

Prediction of stability constants of metal-ligand complexes by machine learning for the design of ligands with optimal metal ion selectivity

Federico Zahariev, Tamalika Ash,[‡] Erandika Karunaratne,[‡] Erin Stender, Mark S. Gordon,

Theresa L. Windus, Marilú Pérez García

Ames National Laboratory, Ames, Iowa

Critical Materials Innovation Hub, Ames, Iowa

Iowa State University, Ames, Iowa

ABSTRACT

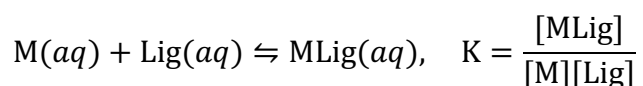
The new LOGKPREDICT program integrates HostDesigner molecular design software with the machine learning (ML) program Chemprop. By supplying HostDesigner with predicted log K values, LOGKPREDICT enhances the computer-aided molecular design process by ranking ligands directly by metal-ligand binding strength. Harnessing the reliable experimental data from a historic National Institute of Standards and Technology (NIST) database, and data from the International Union of Pure and Applied Chemistry (IUPAC) we train message passing neural net algorithms. The multi-metal NIST-based ML model has an RMSE of 0.629 ± 0.044 (R^2 of 0.960

± 0.006), while the two versions of lanthanide-only IUPAC-based ML models have respectively RMSE of 0.764 ± 0.073 (R^2 of 0.976 ± 0.005) and 0.757 ± 0.071 (R^2 of 0.959 ± 0.007). For relative log K predictions on an out-of-sample set of six ligands, demonstrating metal ion selectivity, reaches a commendably low RMSE value of 0.25. We showcase the use of LOGKPREDICT in identifying ligands with high selectivity for lanthanides in aqueous solutions, a finding supported by recent experimental evidence. We also predict new ligands yet to be verified experimentally. Therefore, our ML models implemented through LOGKPREDICT and interfaced with ligand design software HostDesigner pave the way for designing new ligands with predetermined selectivity for competing metal ions in aqueous solution.

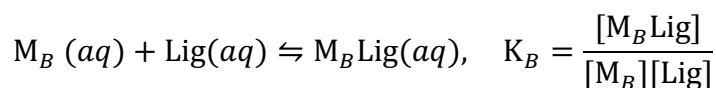
I. INTRODUCTION

Hydrometallurgy is a branch of the extractive metallurgy field, dedicated to recovering metals from natural or recycled sources. Solvent extraction and ion exchange are two common techniques that rely on metal ions selectively binding to a ligand through an electron-donating atom such as oxygen or nitrogen. The binding functional group is typically a part of an organic soluble extractant molecule or attached to a resin. The ligand selectivity is predicated upon several factors such as the binding strength of each electron-donating group to a metal ion, commonly described through hard-soft acid-base theory, the number of electron-donating groups in the molecule, and the compatibility between the molecular structure and its ability to form a multi-dentate metal-ligand complex. With an increasing need to sustainably and responsibly recover the critical elements needed to produce modern technology, it is essential to find new and improved recovery methods to replace the resource-intensive methods currently in use. One way to quickly develop chemicals with improved performance is to design ligands that bind strongly and selectively to a target metal ion using computer-aided molecular design (CAMD).

One can evaluate the relative binding strength of a ligand (Lig) to a metal ion (M) by determining the metal-ligand complex (MLig) stability constant (K), defined as the equilibrium constant for the chemical reaction shown in Scheme 1. The stability constant is the quotient of the MLig concentration ([MLig]) and the product of the M and Lig concentrations, [M] and [Lig]. The selectivity coefficient is the ratio of the K values of competing metals M_A and M_B , $K_A/K_B = K_{AB}$, see Scheme 2.



Scheme 1. Schematic representation of the chemical reaction equilibrium for the binding of a free metal ion (M) with a ligand (Lig) in solution to form a soluble metal-ligand complex (MLig). The equilibrium constant (K) is defined as the ratio of the concentrations of each species involved in the chemical reaction (denoted with square brackets).



Scheme 2. Schematic representation of two equilibria with a fixed ligand complexing to two competing metals, M_A (upper), and M_B (lower) to determine the selectivity coefficient, K_{AB} .

Throughout the years, stability constants, typically reported as log K values, and other thermodynamic properties have been compiled in physical volumes¹ and in digital databases.^{2,3} The availability of stability constant data has led to many efforts to predict the log K values by empirical and computational modeling. For example, empirical models were created to correlate the ligand functional groups^{4,5} or $\text{p}K_a$ ^{6,7} to experimental data. The empirical models often reveal

linear correlations between the number of chemical functional groups in a ligand and the log K for a restricted class of ligands. The lack of a transferable empirical model requires different tools to model selectivity between ligands of different classes.

More accurate but computationally expensive methods exist; however, they bring their own set of challenges.⁸ Computational modeling of metal-ligand complex formation using a thermodynamic cycle, and *ab-initio* calculations have demonstrated that these methods can be challenging and resource-intensive. The complexity in molecular systems requires careful consideration to minimize errors stemming from the need to represent the molecules as they exist in solution, the presence of multiple conformers, the differences in the number of coordinating water molecules, and the choice of the most appropriate level of theory.⁹ In practice, the errors in the computational predictions based on the thermodynamic cycle with respect to the experimental values are often significant, and some additional *ad hoc* empirical corrections are required for meaningful computational results. Unfortunately, the empirical corrections have the same deficiency as the above-mentioned empirical methods: such corrections are devised for a restricted class of ligands and are not easily transferable. As a result, data-driven approaches such as machine learning (ML) via deep neural networks have been used and shown great promise for estimating the metal-ligand log K values.¹⁰

The ML approach to predict the stability constants and metal-ligand selectivity presented in this article is realized through the hybridization of the computer-aided supramolecular design program HostDesigner¹¹ and the ML program Chemprop¹² by the newly created interlinking LOGKPREDICT¹³ software component. Although the prediction of stability constants of metal-ligand complexes was studied by Chaube et al.¹⁰ and Kanahashi et al.,¹⁴ metal selectivity in metal-

ligand binding has yet to be studied. Moreover, we present a tool that combines the power of ML and ligand design to aid in the discovery of new metal binding ligands.

The application of a combined ligand design and stability constant prediction is provided for the ligand selectivity for rare earth elements (REE). Here, along with the direct prediction of the individual stability constants, we also predict the selectivity coefficients. The 17 rare earth elements (REEs) form a group of chemically similar elements, but each simultaneously has distinct and unique electronic structures requiring accurate methods to model their complexes.^{15,16} REEs are typically found together in various ratios and mixed with different metals in a number of mineral forms around the globe.¹⁷⁻²⁰ Both the U.S. Department of Energy and the European Commission have labeled REEs as “critical materials” because of their importance to the high-tech industry and simultaneous supply shortage.²¹⁻²⁴ The ability to extract target REEs through separation processes is of great importance. A combination of liquid-liquid solvent extraction processes using metal ion binding ligands is currently the dominant approach for separating REEs.²⁵⁻²⁸ Thus to demonstrate the functionality of the combined HostDesigner-LOGKPREDICT-Chemprop software, an example of the ligand design process is used.

In Section II, we describe updates to the HostDesigner molecular design program, the Chemprop machine learning program, and LOGKPREDICT, the new software we have developed to interface HostDesigner and Chemprop, enabling the computer-aided molecular design of ligands that are ranked by ML-predicted selectivity values. The datasets used to train the Chemprop ML models is described in Section III. Next, in Section IV, we describe the performance of different ML models. Additionally, we describe the design and ranking of amide-pyridine ligands for selectivity between Gd and La by using LOGKPREDICT. Section V summarizes the results presented in the paper.

II. METHODS

CAMD is a typical modern approach in supramolecular chemistry.^{29,30} The CAMD approach suggests to experimentalists a host molecule, which optimally binds a given target guest molecule or ion, as a most promising candidate to be synthesized in a lab. Studies of ion bindings to simple chemical groups, such as ethers, amines, and arenes, elucidated the basic geometrical patterns of optimal binding.³¹ The development of computationally inexpensive force fields have accelerated the ability to pre-screen for optimal guest-host binding affinity.³² One would usually refine the search with more accurate but computationally more expensive *ab-initio* methods.³³ In all instances, one must start with some plausible molecular structures.

The choice of host and guest molecule or ion is very important to initiate the design. The host molecule could have two or more favorable binding sites for the guest³⁴ and adopt a few low-energy binding conformations with respect to other possible conformations.³⁴⁻³⁸ The process of computer-aided host design can be viewed as a selection of binding sites and geometric structures connecting the sites.

II.1 HostDesigner Ligand Design

The HostDesigner computer program¹¹ constructs candidate supramolecular host complexes based on a list of fragments and linker groups, ranks the resulting hosts by how well they satisfy the user-defined structural constraints and has several optional post-processing capabilities. HostDesigner was initially conceived to design organic molecules that bind a target metal ion. Within two decades HostDesigner has evolved into a mature platform for designing arbitrary supramolecular structures from molecular fragments.

The initial version of HostDesigner was released in 2002.³⁹ In 2006, version 2.0,⁴⁰ extended the capabilities of the code by incorporating multi-atom guests, variable input geometry,

conformational and entropic energy contributions of the host binding and an expanded library of linkers. Version 3.0, released in 2014, eliminated the limitations on the types of input molecules, removed high energy linking fragments from the library, enabled molecular-mechanics post-processing.

The latest version 4.3 has several significant improvements, such as the inclusion of the fusion of atoms or/and bonds, implementation of a greater variety of input file formats (cc1: Chem3D Cartesian Coordinate 1, PDB: Protein Databank, and SDF:⁴¹ molfile) and output (including XYZ, PDB, SDF), the addition of an auxiliary molecular-mechanics engine (mengine) code as optional post-processing of geometric structures, and the inclusion of algorithms for generating polygons, polyhedron edges, or faces.⁴²

To estimate the selectivity of a given host ligand with respect to a target metal ion versus another metal ion, the prediction of one-to-one absolute metal-ligand affinities under a set of standard conditions (aqueous conditions at 25 °C and zero ionic strength) is predicted for each metal ion.

II.2 Log K Predictions with Chemprop

Chemprop¹² is based on a generalization of the Message Passing Neural Network (MPNN), the Directed MPNN (D-MPNN). An MPNN is a neural network that operates on graphs.⁴³ The features typically associated with ML are included as part of the graph components. Consider an undirected graph G that consists of nodes and edges connecting the nodes, and associated features with each node v (denoted by x_v) and each edge (denoted by the feature e_{vw} , for the edge between nodes v and w). The forward pass of the neural network has two phases: a message passing phase and a readout phase. In the message passing phase, information is propagated across the graph in discrete

time steps to build a neural representation of the graph. In the readout phase, the neural representation of the graph is used to make a prediction.

Features associated with the full molecular system can be added separately and used in conjunction with the graph representations. For example, physical characteristics, such as density, or chemical characteristics, such as toxicity.

II.2.1 MPNN Method

In the message passing phase, hidden states h_v^t that are associated with each node undergo a finite number of time steps T (labeled by the upper index t) constructed by means of two functions: a message function M_t and a vertex update function U_t through the intermediacy of messages m_v^t :

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

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}), \quad (1b)$$

where the sum in Eq. (1a) is across the neighboring nodes $N(v)$ of node v . In the readout phase, a readout function R acts on all of the hidden states at the final step T to obtain the prediction \hat{y} .

$$\hat{y} = R(\{h_v^T | v \in G\}). \quad (2)$$

Usually, the readout function, R , is simply a sum of the hidden states followed by a standard feed-forward neural network.

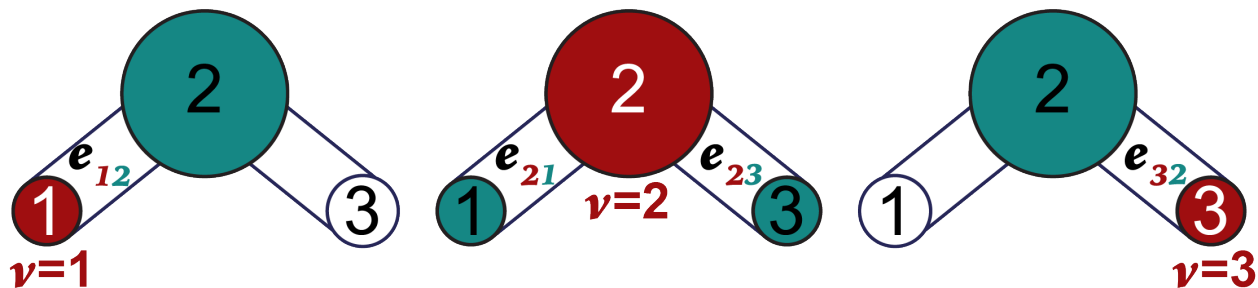


Figure 1. The message passing phase of the forward pass exemplified on a water molecule. The three atoms in water are represented as circles labeled 1, 2, and 3, representing hydrogen, oxygen, and hydrogen, respectively. The three atoms of the water molecule are the nodes, and the molecular bonds are the edges of the graph. The red circle for each v denotes the main node and the green circle(s) correspond to the node(s) neighboring the main node.

If MPNN is applied to molecules, the atoms and bonds correspond to the nodes and edges of an undirected graph. Figure 1 illustrates the graph corresponding to the water molecule. The message passing phase in the case of the water molecule is constructed as a specific version of Eq. (1a) and Eq. (1b), which for node 1 is:

$$m_1^{t+1} = M_t(h_1^t, h_2^t, e_{12}) \quad (3a)$$

$$h_1^{t+1} = U_t(h_1^t, m_1^{t+1}), \quad (3b)$$

for node 2,

$$m_2^{t+1} = M_t(h_2^t, h_1^t, e_{21}) + M_t(h_2^t, h_3^t, e_{23}) \quad (4a)$$

$$h_2^{t+1} = U_t(h_2^t, m_2^{t+1}), \quad (4b)$$

and for node 3,

$$m_3^{t+1} = M_t(h_3^t, h_2^t, e_{32}) \quad (5a)$$

$$h_3^{t+1} = U_t(h_3^t, m_3^{t+1}). \quad (5b)$$

II.2.2 D-MPNN Method

D-MPNN is a generalization of MPNN that addresses the problem of spurious oscillations of the message passing steps,⁴⁴⁻⁴⁶ i.e., $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_n$, where $v_i = v_{i+2}$, i.e. an oscillation between v_i and v_{i+1} . The D-MPNN solution consists of using directed, instead of undirected, graphs and messages associated with directed edges, instead of nodes. The messages in D-MPNN are

associated with directed edges and have two indices instead of just one as in MPNN. Moreover, the order of the two indices matter. For example, the message m_{vw}^t goes from node v to node w , thus it has a directionality. The messages m_{vw}^t and m_{wv}^t differ in directionality. The separation of forward and reverse directions in the D-MPNN results in better graphs than MPNN because there is a greater control over the flow of information in the graph and spurious message-passing oscillations are eliminated.

Chemprop implements D-MPNN with PyTorch and utilizes the convenient molecular representation capabilities of the RDKit Python library. The node features consist of the atom type, number of bonds the atom is involved in, formal charge, chirality, number of bonded hydrogen atoms, hybridization (sp, sp², sp³, sp³d, or sp³d²), aromaticity (whether the atom is part of an aromatic system), and atomic mass. The edge features consist of the bond type (single, double, triple, or aromatic), conjugation (whether the bond is conjugated), “in ring” (whether the bond is part of a ring), and “stereo” (for example, cis/trans).

D-MPNN can be also used with additional, user supplied molecular properties. For convenience, we will call it “D-MPNN+” in the remainder of the document.

II.3 The Link Between HostDesigner and Chemprop by LOGKPREDICT

New metal-ligand complexes and their properties are produced by HostDesigner, passed to LOGKPREDICT to put it in a format that Chemprop understands, and log K values are predicted by Chemprop. These log K values are transformed by LOGKPREDICT and control is passed back to HostDesigner. Figure 2 is a diagram of the stages and information exchanges from HostDesigner to LOGKPREDICT to Chemprop and Chemprop to LOGKPREDICT to HostDesigner, with upper and lower arrows, respectively.

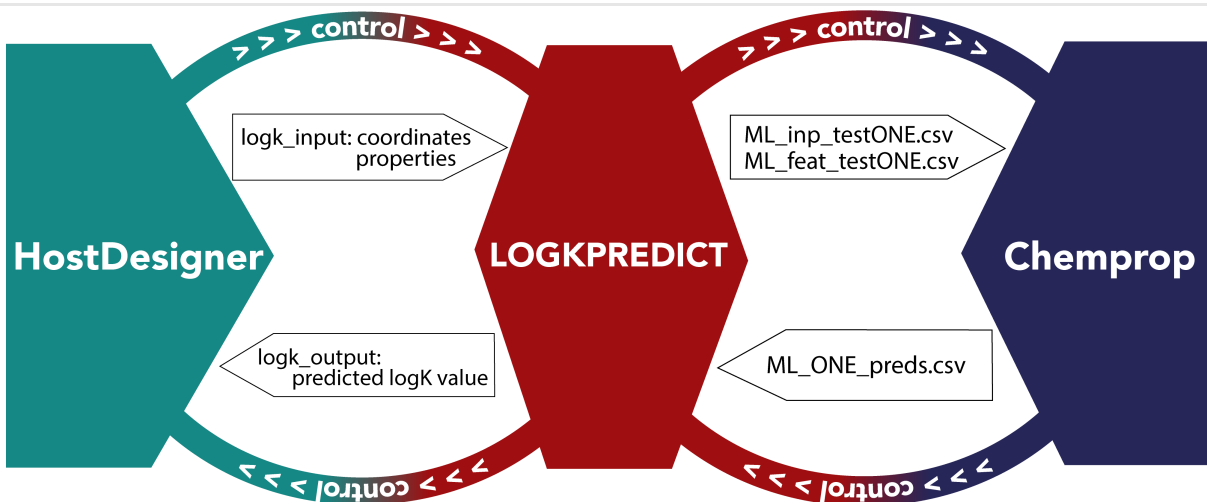


Figure 2. The initial stage of control from HostDesigner to LOGKPREDICT to Chemprop and the exchange of information via files: `logk_input`, `ML_inp_testOne.csv`, and `ML_feat_testOne.csv`. The final stage of control from Chemprop to LOGKPREDICT to HostDesigner and the exchange of information via files: `ML_ONE_preds.csv` and `logk_output`.

Initially, HostDesigner writes the molecular information of the metal-ligand complex in the `logk_input` file and gives control to LOGKPREDICT (see Figure 2). The first two lines of `logk_input` encapsulate the different metal-ligand properties computed by HostDesigner listed in Table 1. The rest of the file contains the MDL Molfile⁴¹ format of the molecular complex.

Table 1. The HostDesigner descriptors and corresponding abbreviations in the “`logK_input`” created by HostDesigner for transfer to LOGKPREDICT, then Chemprop and back to HostDesigner.

Feature No. ^a	Name	Description
1	<code>logK(I=0.0)</code>	log K value at zero ionic strength ^b

2	logK_in	Measured log K value ^e
3	I_in	Measured ionic strength ^c
4	Z_lig	Ligand charge
5	Z_met	Metal formal charge
6	nrot	Number of ligand bond rotations restricted by metal complexation
7	met_r	Effective metal ionic radius ⁴⁷
8	met_CN	Most common coordination number ^d
9	E_strain	MM3 strain energy associated with complex formation
10	G_solv	Metal ion hydration free energy at 25 °C ⁴⁸
11	rdhE	electrostatic descriptor ⁵
12	rdhC	covalent descriptor ⁵

^alisted in order appearing in “logk_input”; ^blog K value corrected using the Davies equation (SI Section II);⁴⁹ ^cexperimental value derived by procedure in SI Section II; ^dbased on a Cambridge Crystal Structure Database⁵⁰ search where inner sphere contains only O and N atoms.

As shown in Figure 3, LOGKPREDICT transforms the MDL Molfile format into a simplified molecular-input line-entry system (SMILES) string⁵¹⁻⁵³ with the RDKit extension. A SMILES string is used to describe the structure of a three-dimensional chemical species in line notation that is easily understood by computer software. The SMILES format used here has “dative bonds” between the metal and the ligand⁵² and inserts dative bonds between the metal and the ligand to the SMILES string and writes it out to the file named “ML_inp_testONE.csv”.

Next, LOGKPREDICT starts Chemprop, which uses “ML_inp_testONE.csv” and “ML_feat_testONE.csv” as basic input and additional feature information, respectively (see Figure 2). Chemprop uses the pre-trained DLNN information stored in the “model.pt” file.

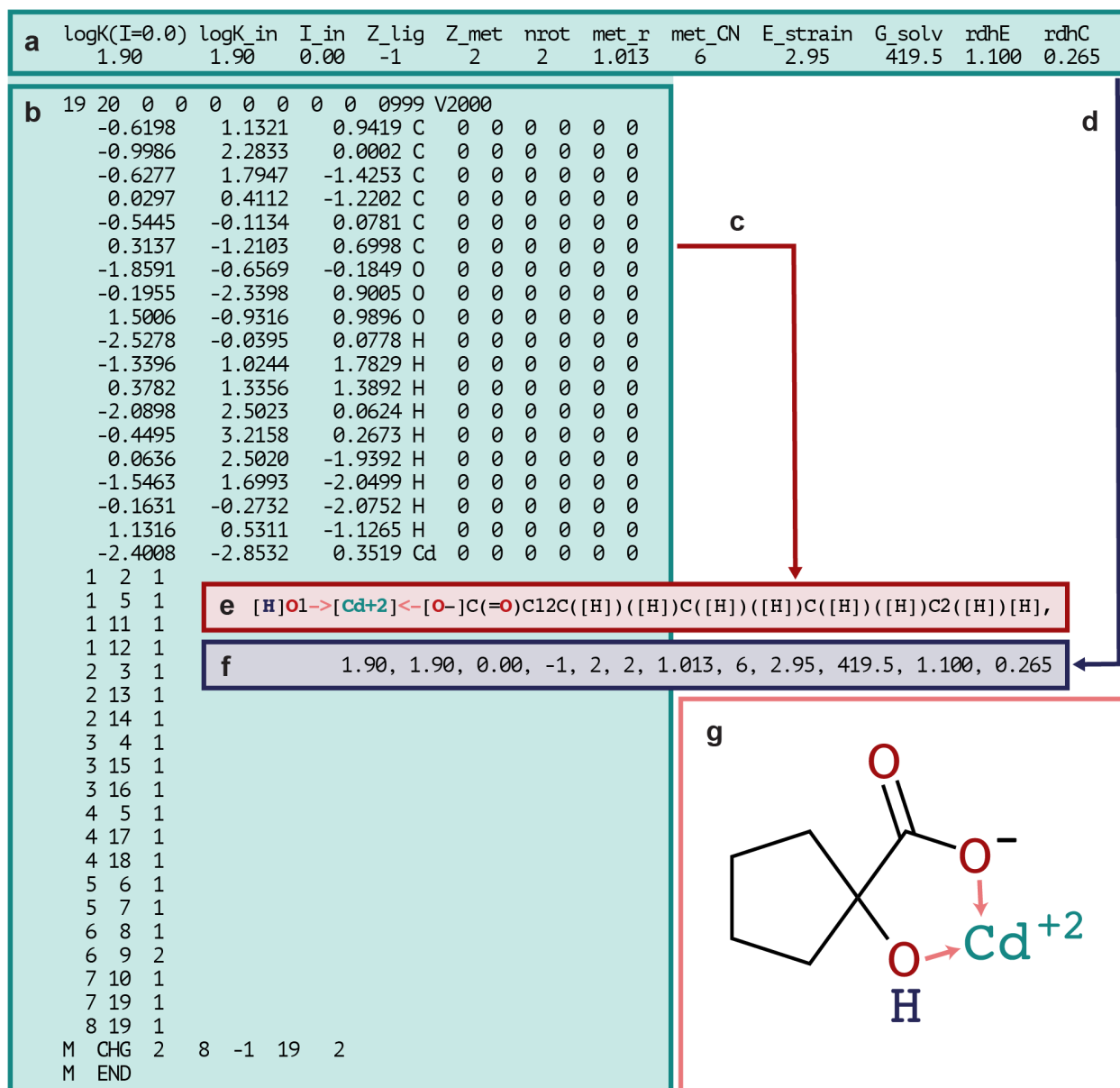


Figure 3. Explicit example of how the metal-ligand information flows from HostDesigner through LOGKPREDICT and back. HostDesigner generates a file named “logK_input” with (a) features and (b) coordinates. LOGKPREDICT converts the coordinates (c) and the features (d) into the respective comma separated value (csv) files with a SMILES string (e), and features (f) for each metal-ligand combination (g).

Once Chemprop has finished its prediction of log K, in less than 10 sec for most computer systems, the predicted log K is written to the “ML_ONE_preds.csv” file, and the control is returned to LOGKPREDICT, Figure 2. LOGKPREDICT subsequently transforms the log K format and writes to the file, “logk_output,” and returns the control back to HostDesigner. The HostDesigner manual¹¹ presents a detailed description of the use of LOGKPREDICT as well as sample cases.

III. DATA

Given that ML training relies heavily on data, all legitimate sources have been duly considered. The NIST and IUPAC databases are widely acknowledged as the foremost authoritative and dependable public sources for stability constants of metal-ligand complexes and have been selected and curated as follows for training the ML models. For this study three datasets derived from the NIST and IUPAC databases are used to train the ML models. For convenience, the datasets are subsequently referred to simply as “NIST”, “IUPAC Version 1”, and “IUPAC Version 2”. NIST includes a wide range of metal ions, whereas the IUPAC Versions 1 and 2 only include rare earth ions. As described in Section IV, the best performing model was obtained with NIST and thus, below it is described in greater detail than IUPAC Versions 1 and 2. Additional details on all three datasets are provided in Section II of the SI.

III.1 NIST Dataset

The National Institute of Science and Technology (NIST) provided the files for an SQL database modified from their software “NIST SRD 46. Critically Selected Stability Constants of Metal Complexes”³ that was based on the multi-volume book of Smith and Martel.¹ A subset of independently verified data from the SQL database was used for the DLNN training. The subset consists of ~1600 log K data points involving only mono- and bi-dentate ligands for which log K

values for at least 5 different metal ions are reported; up to 50 metal ions are reported for some ligands. The entire subset of data was cross-referenced with the Smith and Martell Critical Stability Constants volumes.¹ Figure 4 shows the types of ligands selected for the subset that was used for the ML training. Table 2 contains all of the metal ions from the database subset and the number of ligands binding to each metal ion.

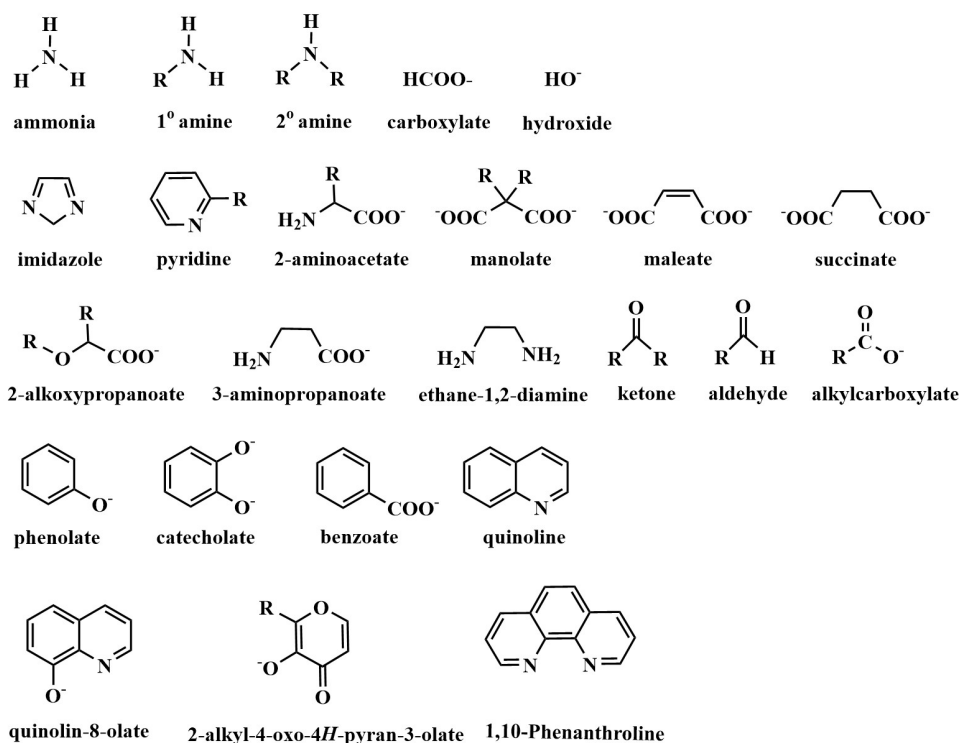


Figure 4. Examples of the ligand types from the NIST dataset.

Table 2. Breakdown of the number of ligands that are bound to a given metal ion in the contracted NIST dataset.

Metal	Charge	No. of Ligands	Metal	Charge	No. of Ligands
Li	1+	10	Cs	1+	2
Be	2+	33	Ba	2+	31

Na	1+	6	Hg	2+	21
Mg	2+	46	Pb	2+	47
Al	3+	22	La	3+	46
K	1+	6	Ce	3+	30
Ca	2+	47	Pr	3+	37
Sc	3+	25	Nd	3+	39
V	2+	33	Pm	3+	5
V	3+	2	Sm	3+	38
Cr	2+	13	Eu	3+	38
Cr	3+	20	Gd	3+	39
Mn	2+	57	Tb	3+	34
Fe	2+	37	Dy	3+	38
Fe	3+	36	Ho	3+	32
Co	2+	74	Er	3+	38
Ni	2+	82	Tm	3+	32
Cu	2+	83	Yb	3+	34
Zn	2+	80	Lu	3+	34
Ga	3+	20	Ac	3+	1
Rb	1+	1	Th	4+	31
Sr	2+	33	U	2+	47
Y	3+	35	U	4+	4
Pd	2+	14	Pu	2+	10
Ag	1+	31	Pu	4+	4
Cd	2+	70	Am	3+	6
In	3+	24	Cm	3+	4

III.2 Lanthanide-Only IUPAC Datasets

III.2.1 IUPAC Version 1

The 571 ligands from the Supplementary Information of the article by Chaube et al.¹⁰ was cross-matched with the IUPAC database by the chemical formulas and abridged molecular names. As a result, CAS numbers of the ligands were identified. The NIH and Pubchem-NIH websites^{54,55} provided the SMILES representation of the molecules by their CAS number. A Python script automated the look-up procedure, which was able to obtain the proper SMILES representations of 336 ligands. The IUPAC Version 1 database contains 4,144 log K metal-ligand values. A detailed list of the number of datapoints per metal ion and the ranges of log K values in this data set are found in Table S3.

III.2.2 IUPAC Version 2

The second version of the lanthanide-only IUPAC dataset was filtered and curated based on the procedure outlined in the SI Section II. Briefly, we retained 20 rare-earth cations including La³⁺, Ce³⁺, Ce⁴⁺, Pr³⁺, Nd³⁺, Pm³⁺, Sm³⁺, Eu²⁺, Eu³⁺, Gd³⁺, Tb³⁺, Dy³⁺, Ho³⁺, Er³⁺, Tm³⁺, Yb²⁺, Yb³⁺, Lu³⁺, Sc³⁺, and Y³⁺. We removed data that was not clearly or consistently labeled, data from experiments performed in non-aqueous media, and at non-standard temperature. The CAS number or IUPAC name of the ligand was utilized to extract the SMILES representation.⁵⁴ The resulting dataset is comprised of 5,096 data points, containing 481 distinct ligand molecules with molecular weights above 350 g/mol, Table S5.

IV. RESULTS AND DISCUSSION

In this section, we show the log K prediction quality by the root mean square error (RMSE), mean error (ME), and the coefficient of determination (R^2). Further, we consider an REE example showcasing the LOGKPREDICT capability to rank structures by selectivity.

IV.1 K-Fold Cross Validation

The K-fold cross validation (with a 5-fold split, i.e. $K=5$) performance of the D-MPNN+ model on the three databases is shown in Table 3. The mean errors are within the uncertainties of calculations and thus the predictions can be deemed unbiased. The NIST dataset gives the smallest RMSE. Technical details concerning hyper-parameter optimization and selection of properties for the D-MPNN+ model are in the SI Section I. Figure 5 shows the parity plot of experimental versus predicted log K values of the D-MPNN+ model on the test set of the NIST database.

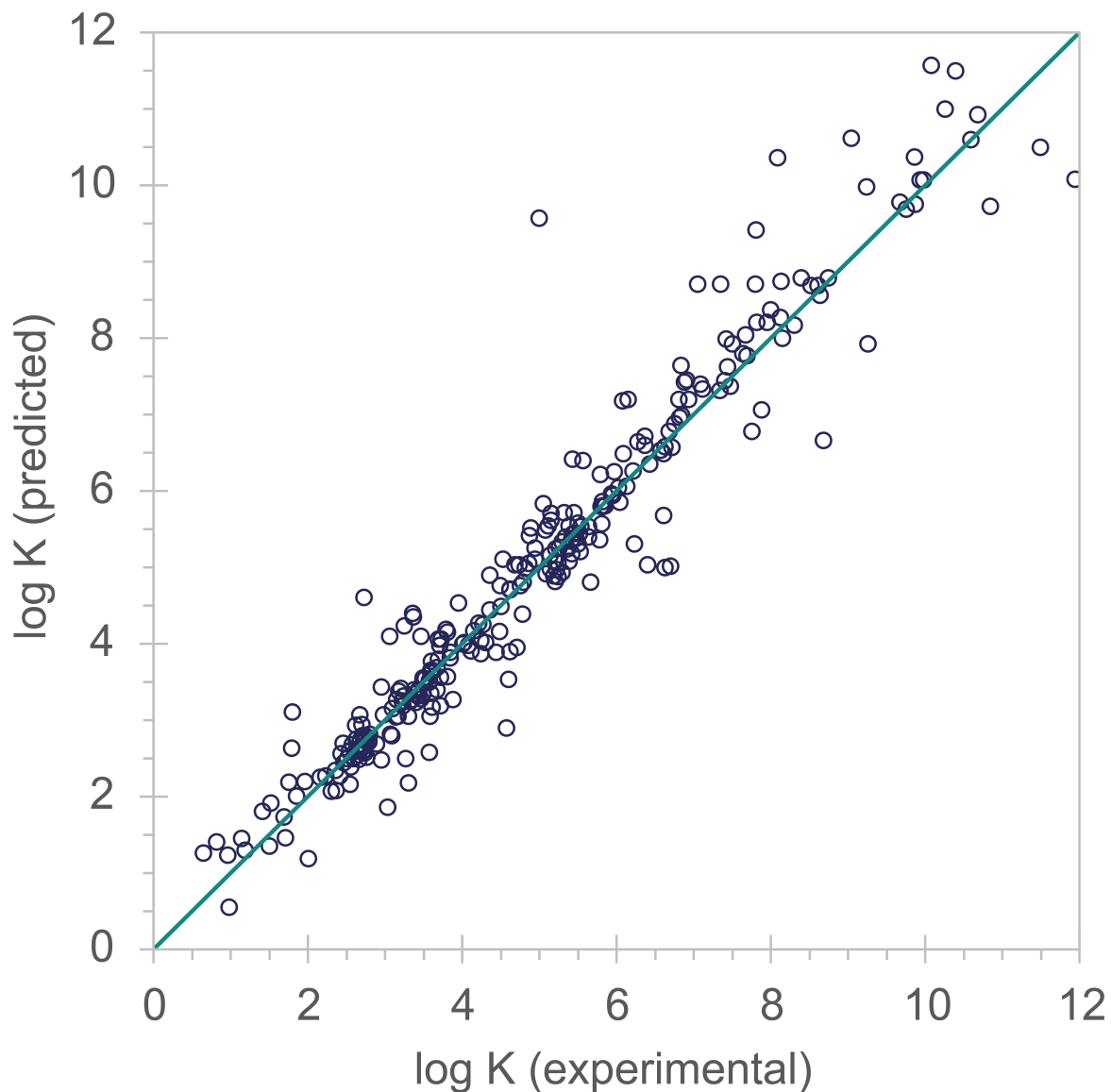


Figure 5. The experimental versus the predicted log K values parity plot using the D-MPNN+ model on the test set of the NIST database, with an added green diagonal line indicating exact predictions.

Table 3. Performance of the D-MPNN+ model on the three databases using 5-fold cross validation for the three datasets described in Section III, NIST, IUPAC Version 1, and IUPAC Version 2.

	NIST	IUPAC Version 1	IUPAC Version 2
RMSE	0.629 ± 0.044	0.764 ± 0.073	0.757 ± 0.071
ME	0.032 ± 0.045	-0.050 ± 0.046	0.006 ± 0.013
R ²	0.960 ± 0.006	0.976 ± 0.005	0.959 ± 0.007

Tables 4 and 5 show the 5-fold cross validation results of the multi-linear and K-nearest neighbors (KNN) regression, respectively, on the three databases. The Scikit-learn library⁵⁶ was used to obtain the results. Tables 4 and 5 also show the results for a full set of properties and with a reduced set of properties selected using the Boruta feature selection method.⁵⁷

Table 4. Performance of the multi-linear model on the three databases using 5-fold cross validation for the three datasets described in Section III, NIST, IUPAC Version 1, and IUPAC Version 2, using the full and reduced set of properties.

	NIST	IUPAC Version 1	IUPAC Version 2
<i>Full set of properties</i>			
RMSE	1.713	N/A ^a	4.606
ME	0.004	N/A ^a	0.690
R ²	0.850	N/A ^a	-0.003
<i>Reduced set of properties</i>			
RMSE	5.390	5.031	5.658
ME	-0.004	0.006	0.000
R ²	0.797	0.801	0.618

^aNot available due to multi-linear regression method of scikit-learn failing to produce results with more than 100 variables.

Table 5. Performance of the K-nearest neighbors (KNN) regression model on the three databases using K-fold cross validation for the three datasets described in Section III, NIST, IUPAC Version 1, and IUPAC Version 2, using the full and reduced set of properties.

	NIST	IUPAC Version 1	IUPAC Version 2
<i>Full set of properties</i>			
RMSE	2.216	1.154	1.647
ME	-0.157	-0.133	0.001
R ²	0.806	0.955	0.873
<i>Reduced set of properties</i>			
RMSE	1.562	1.040	1.647
ME	-0.112	-0.135	0.006
R ²	0.864	0.958	0.889

IV.2 Selectivity Study

To apply the D-MPNN and D-MPNN+ models for the study of metal ion selectivity, six ligands (1-OH-cPn-carboxylate, 2-OH-Me-propanoate, 3-OH-propanoate, 2-2pyr-imidazole, glycinate, and glycolate) with their metal ion counterparts were separated from the rest of the NIST database as a case study set. The remaining part of the dataset was used in training and validating models using both NIST with and without HostDesigner properties, D-MPNN+/NIST and D-MPNN/NIST, respectively.

The six ligands are chosen as illustrative cases for the metal ion selectivity study because they exhibit a variety of structural features and have data available for several different metals. The metal ions represented in the data for the ligands span different groups, alkali earth metals,

transition metals, rare-earth metals, actinides, and multiple oxidation states. In three of the ligands, carboxylate is the coordinating group with the metal in bidentate fashion, where the R groups of $R_2C(=O)(OH)$ are either a cyclic pentyl ring or two methyl groups or H-atoms. In that way we have shown the effect of different substitutions. In the glycinate ligand, the chelating groups are O^- and NH_2 , which represents the chelation of zwitterionic form of amino acids. The sixth ligand is a heterocyclic compound with imidazole bound to pyridine and acts as a bidentate ligand that coordinates with metal ions via the N-centers and shows the effect of aromatic heterocyclic groups on metal chelation. The resulting predictions for each of the six ligands are shown in Figure 6, with additional detail found in Table S2.

Both D-MPNN/NIST and D-MPNN+/NIST predict qualitatively accurate log K values for the six-ligand case, but MPNN+/NIST is noticeably better, which underlines once again the beneficial role of properties for accurate predictions. Table 6 quantifies the errors in terms of RMSEs and MEs of predictions with respect to the experimental results. Since the selectivity compares two different log Ks, relative log K errors as defined below are more relevant than absolute log K errors.

Our definition of 'relative RMSE log K error' is outlined as follows. There are instances where predicted log K values might closely follow the sequence of experimental log K values, but with a consistent bias. In these cases, one could adjust all predicted log K values with a constant shift to align them precisely with the experimental values. We define the relative log K error as the smallest RMSE achievable through a constant shift, effectively compensating for any bias. The optimal shift happens to coincide with the bias as computed by the mean error. Table 6 quantifies the errors of predictions both in terms of absolute and relative RMSEs with respect to the experimental results as well as the biases in terms of mean errors. The relative log K errors are

substantially lower than the corresponding absolute errors, with 0.31 versus 0.53 for D-MPNN/NIST and 0.25 versus 0.46 for D-MPNN+/NIST, after averaging over the six ligands.

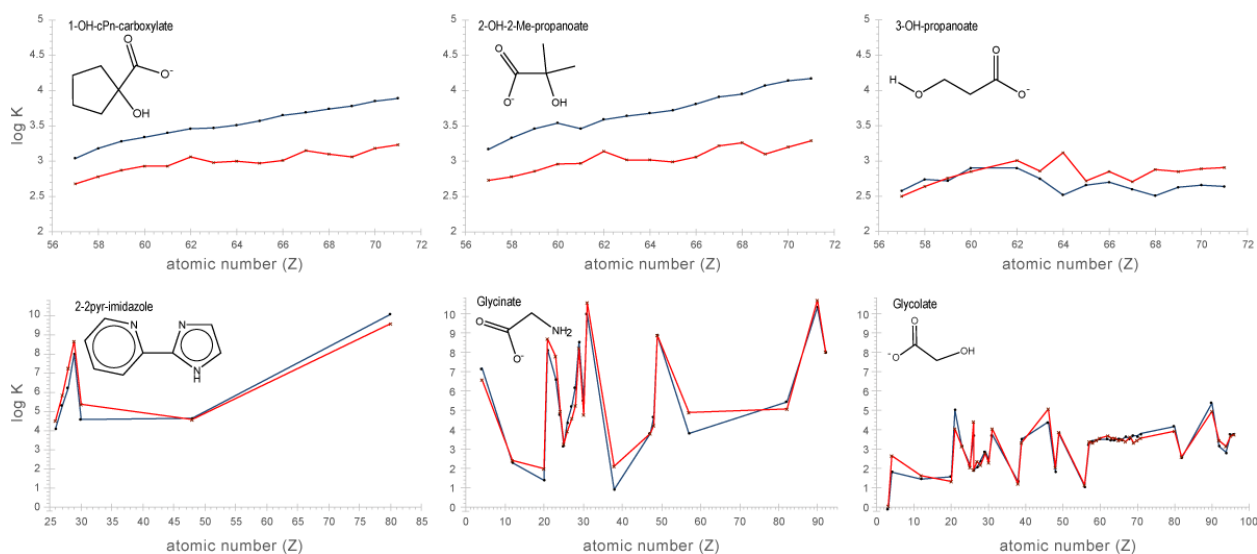


Figure 6. The D-MPNN+/NIST predicted (red) and experimental (blue) log K values are shown for each of the six ligands. The horizontal axis shows the atomic number (Z), and the vertical axis the log K.

Table 6. The ME, and RMSEs, absolute and relative, of predicted log K values using D-MPNN/NIST and D-MPNN+/NIST on the six ligands. Columns 1 through 6 correspond to 1-OH-cPn-carboxylate, 2-OH-Me-propanoate, 3-OH-propanoate, 2-2pyr-imidazole, glycinate, and glycolate, respectively. The last column shows the average error over the six ligands.

Model	Ligand						All
	1	2	3	4	5	6	
	<i>ME</i>						
D-MPNN	0.53	0.67	-0.14	-0.42	-0.06	-0.01	0.10
D-MPNN+	-0.59	-0.13	-0.21	-0.58	-0.23	0.00	-0.29

		<i>Absolute RMSE</i>					
D-MPNN	0.54	0.69	0.23	0.64	0.62	0.32	0.53
D- MPNN+	0.60	0.20	0.24	0.66	0.55	0.33	0.46
		<i>Relative RMSE</i>					
D-MPNN	0.12	0.16	0.18	0.49	0.61	0.32	0.31
D-MPNN+	0.08	0.15	0.12	0.31	0.50	0.33	0.25

IV.3 Ranking of Ligand Structures by Log K and Selectivity

HostDesigner with LOGKPREDICT has the capability of ranking ligands by selectivity. It first gets the predicted log K values for a target metal ion M_A , $\log K_A$ and the log K values for a user-selected competing metal ion M_B , $\log K_B$, and then ranks the candidates in decreasing $\log(K_{AB})$. To showcase LOGKPREDICT, new selective bidentate ligands are generated from N,N-dimethyl formamide and pyridine binding fragments linked by molecular fragments from the HostDesigner library. The target and competing metal ions are Gd and La, respectively. The detailed CAMD procedure is summarized in SI Section III.

The HostDesigner/LOGKPREDICT ranking of the top ten ligands by $\log K_{Gd}$ is provided in Table 7 and the corresponding molecular structures are shown in Figure 7. Notably, the first ranked structure, Figure 7-1, corresponds to the binding site of a recently published water-soluble ligand, of the BLPhen class (Figure 8).⁵⁸ Other ligands from the BLPhen class, which share the same binding site structure, bind to lanthanides in organic solvents.⁵⁹

Table 7. The top ten structures (Figure 7) ranked by $\log K_{Gd}$. Stability constant values are also computed in kcal/mol and shown in a separate column.

Rank	log K _{Gd}	ΔG ^a (kcal/mol)
1	4.27	-5.86
2	3.97	-5.45
3	3.92	-5.38
4	3.83	-5.26
5	3.79	-5.20
6	3.75	-5.15
7	3.74	-5.14
8	3.73	-5.12
9	3.63	-4.99
10	3.58	-4.92

^aUsing the approximation: $\Delta G \approx -2.303RT \times \log K_{Gd}$ where $R = 1.9872 \times 10^{-3} \frac{\text{kcal}}{\text{mol}\cdot\text{K}}$, $T = 298.15 \text{ K}$.

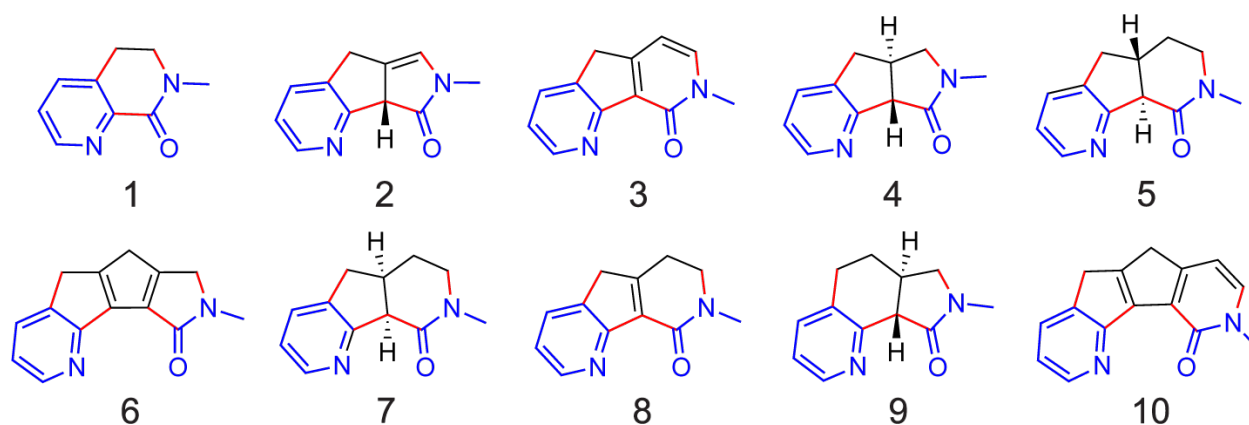


Figure 7. The top 10 ligands ranked by log K_{Gd}. The host fragments, pyridine and amide, are shown in blue. The HostDesigner generated bonds are shown in red and the linking fragments in black. The pyridine nitrogen and the oxygen of the amide are the donor atoms chelating with the Gd (III) ion to form the MLig complexes.

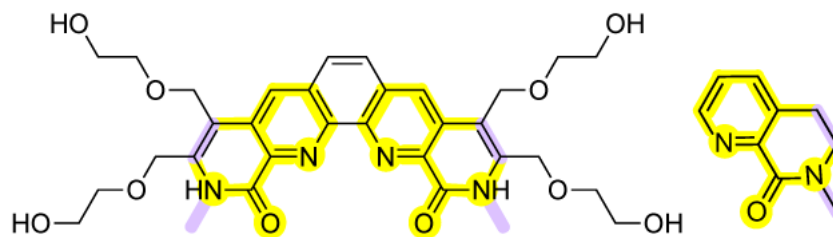


Figure 8. The top-ranking structure (yellow) is consistent with the RE binding site of the ligand (BLPhen, left) which was recently shown by Johnson et al.⁵⁸ to bind to lanthanides in aqueous solution.

The LOGKPREDICT selectivity for Gd over the La ions is shown in Table 8 and the ligand structures are in Figure 9. Note that the top Gd ligand (Figure 7-1) is ranked ten with respect to selectivity for Gd over La. Most of the top-ranking ligands that are selective for Gd over La have one additional carbon separating the pyridine nitrogen and the carbonyl carbon. Thus, we predict that a molecule analogous to BLPhen with an additional carbon between the pyridine and amide functional groups would show greater selectivity for Gd over La than BLPhen.

Table 8. A sample abridged form of the HostDesigner/LOGKPREDICT output. The top 10 ranking structures in terms of decreasing log of the selectivity coefficient (K_{GdLa}) for metal Gd versus metal La. Log K_{Gd} and log K_{La} correspond to metals Gd and La, respectively. The difference between the stability constant values is also computed in kcal/mol and presented, last column.

Rank	log K_{Gd}	log K_{La}	log (K_{GdLa})	$\Delta G_{Gd} - \Delta G_{La}^a$ (kcal/mol)
1	3.79	2.97	0.82	-1.13
2	3.56	2.75	0.81	-1.11

3	3.63	2.84	0.79	-1.08
4	3.74	2.96	0.78	-1.07
5	3.35	2.58	0.77	-1.06
6	3.43	2.66	0.77	-1.06
7	3.83	3.07	0.77	-1.06
8	3.47	2.72	0.76	-1.04
9	3.97	3.26	0.72	-0.99
10	4.27	3.59	0.69	-0.95

^aComputed using the approximation: $\Delta G_{Ln} \approx -2.303RT \times \log K_{Ln}$, where $R = 1.9872 \times 10^{-3} \frac{\text{kcal}}{\text{mol}\cdot\text{K}}$; $T = 298.15 \text{ K}$; and $Ln = \text{Gd, or La}$.

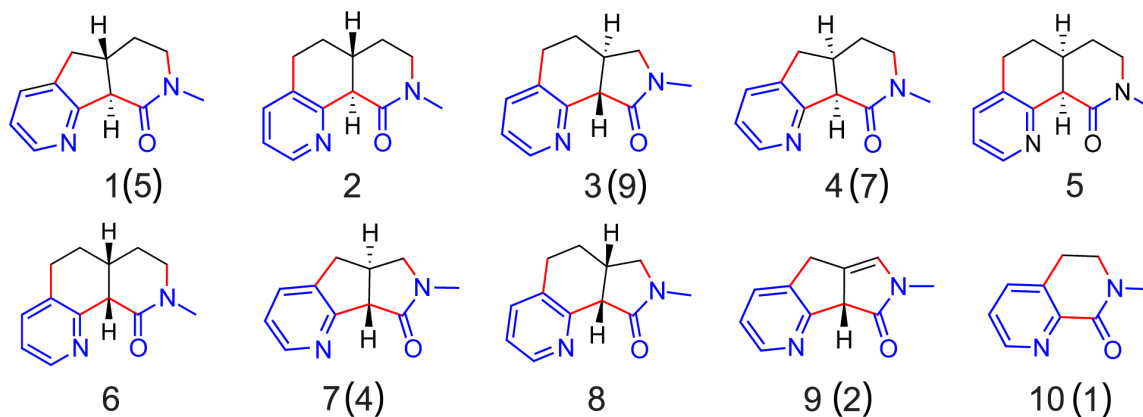


Figure 9. The top 10 ligands ranked by $\log K_{\text{GdLa}}$. The host fragments, pyridine and amide, are shown in blue. The HostDesigner generated bonds are shown in red and the linking fragments in black. The numbers in parentheses correspond to the ranking by $\log K_{\text{Gd}}$.

V. CONCLUSION

The newly created LOGKPREDICT software integrates the molecular design capabilities of HostDesigner with the machine learning (ML) functionalities of Chemprop. LOGKPREDICT

empowers HostDesigner to efficiently rank ligands based on the metal-ligand complex predicted log K values or selectivity coefficients.

The D-MPNN+ model demonstrates an exceptionally high level of accuracy using any of the datasets, NIST, IUPAC Version 1 and IUPAC Version 2. D-MPNN+/NIST has an RMSE of 0.629 ± 0.044 (R^2 of 0.960 ± 0.006). D-MPNN+/ IUPAC Versions 1 has an RMSE of 0.764 ± 0.073 (R^2 of 0.976 ± 0.005) and D-MPNN+/ IUPAC Versions 2 has an RMSE of 0.757 ± 0.071 (R^2 of 0.959 ± 0.007).

Six ligands are used to evaluate the metal ion selectivity of ligands. The six ligands are chosen for their structural diversity and the availability of data for various metals. Both D-MPNN/NIST and D-MPNN+/NIST predict log K values accurately, with D-MPNN+/NIST performing notably better, emphasizing the importance of properties for precision. D-MPNN+/NIST effectively captures the qualitative trends in log K values across various metal ions for each of the six ligand classes, achieving a notably low relative RMSE of just 0.25.

The performance of the LOGKPREDICT software is demonstrated by predicting the log K values for Gd and La. New bidentate ligands were generated using fragments from N,N-dimethyl formamide and pyridine, linked with molecular fragments from the HostDesigner library. Notably, the top-ranked ligand consistent with the binding site of BLPhen, recently demonstrated to experimentally bind to lanthanide in water.⁵⁸ Ligands predicted to have greater selectivity for Gd over La contain an additional carbon between the pyridine nitrogen and the carbonyl carbon, suggesting that a molecule similar to BLPhen, but with an extra carbon could show enhanced selectivity for Gd.

SUPPLEMENTARY MATERIAL

The supplementary material contains further technical details on the three data sets, the training parameters, and ligand design examples.

AUTHOR DECLARATIONS

The authors have no conflicts to disclose.

AUTHOR INFORMATION

Corresponding Authors

*Federico Zahariev, fzahari@ameslab.gov

*Marilú Pérez García, marilu@ameslab.gov

Author Contributions

All authors contributed to writing the manuscript and have given approval to the final version of the manuscript. ‡These authors contributed equally.

ACKNOWLEDGMENT

The authors thank Benjamin Hay for his contributions to curating, formatting, and generating properties for the data used in training the DLNN model, the integration of LOGKPREDICT with HostDesigner, as well as for useful discussions. The authors also thank Kyle Swanson for valuable suggestions on the use of Chemprop. This work was supported by the Critical Materials Innovation Hub funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable

Energy, Advanced Materials and Manufacturing Technology Office. Work at the Ames National Laboratory, which is operated for the U.S. DOE by Iowa State University was conducted under DOE contract # DE-AC02-07CH11358. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC awards ERCAP0017177, ERCAP0020144, and ERCAP0023063.

DATA AND SOFTWARE AVAILABILITY

The data that support the findings in this study are available within the article and on GitHub under the CMI Organization repositories.¹³

REFERENCES

- (1) Martell, A. E.; Smith, R. M. *Critical Stability Constants*; Springer: New York, 1974; Vol. 1–7.
- (2) Petitt, L. D.; Powell, K. J. IUPAC Stability Constant Database (SC-Database), 2016.
- (3) 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>.
- (4) Hancock, R. D.; Marsicano, F. The Chelate Effect: A Simple Quantitative Approach. *J. Chem. Soc. Dalton Trans.* **1976**, No. 12, 1096. <https://doi.org/10.1039/dt9760001096>.
- (5) Hancock, R. D.; Marsicano, F. Parametric Correlation of Formation Constants in Aqueous Solution. 1. Ligands with Small Donor Atoms. *Inorg. Chem.* **1978**, *17* (3), 560–564. <https://doi.org/10.1021/ic50181a009>.
- (6) Carbonaro, R. F.; Di Toro, D. M. Linear Free Energy Relationships for Metal–Ligand Complexation: Monodentate Binding to Negatively-Charged Oxygen Donor Atoms. *Geochim. Cosmochim. Acta* **2007**, *71* (16), 3958–3968. <https://doi.org/10.1016/j.gca.2007.06.005>.
- (7) Carbonaro, R. F.; Atalay, Y. B.; Di Toro, D. M. Linear Free Energy Relationships for Metal–Ligand Complexation: Bidentate Binding to Negatively-Charged Oxygen Donor Atoms. *Geochim. Cosmochim. Acta* **2011**, *75* (9), 2499–2511. <https://doi.org/10.1016/j.gca.2011.02.027>.
- (8) Vukovic, S.; Hay, B. P.; Bryantsev, V. S. Predicting Stability Constants for Uranyl Complexes Using Density Functional Theory. *Inorg. Chem.* **2015**, *54* (8), 3995–4001. <https://doi.org/10.1021/acs.inorgchem.5b00264>.
- (9) Poole, D.; Zahariev, F.; Del Viscio, M.; Windus, T. L.; Pérez García, M. Computational Modeling of Diphosphine Oxide and Diglycolamide Ligand Complexation to Lanthanides

- and Extraction from Acidic Media. In *Computational Science in Lanthanide and Actinide Chemistry*; ACS Symposium Series; ACS Books: Washington, D.C., 2021; Vol. 1388, p 22.
- (10) 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>.
 - (11) Hay, B. P. HostDesigner, 2021. <https://sourceforge.net/projects/hostdesigner/>.
 - (12) Chemprop, 2022. <https://github.com/chemprop/chemprop>.
 - (13) Zahariev, F.; Ash, T.; Pérez García, M. LOGKPREDICT, 2021. <https://github.com/Critical-Materials-Institute/LOGKPREDICT> (accessed 2023-05-30).
 - (14) 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>.
 - (15) Cheisson, T.; Schelter, E. J. Rare Earth Elements: Mendeleev’s Bane, Modern Marvels. *Science* **2019**, *363* (6426), 489–493. <https://doi.org/10.1126/science.aau7628>.
 - (16) García Alejo, A.; De Silva, N.; Liu, Y.; Windus, T. L.; Pérez García, M. Solvent Phase Optimizations Improve Correlations with Experimental Stability Constants for Aqueous Lanthanide Complexes. *Solvent Extr. Ion Exch.* **2023**, 1–11. <https://doi.org/10.1080/07366299.2022.2160646>.
 - (17) Borst, A. M.; Smith, M. P.; Finch, A. A.; Estrade, G.; Villanova-de-Benavent, C.; Nason, P.; Marquis, E.; Horsburgh, N. J.; Goodenough, K. M.; Xu, C.; Kynický, J.; Geraki, K. Adsorption of Rare Earth Elements in Regolith-Hosted Clay Deposits. *Nat. Commun.* **2020**, *11* (1), 4386. <https://doi.org/10.1038/s41467-020-17801-5>.
 - (18) Moldoveanu, G. A.; Papangelakis, V. G. Recovery of Rare Earth Elements Adsorbed on Clay Minerals: I. Desorption Mechanism. *Hydrometallurgy* **2012**, *117–118*, 71–78. <https://doi.org/10.1016/j.hydromet.2012.02.007>.
 - (19) Abbott, A. N.; Löhr, S.; Trethewy, M. Are Clay Minerals the Primary Control on the Oceanic Rare Earth Element Budget? *Front. Mar. Sci.* **2019**, *6*, 504. <https://doi.org/10.3389/fmars.2019.00504>.
 - (20) Feng, X.; Onel, O.; Council-Troche, M.; Noble, A.; Yoon, R.-H.; Morris, J. R. A Study of Rare Earth Ion-Adsorption Clays: The Speciation of Rare Earth Elements on Kaolinite at Basic pH. *Appl. Clay Sci.* **2021**, *201*, 105920. <https://doi.org/10.1016/j.clay.2020.105920>.
 - (21) Grasso, V. B. *Rare Earth Elements in National Defense: Background, Oversight Issues, and Options for Congress*; Congressional Report; Library of Congress Washington DC Congressional Research Service: Fort Belvoir, VA, 2013. <https://apps.dtic.mil/sti/citations/ADA590410>.
 - (22) Nassar, N. T.; Brainard, J.; Gulley, A.; Manley, R.; Matos, G.; Lederer, G.; Bird, L. R.; Pineault, D.; Alonso, E.; Gambogi, J.; Fortier, S. M. Evaluating the Mineral Commodity Supply Risk of the U.S. Manufacturing Sector. *Sci. Adv.* **2020**, *6* (8), eaay8647. <https://doi.org/10.1126/sciadv.aay8647>.
 - (23) Bauer, D.; Diamond, D.; Li, J.; McKittrick, M.; Sandalow, D.; Telleen, P. *U.S. Department of Energy Critical Materials Strategy*; 2; U.S. Department of Energy: Washington, DC, 2011. <https://www.energy.gov/node/349057>.
 - (24) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the 2017 list of critical raw materials for the EU.* <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017DC0490> (accessed 2023-02-10).

- (25) El Maangar, A.; Theisen, J.; Penisson, C.; Zemb, T.; Gabriel, J.-C. P. A Microfluidic Study of Synergic Liquid–Liquid Extraction of Rare Earth Elements. *Phys. Chem. Chem. Phys.* **2020**, *22* (10), 5449–5462. <https://doi.org/10.1039/C9CP06569E>.
- (26) Okamura, H.; Hirayama, N. Recent Progress in Ionic Liquid Extraction for the Separation of Rare Earth Elements. *Anal. Sci.* **2021**, *37* (1), 119–130. <https://doi.org/10.2116/analsci.20SAR11>.
- (27) Arrachart, G.; Couturier, J.; Dourdain, S.; Levard, C.; Pellet-Rostaing, S. Recovery of Rare Earth Elements (REEs) Using Ionic Solvents. *Processes* **2021**, *9* (7), 1202. <https://doi.org/10.3390/pr9071202>.
- (28) Dewulf, B.; Riaño, S.; Binnemans, K. Separation of Heavy Rare-Earth Elements by Non-Aqueous Solvent Extraction: Flowsheet Development and Mixer-Settler Tests. *Sep. Purif. Technol.* **2022**, *290*, 120882. <https://doi.org/10.1016/j.seppur.2022.120882>.
- (29) Steed, J. W.; Atwood, J. L. *Supramolecular Chemistry*, 2nd ed.; Wiley: Chichester, UK, 2009.
- (30) Schneider, H.-J.; Yatsimirsky, A. K. *Principles and Methods in Supramolecular Chemistry*; Wiley: New York, 2000.
- (31) Hudlicky, T.; Reed, J. W. *The Way of Synthesis: Evolution of Design and Methods for Natural Products*; Wiley-VCH: Weinheim, 2007.
- (32) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discov.* **2010**, *9* (4), 273–276. <https://doi.org/10.1038/nrd3139>.
- (33) Austin, N. D.; Sahinidis, N. V.; Trahan, D. W. Computer-Aided Molecular Design: An Introduction and Review of Tools, Applications, and Solution Techniques. *Chem. Eng. Res. Des.* **2016**, *116*, 2–26. <https://doi.org/10.1016/j.cherd.2016.10.014>.
- (34) Cram, D. J.; Lein, G. M. Host-Guest Complexation. 36. Spherand and Lithium and Sodium Ion Complexation Rates and Equilibria. *J. Am. Chem. Soc.* **1985**, *107* (12), 3657–3668. <https://doi.org/10.1021/ja00298a041>.
- (35) Chemical Foundations for the Understanding of Natural Macrocyclic Complexes. In *Bioinorganic Chemistry*; Busch, D. H., Farmery, K., Goedken, V., Katovic, V., Melnyk, A. C., Sperati, C. R., Tokel, N., Eds.; Advances in Chemistry; American Chemical Society: Washington, D.C., 1971; Vol. 100, pp 44–78. <https://doi.org/10.1021/ba-1971-0100>.
- (36) McDougall, G. J.; Hancock, R. D.; Boeyens, J. C. A. Empirical Force-Field Calculations of Strain-Energy Contributions to the Thermodynamics of Complex Formation. Part 1. The Difference in Stability between Complexes Containing Five- and Six-Membered Chelate Rings. *J. Chem. Soc. Dalton Trans.* **1978**, No. 11, 1438. <https://doi.org/10.1039/dt9780001438>.
- (37) Anichini, A.; Fabbri, L.; Paoletti, P.; Clay, R. M. A Microcalorimetric Study of the Macrocyclic Effect. Enthalpies of Formation of Copper(II) and Zinc(II) Complexes with Some Tetra-Aza Macrocyclic Ligands in Aqueous Solution. *J. Chem. Soc. Dalton Trans.* **1978**, No. 6, 577. <https://doi.org/10.1039/dt9780000577>.
- (38) Stack, T. D. P.; Hou, Z.; Raymond, K. N. Rational Reduction of the Conformational Space of a Siderophore Analog through Nonbonded Interactions: The Role of Entropy in Enterobactin. *J. Am. Chem. Soc.* **1993**, *115* (14), 6466–6467. <https://doi.org/10.1021/ja00067a094>.
- (39) Hay, B. P.; Firman, T. K. HostDesigner: A Program for the de Novo Structure-Based Design of Molecular Receptors with Binding Sites That Complement Metal Ion Guests. *Inorg. Chem.* **2002**, *41* (21), 5502–5512. <https://doi.org/10.1021/ic0202920>.

- (40) Hay, B. P.; Jia, C.; Nadas, J. Computer-Aided Design of Host Molecules for Recognition of Organic Guests. *Comput. Theor. Chem.* **2014**, *1028*, 72–80. <https://doi.org/10.1016/j.comptc.2013.12.003>.
- (41) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244–255. <https://doi.org/10.1021/ci00007a012>.
- (42) Young, N. J.; Hay, B. P. Structural Design Principles for Self-Assembled Coordination Polygons and Polyhedra. *Chem Commun* **2013**, *49* (14), 1354–1379. <https://doi.org/10.1039/C2CC37776D>.
- (43) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; Precup, D., Teh, Y. W., Eds.; Proceedings of Machine Learning Research; PMLR, 2017; Vol. 70, pp 1263–1272.
- (44) Dai, H.; Dai, B.; Song, L. Discriminative Embeddings of Latent Variable Models for Structured Data. In *Proceedings of The 33rd International Conference on Machine Learning*; Balcan, M. F., Weinberger, K. Q., Eds.; Proceedings of Machine Learning Research; PMLR: New York, New York, USA, 2016; Vol. 48, pp 2702–2711.
- (45) Mahé, P.; Ueda, N.; Akutsu, T.; Perret, J.-L.; Vert, J.-P. Extensions of Marginalized Graph Kernels. In *Twenty-first International Conference on Machine Learning - ICML '04*; ACM Press: Banff, Alberta, Canada, 2004; p 70. <https://doi.org/10.1145/1015330.1015446>.
- (46) 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.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>.
- (47) Shannon, R. D. Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides. *Acta Crystallogr. Sect. A* **1976**, *32* (5), 751–767. <https://doi.org/10.1107/S0567739476001551>.
- (48) Marcus, Y. Thermodynamics of Solvation of Ions. Part 5.—Gibbs Free Energy of Hydration at 298.15 K. *J Chem Soc Faraday Trans* **1991**, *87* (18), 2995–2999. <https://doi.org/10.1039/FT9918702995>.
- (49) Davies, C. W. *Ion Association*; Butterworths, 1962.
- (50) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72* (2), 171–179. <https://doi.org/10.1107/S2052520616003954>.
- (51) Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (3), 237–243. <https://doi.org/10.1021/ci00067a005>.
- (52) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36. <https://doi.org/10.1021/ci00057a005>.
- (53) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101. <https://doi.org/10.1021/ci00062a008>.
- (54) Nicklaus, M. *CADD Group Chemoinformatics Tools and User Services*. <https://cactus.nci.nih.gov/index.html> (accessed 2023-01-15).

- (55) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373–D1380. <https://doi.org/10.1093/nar/gkac956>.
- (56) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (57) Kursu, M. B.; Rudnicki, W. R. Feature Selection with the **Boruta** Package. *J. Stat. Softw.* **2010**, *36* (11). <https://doi.org/10.18637/jss.v036.i11>.
- (58) Johnson, K. R.; Driscoll, D. M.; Damron, J. T.; Ivanov, A. S.; Jansone-Popova, S. Size Selective Ligand Tug of War Strategy to Separate Rare Earth Elements. *JACS Au* **2023**, *3* (2), 584–591. <https://doi.org/10.1021/jacsau.2c00671>.
- (59) Healy, M. R.; Ivanov, A. S.; Karslyan, Y.; Bryantsev, V. S.; Moyer, B. A.; Jansone-Popova, S. Efficient Separation of Light Lanthanides(III) by Using Bis-Lactam Phenanthroline Ligands. *Chem. – Eur. J.* **2019**, *25* (25), 6326–6331. <https://doi.org/10.1002/chem.201806443>.