Using Quantum Chemical Features in a Neural Network to Improve Aqueous Solubility Prediction

by

Brett Jeffrey Ondich

Bachelor of Science, University of Pittsburgh, 2018

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Brett Jeffrey Ondich

It was defended on

December 6, 2022

and approved by

Dr. Kenneth Jordan, Distinguished University Professor, Department of Chemistry

Dr. Peng Liu, Professor, Department of Chemistry

Thesis Advisor: Dr. Geoffrey Hutchison, Associate Professor, Department of Chemistry

Using Quantum Chemical Features in a Neural Network to Improve Aqueous Solubility Prediction

Brett Jeffrey Ondich, MS

University of Pittsburgh, 2022

Aqueous solubility is a vital molecular property in numerous fields, such as drug discovery and material design. Accurate prediction of molecular aqueous solubility can reduce the number of potential candidates prior to experimental analysis. Shrinking the chemical search space can result in streamlining the selection process, saving valuable time and resources. Recent developments have increased interests in utilizing machine learning techniques to computationally predict aqueous solubility rather than experimentation. One such technique is the Molecular Attention Transformer (MAT). Transformers are a special case of graph neural networks (GNN). GNNs utilize inputs in the form of graphs that have data stored as nodes and edges, which can be thought of as atoms and bonds, respectively. An important aspect of building a GNN is determining which features to use as descriptors for the nodes and edges. This paper investigates the effects of including quantum chemical data as node features in a GNN model. The hypothesis was that by including this quantum data, the model will be able to better discriminate between compounds of high similarity and more accurately predict their aqueous solubility. However, there was no significant improvement in model performance when the quantum data was included in the model. The accuracy of the quantum data was analyzed to determine if the performance did not improve due to the data or the model. It was determined that the solvation models being used to compute the quantum data were unable to produce data at a level of accuracy to enable the model to benefit from the inclusion of the quantum features. Furthermore, a recently published model pretrained on

quantum data was compared to the base model being used to determine if including quantum features improves performance. The quantum model outperformed the base model, further showing that including quantum features should improve model performance but requires quality quantum data.

Table of Contents

List of Tables

List of Figures

Preface

I would like to first and foremost thank each of my committee members for their guidance and support in helping me to complete my Master's thesis. Thanks go out to the faculty and staff of the Department of Chemistry whose support and encouragement went above and beyond during my time in the program. I would also like to thank my fellow group members who helped me during my time including Omri Abarbanel, Caroline Chun, Brianna Greenstein, Maya Hayden, Danielle Heiner, Keren Lee, and Annaliese Schmidt, as well as Paul Francoeur for assistance with his neural network model. Thank you to Emily Stickney for her love and support during this time. Special thanks to the United States Air Force for funding me in this endeavor.

1.0 Introduction

In this work, the use of quantum chemical data as node features in a graph neural network to improve the accuracy of an aqueous solubility model will be tested. It is believed that by including more relevant node features, the graph neural network will be able to better discern physically similar compounds and thus more accurately predict the aqueous solubility of the compound. The improved model can then be used to effectively screen potential candidates based on molecular aqueous solubility, which is important in fields such as drug discovery and material design.

1.1 Importance

The total chemical space has been estimated to be 10^{180} compounds.[1] This means that the number of possible compounds is more than twice the number of atoms in the universe. Current molecular screening libraries are nowhere close to reaching this number but continue to rapidly increase. As the size of molecular screening libraries increase, the ability to accurately predict molecular properties becomes vital for fields such as drug discovery and material design.

Based on Lipinski's rule-of-five for oral bioavailability, the "drug-like" chemical search space has been estimated at 10^{60} organic molecules.[1] Ideally, one would be able to directly measure the molecular properties of a given compound. However, this approach is slow and expensive. Pharmaceutical research and development of new molecules entails substantial investment with usually over 10 years until patients can access the new products.[2] Couple this slow approach with

the fact that the overall failure rate in drug development is over 96%, including a 90% failure rate during clinical development[3], and it is not a surprise that the mean capitalized research development investment to bring a new drug to market is estimated to be around $1.3 billion[4].

A few of the unfortunate consequences of this high cost of research and development are significantly inflated prices of the few successful drugs, which are priced in order to recoup the incurred cost of historical failures, and the discouraging of real innovation where the developmental risk is greater.[3] Many of these failures could have been potentially avoided by accurately predicting a clinically relevant property of the compound, such as aqueous solubility. Thereby, narrowing down the vast chemical search space to candidates that are more likely to succeed.

1.2 General Solubility Equation

The solubility of a solid in water depends on two factors: the crystallinity of the solute and the interaction of the solute with water. An early attempt at predicting the aqueous solubility of a molecule is the general solubility equation (GSE)[5], which can be used to estimate the aqueous solubility of a set of organic nonelectrolytes. The GSE is a simple way of estimating the aqueous solubility since the only inputs used are the Celsius melting point (MP) and the octanol water partition coefficient (K_{ow}).[5] The GSE does not use any fitted parameters and thus does not require a training set containing analogs of test compounds.[5]

The revised GSE proposed by Jain and Yalkowsky[6] utilizes five fitted parameters, decreasing the average absolute error from 0.56 to 0.43, resulting in a more accurate version than the original GSE. However, it is clear that until an adequate description of the lattice energy (or

the crystalline state) of the material is available, progress on predicting solubilities ab initio will be limited.[7] The differences between crystalline and amorphous solubility can be large, which can have significant effects on the observed pharmacokinetics of the formulation.[7] This effect is often used to increase the solubility of a compound during drug design.[8]

1.3 Solubility Challenges

The motivation for improving prediction capability of a molecule's aqueous solubility for pharmaceutical companies can be clearly seen when Pfizer Institute for Pharmaceutical Materials Science & Unilever Centre for Molecular Informatics issued a challenge to the cheminformatics community to develop a method to better predict aqueous solubility.[7] The challenge believed that serious deficiencies in the consistency and reliability of solubility data found in literature was one of the main reasons solubility is such a difficult property to predict.[7]

Therefore, a training set containing the solubility values of 100 druglike molecules measured using a technique called chasing equilibrium (CheqSol). CheqSol produces a precipitate after several cycles, switching back and forth between a supersaturated and a subsaturated solution. The final precipitate obtained is thermodynamically driven and the solubility data are highly reproducible with an associated error of approximately 0.05 log units.[7]

Using this high-precision set of 100 molecules as a training set, contestants attempted to predict the aqueous solubility of 32 novel druglike molecules. Contestants employed the entire spectrum of approaches, including multiple linear regression (MLR) and random forest regression (RFR), available at the time (2008). However, no one approach distanced itself from the other methods.[9]

Ten years after the initial solubility challenge, another challenge was issued to examine the extent to which computational methods had improved.[10] One of the main differences between the first and second challenge was that participants were allowed to use their own training sets, as long as the training set did not contain any of the test molecules. The findings of the second challenge concluded that no improvement in the prediction of solubility is recognizable and that the new methods perform equally well as older ones [11], clear indication there is more work to be done. The challenge did not limit the participants to any particular model, but all competitors did submit predictions based on quantitative structure-property relationship (QSPR) approaches. The main type of model used were artificial neural networks, which accounted for 30% of the models submitted.[11]

1.4 Neural Networks

While the solubility challenge is no longer accepting entries, the challenge of predicting aqueous solubility is still resulting in new methods being created. A recently published machine learning algorithm for predicting a molecule's aqueous solubility is SolTranNet.[12] SolTranNet is an optimized fork of the molecule attention transformer (MAT).[13] MAT is a transformer that is adapted to chemical molecules by augmenting the self-attention with inter-atomic distances and molecular graph structure.[13] Transformers are a special case of graph neural network (GNN).

A GNN utilizes graphs as inputs with nodes and edges depicting the relationship between a group of entities. SolTranNet is designed to create a 2D graph representation of a molecule from the molecule's simplified molecular input line entry system (SMILES) representation.[12] Molecules are transformed into a graph representation by treating the atoms as nodes and the bonds between

the nodes as edges. These nodes and edges can be further described using an array of descriptors, which can help the neural network further learn about the graph.[14] Descriptors resulting from quantum chemical calculations could be of use improving the ability of SolTranNet to predict a compound's aqueous solubility.

1.5 Solvation Models

Quantitative prediction of thermodynamics properties of solute molecules requires an accurate description of the solvent. To accomplish this, a solvation model may either have explicit solvent molecules or an implicit description of the solvent environment.[15] Implicit, or continuum, denotes that the solvent is not represented explicitly but rather as a dielectric medium with surface tension at the solute-solvent boundary.[16] It is because of this structureless continuum, that the number of interacting particles and the number of degrees of freedom of a system are significantly reduced, considering that explicit solvent molecules can contribute over 90% of atoms in a simulated system. The relatively high computational cost of explicit solvent models has resulted in implicit solvent models remaining popular.[15] The solvation model allows the quantum chemistry calculations to include the interactions between solvents and the quantum solute.

Three implicit solvation models that can be used for quantum chemical calculations are the analytical linearized Poisson-Boltzmnn (ALPB) model, generalized Born model with surface area contributions (referred to as GBSA), and the solvent model based on density (SMD). The solvation model ALPB is a robust and efficient method to implicitly account for solvation effects in modern semiempirical quantum mechanics and force fields. When used to calculate hydration free energies of small molecules, ALPB is nearing the accuracy of more sophisticated explicitly solvated

approaches, with a mean absolute deviation of 1.4 kcal/mol compared to the experiment.[17] The generalized Born models are widely used for molecular dynamics simulations of proteins and nucleic acids. These approaches model hydration effects and provide solvent-dependent forces with efficiencies comparable to molecular mechanics calculations on the solute alone[18].

The remainder of this paper will detail the methodology that was utilized to test the hypothesis that including quantum chemical data as node features in a neural network will improve the model's ability to predict aqueous solubility. The results of the work will then be shown along with a discussion of the results. Finally, the conclusions that can be drawn from the results and the potential future work will be stated.

2.0 Methods

2.1 Quantum Chemical Data Generation

For this project, the AqSolDB [19] was the primary data set utilized for training the machine learning models, as it was the largest publicly available data set. AqSolDB spans a wide range of solubility values and is collated from differing data sets, however it was only screened for identical molecules and did not verify whether those solubilities were measured in buffered conditions or water or at what pH the measurement was taken. This is especially noteworthy as these differing conditions can change the measurement by orders of magnitude. Nonetheless, this dataset was used since it has been observed that neural network models tend to perform better with larger data sets, even if the data contains more noise[20]. Only the SMILES strings and reported solubilities (log S, S in mol/L) were utilized.

From the compounds' SMILES, a 3-dimensional structure was generated by using RDkit [21]. The conformer was initially optimized by minimizing the geometry by the application of a molecular mechanics force field. RDkit [21] uses Merck molecular force field (MMFF) family of force fields [22]. After the conformers are generated using distance geometry, the ETKDG method of Riniker and Landrum [23], which uses torsion angel preferences from the Cambridge Structural Database, is used to correct the conformers. Since RDkit merely provides quick 3D structures, it is not intended to be a replacement for a "real" conformer analysis tool. For this reason, the Conformer-Rotamer Ensemble Sampling Tool (CREST) [24] was utilized to generate the favored conformation. CREST utilizes GFN2-xTB, which is an extended semiempirical tight-binding

model, to provide the thermally accessible ensemble of minimum-energy structures [25], which is the most likely form the compounds will be in once in solution.

Once the minimum-energy structure of the compound was determined by CREST, it was introduced into an implicit solvent model using the quantum chemistry program xTB [26]. The geometry of the conformer was further optimized using xTB, which has a built-in geometry optimizer called approximate normal coordinate rational function optimizer (ANCopt), which uses a Lindh-type model Hessian to generate an approximate normal coordinate system [26]. Using water as the solvent, the generalized Born with solvent accessible surface area contributions (GBSA) [26] solvation model was used to calculate the following atomic quantum chemical variables: partial charge, coordination number, dispersion coefficient, and polarizability. The values for the atomic quantum chemical variables were then included as node features during training for the aqueous solubility prediction tool, SolTranNet [12].

2.2 Quantum Data Accuracy Analysis

Different combinations of solvation models and density functional were used to compute solvation free energies to compare to experimentally determined solvation free energies. This was done to determine whether the quantum chemical calculation data being used was accurate enough to enable the machine learning algorithm to learn and thus more accurately predict a compound's aqueous solubility. Three different solvation models were compared to investigate the accuracy of the computed solvation free energy, GBSA and ALPB along with the universal solvation model based on density (SMD) [16]. GBSA and ALPB were utilized in xTB, while SMD was available in the quantum chemistry program ORCA [27]. For this, both the geometries optimized by CREST and

the geometries provided in the MNSOL database [28] were used to represent the two tested functionals, GFN2-xTB and M06-2X, respectively. The MNSol database consists of a collection of 3037 experimental free energies of solvation for 790 unique solutes in 92 solvents, including water. For SMD in ORCA, the solvation free energy was calculated by taking the difference between the gas-phase energy and the SMD energies.

2.3 Published Model Comparison

The generalization of two published models, SolTranNet and SolProp_ML [29] was also investigated. SolProp_ML is said to be a more robust model since it is the first modeling tool that can predict the solid solubility for a broad range of solvents and temperatures [29]. Vermeire et al. utilize the ability of machine learning to transfer learn to pretrain the deep neural network on two databases, CombiSolv-QM and CombiSolv-Exp, and then fine tune the network with experimental data [30]. They argue that the transfer learning approach improves the performance on higher molar mass solutes compared to direct training of the deep neural network on experimental data [30]. The comparison consisted of predicting the aqueous solubility of all the compounds in the AqSolDB dataset and the MNSOL dataset using the two models and then comparing the computed aqueous solubilities to the experimentally determined aqueous solubilities.

3.0 Results

3.1 Quantum Data Node Features

The performance of SolTranNet with and without the additional quantum chemical node features was measured using the coefficient of determination and the root-mean-square-error (RMSE). The model with quantum chemical node features is referred to as quantum SolTranNet, while the model without quantum chemical node features is referred to as SolTranNet. The two models were trained for 2000 training epochs and then used to predict the aqueous solubility of a withheld training set. Quantum SolTranNet predicted the testing set at an RMSE of 0.926 and a coefficient of determination is 0.838. SolTranNet predicted with an RMSE of 0.927 and coefficient of determination of 0.848. The linear correlation between the computed aqueous solubilities and experimental aqueous solubilities of Quantum SolTranNet and SolTranNet can be seen in Figures 1 and 2, respectively.

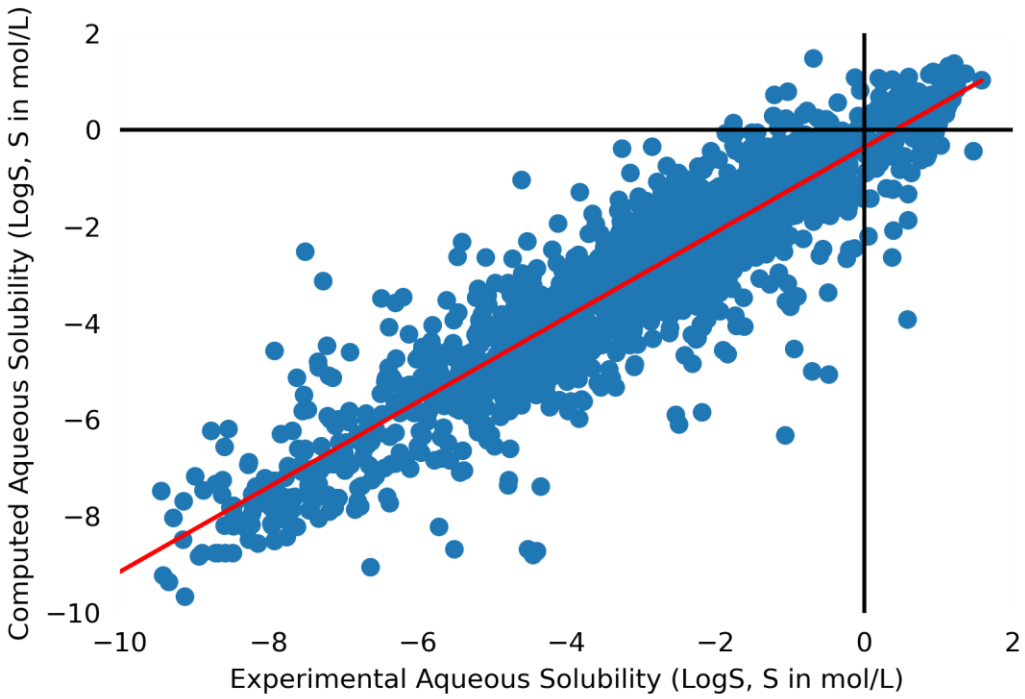

Figure 1. Comparison of aqueous solubility values predicted by Quantum SolTranNet to experimentally derived values.

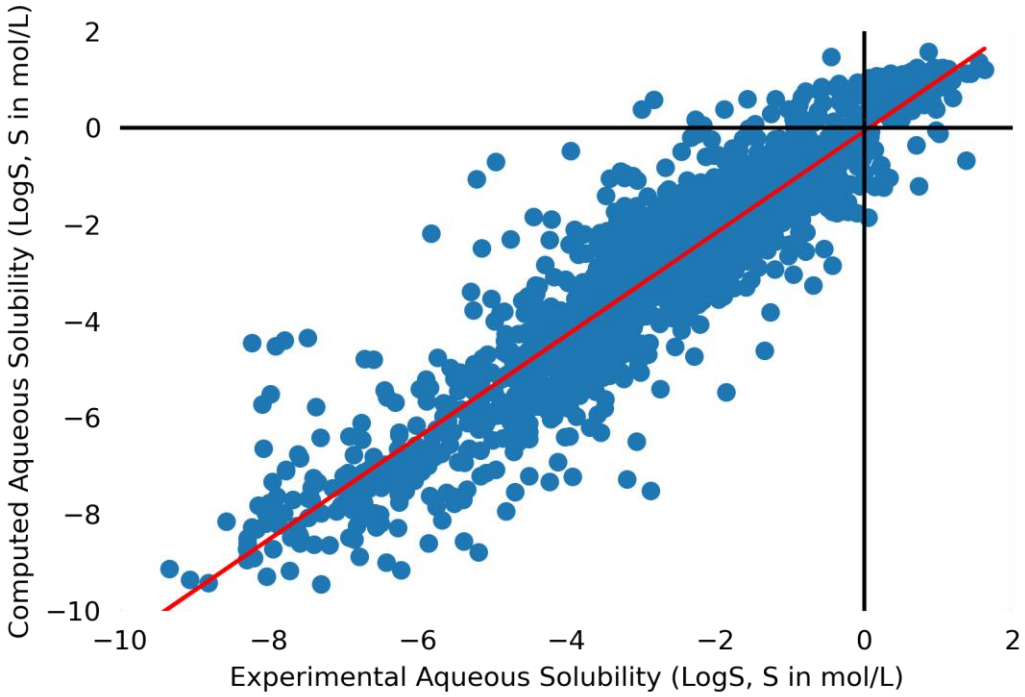

Figure 2. Comparison of aqueous solubility values predicted by SolTranNet to experimentally derived values.

Two other versions of SolTranNet were trained for 2000 training epochs and then used to predict the aqueous solubility of a withheld testing set. One version of SolTranNet included an additional node feature that determined how many of an atom's heavy neighbors were halogens. The halogen SolTranNet model had an RMSE of 0.970 and a coefficient of determination of 0.832. The other version of SolTranNet was a result of a modification to the existing identity feature node. This node identified what element the atom was in the molecule. One element that was missing from the identity list was silicon, which is present in the AqSol database. The silicon SolTranNet

model had an RMSE of 0.949 and a coefficient of determination of 0.838. A summary of the performances of the four different versions of the SolTranNet model can be found in Table 1.

Table 1. Performance of different versions of SolTranNet.

Model	Training		Testing	
	RMSE	R^2	RMSE	R^2
SolTranNet	0.953	0.829	0.927	0.848
Quantum SolTranNet	0.967	0.824	0.926	0.838
SolTranNet w/ Silicon Identity	0.908	0.851	0.949	0.838
SolTranNet w/ Halogen Node	0.919	0.849	0.970	0.832

3.2 Quantum Data Accuracy Analysis

The accuracy of the six different combinations of continuum solvation model and density functional was compared using RMSE, coefficient of determination, and mean absolute deviation. The solvation free energy of 291 compounds from the MNSOL database was predicted using each combination of solvation model and density functional. The computed solvation free energy was then compared to the experimentally determined solvation free energy provided in the MNSOL database. The SMD model performed the best when compared to the other two models, ALPB and

GBSA. The density functional did not seem to affect the performance of the solvation model based on the results shown in Table 2.

Table 2. RMSE and R2 values for different combinations of solvation model and density functional.

Solvation Model	*Density Functional*	*RMSE*	R^2	*MAD*
ALPB	GFN2	4.80	0.56	4.01
ALPB	M062X	4.66	0.66	5.04
GBSA	GFN2	3.48	0.51	3.29
GBSA	M062X	3.48	0.64	4.18
SMD	GFN2	1.71	0.84	3.08
SMD	M062X	1.68	0.86	3.31

3.3 Published Model Analysis

SolTranNet was compared to another recently published model, SolProp. Both models are able to predict a molecule's aqueous solubility, but SolProp is also able to predict a molecule's solvation free energy. The first test of SolProp was to predict the solvation free energy of the compounds from the MNSOL database. The results can be seen in Figure 4. SolProp has excellent correlation between the predicted and experimental value with a coefficient of determination of 0.987 and a great RMSE of 0.486.

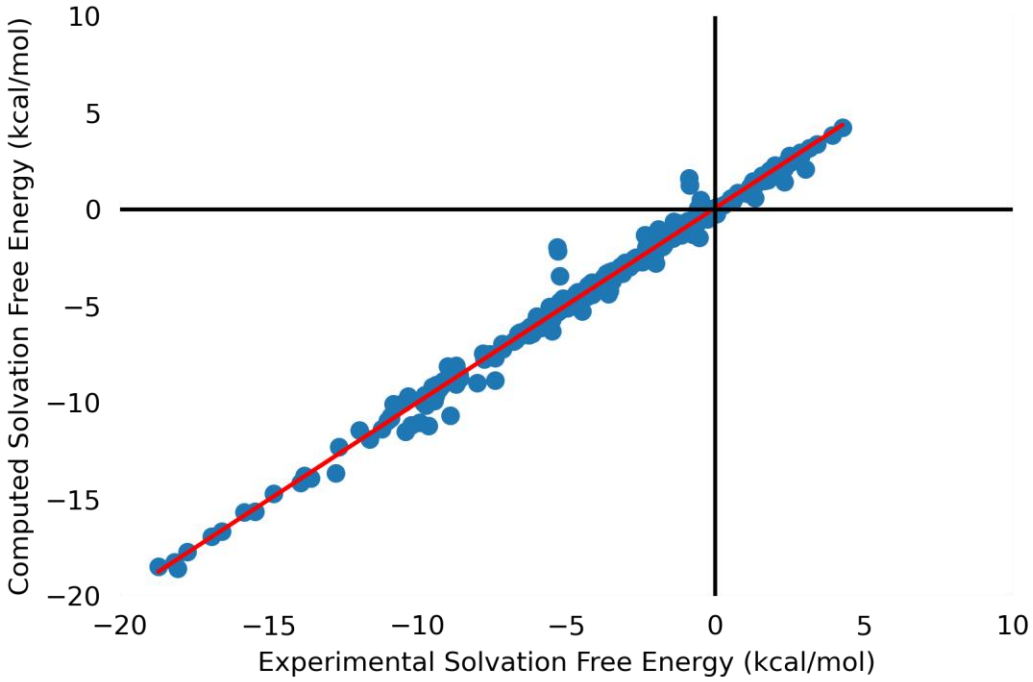

Figure 3. Comparison of solvation free energies predicted by the SolProp model and the experimentally determined solvation free energies.

As stated above, both models are able to predict a molecule's aqueous solubility. Therefore, both models were used to predict the aqueous solubility of the molecules in the AqSol database. SolProp is only able to predict aqueous solubility on neutral solutes, therefore the charged solutes were removed from the AqSol database. SolProp predicted the aqueous solubility at a rate of 618 ms per molecule and had a RMSE of 0.460 and coefficient of determination of 0.961. SolTranNet had a lower RMSE and coefficient of determination, 0.962 and 0.834, respectively, however it did predict at a faster rate of 5.62 ms per molecule. The strong correlation of SolProp and SolTranNet can be seen in Figures 5 and 6, respectively.

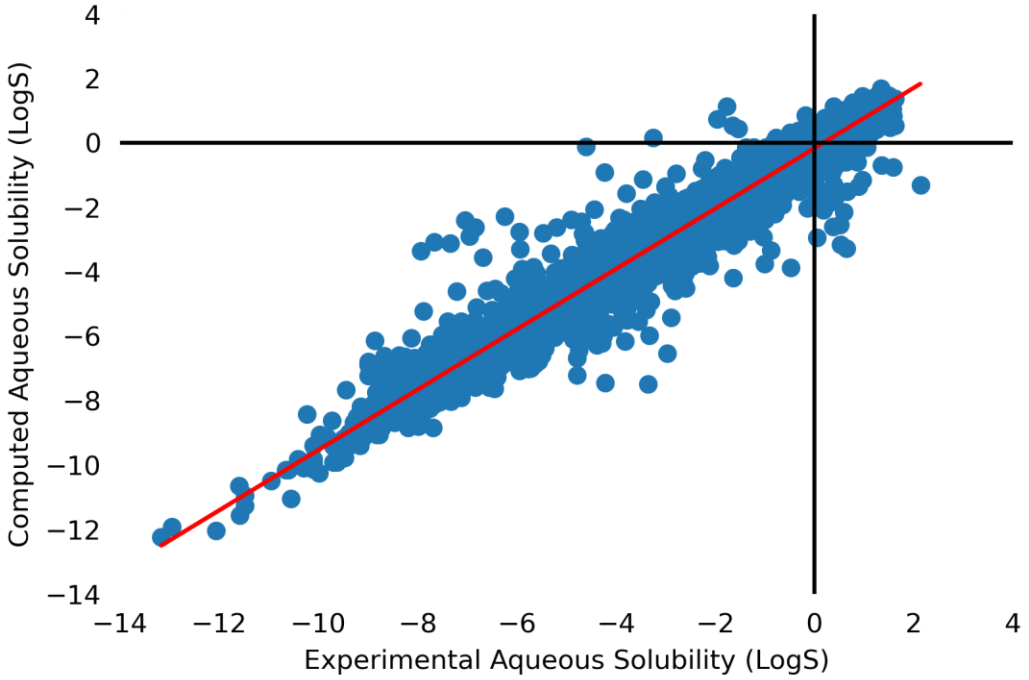

Figure 4. Comparison of aqueous solubility predicted by SolProp to experimentally derived aqueous solubility.

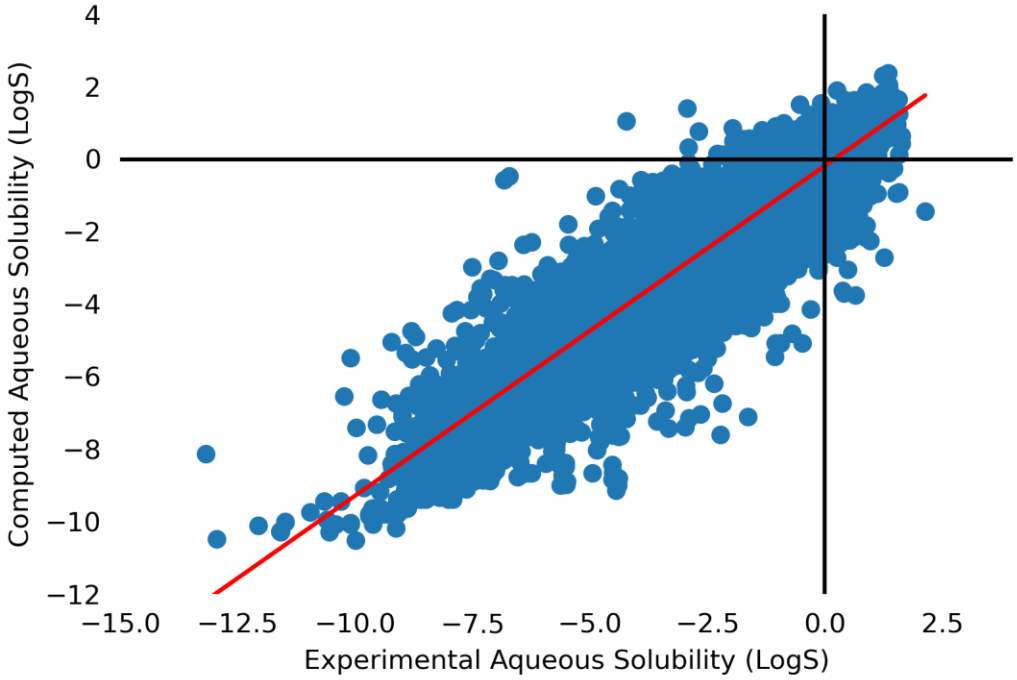

Figure 5. Comparison of aqueous solubility predicted by SolTranNet to experimentally derived aqueous solubility.

4.0 Discussion

The similar performance of the different SolTranNet based models shown in Table 1 seems to indicate that including the quantum chemical node features (dispersion coefficient, coordination number, polarizability, and partial charge) did not improve the performance of SolTranNet. As shown by the solvation model analysis and the analysis conducted by others[31], the likely reason for this is that the quantum data being generated by the solvation models is not accurate enough. Inaccurate data may prevent the algorithm from learning any trend in the data, thus leading to the algorithm to put less weight or even ignoring the feature in the vector.

Including silicon in the identity feature also did not result in significant improvement of the model. There are very few elements present in the dataset that are not included in the identity feature. Therefore, it is very possible that the neural network had been able to learn the identity of these nodes without being explicitly told the identity via the feature vector. Identifying the number of heavy atoms that were halogens also did not improve the performance of the model. As with adding silicon to the identity feature, this could be because we were not telling the neural network anything it did not learn after numerous training epochs.

Due to the lack of performance improvement when including quantum node features, we wanted to see if the quantum data that was being provided was of good quality. If the quantum data was not of good quality, then we should not expect to see an improvement in the model. To test this, we compared the solvation free energy computed by three different solvation models using geometries optimized by two different functionals to the experimentally determined values. The results of this analysis can be seen in Table 2. The hypothesis for this experiment is that if the solvation models are unable to accurately compute the solvation free energies, then they more than

likely are unable to accurately predict other quantum values, such as the ones added to the node features. The results in Table 2 agree with previously reported results in the sense that the SMD solvation model performed the best but is more than likely still not accurate enough to provide quantum data to improve the performance of SolTranNet.

During this project, a model pretrained on quantum data called SolProp was published. SolProp is comparable to the model that we intended to develop and for that reason we wanted to compare it to SolTranNet. SolProp is able to predict not only aqueous solubility but also solvation free energy. Therefore, we first predicted the solvation free energy for the compounds in MNSOL. SolProp performed very well when compared to the experimental solvation free energy with excellent correlation.

Then we did a comparison of the two published models, SolTranNet and SolProp_ML. Both models were trained on either part of or all of the AqSol database. SolProp is only able to predict aqueous solubility on neutral solutes and for this reason, charged solutes were removed from AqSol. Both models performed well, as shown in figures 3 and 4. SolTranNet ran at a speed 100x that of the speed of SolProp, operating at a speed of 5.92 ms per molecule and SolProp operating at a speed of 618 ms per molecule. It was expected that SolTranNet would have a faster run-time performance since it was designed to be a quick tool to predict aqueous solubility but the overall speed increase was unexpected.

5.0 Conclusions

Based on the results comparing SolTranNet to quantum SolTranNet, it seems like the hypothesis that inclusion of quantum data will enable the model to perform better should be rejected. However, looking at the results of the analysis of the solvation models, it seems likely that the quantum data that was used for the quantum node features was not very accurate. Coupling this analysis with the results provided by the comparison between SolTranNet and SolProp, it seems like including quantum data in an aqueous solubility model should increase performance as long as the quantum data is sufficiently accurate.

While SolProp seems to do very well predicting the aqueous solubility of neutral solutes, it does not allow predictions for charged solutes. Salts pose a unique problem to graph neural networks because they and their corresponding compound have high similarity between descriptors but greatly vary in solubility. This is not surprising since salinization is often used to increase the solubility of a compound in drug design. Therefore, since salts are of interest in fields such as drug design, further research should be done to determine more node features to be included in GNN and other neural network models to increase model performance for this class of compound.

Future iterations of SolTranNet may also want to implement graph level descriptors as well. It is at the graph level that the model may be modified to enable it to identify whether the compound is a salt or not. All of the current descriptors are atomic level descriptors; however, it could be useful to look at the molecule as a whole or look for functional groups that may not be identified by the current feature vector. Including MACCS keys fingerprints for example may improve the performance of the model. Implementing graph level descriptors would also enable

the model to utilize thermodynamic properties of molecules such as solvation free energy and solvation enthalpy, which are both used by SolProp.

Bibliography

1.Reymond, J.-L., The Chemical Space Project. *Accounts of Chemical Research* **2015,** *48* (3), 722-730.

2.Kaló, Z.; Petykó, Z. I.; Fricke, F.-U.; Maniadakis, N.; Tesař, T.; Podrazilová, K.; Espin, J.; Inotai, A., Development of a core evaluation framework of value-added medicines: report 2 on pharmaceutical policy perspectives. *Cost Effectiveness and Resource Allocation* **2021,** *19* (1), 42.

3.Hingorani, A. D.; Kuan, V.; Finan, C.; Kruger, F. A.; Gaulton, A.; Chopade, S.; Sofat, R.; MacAllister, R. J.; Overington, J. P.; Hemingway, H.; Denaxas, S.; Prieto, D.; Casas, J. P., Improving the odds of drug development success through human genomics: modelling study. *Scientific Reports* **2019,** *9* (1), 18911.

4.Wouters, O. J.; McKee, M.; Luyten, J., Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *Jama* **2020,** *323* (9), 844-853.

5.Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H., Estimation of Aqueous Solubility By The General Solubility Equation (GSE) The Easy Way. *QSAR & Combinatorial Science* **2003,** *22* (2), 258-262.

6.Ran, Y.; Yalkowsky, S. H., Prediction of Drug Solubility by the General Solubility Equation (GSE). *Journal of Chemical Information and Computer Sciences* **2001,** *41* (2), 354-357.

7.Llinàs, A.; Glen, R. C.; Goodman, J. M., Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *Journal of Chemical Information and Modeling* **2008,** *48* (7), 1289-1303.

8.Deng, C.; Liang, L.; Xing, G.; Hua, Y.; Lu, T.; Zhang, Y.; Chen, Y.; Liu, H., Multi-channel GCN ensembled machine learning model for molecular aqueous solubility prediction on a clean dataset. *Molecular Diversity* **2022**.

9.Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M., Findings of the Challenge To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **2009,** *49* (1), 1-5.

10.Llinas, A.; Avdeef, A., Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets. *Journal of Chemical Information and Modeling* **2019,** *59* (6), 3036-3040.

11.Llinas, A.; Oprisiu, I.; Avdeef, A., Findings of the Second Challenge to Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **2020,** *60* (10), 4791-4803.

12.Francoeur, P. G.; Koes, D. R., SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction. *Journal of Chemical Information and Modeling* **2021,** *61* (6), 2530-2536.

13.Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S., Molecule attention transformer. *arXiv preprint arXiv:2002.08264* **2020**.

14.Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M., Graph neural networks: A review of methods and applications. *AI Open* **2020,** *1*, 57-81.

15.Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D., Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation* **2017,** *13* (3), 1034-1043.

16.Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric

Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009,** *113* (18), 6378-6396.
17.Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S., Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. *J Chem Theory Comput* **2021,** *17* (7), 4250-4261.
18.Onufriev, A. V.; Case, D. A., Generalized Born Implicit Solvent Models for Biomolecules. *Annual review of biophysics* **2019,** *48*, 275-296.
19.Sorkun, M. C.; Khetan, A.; Er, S., AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data* **2019,** *6* (1), 143.
20.Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R., Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling* **2020,** *60* (9), 4200-4215.
21.(RRID:SCR_014274), R. O.-S. C. S.
22.Tosco, P.; Stiefl, N.; Landrum, G., Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of Cheminformatics* **2014,** *6* (1), 37.
23.Riniker, S.; Landrum, G. A., Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015,** *55* (12), 2562-2574.
24.Pracht, P.; Bohle, F.; Grimme, S., Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020,** *22* (14), 7169-7192.
25.Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019,** *15* (3), 1652-1671.
26.Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **2021,** *11* (2), e1493.
27.Neese, F., The ORCA program system. *WIREs Computational Molecular Science* **2012,** *2* (1), 73-78.
28.Marenich, A. V. K., Casey P; Thompson, Jason D; Hawkins, Gregory D; Chambers, Candee C; Giesen, David J; Winget, Paul; Cramer, Christopher J; Truhlar, Donals G., Minnesota Solvation Database (MNSOL) version 2012. Retrieved from the Data Repository for the University of Minnesota: 2020.
29.Vermeire, F. H.; Chung, Y.; Green, W. H., Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *Journal of the American Chemical Society* **2022,** *144* (24), 10785-10797.
30.Vermeire, F. H.; Green, W. H., Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021,** *418*, 129307.
31.Patel, C.; Roy, D., Octanol–Water Partition Coefficients of Fluorinated Drug Molecules with Continuum Solvation Models. *The Journal of Physical Chemistry A* **2022,** *126* (26), 4185-4190.