Machine learning based QSPR approaches to predict solvation free energy of Quinone molecules for flow battery applications

Sivadurgaprasad Chinta, Shania Mitra and Raghunathan Rengaswamy *

Department of Chemical engineering, Indian Institute of Technology Madras, India - 600036

Abstract

In the recent years, flow battery technology has been gaining interest due to its ability to decouple power and energy. However, a major disadvantage of flow batteries over other storage devices is their low energy density. Identifying new electrolyte chemistries with reasonable energy densities can make flow batteries economically viable. Quinone redox couples are gaining interest due to their relatively higher energy densities compared to vanadium redox couples. Machine learning (ML) approaches are widely being used lately for material discovery applications due to their ability to capture complex patterns. Quantitative Structural Property Relationships (QSPRs) are well known in computational chemistry and drug discovery areas to correlate the structural features of molecules to their properties using different mathematical models. The present work focuses on establishing a robust ML based QSPR framework for predicting solvation free energy of Quinone derivatives and compare the performance with traditional approaches like group contribution methods. A new set of explainable input features are used to interpret the predictions of ML methods, which are further used in designing a set of new molecules and validate the performance of the framework. The proposed ML framework enables quick and efficient exploration of the search space for material discovery compared to experimental and traditional approaches.

Keywords: flow batteries, solubility prediction, group contribution approach, QSPR approach, multiple model learning

1 Introduction

Flow battery (FB) is an electrochemical device in which the electrical energy is derived from the chemical energy stored in the electrolytes. This chemical energy is converted to electrical energy during discharge. The electrolytes are circulated through the cell during both charge and discharge. In general, FBs contain two electrolytes, one to store the active materials for negative electrode reactions and the other to store active materials for positive electrode reactions [1]. Electrolyte solutions contain both reduced and oxidized form of reactants in the same phase, where the relative concentrations of oxidized and reduced forms vary over the course of charge or discharge. Due to its scalable nature, easy decoupling and refuelling, flow batteries are more efficient than other storage devices [2]. The low energy density values possessed by various redox chemistries are the major impediments for flow battery commercialization. Identifying new electrolyte chemistries with reasonable energy densities can make flow batteries economically viable. Energy densities of existing electrolytes can be improved by increasing their solubility by selecting suitable solvents. Soloveichik [3] reviewed various flow battery technologies in detail along with the technical and economic challenges and possible remedies for the same.

Quinones are gaining interest as electrolytes for flow batteries in the past few years due to their ability to transfer two electrons for a single molecule and impressive solubility characteristics, which results in relatively

*Corresponding author: Raghunathan Rengaswamy, Department of Chemical engineering, Robert Bosch centre for Data Science and Artificial Intelligence, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India-600036. Ph.no. +91-44-22574159 Email: raghur@iitm.ac.in

high energy densities as compared to other flow battery chemistries. Quinones also exhibit minimal membrane crossover due to their large molecule size and can be produced on large scale with less expenditure compared to vanadium. Suleyman et. al. [4] showed that solubility of Quinones and the reduction potential of Quinone redox couples can be tuned by substituting various functional groups. They used Perdew–Burke–Ernzerhof (PBE) based Density Functional Theory (DFT) calculations to obtain reduction potential and solvation free energy values. Quinones can be used as electrolytes on both sides of a flow battery, hence a rigorous exploration of Quinone derivatives space to identify more efficient electrolytes is of interest. Quick and robust mathematical models are required for further exploration of Quinones either by substituting with a new set of functional groups or by substituting with two or more functional groups on a single molecule. These models can be useful for computationally tractable search of potential molecules in the derivatives space avoiding computationally expensive DFT simulations.

Group contribution (GC) approaches are well established to estimate a wide variety of physical and chemical properties ranging from melting point to toxicity of organic molecules [5, 6, 7, 8, 9]. These approaches assume that the organic molecules are constructed using fragments from a predefined set of fragments and the properties of these compounds are linearly dependent on the occurrences of each fragment. Marrero and Gani [10] proposed an efficient multilevel group contribution approach, in which the property of interest is initially regressed with the occurrences of first-order groups. Then the residuals are regressed with the occurrences of second order groups. Finally, the remaining residuals are related to the occurrences of third order groups. First order groups are simple functional groups that can form a molecule structure such that no atom will be counted twice. Second order groups are used to distinguish between isomers effectively. Third order groups are usually fused and non-fused rings. Using this multilevel group contribution approach, Marrero and Gani [11] estimated octanol/water partition coefficient and aqueous solubility of a broad range of compounds ranging from C3 to C70. Correa et al.[12] proposed Analytical Solutions of Groups (ASOG) group contribution approach to predict water activities in aqueous electrolytes.

Quantitative structure activity or property relationships (QSAR/QSPRs) are the mathematical representations of the functional behavior between the biological activity or chemical response of a component and its quantifiable structural information. This structural information is denoted in the form of structural descriptors/features such as atom counts, surface area, refractivity etc. QSPR/QSARs are widely used in the fields of molecule design, predictive toxicity and drug design [13] for identifying various properties of organic molecules such as flash point[14], vapor pressure, water-air partition coefficients[15], water-octanol partition coefficients[16], solubility[17] and toxicity[18] etc. Any QSPR/QSAR study involves three major steps, i.e. calculating structural features (descriptors) for the predefined molecules set, identifying suitable descriptors and obtaining an efficient correlation between structural features and the property of interest. Structural features can be obtained using first principles, theoretical models and platforms like PaDEL-Descriptor, DRAGON, OpenBabel etc.[13], which are specifically designed for calculation of structural features. Selecting suitable descriptors and obtaining robust models involve a wide range of chemometrics such as PCA, regression tools and neural networks etc. Yousefinejad and Hemmateenejad [19] consolidated various chemometric methods used in both feature selection and model development phases of QSPR studies. Once a robust QSPR is identified, the structural features for a specified objective can be obtained in an inverse QSPR frame work[20]. Various kinds of feature selection algorithms are proposed in literature i.e. classical methods such as forward selection and backward selection[21], artificial intelligence based methods such as genetic algorithm (GA)[22], particle swarm optimization (PSO) based approaches[23] and dimensionality reduction based approaches such as principal component analysis etc. Forward selection approach starts with zero descriptors and in each step, one new descriptor is added based on predefined criteria until a stopping criterion is satisfied. Backward selection approach starts with the complete set of descriptors and in each step, a new descriptor will be removed based on predefined criteria until satisfies stopping criterion is satisfied. In stepwise selection, a combination of both forward and backward selection at each step is shown to be more robust. GA and PSO approaches formulate feature selection as an optimization problem with binary variables i.e. each variable corresponds to the decision of whether a feature should be considered or not. Principal component based approaches obtain few linear combinations of original descriptors, which can explain maximum variability in the data. In the era of machine learning, due to the availability of a wide range of modelling techniques, selecting a suitable modelling method is also crucial to obtain a robust structure property relationship. Each modelling technique has its own advantages and disadvantages. Multivariate linear regression can be used if the dependency of the property

of interest on selected features is anticipated to be linear[24]. Principal component regression[25], in the case of dependencies among inputs, polynomial regression and artificial neural networks (ANN)[26] in cases where nonlinear relationship exist between selected features and property of interest can be explored. Though ANN models can fit very complex nonlinear behaviour, interpretability and overfitting are major issues. Piecewise linear models have been shown to mimic nonlinear behaviour using piecewise linear assumptions[27]. In our earlier work[28], piecewise linear models were used to fit the non-linear behaviour in order to obtain a robust QSPR to predict drug solubility in binary systems. We proposed a prediction error based clustering approach in our previous work [29], which can identify the significant features as well as operating models in a single framework.

Traditionally, K-fold cross-validation is done on the whole data set or train and test data sets are derived from the same/similar distribution to gauge the performance of models. In this study, we carry out two levels of performance analysis and design a radically different test set to demonstrate the robustness of our models. *This is an important step to establish the generalisability of our model to find the solubility of molecules reliably without experiments or complex computation.* Starting with a test set close to the training set in the sample space, similar to the ones customarily chosen, we observe the predictions of the model on these instances. Following this, a new feature set for these test molecules, based on their physical and chemical properties is derived, for ease of interpretability. Based on the new feature set, insights about the effect of these properties on solubility are drawn, using which, a diverse test set with new quinones (and same functional groups) is designed aiming at different solubility ranges. Following this, a vastly different test set with a new functional group, O-, is designed to further validate the inferences drawn in the previous steps and demonstrate the generalisability of our model.

In this section, a brief overview of group contribution approaches, QSPR approaches and useful chemometric approaches are provided, along with an outline of the procedures adopted in the study. In the following section, a problem specific group contribution approach is described to obtain the solvation free energy. In section 3, three different QSPR approaches i.e. linear, neural network and piecewise linear models are employed to obtain a robust QSPR model. In section 4, we use the obtained models to predict the solvation free energy of Quinone derivatives with two different functional groups (on same structure at different positions) and study the effect of newly derived interpretable features such as aromaticity, inductive effect etc. on solvation free energy. In section 5, we design molecules aimed at different solubility ranges based on the observed insights and validate them. In Section 6, we use the estimated ML models to predict solvation free energy of Quinone derivatives substituted with a new functional group, i.e., O^-, which is absent in the data used to build ML models. Finally, this paper concludes with comments and discussions on the efficiency of proposed ML framework.

2 Group contribution approach

Group contribution (GC) approaches assume that the property of interest of a compound is a function of predefined set of structural fragments and it is computed by summing the frequency of each group occurring in the molecule times its contribution [10]. The group contribution framework to obtain the properties of interest of organic molecules is shown in Figure 1. GC approach involves two steps, initially, the occurrences of each fragment are counted and then a linear correlation is obtained between the property of interest and the occurrences of each fragment to obtain the contribution of each fragment. The contributions (C) of all fragments are calculated using equation 1, where $f(X)$ is the property of interest of molecule X, N_i is the number of times fragment i occurred in molecule X and C_i is the contribution of fragment i. Now, for any new test molecule, the occurrences of each fragment are evaluated and substituted in the linear relationship (in equation 1) to obtain the property of interest of the test molecule.

$$f(X) = \sum_{i=1}^{n} N_i C_i$$

In this case study, data set [4] for the solvation free energy estimation includes three variants of Quinones i.e. benzoquinone, naphthoquinone, and anthraquinone substituted with 18 functional groups. To differentiate the three variants of Quinones and the 18 functional groups, we considered 41 different types of fragments, which

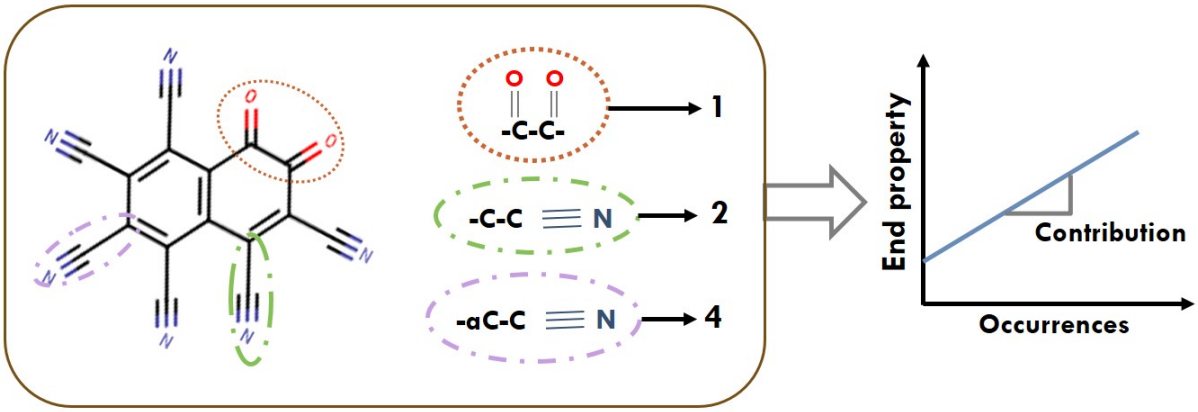

Figure 1: Group contribution approach framework

are specific for this case study. 38 out of these 41 fragments are first-order groups, whereas the remaining 3 groups belong to the second-order, which are useful to differentiate between the Quinone types. The data set contains 407 data samples. Initially, the data is randomly divided into 'model' and 'global test' data sets with 80% and 20% of data samples respectively. Model data set is used to obtain contributions of each group (i.e. model parameters) in association with K-fold (K as 5) validation approach. In each run, the model data set is again randomly divided into K-equal partitions and each time data in K-1 folds are used to train the model and the remaining to test. This procedure is repeated for 100 random runs and the model parameters (contributions of all groups) are averaged and reported as the final set of parameters in Table 1 along with the performance metrics.

It can be observed from the contributions values (bolded values in Table 1) that substituting with PO_3H_2 can increase the solubility (low solvation free energy) followed by COOH, SO_3H, and NH_2 as suggested in the literature[4]. It is also interesting to note that from the second order functional group contributions (italic values in Table 1) having two C=0 groups side by side in a ring can increase the solubility than having two C=0 groups opposite to each other. This can be validated by comparing the solvation free energy values of 1,2-BQ, 1,2-NQ and 1,2-AQ variants (i.e. substituted with the functional groups) with 1,4-BQ, 1,4-NQ and 1,4-AQ variants respectively. The performance metrics of the GC approach to estimate solvation free energy can be obtained in Table 2. Though considering more fragments to differentiate isomers effectively can improve the performance of the GC approach, the size of the data is an impediment in this case study. The major setback of GC approaches is that the property of interest of any new molecule which contains the fragments, which are not included in the training set cannot be evaluated. For example, in this work, we have designed a set of Quinone derivatives with O^- functional group. GC methods can not be used to predict solvation free energy for them since the corresponding fragment contribution is not available.

3 QSPR based approaches

Identifying QSPR consists of three phases, i.e., data generation, feature selection, and model prediction. In data generation phase, chemical structures are converted into an accessible form such as .mol, .smi etc. to calculate structural feature values. Feature selection involves both domain and mathematical knowledge to identify the significant and independent features set that affects the property of interest. Feature selection algorithms such as forward selection, backward selection, stepwise regression, and evolutionary optimization approaches are mathematical ways of exploring the most suitable feature subset to reduce the model complexity, thus avoiding overfitting of models[30]. Model prediction is the process of identifying a robust model between the significant features set and the property of interest. QSPR framework to estimate any end property of organic molecules is depicted in Figure 2.

PaDEL-Descriptor[31] is an openly available software to compute various kinds of structural features of molecules varying from topological parameters to chemical fingerprints. In this case study, for QSPR estimation, the solvation free energy data of 407 Quinone variants provided in the literature [4] is used. Structure files of all 407 Quinone variants are generated in smiles (.smi) format and processed to obtain the structural features using PaDEL-Descriptor. McGowan characteristic volume (McG_Vol), Molecular weight (MW), Van der Waals volume (VABC), first ionization potentials (Si), sum of atomic polarizabilities (Apol), solvent accessible surface area (TSA), topological polar surface area (TopoPSA), combined polarizability (MLFER_S), excessive molar refraction (MLFER_E), Molar refractivity (AMR), overall hydrogen bond basicity (MLFER_BH) and acidity (MLFER_A) values of the molecule are found to affect solvation free energy values[28, 32, 33] hence these are considered as structural features for this case study. Since the structural features can be of different magnitudes, to avoid the influence of any particular variable on the model parameters, features are scaled individually by mean centric scaling using mean and standard deviation of a particular feature.

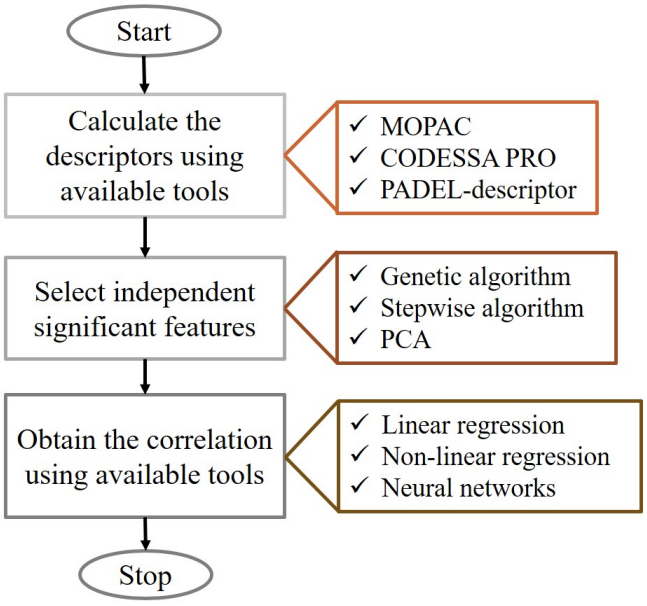

Figure 2: QSPR framework to estimate the property of interest of organic molecules

3.1 Single linear model based QSPR

In this case study, initially, we identify the significant features using K-fold validation (K value as 10) in conjunction with F-test. For feature selection, a linear relationship is assumed between the 12 descriptors set and solvation free energy. The model parameters obtained in each fold are averaged and each parameter is tested with F-test to check if it is significant or not. It is identified that out of 12 variables, 7 variables i.e. McG_Vol, MLFER_A, MLFER_BH, MW, MLFER_S, MLFER_E, and TopoPSA are significant. Later, a linear correlation is obtained between the above identified significant features set and solvation free energy using ordinary least squares associated with K-fold validation (K value as 5).

Initially, the data is randomized and divided into 'model' and 'global test' data sets with 80% and 20% of data samples respectively. Data in 'model' data set is randomized and equally divided into K-partitions and each time data in K-1 partitions are used to train the model and the trained model is tested on remaining data. Model coefficients obtained in all K-folds are averaged for 100 random runs and reported as final model parameters. Performance metrics of linear relationship obtained on the model data set, global test set and overall data set can be obtained in Table 2. The poor performance (R^2 value as 0.6395) of a single linear model suggests that the linear behavior assumption may not be valid hence a non-linear model can be anticipated to increase the prediction accuracy. In following subsections, neural network and piecewise linear based models

are tested to obtain robust non-linear models.

3.2 Neural network based QSPR

Neural networks are highly recommended to mimic non-linear behavior due to their ability to capture complex functions. Jalali-Heravi et al.[34] concluded that the Levenberg-Marquardt algorithm is more suitable for QSPR prediction compared to other training approaches such as backpropagation and conjugate gradient algorithms. In this case study, a Levenberg-Marquardt algorithm based neural network is trained to identify the relationship between structural features and solvation free energy. In this case study, to identify significant features, a backward stepwise approach[35] is used. In this approach, if n variables have to be ranked, initially n networks with *n-1* different variables have to be trained on training data set. The n^{th} missed out variable for which the network results in the largest error on test data set is considered to be the most important. To identify the next most significant variable, the current important variable is removed and the above procedure is repeated. This procedure is continued until all n variables or the first m (n) important variables get individual rankings. In this case study, feature selection is carried on 12 input variables (structural features) with 1 hidden layer architecture. The ranking procedure described above is repeated for 100 random runs and the consolidated rankings are reported in Table 3. The first seven important variables are used to obtain the final network to estimate solvation free energy. Data is randomly divided into 'model' and 'global test' data sets with 80% and 20% of data samples respectively. Initially, to obtain optimal network architecture, networks with hidden layer sizes ranging from 1 to 7 are tested for 10 random runs. In each run, a neural network is trained for all 7 configurations on the model data set and tested on testing data set. The adjusted root mean squared error values of all 7 networks are averaged for all 10 runs and the network architecture with the least averaged error on the test data set is assumed to be the optimal network. The mean adjusted RMSE values of networks with 1 to 7 hidden layers on test data set are 30.2237, 30.8945, **29.0444**, 30.1465, 33.3646, 36.7107 and 44.5868 respectively, hence the network with 3 hidden layers is considered to be optimal. Once the optimal network architecture is obtained, then a neural network with a similar architecture is trained on model data set for 100 random runs. The network with the least root mean squared error on the test data set is considered to be the final network to predict solvation free energy. Prediction accuracy of the final neural network model on model data set, test data set and on overall data is given in Table 2. It can be observed from the table that neural network based QSPR performs better than the single linear model based QSPR due to the ability of NN to mimic the non-linear behavior.

3.3 Multiple model based QSPR

Piecewise linear models have been shown to identify non-linear behavior[36] and are also easily interpretable. Identifying piecewise linear models and their operating regions is known as multiple model learning. In this case study, we assume solvation free energy values are linearly dependent on the structural features with different hyperplanes in different regions. A fuzzy clustering approach based on prediction error is used to obtain the operating models [29]. The major advantage of this approach is that both feature selection and model identification are included in a single framework. It is interesting to note that in different operating regions different structural features can be significant. The data is randomly divided into 'model' and 'global test' data sets with 80% and 20% of data samples respectively. In this approach, initially, the number of underlying models and their true orders (i.e. significant features in each operating region) are identified using the clustering approach [29] on model data set. Later, the clustering procedure is initiated with true models and their orders associated with K-fold (K value as 4) validation on model data set to obtain the final model parameters. The details of the clustering approach[29] are provided below.

Details of the multiple model clustering approach based on prediction error:

The objective for a fuzzy clustering algorithm to group M data samples into N piecewise linear models based on prediction error is as follows:

$$f = \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\mu_{ij}^{q}\left\|y_j - C_i x_j\right\|_2^2\right)$$

Where x_j, y_j are the input features vector and response (output) of sample j respectively and the linear model parameters vector of cluster i is denoted with C_i such that $\hat{y}_j = C_i x_j$.

The prediction error for a data sample j with respect to model i is estimated as:

$$PE_y = \|y_j - C_l x_j\|_2$$

1. Initially, randomly generate N vectors of predefined orders with different parameter values. Each vector with different set of parameter values denotes a different cluster.
2. Obtain the prediction error of sample j with respect to each randomly generated cluster i using Equation
3. Calculate the fuzzy membership of sample j with respect to cluster i as follows:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{N} \left(\frac{PE_{ij}}{PE_{ij}}\right)^{2/q-1}}$$

$\forall j = 1, 2, \ldots M; \quad \forall i = 1, 2, \ldots N$ M - samples, N - clusters

4. Update the model parameters (cluster centers) using the gradient descent algorithm as follows:

$$C_i^{r+1} = C_i^r - \alpha^r \nabla g^r$$

Where the step length ( ) and gradient are calculated as:

$$\alpha^r = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} \mu_j^q \left(\nabla g^\top x_j\right)^T (y_j - C_i^\tau x_j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} \mu_i^q \left(\nabla g^T x_j\right)^T (\nabla g^T x_j)}$$

where,

$$\nabla g^r = \frac{\partial f}{\partial C_i} = \left(\sum_{j=1}^{M} \mu_i^q \left(y_j - C_i^T x_j\right) x_j^T \right)$$

5. Calculate the prediction errors for all samples with respect to the updated models
6. Compute root mean square error as follows (where b is the best model fit for sample j):

$$RMSE = \sqrt{\frac{\sum_{j=1}^{M} PE_{bj}^2}{M}}$$

7. Terminate if any pre-specified criteria is satisfied (number of iterations exceeds the limit or RMSE less than predefined limit) and go to next step else go to step 3
8. Assign each data sample to respective clusters based on prediction errors
9. Calculate the cosine angle between each model to the others and merge like models[37]

$$\theta_{ik} = \cos^{-1}\left(\frac{C_i C_k^T}{\|C_i\| \, \|C_k\|}\right)$$

Cosine angle between models i and k:

10. The models that have fewer data samples ($0.05M$) are discarded and the data points are reassigned to models that fit them best
11. Calculate the final model parameters using ordinary least squares (OLS).

Once the final set of models are obtained at the end of an iteration (step 1 to step 11), each model is tested with F-test[38] to identify whether a particular variable is significant or not.

12. Each model is tested using F-test, and if any variable in a particular model is identified as insignificant then that variable will be removed thus reducing the model order.

13. If any of the models contain insignificant variables then the whole clustering approach is restarted with a new number of models and their orders i.e. go to step 1 with modified 'N' and their individual orders, else report the final model orders and parameters.

In this study, the multiple model learning approach contains two stages. In the first stage, to obtain the true number of models and their individual orders, the clustering approach described above is used on model data set. In the first iteration, in step 1, five models are initiated with 13 input variables (12 scaled structural features and intercept) with random model parameters. From the second iteration onwards, the models are initiated with the final set of models that are obtained in the previous iteration. Once the true models and corresponding model orders are obtained, in the second stage, to obtain robust model parameters we use a K-fold (K value as 5) validation-based approach. Model data set is equally divided into K equal partitions. In each fold, data in *K-1* partitions is used to build the model and remaining data to test the obtained models. In K-fold validation, switching 20% of data samples each time for a new fold results in a substantial deviation in model parameters. Hence, to obtain robust multiple models, an iterative weight based optimization approach is used[28], in which the models information in previous fold is included in form of weight based objective for next fold. In this approach, the final set of models in step 11 are obtained using a weight based optimization approach for the objective specified in Equation with λ value as 10 and the final model parameters (C_{pev}) obtained in the previous fold as initial guess. In case of first fold of the first phase, both in step 1 (initiation of the models) and in step 11 (for the optimization problem) the model parameters obtained from the first stage (i.e. identification of the true number of models and model orders) are used. This procedure is repeated until the respective models in all the folds are relatively close, which can be measured using the similarity metric proposed in our earlier work[29]. In the second stage, models are neither merged nor discarded (i.e. step 9 and 10) and statistical testing (i.e. step 12 and 13) is also avoided since clustering procedure in this stage is assumed to start with the true number of models and their orders.
Weight based objective:

$$\min_{C} \left(\left(\sum_{i=1}^{Nt} \left(y_i - y_i^{prd} \right)^2 \right) + \lambda \left(C - C_{pm} \right)^2 \right)$$

Similarity metric:

$$\varphi = \max\left(\phi_{i,k}\right); \quad \forall i \in N; k \in K$$
$$\text{where} \quad \phi_{1,k} = \max\left(\min\left(\theta^k\right); \quad \forall j \in N\right)$$
$$\forall k' \in K \wedge k' \neq k;$$

The pseudo code for the identification of multiple model parameters (second stage) as follows:

1. **Do while:**

   For fold in 1 to *K*-folds:

   (a) Divide the whole data into respective training and testing data sets

   (b) Initialize the clusters with the final model parameters provided in the previous fold.

   (c) Follow the clustering procedure provided above from step 2 to step 8

   (d) Obtain the final model parameters using a weight based optimization approach for the objective specified in Equation with the final model parameters obtained in the previous fold. **End**

2. Obtain the similarity metric of the multiple models obtained in all folds

3. If the similarity metric is smaller than the tolerance or larger than the similarity metric in the previous iteration, then terminate and report the averaged model parameters respectively over all the K-folds as the final set of model parameters, else, continue. **End**

In our earlier work [28], a prediction error-based K-nearest neighbours testing approach is proposed in order to identify a appropriate model for a new test sample. *In this case study, we used a weighted prediction error based* K*-nearest neighbours method.* The weights provided are inversely proportional to the distance from the test sample to the neighbour i.e. a neighbour, which is nearest to the test sample will have more impact than a neighbour which is far from the test sample. The clustering procedure is started assuming five linear models of order 13 (12 scaled structural features and intercept) and converged to three models of different orders. It is interesting to note that in different operating regions of feature space, different features are found to be significant. The details of the converged models along with the number of data points (of model data set) that belong to each model is given in Table 4. It is interesting to note that VABC, MW, and AMR are found to be insignificant in the whole feature space, which is also validated by the neural network. It can be observed from Table 4 that though MLFER_A, MLFER_BH, MLFER_S, and MLFER_E are found to be more significant in the features set, no feature is found to be significant in the complete feature space. It is interesting to note from the coefficients reported in Table 4 that solubility is directly proportional to hydrogen bond acidity and combined polarizability and inversely proportional to the excessive molar refraction in the complete feature space but with different proportionality constants.

All the data samples are associated with the final set of models obtained in the above iterative procedure based on the final prediction error. Association of these data samples is further useful to select a suitable model for any new test sample (i.e. for global test set or for any novel molecule). Prediction accuracy of the final set of multiple models on model data set, test data set and on overall data is given in Table 2. It can be seen from the table that multiple models perform better than any other approach. The adjusted RMSE and R^2 values demonstrate that the final set of multiple linear models can be used for estimating the solvation free energy of Quinone molecules. Due to the capability of neural networks in identifying the non-linear dynamics, the neural network (NN) based QSPR approach is shown to be better than the OLS approach; however, both approaches were not robust to estimate solvation free energy for a wide range of Quinone derivatives. It is interesting to note that though the estimates of the GC method on model data set is slightly better than the NN based approach, NN estimates are better on the global test set. This can be attributed to the overfitting of contributions due to highly sparse input data in the group contribution approach.

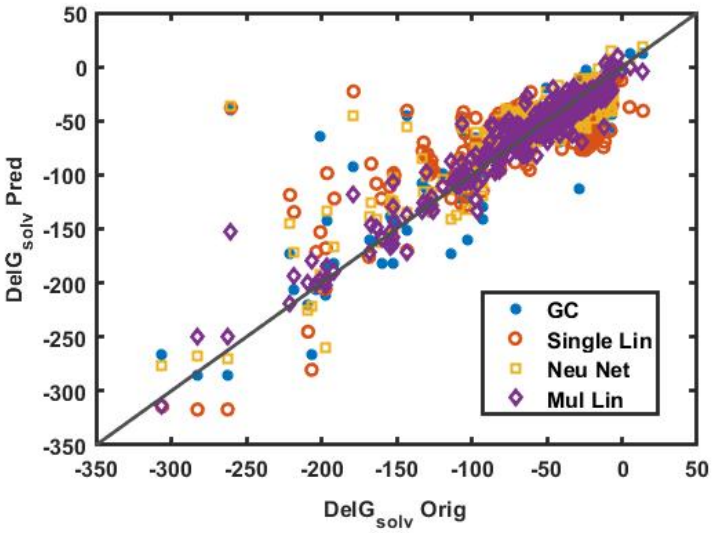

Figure 3: Solvation free energy original vs predicted using several approaches

The solvation free energy values obtained using several approaches are plotted against their original values in Figure 3. It can be verified from the figure that multiple linear models have better estimates compared to other approaches throughout the range of solvation free energy values, especially for high solubility molecules (samples having solvation free energy values ranging between -150 to -310 kJ/mol).

4 Exploration of two functional group substituted quinone molecules

In any machine learning based study, interpretation and verification of the results of models is of utmost importance. In this section, we attempt to interpret and derive insights from the results of the machine learning models on the existing Quinone derivatives data set. These insights are further validated by constructing a new set of Quinone derivatives with two different functional groups, using the insights.

In this study, one, two and three ringed quinones with NH_2 and PO_3H_2 substituents are explored as these functional groups possess relatively high solubility values in the data. The models proposed in the previous sections are used to predict solvation free energy values which are then analysed, to come to conclusions on what factors could help increase the solubility of quinone derivates. Based on these insights, new molecules are designed aiming at different solubility ranges. Following this, predictions are made on these new molecules using both the models to see if they match our expectation.

For this purpose, a new dataset with 61 structures consisting of anthraquinone, benzoquinone and naphthoquinones substituted by NH_2 and PO_3H_2 at single position or multiple positions was formulated. The 12 descriptors which are used by the proposed models were generated and fed to the models in order to obtain their prediction. From Table 5 it is observed that the values of predictions from both the models have different ranges. However, the predictions have a high Pearson Correlation Coefficient of 0.9405, indicating a strong linear correspondence between the predictions. Figure 4 plots the solvation free energy predictions from the neural network against those from multiple linear models, in a parity plot equalising the ranges, visually exhibiting the strong linear correspondence. It can further be observed that the structure with lowest solvation free energy according to multiple linear models, is also the one with lowest solvation free energy according to the neural network predictions.

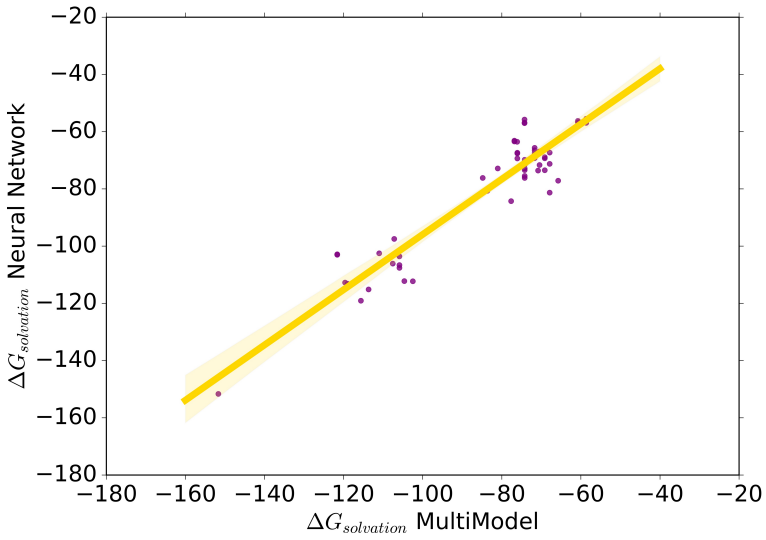

Figure 4: Parity plot of solvation free energy predictions from neural net and multiple linear models

Since the values of solvation free energy from multiple linear models and neural network are highly correlated, and as we observed from section 3 (Table 4), that multiple linear models have slightly better performance compared to neural network models with respect to adjusted R2 and RMSE values, we proceed to further analysis using the solvation free energy values obtained from multiple linear models.

Analysing the structures in the dataset, it is seen that most of the 61 chosen structures with NH_2 and PO_3H_2 have solubility in the range of [-80, -60], which can be verified from Figure 5 which plots the frequency distribution of the solvation free energy values from multiple linear models. It is to be noted that lower the solvation free energy value, higher is the solubility.

In this section, we derived new features for easy interpretation of models. The derived new features, which help in visualising and deriving insights on the molecules, are: 'Aromaticity', 'NH2', 'PO3H2', 'O-O', 'NH2-O',

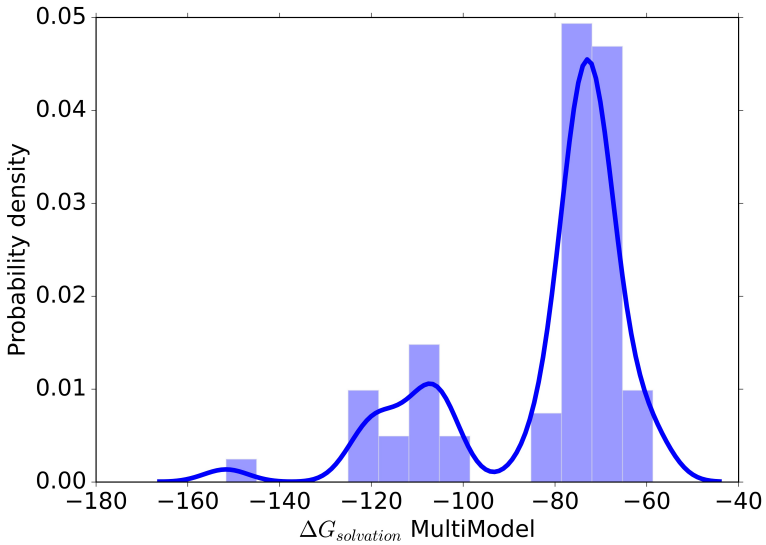

Figure 5: Probability distribution of solvation free energy values in the dataset

'PO3H2-O','Rings','Structure','Base Structure'. These have been explained as follows:

1. 'Aromaticity': explains whether the structure is aromatic or not
2. 'delocal NH_2': number of delocalised NH_2 in the structure
3. 'delocal PO3H2': number of delocalised PO_3H_2 in the structure
4. 'Rings': number of rings in the structures
5. 'Structure': type of isomer, on which functional groups are attached. For example: 1,2 AQ
6. 'Base Structure': type of parental quinone – AQ, BQ, NQ on which substituents are added, derived from the number of rings
7. 'O-O', 'NH2-O', 'PO3H2-O': minimum bond distances between the two groups, used to account for inductive effect. The method to calculate these distances are explained below.

Procedure used to obtain O-O', 'NH2-O', 'PO3H2-O':

1. 'O-O': Each structure has two double bond Os. Start at one O. Minimum number of positions to reach next double bond O is the value of the feature.
2. 'NH2-O': For each NH_2 present in the structure, locate the nearest double bond O and fill in the number of positions required to reach it starting from that particular NH_2.
3. 'PO3H2-O': Calculated in a manner similar to 'NH2-O'

These bond distances were added to analyse the effect of -I effect (Inductive effect) of various substituents on the solubility of the molecule. Among new features, it can be seen that the solubility depends on number of rings. Higher the number of rings, more the solubility.

4.1 Effect of Quinone Base Structure on Aqueous Solubility of Quinones:

The base structures taken were anthraquinone (AQ) having three rings, naphthoquinone (NQ) having two rings and benzoquinone (BQ) having one ring. On plotting the distribution of solvation free energy predictions, of

each of these three categories of quinone derivatives, i.e., one ringed two ringed and three ringed structures, as in Figure 6, it can be seen that the as the number of rings increase, the mean solvation free energy decreases, indicating increasing solubility. Thus, anthraquinone structures are more soluble in water followed by naphthoquinone, followed by benzoquinone.

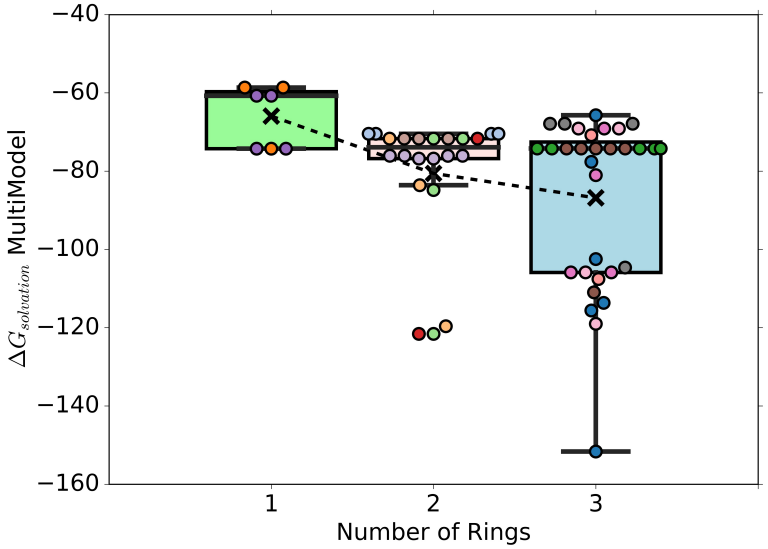

Figure 6: Solvation free energy distributions of molecules grouped by number of rings in the structure

4.1.1 Trends of Solubility among Anthraquinone Isomers:

Figure 7 analyses the variation of solubility among the various anthraquinone isomers. It is observed that 1,2-anthraquinone derivatives have a lower mean and median solvation free energy values as compared to the other anthraquinone isomers, while 9,10-anthraquinone derivatives tend to have the least negative median solvation free energy values.

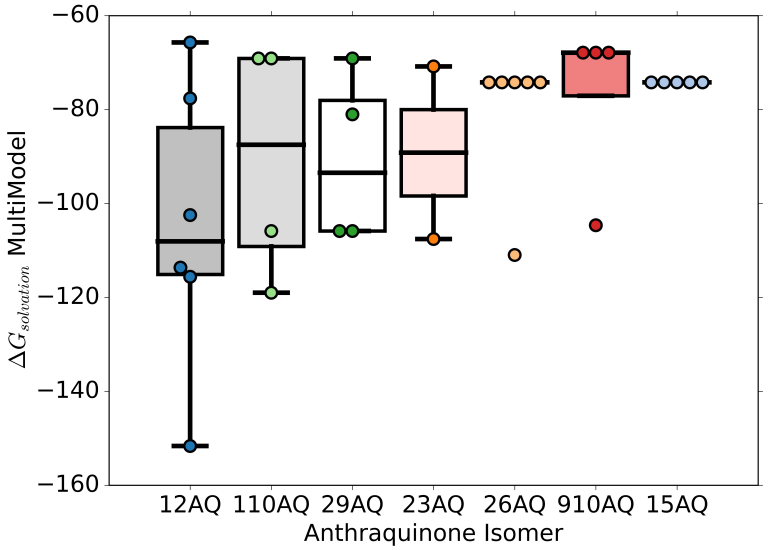

Figure 7: Solvation free energy distributions of Anthraquinone molecules

4.1.2 Trends of Solubility among Naphthoquinone Isomers:

The solvation free energy values for the various naphthoquinone isomers reveal that the 1,7-NQ derivatives are more likely to be soluble than others, while the 1,2-NQ derivatives are less likely to be soluble as compared to the others. This can be verified from Table 7, which lists the median solvation free energy values for each of the naphthoquinone isomers, as well as Figure 8, which plots the distribution of solvation free energy values for each naphthoquinone isomer in a boxplot.

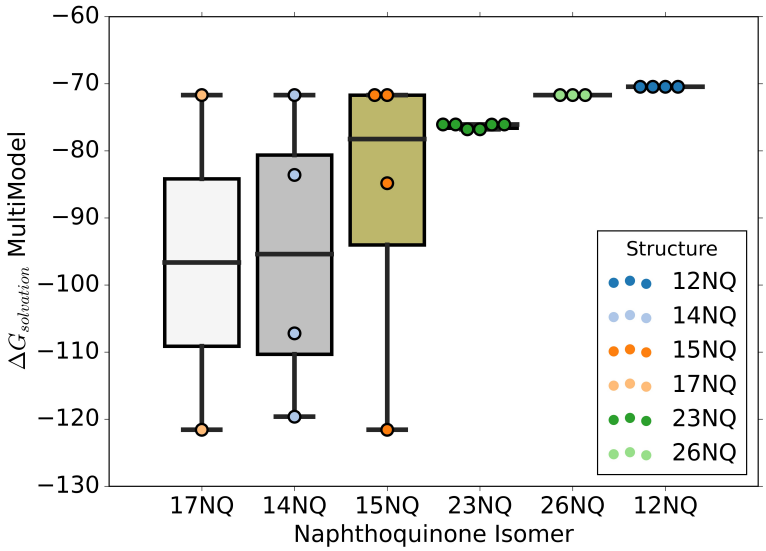

Figure 8: Solvation free energy distributions of Napthoquinone molecules

4.1.3 Trends of Solubility among Benzoquinone Structures

Benzoquinones are one ring structures with only two possible isomers, i.e., 2,3-benzoquinone and 1,4-benzoquinone. Among the two, 2,3-benzoquinone is seen to have the lowest median solvation free energy value, indicating greater solubility as compared to their 1,4-benzoquinone isomer. This can be inferred from Table 8, which lists the median solvation free energy values as well as Figure 9, which plots their distribution.

4.1.4 Combined Trends of Solubility among all Quinone Derivatives

On putting all derivatives and isomers of the derivatives together, it can be observed that 1,2-anthraquinone structures have the lowest median solvation free energy, while 1,4- benzoquinone structures have the highest median solvation free energy values, which matches our expectation from section 4.1, 4.2 and 4.3. It must be noted, however, that there are slight deviations from the expected trend such as 1,7-napthaquinone having lower median solvation free energy than 2,9-anthraquinone even though anthraquinones have lower median solvation free energy than naphthoquinones. These exchanges may be attributed to resonance and inductive effects prevalent more in some structures as compared to others, discussed in the upcoming sections. Figure 10 collates the distributions of all the isomers discussed in section 4.1, 4.2 and 4.3 arranged in increasing order of solvation free energy.

4.2 Impact of Inductive Effect on the Solubility of Quinones

4.2.1 Minimum Bond Distance between NH_2-O

It is known that NH_2 is a +M, -I group. On plotting the minimum number of positions of NH_2 from O for various structures vs their distribution of solvation free energy, as in Figure 11(a), it can be observed from

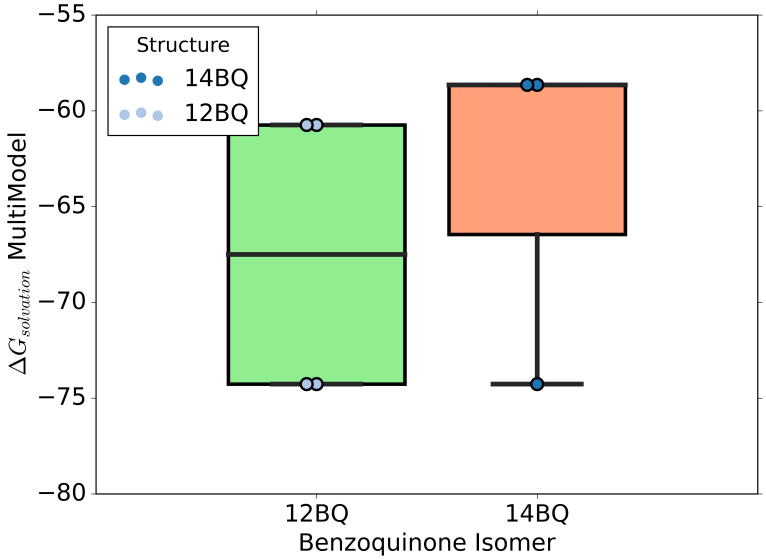

Figure 9: Solvation free energy distributions of Benzoquinone molecules

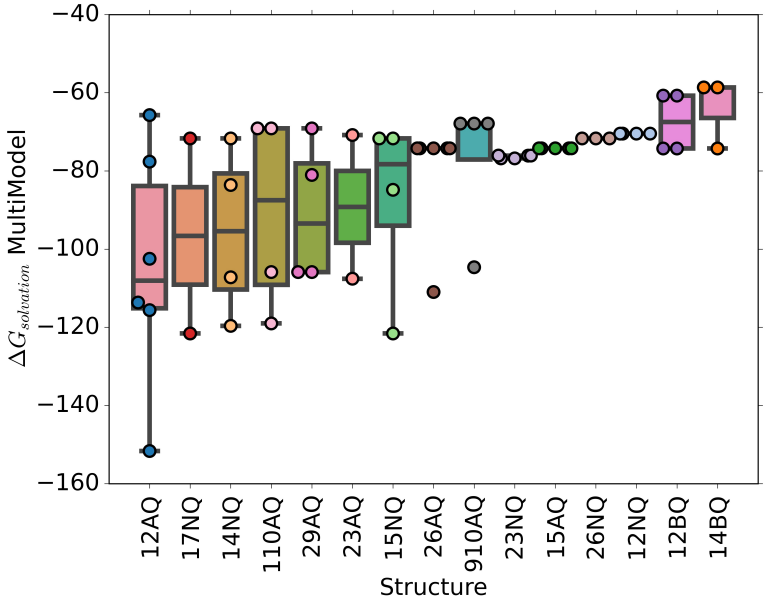

Figure 10: Solvation free energy distributions for all Quinone structures in the dataset

the curve connecting the means of each group that, as the NH_2 group comes closer to the double-bond Os, the solubility increases and solvation free energy decreases. As the distance begins to increase, solvation free energy increases as well, impacting the solubility negatively. At large distances however, the inductive effect of the group on the double bond Os decays rapidly, losing its impact on solubility. It must be noted that the minimum distance between NH_2 and O is taken only because inductive effect decays rapidly with increasing bond distance. The closest NH_2 group would have the largest impact since, greater the distance, the weaker the effect.

It must be pointed out that even though structures having NH_2-O distance as 1 have low median solubility, they have a wide range. This may be attributed to the fact that in all the structures, the effect of NH_2 alone

could not be isolated. These structures also contain NH_2 and PO_3H_2 at other positions which also affect solubility, leading to the wide distribution of free solvation energies.

4.2.2 Minimum Bond Distance between PO_3H_2-O

It is known that PO_3H_2 is a -M, -I group. On plotting the minimum number of positions of PO_3H_2 from O for various structures vs their distribution of solvation free energy, as in Figure 11(b), it can be seen that the curve connecting the means of each group has a decreasing trend, such that, on increasing the distance between PO_3H_2 groups and double bond O, solvation free energy of the molecule decreases indicating increasing solubility.

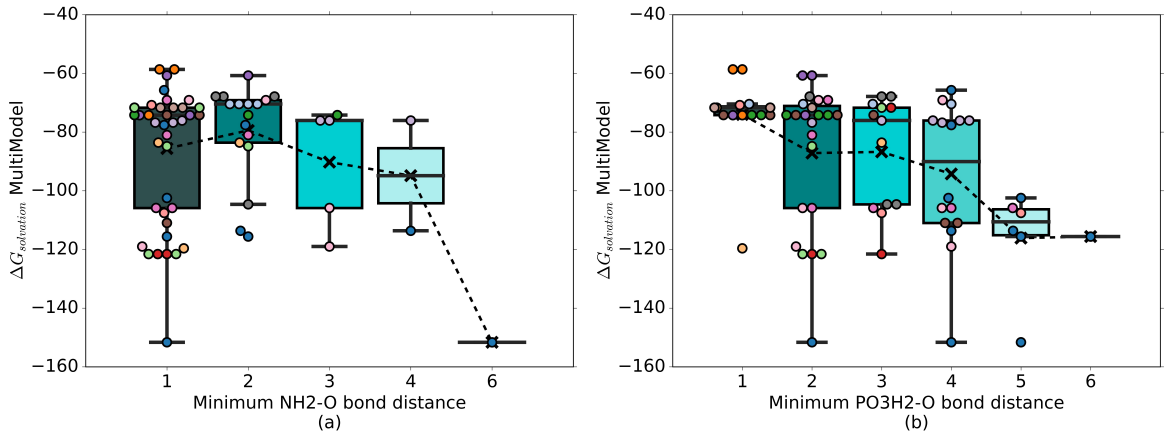

Figure 11: Solvation free energy distributions of molecules grouped by minimum distance between (a) NH_2 and O in the molecule (b) PO_3H_2 and O in the molecule

4.2.3 Impact of Resonance Effect on Solubility of Quinones

In the previous section we studied the impact of inductive effect on the solvation free energies of quinones. Resonance effect, on the other hand, does not decay with bond distance. The degree of delocalisation, however, depends on the electrophilic/nucleophilic nature of the functional group as well as the electron density in the rings. In the case presented above, consisting of NH_2 and PO_3H_2 derivatives of quinones, since NH_2 is a +M group and PO_3H_2 is a -M group, no cases of cross-conjugation arise.

Figures 12(a) and (b) plot the solvation free energy values against number of NH_2 and PO_3H_2 groups present in the molecule. It can be deduced that as the number of NH_2/ PO_3H_2 groups increase, the molecule becomes more polar and the solvation free energy increases, thus increasing aqueous solubility.

4.3 Effect of Aromaticity on Solubility of Quinones

Figure 13 shows the plot of solvation free energy values, grouped by their aromaticity. It must be noted that the aromatic structures have a wider range of solubility values as compared to the non-aromatic structures. The median as well as mean solvation free energy values of aromatic structures is lower. Hence, it can be concluded that, on being given a quinone derivative with high solubility, it is more likely to be aromatic than non-aromatic. In this case study, all boxplots shown have been realised as swarm plots, in which points belonging to the same category in the x-axis, having the same y-axis value are plotted in a spread-out manner instead of overlapping with each other. It must be pointed out that even though it appears to be a 2-D plot, it remains a 1-D plot with a category plotted in the x-direction and a continuous feature on the y-axis. The points in each box of the boxplot are spread out only to prevent overlap, not to indicate a second dimension.

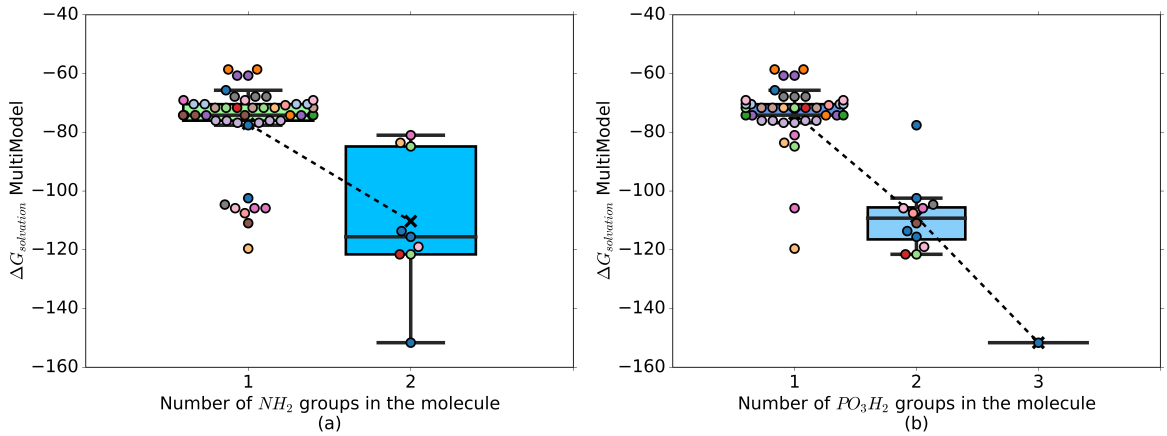

Figure 12: Solvation free energy distributions of molecules grouped by number of (a) NH_2 substituents (b) PO_3H_2 substituents in the molecule

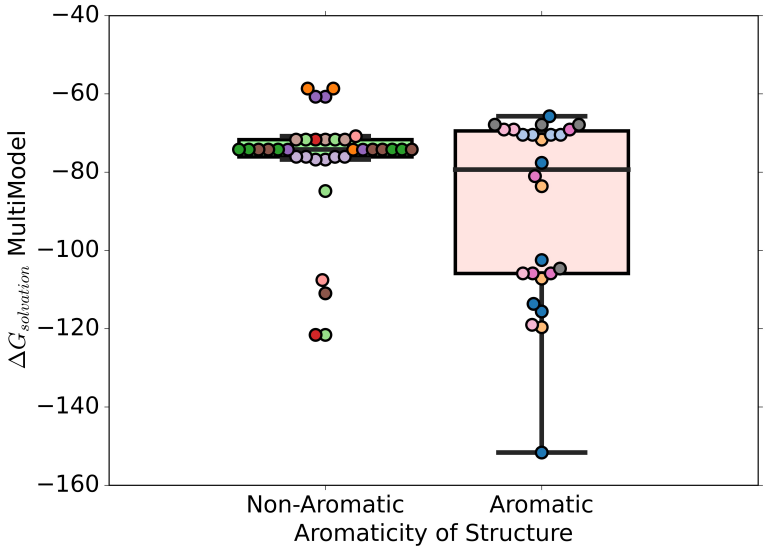

Figure 13: Solvation free energy distributions of molecules grouped by aromaticity

4.4 Effect of Number of Aromatic Rings on Solubility of Quinones

The 61 structures in the dataset consist of one, two and three-ringed structures which have varying number of aromatic rings. Since quinones must have two double bond Os, at least one ring must be non-aromatic. Observing Figure 14 which plots the solvation free energy distributions of the molecules grouped by the number of aromatic rings in their structure, it can be stated that as number of adjacent aromatic rings increases, solubility increases since the solvation free energy decreases, as extended resonance takes place. It must be noted, however, that two non-adjacent aromatic rings do not contribute much to solubility since there is a lack of extended conjugation. To explore this idea further, the ratio of aromatic carbons to the total number of carbons for each molecule is plotted against their solvation free energy values, in Figure 15. This is done because considering solely the absolute number of rings may not be enough to capture the trend. For example, the presence of one aromatic ring in a molecule with three rings must be treated differently from the presence of one aromatic ring in a molecule with two rings, due to their differing molecular weights. This figure further stresses on the fact that two non-adjacent aromatic rings, with a ratio of $12/14 = 0.86$ does not contribute any

more solubility than the structures with a single aromatic ring. Further, this plot reveals that for the same number of aromatic carbons, if total number of carbons is higher, solvation free energy is lower, as the molecular weight increases with increasing number of rings. This can be deduced from the fact that molecules with one aromatic ring in a two-ring structure having less negative median solvation free energy than those with one aromatic ring in a three ringed structure.

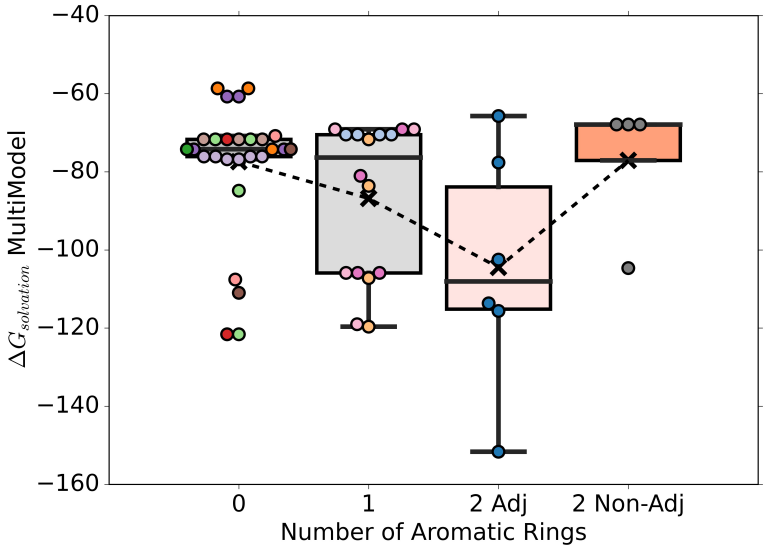

Figure 14: Solvation free energy distributions of molecules grouped by number of aromatic rings in the molecule

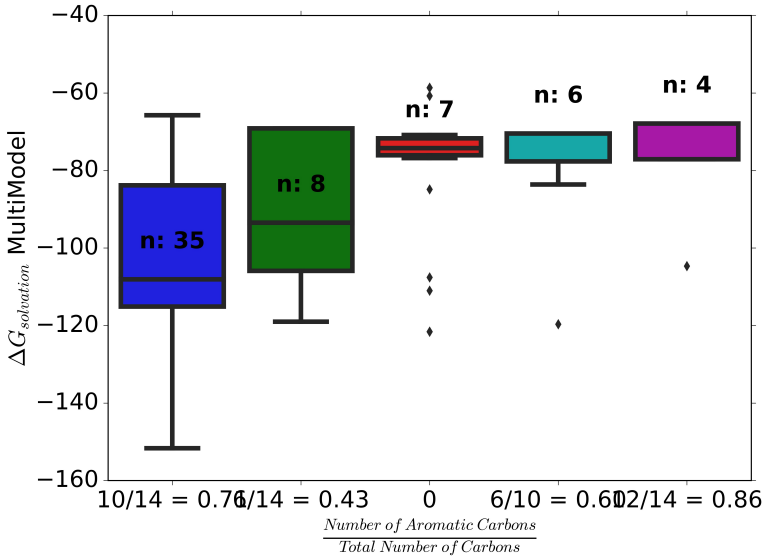

Figure 15: Solvation free energy distributions of molecules grouped by ratio of number of Aromatic carbons to the total number of carbons. Each box is labelled by the number of samples of each category

4.5 Combined Effect of Substituents on Solubility of Quinones

In the earlier sections, the effects of each of the substituent one at a time were analysed even though it was not possible to isolate their effects. In this section, the combined effects of these substituents are analysed

with the help of Figure 16, in which solvation free energy values are plotted against the number of NH_2 and PO_3H_2 groups with labels of the format *(number of* NH_2 *groups) | (number of* PO_3H_2 *groups)*. It can be seen, in Figure 17, addition of PO_3H_2 and NH_2 increases solubility in all structures except in 2,9-AQ (17(e)) and 1,4-NQ (17(g)) , addition of NH_2 reduces solubility. This anomaly may have multiple reasons. Firstly, we are considering the predictions from the models for the analysis, however these predictions may differ from the true values, leading to the anomalous trend. Further, it may be the case that in these structures, the electron density in the ring is high and thus, when another NH_2 substituent is added, its electron pair upon delocalising adds to the electron density of the ring, reducing its stability, making it a less favourable arrangement. Figure 18, shows the combined effect of the substituents on all the anthraquinone isomers. From this it can be stated that as the number of substituents increases, the median solvation free energy decreases indicating an increase in solubility. However, the addition of a PO_3H_2 group decreases solvation free energy much more than the addition of an NH_2 group.

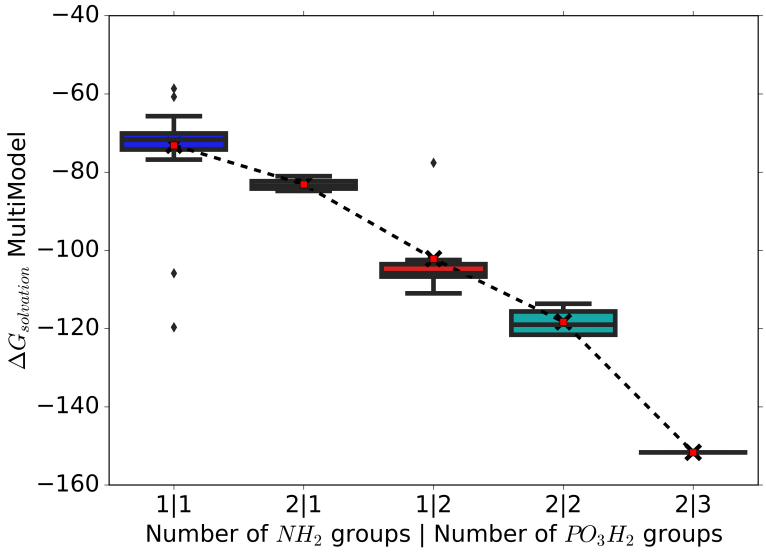

Figure 16: Plots indicating change solvation free energy with number of NH_2 and PO_3H_2 groups for all structures combined. Labels are of the form: *(number of* NH_2 *groups) | (number of* PO_3H_2 *groups)*

On plotting the effect for naphthoquinone structures, as in Figure 19, it can be seen that solubility increases with increase in number of substituents, matching the expectations built by the previous experiment. However, remarks on the effect of PO_3H_2 vs NH_2 cannot be made since there are no structures having 1 NH_2 and 2 PO_3H_2 in the dataset, to compare with the existing structure of 2 NH_2 and 1 PO_3H_2. In case of benzoquinone, all structures in the dataset have 1 NH_2 and 1 PO_3H_2 group and hence this comparison cannot be carried out.

5 Validation of new molecules derived from ML model insights

The key takeaways from the analysis in the previous section are discussed below. On considering quinones with two substituents one with +M effect and another with -M effect the following can be said:

1. A molecule with larger number of rings is more likely to have higher solubility than one with fewer rings
2. An aromatic molecule is more likely to have higher solubility than a non-aromatic one.
3. Among aromatic quinones, one with larger extended conjugation is likely to be more soluble than one with non-adjacent aromatic rings.
4. Smaller the minimum bond distance between -I substituent and double bond O, more the probability of it having high solubility.

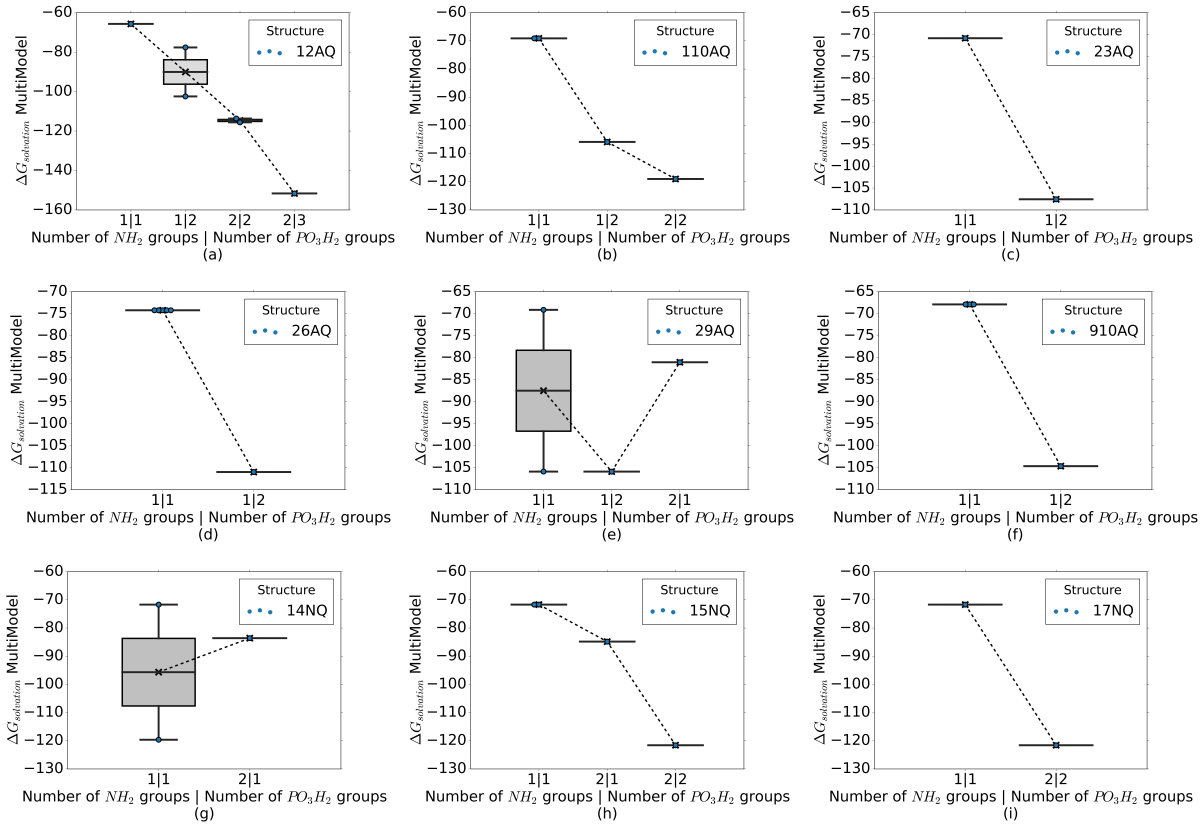

Figure 17: Plots indicating change in solvation free energy with number of NH_2 and PO_3H_2 groups for each individual structure. Labels are of the form: *(number of NH_2 groups)/(number of PO_3H_2 groups)*

5. Increasing the number of substituents may help in increasing the solubility of quinones

Using these insights, new molecules were designed with different solubility ranges in mind. The results can be seen, listed in Table 9. In can be observed that the solubility predictions of most of the molecules are in line with their expected ranges. The two molecules highlighted are the ones that do not conform to the prediction. It must also be mentioned that new base structures – Tetracenequinone (TQ), Phenanthrenequinone (PQ), Chrysenequinone (CQ), which are not a part of any of the previous datasets. These have been included to demonstrate the generalisability of the inferences drawn from the previous case study.

6 Exploration of new functional group substituted quinone molecules

This case study differs from the previous exploration in multiple ways. In the previous study, we studied the combined effects of Electron Donating Group (NH_2) and Electron Withdrawing Group (PO_3H_2) groups. In this study, we isolate the effect of Electron Donating Group only. The earlier study experimented with number and position of both substituents. However, in this study, we place the Electron Donating Group at either one position or all positions and see its effect on different kinds of structures - with different number of rings, aromaticity etc. This study uses O- as the Electron Donating Group in all the structures. O- like NH_2 shows weak inductive effect (-I) and strong resonance effect (+M), stronger than (NH_2). Most importantly, this is a crucial step to establish the generalisability of our model, being extended to newer functional groups with less information, reliably. In the previous study we looked at functional groups that are present in the training set. However, in this section we choose to dope using functional groups that are not present in the training set, thus nullifying the use of Group Contribution Methods. This helps us in establishing the use of multiple models and

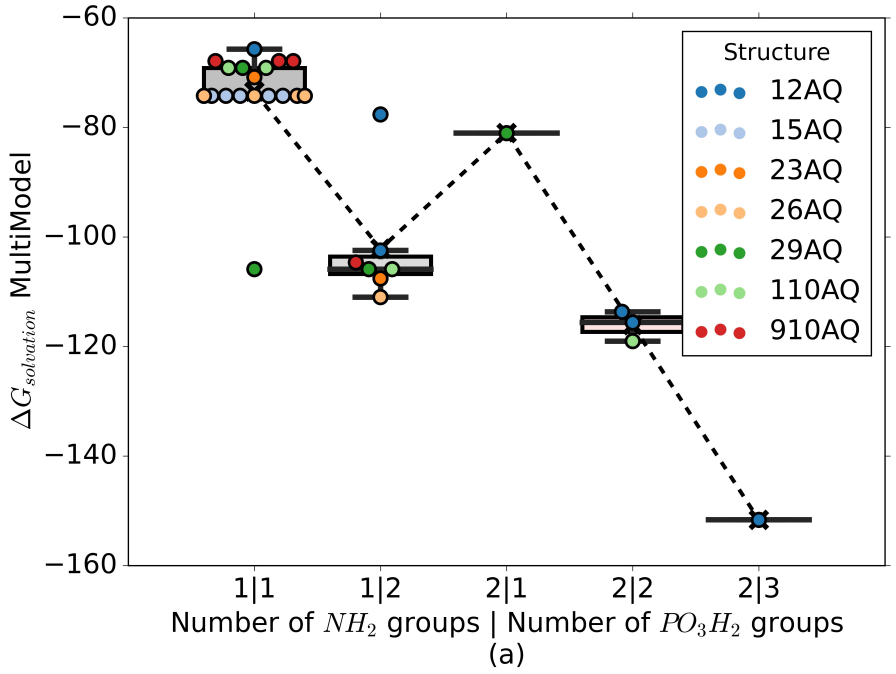

Figure 18: Solvation free energy distribution for Anthraquinone molecules groups by number of NH_2 and PO_3H_2 groups

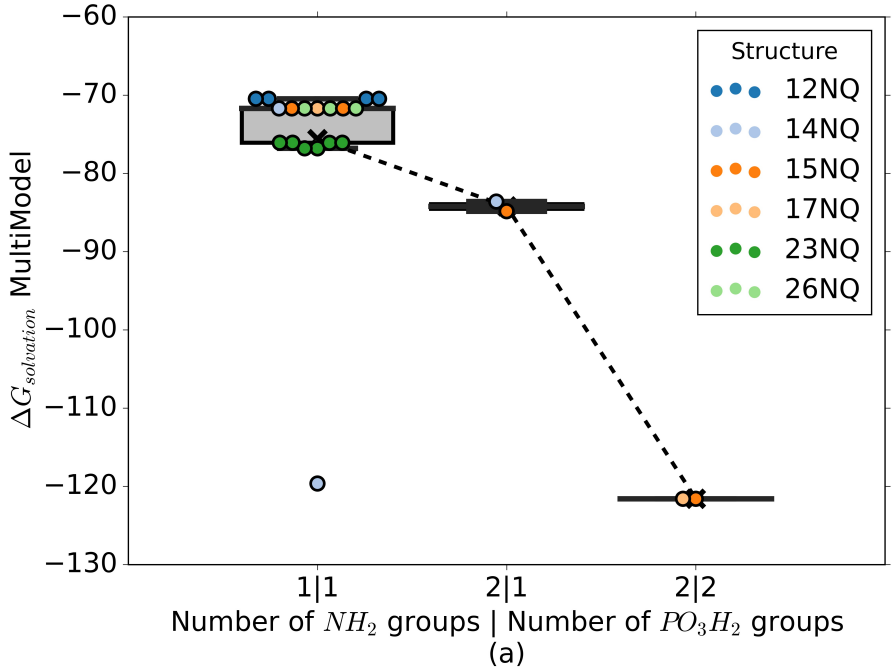

Figure 19: Solvation free energy distribution for Naphthoquinone molecules grouped by number of NH_2 and PO_3H_2 molecules

neural networks for datasets that are reasonably far from the training set in the sample space, allowing us to

use these methods for obscure molecules for which pre-existing data is not easily found

The dataset used in this case study consists of 49 molecules with O- as the substituent. The descriptors used by the models are generated, following which the samples are fed to the models to generate predictions for solvation free energy, from both the models.

In Figure 20 which shows the distribution of solvation free energy values in the dataset, it can be observed that most of the samples lie in the region of medium solubility, while a few have high solubility, with highly negative solvation free energy values. Table 10 highlights the range of predictions of solvation free energy from both the aforementioned models.

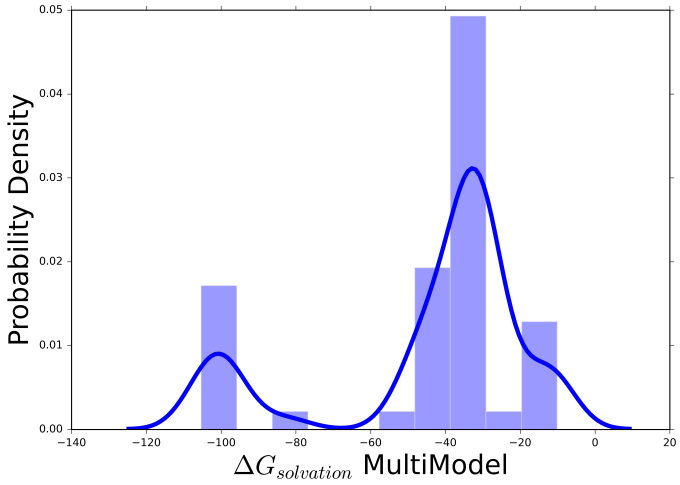

Figure 20: Probability distribution of solvation free energy values in the new test set

Following are the inferences from this study:

6.1 *Effect of Number of Rings on Solubility of Quinones: A*s the number of fused rings increases, solvation free energy decreases

6.2 *Effect of Aromaticity of the Structure on Solubility of Quinones:* No conclusive inference can be drawn

6.3 *Effect of Number of Aromatic Rings on Solubility of Quinones:* As delocalisation in the molecule increases, solvation free energy decreases

6.4 *Effect of Number of Substitutions on Solubility of Quinones: A*s the number of electron donating substituents increases, solvation free energy first increases and then decreases.

6.5 *Effect of Bond Distances on Solubility of Quinones:* As the distance increases, solvation free energy increases, until the impact of Inductive effect completely wears out

6.6.1 *Insights on Solubility of Anthraquinone Isomers:* Trends in all the features, except trend in solubility with respect to aromaticity, remain the same as the previous case study. For aromaticity, however, no conclusion can be drawn

6.6.2 *Insights on Solubility of Naphthoquinone Isomers:* Trends are similar to Anthraquinone Isomers

6.6.3 *Insights on Solubility of Benzoquinone Isomers:* Trends in all the features, except trend in solubility with respect to aromaticity, remain the same as the previous case study. For aromaticity, however, no conclusion can be drawn. For more insights and details on each of these inferences, please refer to the Supplementary Material.

7 Conclusion

In this study, different machine learning based QSPR approaches along with the GC approach have been tested to predict solvation free energy of Quinone derivatives. For an unbiased comparison of these approaches, we used adjusted root mean squared error and adjusted R^2 values since ML approaches are parameter sensitive. It is observed from the reported metrics in a train-test validation approach that multiple model based QSPR

approach performs better than the other approaches. It should be noted that GC approach estimates are restricted by the variety of groups or fragments considered in the training data set hence it can not be used to predict for the molecules that are designed using a different set of functional groups.

It is identified from GC approach that substituting hydrogen atoms with groups like PO_3H_2, COOH etc. can increase the solubility of Quinones. Using feature selection methods in ML models - it is observed that structural features like overall hydrogen bond basicity (MLFER_BH) and acidity (MLFER_A), combined polarizability (MLFER_S) and excessive molar refraction (MLFER_E) of the Quinone derivative are significant in predicting the aqueous solubility. Proposed ML framework has accurately classified the newly designed molecules into low, medium and high solubility categories as per the expectations thus enabling the ML insights to design novel set of molecules with higher solubility values.

This work can be further extended to obtain a robust correlation to estimate reduction potential of Quinone molecules and by using these correlations, Quinone variants search space can be explored in an inverse multi-objective optimization framework or design potential molecules based on insights for flow battery applications.

Acknowledgments:

We would like to thank Robert Bosch centre for Data Science and Artificial Intelligence for providing computational facilities.

Supporting Information:

Provided supporting information contains three tables. Table S1, contains the number of occurrences of predefined groups in each Quinone derivative in the data along with estimated contributions of each group. Table S2, contains details of structural descriptors of all the Quinones derivatives considered for this case study. Table S3, consists of mean solvation free energy values for quinones, substituted by O-, from neural network and multiple linear models for Aromatic and Non-Aromatic molecules, as detailed in section 6.

References

[1] Trung Nguyen and Robert F Savinell. Flow batteries. *The Electrochemical Society Interface*, 19(3):54, 2010. 1

[2] Puiki Leung, Xiaohong Li, Carlos Ponce De León, Leonard Berlouis, CT John Low, and Frank C Walsh. Progress in redox flow batteries, remaining challenges and their applications in energy storage. *Rsc Advances*, 2(27):10125–10156, 2012. 1

[3] Grigorii L Soloveichik. Flow batteries: current status and trends. *Chemical reviews*, 115(20):11533–11558, 2015. 1

[4] Süleyman Er, Changwon Suh, Michael P Marshak, and Alán Aspuru-Guzik. Computational design of molecules for an all-quinone redox flow battery. *Chemical science*, 6(2):885–893, 2015. 1, 2, 3

[5] Leonidas Constantinou and Rafiqul Gani. New group contribution method for estimating properties of pure compounds. *AIChE Journal*, 40(10):1697–1710, 1994. 1

[6] Elisa Conte, Ana Martinho, Henrique A Matos, and Rafiqul Gani. Combined group-contribution and atom connectivity index-based methods for estimation of surface tension and viscosity. *Industrial & Engineering Chemistry Research*, 47(20):7940–7954, 2008. 1

[7] Chao Gao, Rakesh Govind, and Henry H Tabak. Application of the group contribution method for predicting the toxicity of organic chemicals. *Environmental Toxicology and Chemistry: An International Journal*, 11(5):631–636, 1992. 1

[8] Kevin G Joback and Robert C Reid. Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications*, 57(1-6):233–243, 1987. 1

[9] KM Klincewicz and RC Reid. Estimation of critical properties with group contribution methods. *AIChE Journal*, 30(1):137–142, 1984. 1

[10] Jorge Marrero and Rafiqul Gani. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*, 183:183–208, 2001. 1, 2

[11] Jorge Marrero and Rafiqul Gani. Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Industrial & Engineering Chemistry Research*, 41(25):6623–6633, 2002. 1

[12] A Correa, JF Comesana, JM Correa, and AM Sereno. Measurement and prediction of water activity in electrolyte solutions by a modified asog group contribution method. *Fluid phase equilibria*, 129(1-2):267–283, 1997. 1

[13] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, 2015. 1

[14] Suhani J Patel, Dedy Ng, and M Sam Mannan. Qspr flash point prediction of solvents using topological indices for application in computer aided molecular design. *Industrial & Engineering Chemistry Research*, 48(15):7378–7387, 2009. 1

[15] Alan R Katritzky, Yilin Wang, Sulev Sild, Tarmo Tamm, and Mati Karelson. Qspr studies on vapor pressure, aqueous solubility, and the prediction of water- air partition coefficients. *Journal of Chemical Information and Computer Sciences*, 38(4):720–725, 1998. 1

[16] Markus Muehlbacher, Ahmed El Kerdawy, Christian Kramer, Brian Hudson, and Timothy Clark. Conformation-dependent qspr models: logpow. *Journal of chemical information and modeling*, 51(9):2408–2416, 2011. 1

[17] Pablo R Duchowicz and Eduardo A Castro. Qspr studies on aqueous solubilities of drug-like compounds. *International journal of molecular sciences*, 10(6):2558–2577, 2009. 1

[18] Feng Luan, Ting Wang, Lili Tang, Shuang Zhang, and MNDS Cordeiro. Estimation of the toxicity of different substituted aromatic compounds to the aquatic ciliate tetrahymena pyriformis by qsar approach. *Molecules*, 23(5):1002, 2018. 1

[19] Saeed Yousefinejad and Bahram Hemmateenejad. Chemometrics tools in qsar/qspr studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149:177–204, 2015. 1

[20] Tomoyuki Miyao, Hiromasa Kaneko, and Kimito Funatsu. Inverse qspr/qsar analysis for chemical structure generation (from y to x). *Journal of chemical information and modeling*, 56(2):286–299, 2016. 1

[21] Lu Xu and Wen-Jun Zhang. Comparison of different methods for variable selection. *Analytica Chimica Acta*, 446(1):475–481, 2001. 7th International Conference on Chemometrics and Analytical Chemistry Antwerp, Belgium, 16-20 October 2000. 1

[22] Yi-Leh Wu, Cheng-Yuan Tang, Maw-Kae Hor, and Pei-Fen Wu. Feature selection using genetic algorithm and cluster validation. *Expert Systems with Applications*, 38(3):2727–2732, 2011. 1

[23] Dimitris K Agrafiotis and Walter Cedeno. Feature selection for structure- activity correlation using binary particle swarms. *Journal of medicinal chemistry*, 45(5):1098–1107, 2002. 1

[24] Saeed Yousefinejad, Fatemeh Honarasa, and Hanieh Montaseri. Linear solvent structure-polymer solubility and solvation energy relationships to study conductive polymer/carbon nanotube composite solutions. *RSC Advances*, 5(53):42266–42275, 2015. 1

[25] Bahram Hemmateenejad. Optimal qsar analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based pcr. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(11):475–485, 2004. 1

[26] David J Livingstone, David T Manallack, and Igor V Tetko. Data modelling with neural networks: advantages and limitations. *Journal of computer-aided molecular design*, 11(2):135–142, 1997. 1

[27] Shuning Wang and Masahiro Tanaka. Nonlinear system identification with piecewise-linear functions. *IFAC Proceedings Volumes*, 32(2):3796–3801, 1999. 1

[28] Sivadurgaprasad Chinta and Raghunathan Rengaswamy. Machine learning derived quantitative structure property relationship (qspr) to predict drug solubility in binary solvent systems. *Industrial & Engineering Chemistry Research*, 58(8):3082–3092, 2019. 1, 3, 3.3, 3.3

[29] Sivadurgaprasad Chinta, Abhishek Sivaram, and Raghunathan Rengaswamy. Prediction error-based clustering approach for multiple-model learning using statistical testing. *Engineering Applications of Artificial Intelligence*, 77:125–135, 2019. 1, 3.3, 3.3

[30] Mohammad Goodarzi, Bieke Dejaegher, and Yvan Vander Heyden. Feature selection methods in qsar studies. *Journal of AOAC International*, 95(3):636–651, 2012. 3

[31] Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474, 2011. 3

[32] George R Famini, Carl A Penski, and Leland Y Wilson. Using theoretical descriptors in quantitative structure activity relationships: Some physicochemical properties. *Journal of physical organic chemistry*, 5(7):395–408, 1992. 3

[33] Abolghasem Jouyban and Mohammad AA Fakhree. Experimental, computational methods pertaining to drug solubility. *Toxicity and drug testing*, 1, 2012. 3

[34] M Jalali-Heravi, M Asadollahi-Baboli, and P Shahbazikhah. Qsar study of heparanase inhibitors activity using artificial neural networks and levenberg–marquardt algorithm. *European journal of medicinal chemistry*, 43(3):548–556, 2008. 3.2

[35] Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264, 2003. 3.2

[36] Ahmed Adebowale Adeniran and Sami El Ferik. Modeling and identification of nonlinear systems: A review of the multimodel approach—part 1. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(7):1149–1159, 2017. 3.3

[37] Vidyashankar Kuppuraj and Raghunathan Rengaswamy. Evaluation of prediction error based fuzzy model clustering approaches for multiple model learning. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 4(1):10–21, 2012. 9

[38] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John Wiley & Sons, 2010. 11

8 Tables:

Group	Contribution	Group	Contribution	Group	Contribution
aC-H	-0.2443	aC-COOH	-20.3895	C-CHO	-6.4206
aC-$N(CH_3)_2$	-4.4997	aC-PO_3H_2	-27.7461	C-$COOCH_3$	-5.8861
aC-NH_2	**-14.2381**	aC-SO_3H	-17.346	C-CF_3	3.2318
aC-OCH_3	1.2739	aC-NO_2	-0.4057	C-CN	-9.2176
aC-OH	**-15.0724**	C-H	-2.3058	C-COOH	-22.6403
aC-SH	0.4298	C-$N(CH_3)_2$	-6.6263	C-PO_3H_2	-32.5914
aC-CH_3	2.2893	C-NH_2	**-13.2065**	C-SO_3H	-19.5789
aC-SiH_3	4.2378	C-OCH_3	3.8392	C-NO_2	0.5313
aC-F	4.1042	C-OH	-8.9845	*both C=0 on different rings*	*-25.3076*
aC-Cl	2.0636	C-SH	0.7729		
aC-C_2H_3	1.495	C-CH_3	-3.6894	*both C=0 side by side of same ring*	*-29.7715*
aC-CHO	-5.6467	C-SiH_3	3.9062		
aC-$COOCH_3$	-4.3014	C-F	1.687	*both C=0 in opposite sides of same ring*	*-21.0962*
aC-CF_3	2.9312	C-Cl	-10.4269		
aC-CN	-9.6447	C-C_2H_3	-1.5827		

Table 1: All 41 groups that are considered for the case study along with contributions

		RMSE	Adj. RMSE	R^2 value	Adj. R^2
Group Contribution	Model	21.5371	23.0393	0.7806	
	G test	17.2844	24.4439	0.8728	
	Over all	20.7505	21.8820	0.8019	
Single linear	Model	28.5975	28.9561	0.6132	
	G test	25.4517	26.7921	0.7242	
	Over all	27.9921	28.2714	0.6395	
Neural network	Model	22.3767	23.4077	0.7632	
	G test	16.2358	20.0070	0.8878	
	Over all	21.2825	22.0546	0.7916	
Multiple linear	Model	12.1045	12.4340	0.9307	
	G test	15.4585	17.3628	0.8983	
	Over all	12.8508	**13.1279**	0.9240	

Table 2: Performance metrics of several approaches for solvation free energy estimation

Variable	Rank	Variable	Rank	Variable	Rank	Variable	Rank
MLFER_A	1	MLFER_E	4	AMR	7	Apol	10
TopoPSA	2	TSA	5	MW	8	Si	11
MLFER_BH	3	MLFER_S	6	McG_Vol	9	VABC	12

Table 3: Features ranking obtained using stepwise approach for NN-QSPR

Model	Active features and their coefficients in final averaged models	No.of samples
1	Apol, Si, McG_Vol, MLFER_BH, MLFER_S, MLFER_E [-153.670 51.156 101.901 -30.166 -8.982 24.902]	127
2	MLFER_A, MLFER_BH, MLFER_S, TSA [-22.249 35.523 -56.263 57.925]	22
3	Si, MLFER_A, MLFER_E, TopoPSA [-9.876 -5.696 16.634 -56.429]	176

Table 4: Details of the final set of multiple models converged

Model	Minimum value (kJ/mol)	Maximum value (kJ/mol)
Neural Network	-192.329	-70.379
Multiple Linear Models	-151.658	-58.650

Table 5: Range of $\Delta G_{solvation}$ values from neural network and multiple linear models

Structure	Median $\Delta G_{solvation}$
1,2-AQ	**-108.07535**
2,9-AQ	-93.45825
2,3-AQ	-89.20515
1,10-AQ	-87.50450
2,6-AQ	-74.22910
1,5-AQ	-74.22910
9,10-AQ	-67.89440

Table 6: Median Solvation free energy values for various Anthraquinone structures

Structure	Median $\Delta G_{solvation}$
1,7-NQ	**-96.64655**
1,4-NQ	-95.41375
1,5-NQ	-78.26925
2,3-NQ	-76.09135
2,6-NQ	-71.69910
1,2-NQ	-70.46630

Table 7: Median Solvation free energy values for various Napthoquinone structures

Structure	Median $\Delta G_{solvation}$
1,2-BQ	-67.5067
1,4-BQ	-58.6499

Table 8: Median Solvation free energy values for benzoquinone isomers

Table 9: New molecules designed based on inferences

Solubility Expectation~	**Postion-based Nomenclature of Molecule**	**Reason for Expectation**	**DelG-Solv_NN**	**DelG-Solv_Multi-Model**
Low Solubility	R2-PO3H2_1,4-BQ	Number of Rings = 1 Number of Aromatic Rings = 0 Number of PO3H2 = 0 Number of NH2 = 1 NH2_O distances = 1	-31.604	-37.5163
	R2-PO3H2_R6-NH2_1,4-BQ	Number of Rings = 1 Number of Aromatic Rings = 0 Number of PO3H2 = 1 PO3H2_O distance = 1 Number of NH2 = 1 NH2_O distance = 1 Added a PO3H2 to previous structure	-70.3792	-58.6499
	R2-PO3H2_R5-NH2_1,4-BQ	Number of Rings = 1 Number of Aromatic Rings = 0 Number of PO3H2 = 1 PO3H2_O distance = 1 Number of NH2 = 1 NH2_O distance = 1 Changed position of PO3H2 in previous structure	-72.23	-58.6499
	R2-PO3H2_R3-NH2_1,5-NQ	Number of Rings = 2 Number of Aromatic Rings = 0 Number of PO3H2 = 1 PO3H2_O distance = 1 Number of NH2 =~ 1 NH2_O distances = 2	-87.125	-71.6991
	R2-PO3H2_R6-NH2_1,5-NQ	Number of Rings = 2 Number of Aromatic Rings = 0 Number of PO3H2 = 1 PO3H2_O Distances = 1 Number of NH2 = 1 NH2_O distances = 1	-88.1262	-71.6991
	R3-PO3H2_R7-NH2_2,6-NQ	Number of Rings = 2 Number of Aromatic Rings = 0 Number of PO3H2 = 1 PO3H2_O distance = 1 Number of NH2 = 1 NH2_O distance = 1	-88.7794	-71.6991
Medium Solubility	R2,3-NH2_R10-PO3H2_5,12-TQ	Number of Rings = 4 Number of Aromatic Rings = 3 Extended Conjugation = 2 Number of PO3H2 = 1 PO3H2_O distance = 4 Number of NH2 = 2 NH2_O distances = 3,3	-81.5747	-74.8587

	R3-PO3H2_R4,8-NH2_2,9-AQ	Number of Rings = 3 Number of Aromatic Rings = 1 Number of PO3H2 = 1 PO3H2_O distance = 1 Number of NH2 = 2 NH2_O distances = 2,2	*-92.8386*	*-81.0347*
	R7,8-NH2_R5,6-PO3H2_1,4-AQ	Number of Rings = 3 Number of Aromatic Rings = 2 Extended Conjugation = 2 Number of PO3H2 = 2 PO3H2_O distance = 4,5 Number of NH2 = 2 NH2_O distances = 4,5	-158.217	-114.1852
	R10-NH2_R1,9,8-PO3H2_2,6-AQ	Number of Rings = 3 Number of Aromatic Rings = 0 Number of PO3H2 = 3 PO3H2_O distances = 1,2,3 Number of NH2 = 1 NH2_O distance = 3	-170.1693	-147.0206
	R2,6-NH2_R3,4,5-PO3H2_1,7-NQ	Number of Rings = 2 Number of Aromatic Rings = 0 Number of PO3H2 = 3 PO3H2_O distance = 2,2,3 Number of NH2 = 2 NH2_O distances = 1,1	-163.9477	-158.3487
	R4-NH2_R5,6,7,8-PO3H2_2,3-NQ	Number of Rings = 2 Number of Aromatic Rings = 0 Number of PO3H2 = 4 PO3H2_O distances = 3,3,4,4 Number of NH2 = 1 NH2_O distances = 1 Added a PO3H2 and removed an NH2 from previous structure	*-158.9288*	*-187.065*
High Solubility	*R3,6-NH2_R9,11-PO3H2_1,4-TQ (tetracenequinone)*	Number of Rings = 4 Number of Aromatic Rings = 3 Extended Conjugation = 3 Number of PO3H2 = 2 PO3H2_O distance = 6,7 Number of NH2 = 2 NH2_O distances = 1,4	*-154.9984*	*-108.9298*
	R3-NH2_R5,6,9,11-PO3H2_1,2-CQ (Chrysenequinone)	Number of Rings = 4 Number of Aromatic Rings = 3 Extended Conjugation = 3 Number of PO3H2 = 4 PO3H2_O distance = 4,6,7,8 Number of NH2 = 1 NH2_O distances = 3	-227.4058	-169.2988

R4-NH2_R5,6,7,8,9-PO3H2_1,2-PQ (Phenanthrenequinone)	Number of Rings = 3 Number of Aromatic Rings = 2 Extended Conjugation = 2 Number of PO3H2 = 5 PO3H2_O distance = 3,5,5,6,6 Number of NH2 = 1 NH2_O distances = 1	-238.6056	-151.9886
R3-NH2_R5,6,7,8,9-PO3H2_1,2-PQ	Number of Rings = 3 Number of Aromatic Rings = 2 Extended Conjugation = 2 Number of PO3H2 = 5 PO3H2_O distance = 3,5,5,6,6 Number of NH2 = 1 NH2_O distances = 2 Increased NH2_O distance by 1 from previous structure	-231.4838	-212.7444

Model	**Minimum value (kJ/mol)**	**Maximum value (kJ/mol)**
Neural Network	-78.16	-13.18
Multiple Linear Models	-105.35	-10.14

Table 10: Range of $\Delta G_{solvation}$ values from neural network and multiple linear models