A Machine Learning Approach to Cation Identification for Solubility Optimization Of Redox Active Material for Non-aqueous Redox Flow Batteries

Yanni Etchi and Dr. Maricris L. Mayes

UMass | Dartmouth

National Science Foundation Research Experiences for Undergraduates (NSF-REU) site for Advanced Interdisciplinary Materials Research for Maritime Applications

Motivation

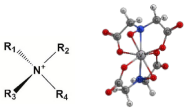
As renewable energy sources like wind and solar gain traction, efficient energy storage becomes crucial for grid stability. Redox flow batteries (RFBs) offer promising solutions, with non-aqueous RFBs (NRFBs) showing distinct advantages like wider temperature range and higher energy density.[3] The solubility of active materials is a critical factor in determining the energy density of NRFBs. However, synthetically exploring a large parameter space to identify soluble species is an inefficient approach. To expedite NRFB material discovery and optimization, this work explores the transformative role of computation and machine learning.[1]

Objectives

1. Develop a quantum chemistry-informed machine learning (ML) model to predict the solubility of ionic redox active materials.
2. Identify important design features for optimizing electrolyte solubility and use ML model to predict solubility of new possible cations.

Methods

- *Model system*: alkylammonium vanadium bis-hydroxyiminodiacetate
  (V^{+4}:[cation]$_2$[VBH]; V^{+5}: [cation][VBH])
  Solvent: Acetonitrile



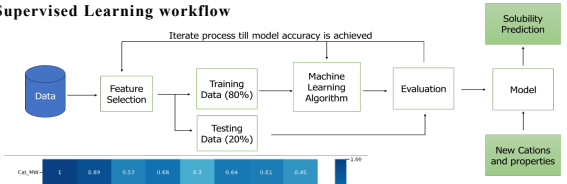
[VBH]$^{2-}$

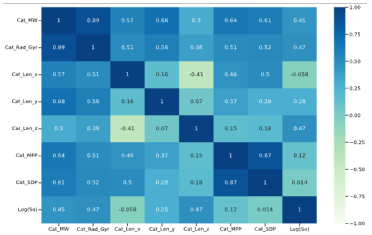
- *VBH solubility dataset:* calculated using

$$\Delta G^{*}_{dis} = \Delta G^{*}_{sub} + \Delta G^{*}_{sol} = -RT\ln(S_o V_m)$$

  Each thermodynamic term is taken from quantum chemistry calculations. Contains 119 crystals for each oxidation state.
- Selection of relevant cation features for $[VBH]^{2-}$ and $[VBH]^{-}$ subsets by variance and correlation analysis.
- Training of various regression models with 80% of each subset, and evaluation of these models with 20% of each subset making use of R^2 and Root Mean Squared Error(RMSE) metrics. Models were created using scikit-learn.[2]
- Comparison of relevant features for solubility prediction by importance, for tree-based models.
- Calculation of relevant features for 500 unexplored quaternary ammonium cation molecules from ChemBL database[4] using multiwfn[5] and prediction of their solubility with $[VBH]^{2-}$and $[VBH]^{1-}$ as anions using the most accurate model.

Supervised Learning workflow

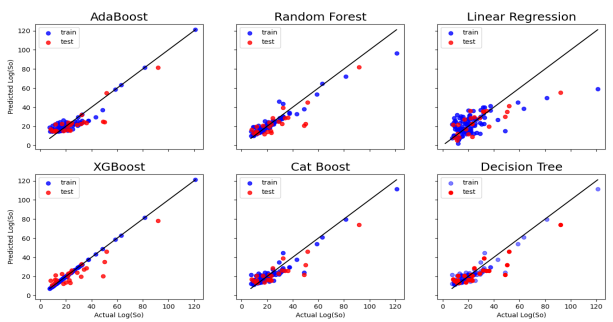




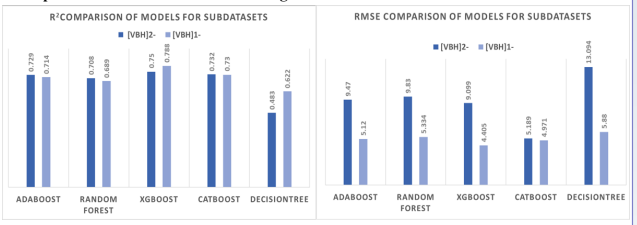
Heatmap of selected cation features (correlation threshold(0.97)

Results

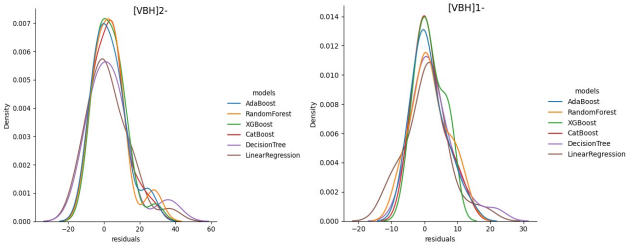
Cross plots of models for $[VBH]^{2-}$ subset (similar for $[VBH]^{-}$ subset)



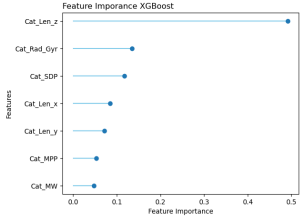
Comparison of R^2 and RMSE for regression models



Residual distribution of models for sub-datasets

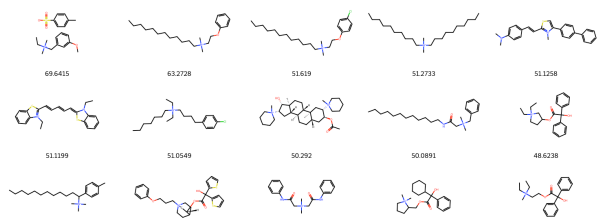


Feature comparison by importance for most accurate model(XGBoost)

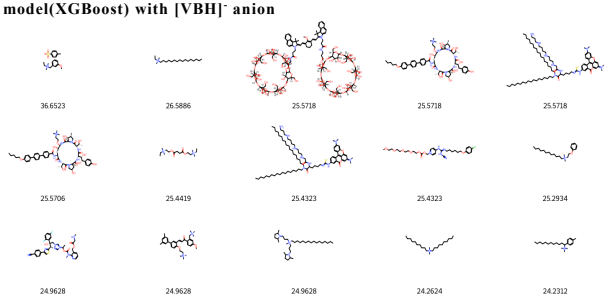


Feature	Description
Cat_MW	Molecular Weight of Cation, amu
Cat_Rad_Gyr	Radius of Gyration of the Cation, Ang
Cat_Len_x	Length of the system along x, Ang
Cat_Len_y	Length of the system along y, Ang
Cat_Len_z	Length of the system along z (Ang)
Cat_MPP	Molecular Planarity Parameter of Cation (MPP, Multiwfn), Ang
Cat_SDP	Span of Deviation from Plane (SDP, Multiwfn), Ang

Molecules with highest solubility prediction using most accurate model(XGBoost) with $[VBH]^{2-}$ anion



Molecules with highest solubility prediction using most accurate model(XGBoost) with $[VBH]^{-}$ anion



Conclusions

- Accurate models were trained and used to predict the solubility of $[VBH]^{2-}$ and $[VBH]^{-}$ based compounds depending on the cation properties. Extreme gradient boost (XGBoost) was found to be the most accurate at predicting solubility for our specific dataset.
- Unexplored quaternary ammonium cations predicted with high solubility were identified.
- Target cation features were identified and ranked based on their importance in solubility prediction for tree-based regression models. The length of the molecule along the Z direction was found to be the most relevant when predicting solubility for our particular model system.

References

1. Chalamala, B. R., Soundappan, T., Fisher, G. R., Anstey, M. R., Viswanathan, V. V., & Perry, M. L. (2014). *Proceedings of the IEEE*, *102*(6), 976–999.
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
3. Gong, K., Fang, Q., Gu, S., Li, S. F., & Yan, Y. (2015). *Energy & Environmental Science*, *8*(12), 3515–3530.
4. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A. Overington, J. P. (2011).
5. Lu, T. and Chen, F. (2012) Multiwfn: A Multifunctional Wavefunction Analyzer. Journal of Computational Chemistry, 33, 580-592.

Acknowledgements

This research work is supported by National Science Foundation' Division of Materials Research (DMR) and Ocean Sciences (OCE) under the grant number 2149893.