Data-Driven Design of Ion-Selective Polymeric Materials

Wylie Kau
Stanford Chemical Engineering
wyliekau@stanford.edu

Winter 2024-25 CHEMENG 277 Final Project Paper

Project Github Repository: wkau/chemeng277-project-WFK

Abstract

Ion-selective ligand functionalized polymers (LFPs) are the next generation of materials under development for the separation and recovery of critical minerals from alternative feedstocks, such as lithium and copper ions from aqueous battery and electronics waste. Edisonian approaches to LFP development cannot keep pace with the increasing demand for critical minerals required for electrification. Precise design of LFPs requires exhaustive knowledge of the reactivity between a critical or contaminant mineral ion and organic ligand. Available datasets that tabulate the reactivity between a mineral ion and ligand molecule (ion-ligand) are incomplete and therefore have limited practical use in material design objectives. Beyond the issue of data sparsity, direct translation of the reported reactivity trend from the ideal/simplified ion-ligand system to the real-world environment in the LFP under study (ion-LFP) is impossible. Motivated to address these issues, I demonstrate an experimentally-light, data-driven design framework to rapidly accelerate the development of ion-selective LFPs through combination of supervised learning models and empirically determined Linear Free Energy Relationships (LFER). This method relies on 1) widely available datasets reporting ion-ligand reactivity, 2) supervised learning models to interpolate missing reactivities in the datasets, and 3) performing few experiments in order to determine LFP material-specific LFER parameters. I show that this method is applicable to predicting competitive ion-sorption in LFP materials under real-world conditions.

1. Introduction

The growing demand for copper (Cu) and lithium (Li) to support the clean energy transition motivates development of novel separations that enable critical mineral recovery from complex aqueous feedstocks. By 2040, traditional supply of Cu and Li from mining is projected to only meet 60% and 50% of demand, respectively.[1] Separation processes based on ion-selective LFP materials, such as electrochemical membrane separations or ion exchange sorption, show promise for the extraction of critical mineral products from impure aqueous feedstocks. LFPs affix an ion-binding ligand molecule to an inert polymer backbone support material. The choice of ligand depends on the composition of the aqueous feedstock, e.g. the target critical mineral and contaminant profile.[2] Therefore, design of LFP material is gated by knowledge of the reactivity between the a given mineral ion and a candidate ligand (ion-ligand). Existing datasets that

tabulate known reactivities are critically useful, yet incomplete, limiting direct comparison between ion reactivities for a single ligand. In addition, once a ligand molecule is affixed to a polymer framework, the reaction between the ion and ligand occurs in a different environment than that reported. Therefore, a methodology to translate reported reactivities to observed reactivities in real-world material is needed. In this work I address these issues by combining machine learning models with empirically determined reactivity translation relationships. First, I overcome data sparsity by applying supervised learning models to accurately interpolate missing ion-ligand interaction strength data with an average prediction root-mean-squared-error (RMSE) of 0.177 kcal/mol and to extrapolate interaction strength trends to novel, un-observed ligand chemistries with a prediction RMSE of 0.285 kcal/mol. Second, I apply well-described Linear Free Energy Relationships (LFERs) to translate the reported/predicted ion-ligand reactivity trend to the experimentally observed ion-LFP reactivity trend. Third, I demonstrate how the combination of the interpolated ion-ligand dataset and the experimentally determined LFER can be used to rapidly screen for promising critical mineral separations unique to an LFP chemistry, without performing wet-lab experiments. Finally, I present future opportunities for data-driven approaches to further reduce experimental time in LFP research and development, through learning relationships between LFER parameters and predicted LFP material properties. In summary, by combining data interpolation via supervised learning models with empirically determined reactivity translation relationships, I demonstrate a data driven design method to rapidly accelerate the design of polymer materials key to secure critical mineral supply chains.

2. Related Work

There have been several studies published recently using machine learning approaches to explore prediction of ion-ligand binding affinity. Kanahashi, et al. demonstrate a Gaussian Process Regression (GPR) for predicting the stability constants of single- and multi-dentate ion-ligand complexes using molecular descriptor featurization and ion properties.[3] Zahariev, et al. demonstrate a Message Passing Neural Network (MPNN) approach for prediction ion-ligand interaction strength using molecular graphs and incorporation into a novel ligand design workflow.[4] Chaube, et al. report investigating a variety of supervised learning techniques ranging from Support Vector Machines (SVM) to Multi-Layered Perceptrons (MLP) to predict ion-ligand interaction strength specifically for lanthanide ions and the use of the trained model to screen out-of-training-set ligand candidates.[5] However, across all these studies there is no experimental validation of generated predictions for test or new ligands, and no discussion of how to translate the results of such workflows beyond the ion-ligand system to additional systems of interest.

3. Methods

3.1 Supervised Learning for Dataset Interpolation

The NIST SRD 46. Critically Selected Stability Constants of Metal Complexes[6] dataset includes a vast number of recorded equilibrium constant (K) values for organic ligands and metal ions. To start, the dataset is filtered to only include observations of single ligand-single ion interactions and 23 ions spanning the Alkali, Transition, and Rare Earth elements. The filtered dataset includes a total of 7007 datapoints, 23 ions, and 1091 unique ligands (Figure S1). Reported equilibrium

constants are converted to reaction free energies via $\Delta G_{Ion-Ligand} = -RT\ln(K)$ to use as the target for prediction tasks, assuming an ambient temperature of 25 C.

Each database entry is featurized as follows: (1) The ion is represented by collection of standardized aqueous ion properties collected from literature, for a total of 16 features. (2) The ligand is first converted into a canonical smiles string, then encoded as 191 quantitative features using RDKit's Lipinski, Crippen, and Chem.Descriptors packages. In total, 207 features encoding ligand and ion identity are used as inputs into the supervised learning task to predict $\Delta G_{Ion-Ligand}$. No explicit dimensionality reduction or feature selection techniques are utilized.

Two regression models are compared for $\Delta G_{Ion-Ligand}$ prediction: an XGBoost Regressor (eXtreme Gradient Boosted Trees) and an ElasticNet Regressor (L1&L2 regularized linear regression). Model training and hyperparameter optimization is performed using the HalvingGridSearchCV methodology on an 80%/20% train-test-split.

Model prediction error evaluation is performed for both out-of-training-set ligands (OOTSL error) and for each specific ion (Ion-specific error). OOTSL $\Delta G_{Ion-Ligand}$ error is evaluated by the average R^2 and RMSE across ten repetitions of holding out 10% of all ligands from training, fitting a model with the optimal hyperparameters, and evaluating the prediction on the holdout data. Ion-specific error is evaluated by iterating across all 23 ions and performing three repetitions of holding out 50% of the ion-specific observed data from training and evaluating the prediction on the holdout data. In this way, we can assess the model's predictive accuracy for new ligand chemistries not tabulated in training dataset and the predictive accuracy for different electrophiles, which have varying numbers of observations in the training dataset.

Finally, the model with optimized hyperparameters is re-trained on the entire input dataset. Then, the trained model is used to predict $\Delta G_{Ion-Ligand}$ for all ion-ligand pairs for which there is not an observation in the training dataset and the prediction error tabulated on a per-ion basis as determined above, to construct an "augmented" dataset which combines reported and predicted $\Delta G_{Ion-Ligand}$ values.

3.2 Linear Free Energy Relationships for Reactivity Translation

To measure $\Delta G_{Ion-LFP}$ in a real-world polymer system, a model system 50 mol% 4-vinylpyridine (4VP) LFP membrane is synthesized as reported previously.[2] $\Delta G_{Ion-LFP}$ is measured by performing single-salt sorption experiments (appendix) to measure ion-specific equilibrium sorption capacity, q_e, and $\Delta G_{Ion-LFP}$ calculated using the Langmuir Adsorption thermodynamic model with a monolayer maximum sorption capacity assumption (with error from triplicate measurements) (Figure S3).

A Linear Free Energy Relationship of the form $\Delta G_{Ion-LFP} = \beta_1 * \Delta G_{Ion-Ligand} + \beta_0$ is then applied to the data and parameters fit using Orthogonal Distance Regression, which is a regression tool which accounts for error in both the reported/predicted $\Delta G_{Ion-Ligand}$ and measured $\Delta G_{Ion-LFP}$.[7]

Finally, the fit LFER is used to generate predicted partitioning factors between two ions for the 50 mol% 4VP system and compared directly to binary salt sorption separation factors to assess the practicality/use of the framework in assessing real-world material performance (Figure S4).

4. Results and Discussion

Results of supervised model training and error evaluation are shown in Figure 1. The XGBoost Regression model outperforms the ElasticNet Regression model in both predictive accuracy (R2 Score) and prediction RMSE. In the case of the XGBoost Regressor, a monotonic increase in prediction performance is observed over 5 training epochs. A test R2 score of 0.92 is achieved with a RMSE prediction error of 0.170 kcal/mol and Symmetric Mean Absolute Percent Error (SMAPE) of 24.7%. However, because the original train-test-split of the training data is agnostic to ligand identity, it is possible that ligand molecules may appear in both the training and test sets, necessitating separate error evaluation. Comparing OOTSL error and ion-specific prediction error, the XGBoost regressor outperforms the ElasticNet regressor. Therefore, for all future discussion of predicted $\Delta G_{Ion-Ligand}$, the XGBRegressor with optimal hyperparameters determined was used and referred to as "Final Model". The final model was re-fit to all available training data and then used to interpolate missing observations in the NIST46 dataset to construct the augmented $\Delta G_{Ion-Ligand}$ dataset. This corresponds to expanding the available data from 4949 ion-ligand pairings to 25093 ion-ligand pairings.

However, the performance of the superior XGBoost Regressor model does not surpass the highest performance reported in literature for similar prediction tasks. Kanahashi, et al. report a an out-of-training-set ligand predictive accuracy R2 score of 0.84 using a Gaussian Process Regression model (compare to a mean R2 score of 0.70 observed in OOTSL error evaluation for the XGBoost Regressor model reported here).[3] The supervised learning approach demonstrated here does not include any dimensionality reduction or feature selection and is relatively simple.

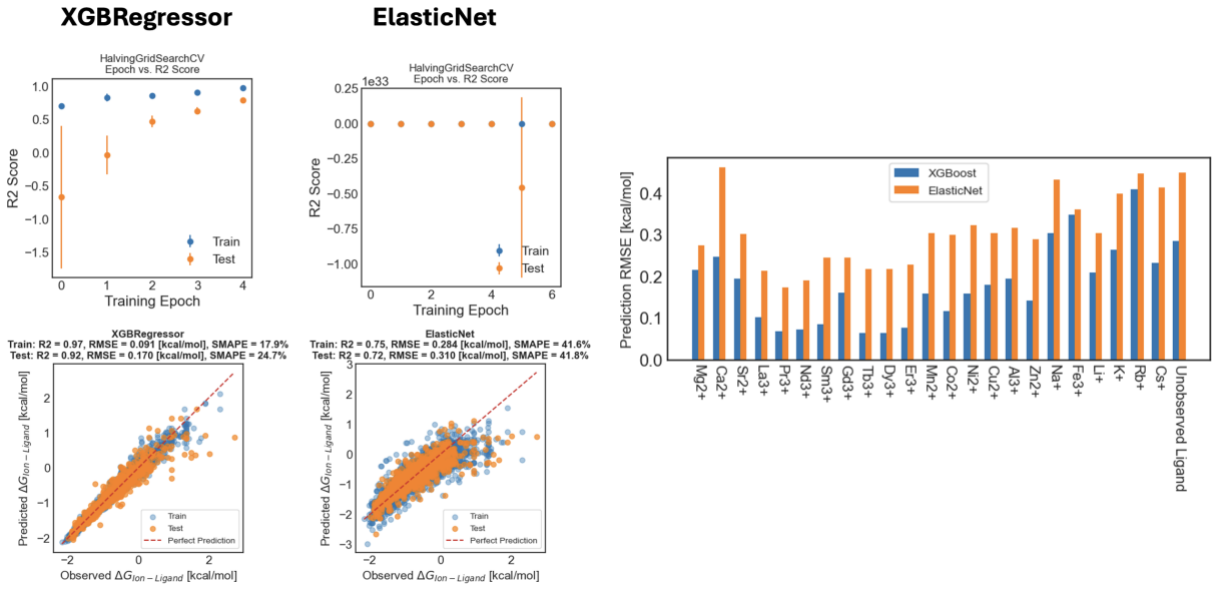

Figure 1. (Left) R2 Score vs. Training Epoch and parity plots for both XGBoost and ElasticNet regression models. (Right) Results of OOTSL and Ion-specific error evaluation for both XGBoost and ElasticNet regression models.

To demonstrate the application of exhaustive $\Delta G_{Ion-Ligand}$ knowledge and transfer reactivity trends to a real-world system under study in our lab, the experimentally observed $\Delta G_{Ion-LFP}$ for

a 50 mol% 4VP membrane is plotted against the reported/predicted $\Delta G_{Ion-Ligand}$ tabulated in the augmented NIST46 dataset for the pyridine molecule and 6 ions of interest. Of the 6 ions experimentally measured, 4 $\Delta G_{Ion-Ligand}$ values are real observations in the NIST46 dataset (Cu2+, Mn2+, Co2+, and Ni2+) while 2 are predicted values (Li+, and Mg2+) (Figure S2). Fitting a linear relationship via Orthogonal Distance Regression, the resulting fit achieves at R2 score of 0.83 (Figure 2). The linear transferability between $\Delta G_{Ion-Ligand}$ and $\Delta G_{Ion-LFP}$ is a promising result, as referencing the idealized ion-ligand system will allow comparison between un-observed ions of interest given a ligand identity. To further probe the usefulness of the LFER approach, the LFER predicted partitioning coefficient, or ratio of equilibrium constants between ions (K_i/K_j), for 3 ion pairs are compared to the separation factors observed in binary salt separation experiments (Figure 2). For all ion pairs, the LFER prediction falls within error of the observed separation factors. This result demonstrates the promise of the approach, as knowledge of the separation factor can help guide assessment of LFP material performance in different separation conditions, such as a membrane or ion-exchange separation. To my knowledge, this is the first demonstration of applying LFERs to translate reactivity and selectivity trends from idealized metal-ligand complexation data to a LFP system.

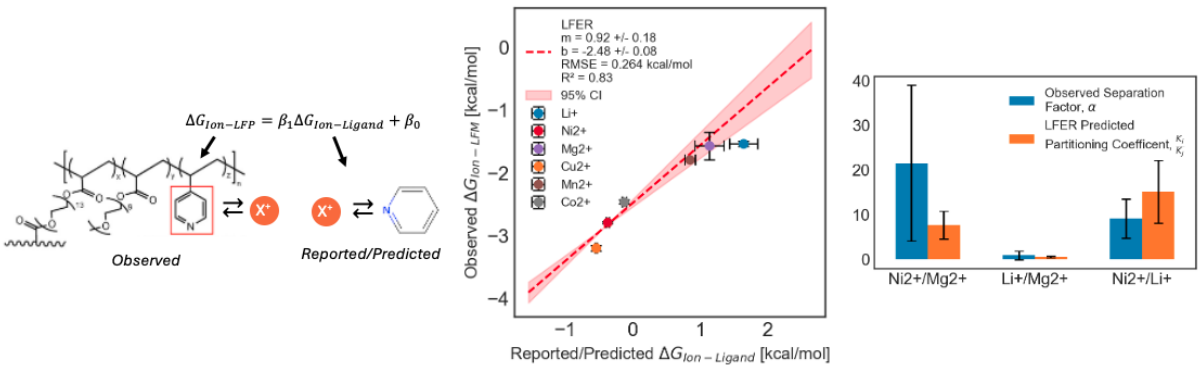

Figure 2. (Left) LFER demonstrating reactivity translation between Ion-Ligand and Ion-LFP systems. (Center) Applied LFER for a pyridine molecule and a 4-vinylpyridine functionalized LFP. (Right) Comparison of LFER predicted partitioning coefficients to experimentally observed separation factors for 3 ion pairings.

5. Conclusion

This work presents a novel framework for polymer material design relying on supervised learning and empirically determined free energy relationships. Future work can be split into three thrusts: 1) improving the supervised learning model used for NIST46 dataset augmentation to improve $\Delta G_{Ion-Ligand}$ prediction accuracy; 2) synthesizing additional ligand functionalized polymer materials, such as acrylic-acid functionalized monomers and applying a LFER to the data; and 3) exploring prediction of LFER parameters based on LFP structure, to completely eliminate the need to perform sorption experiments to extract $\Delta G_{Ion-LFP}$. If successful, this framework will rapidly accelerate LFP design for critical mineral separation applications and bolster secure critical mineral supply chains to support electrification.

References

(1) Kim, T.-Y.; Gould, T. Global Critical Minerals Outlook 2024. *Int. Energy Agency* **2024**.

(2) Abels, K.; Botelho Junior, A. B.; Chen, X.; Tarpeh, W. A. Ligand Content and Driving Force Effects on Ion-Ion Permselectivity in Ligand-Functionalized Membranes. *J. Membr. Sci.* **2025**, *714*, 123418. https://doi.org/10.1016/j.memsci.2024.123418.

(3) Kanahashi, K.; Urushihara, M.; Yamaguchi, K. Machine Learning-Based Analysis of Overall Stability Constants of Metal–Ligand Complexes. *Sci. Rep.* **2022**, *12* (1), 11159. https://doi.org/10.1038/s41598-022-15300-9.

(4) Zahariev, F.; Ash, T.; Karunaratne, E.; Stender, E.; Gordon, M. S.; Windus, T. L.; Pérez García, M. Prediction of Stability Constants of Metal–Ligand Complexes by Machine Learning for the Design of Ligands with Optimal Metal Ion Selectivity. *J. Chem. Phys.* **2024**, *160* (4), 042502. https://doi.org/10.1063/5.0176000.

(5) Chaube, S.; Goverapet Srinivasan, S.; Rai, B. Applied Machine Learning for Predicting the Lanthanide-Ligand Binding Affinities. *Sci. Rep.* **2020**, *10* (1), 14322. https://doi.org/10.1038/s41598-020-71255-9.

(6) Burgess, D. R. NIST SRD 46. Critically Selected Stability Constants of Metal Complexes: Version 8.0 for Windows, 2020. https://doi.org/10.18434/M32154.

(7) *Orthogonal distance regression (scipy.odr) — SciPy v1.15.2 Manual*. https://docs.scipy.org/doc/scipy/reference/odr.html (accessed 2025-03-12).

Appendix

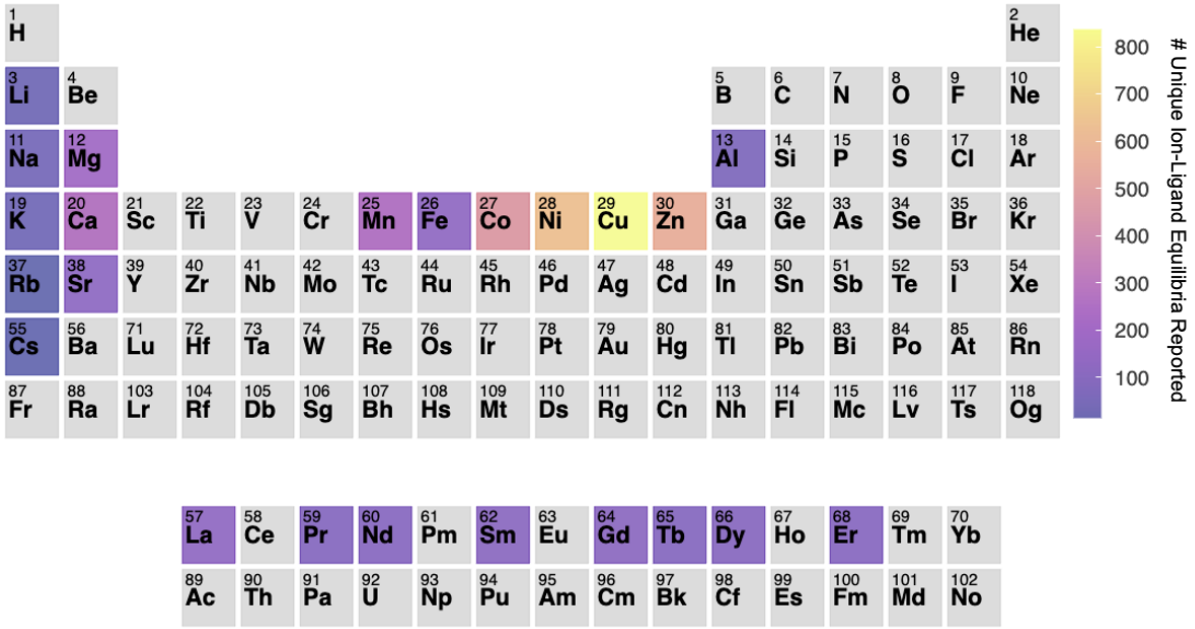

Figure S1. Filtered NIST46 Dataset used in study, showing observation distribution across ion elements. Single ligand-single ion equilibria, e.g. K=[ML]/[M][L] only.

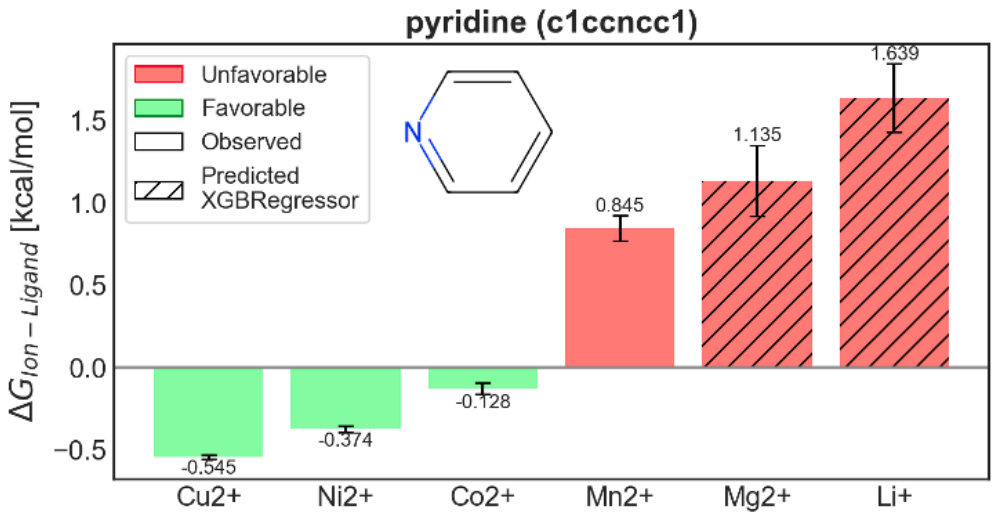

Figure S2. Query results of the augmented NIST46 dataset for the pyridine molecule and 6 ions of interest.

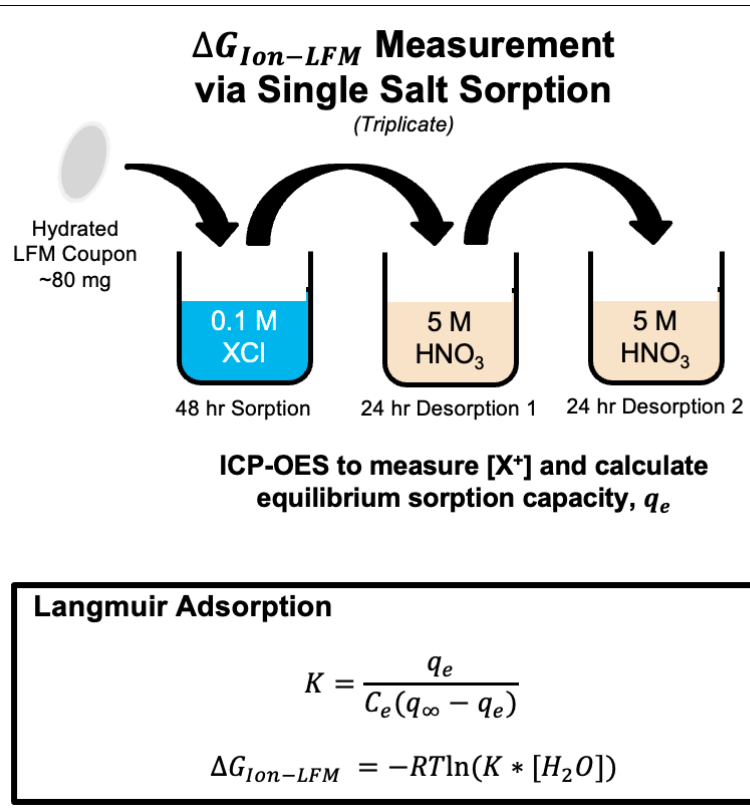

Figure S3. Single salt sorption experiment to determine $\Delta G_{Ion-LFP}$. In the Langmuir Adsorption equations: C_e = the ion concentration in the sorption solution, q_∞ = the ligand loading in the LFP.

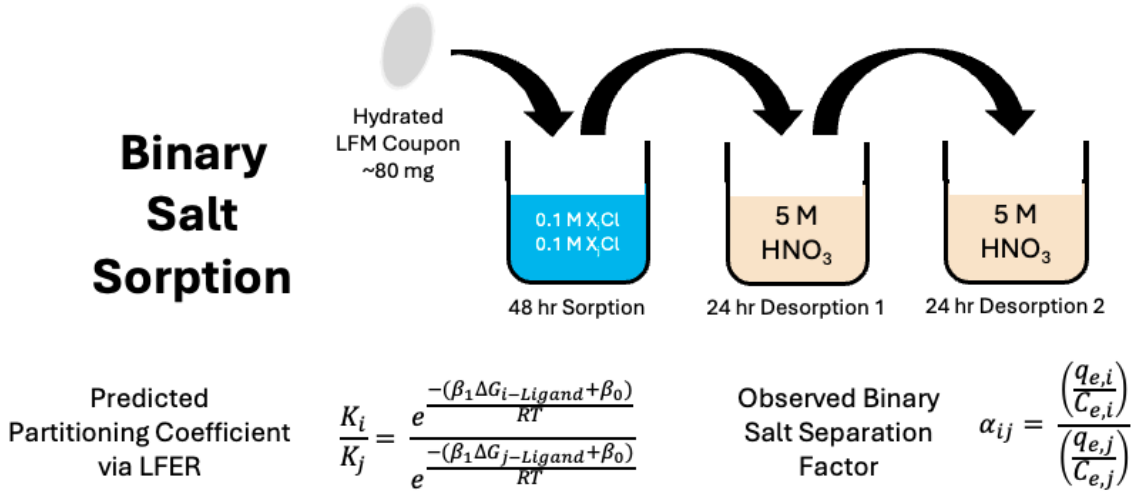

Figure S4. Binary salt sorption experiment, equations for predicted partitioning coefficient and calculating observed binary salt separation factors.