# Addressing uncertainty in sub-cooled liquid property estimation

Trevor N. Brown[1,2], James M. Armitage[2,3], Jon A. Arnot[2,3]

1 trevor.n.brown@gmail.com, TNB Research, Halifax, Canada
2 ARC Arnot Research & Consulting, Inc., Toronto Canada
3 University of Toronto Scarborough, Toronto, Canada

## Introduction

Chemical partitioning properties, e.g. solubilities and equilibrium partition coefficients, are important parameters for multimedia modelling and other risk assessment applications.

For chemicals that are solids at system temperature, the appropriate reference state for internally consistent chemical partitioning properties are sub-cooled liquid properties.

Partitioning properties of the solid have to be adjusted to sub-cooled liquid property values using the Fugacity Ratio (F). For example, solubility in water: $S_{W(L)} = S_{W(S)} / F$

The rigorous equation for calculating F has terms that include the entropy of fusion at the melting point ($\Delta S_m$) and heat capacity. Two different assumptions about the heat capacity can be made resulting in the approximations below.

If Walden's Rule ($\Delta S_m$ = 56.5 kJ/mol) is assumed to be valid, F can be estimated using only the melting point ($T_m$).[1]

Another important application of the Fugacity Ratio is in the calculation of chemical activity, which has uses in chemical risk assessment. **For example see Posters WP132 and WP138.**

### Van't Hoff Approximation

$$\log F = -\frac{\Delta S_m}{2.303 \cdot R} \cdot \frac{(T_m - T)}{T}$$

### Hildebrand Approximation

$$\log F = -\frac{\Delta S_m}{2.303 \cdot R} \cdot \ln\left(\frac{T_m}{T}\right)$$

## QSARs - Methods, Results and Discussion

**$T_m$ - Melting Point QSARs**

A dataset of 2884 chemicals from the currated Bradley $T_m$ database[2,3] was split into the primary training and validation datasets. The much larger Enamine database[2,4] was used as an additional external validation dataset.

A fragment-based QSAR was constructed as described in the literature[5,6,7]. Two thirds of the Bradley dataset (1922 chemicals) were used as a training dataset, and the remaining 962 chemicals were used as an external validation dataset. For comparison the USEPA program EPIuite was also used to predict $T_m$ for the external validation dataset. These results are shown in Figure 1.

The much larger Enamine dataset was used to assess the models performance in the context of high throughput screening. The results are shown in Figure 2.

For the validation chemicals from Bradley dataset the IFS predictions are superior to the EPISuite predictions. For the much more diverse Enamine dataset both QSARs give fairly poor results, though IFS has a better correlation.

**$\Delta S_m$ - Entropy of Fusion and Melting Point QSARs**

Jain *et al* 2004[1] transcripted a database of experimental $\Delta S_m$ values, 1585 of these were extracted with structural information in the form of SMILES.

A fragment-based QSAR for $\Delta S_m$ was constructed following the same procedures as for $T_m$. 1056 and 529 chemicals were assigned to the training and validation datasets respectively.

Figure 3 shows the comparison of the new IFS predictions, compared with the predictions summarized in the supporting information of Jain *et al* 2004.

The predictive power of the two QSARs superficially appears to be comparable. However, Jain *et al* 2004 do not discuss training and validation datasets, so it is likely that the model results shown in Figure 3A were fitted to the data, whereas Figure 3B shows the results of an external validation.

The simple QSAR presented by Jain *et al* 2004 was reimplemented as a fragment based model, but neglecting molecular symmetry effects. This implementation gave results slightly better than those presented in the original publication.

## Conclusions and Future Work

The poor performance of both QSARs for the Enamine $T_m$ dataset is because the database contains many chemicals that are outside of the chemical domain of their training datasets.

The Bradley dataset contains is considered high quality, but this usually also means well-studied and therefore the chemical domain covered by these data are limited.

Future work will focus on integrating high quality and lower quality data in a weighted regression model. This will use all available data and expand the model domain of applicability.

Additional future work will attempt to combine important physicochemical properties such as F, $T_m$, $\Delta S_m$ and possibly $S_W$ within a single integrated QSAR.

Finally, future work will also include case study applications and comparisons of methods for estimating chemical activity and for developing internally consistent QSARs for predicting chemical solubilities and partition coefficients.

Figure 1: Bradley $T_m$ External Validation Dataset.
A: EPISuite Predictions; B: IFS Predictions.

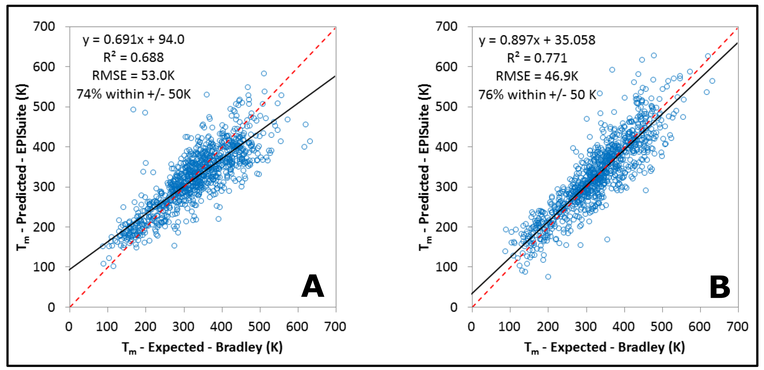



Figure 2: Enamine $T_m$ Dataset.
A: EPISuite Predictions B: IFS Predictions

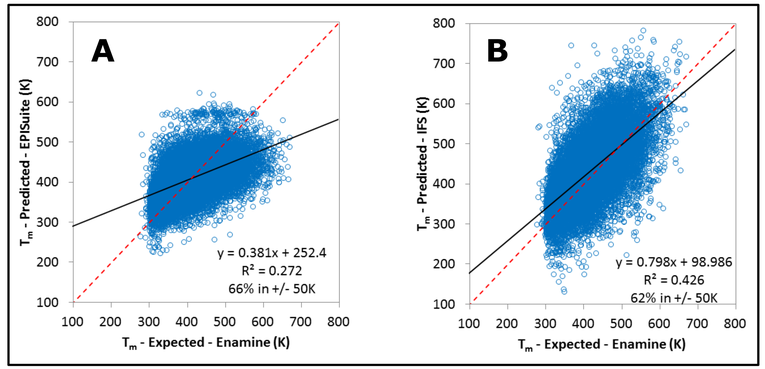
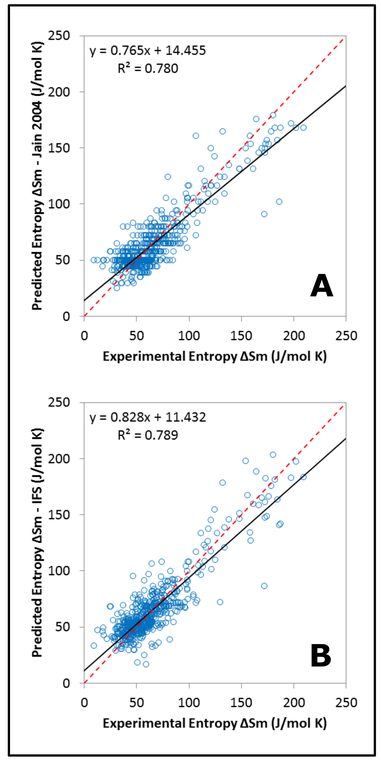



Figure 3: Jain *et al* 2004 $\Delta S_m$ Dataset
A: Jain *et al* 2004 Predictions;
B: IFS Predictions.




## Fugacity Ratio

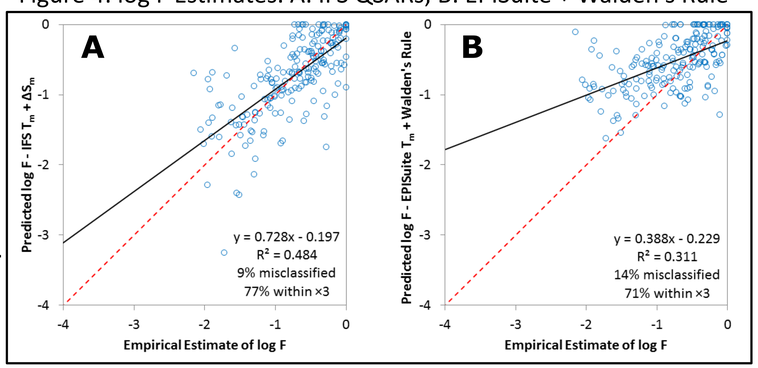
361 chemicals from the $\Delta S_m$ validation dataset also had empirical $T_m$ values which were used together in the Van't Hoff approximation for an empirical estimate of F.

199 of these chemicals were solids or predicted to be solids, and F values were predicted by one of two methods: 1) Using the $T_m$ and $\Delta S_m$ IFS QSARs 2) Using EPISuite $T_m$ and Walden's Rule.

Both QSARs misclassified some solids as liquids or liquids as solids. 77% and 71% of predictions were within a factor 3 of the empirical F estimate. Maxiumum error was a factor of 34 for IFS and a factor of 111 for EPISuite + Walden's Rule.

The IFS QSARs show less bias in the predictions, with fewer misclassifications and fewer poor predictions.

Figure 4: log F Estimates. A: IFS QSARs; B: EPISuite + Walden's Rule




## References

[1] Jain, A., G. Yang, and S.H. Yalkowsky, Estimation of Total Entropy of Melting of Organic Compounds. Industrial & Engineering Chemistry Research, 2004. 43(15): p. 4376-4379.
[2] Tetko, I.V., et al., How Accurately Can We Predict the Melting Points of Drug-like Compounds? Journal of Chemical Information and Modeling, 2014.
[3] Bradley, J.-C.; Lang, A.; Williams, A. Jean-Claude Bradley Double Plus Good (Highly Curated and Validated) Melting Point Dataset. http://dx.doi.org/10.6084/m9.figshare.1031638 (accessed November 15, 2014).
[4] ENAMINE Ltd. http://www.enamine.net (accessed November 15, 2014).
[5] Brown, T.N., J.A. Arnot, and F. Wania, Iterative fragment selection: A group contribution approach to predicting fish biotransformation half-lives. Environmental Science & Technology, 2012. 46(15): p. 8253-60.
[6] Brown, T.N., Predicting hexadecane-air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. SAR and QSAR in Environmental Research, 2014. 25(1): p. 51-71.
[7] Arnot, J.A., T.N. Brown, and F. Wania, Estimating screening-level organic chemical half-lives in humans. Environmental Science & Technology, 2014. 48(1): p. 723-730.

## Acknowledgements

Funding from UNILEVER and the American Chemistry Council Long-Range Research Initiative (ACC-LRI) for this project are gratefully acknowledged.