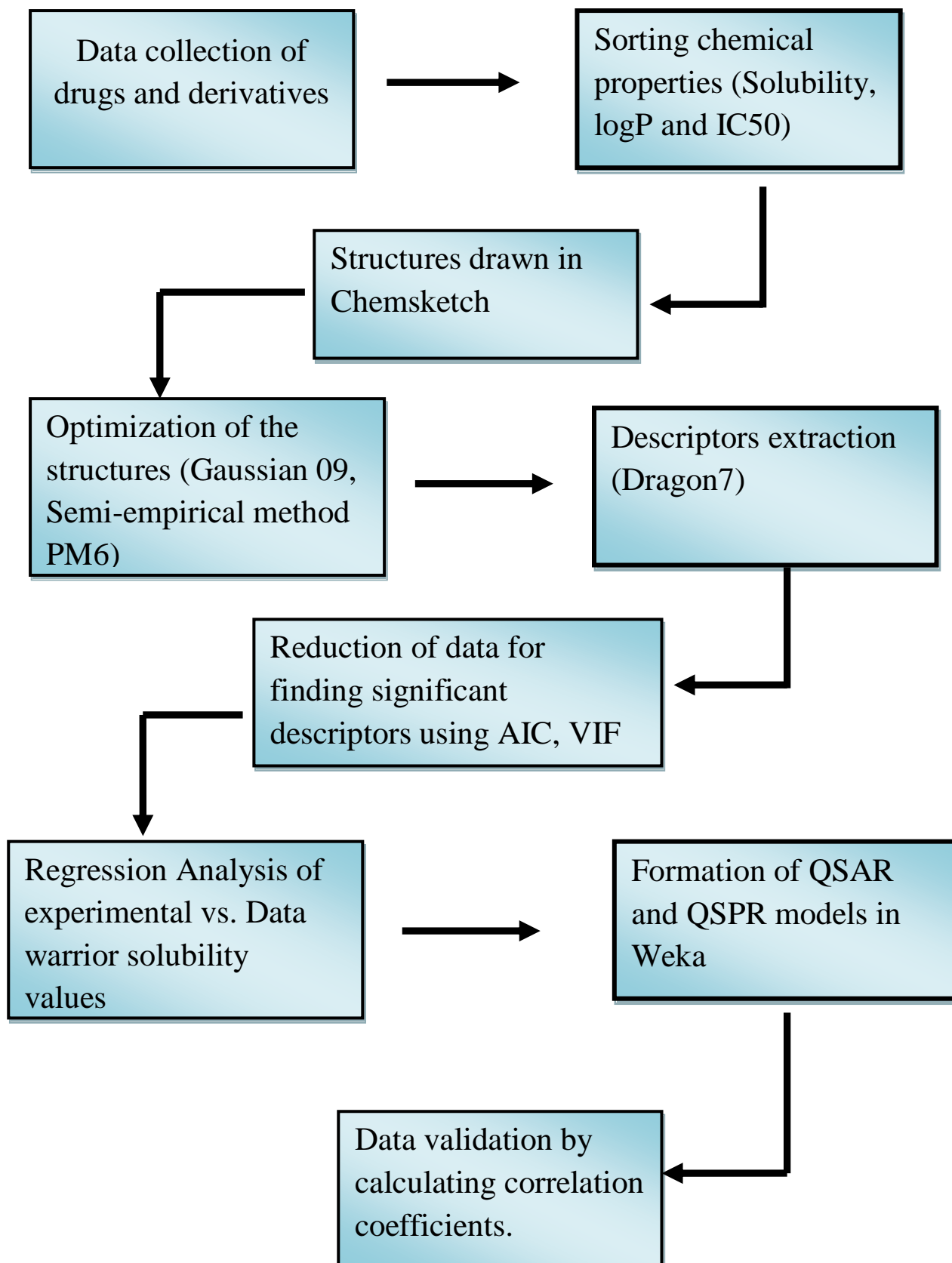
In Bioinformatics, materials are the computational tools which aid us in data analysis and annotation. Methods are the implication of such tools onto a specific dataset (structural data in this case). The materials used for this study are either online or offline downloadable versions of the bioinformatics softwares and databases. All of these are easily accessible and are documented to be giving reliable results.[11]

- **Drug bank:** Drug Bank consolidates definite medication information like concoction, pharmacological and pharmaceutical information with complete medication target data. Its broad medication and medication target information has empowered the disclosure and repurposing of various existing medications to treat uncommon and recently distinguished diseases. Also, 4270 non-repetitive protein (i.e. medicate target/compound/transporter/bearer) groupings are connected to these medication passages.
- **ChemSketch:** A molecular modelling program used to make and adjust pictures of substance structures. It is programming that permits atoms and sub-atomic models showed in two and three measurements, to comprehend the structure of concoction bonds and the idea of the utilitarian gatherings. It has a few formats with particles and useful gatherings with the likelihood to include content and utilize different devices to enhance preparations made by the product.
- **Open Babel:** Open Babel is a chemical tool stash intended to talk the numerous dialects of concoction information. It's an open, community oriented task enabling anybody to seek, change over, examine, or store information from sub-atomic demonstrating, science, strong state materials, natural chemistry, or related territories
- **Data Warrior:** For the calculation of logP and logS i.e. water octanol and the solubility values respectively.
- **Molinspiration:** For the calculation of logP i.e. water octanol value.

- **Dragon 7:** For calculation of descriptors. Descriptors extracted in 30 different groups.
- **R-3.2.2:** A language and environment for statistical computing and graphics. It gives a wide assortment of measurable and graphical methods
- **Weka:** For classification of descriptors

3.2. Methods

❖ The Protocol



3.2.1 Data Collection:

- Structure data was collected from literature. Also activity data and other biological features of the molecules (Prodrugs) were collected from literature.

3.2.1 Data Analysis:

- **Drug bank:** Structures of the parent drugs were downloaded.
- **Chemsketch:** The substitution of functional groups, alkyl, aryl groups and to the core groups are done so as to prepare the structures for prodrugs.

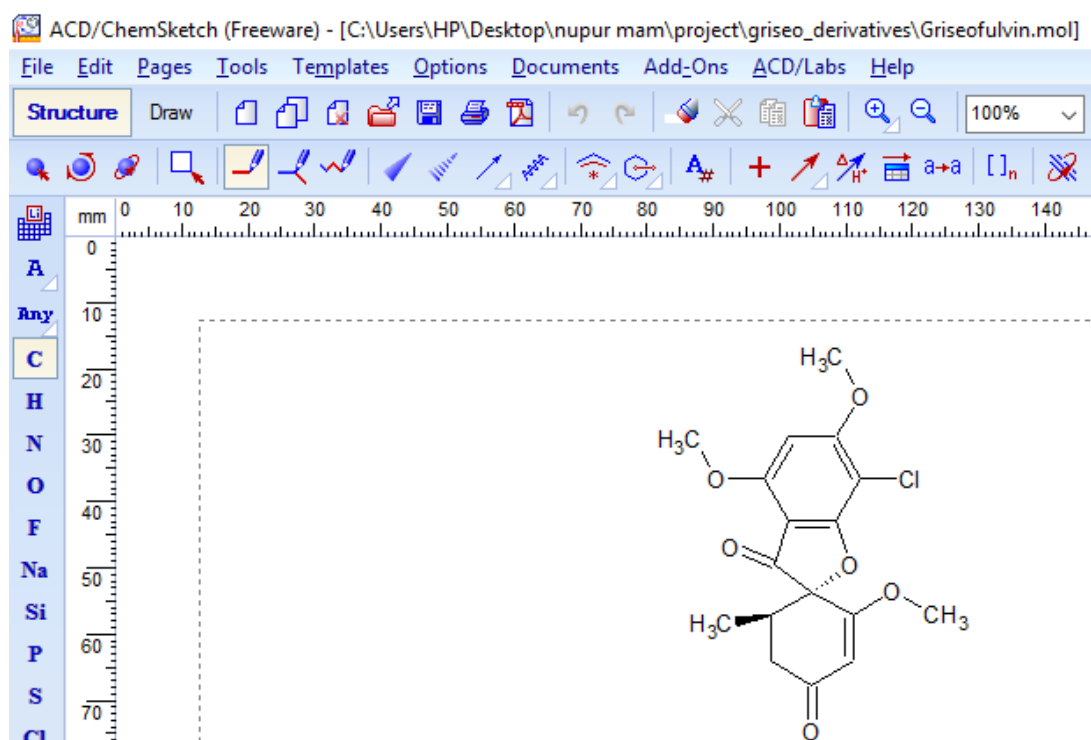


Figure 2: Chemsketch

- **Open babel:** Changed the file formats. Initially chemsketch produced .Mol files, then in optimisation .xyz files were used, datawarrior and molinspiration needed .sdf files and dragon7 used .mol2 file.

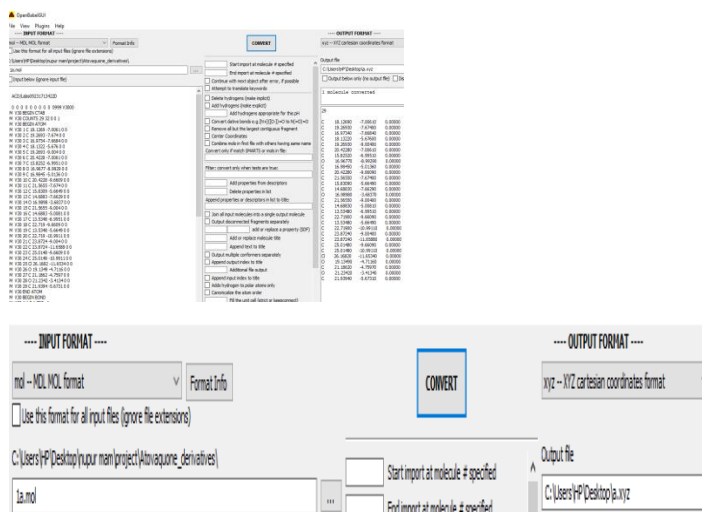


Figure 3: Open Babel

- **Data warrior:** log S and logP values were determined by taking .sdf files as input.

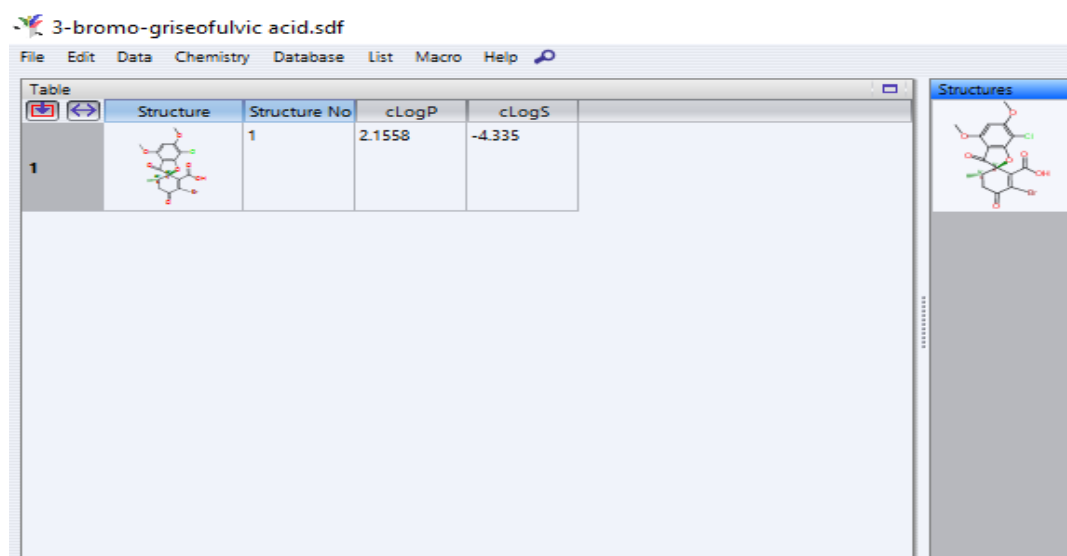
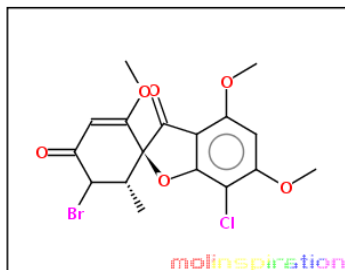


Figure 4: Data Warrior

- **Molinspiration:** Again water octanol coefficient i.e. log P was calculated and validated with the data warrior results.

molinspiration

originalSMILES O=C2c3c(O[C@]12C(=CC(=O)C(Br)[C@H]1C)OC)c(Cl)c(OC)cc3OC
 miSMILES: O=C2c3c(O[C@]12C(=CC(=O)C(Br)[C@H]1C)OC)c(Cl)c(OC)cc3OC



[Molinspiration property engine](#) v2016.10

miLogP	2.03
TPSA	71.08
natoms	25
MW	431.67
nON	6
nOHNH	0
nviolations	0
nrotb	3
volume	311.51

[Get data as text](#) (for copy / paste).

[Get 3D geometry](#) BETA

This was request 1 out of 1000 available this month for your site 128.199.33.160
 With technology from Molinspiration you can easily setup similar service also directly on your intranet.
 Comments or questions ? See our [FAQ](#) and do not hesitate to provide feedback or contact us by email !

[New molecule](#) [Predict bioactivity](#) [About properties](#) [MyMolecules](#) [Molinspiration home](#)

©2017 Molinspiration Cheminformatics [Terms of service](#)

Figure 5: Molinspiration

- **Dragon 7:** The descriptors were calculated from .mol2 file. We got 30 descriptor files for each of the 36 files.

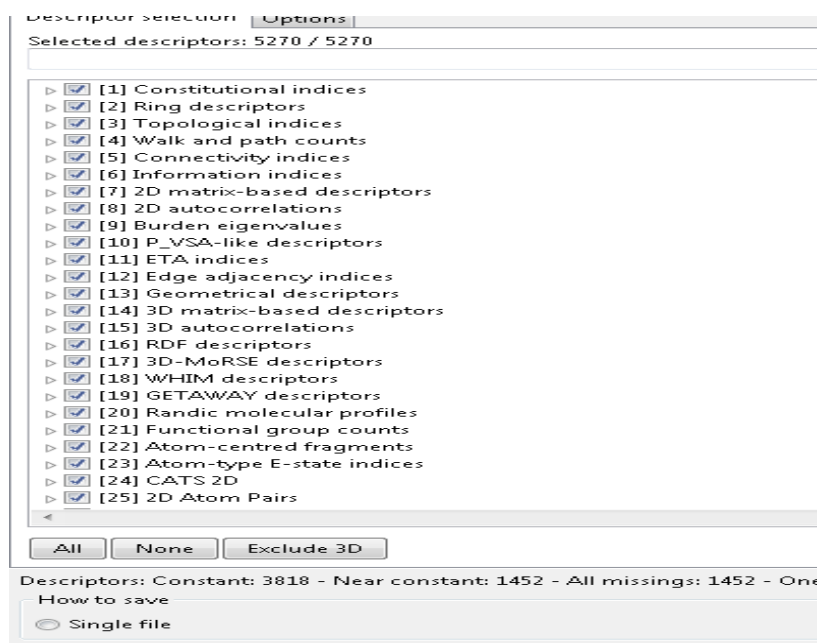


Figure 6: Dragon 7

- **R-3.2.2:** R script for the formation of subgroups of the descriptor groups was compiled and executed in R-3.2.2. that divides each of the group into eight subgroups on the basis of correlation and covariance among the descriptors are pvhchv, pvhclv, phchv, phchv, pmchv, pmclv, zchv, zclv.

Independent descriptor selection was done in R-3.2.2 by applying the multicollinearity indicators AIC (akiake information criteria), VIF (Variance inflation factor). QSPR model was developed with the those independent descriptors

```

99- {
100-   xm2<-xm1[, -id2]
101- }
102-
103- xm2
104- # If number of column in xm2 is zero then stop here.
105-
106- dat1 <- cbind(data.frame(Y=y),as.data.frame(xm2))
107- fit1 <- lm( Y~. , data=dat1 )
108- summary(fit1)
109- summary(fit1)$coefficient[,1]
110- summary(fit1)$coefficient[,4]
111- summary.aov(fit1)
112- fit2<-step(fit1, direction="backward")
113- summary(fit2)
114- summary.aov(fit2)
115- #step(fit2, direction="backward")
116- anova(fit2,fit1)
117-
118- #####
6:17 (Top Level) :
R Script :
insolve C:/Users/HP/Desktop/nupurmani/
5.211 -4.136 -0.131 4.291 20.115

efficients:
Estimate std. Error t value Pr(>|t|)
(Intercept) -84.844 41.738 -2.033 0.0502 .
x1 -18.905 4.139 -4.567 6.56e-05 ***
x2 531.764 153.616 3.462 0.0015 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.285 on 33 degrees of freedom
Multiple R-squared: 0.5407 Adjusted R-squared: 0.5210

```

Figure 7: R 3.2.2

- **Weka-3.6.11:** Linear regression-S 0 –R 1.0E-8 module was used for the formation of regression model. CSV files for the independent descriptors for each of the descriptor group was done and the R² and Q² values were obtained and the equations and RMSE values were calculated for each of the descriptor groups.

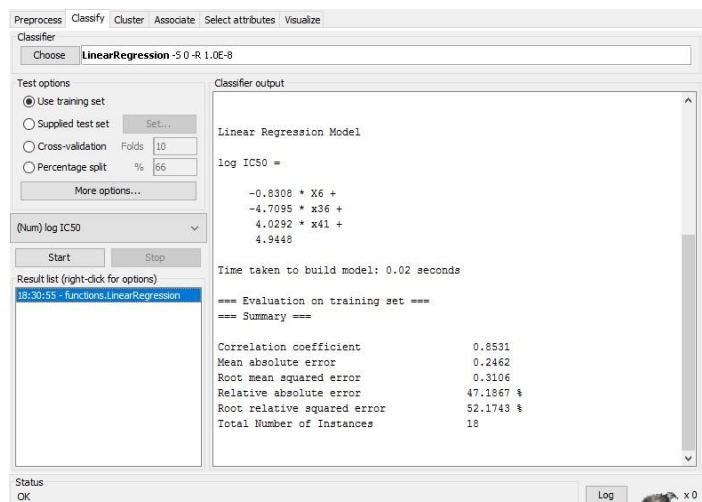


Figure 8: Weka

- **Regression Analysis:**

I have done regression analysis for the solubility values obtained from model and those calculated from datawarrior. This would validate the datawarrior results as well as give a regression equation for the calculation of actual solubility value.

Similarly, for QSAR data I calculated regression values for the experimental IC50 and the one obtained from model.

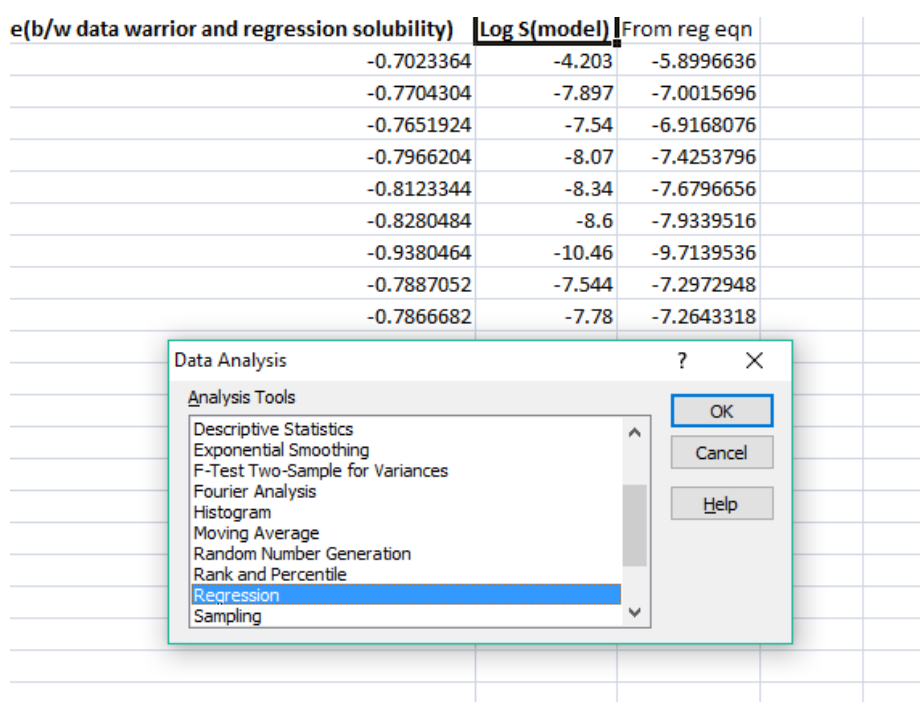


Figure 9: Regression Analysis

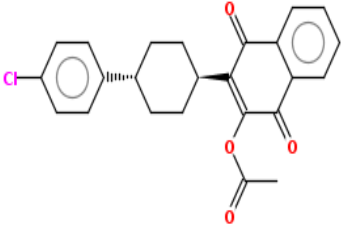
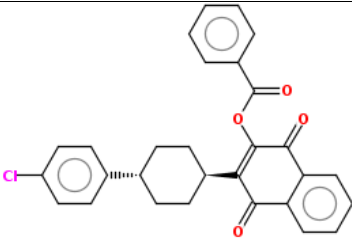
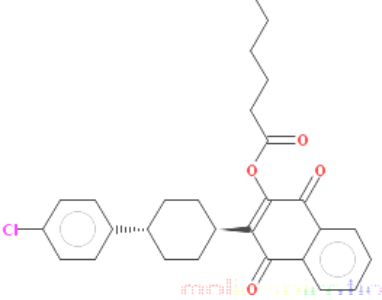
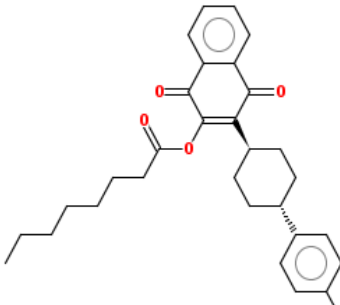
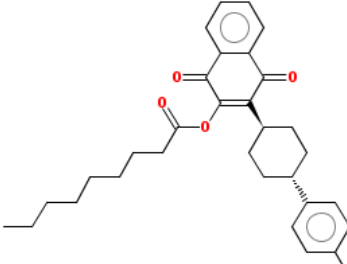
Chapter 4- Results and Observations

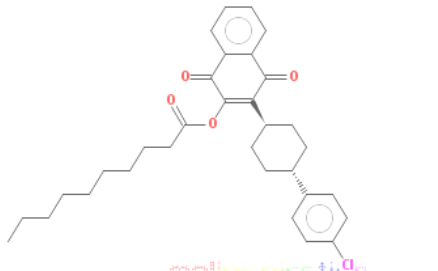
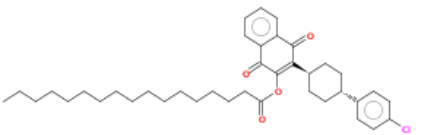
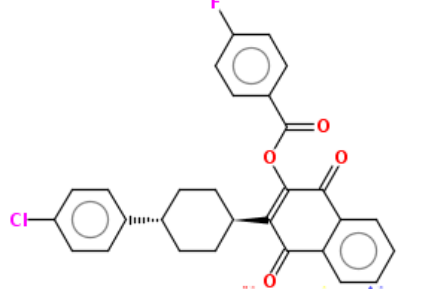
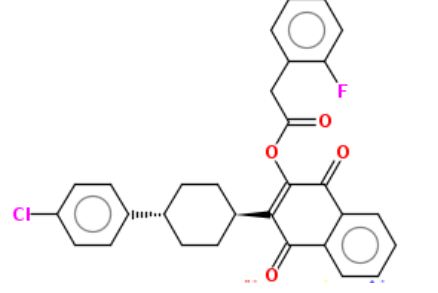
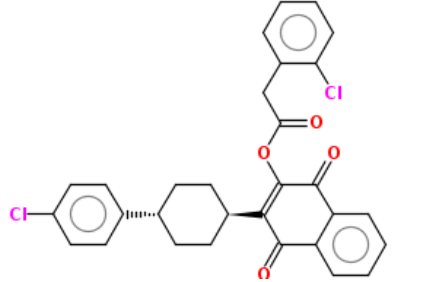
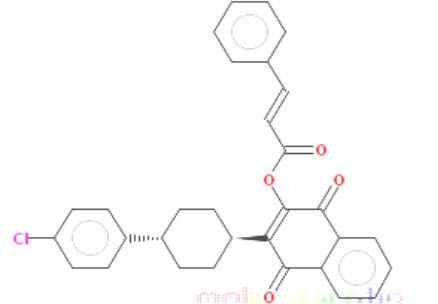
The drugs selected for the study were:

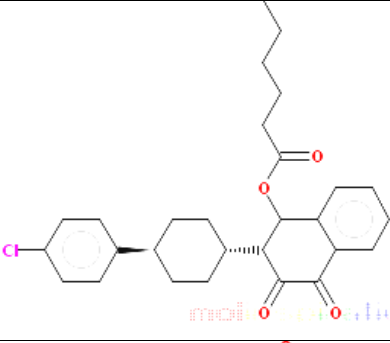
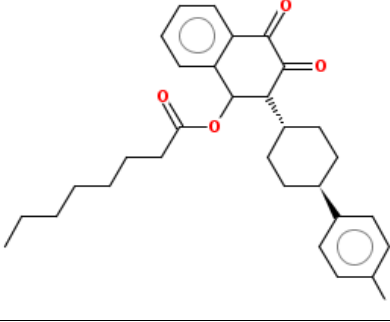
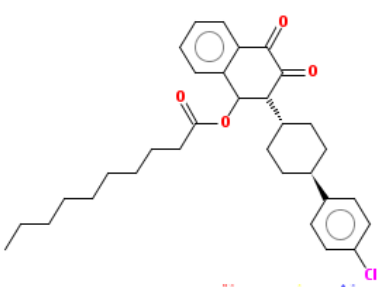
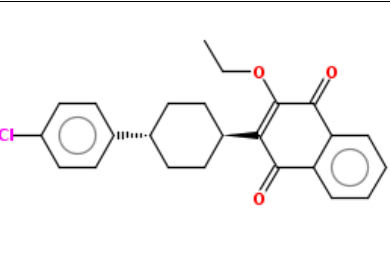
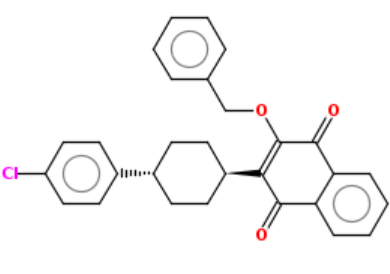
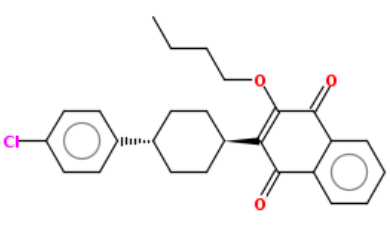
Therapeutic class	Drug	Solubility value (mg/ml)
Antifungal	Griseofulvin, Atovaquone	0.0504, 0.000796
Antibacterial	Ofloxacin	1.44
Antiparasitic	Albendazol	0.0228
Antibiotic	Aceclofenac	0.015

Table 2

The Data set used:

Name	Structure	IC50(activity value)
1a		1.5nm
1b		1.25nm
1c		1.25nm
1d		1.5nm
1e		3.1nm

1f		1.9nm
1g		52nm
1h		1.6nm
1i		1.75nm
1j		1.65nm
1k		1.45nm

2c		0
2d		0
2f		>100nm
3a		5.55nm
3b		7.65nm
3c		13.5nm

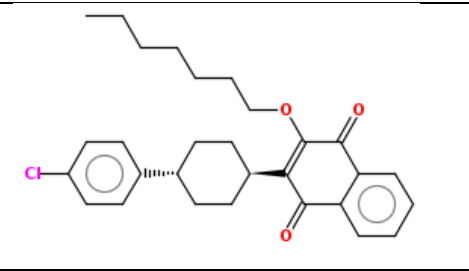
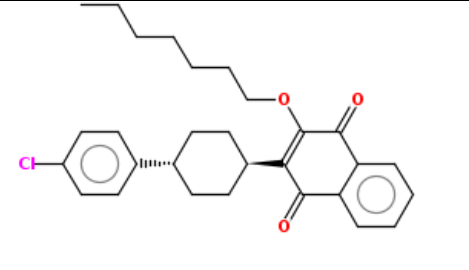
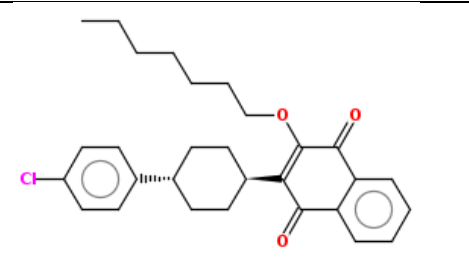
3e		14nm
3f		11nm
3g		32nm

Table 3

Results from R-3.2.2

For eg: For atavaquone derivatives, the PVSA descriptor showed following results for gpcorbreak:

> Positive Very High Correlation, High Variance# t(pvhchv)

[,1]

[1,] NA

> Positive Very High Correlation, Low Variance# t(pvhclv)

[,1]

[1,] NA

> Positive High Correlation, High Variance# t(phchv)

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]

[1,] 9 23 27 30 32 34 20 50 NA

> Positive High Correlation, Low Variance# t(phclv)

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]

[1,] 6 19 18 10 31 24 8 33 NA

> Positive Moderate Correlation, High Variance# t(pmchv)

[,1] [,2] [,3] [,4] [,5]

[1,] 55 46 14 53 NA

> Positive Moderate Correlation, Low Variance #t(pmclv)

[,1] [,2] [,3] [,4] [,5]

[1,] 45 37 5 15 NA

> Positive No Correlation, High Variance #t(zchv)

[,1] [,2] [,3] [,4] [,5]

[1,] 42 29 52 11 NA

> Positive No Correlation, Low Variance# t(zclv)

[,1] [,2] [,3] [,4] [,5]

[1,] 4 38 54 44 NA

And for Final.r, the significant outcomes were:

(Intercept) -84.844 41.738 -2.033 0.0502.

X18 -18.905 4.139 -4.567 6.56e-05 ***

X20 531.764 153.616 3.462 0.0015 **

Weka Outputs for QSAR:

Method	No. of folds	R ²	Q ²	Root mean squared error
Linear regression	9	0.9055	0.7925	2.0765

Table 4

QSAR Model formed: $\log IC_{50} = -0.8308 * X_6 + -4.7095 * x_{36} + 4.0292 * x_{41} + 4.9448$

Here, X6 = Connectivity descriptor

X36= Walk count descriptor

X41= Walk count descriptor

Results from Weka for QSPR:

Method	No. of folds	R ²	Q ²	Root mean squared error
Linear regression	9	0.9702	0.9806	0.3065

Table 5

QSPR Model formed: $\log s \text{ (solubility)}_{\text{datawarrior}} = 0.2783 * x_{38} + -0.3331 * x_{46} + 0.0469 * x_{58} + -0.0159 * x_{62} + -0.92$

Where ,

X38 -> Value from Coloumn 5 of connectivity indices descriptor calculated from Dragon7 software.

X46 -> Value from Coloumn 7 of ETA indices descriptor calculated from Dragon7 software.

X58 -> Value from Coloumn 5 of PVSA descriptor calculated from Dragon7 software.

X62 -> Value from Coloumn 46 of PVSA descriptor calculated from Dragon7 software.

Results from regression analysis:

• **Regression calculation for QSPR**

Name	Log S (datawarrior)	Log S (model)
1a	-6.602	-4.203
1b	-7.772	-7.897
1c	-7.682	-7.54
1d	-8.222	-8.07
1e	-8.492	-8.34
1f	-8.762	-8.6
1g	-10.652	-10.46
1h	-8.086	-7.544
1i	-8.051	-7.78
1j	-8.473	-8.41
1k	-8.142	-8.68
2f	-8.018	-8.28
3a	-6.62	-6.56
3b	-7.643	-7.54
3c	-7.16	-7.11
3e	-7.97	-7.9
3f	-8.51	-8.43
		-
3g	-10.67	10.55

Table 6

Regression equation obtained:

$$Y=0.9418*x+0.3181$$

$R^2 = 0.9703$ i.e. 97% accurate.

Where, x is the data warrior solubility value. And Y would be then the experimental solubility value.

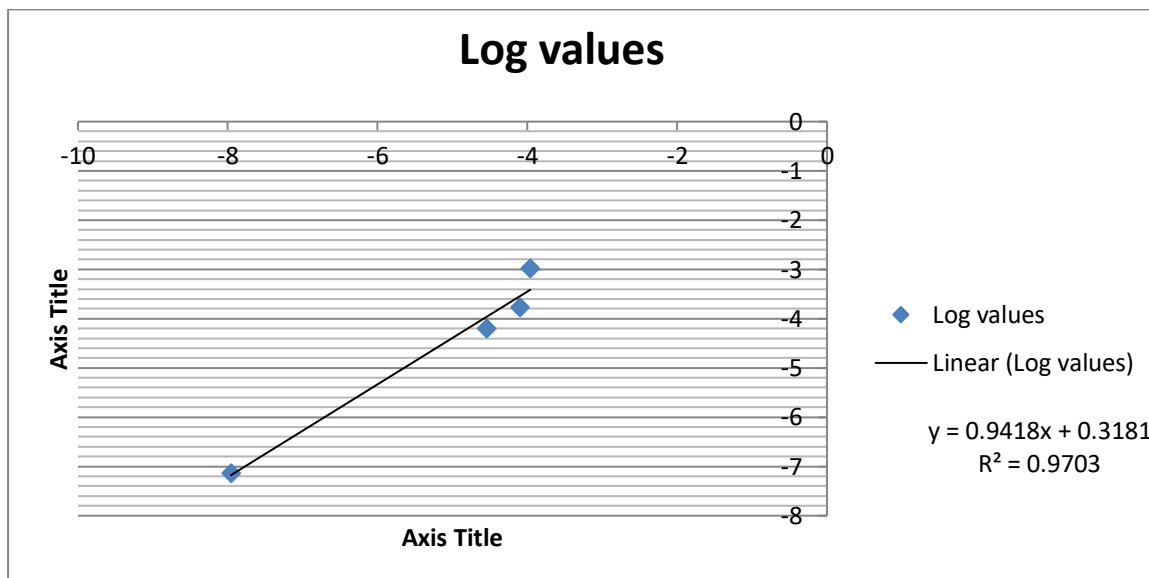


Figure 10:Scatter plot

Now from the above equation values of Y are calculated and then again the regression analysis is done so as to validate the results.

Name	Log S (model)	Y (solubility from regression equation)
1a	-4.203	-5.8996636
1b	-7.897	-7.0015696
1c	-7.54	-6.9168076
1d	-8.07	-7.4253796

1e	-8.34	-7.6796656
1f	-8.6	-7.9339516
1g	-10.46	-9.7139536
1h	-7.544	-7.2972948
1i	-7.78	-7.2643318
1j	-8.41	-7.6617714
1k	-8.68	-7.3500356
2f	-8.28	-7.2332524
3a	-6.56	-5.916616
3b	-7.54	-6.8800774
3c	-7.11	-6.425188
3e	-7.9	-7.188046
3f	-8.43	-7.696618
3g	-10.55	-9.730906

Regression analysis for Table7

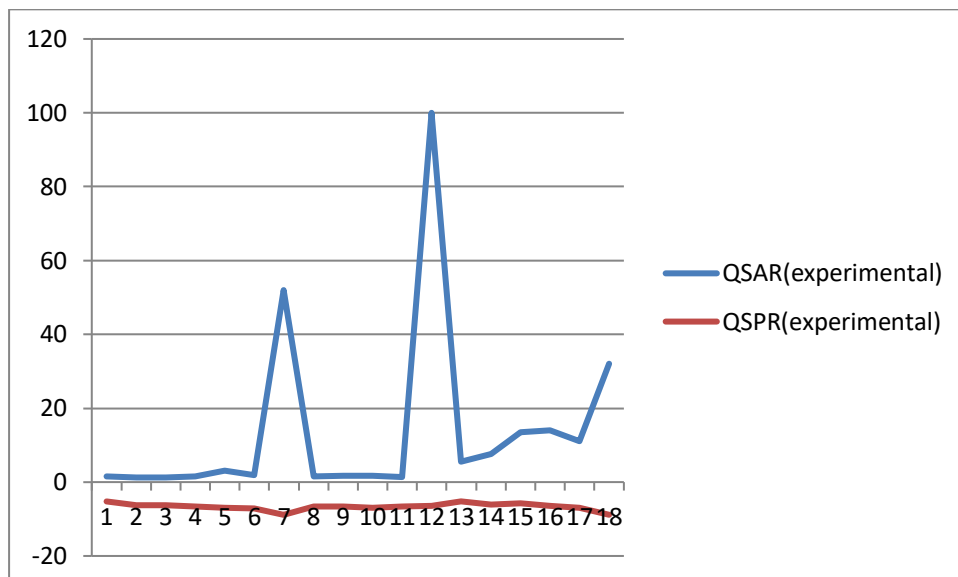
SUMMARY OUTPUT									
<i>Regression Statistics</i>									
Multiple R	0.99753155								
R Square	0.995069194								
Adjusted R Square	0.936245664								
Standard Error	0.585575413								
Observations	18								
<i>ANOVA</i>									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>gnificance F</i>				
Regression	1	1176.386	1176.386	3430.712	4.24E-20				
Residual	17	5.829276	0.342899						
Total	18	1182.215							
	<i>Coefficients</i>	<i>andard Err</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>ower 95.0%</i>	<i>pper 95.0%</i>	
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	
From reg eqn	1.082784857	0.018486	58.57228	4.69E-21	1.043782	1.121788	1.043782	1.121788	

Figure 11:
Regression output

Thus, from this encircled R^2 value I hereby conclude that there is 99% similarity of data i.e. solubility values from regression equation validates the solubility values of the model developed and hence model is 99% accurate.

Trend between QSAR and QSPR:

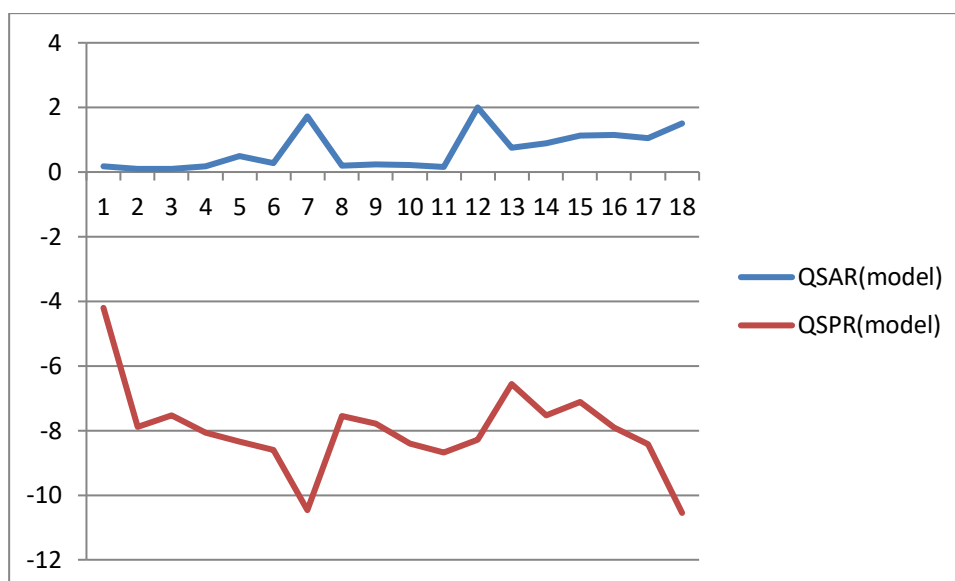
- Between experimental values



**Figure 12: Trend
(experimental)**

Ideally, with the increase in activity the solubility also increases. The same trend is observed here. With the increase in activity value from 0 to 25 the log experimental values which is calculated from regression equation also increases and hence both have a high correlation value of 0.97.

- Between values obtained from models



**Figure 13: Trend
(model)**

Similar trend is observed in the data values obtained from the model. All the log solubility values were negatively skewed and the log IC50 values were positively skewed thus the variation is depicted in the plot. The correlation between the two is calculated to be 0.99 i.e. 99% and thus proves to be very robust model.

Chapter 5- Conclusion

Prodrugs keep on being an energizing region of research. Disintegration of medication is the rate deciding advance for oral retention of the ineffectively water dissolvable medications and dissolvability is the fundamental prerequisite for the assimilation of the medication from GIT.

The QSPR model was created utilizing log dissolvability and all the free descriptors of Atovaquone subordinates. Measurable multicollinearity markers, for example, AIC and VIF were used to dispose of between related descriptors and just huge descriptors were kept and utilized for display determination. If there should be an occurrence of model determination, the impact of every factor on display is assessed. The proposed QSAR and QSPR models are of high factual quality.

Further, the powerful measurable techniques AIC and VIF were actualized for descriptor lessening and choice to decide noteworthy three or four highlights from initial 5250 descriptors gave by Dragon 7 programming. The QSAR and QSPR models shaped can be actualized for anticipating the fluid dissolvability of obscure derivatives. Solubility improvement relies on medicate qualities like solvency, substance nature, liquefying point, ingestion site, physicochemical nature, pharmacological and energy conduct etc. administrative preferables like most extreme every day intake of excipients or potentially tranquilize, affirmed excipients, explanatory exactness etc.

Chapter 6 - Publication

- Nupur S. Munjal, Sumita Dutta, Manu Sharma, Chittaranjan Rout, “QSAR and QSPR Model Development and Comparison for Drugs Having Low Solubility”, International Journal of Engineering Technology Science and Research, Volume 4, Issue 12, 2017.

Chapter 7 - References

1. Mitchell, B. E.; Jurs, P. C., "Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure", *J. Chem. Inf. Comput. Sci.* 1998, 38, 489-496.
2. Murray, J.S.; Lane, P.; Brinck, T.; Politzer, P. "Relationships between Computed Molecular Properties and Solute-Solvent Interactions in Supercritical Solutions." *J. Phys. Chem.* 1993, 97, 5144-514
3. Yalkowsky, S. H.; Banerjee, S., "Aqueous Solubility, Methods of Estimation for Organic Compounds": Marcel Dekker: New York, Basel and Hong Kong, 1992; pp 128-148.
4. Ran, Y., Jain, N. and Yalkowsky, S.H. (2001) "Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE)". *J Chem Inf Comput Sci.* 41, 1208-1217.
5. Wang, R.; Gao, Y.; Lai, L. "Calculating Partition Coefficient by AtomAdditive Methodol". *Persp. Drug Design* 2000, 19, 47-66.
6. Tareq, Hassan and Khan, M. (2010) Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr. Drug Metab.* 11, 285-295.
7. Valentino J. Stella, Kwame W. Nti-Addae, "Prodrug strategies to overcome poor water solubility", in *Advanced Drug Delivery* , 677–694, 2007.
8. C. Mueller, H. Wollmann, "Stability of drugs and pharmaceuticals. 26 studies on the stability of drugs with medium stability", *Pharmazie* 41 (1) (1986) 53.
9. H. Wollmann, "Stability of drugs and pharmaceuticals. 26 studies on the stability of drugs with medium stability", *Pharmazie* 41 (1) (1986) 53.
10. Remya, K. and Suresh, C.H. (2013) "Which density functional is close to CCSD accuracy to describe geometry and interaction energy of small noncovalent dimers? A benchmark study using Gaussian09." *J. Comput. Chem.* 34, 1341-1353.
11. N S Munjal, N Kumar, M Sharma, C Rout, " QSAR Model development for the prediction of solubility" *IEEE explorer* 2016 (<http://ieeexplore.ieee.org/document/7552139/>)