

# Computational Insights into Solubility Prediction: A Comparative Analysis of Predictive Models

Student number: 20067973

UCL School of Pharmacy, 29-39 Brunswick Square, London WC1N 1AX

*Keywords: Solubility prediction, computational tools, log S, log P, structural analysis, correlation analysis, data analysis, salt compounds, software accuracy*

---

**ABSTRACT:** Aqueous solubility and hydrophobicity significantly influence the drug-like properties of a molecule, dictating its bioavailability or ability to reach the systemic concentration for pharmacological efficacy. Poor drug solubility poses a substantial challenge in drug discovery, with 10 to 15% of development failures attributed to inadequate drug-like properties, including solubility and permeability. Besides, hits from high throughput screening (HTS) often exhibit characteristics like high molecular weight, elevated lipophilicity, and low aqueous solubility. Hence, predictive computational tools play important roles during the early stage of drug discovery by identifying compounds with poor aqueous solubility. This study evaluates the predictive accuracy of prominent computational tools (Stoplight, ChemAxon, DataWarrior, ALOGPS 2.1, and ChemDraw) in estimating the solubility of structurally diverse nitrogen-containing compounds. In this study, we compared the calculated solubility values obtained from the computational tools with experimental solubility data. Our analysis reveals a notable lack of reasonable accuracy of these tools across the tested compounds, accompanied by a surprising proficiency in predicting less polar or less soluble substances. We found that ChemAxon consistently stands out as the most reliable tool for predicting log S, while DataWarrior shows dependable performance in predicting log P. However, both tools exhibit relatively low correlation coefficients, highlighting areas for improvement. Despite an exhaustive investigation, the study does not conclusively identify clear reasons for the overall inaccuracy observed in these tools.

---

## 1.0 INTRODUCTION

### 1.1 Drug-Like Properties

Drug-like molecules can be defined as molecules or compounds exhibiting specific characteristics and physicochemical properties which are crucial to the development of potential pharmaceutical drugs (1). In the past, early assessment of pharmacokinetics and physicochemical properties depended on molecular characteristics such as polarity, aromatic ring count, lipophilicity, and molecular mass. However, there has been a recent expansion of these assessments to encompass more intricate properties like metabolic stability, permeability, solubility, and pharmacokinetic/pharmacodynamic modeling (2–4). These properties determine the pharmacokinetics of the drug, where favorable absorption, distribution, metabolism, and excretion (ADME) properties are required. Among these properties, aqueous solubility and hydrophobicity have the most profound impact on the drug-like properties of a molecule as they determine a drug's bioavailability or its ability to dissolve in physiological fluids and achieve its desired concentration in systemic circulation for pharmacological response (1). The solubility of a drug-like molecule depends on its lipophilicity and the number of hydrogen bonds that can be formed with the solvent. The golden rule of solubility states that like dissolves like in which polar materials will dissolve in

polar solvents while non-polar materials tend to dissolve in non-polar solvents.

The Lipinski's rule of 5 formulated by Christopher Lipinski and colleagues in 1997, is valuable in determining the drug-likeness of a molecule for oral administration. High-throughput screening (HTS) is useful as a guide in drug development as it helps determine if a hit or a potential compound is likely to encounter challenges related to oral absorption, mainly due to issues like poor solubility or inadequate membrane permeability (5). According to the Lipinski's rule of 5, poor absorption and permeation occurs when there are more than 5 hydrogen bond donors and 10 hydrogen bond acceptors, the molecular weight is greater than 500 Daltons, and the calculated log P is greater than 5 (6,7). Log P is defined as the logarithm of the partition coefficient of unionised compound between organic solvent, typically octanol, and water (8). It is a measure of the hydrophobicity of a compound by assessing the distribution of a molecule between hydrophobic and hydrophilic environments. Compounds with a positive log P value are hydrophobic and tend to partition into non-polar environments whereas compounds with a negative log P value are hydrophilic and has a higher affinity for aqueous phase. Highly lipophilic compounds often have low aqueous solubili-

This project is a **Data-analysis project** with a word count of 5966 words formatted in the printed style of Journal of Medicinal Chemistry.

ty and show increased risk of toxicity as they tend to bind to hydrophobic targets instead of the desired targets (9). However, some degree of lipophilicity is necessary for drug absorption from the small intestine into the bloodstream through cell membrane penetration. Hence, orally active compounds which show optimal physicochemical and ADME properties are more likely to have a log P value of greater than 1 or less than 4 (9).

Log P is calculated using the formula:

$$\text{Log P} = \text{Log}_{10} ([\text{solute}] \text{ in organic solvent} / [\text{solute}] \text{ in water})$$

## 1.2 General Introduction to Solubility

Solubility is defined as the ability of a solute to dissolve in a solvent to form a solution. Log S, or logarithm of solubility, is a crucial physicochemical parameter used to quantify the solubility of a compound in a solvent. The two distinct concepts related to the solubility of a substance in a solvent are kinetic and thermodynamic solubility. Kinetic solubility refers to the rate at which the solute dissolves, typically in the early stages of dissolution (10). In other words, Kinetic solubility is the solubility observed when a precipitate is first induced in a solution (10). In most cases, the kinetic solubility of a compound is often higher than the measured thermodynamic solubility, and it is determined using a stock solution dissolved in an organic solvent, typically dimethyl sulfoxide (DMSO). In thermodynamic solubility, also known as equilibrium solubility, the dissolved compound is in equilibrium with the undissolved material in excess, and this usually happens at the end of the dissolution process. Both kinetic and thermodynamic solubilities of a compound can be measured using shake flask method and HPLC detection. However, thermodynamic solubility measurements are more time-consuming as they require the system to reach an equilibrium. As well as generating a reasonable throughput with a high sensitivity, HPLC detection also helps determine the purity of compounds, which is critical to obtaining an accurate reading of solubility (11).

**Figure 1**

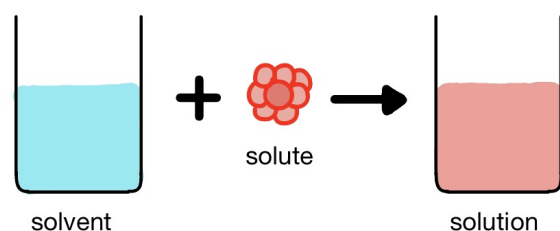


Figure 1. Visualisation of dissolution.

Solubility often dictates the fate of drug candidates by determining the route of drug administration and thus it is an important parameter throughout all stages of drug discovery. Following oral ingestion, drug solubility and intestinal permeability are the two main factors that determine drug absorption (12). The Biopharmaceutics Classification System (BCS) clas-

sifies drug substances into four types based on their aqueous solubility and intestinal permeability (13).

**Table 1: BCS Classification System**

Class	Permeability	Solubility
I	High	High
II	High	Low
III	Low	High
IV	Low	Low

According to **Table 1**, drugs that are categorised into class II and IV have limited solubility so solubility enhancing techniques such as physical and chemical modifications must be in place. Poor drug solubility is one of the major challenges for formulation scientist in the drug discovery process. Research shows that over 40% of new chemical entities developed in pharmaceutical industry are water insoluble (12). A multitude of strategies such as nanoparticle formulation exists to enhance drug solubility and bioavailability following oral administration (14). Particle technologies involving high pressure homogenisation, jet milling, and cryogenic spray processes have been used to reduce particle size (15,16). By doing so, they increase the surface area to drug ratio, thereby increasing drug dissolution rate. Yet, these technologies often lead to high production costs and potential for agglomeration. Hence, the existence of reliable solubility predictive tools can help identify and eliminate poorly soluble compounds, minimize the need for extensive experimental work to improve solubility, thereby saving time and resources.

## 1.3 General Introduction to Aggregation

Aggregation, on the other hand, refers to clusters of small molecules coming together to form aggregates. In pharmaceutical drug discovery, HTS is commonly used to screen libraries of small molecules and identify lead compounds for further development as potential drug candidates. It is followed by lead optimisation to improve their potency and selectivity. However, certain compounds within the small molecule libraries have the potential to aggregate, and such aggregates can lead to non-specific enzymatic activation or assay interference. Large aggregates formed are usually known as promiscuous inhibitors as they inhibit the target protein and cause non-specific alterations to many different proteins (17). As a result, they are a primary source of false positive hits due to their ability to interact non-specifically with many proteins. Colloidal aggregation contributed to 88% of false positives in hits screening while fluorescence interference and promiscuous covalent inhibition are responsible for the rest of the false positive hits identified by the quantitative HTS (18). Hence, colloidal aggregates must be identified and eliminated prior to screening in the early phase of drug discovery. Dynamic light scattering is a low throughput method that determines the size of small molecule aggregates by analysing the correlation between the intensity of scattered light over time (19). As shown in **Figure 2**, A larger aggregate will maintain a higher correlation over a longer delay period as compared to a smaller aggregate.

Figure 2

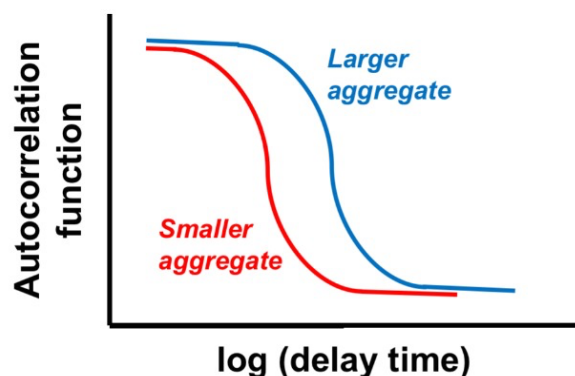


Figure 2. Dynamic light scattering for aggregate characterisation (this figure is taken from (20) <https://www.ncbi.nlm.nih.gov/books/NBK442297/> reproduced in accordance with the license (CC BY-NC-SA 3.0)). Autocorrelation function represents how the intensity of scattered light varies with time.

In some cases, the formation of aggregates can impact the solubility of a substance. Past research indicates that aggregates often adopt configurations that reduce their interaction with water, consequently leading to diminished aqueous solubility (21). It is widely acknowledged that compounds with lower solubility may demonstrate a heightened tendency to aggregate, seeking to minimize their contact with the solvent. The findings from our study underscore a strong correlation between aggregation and poor aqueous solubility, though it is essential to note that not all poorly soluble compounds necessarily undergo aggregation (22). Instead of forming larger aggregates or precipitates, the molecules may remain in a state of fine suspension or dispersion. Given that log P is commonly utilised to predict a molecule's lipophilicity, it offers valuable insights into the molecule's aggregation behavior. According to a research, molecules with log P values of higher than 3 exhibit an increased propensity for aggregation (23).

#### 1.4 The Significance and Challenges of Predictive Tools

The methods used to measure solubility and aggregation, while valuable for understanding a compound's behaviour in a specific solvent, have certain drawbacks which include being time-consuming and labour-intensive. These limitations have led to the emergence of predicting solubility and potential aggregation through computational and modelling approaches. According to the current research, 10 to 15% of drug development failures are attributed to poor drug-like properties, such as solubility and permeability (24). Therefore, accurate solubility prediction plays a role in reducing the risk of costly clinical trials failures by identifying potential issues during the early stage of drug development process. Accurate solubility prediction also reduces the need for experimental measurements. The hits generated through HTS often exhibit characteristics such as high molecular weight, high lipophilicity, and low aqueous solubility (25) (26). Therefore, it is crucial to leverage predictive computational tools in identifying hits with

low aqueous solubility as these may need optimisation for better solubility, or they might be eliminated to focus efforts on more promising candidates. Besides, computational tools for predicting potential aggregation can effectively aid HTS triage by flagging colloidal aggregates and reducing the probability of false positives. As a result, pharmaceutical companies can minimize the waste of precious resources, time, and cost associated with drug development.

While computational tools and high-throughput screening have advanced greatly in recent years, accurate prediction of the solubility of drug-like molecules remains a mystery in the field of pharmaceutical research. There are various software tools available that can predict physicochemical and pharmacokinetic properties of molecules. This study aims to achieve three key objectives: firstly, to evaluate the precision of software tools in predicting solubility by comparing measured solubility data with predictions generated by these tools; secondly, to investigate the underlying reasons behind any inaccuracies observed in the tools; and finally, to conduct a comprehensive comparison of the performance exhibited by various computational tools. These pursuits collectively contribute to revisiting the research question: How accurately can computational software predict the solubility of diverse chemical compounds?

## 2.0 MATERIALS AND METHODS

### 2.1 Computational Tools

A comprehensive set of computational tools including Stoplight, ChemAxon, DataWarrior, ALOGPS 2.1, and ChemDraw, was employed to predict the physicochemical properties, specifically the solubility and log P values of the compounds. The mentioned webtools are designed for predicting various physicochemical and pharmacokinetic properties of molecules which have significant impacts on the ADME-related properties of the molecules. Among these computational tools, Stoplight and ALOGPS 2.1 assess the intrinsic solubility of compounds, while DataWarrior calculates solubility at a fixed pH of 7.5. Notably, ChemAxon stands out as the sole software capable of computing pH-dependent solubilities for compounds, allowing for calculations across different pH values. Notably, all the mentioned tools are freely available and do not require any payment or subscription for usage.

### 2.2 Kinetic Solubility Assay

Two diverse sets of screening compounds from University of North Carolina-Chapel Hill were used in this study. The master dataset consists of 354 compounds whereas the additional dataset consists of 138 compounds to be tested. The two datasets of experimental solubility data were measured and collected by Analiza Inc., a US-based contract research organisation (CRO) specialising in high-throughput physicochemical property and in-vitro ADME testing. Automated miniaturized gold standard shake flask method with chemiluminescent nitrogen detection (CLND) was employed to measure the kinetic solubility of the compounds from DMSO stocks in phosphate buffered saline (pH 7.4) with a final DMSO con-

centration of 2%. Due to confidentiality constraints, specific details about the experimental procedures cannot be disclosed.

## 2.3 Data Extraction and Collection

The predicted intrinsic solubility and partitioning character (log P) values were calculated and obtained from each software tool. Simplified Molecular Input Line Entry System (SMILES) is a notation to represent a chemical structure that is easily understood by computational software. The SMILES string of each compound was input into the computational tools, specifically Stoplight, ChemAxon, DataWarrior, and ALOGPS 2.1, to obtain predicted solubility data. Unfortunately, ChemAxon faced challenges in predicting certain salt compounds in both datasets, opting to predict the acid and base components individually rather than providing predictions for the complete salt entities. Stoplight provided solubility values in mg/L, while ChemAxon, DataWarrior, and ALOGPS 2.1 expressed predictions in the unit of log S. Furthermore, Log P values were gathered from the same software tools using the SMILES string of each compound. Notably, a software named ChemDraw was incorporated to predict log P, examining whether its performance surpassed that of other tools.

## 2.4 Unit Conversion

To facilitate comparison with experimental solubility data, which was measured in  $\mu\text{M}$ , unit conversion of mg/L and log S to  $\mu\text{M}$  was performed. 1 molar solution is defined as 1 mole of a compound dissolved in a total volume of 1 litre and is expressed in the unit of 1 mol/L. Solubility is often expressed in the unit of mol/L so the values were multiplied by 1000,000 to convert into  $\mu\text{mol/L}$  or  $\mu\text{M}$ . To transform the solubility data into log S values, both experimental and predicted solubilities in  $\mu\text{M}$  were logarithmically scaled using base 10.

The conversions were done using the formulas:

Concentration ( $\mu\text{M}$ ) = [Concentration (mg/L) / Molecular Weight (g/mol)] x 1000

Concentration ( $\mu\text{M}$ ) =  $\log^{-1} S \times 1000,000$

## 2.5 Graph Plotting

Various scattered plots were generated to analyse the correlation between predicted solubility or log P and experimental solubility, employing Microsoft Excel for visualisation. For each computational tool in every plot, a trendline accompanied by the coefficient of determination,  $R^2$ , was incorporated to provide insights into the goodness of fit of linear regressions when comparing the computational tools. Additionally, correlation coefficient,  $r$ , and standard deviation,  $s$ , for each software in predicting log S and log P were computed using the correlation and standard deviation functions within Microsoft Excel. Mean absolute error, MAE was calculated for each solubility prediction software in forecasting log S.

MAE was calculated using the formula (27):

$$MAE = \frac{1}{n} \sum_{i=1}^n |(\gamma_i - \chi_i)|$$

Where:

MAE = mean absolute error

$n$  = total number of data points

$\gamma_i$  = predicted solubility for the  $i$ -th data point

$\chi_i$  = experimental solubility for the  $i$ -th data point

## 3.0 RESULTS AND DISCUSSION

### 3.1 Solubility Analysis for the Master Dataset

Figure 3a shows the correlation between predicted and experimental solubility of the compounds from the master dataset. The examination of the relationship between experimental and predicted solubilities revealed a notable absence of correlation. The experimental solubility of the compounds clustered within the range of 0 to 500  $\mu\text{M}$ . In contrast, the predicted solubilities exhibited a much broader distribution of between 0 to 1,300,000  $\mu\text{M}$ . Noteworthy outliers were identified, particularly those with predicted solubilities exceeding 200,000  $\mu\text{M}$ . Each computational tool played a role in predicting these outliers, and among them, Stoplight emerged as the tool predicting the most substantial outlier with a predicted solubility of about 1,300,000  $\mu\text{M}$ . The presence of these outliers, particularly those exceeding 200,000  $\mu\text{M}$ , had a significant impact on the visibility of trends within the data. This phenomenon prompts a closer examination of the specific compounds associated with these outliers. Given the lack of a discernible trend in the presence of these outliers, we conducted a focused analysis by zooming in on the dataset to determine if there is an underlying trend when these extreme values are excluded, and the results are shown below in section 3.1.2.

Figure 3a

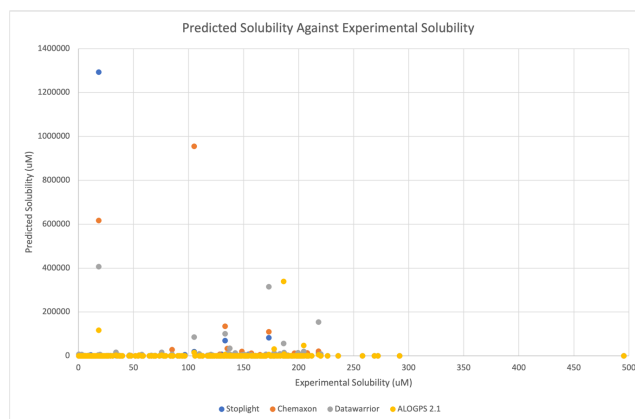


Figure 3a. Correlation between predicted and experimental solubility in  $\mu\text{M}$ .

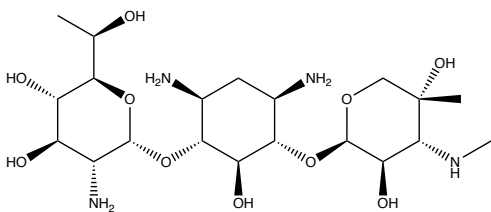
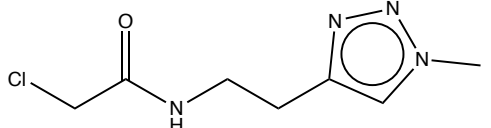
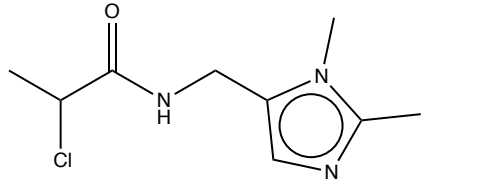
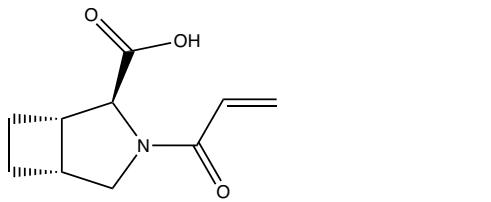
### 3.1.1 Structural Analysis of the Outliers

This structural analysis aims to discover any distinctive features or patterns that may contribute to the atypical behavior observed in solubility predictions. **Table 2** provides information about the chemical structures of the four significant outliers and their corresponding solubility data. These outliers are sourced from the master dataset, and they exhibit predicted solubilities exceeding 300,000uM, which are significantly higher than their experimental solubilities.

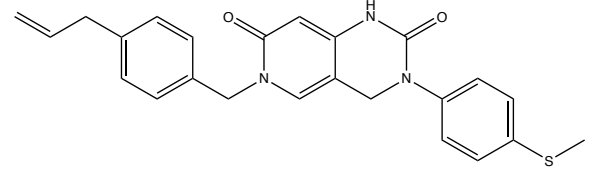
As shown in **Table 2**, compound 1 stands out as an extreme case with its predicted solubility from Stoplight higher than what was measured. When we look at its structure, we notice it has a high density of polar functional groups like amines and hydroxides. This abundance of polar groups makes us wonder if Stoplight, the predictive tool we used, struggles when dealing with compounds that have lots of these polar parts. We were intrigued by the relatively low experimental solubility, especially considering the abundance of polar groups in the molecular structure. Moving on to compound 2, 3, and 4, each contains reactive chemical motifs which can react with water, and that explains their moderately high experimental solubilities but not the predicted ones. Additionally, they all consist of relatively lower molecular weights in comparison to compound 1. However, their predicted solubility values appear unusually elevated. Notably, compounds 2 and 3 share analogous structures, featuring common functional groups like an amide group, a chlorine moiety, and a five-membered heterocycle. Compound 4 not only possesses a low molecular weight but also exhibits a structurally simpler composition compared to most compounds in the master dataset. These observations raise questions about the interplay between molecular complexity, weight, and the predictive performance of different software tools.

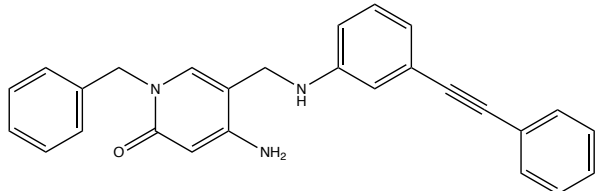
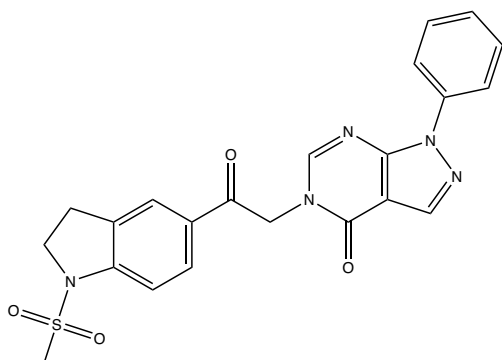
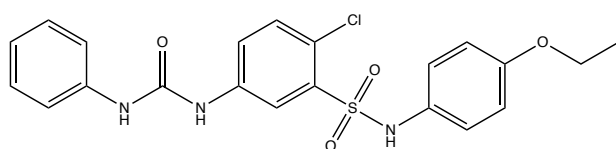
To delve deeper into structural analysis, **Table 3** presents the chemical structures and solubility data of four compounds accurately predicted by each computational tool. These compounds demonstrated predicted solubilities within 1uM of the experimental values. Upon examination of their structures, these compounds share common features of being large, non-polar, and containing multiple aromatic rings. Additionally, they exhibit high molecular weights in contrast to the outliers highlighted in **Table 2**. We observed a trend indicating that the computational tools demonstrate improved performance for less polar compounds or those with higher molecular weights, characteristics typically associated with poor solubility. This suggests that the tools exhibit better predictive accuracy for compounds with lower solubility.

**Table 2. Chemical structures and corresponding solubility data of some outliers in *master dataset***

Compounds	Chemical Structure and Molecular Weight (MW)	Experimental Solubility (uM)	Predicted Solubility (uM)	Predicted by
1	 MW = 496.56	18.5	1292769.5	Stoplight
2	 MW = 202.64	105.1	954992.6	ChemAxon
3	 MW = 215.68	172.9	315500.46	DataWarrior
4	 MW = 195.22	186.3	338844.2	ALOGPS 2.1

**Table 3. Chemical structures and corresponding solubility data of some accurately predicted compounds in *master dataset***

Compounds	Chemical Structure and Molecular Weight (MW)	Experimental Solubility (uM)	Predicted Solubility (uM)	Predicted by
1	 MW = 496.56	11.5	10.85	Stoplight

2	 <p>MW = 202.64</p>	2.2	2.15	ChemAxon
3	 <p>MW = 215.68</p>	5.45	5.3	DataWarrior
4	 <p>MW = 195.22</p>	4.2	4.27	ALOGPS 2.1

### 3.1.2 Solubility Analysis Within 1000uM for the Master Dataset

To visually discern patterns within the solubility data, we performed a close examination of the master dataset, zooming in to explore the presence of any discernible trends. The scattered plot in **Figure 3b** was generated, displaying both experimental and predicted solubilities within the range of 0 to 1000uM. The application of trendlines to the data points yielded  $R^2$  values that nearly approach zero across all software. This outcome suggests that the trendline does not effectively capture the variability in the solubility data, supporting the assertion that there is almost no discernible trend at all, and this is especially evident for Stoplight and ChemAxon, where the  $R^2$  values are almost 0. Notably, most of the data points were clustered below 200uM, and the outliers were consistently predicted by each computational tool. This trend raises concerns about the accuracy of these models, indicating a potential systematic bias towards overestimating solubility. Regardless of the software used, the data points exhibited substantial scattering, as indicated by the very poor  $R^2$  values.

**Figure 3b**

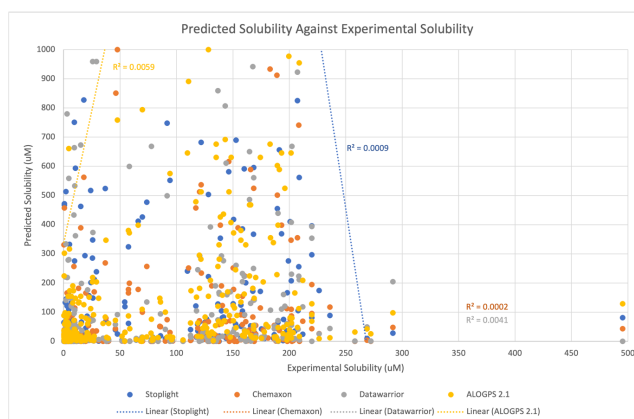


Figure 3b. Correlation between predicted and experimental solubility of within 1000uM.

### 3.2 Solubility Analysis for the Second Dataset

To further validate and support the findings obtained from the master dataset, a similar solubility analysis was conducted on the second dataset and the results is shown in **Figure 4a**. Notably, the second dataset, comprising fewer compounds than the master dataset, presented a more straightforward de-



piction of trends. Yet, the correlation between predicted and experimental solubility remained notably poor, approaching zero. It is essential to note that experimental solubility in the second dataset remained within the narrow range of 200uM, while predicted solubility spanned up to 2500uM. This contrasts with the much broader range observed in the master dataset. The graph for the second dataset exhibited patterns like those observed in the master dataset, reinforcing the consistency of trends across datasets. Despite the reduced complexity in the second dataset, the correlation between predicted and experimental solubility values remained remarkably low, mirroring the challenges encountered in the master dataset. This indicates that the limited size of the dataset did not translate into a significant improvement in predictive accuracy. Consistent with the master dataset, outliers in the second dataset were identified, with a notable recurrence of predictions from Stoplight and ALOGPS 2.1 surpassing experimental solubility values. Stoplight and ChemAxon were accounted for the poorer  $R^2$  values among all the software tools.

**Figure 4a**

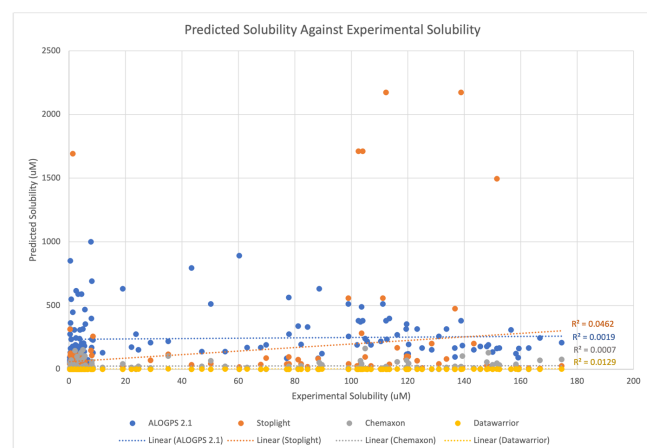


Figure 4a. Correlation between predicted and experimental solubility in uM.

### 3.2.1 Solubility Analysis Within 500uM for the Second Dataset

**Figure 4b** shows a focused analysis that was conducted by narrowing the predicted solubility range to below 500uM, mirroring the approach taken with the master dataset. Upon narrowing the predicted solubility range, distinct patterns emerged within the second dataset. Notably, most predicted solubilities from ALOGPS 2.1 were consistently higher than the experimental solubilities. In contrast, Stoplight predictions demonstrated closer alignment with experimental values. Stoplight exhibited the best correlation among all software tools, although the correlation remained poor overall. Furthermore, we found that DataWarrior consistently underestimated the solubility of most compounds when compared to the experimental values. This systematic bias raises questions about the reliability of DataWarrior in predicting solubility within the specified concentration range. The complex relationships observed emphasize the need for continued refinement and evaluation of these models in solubility prediction.

**Figure 4b**

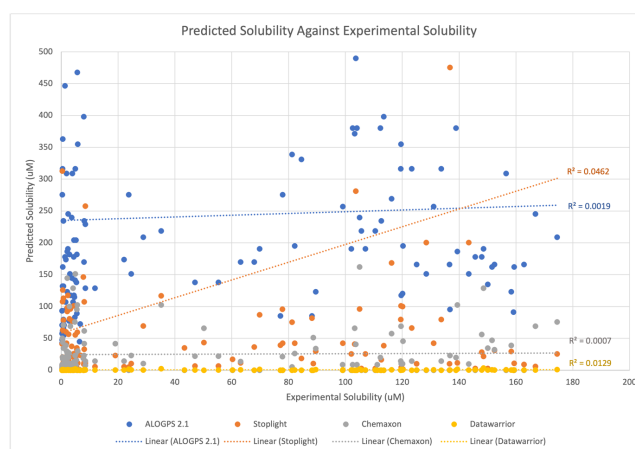


Figure 4b. Correlation between predicted and experimental solubility of within 500uM.

## 3.3 Comparative Analysis of Predictive Accuracy Across Computational Tools

As it was difficult to spot a clear trend in the solubility graphs above, we turned to a different approach. To assess the accuracy of the computational tools, we compared the experimental solubility data with the predicted solubility data for all 354 compounds in the master dataset. Compounds with differences exceeding 50uM were categorised as 'non-compliant,' while those with differences less than or equal to 50uM were labelled 'compliant.' This approach was inspired by a case study evaluating the accuracy of Stoplight predictions, which reported a 65% alignment with the actual results (28). In our study, both Stoplight and DataWarrior exhibited 40% compliance, while ChemAxon and ALOGPS 2.1 demonstrated similar predictions with around 45% compliance as displayed in **Table 4a**. These findings suggest that the computational tools, including Stoplight, do not exhibit satisfactory accuracy in solubility prediction. Comparing Stoplight's performance in our study with the case study, we observed a 25% decrease in predictive accuracy, despite the increased number of compounds in our master dataset (354 compounds) compared to the case study (185 compounds).

**Table 4a. Number and percentage of compounds with solubility differences within 50uM for the master dataset**

Computational tool	Number of compounds with solubility differences within 50uM	Percentage (%) of compounds with solubility differences within 50uM
Stoplight	142	40
ChemAxon	164	46
DataWarrior	143	40
ALOGPS 2.1	159	45

We extended our analysis to the second dataset comprising 138 compounds, a number closer to that examined in the case



study for Stoplight performance. According to **Table 4b**, our findings reveal that Stoplight, ChemAxon, and DataWarrior demonstrate compliance rates of 50% and above, while ALOGPS only achieves 19%. We were not persuaded by the highest compliance percentage of DataWarrior due to its consistently low predicted solubility values below 1uM across all 138 compounds, indicating potential unreliability. Notably, we observed a modest 13% decrease in the predictive accuracy of Stoplight compared to the case study. This subtle decrease, alongside the significantly smaller size of this dataset compared to the master dataset, raises questions about the impact of dataset size on prediction errors.

**Table 4b. Number and percentage of compounds with solubility differences within 50uM for the *second* dataset**

Computational tool	Number of compounds with solubility differences within 50uM	Percentage (%) of compounds with solubility differences within 50uM
Stoplight	72	52
ChemAxon	69	50
DataWarrior	79	57
ALOGPS 2.1	26	19

As we consolidate the findings across both datasets in **Figures 5a and 5b**, we realised that a significant majority of compounds labelled as 'compliant' by each software demonstrate low experimental solubility, typically less than 50uM. This observation is in line with the findings from the structural analysis of outliers in **3.1.1**, suggesting a potential strength of these computational tools in handling compounds with poor solubility, a critical aspect in pharmaceutical and chemical research where predicting the behaviour of less soluble compounds is of paramount importance.

**Figure 5a**

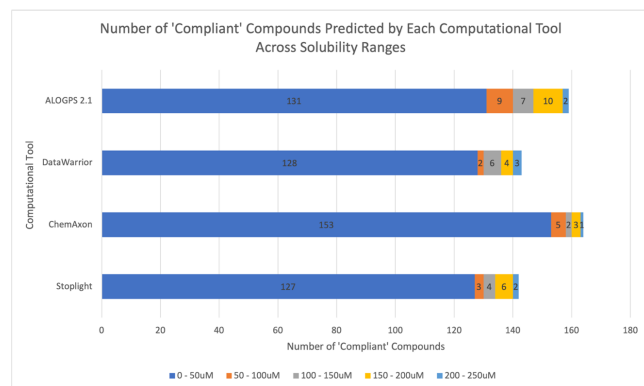


Figure 5a. Number of 'compliant' compounds predicted by each computational tool across different experimental solubility ranges for the *master* dataset.

**Figure 5b**

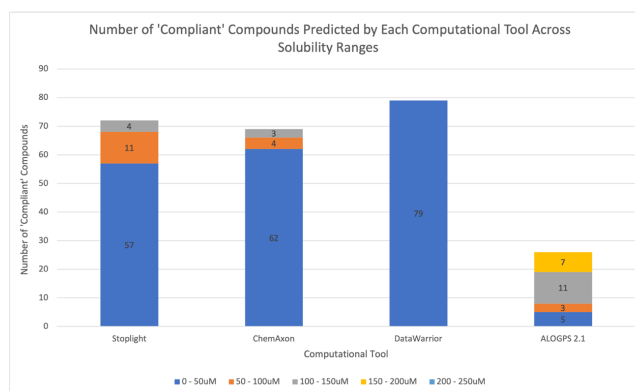


Figure 5b. Number of 'compliant' compounds predicted by each computational tool across different experimental solubility ranges for the *second* dataset.

### 3.4 Log S Transformation for Improved Trend Visibility

Recognising the challenges in establishing trends within the original solubility data, a log S transformation was implemented across all concentration ranges (in uM) in **Figure 6a**. The resulting graph revealed a markedly clearer trend between predicted and experimental log S values. The predicted log S values were in the range between -3 and 7 while the experimental log S values fell into a narrower range of -1 to 3. The application of log S transformation led to a substantial improvement in the  $R^2$  values for all software, indicating a more robust fit of the data to the trendline compared to the initial solubility graphs. However, despite the improvements in  $R^2$  values, it is crucial to note that the overall correlation remains poor. Encouragingly, the log S transformation revealed that predictions from ALOGPS 2.1 and ChemAxon were much closer to experimental solubilities. This shift implies a more accurate prediction alignment, demonstrating the potential utility of ALOGPS 2.1 and ChemAxon in solubility estimation.

**Figure 6a**

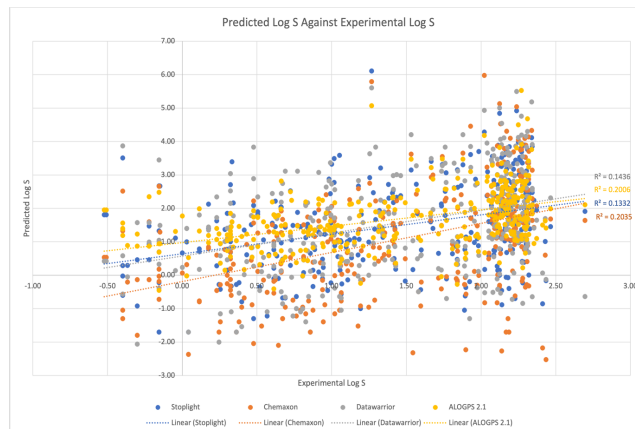


Figure 6a. Correlation between predicted and experimental log S for the *master* dataset.

**Figure 6b** shows the log S transformation from solubility in  $\mu\text{M}$  for the second dataset. The predicted log S values fell within the range of -3 to 4, whereas the observed experimental log S values spanned from -1 to 2.5. In **Figure 6b**, the log S transformation revealed that predicted values from ALOGPS 2.1 are much higher than expected. Its predictions became more accurate for compounds with higher experimental log S values. This suggests that ALOGPS 2.1 predicts better for compounds with higher solubility. Despite the observed improvements in ALOGPS 2.1 predictions for higher log S values, the overall correlation across all software tools remained poor. ALOGPS 2.1, however, exhibited the highest  $R^2$  value among the tools, indicating a higher degree of consistency in the data points. Contrastingly, both ChemAxon and DataWarrior exhibited scattered predictions, resulting in lower  $R^2$  values. Similar to the observations in **Figure 4a**, **Figure 6b** further illustrates the performance of DataWarrior. Predictions from DataWarrior were consistently lower than expected, yet the model demonstrated more accurate predictions at lower concentrations.

**Figure 6b**

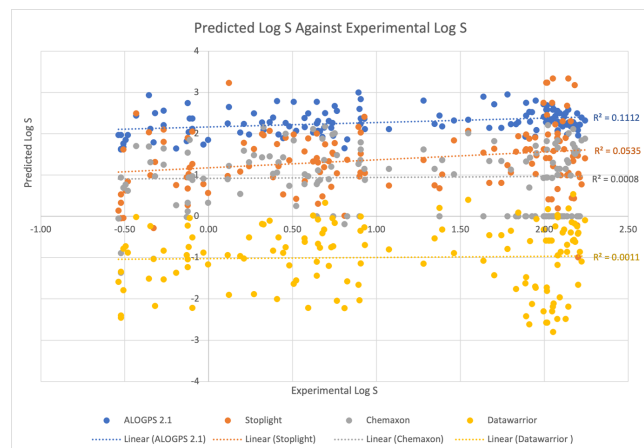


Figure 6b. Correlation between predicted and experimental log S for the *second dataset*.

### 3.4.1 Correlation Analysis of Computational Tools in Log S Prediction

We conducted this analysis by following the approach outlined in (5). **Tables 5a and 5b** provides an overview of the predictive abilities of the computational tools employed in this study for the master and second dataset respectively. The tables list each computational tool along with the number of compounds for which solubilities were predicted within  $\pm 1.0$  log unit, as well as the corresponding correlation coefficient, standard deviation, and mean absolute error for each software. According to **Table 5a**, we observed that each computational tool demonstrated the capability to predict solubilities within  $\pm 1.0$  log unit for over 50% of the compounds. However, we found that correlation coefficients are similar for all computational tools, with no significant differences. This suggests that the tools exhibit comparable predictive abilities across the main dataset. Similar to the master dataset, ChemAxon exhibited the highest correlation coefficient in the second dataset as

shown in **Table 5b**, indicating its consistent predictive performance across different datasets. Despite predicting the highest number of compounds within  $\pm 1.0$  log unit, Stoplight displayed a poor correlation coefficient of 0.23 in the second dataset, contrasting with the higher-performing ChemAxon. Besides, we realised the exceptionally low correlation coefficient of DataWarrior for the second dataset, recorded at 0.03. That is significantly lower than not only its own correlation coefficient for the master dataset, notably 0.38 as shown in **Table 5a**, but also the coefficients observed for other computational tools in the second dataset.

Among the tools assessed, ALOGPS 2.1 emerged as the best-performing tool, with the lowest standard deviation and highest number of compounds predicted within  $\pm 1.0$  log unit and  $\pm 0.5$  log unit across both datasets. In addition, ALOGPS 2.1 and ChemAxon demonstrated higher correlation coefficients, indicating a moderately positive linear relationship between the predicted and experimental solubilities for these tools. However, we noticed that the computational tools within the master dataset exhibited a higher standard deviation and MAE, despite the larger number of compounds included in this dataset. This observation suggests that, contrary to expectations, the increased dataset size did not lead to a reduction in variability. It is plausible that the master dataset, encompassing a larger and potentially more diverse array of data points, along with a greater number of outliers, may contribute to the observed higher standard deviation.

**Table 5a. Performance or predictive abilities of each computational tool for aqueous solubility prediction, based on the *master dataset* which consists of 354 compounds**

Computational tool	Number of compounds predicted within $\pm 0.5$ log unit	Number of compounds predicted within $\pm 1.0$ log unit	Correlation coefficient, $r$	Standard deviation, $s$	MAE
Stoplight	97	201	0.36	1.30	1.01
ChemAxon	100	184	0.45	1.57	1.15
DataWarrior	89	178	0.38	1.47	1.13
ALOGPS 2.1	147	262	0.45	0.89	0.73

**Table 5b. Performance or predictive abilities of each computational tool for aqueous solubility prediction, based on the *second dataset* which consists of 138 compounds**

Computational tool	Number of compounds predicted within $\pm 0.5$ log unit	Number of compounds predicted within $\pm 1.0$ log unit	Correlation coefficient, $r$	Standard deviation, $s$	MAE
Stoplight	45	88	0.23	0.77	0.88
ChemAxon	47	74	0.43	0.65	0.59
DataWarrior	11	27	0.03	0.79	2.10
ALOGPS 2.1	46	64	0.33	0.30	1.19

### 3.5 Comparison with Log P Predictions

To assess the predictive capabilities of computational tools in a different context, we investigated their performance in predicting log P and compared these predictions with experimental log S values. Our initial hypothesis suggested that as log S values increased, the corresponding log P values would exhibit a decreasing trend. This idea is based on the common understanding that compounds which dissolve well in water are usually less likely to be hydrophobic. We wanted to explore if this expected relationship held true by plotting and analysing the log P values against the log S values in our datasets. Besides, we included ChemDraw as an additional software in our analysis to explore whether ChemDraw outperforms other computational tools in capturing the correlation between predicted log P and experimental log S.

**Figures 7a and 7b** represents the correlation between predicted log P and experimental log S for the master and second dataset respectively. Our initial expectation was confirmed, as increasing log S indeed corresponds with decreasing log P, aligning with the general understanding of the relationship between a compound's solubility and hydrophobicity. We observed that all the computational tools exhibited similar and relatively low  $R^2$  values. This indicates that, while the expected trend between log P and log S is present, the data points are scattered, and the models have limitations in accurately capturing the variation in the relationship. Despite this consistency, DataWarrior exhibited the highest correlation coefficient across both datasets, suggesting a relatively stronger alignment between its predicted log P values and the experimental log S values in comparison to other tools. This sustained performance across datasets indicates DataWarrior's proficiency in predicting log P. In contrast, ChemDraw show-

cased the lowest  $R^2$  value, indicating a relatively lower consistency of its data points along the trendline.

The reciprocal relationship between log P and log S is further proven by the negative correlation coefficient values as shown in **Tables 6a and 6b**. Upon detailed analysis, it was notable that all computational tools demonstrated correlation coefficients that were quite similar. The relatively minor variations in coefficients indicate a comparable predictive performance across the board, with no single tool significantly outperforming the others. Similar to the results observed in the log S prediction analysis in **3.4.1**, the standard deviation is higher in the master dataset in the context of log P predictions.

**Figure 7a**

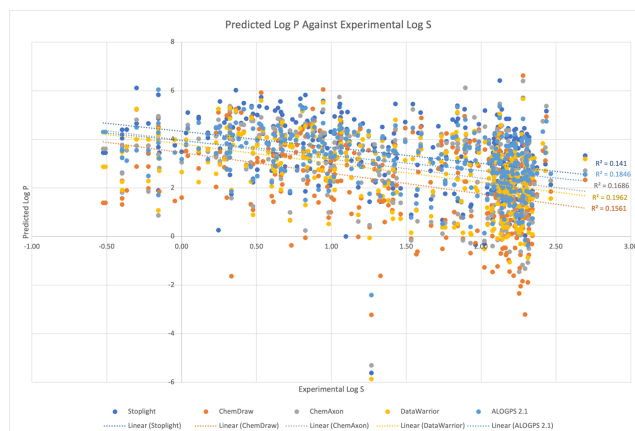


Figure 7a. Correlation between predicted log P and experimental log S for the *master dataset*.

**Table 6a. Correlation coefficient,  $r$ , and standard deviation,  $s$ , of each computational tool in predicting log P for the master dataset**

Computational tool	Correlation coefficient, $r$	Standard deviation, $s$
Stoplight	-0.38	1.42
ChemDraw	-0.40	1.74
ChemAxon	-0.41	1.54
ALOGPS 2.1	-0.42	1.16
DataWarrior	-0.44	1.45

**Figure 7b**

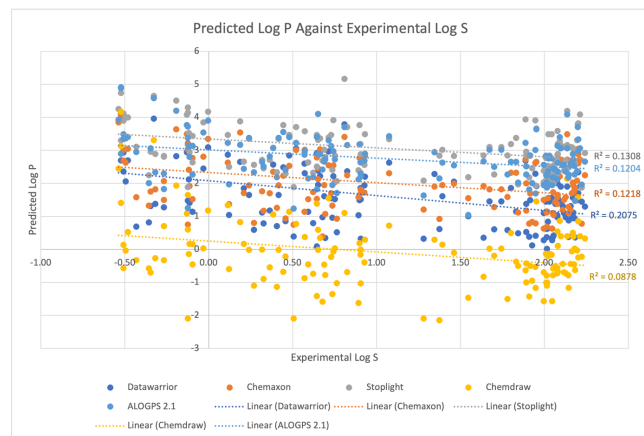


Figure 7b. Correlation between predicted log P and experimental log S for the second dataset.

**Table 6b. Correlation coefficient,  $r$ , and standard deviation,  $s$ , of each computational tool in predicting log P for the second dataset**

Computational tool	Correlation coefficient, $r$	Standard deviation, $s$
Stoplight	-0.36	0.70
ChemDraw	-0.30	1.03
ChemAxon	-0.35	0.82
ALOGPS 2.1	-0.35	0.68
DataWarrior	-0.46	0.92

### 3.6 Effect of Salt Compounds on Predictive Accuracy of Computational Tools

Apart from the chemical structure of the outliers, we wonder if the salt compounds in the master dataset contribute to the inaccuracy of the computational tools. As shown in **Figures 8a – 8c**, three salt compounds were randomly selected from a pool of 11 identified in the master dataset, aiming to systematically evaluate the correlation between their experimental and predicted log S values. Each salt compound underwent a comprehensive evaluation in four distinct forms: the neutral pair,

salt (ionised pair), neutral organic part, and ionised organic part, to explore the impact of varying molecular representations on the predictive accuracy of computational tools. The predicted solubility values for each form were compared against the corresponding experimental solubility, allowing for a detailed assessment to determine which form best aligns with the observed experimental solubility.

**Tables 7a – 7c** show the predicted log S values across different forms of the compounds, as generated by various computational tools, exhibit minimal variations from the corresponding experimental log S values. Notably, DataWarrior consistently shows similar log S values for all forms of each compound. Conversely, ChemAxon faces limitations in calculating log S for compounds that exist in pairs. As a result, both DataWarrior and ChemAxon are identified as less reliable in predicting solubility for salt compounds. It becomes evident that the experimental log S values align best with the predicted log S for the neutral pair form when Stoplight and ALOGPS 2.1 were employed. This observation implies that Stoplight and ChemAxon are particularly good at predicting the solubility of salt compounds when they exist in their neutral form. It is important to note, however, that the compounds under study were originally tested in the form of neutral pairs. Therefore, any inaccuracies observed in the predictive tools within this study are not attributed to the nature of salts. The varying predictions for each form in Stoplight and ALOGPS 2.1 imply that the way molecules are represented can impact how accurately these tools make predictions.

**Figure 8a – 8c**

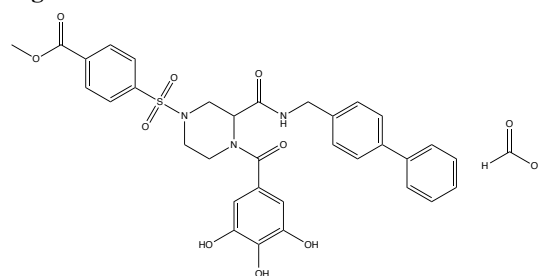


Figure 8a. Chemical structure of compound X.

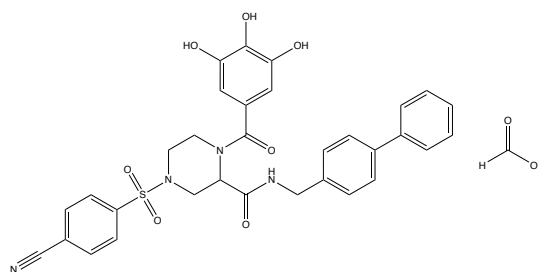


Figure 8b. Chemical structure of compound Y.

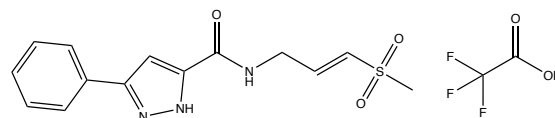


Figure 8c. Chemical structure of compound Z.

**Table 7a. Experimental and predicted log S of different forms of compound X**

Forms	Experimental Log S	Predicted Log S (Stoplight)	Predicted Log S (ALOGPS 2.1)	Predicted Log S (DataWarrior)	Predicted Log S (ChemAxon)
Neutral pair	<b>0.86</b>	<b>0.65</b>	1.38	1.72	N/A
Salt (ionised pair)	N/A	3.36	<b>0.94</b>	1.72	N/A
Neutral organic part	N/A	1.82	1.38	1.72	-0.34
Ionised organic part	N/A	1.91	0.49	1.72	-0.34

**Table 7b. Experimental and predicted log S of different forms of compound Y**

Forms	Experimental Log S	Predicted Log S (Stoplight)	Predicted Log S (ALOGPS 2.1)	Predicted Log S (DataWarrior)	Predicted Log S (ChemAxon)
Neutral pair	<b>2.07</b>	1.77	<b>1.24</b>	1.09	N/A
Salt (ionised pair)	N/A	3.70	0.82	1.09	N/A
Neutral organic part	N/A	1.88	<b>1.24</b>	1.09	-0.53
Ionised organic part	N/A	<b>2.14</b>	0.51	1.09	-0.53

**Table 7c. Experimental and predicted log S of different forms of compound Z**

Forms	Experimental Log S	Predicted Log S (Stoplight)	Predicted Log S (ALOGPS 2.1)	Predicted Log S (DataWarrior)	Predicted Log S (ChemAxon)
Neutral pair	<b>2.18</b>	<b>1.77</b>	<b>2.39</b>	3.67	N/A
Salt (ionised pair)	N/A	3.17	1.07	3.67	N/A
Neutral organic part	N/A	2.68	<b>2.39</b>	3.67	2.54
Ionised organic part	N/A	3.97	1.89	<b>2.97</b>	2.54

### 3.7 Limitations of the Study and Future Work

An identified weakness lies in the inability to conclusively determine the reasons behind the tools' inaccuracy in predicting solubility, even after structural and salt analyses. The reliability of the solubility predictions heavily relies on the accuracy of experimental solubility data. A thorough examination of the assay methodology is necessary to ensure the generation of reliable and high-quality data. Besides, future work could assess other computational tools or involve more in-depth mechanistic studies to discover the specific reasons behind the inaccuracy of computational tools in solubility prediction. Further research could also focus on calibrating or refining existing computational tools by incorporating advanced machine learning techniques or including compounds with greater structural diversity in the training sets. Additionally, considering the theory proposing melting point as a promising predictor of solubility, future investigations could explore this relationship to enhance predictive accuracy. In future research, it would be insightful to investigate the impact of specific structural motifs on solubility predictions. While the focus of this study centred on solubility predictions, the limitations

observed suggest a need for exploration into the tools' capability to predict aggregation behaviours accurately.

## 4.0 CONCLUSION

In our analysis, we have determined that the computational tools employed in this study (Stoplight, ChemAxon, DataWarrior, ALOGPS 2.1, and ChemDraw) do not show reasonable accuracy in predicting solubility of the tested compounds. However, these tools demonstrate a higher degree of accuracy in predicting less polar or less soluble compounds. Surprisingly, no clear reasons for the overall inaccuracy of these tools have been identified in this study. ChemAxon emerges as the most consistent tool for predicting log S, while DataWarrior exhibits reliability in predicting log P. However, it is crucial to note that despite these observations, the correlation coefficients for both tools remain relatively low, suggesting room for improvement. Addressing these unanswered questions should be a focus of future research, along with efforts to enhance these tools by incorporating a more diverse range of compounds in the training sets. Importantly, it should be acknowledged that there is an additional dataset of experi-



mental values for aggregation. However, it was not available in time for the completion of this project. Given the inadequacies in solubility prediction, a promising direction for future research is to check how well they predict aggregation. In the spirit of open science, the data from this study have been posted online on GitHub under The Structural Genomics Consortium to facilitate the involvement of others in advancing and extending this project. Future research is required to pave the way for more robust and reliable computational models, fostering advancements in drug discovery and development.

## REFERENCES

1. Schneider G. Prediction of Drug-Like Properties. In: Adaptive Systems in Drug Design. 2020.
2. Rowland M, Peck C, Tucker G. Physiologically-based pharmacokinetics in drug development and regulatory science. *Annu Rev Pharmacol Toxicol*. 2011;51.
3. Gabrielsson J, Green AR. Quantitative pharmacology or pharmacokinetic pharmacodynamic integration should be a vital component in integrative pharmacology. Vol. 331, *Journal of Pharmacology and Experimental Therapeutics*. 2009.
4. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond rules: The development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem Neurosci*. 2010;1(6).
5. Caldwell GW. In silico tools used for compound selection during target-based drug discovery and development. Vol. 10, *Expert Opinion on Drug Discovery*. 2015.
6. Benet LZ, Hosey CM, Ursu O, Oprea TI. BDDCS, the Rule of 5 and drugability. Vol. 101, *Advanced Drug Delivery Reviews*. 2016.
7. Walters WP. Going further than Lipinski's rule in drug design. Vol. 7, *Expert Opinion on Drug Discovery*. 2012.
8. Berger TA, Berger BK, Kogelman K. Supercritical Fluid Chromatography for Chiral Analysis and Semi-preparative Purification. In: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. 2022.
9. Gao Y, Gesenberg C, Zheng W. Oral Formulations for preclinical studies: Principle, design, and development considerations. In: *Developing Solid Oral Dosage Forms: Pharmaceutical Theory and Practice: Second Edition*. 2017.
10. Box KJ, Völgyi G, Baka E, Stuart M, Takács-Novák K, Comer JEA. Equilibrium versus kinetic measurements of aqueous solubility, and the ability of compounds to supersaturate in solution - A validation study. *J Pharm Sci*. 2006;95(6).
11. Saal C, Petereit AC. Optimizing solubility: Kinetic versus thermodynamic solubility temptations and risks. *European Journal of Pharmaceutical Sciences*. 2012;47(3).
12. Savjani KT, Gajjar AK, Savjani JK. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm*. 2012;2012.
13. Samineni R, Chimakurthy J, Konidala S. Emerging Role of Biopharmaceutical Classification and Biopharmaceutical Drug Disposition System in Dosage form Development: A Systematic Review. Vol. 19, *Turkish Journal of Pharmaceutical Sciences*. 2022.
14. Merisko-Liversidge EM, Liversidge GG. Drug Nanoparticles: Formulating Poorly Water-Soluble Compounds. *Toxicol Pathol*. 2008;36(1).
15. Khadka P, Ro J, Kim H, Kim I, Kim JT, Kim H, et al. Pharmaceutical particle technologies: An approach to improve drug solubility, dissolution and bioavailability. Vol. 9, *Asian Journal of Pharmaceutical Sciences*. 2014.
16. Merisko-Liversidge E, Liversidge GG. Nanosizing for oral and parenteral drug delivery: A perspective on formulating poorly-water soluble compounds using wet media milling technology. Vol. 63, *Advanced Drug Delivery Reviews*. 2011.
17. Chan LL, Lidstone EA, Finch KE, Heeres JT, Hergenrother PJ, Cunningham BT. A Method for Identifying Small-Molecule Aggregators Using Photonic Crystal Biosensor Microplates. *J Lab Autom*. 2009;14(6).
18. Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, Keiser MJ, et al. Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J Med Chem*. 2010;53(13).
19. Chan LL, Lidstone EA, Finch KE, Heeres JT, Hergenrother PJ, Cunningham BT. A method for identifying small molecule aggregators using photonic crystal biosensor microplates. In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*. 2009.
20. Auld DS, Inglese J, Dahlin JL. Assay Interference by Aggregation. *Assay Guidance Manual*. 2004.
21. Roberts CJ. Protein aggregation and its impact on product quality. Vol. 30, *Current Opinion in Biotechnology*. 2014.
22. Guha R, Dexheimer TS, Kestranek AN, Jadhav A, Chervenak AM, Ford MG, et al. Exploratory analysis of kinetic solubility measurements of a small molecule library. *Bioorg Med Chem*. 2011;19(13).
23. Irwin JJ, Duan D, Torosyan H, Doak AK, Ziebart KT, Sterling T, et al. An Aggregation Advisor for Ligand Discovery. *J Med Chem*. 2015;58(17).
24. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Vol. 12, *Acta Pharmaceutica Sinica B*. 2022.
25. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods*. 2000;44(1).
26. Clark DE, Pickett SD. Computational methods for the prediction of 'drug-likeness'. Vol. 5, *Drug Discovery Today*. 2000.
27. Karunasingha DSK. Root mean square error or mean absolute error? Use their ratio as well. *Inf Sci (N Y)*. 2022;585.
28. Wellnitz J, Martin H-J, Hossain M, Rath M, Fox C, Popov K, et al. STOPLIGHT: A Hit Scoring Calculator. *ChemRxiv*. Cambridge: Cambridge Open Engage; 2023; This content is a preprint and has not been peer-reviewed.

## ASSOCIATED CONTENT

### Supporting Information

The raw data for the master dataset can be found here ([https://docs.google.com/spreadsheets/d/1Rk95TxfA-ei3lvsprS-moFI\\_4UtlamfOXqzjRSPMjpk/edit#gid=0](https://docs.google.com/spreadsheets/d/1Rk95TxfA-ei3lvsprS-moFI_4UtlamfOXqzjRSPMjpk/edit#gid=0)) while the raw data for the second dataset can be found here (<https://docs.google.com/spreadsheets/d/1Gp1VEvxSSArujaj3cgzmx0-Ku7ivrTLIljSKGJDgt4/edit#gid=0>).