DeepGamma: A deep learning model for activity coefficient prediction

Kobi C. Felton, Hashem Ben-Safar, Alexei A. Lapkin [*]

Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Road, Cambridge, UK

Abstract

Vapor liquid equilibrium is a ubiquitous aspect of designing industrial chemical processes that make products ranging from pharmaceuticals to petrochemicals. To predict vapor liquid equilibrium for a wide variety of industrially relevant mixtures, activity coefficients are often used. However, calculating activity coefficients experimentally is time and labor-intensive, and existing methods for predicting activity coefficients are limited in scope or computationally expensive. Herein, we introduce DeepGamma, a deep learning method for predicting activity coefficients of binary mixtures directly from the molecular structures of their components. DeepGamma is demonstrated to have strong performance on a variety of mixtures with extremely fast prediction times.

Introduction

Vapor-liquid equilibrium thermodynamics play an essential role in a wide variety of fields in the basic and applied sciences. For example, chemical reactions often happen in between the vapor and liquid phase, where vapor produced by a reaction is extracted to drive conversion or a reactant is introduced as a vapor. Similarly, large scale chemical processes require separation of mixtures into components to purify a valuable product. One ubiquitous method for achieving such separations is distillation, which relies on differences in boiling points of mixtures to achieve separation.

In order to design and engineer systems which contain vapor-liquid equilibrium, thermodynamic equations are utilized. Thermodynamic equations describe the relationship between the composition of the liquid and the vapor at a given temperature and pressure. One well-known thermodynamic equation is Raoult's Law:

$$y_i P = x_i P_i^{sat}(T) \tag{1}$$

where y_i and x_i are the vapor and liquid compositions of component i of the mixture respectively, P is the absolute pressure, and $P_i^{sat}(T)$ is the vapor pressure at temperature T. Raoult's law describes non-interacting ideal systems, yet it fails to properly predict vapor-liquid equilibrium the wide variety of mixtures used in industrial chemical processes. Therefore, the simplified gamma-phi equation was developed to describe deviations from ideality:

$$y_i P = x_i \gamma_i(x, T, P) \exp\left(\frac{V_i^L (P - P_i^{sat}(T))}{RT}\right) \tag{2}$$

where $\gamma_i(\mathbf{x}, T, P)$ is the activity coefficient at liquid composition $\mathbf{x}$, temperature T, and pressure P; V_i^L is the specific volume and P_i^{sat} is the vapor pressure.

Activity coefficients are the unknown parameters in the gamma-phi thermodynamic equation that correct for deviations from ideality. These activity coefficients must be predicted for each new mixture using thermodynamic models such as the non-random two liquid model (Renon and Prausnitz 1968).

Unfortunately, the vapor-liquid equilibrium experiments required to parameterize thermodynamic models for predicting activity coefficients are notoriously time-intensive to complete. Classical thermodynamic measurement apparatus such as stills require equilibrium to be reached at each temperature and pressure combination before a measurement is taken (Ronc and Ratcliff 1976; Dechambre et al. 2014). This equilibration process can take anywhere from minutes to hours, meaning a full set of experiments for a new mixture can take weeks to months. As a result, vapor liquid equilibrium experiments are executed very selectively.

To reduce the time required for parameterizing thermodynamic models for predicting activity coefficients, there is a large body of research focused on directly predicting the parameters from molecular structures. This work ranges from simple counting of functional groups (Fredenslund, Jones, and Prausnitz 1975) to density functional theory (Klamt, Eckert, and Arlt 2010) and machine learning (Urata et al. 2002; Nami and Deyhimi 2011; Jirasek et al. 2020). However, these prediction approaches often only cover a small subset of all mixture types (e.g., only hydrocarbons) or require significant calculation times. Thus, there is a need for a general, fast and accurate method for predicting activity coefficients *a priori*.

Herein, we introduce *DeepGamma*, a deep learning model for fast predictions of activity coefficients directly from molecular structures. We leverage message passing neural networks (MPNN) to reproduce the results of quantum simulations for predicting activity coefficients without compu-

[*] Email: aal35@cam.ac.uk

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tational expense. Specifically, our model decreases the time required for calculations by a factor of 1900. Our work focuses on predicting activity coefficients of binary mixtures as a first proof of concept.

Models

DeepGamma, a message passing neural network: Molecules can be treated as graphs with atoms as nodes and bonds as edges. Therefore, message passing neural networks (MPNN) that operate on graphs can be used for end-to-end prediction of molecular properties (Gilmer et al. 2017).

One type on MPNN is a directed message passing neural network (D-MPNN) in which the encoder acts on edges (bonds) instead of nodes (atoms) to improve stability of training (Yang et al. 2019). Yang, Swanson and colleagues showed D-MPNNs have superior performance to other tools such as gradient boosted trees for property prediction tasks (2019). Formally, a molecule in a D-MPNN is considered to be a graph G with edges e_{vw} and nodes v and w with atom features x_v. A message passing update m_v^t is as follows:

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{3}$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{4}$$

where M_t is the message function, U_t is the update function and h_v^t is the hidden state at step t. To obtain predictions $\hat{y}$, the outputs of the last message passing step T are passed through a feed forward network R in a readout phase:

$$\hat{y} = R(h_v^T \in G) \tag{5}$$

In addition to using outputs of the message passing steps as input to the feed forward network, additional features f can also be added:

$$\hat{y} = R(h_v^T \in G, f) \tag{6}$$

In our case of predicting activity coefficients of binary mixtures at atmospheric pressure, we treat the temperature and composition as additional features. Therefore, the feed forward network can be written as:

$$\ln \gamma(x, T) = R(h_v^T \in G, x, T) \tag{7}$$

Note that we predict the natural logarithm of the activity coefficient since these values can vary over an order of magnitude.

DeepGamma Polynomial: Often, adding chemical knowledge into machine learning models can make them more accurate. In the case of activity coefficient prediction, it would be best to fit a thermodynamically consistent model such as the the non-random two liquid model (NRTL) (Renon and Prausnitz 1968). However, we found that many of the mixtures in our training set (see ”Datasets”) were difficult to fit using NRTL, often because the large correlations between its parameters (Höller et al. 2019). We found empirically that the activity coefficient curves fit well to a fourth order polynomial model. While this polynomial form does not necessarily ensure Gibbs-Duheim thermodynamic consistency[1], it should work well for the sole purpose of vapor liquid equilibrium activity coefficients (not liquid-liquid equilibrium for example). Therefore, we fitted fourth order polynomials to the activity coefficient data and used the D-MPNN to predict the coefficients of the polynomial model:

$$\ln \gamma_i(x_i, T) = \sum_{j=0}^{4} c_{ij}(T) x^j \tag{8}$$

$$\mathbf{c_{ij}} = R(h_v^T \in G, \mathbf{x}, T) \tag{9}$$

In our results we compare direct prediction of activity coefficients and predicting polynomial coefficients.

Datasets

Previous work has demonstrated the power of transfer learning for improving predictions of D-MPNNs (Vermeire and Green 2021). We aim to obtain similar results for activity coefficient prediction, relying on two datasets.

Combisolv Solvation Energy Dataset: Solvation energies describe the change in free energy when a gas molecule of a solute is placed into a solvent. Solvation energies are closely related to activity coefficients at infinite dilution (Moine et al. 2017), so encoder representations learned on this task could be useful for the downstream task of activity coefficient prediction. We utilize the Combisolv dataset, which contains the largest number of solvation energies publicly available to date: one million binary pairs of molecules calculated using COSMO-RS (Vermeire and Green 2021).

COSMO-RS Activity Coefficient Dataset: We executed DFT calculations for over 18 million activity coefficients by taking all the binary pairs of 460 common solvent molecules from a previously published dataset (Amar et al. 2019). The activity coefficients were calculated in 5K temperature increments between the normal boiling points of the each binary pair, and 0.1 mol/mol composition grid was used. All activity coefficients were calculated at atmospheric pressure. COSMOtherm 2020 at the TZVPD fine fidelity level was utilized, and parameters were taken the 2020 COSMObase database (Klamt, Eckert, and Arlt 2010). These calculations took 41 days on a 24 core machine.

Training

Holdout data: One of the important aspects of evaluating the applicability of machine learning models for property prediction is holding out data from training to evaluate the model fairly. Random splits can artificially inflate model accuracy scores because similar or identical molecules can be placed in the train and holdout sets (Kovács, McCorkindale, and Lee 2021). Therefore, we use the Butina clustering algorithm (Butina 1999) to group similar molecules and allocate

[1] The Gibbs-Duheim equation relates changes in chemical potential to changes in temperature and pressure. $\sum_{i=1}^{I} N_i d\mu_i$ = -SdT + VdP. At equilibrium, the right hand side becomes zero, so traditional thermodynamic models like NRTL obey the equation $\sum_{i=1}^{I} N_i d\mu_i = 0$

10% of clusters for holdout sets (5% of the clusters for validation and 5% for test set). Since we are working with binary mixtures, we create three types of holdout sets to represent common use cases:

- **MIX**: One of the molecules in a mixture is in the train and the other is in the holdout set.
- **INDP**: Both of the molecules in a mixture are not in the train set (i.e., only in the holdout set).
- **CONT**: A random split on molecules in the train clusters. This represents when there are some measurements of activity coefficients of both of the molecules in a mixture in the training set but not at the temperature or composition being queried in the holdout set.

Training Details: We leveraged the implementation of D-MPNNs in the python package chemprop (Yang et al. 2019). A model was trained on the Combisolv dataset using the same hyperparameters as in the paper by Vermeire and Green (batch size 50, encoder hidden size 200, and feed forward network hidden size of 500) except that a shared encoder was used instead of one for the solute and solvent. Using a shared encoder did not have a significant impact on the results. Since the paper did not report learning rate, a small set of experiments were conducted to find that an initial learning rate of 1e-4 and max learning rate of 2e-4 were optimal for the Noam learning rate scheduler (Vaswani et al. 2017). Training on a single Tesla T4 GPU (AWS EC2 g4dn.xlarge) required 43 hours. For the COSMO-RS models, we used a feedforward network hidden size of 380, a batch size of 4550, and a max learning rate of 6e-3. The COSMO-RS models were trained for 20 epcohs using instances equipped with a single Tesla V100 GPU (16 GB) and 61 GB of RAM (the COSMO-RS models required 30+ GB of space in memory). For transfer learning on the COSMO-RS dataset, we froze the encoder learned for the Combisolv model and only adjusted the parameters of the feed forward network. The polynomials were fit with LMFit (Newville et al. 2014) parallelized on a 24 core machine using Ray (Moritz et al. 2018).

Results

DeepGamma is rapidly trained to reproduce results of COSMO-RS for activity coefficient prediction. It takes 2.5 days to train our best performing model, and predictions on more than 1M activity coefficients takes less than 30 minutes on a modern GPU. This is significant because generating the original data required over one month, representing a 1900x speed up.

Results of training on the COSMO-RS activity coefficient dataset are shown in Table 1, the base DeepGamma model achieves the lowest mean absolute error on all holdout sets. Similar mean absolute errors are achieved on the validation and test datasets for all models. As expected, the Valid CONT holdout set has the lowest MAE due to the same molecules being in the training set and holdout set. For the best performing base model, the INDP and MIX datasets have similar error profiles, indication that model has learned a good encoder representation. Transfer learning models have slightly worse performance than the base models, but the transfer learning models take 50% less time to train. Part of the reason for poorer performance is that the Combisolv model was trained with a depth of four (i.e., number of message passing steps), while the COSMO-RS model only utilized three. Since each message passing step can cause changes in the atom representation, this difference in depth could effect quality of the final fingerprint formed. Interestingly, the DeepGamma Polynomial models have the worst performance of all models. Furthermore, there are significant differences in error between the different activity coefficients predicted by the DeepGamma Polynomial. It is possible that further hyperparameter tuning could improve the accuracy of the model, particularly through changes in the encoder and feedforward network hidden size (Yang et al. 2019; Vermeire and Green 2021).

Related Work

We classify existing methods for predicting activity coefficients into three categories: group-contribution methods, quantum chemistry based methods and machine learning methods. Group contribution methods such as UNIFAC were originally developed in the mid 1970s and predict the activity coefficient as a weighted sum of the occurrence of a predetermined set of chemical functional groups (Fredenslund, Jones, and Prausnitz 1975). While UNIFAC provides fast predictions, the accuracy of the model is limited by the interactions explicitly accounted for in each basis function. It can be challenging to hand-craft all such interactions as they can often extend between several atoms.

The second set of methods are quantum chemistry methods. These simulations often use a combination of molecular dynamics and density functional theory (DFT) calculations to predict activity coefficients for each mixture component (Constantinescu, Klamt, and Geană 2005). COSMO-RS is one of the most reliable computational methods for liquid-phase thermodynamic predictions (Klamt 1995; Klamt, Eckert, and Arlt 2010). It relies on the theory of screening charges, which states that every element of a surface of a dissolved solute must be complemented by an opposite charge in the solvent. The charge surface around a solute is divided into infinitesimally small pieces, which are then integrated to find the charge density. This charge density can then be used to calculate activity coefficients. The charge density is calculated via rigorous DFT calculations or a much faster quantitative structure-property relationship based on a database of over 65,000 pre-calculated compounds. COSMO-RS and a related method named COSMO-SAC have been applied in calculations of VLE (Constantinescu, Klamt, and Geană 2005), liquid-liquid equilibrium (LLE) (Dechambre et al. 2014) and vapor-liquid-liquid-equilibrium (VLLE) curves (Kundu and Banerjee 2011).

The downside of COSMO methods is that they are often not accurate for polar compounds (Constantinescu, Klamt, and Geană 2005; Kundu and Banerjee 2011). This is often due to the lack of theory for the hydrogen bonding present in these systems (Kundu and Banerjee 2011). Another challenging aspect of COSMO-RS is the computational intensity of DFT calculations for new mixtures, though this can be

Table 1: Validation and test mean absolute error of DeepGamma (DG) models on the COSMO-RS Activity Coefficient Dataset. Best results are bolded. TLCB stands for transfer learning, where the encoder from the model trained on the Combisolv dataset is frozen and only the feedforward network is tuned. DGP stands for models which predict polynomial coefficients instead of activity coefficients directly. Time is training time in hours.

		Valid CONT		Valid INDP		Valid MIX		Test MIX		Test INDP	
	Training (h)	$\ln\gamma_1$	$\ln\gamma_2$	$\ln\gamma_1$	$\ln\gamma_2$	$\ln\gamma_1$	$\ln\gamma_2$	$\ln\gamma_1$	$\ln\gamma_2$	$\ln\gamma_1$	$\ln\gamma_2$
DG	62	**0.02**	**0.02**	**0.07**	**0.07**	**0.07**	**0.07**	**0.06**	**0.06**	**0.06**	**0.05**
DG-TLCB	36	0.04	0.04	0.09	0.09	0.08	0.08	0.07	0.07	0.07	0.07
DGP	6.2	0.10	0.29	0.16	0.36	0.14	0.33	0.14	0.28	0.13	0.28
DGP-TLCB	3.5	0.10	0.29	0.16	0.36	0.14	0.33	0.14	0.28	0.13	0.28

somewhat alleviated by a less accurate method for predictions of charge surfaces from molecular structures (Loschen and Klamt 2012). Our method is distinct because it directly predicts activity coefficients instead of charge surfaces.

Previous machine learning work has primarily focused on predicting activity coefficients at infinite dilution. A large proportion of studies combined simple descriptors of molecules as input features with artificial neural networks (Urata et al. 2002; Ramírez-Beltrán et al. 2009; Nami and Deyhimi 2011; Behrooz and Boozarjomehry 2017). For example, Nami et al. used neural networks to predict activity coefficients at infinite dilution of organic solutes in ionic liquids (2011). Their work leveraged hand-crafted descriptors of molecules and achieved a RMSE of 0.128 on a limited set of compounds. Other work has demonstrated that matrix completion can be used for activity coefficient prediction (Jirasek et al. 2020). The aforementioned studies focused on predicting activity coefficients at infinite dilution or with a limited number of molecules. Our work is the first to consider a wide range of mixtures at different compositions and temperatures. Furthermore, we predict activity coefficients directly from molecular structures.

Conclusion

Herein, we develop *DeepGamma*, an approach to predicting activity coefficients directly from molecular structures using directed messasge passing neural networks. Our approach offers an over 1900x speed-up compared to the original quantum simulations. Furthermore, our model is accurate at predicting activity coefficients of unseen molecules.

The main weakness of our approach is that it relies solely on quantum simulation data. As a next step, we plan to leverage data from experiments to improve the accuracy of the model using transfer learning and complete further benchmarking on experimental data. Additionally, we plan to utilize uncertainty quantification to inform practicioners of the quality of model predictions (Soleimany et al. 2021).

References

Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; and Lapkin, A. 2019. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science*, 10(27): 6697–6706.

Behrooz, H. A.; and Boozarjomehry, R. B. 2017. Prediction of limiting activity coefficients for binary vapor-liquid equilibrium using neural networks. *Fluid Phase Equilibria*, 433: 174–183.

Butina, D. 1999. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Science*, 39(4): 747–750.

Constantinescu, D.; Klamt, A.; and Geană, D. 2005. Vapor-liquid equilibrium prediction at high pressures using activity coefficients at infinite dilution from COSMO-type methods. *Fluid Phase Equilibria*, 231(2): 231–238.

Dechambre, D.; Pauls, C.; Greiner, L.; Leonhard, K.; and Bardow, A. 2014. Towards automated characterisation of liquid-liquid equilibria. *Fluid Phase Equilibria*, 362: 328–334.

Fredenslund, A.; Jones, R. L.; and Prausnitz, J. M. 1975. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AiChE Journal*, 21(6): 1086–1099.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1263–1272. JMLR.org.

Höller, J.; Bickert, P.; Schwartz, P.; von Kurnatowski, M.; Kerber, J.; Künzle, N.; Lorenz, H.-M.; Asprion, N.; Blagov, S.; and Bortz, M. 2019. Parameter Estimation Strategies in Thermodynamics. *chemengineering*, 3(2): 56.

Jirasek, F.; Alves, R. A. S.; Damay, J.; Vandermeulen, R. A.; Bamler, R.; Bortz, M.; Mandt, S.; Kloft, M.; and Hasse, H. 2020. Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion. *Physical Chemistry Letters*, 11(3): 981–985.

Klamt, A. 1995. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry*, 99(7): 2224–2235.

Klamt, A.; Eckert, F.; and Arlt, W. 2010. COSMO-RS: An Alternative to Simulation for Calculating Thermodynamic Properties of Liquid Mixtures. *Annual Review of Chemical and Biomolecular Engineering*, 1(1): 101–122.

Kovács, D. P.; McCorkindale, W.; and Lee, A. A. 2021. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. 12(1).

Kundu, D.; and Banerjee, T. 2011. Multicomponent vapor-liquid-liquid equilibrium prediction using an a priori segment based model. *Industrial and Engineering Chemistry Research*, 50(24): 14090–14096.

Loschen, C.; and Klamt, A. 2012. COSMO quick: A novel interface for fast σ-profile composition and its application to COSMO-RS solvent screening using multiple reference solvents. *Industrial and Engineering Chemistry Research*, 51(43): 14303–14308.

Moine, E.; Privat, R.; Sirjean, B.; and Jaubert, J.-N. 2017. Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes. 46(3): 033102.

Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul, W.; Jordan, M. I.; and Stoica, I. 2018. Ray: A Distributed Framework for Emerging AI Applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 561–577. Carlsbad, CA: USENIX Association. ISBN 978-1-939133-08-3.

Nami, F.; and Deyhimi, F. 2011. Prediction of activity coefficients at infinite dilution for organic solutes in ionic liquids by artificial neural network. 43(1): 22–27.

Newville, M.; Stensitzki, T.; Allen, D. B.; and Ingargiola, A. 2014. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python.

Ramírez-Beltrán, N. D.; Vallés, H. R.; Estévez, L. A.; and Duarte, H. 2009. A neural network approach to predict activity coefficients. 87(5): 748–760.

Renon, H.; and Prausnitz, J. M. 1968. Local compositions in thermodynamic excess functions for liquid mixtures. 14(1): 135–144.

Ronc, M.; and Ratcliff, G. R. 1976. Measurement of vapor-liquid equilibria using a semi-continuous total pressure static equilibrium still. *The Canadian Journal of Chemical Engineering*, 54(4): 326–332.

Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; and Coley, C. W. 2021. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Central Science*, 7(8): 1356–1367.

Urata, S.; Takada, A.; Murata, J.; Hiaki, T.; and Sekiya, A. 2002. Prediction of vapor–liquid equilibrium for binary systems containing HFEs by using artificial neural network. 199(1-2): 63–78.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. 30.

Vermeire, F. H.; and Green, W. H. 2021. Transfer learning for solvation free energies: From quantum chemistry to experiments. 418: 129307.

Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; and Barzilay, R. 2019. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.*, 59(8): 3370–3388.