Fast and Accurate log K Prediction

Progress Report for Period 7/1/2020 thru 11/30/2020

A Letter Report of Work Performed Under the Auspices of:

The Critical Materials Institute, an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Advanced Manufacturing Office.

Benjamin P. Hay

Principal Investigator, Supramolecular Design Institute, 127 Chestnut Hill Road, Oak Ridge, TN 37830-7185, Phone: 865-481-3237, Email: hayben@comcast.net

November 30, 2020

Submitted to Dr. Marilu Perez Garcia, Ames Laboratory in accordance with Subcontract No. SC-20-528, Appendix A. This report documents work performed to accomplish Year 8, Q9 and Q10 Progress Measures.

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Ames Laboratory, nor the Supramolecular Design Institute, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, product, or process disclosed. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Ames Laboratory, or the Supramolecular Design Institute. The views and opinions of the author expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Executive Summary

Since the inception of the Critical Materials Institute in June 2013, the overarching objective at the Supramolecular Design Institute has been the creation of improved computer-aided molecular design software to facilitate the development of effective ligands for use in metal ion sequestration, separation, and purification. In Phase 1 the primary objective was to generalize and expand the molecule building algorithms to allow a wider range of input structures and build a more complete set of possible structures. This objective was achieved with the publication of HostDesigner, Version 4.0. In Phase 2, the primary objective has shifted to a new focus, the development of machine learning software to predict aqueous phase formation constants for metal-ligand complexes that exists in aqueous solution. This novel capability will allow the prediction of a ligand's affinity and selectivity for a targeted metal ion, key properties for the development of technologies used to obtain, recycle, and purify metals.

This letter report documents work performed at the Supramolecular Design Institute in support of development and testing of machine learning software being performed at Ames Laboratory. The primary goal over this reporting period was to create a training data set that contains the following information for each entry: log K value, 3D description for the MM3-optimized metal-ligand complex, and user-defined descriptors. A training data set that contains entries for 1,662 complexes between 85 representative unidentate and bidentate ligands with up to as many as 50 different metal ion species was provided to Ames Laboratory. A secondary goal was to define the mechanism and the data format that will be used to pass information between HostDesigner and the Ames code. The mechanism and format, defined in collaboration with Ames Laboratory collaborators, are described.

Table of Contents

Table of Figures

Table of Tables

1.0 Introduction

The design molecular architecture with desired physical and chemical properties is a challenge that cuts across scientific disciplines. This is particularly true in supramolecular chemistry [1,2], where ion and molecular recognition by organic substrates are topics of fundamental importance. For example, effective control of host-guest binding affinity is critical in the development of separating agents (liquid-liquid extraction, ion exchange), improved analytical techniques (separation and concentration of dilute analytes), sensors (detection of species in groundwater and chemical process streams), medicine (drug design), and homogenous catalysts (control of selectivity and activity). Molecular recognition plays a key role in controlling the self-assembly of nanoscale structures. Here the underlying hypothesis that the morphology of nanoscale assemblies can be controlled and directed at the molecular-level through the design of building blocks that are shaped and functionalized to recognize one another in specific ways.

With the tremendous advances in computational hardware, computer-aided molecular design represents a promising approach for addressing these challenges. Computational chemistry has advanced to the point where it is possible to calculate accurate structural and energetic features of molecules and molecular complexes. Although computational methods exist for the prediction of both molecular structures and their chemical properties prior to synthesis, fully automated computer-aided design requires additional methodology – a technique for generating the molecules that will be evaluated. Outside the pharmaceutical industry, general-purpose design software for the automated generation and evaluation of 3D molecular structures did not exist.

Given the size and complexity of molecular space, building molecular structures by covalently linking sets of disconnected molecular fragments in three dimensions is not a trivial exercise. Generating trial structures by hand with a graphical user interface is very time-consuming. It is not readily obvious which linkage structures might be best used to connect the fragments. Researchers often default to linkages chosen for their ease of synthesis rather than for the molecular architectures that will result. To attain desired chemical or physical properties one needs a way to go beyond just informed guess or chemical intuition and escape the cycle of trial-and-error research.

Extensive research in the field of drug design has led to the development of algorithms for rapidly constructing and evaluating large numbers of novel molecules on the computer. A common approach, called *de novo* structure-based design, has been implemented in numerous drug discovery codes [3-16]. The term *de novo* indicates that new molecules are constructed by covalently linking molecular fragments to one another. The term *structure-based* refers to the fact that this method requires knowledge of the 3D structure of the protein binding site and the 3D structural information regarding the docking of fragments within the binding site. The design process involves two main steps: (i) constructing candidate molecular geometries and (ii) scoring the candidates to rank-order them with respect to desired properties.

In principle, the *de novo* structure-based design approach is not limited to drug design. Any time it is possible to define the relative orientation of two bonding vectors, this approach can be used to identify linkages that span them. For example, this method should readily be applied to the design of host molecules, for example, a sequestering agent for a rare earth metal ion. Unfortunately, computer programs developed to perform *de novo* structure-based drug design

[3-16] are not generally applicable in chemistry. Drug design programs require input of atomic coordinates for a protein-binding site and are highly specialized to address protein-organic interactions. Thus, although computer hardware has developed rapidly over the past two decades, a lack of computer software has prevented the application of computer-aided molecular design methods to chemical and material science challenges outside the pharmaceutical industry. Adapting *de novo* structure-based design methods for more general application would provide a pathway to accelerated discovery and go a long way toward eliminating the costly Edisonian research that dominates chemistry and material science today.

The author began to address this issue in 2000 with the development of software that became known as HostDesigner (HD). Although originally developed for the design of metal ion hosts [17-21], this software has been adapted to handle a wide range of host-guest interactions and has been used successfully in the design of anion hosts [22-25], receptors for small organic molecules [26], and components that direct the formation of high-symmetry molecular assemblies [27-29]. HD is designed to cover a large region of molecule space in a very efficient manner. In the time that it would take someone to create just one molecule with a molecular graphics interface, HD is capable of generating and evaluating millions of 3D molecular structures. For example, in a recent test run HD was observed to evaluate and rank 8,193,542 molecular geometries in just 61.2 seconds, a rate of one geometry every 0.0000075 seconds (MacPro, 2.8 GHz Intel Xeon processor). The latest version of HD can be downloaded at no cost from the https://sourceforge.net/projects/hostdesigner/ website.

In Phase 1 of this project (2013-2018) efforts were focused on the first step of the design process, in other words, on generalizing and enhancing the algorithms used to build 3D molecular structures. Major improvements to HD included the capabilities (i) to accept any organic molecular fragment as an input structure and (ii) to connect fragments by fusing atoms or bonds [30]. In Phase 2, the objective has shifted to the second step of the design process – improving the scoring methods. Specifically, the objective focuses on improving the scoring methods that are used to rank-order organic ligands that bind metal ions.

HD currently supports two scoring methods. The first method, which is always applied by default, uses geometric parameters and conformational energy increments to estimate how well organized the host is for binding the guest [17-21]. Although approximate in nature, this method provides a very rapid means of identifying potential candidates from a large group of structures. The best candidates can then be re-prioritized using more accurate evaluation methods to estimate relative binding affinity. At this time the second method involves the use of molecular mechanics models to rank the candidates based on intra-ligand steric strain. This computational method has been demonstrated to correlate the relative binding affinity for a single metal ion with series of ligands bearing a constant set of donor atoms [31-38], which describes the series of structures generated by each individual HostDesigner run.

The Phase 2 goal is to address two limitations associated with the current molecular mechanics scoring methods. The default molecular mechanics model used by HostDesigner is known as MM3, developed by N. L. Allinger [39,40] and the most accurate force field for organic compounds. The first limitation is that the molecular mechanics parameters needed to perform the calculations are often not available for the system under design. This necessitates parameter development activities before MM3 calculations can be done. Missing MM3 parameters represents a significant bottleneck that historically had added ≥ 6 months of delay to design campaigns. For this reason, significant effort was expended to facilitate this process in Phase 1 of the project and two approaches were developed to assign parameters for organic molecules.

In the first approach to parameter development, empirical algorithms that were implemented in the original MM3 software have now been implemented in the MM code distributed with HD, MENGINE. These algorithms apply a simple set of rules to estimate missing MM3 parameters for interactions that occur in organic molecules [41]. At this time, the MENGINE code will estimate all missing parameters for any organic molecule, always allowing a calculation to proceed. When this happens, the code outputs a file with a list of all estimated parameters, providing an excellent starting point for the application of the second approach.

The second approach to parameter development involves the time-intensive process of generating, collecting, and fitting data obtained from QM calculations. In order to automate and streamline this process, the ParFit software was developed at Ames Laboratory [42]. ParFit can generate QM computation input files for multiple jobs, pull the information from the output files to generate the QM energy profiles, and fit these QM energy profiles via simultaneous adjustment of multiple MM parameters, representing a savings of at least an order of magnitude over having to do all of this by hand. In addition, ParFit has been developed with multiple optimization methods for both local and global optimization using the same MENGINE back-end software for MM calculations and uses constraints based on symmetry and physical reasonableness to obtain the best parameter fits using the QM data. ParFit can be downloaded at no cost from the https://github.com/fzahari/ParFit website.

Although the assignment of MM3 parameters for interactions that occur within organic molecules was addressed in Phase 1, the assignment of MM3 parameters for interactions found in metal complexes remained problematic. The empirical algorithms used for organic interactions are not generally applicable to interactions involving metal ions and the automated application of QM to generate energy profiles for metal ions has not been generalized. For these reasons, at the onset of Phase 2 it remained necessary to invest the effort to develop the handful of metal-dependent parameters required to apply the MM3 model to a metal-ligand complex. Thus, one of the Phase 2 objectives was to develop algorithms that generalize and automate the assignment of metal-dependent MM3 parameters needed to model metal complexes that occur in aerobic, aqueous solution with common ligand functional groups.

This objective was realized in May 2020 after the MM3 molecular mechanics model was successfully extended to handle aqueous metal complexes with ligands containing N donor groups, such as amines, azines, azoles, and imines, and/or O donor groups such as water, alcohols, ethers, aldehydes, ketones, amides, phosphine oxides, sulfoxides, alkoxides, phenoxides, pyridine N-oxides, carboxylates, and nitrate. Using empirical relationships and crystal structure data, algorithms were developed to generate the requisite metal-ligand MM3 parameters using simple correlations with metal ion radii. These algorithms were implemented within the MENGINE software to automate parameter assignment for these types of metal complexes. The extended MM3 model performance was validated by comparison of optimized versus experimental geometries for 17,426 metal complexes taken from the CSD.   On average the model reproduces bond lengths to ± 0.026 Å, bond angles to ± 2.2°, and dihedral angles to ± 4.8° [43]. This upgraded version of MENGINE is available at no cost as part of the HostDesigner, Version 4.2 download https://sourceforge.net/projects/hostdesigner/ .

Having significantly reduced the first limitation associated with the use of MM3-based scoring methods, attention was next directed toward eradicating the second limitation – the inability to predict selectivity. Selectivity prediction requires the ability to predict absolute log K values for the same ligand with two or more metal ions. Unfortunately, the current MM3 scoring method uses ligand strain energies to predict relative, rather than absolute, binding affinity. Although well-suited for identifying complementary architectures for hosts bearing fixed sets of donor groups, the current scoring method does not allow comparison of ligands that contain different sets of donor groups nor does it allow for comparison of the binding affinity for one metal ion over another one, in other words, selectivity. What is needed is a fast and accurate method for predicting the absolute binding affinity, for example, the aqueous formation constant, for the complex formed between any ligand and any metal ion.

During the period from 1950 through 1990, a large number of aqueous metal-ligand formation constants, in other words, log K values for the equilibrium M + L <=> M–L, were experimentally determined. This data has been critically reviewed and is now available in the form of a six volume set named Critical Stability Constants [44], an electronic database from the National Institute of Standards [45], and an electronic database from the International Union of Pure and Applied Chemistry [46]. As the volume of data grew, scientists identified ligand and metal properties that controlled the strength of the metal-ligand interaction. Because it is not possible to measure the formation constant for every metal-ligand combination, there were many attempts to develop quantitative correlations to allow the prediction of unknown log K values. This research effort had largely ended by the mid-1990's.

Methods for the prediction of log K values using various empirical relationships have been reviewed [47-51]. The majority of these relationships use known log K data to predict unknown log K data. This allows the estimation of unknown log K values when sufficient data exists for related metal complexes. For series of metal ions with one ligand, experimental log K values are in some cases correlated with physical properties of the metal ion. A variety of functional dependencies have been observed with metal properties such as the charge, ionic radius, electronegativity, and ionization potential [52-56]. Linear correlations between experimental log K values for one metal with a series of ligand structures have been obtained using descriptors such as the number and type of ligand donor atom, the number of chelate rings, and the size of the chelate rings [57,58]. The MM3-based scoring method used by HostDesigner [17-21] is based on two descriptors associated with the chelation event - the ligand strain energy [32-38] and the number of restricted rotatable bonds [59-63]. One of the more impressive relationships involved the extension of a dual basicity scale, with three parameters for each metal ion and three parameters for each ligand where log K values for 1:1 complexes formed between 35 metal ions and 19 unidentate ligands were predicted to within an accuracy of ± 0.2 log units [49,50].

Although such studies have identified chemical properties that influence the strength of the metal-ligand interaction, *a priori* predictions of unknown log K values in the absence of related thermodynamic data are not yet possible. Thus, one of the goals of this project is to develop a method that will allow us to predict the aqueous log K value for a metal-ligand complex. Building upon tremendous advances in cheminformatics and computational science over the past three decades, machine learning methods have recently emerged as the logical next approach toward making such predictions. Coupling large thermodynamic property data sets with molecular descriptors common in traditional QSAR models and/or learned molecular representations using graph convolutions, machine learning methods have recently been investigated for their ability to predict thermodynamic properties such as water solubility, hydration free energy, octanol/water distribution coefficients, and protein binding affinities [64].

A key objective in this project is to evaluate whether analogous machine learning approaches can be used to predict metal-ligand log K values. After extensive review and preliminary testing of several possible machine learning frameworks, collaborators at Ames Laboratory selected the open source code named CHEMPROP [65] as the best machine learning software for this application. This code generates its own set of 2D graphic descriptors for each molecule requiring only a SMILES string description for each metal-ligand complex. CHEMPROP also has the ability to combine these 2D graphic descriptors with user-defined descriptors, providing a pathway to improve log K prediction accuracy by adding user-defined descriptors known to correlate subsets of log K data.

This letter report documents work performed at the Supramolecular Design Institute to support CHEMPROP modifications and testing currently underway at Ames Laboratory. An initial goal was to create a training data set that contains the following information for each entry: log K value, 3D description for the MM3-optimized metal-ligand complex, and user-defined descriptors. A training data set that contains entries for 1,662 complexes between 85 representative unidentate and bidentate ligands with up to as many as 50 different metal ion species was provided to Ames Laboratory. A second goal was to define the mechanism and the data format that will be used to pass information between HostDesigner and the Ames code. The mechanism and format, defined in collaboration with Ames Laboratory, are described.

2.0 Methods

Molecular mechanics calculations were performed using the MM3 model [39,40] as implemented within either PCModel [66] or MENGINE, which is the open source version of the computational component within PCModel. The MENGINE software, provided at no cost as part of the HostDesigner download package [30], has been expanded to automatically estimate the MM3 parameters needed for a wide variety of metal complexes. In this extended MM3 model the points-on-a-sphere method [67] is used to treat the angles subtended at the metal center. During all MM3 calculations the gas phase dielectric constant, which by default is set to 1.5, was increased to a value of 4.0 for use in condensed phases. This was done to weaken contributions from electrostatic interactions and hydrogen bonding.

As shown in prior studies involving a variety of metal complexes [68-71], the steric influence of inner-sphere aquo ligands can be successfully modeled by using single oxygen atoms and this approximation was implemented in this study. The following process was used to provide a measure of the steric strain associated with the formation of each metal-ligand complex. Using the most common solid-state coordination number for the metal ion, initial coordinates for the metal aquo ion were created by adding oxygen atoms to saturate the coordination sphere. Next, initial coordinates were generated for the metal-ligand complex by either replacing one oxygen atom with a unidentate ligand or two oxygen atoms with a bidentate ligand, and then conformationally searching to locate the lowest energy form. A final conformational search was performed to identify the lowest energy form for the free ligand.

All calculations were performed an iMac computer running MacOS 10.13.6 and research software was produced using GNU gcc and gfortran compilers [72].

3.0 Results and Discussion

3.1 Training data set for CHEMPROP

3.1.1 Overview After extensive review and preliminary testing of several possible machine learning frameworks, collaborators at Ames Laboratory selected the open source code named CHEMPROP [65] as the most promising machine learning software for log K prediction. Before using any machine learning software to predict a chemical property, for example, the log K value for the formation of a metal-ligand complex, it must first be trained to identify relationships between a chemical property and molecule descriptors. The relationships identified by this training process can be stored in a file so that the process only needs to be done once. A prerequisite of the training process is the existence of a set of data in which each entry includes measurement of the chemical property and molecular information. The user must decide which molecular information to include and the success of the training process depends heavily upon this information. A major objective during this funding cycle was to create a data set that can be used to train CHEMPROP [64] to predict log K values.

This objective has been achieved by following several distinct steps. The first step, discussed in Section 3.1.2, was to identify the metal-ligand complexes that will be included in the training data set. The second step, discussed in Section 3.1.3, was to decide which molecular information and descriptors to include for each metal-ligand complex. The final step, discussed in Section 3.1.4, was to define the format that will be used for each entry in the training data set.

3.1.2 Selection of log K data As noted in Section 1.0 there exist several large databases containing aqueous log K data for a wide variety of metal-ligand complexes including the Critical Stability Constants books [44], an electronic database from the National Institute of Standards [45], and an electronic database from the International Union of Pure and Applied Chemistry [46]. The amount of this data that can be used to train CHEMPROP directly depends upon the molecular descriptors that are used the training process. In other words, it must be possible to generate the selected molecular descriptors for each metal-ligand complex in the set. Several promising molecular descriptors require knowledge of the 3D geometry of the metal-ligand complex (see Section 3.1.3). Using such descriptors imposes a significant constraint on the log K data that can be used. The ability to know or predict the 3D geometry for a metal-ligand complex is most likely in cases where the ligand has only one or two donor groups.

When the ligand contains only one donor group, it can be assumed that the unidentate ligand forms a metal-ligand complex by displacing one aquo ligand from the metal aquo ion. Examples of such ligands include ammonia, pyridine, and phenoxide. When the ligand contains two donor groups, the situation immediately becomes more complex. A bidentate ligand can form a metal-ligand complex by displacing two aquo ligands to yield a chelated structure where both donor groups contact the metal ion. Examples include oxalate, bipyridine, and ethylenediamine. Alternatively, a formally bidentate ligand may act like a unidentate ligand, displacing only one aquo ligand with either of the two donor groups. This situation may arise when a rigid ligand architecture does not allow both donor groups to simultaneously contact a metal ion, as with fumaric acid, or the donor groups are separated by ≥ 4 methylene groups, as with 1,4-diaminopentane, and adipic acid. When three donor groups are present, then there are seven possible binding motifs - one tridentate form, three bidentate forms, and three unidentate forms. In general, the number of possible binding motifs and the connectivity in the resulting metal-ligand complex becomes less certain as the denticity of the ligand increases. For this reason, it was decided that the initial training data set would be restricted to log K data for unidentate and bidentate ligands.

It was also decided to restrict the training data set to complexes with 1:1 metal-ligand stoichiometries. This is because additional ambiguity regarding the 3D structure of metal-ligand complexes arises when two or more ligands are present. This ambiguity exists with bidentate ligands for the same reasons given above in that it is not always clear which donor groups are bound to the metal. In addition, when a metal complex forms with two or more bidentate ligands, stereochemistry can become an issue. A well-known example are the *facial* and *meridial* isomers that can occur with octahedral metal ions [73]. Thus, an additional constraint for selecting log K values to include in the training data set was to use only data for 1:1 metal-ligand complexes where structural ambiguities are minimized.

The initial set of ligands was identified by manually browsing the Smith and Martell volumes [44]. Only unidentate and bidentate ligands with log K values for the formation of 1:1 metal-ligand complexes were considered. In addition, another consideration was the number of metal ions for which the log K data had been determined for the ligand. Cases where there were log K values for less than five metal ions were not added. Figures 1 – 3 summarize the 85 ligands that meet the above criteria.

The log K values presented in Smith and Martell compilation [44] were critically reviewed and identified as the most reasonable and accurate values at the time of printing. Critically reviewed means that the values presented in Smith and Martell represent the authors' selection of the most reliable values among those available in the literature. When several workers were in close agreement for a particular value, the average of their results was selected for that value. In cases where agreement is poor and few results are available for comparison, more subtle methods were applied to select the best value. This selection was often guided by a comparison with values obtained for other metal ions with the same ligand and/or with values obtained for the same metal ion with similar ligands.

It was recognized that many of the data listed in the Smith and Martell series, were based on only one of a very few literature references and are subject to change when better data become available. Thus, such data compilations need to be revised and updated from time to time. Smith and Martell Vol. 5 reports data updated in 1986 and Smith and Martell Vol. 6 reports data

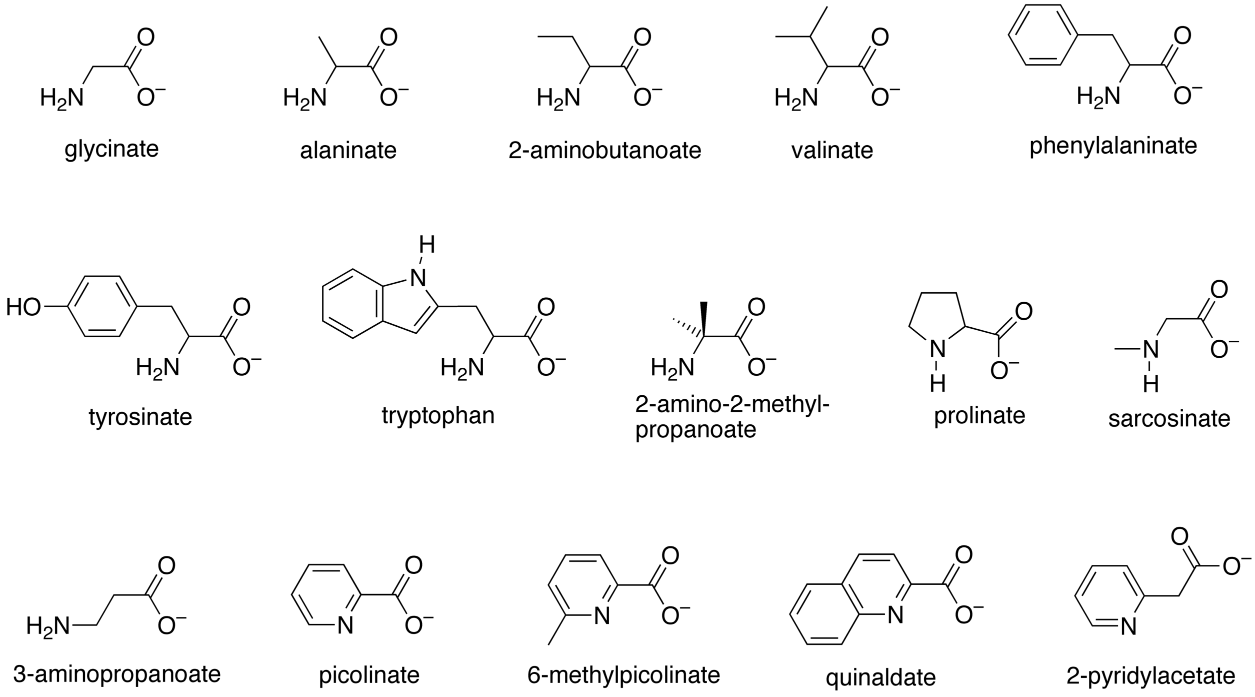

Figure 1. Ligands from Smith and Martell Vol 1 (aminocarboxylates).

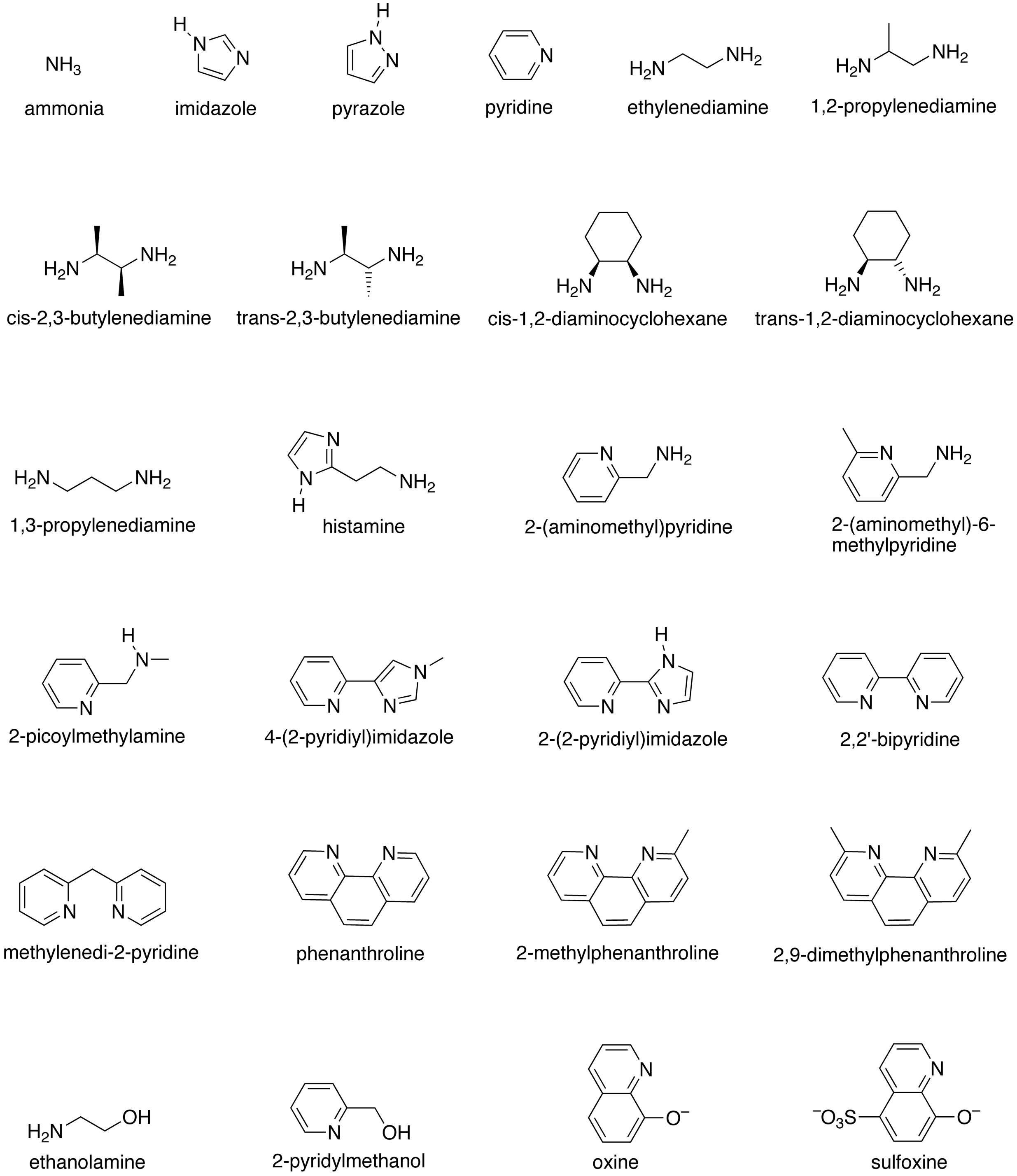

Figure 2. Ligands from Smith and Martell Vol 2 (amines) and Vol 4 (inorganic ligands).

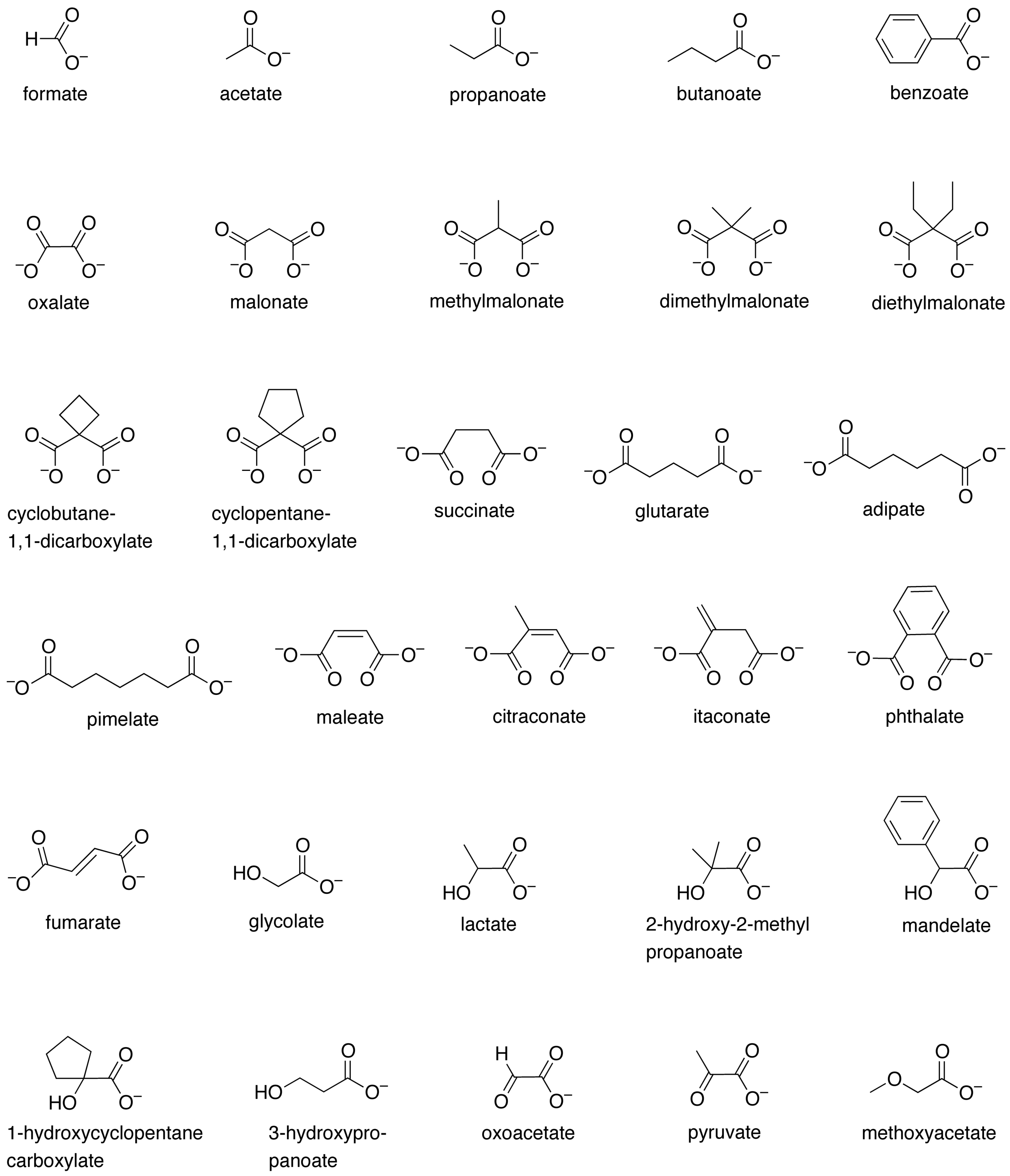

Figure 3. Ligands from Smith and Martell Vol 3 (other organic ligands) and Vol 4 (inorganic ligands).

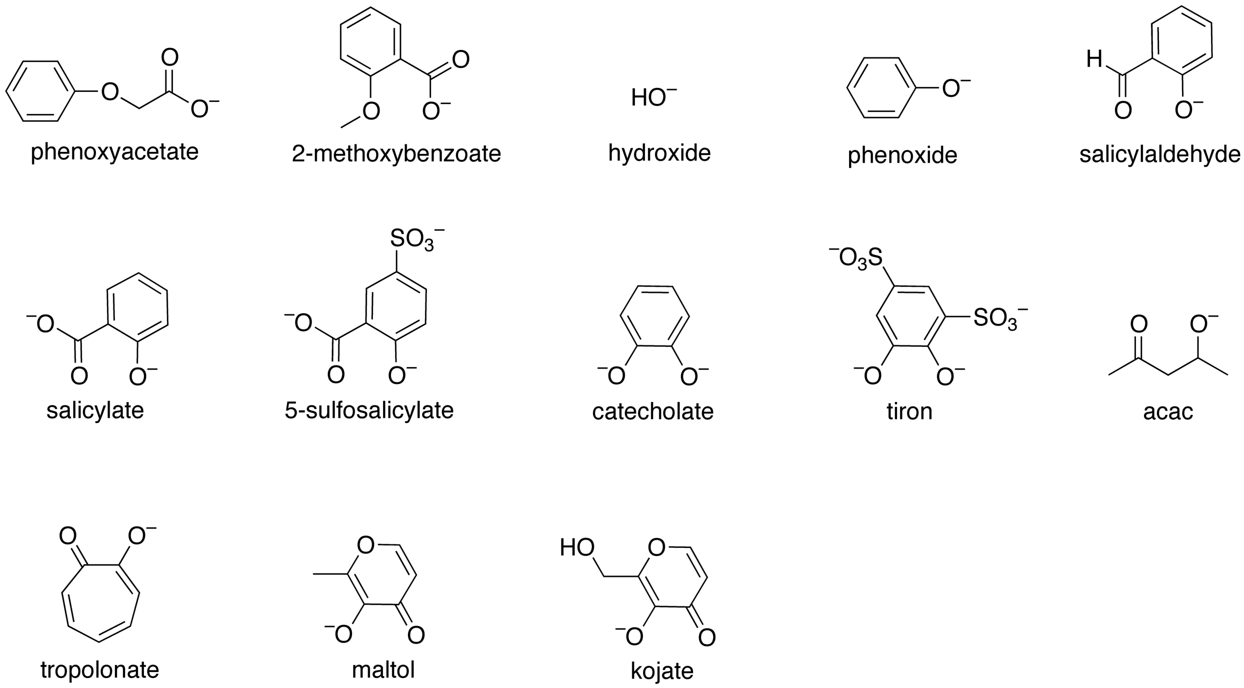

Figure 3 (cont). Ligands from Smith and Martell, Vol 3 (other organic ligands) and Vol 4 (inorganic ligands).

updated in 1989. Thus, any log K values taken from Smith and Martell are first taken from Vol. 6, then if not available in Vol. 6 they are taken from Vol. 5, and finally, if not available in Vol. 6 or 5, they are taken from Vol. 1 – 4. The NIST database [45], created in 1993 and abandoned by 2004, is essentially an electronic copy of the Smith and Martell data. In addition, the NIST interface is extremely cumbersome to use. Therefore, the NIST database was not used to augment the log K data. Instead, the IUPAC database [46], which was created in 1993 and abandoned by 2008, was used to verify, validate, and add to the log K values selected for use in the training data set.

Available log K values for a given metal-ligand complex are often presented at several ionic strengths. For each metal-ligand complex, all log K data at ionic strengths less than 1.0 M from the most recent volume of Smith and Martell and from the IUPAC database were combined to produce a single average log K value at ionic strength 0 M. Before log K values were combined they were corrected to ionic strength 0 M using the Davies equation, Eq. 1 and 2. The Davies equation [74] yields the log base 10 value for the activity coefficient of an ionic species, γ, as a function of the charge, *z*, and the ionic strength, *I*. This is a simple extended Debye-Hückel model (it reduces to the Debye-Hückel equation if the "– 0.20 I" part is removed). After computing log γ values for the complex, the metal ion, and the ligand at the ionic strength used to measure the experimental log K value, then the log K at ionic strength 0 M is obtained using Eq. 2.

$$\log \gamma_i = -0.51 z_i^2 \left[\left(\frac{\sqrt{I}}{\sqrt{I} + 1} \right) - 0.20\, I \right] \tag{1}$$

$$\log K\ (I=0\ M) = \log K_{exp} + \log \gamma_{complex} - \log \gamma_{metal} - \log \gamma_{ligand} \tag{2}$$

Although it was originally proposed that the Davies equation should be limited for use at ionic strengths below 0.1 M [74], it often appears to be valid at ionic strengths up to 1.0 M [75,76]. Table 1 provides an example of how log K values measured for the formation of the 1:1 complex between Ni^{2+} and the phthalate dianion at four different ionic strengths can

Table 1. Example use of Davies equation to adjust log K data to ionic strength 0 M.

I, M	log K_{exp}	log K°	
0.0	2.95	2.95	Example data for $[Ni(phthalate)(OH_2)]^0$ was
0.1	2.14	3.04	corrected to ionic strength = 0 M using
0.5	1.72	3.00	Eq. 1 and 2 yielding an average log K° value
1.0	1.57	2.79	= 2.95 +/- 0.11

be combined to obtain an average value at 0 M ionic strength. Any log K values at ionic strengths ≤ 1.0 M were corrected to an ionic strength of 0 M and considered when computing the mean observed log K values that were included in the training data set. Combining data from the Smith and Martell volumes with those in the IUPAC database yielded log K values for a total of 1662 different metal-ligand complexes.

3.1.3 Molecular descriptors In addition to the log K value, a training data set entry must also include one or more molecular descriptors. At a minimum the CHEMPROP software requires a SMILES string description of the molecule. The SMILES string is based on the valence model of chemistry, which uses a mathematician's graph to represent a molecule. In a chemical graph, the nodes are atoms and the edges are semi-rigid bonds that can be single, double or triple according to the rules of valence bond theory. During training CHEMPROP uses the SMILES string input to learn its own task-specific descriptors through a process called graph convolution [64].

From the foregoing discussion, each entry in a training data set for CHEMPROP must at the minimum contain a log K value and a SMILES string for the metal-ligand complex. This requirement brings up a potential issue. There are several open source codes that are able to convert other file formats into SMILES strings, but there is no standardization of SMILES string specifications for handling the types of bonds and stereochemistries that occur in metal-ligand complexes [77]. After consultation with Ames collaborators, it was decided that in order to ensure that the SMILES string representation used is compatible with the CHEMPROP code, AMES collaborators would create CHEMPROP preprocessor code to convert metal-ligand complex geometries given in SDF molfile format into a viable SMILES string format. The standardized SDF molfile format, [78] which includes atom labels, Cartesian coordinates, and a bond list with bond orders, is one of the most common formats in use today and is recognized by most molecular modeling software.

In addition to the requisite SMILES string input, CHEMPROP is also able to employ user-defined descriptors. In the current application several metal ion descriptors that are expected to show significant correlation with log K values are included in each training data set entry. These descriptors are given in Table 2 for common metal ion species in aqueous solution that are treated by the extended MM3 model included in MENGINE [43]. Basic descriptors include the formal charge, *z*, the most common coordination number observed in the Cambridge Structural Database for complexes containing O and N donor ligands, *CN*, and the effective ionic radius assigned to this *CN* [43]. Another descriptor that provides an implicit measurement of the charge/size ratio is the Gibb's free energy of hydration for the metal ion at 25°C [79], $-\Delta G_{hyd}$. In the few cases where experimental $-\Delta G_{hyd}$ values are not available, then plots of $-\Delta G_{hyd}$ versus the effective ionic radius were used to interpolate/extrapolate the missing value.

Also included in Table 2 are Hancock's parametric descriptors, C_A and E_A, which are used in the parametric relationship given by Eq. 3 [80]. These values are defined using the formation constants for 1:1 hydroxide and fluoride metal complexes: C_A = log K (HO^-)/14.00 and E_A = log K (F^-). The *C* and *E* parameters in Eq. 3 are interpreted as the tendency for a Lewis acid A or Lewis

$$\log K = C_A \bullet C_B + E_A \bullet E_B \tag{3}$$

base B to undergo either covalent (C_A and C_B) or electrostatic (E_A and E_B) bonding. Eq. 3 was found to predict log K values for metal complexes with F^-, HO^-, acetate, NH_3, and pyridine to within +/- 0.2 log units. This finding suggests that if C_B and E_B values could be assigned to other simple prototype donor groups, Eq. 3 might provide a method to predict the log K contribution

Table 2. Metal ion descriptors[a]

MM3 type	atom label	z	CN	ionic radius (Å)	$-\Delta G_{hyd}$ (kcal/mol)	C_A	E_A
301	Li	1	4	0.622	113.5	-0.017	0.57
302	Be	2	4	0.231	572.4	0.495	5.43
303	Na	1	6	1.056	87.2	-0.022	-0.20
304	Mg	2	6	0.712	437.4	0.191	2.05
305	Al	3	6	0.483	1081.5	0.525	7.01
306	K	1	8	1.561	70.5	-0.360	-0.89
307	Ca	2	7	1.075	359.7	0.136	1.24
308	Sc	3	6	0.759	907.0	0.735	7.10
309	Ti	3	6	0.674	959.6	0.907	6.65
310	Ti	4	6	0.600	1749.0	1.678	10.32
311	V	3	6	0.644	1008.6	0.836	6.20
312	V	2	6	0.636	478.6	0.523	3.34
313	Cr	2	6	0.800	442.2	0.375	1.22
314	Cr	3	6	0.596	958.4	0.751	5.20
315	Mn	2	6	0.850	420.7	0.218	1.30
316	Mn	3	6	0.675	989.5	1.029	5.60
317	Fe	2	6	0.770	439.8	0.230	1.41
318	Fe	3	6	0.679	1019.4	0.890	6.00
319	Co	2	6	0.741	457.7	0.240	1.11
320	Co	3	6	0.534	1074.3	0.864	3.30
321	Ni	2	6	0.699	473.2	0.264	1.30
323	Cu	2	6	0.621	480.4	0.455	1.26
324	Zn	2	6	0.745	467.3	0.268	1.43

[a] MM3 type = atom type number used in extended MM3 model; z = formal charge of metal species; CN = coordination number; effective ionic radii from ref 43; $-\Delta G_{hyd}$ values from ref 79; C_A and E_A (see ref 49 and 50) measure covalent and electrostatic character of metal species.

Table 2. Metal ion descriptors[a]

MM3 type	atom label	z	CN	ionic radius (Å)	$-\Delta G_{hyd}$ (kcal/mol)	C_A	E_A
325	Ga	3	6	0.596	1079.1	0.770	5.90
326	Ge	4	6	0.513	1776.0	1.714	9.92
327	Rb	1	9	1.671	67.7	-0.040	-1.05
328	Sr	2	8	1.278	329.8	0.120	0.75
329	Y	3	8	0.990	824.6	0.455	4.80
330	Zr	4	8	0.761	1622.9	0.950	9.80
334	Ru	2	6	0.697	458.3	0.293	1.35
338	Pd	2	4	0.610	456.5	1.050	1.72
339	Ag	1	6	1.150	102.8	0.112	-1.52
340	Cd	2	6	1.013	419.5	0.265	1.10
341	In	3	6	0.843	951.2	0.670	4.60
342	Sn	2	6	0.887	356.1	0.757	5.65
344	Cs	1	9	1.942	59.8	-0.050	-1.43
345	Ba	2	9	1.470	298.8	0.127	0.41
346	La	3	9	1.213	751.7	0.410	3.85
347	Ce	3	9	1.183	764.8	0.427	4.19
348	Ce	4	9	1.017	1462.7	0.854	8.79
349	Pr	3	9	1.163	775.6	0.444	4.25
350	Nd	3	9	1.142	783.9	0.456	4.34
351	Pm	3	9	1.131	795.8	0.446	4.38
352	Sm	3	9	1.115	794.7	0.473	4.39
353	Eu	2	8	1.281	331.0	0.057	0.73
354	Eu	3	9	1.092	803.1	0.480	4.45

[a] MM3 type = atom type number used in extended MM3 model; z = formal charge of metal species; CN = coordination number; effective ionic radii from ref 43; $-\Delta G_{hyd}$ values from ref 79; C_A and E_A (see ref 49 and 50) measure covalent and electrostatic character of metal species.

Table 2. Metal ion descriptors[a]

MM3 type	atom label	z	CN	ionic radius (Å)	$-\Delta G_{hyd}$ (kcal/mol)	C_A	E_A
355	Gd	3	9	1.081	806.6	0.475	4.57
356	Tb	3	8	1.053	812.6	0.470	4.66
357	Dy	3	8	1.043	818.6	0.466	4.69
358	Ho	3	8	1.030	829.4	0.471	4.74
359	Er	3	8	1.016	835.3	0.474	4.76
360	Tm	3	8	1.004	840.1	0.477	4.78
361	Yb	3	8	0.988	853.3	0.472	4.80
362	Lu	3	8	0.974	840.1	0.484	4.83
363	Hf	4	8	0.789	1664.7	0.979	9.73
369	Pt	2	4	0.638	478.6	0.933	1.40
371	Hg	2	7	1.080	420.7	0.825	1.60
372	Tl	1	8	1.590	71.7	0.056	0.10
373	Tl	3	8	0.980	948.9	0.957	2.55
374	Pb	2	8	1.316	340.1	0.395	2.05
375	Bi	3	9	1.133	831.7	0.921	5.91
376	Fr	1	9	2.000	57.8	-0.060	-1.53
377	Ra	2	8	1.530	298.8	0.047	0.28
378	Ac	3	9	1.238	761.7	0.592	4.06
379	Th	4	9	1.080	1389.8	0.925	8.44
380	Pa	4	9	1.077	1520.1	1.057	8.46
381	U	4	9	1.047	1520.1	0.950	8.66
382	U	2	7	1.021	379.8	0.625	5.11
384	Np	2	7	1.027	377.9	0.555	5.20

[a] MM3 type = atom type number used in extended MM3 model; z = formal charge of metal species; CN = coordination number; effective ionic radii from ref 43; $-\Delta G_{hyd}$ values from ref 79; C_A and E_A (see ref 49 and 50) measure covalent and electrostatic character of metal species.

Table 2. Metal ion descriptors[a]

MM3 type	atom label	z	CN	ionic radius (Å)	$-\Delta G_{hyd}$ (kcal/mol)	C_A	E_A
385	Pu	4	9	0.976	1567.9	1.260	8.98
387	Pu	2	7	1.032	377.9	0.545	5.30
388	Am	3	9	1.156	786.1	0.510	4.35
389	Cm	3	9	1.147	789.3	0.570	4.30
390	Bk	3	9	1.134	795.8	0.634	4.38
391	Cf	3	9	1.122	799.2	0.634	4.42
392	Es	3	9	1.111	802.5	0.629	4.47
393	Fm	3	9	1.103	806.0	0.631	4.51
394	Md	3	8	1.059	820.0	0.644	4.65
395	No	3	8	1.044	827.3	0.650	4.70
396	Lr	3	8	1.030	831.0	0.653	4.73

[a] MM3 type = atom type number used in extended MM3 model; z = formal charge of metal species; CN = coordination number; effective ionic radii from ref 43; $-\Delta G_{hyd}$ values from ref 79; C_A and E_A (see ref 49 and 50) measure covalent and electrostatic character of metal species.

from each metal-donor interaction. It is noted that the relationship in Eq. 3 was subsequently modified by adding another term, $D_A \bullet D_B$, in order to account for steric crowding encountered with larger donor atoms and smaller metal ions allowing the relationship to be extended to other ligands [49,50]. As will be discussed below, the MM3-based strain energies should provide an alternate and more accurate descriptor to account for such steric effects.

In addition to the user-defined metal ion descriptors discussed above, it is also possible to include user-defined ligand descriptors as part of the training input. During prior applications of the CHEMPROP software it was observed that the ability to predict molecular properties can often be improved by augmenting learned molecular representations with additional user-defined fragment-based descriptors [64,81]. CHEMPROP can interface with RDKit, which is an open source toolkit for cheminformatics [82], to generate such fragment descriptors. For example, the RDKit Chem.Fragments module will generate 200 fragment descriptors, including values such as number of aliphatic carboxylic acids, number of aromatic nitrogen atoms, number of hydroxylamine groups. A full list of these descriptors is provided in Appendix A of this report. Because the inclusion of such RDKit-generated ligand fragment descriptors can readily be implemented as an option in the CHEMPROP preprocessor, such user-defined ligand fragment descriptors were not included in the training data set entries.

The training data set entries do, however, include two user-defined ligand descriptors. The first descriptor is the formal charge of the ligand species that forms the complex. For example, ammonia is assigned a charge of 0, acetylacetonate assigned a charge of –1, and phthalate is assigned a charge of – 2. The second descriptor is the number of freely rotating bonds in the ligand that are frozen on metal chelation. Studies suggest that each restricted bond rotation weakens ΔG for complex formation by 0.31 kcal/mol, on average [59-63].

The final user-defined descriptor added to each training data set entry, *ΔU*, is a measure of the steric strain associated with metal-ligand complex formation. This descriptor is obtained by performing several MM3 calculations. The procedure used to compute this value is depicted in Figure 4 using $[Co(bipy)(OH_2)_4]^{2+}$ as an example. The procedure starts by creating an input structure for the metal-ligand complex is created by connecting the ligand to the metal ion and adding O atoms, which are used to model the aquo ligands [68-71], to reach the most common coordination number (see CN values in Table 2). A conformational search is performed to identify the lowest energy conformer for the complex, Figure 4a.

The optimized geometry for the metal-ligand complex is then fragmented in two different ways. In the first fragmentation the metal and aquo ligands are removed to give the ligand in its bound geometry, Figure 4b. A single point energy calculation on this fragment gives the steric energy for the bound ligand, U_{bound}. In the second fragmentation, all atoms not attached to the metal are removed to give the inner sphere geometry for the complex, Figure 4c. A single point energy calculation on this fragment gives the steric energy for the inner sphere, U_{inner}.

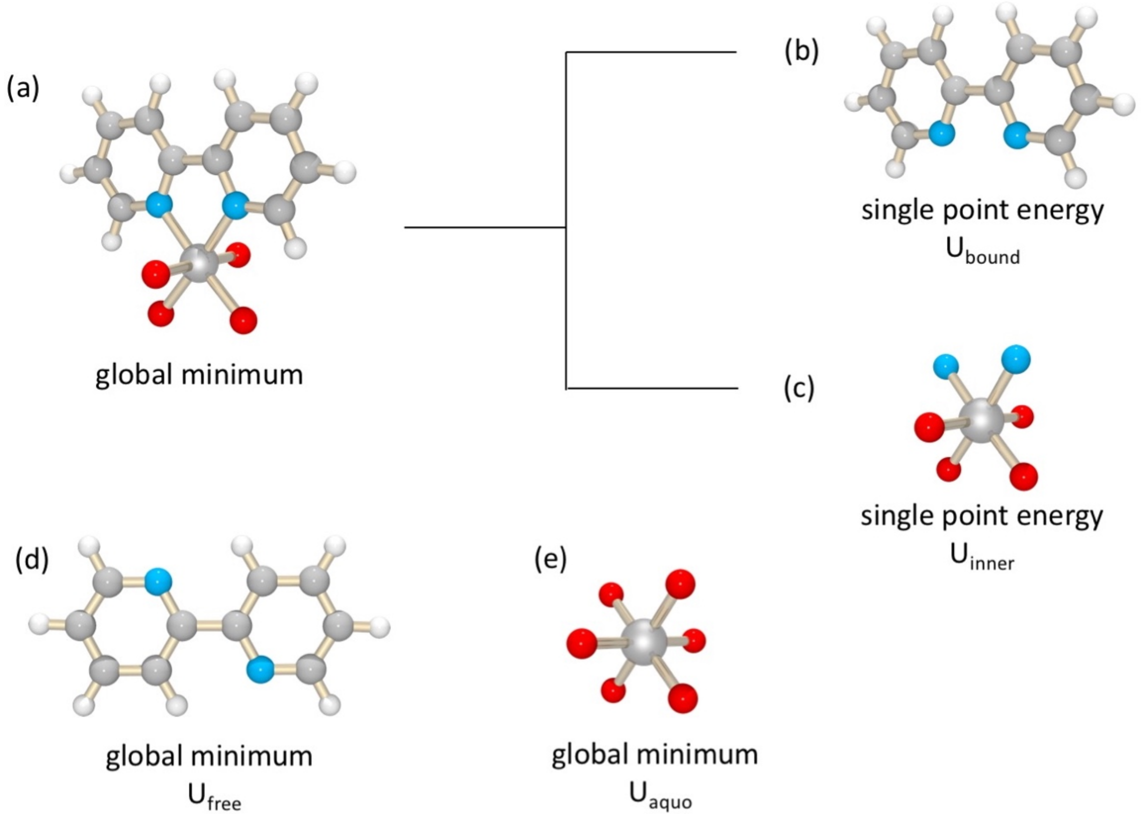

Figure 4. Graphic example of the MM3 calculations used to obtain the steric strain descriptor $\Delta U = (U_{bound} - U_{free}) + (U_{inner} - U_{aquo})$

Second, conformational search is performed on the ligand in order to obtain the steric energy for lowest energy conformer of the free ligand, U_{free}, Figure 4d. Third, an input structure for the metal aquo ion is created by adding O atoms to the metal to reach the most common coordination number (see CN values in Table 2). Optimization of this structure gives the steric energy for the aquo ion, U_{aquo}, Figure 4e. Finally, the strain energy descriptor is given by Eq. 4, where the first term ($U_{bound} – U_{free}$) measures the strain in the ligand on going from the free state to the metal-bound state and the second term ($U_{inner} – U_{aquo}$) gives the increase in inner sphere strain in the metal-ligand complex versus the inner sphere strain in the aquo ion.

$$\Delta U = (U_{bound} - U_{free}) + (U_{inner} - U_{aquo}) \tag{4}$$

3.1.4 Format of training data entry files It was decided that the training data set for CHEMPROP would take the form of a collection of individual files, with one file for each metal complex. This approach has several benefits. The first benefit is that entries can be added or removed from the training set simply by adding or removing entry files from the training data set directory. This facilitates the machine learning process wherein a large part of the data will be used train the code and a small part of the data will be used to validate the ability of the model to predict values not present in the training set. The ability to alter the entries used in the training data set simply by moving files in and out of a directory allows the trainer to perform multiple training runs with different sets of data used for training and validation. A second benefit of this approach is that it is simple to increase the size of the training data set entries by including entry files for additional metal-complexes. Finally, as will be discussed in the next section, the preprocessing code that will be developed to create requisite CHEMPROP input by reading a series of individual entry files from the training set can easily be adapted to read input files for predicting log K values for HostDesigner candidates one at a time.

In accord with the discussion of descriptors presented in the preceding section, the development of a format for a training data set entry was straightforward. An example of this format is given in Figure 5. The first line is a comment line that identifies the numeric descriptors given in order of their appearance on the second line. These are the log K value at ionic strength zero, a log K value at a specified ionic strength (logK_in and I_in), the formal charge of the ligand (Z_lig), the formal charge of the metal ion (Z_met), the number of freely rotating bonds in the ligand that are frozen on metal chelation (nrot), the effective ionic radius for the metal ion (met_r) at the most common coordination number (met_CN), the MM3 strain energy descriptor (E_strain), the negative Gibb's free energy of hydration (G_solv), the Hancock electrostatic (rdhE) and covalent (rdhC) descriptors. The second line contains the values of these descriptors in the same order. The remainder of the file consists of an SDL mol formatted description of the metal-ligand adduct with the inner sphere aquo ligands deleted.

```
logK(I=0.0) logK_in   I_in   Z_lig   Z_met   nrot    met_r   met_CN   E_strain  G_solv  rdhE     rdhC
    1.14      1.14    0.00    -1       2       0     0.712      6       0.86     437.4  2.050    0.191
STRUCTURE: Mg_2+_acetate
Comment: MM3 optimized geometry in MOL format, inner sphere aquo O atoms deleted
Comment: optimization performed with MENGINE (HostDesigner, Version 4.2)
  8  7  0  0  0  0  0  0  0  0999 V2000
   -2.6659    0.0335   -1.0249 C   0  0  0  0  0  0
   -1.3316   -0.5235   -0.5906 C   0  0  0  0  0  0
   -0.3851    0.3071   -0.3935 O   0  0  0  0  0  0
   -1.2138   -1.7435   -0.4462 O   0  0  0  0  0  0
   -3.4204   -0.7659   -1.1696 H   0  0  0  0  0  0
   -2.5715    0.5813   -1.9839 H   0  0  0  0  0  0
   -3.0655    0.7368   -0.2670 H   0  0  0  0  0  0
    1.5772    0.1503    0.1853 Mg  0  0  0  0  0  0
  1  2  1
  1  5  1
  1  6  1
  1  7  1
  2  3  1
  2  4  2
  3  8  1
M  END
$$$$
```

Figure 5. Example training data set entry file for $[Mg(acetate)(OH_2)_5]^{+1}$

Training data entry files like this one were created for 1662 different metal complexes. Utility software was developed to facilitate this process. This software, named MAKEDATA, requires two input files for each of the 85 ligands included in this set (see Figures 1-3). One of these input files is an MENGINE formatted input file for the metal-ligand adduct without any aquo ligands. The other input file defines the name of the MENGINE formatted input file, charge of the ligand, number of ligand bonds frozen on chelation, and contains a list of log K values for 1:1 metal-ligand complexes with different metal ion species. Using these input files, MAKEDATA creates MENGINE files and performs the set of MM3 calculations needed to produce the strain energy descriptor (see Figure 4), assigns the other metal ion descriptors from data tables, and writes a training data set entry for each of the log K values in the list.

The entire training data set, which was transmitted by email to Ames collaborators on 11/13/2020, a copy of the input files used to create it, and a copy of the MAKEDATA source code are available upon request from the author.

3.2 Interaction between HostDesigner and CHEMPROP The overarching project objective is to enhance HostDesigner scoring functions to allow ranking candidates by (1) absolute log K values for the target metal and (2) selectivity for one metal over a competing second metal, which can be expressed as $\log K_1 - \log K_2 = \log (K_1/K_2)$. Assuming that it is possible to train CHEMPROP to predict log K values within an acceptable level of accuracy, then achieving this objective will require modifications to the HostDesigner software. An important part of these modifications is the ability to interface HostDesigner with the as-yet-to-be-coded preprocessor software that will run the CHEMPROP model to predict a log K value.

In order to discuss how HostDesigner will interact with this as-yet-to-be-coded preprocessor software, it is necessary to first assume a couple characteristics of this preprocessor. It is assumed (1) that after training the Ames collaborators create a code named LOGKPREDICT by combining a preprocessor with the existing CHEMPROP framework, (2) this code can be launched from the command line by entering its name, (3) it always attempts to read a file named *logk_input*, and, (4) after it has predicted a log K value, it always writes a file named *logk_output*. The file named *logk_input* would contain the same set of descriptors and be formatted in the same way as a training data set entry file (see Section 3.1.4). In this case a logK(I=0) value of -999 could be used to indicate that the value is not known and is to be predicted. The file named *logk_output* would contain a single real number representing the predicted log K value.

A simplest method to interface HostDesigner with LOGKPREDICT is to use the library function named *system*. This function takes a string as input and allows HostDesigner to execute this string as a command line command. Given that the assumptions described in the preceding paragraph are true, then HostDesigner would obtain a log K value using the following process:

(1) Generate the descriptors and write the *logk_input* file for a ligand candidate

(2) Run this line of code: call system('LOGKPREDICT')
This *system* function call will temporarily halt execution of HostDesigner and launch LOGKPREDICT, which then reads the *logk_input* file, predicts a log K value, writes the predicted log K value to the file named *logk_output*, and stops.

(3) Control returns to HostDesigner, which will then read the predicted log K value from *logk_output*

Going forward, the next task at the Supramolecular Design Institute will be to create the new HostDesigner scoring modules that will allow candidate rankings by absolute log K and by selectivity using the methods described above to interface with the LOGKPREDICT code. This task will take place concurrently with Ames Laboratory collaborators efforts to create the LOGKPREDICT code.

4.0 Summary

This letter report has documented work performed at the Supramolecular Design Institute to support the CHEMPROP modifications and testing activities that are currently being performed at Ames Laboratory. The first goal was to create a training data set that contains the following information for each entry: log K value, 3D description for the MM3-optimized metal-ligand complex, and user-defined descriptors. As described in Section 3.1, this goal was achieved and a training data set that contains entries for 1,662 complexes between 85 representative unidentate and bidentate ligands with up to as many as 50 different metal ion species per ligand was provided to Ames Laboratory. The second goal was to define the mechanism and the data format that will be used to pass information between HostDesigner and the Ames code. The mechanism and format are described in Section 3.2.

5.0 Bibliography

[1] J. W. Steed, J. L. Atwood, Supramolecular Chemistry, 2nd edition, John Wiley & Sons, Ltd. (2009).

[2] H.-J. Schneider, A. Yatsimirski, Principles and Methods in Supramolecular Chemistry; John Wiley & Sons, Ltd: Chichester (2000).

[3] H.-J. Böhm, The Computer Program LUDI – A New Method for the De Novo Design of Enzyme Inhibitors, J. Comput.-Aided Mol. Des. 6 (1992) 61-78.

[4] M. C. Lawrence, P. C. Davis, CLIX – A Search Algorithm for Finding Novel Ligands Capable of Binding Proteins of Known 3-Dimensional Structure, Proteins: Struct. Funct. Genet. 12 (1992) 31-41.

[5] J. Sadowski, J. Gasteiger, From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Molecule Builders, Chem. Rev. 93 (1993) 2567-2581.

[6] C. M. W. Ho, G. R. Marshall, SPLICE – A Program to Assemble Partial Query Solutions from 3-Dimensional Database Searches into Novel Ligands, J. Comput.-Aided Mol. Des. 7 (1993) 623-647.

[7] S. H. Rotstein; M.A. Murcko, GROUPBUILD – A Fragment-Based Method for De Novo Drug Design, J. Med. Chem. 36 (1993) 1700-1710.

[8] M. B. Eisen, D. C. Wiley, M. Karplus, R. E. Hubbard, HOOK: A Program for Finding Novel Molecular Architectures that Satisfy Chemical and Steric Requirements of a Macromolecule Binding Site, Proteins: Struct. Funct. Genet. 19 (1994) 199-221.

[9] V. Tschinke, N. C. Cohen, The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses, J. Med. Chem. 36 (1993) 3863-3870.

[10] V. J. Gillet, W. Newell, P. Mata, G. J. Myatt, S. Sike, Z. Zsoldos, A. P. Johnson, SPROUT – Recent Developments in the De-Novo Design of Molecules, J. Chem. Inf. Comput. Sci. 34 (1994) 207-217.

[11] A. R. Leach, S. R. Kilvington, Automated Molecular Design – A New Fragment Joining Algorithm, J. Comput.-Aided Mol. Des. 8 (1994) 283-298.

[12] P. Mata, V. J. Gillet, A. P Johnson, J. Lampreia, G. J. Myatt, S. Sike, A. L. Stebbings, SPROUT: 3D Structure Generation Using Templates, J. Chem. Inf. Comp. Sci. 35 (1995) 479-493.

[13] D. C. Roe, I. D. Kuntz, BUILDER V2 – Improving the Chemistry of a De-Novo Design Strategy, J. Comput.-Aided Mol. Des. 9 (1995) 269-282.

[14] H.-J. Böhm, Computational Tools for Structure-Based Ligand Design, Prog. Biophys. Molec. Biol. 3 (1996) 197-210.

[15] R. X. Wang, Y. Gao, L. H. Lai, LigBuilder: A Multipurpose Program for Structure-Based Drug Design, J. Mol. Mod. 6 (2000) 498-516.

[16] G. Schneider, U. Fechner, Computer-Based De Novo Design of Drug-Like Molecules, Nature Rev. Drug. Disc. 4 (2005) 649-663.

[17] B. P. Hay, T. K. Firman, HostDesigner: A Program for the De Novo Structure-Based Design of Molecular Receptors with Binding Sites that Complement Metal Ion Guests, Inorg. Chem. 41 (2002) 5502-5512.

[18] B. P. Hay, T. K. Firman, G. J. Lumetta, B. M. Rapko, P. A. Garza, S. I. Sinkov, J. E. Hutchison, B. W. Parks, R .D. Gilbertson, T. J. R. Weakley, Toward the Computer-Aided Design of Metal Ion Sequestering Agents, J. Alloys. Comp. 374 (2004) 416-419.

[19] B. P. Hay, A. A. Oliferenko, J. Uddin, C. G. Zhang, T. K. Firman, Search for Improved Host Architectures: Application of De Novo Structure-Based Design and High-Throughput Screening Methods to Identify Optimal Building Blocks for Multidentate Ethers, J. Am. Chem. Soc. 127 (2005) 17043-17053.

[20] S. Vukovic, B. P. Hay, De Novo Structure-Based Design of Bis-Amidoxime Uranophiles, Inorg. Chem. 52 (2013) 7805-7810.

[21] B. W. McCann, N. De Silva, T. L. Windus, M. S. Gordon, B. A. Moyer, V. S. Bryantsev, B. P. Hay, Computer-Aided Molecular Design of Bis-Phosphine Oxide Lanthanide Extractants, Inorg. Chem. 55 (2016) 5787-5803.

[22] V. S. Bryantsev, B. P. Hay, De Novo Structure-Based Design of Bis-Urea Hosts for Tetrahedral Oxoanion Guests, J. Am. Chem. Soc. 128 (2006) 2035-2042.

[23] C. Reyheller, B. P. Hay, S. Kubik, Influence of Linker Structure on the Anion Binding Affinity of Biscyclopeptides, New. J. Chem. 31 (2007) 2095-2102.

[24] B. P. Hay, V. S. Bryantsev, In *Computational Methods for Sensor Material Selection, Series: Integrated Analytical Systems*, M.A. Ryan, A.V. Shevade, C.J. Taylor, M.L. Homer, and M. Blanco, Eds.; Springer: New York (2009).

[25] B. P. Hay, De Novo Structure-Based Design of Anion Receptors, Chem. Soc. Rev. 39 (2010) 3700-3708.

[26] B. P. Hay, C. Jia, J. Nadas, Computer-Aided Design of Host Molecules for Recognition of Organic Guests, Comp. Theor. Chem. 1028 (2014) 72-80.

[27] R. Custelcean, J. Bosano, P. V. Bonnesen, V. Kertesz, B. P. Hay, Computer-Aided Design of a Sulfate-Encapsulating Receptor, Angew. Chem. Int. Ed. 48 (2009) 4025-4029.

[28] N. J. Young, B. P. Hay, Structural Design Principles for Self-Assembled Coordination Polygons and Polyhedra, Chem. Commun. 49 (2013) 1354-1379.

[29] C. Jia, B. P. Hay, R. Custelcean, De Novo Structure-Based Design of Ion-Pair Triple-Stranded Helicates, Inorg. Chem. 53 (2014) 3893-3898.

[30] B. P. Hay, HostDesigner, Version 4.2 User's Manual, https://sourceforge.net/projects/hostdesigner/, Supramolecular Design Institute, Oak Ridge, TN, 2020.

[31] B. P. Hay, J. R. Rustad, J.C. Hostetler, Quantitative structure-activity relationship for potassium Ion complexation by crown ethers. A molecular mechanics and ab initio study, J. Am. Chem. Soc. 115 (1993) 11158-11164.

[32] B. P Hay, D. Zhang, J. R. Rustad, Structural Criteria for the Rational Design of Selective Ligands. 2. Effect of Alkyl Substitution on Metal Ion Complex Stability with Ligands Bearing Ethylene-Bridged Ether Donors, Inorg. Chem. 35 (1996) 2650-2658.

[33] B. P. Hay, A Molecular Mechanics Method for Predicting the Influence of Ligand Structure on Metal Ion Binding Affinity, in *Metal Ion Separation and Preconcentration: Progress and Opportunities*; A. H. Bond, M. L. Dietz, R. D. Rogers, Eds.; ACS Symposium Series 716, American Chemical Society: Washington, DC, 1999.

[34] M. L. Dietz, A. H. Bond, B. P. Hay, R. Chiarizia, V. J. Huber, A. W. Herlinger, Ligand Reorganization Energies as the Basis for the Design of Synergistic Metal Ion Extractants, Chem. Commun. 13 (1999) 1177-1178.

[35] B. P. Hay, The Use of Molecular Mechanics in the Design of Metal Ion Sequestering Agents, in *Metal Separation Technologies Beyond 2000: Integrating Novel Chemistry with Processing*; K.C. Liddell, D.J. Chaiko, Eds.; The Minerals, Metals, and Materials Society: Warrendale, Pennsylvania, 1999.

[36] B. P. Hay, D. A. Dixon, R. Vargas, J. Garza, K. N. Raymond, Structural Criteria for the Rational Design of Selective Ligands. 3. Quantitative Structure-Stability Relationship for Iron(III) Complexation by Tris-Catecholamide Siderophores, Inorg. Chem. 40 (2001) 3922-3935.

[37] B. P. Hay, R. D. Hancock, The Role of Donor Group Orientation as a Factor in Metal Ion Recognition by Ligands, Coord. Chem. Rev. 212 (2001) 61-78.

[38] G. J. Lumetta, B. M. Rapko, P. A. Garza, B. P. Hay, R. D. Gilbertson, T. J. R. Weakley, J. E. Hutchison, Deliberate Design of Ligand Architecture Yields Dramatic Enhancement of Metal Ion Affinity, J. Am. Chem. Soc., Comm. Ed. 124, (2002) 5644-5645.

[39] N. L. Allinger, Y. H. Yuh, J. H. Lii, Molecular mechanics - The MM3 Force-Field for Hydrocarbons, J. Am. Chem. Soc. 111 (1989) 8551-8566.

[40] J. H. Lii, N. L. Allinger, Molecular mechanics - The MM3 Force-Field for Hydrocarbons. 2. Vibrational Frequencies and Thermodynamics, J. Am. Chem. Soc. 111 (1989) 8566-8575.

[41] N. L. Allinger, X. Zhou, J. Bergsma, Molecular Mechanics Parameters, J. Mol. Struct. THEOCHEM 312 (1994) 69-83.

[42] F. Zahariev, M. Dick-Perez, M. Gordon, T.L. Windus, ParFit, Version 1.1, https://github.com/fzahari/ParFit, Iowa State University and Ames Lab, Ames, IA, 2018.

[43] (a) B. P. Hay, *Fast and Accurate log K Prediction, YR 6 (2018-2019) Progress Report*, Letter Report, Supramolecular Design Institute, Oak Ridge, TN, April 18, 2019. (b) B. P. Hay, *Fast and Accurate log K Prediction, YR 7 (2019-2020) Progress Report*, Letter Report, Supramolecular Design Institute, Oak Ridge, TN, June 12, 2020.

[44] A. E. Martell, R. M. Smith, Critical Stability Constants, Plenum Press: New York, Vol. 1-6, 1974 – 1989.

[45] R. J. Motekaitis, National Institute of Standards and Technology, Standard Reference Database 46, NIST Critical Stability Constants of Metal Complexes, PC-based Database, Gaithersburg, MD 20899, USA, 2004.

[46] L. D. Pettit, K. J. Powell, Stability Constants Database, IUPAC and Academic Software, Timble, Otley, Yorks, UK, 1993-2008.

[47] P. W. Dimmock, P. Warwick, R. A. Robbins, Approaches to Predicting Stability Constants, Analyst 120, (1995) 2159 - 2169.

[48] B. P. Hay, K. J. Castleton, J. R. Rustad, Stability Constant Estimator User's Guide, PNNL-11434, Pacific Northwest National Laboratory, Richland, Washington 99352, USA, 1996.

[49] R. D. Hancock, Approaches to Predicting Stability Constants, a Critical Review, Analyst 122 (1997) 51R-58R.

[50] A. E. Martell, R. D. Hancock, Metal Complexes in Aqueous Solution, Modern Inorganic Chemistry Series, Editor J. P. Fackler, Jr; Plenum Press: New York, 1996.

[51] B. P. Hay, A. Chagnes, G. Cote, On the Metal Ion Selectivity of Oxoacid Extractants, Solv. Extr. Ion Exch. 31 (2014) 95-105.

[52] C. W. Davies, The Electrolytic Dissociation of Metal Hydroxides, J. Chem. Soc. (1951) 1256-1267

[53] R. M. Izatt, W. C. Fernelius, B. P. Block, Studies on Coordination Compounds. XIII. Formation Constants of Bivalent Metal Ions with the Acetylacetonate Ion, J. Phys. Chem. 59 (1955) 80-84.

[54] R. M. Izatt, W. C. Fernelius, C. G. Haas, Jr., B. P. Block, Studies on Coordination Compounds. XI. Formation Constants of Some Tervalent Ions and the Thorium(IV) Ion with the Acetylacetonate Ion, J. Phys. Chem. 59 (1955) 170-174.

[55] E. Nieboer, W. A. E. McBryde, Free-Energy Relationships in Coordination Chemistry. II. A Comprehensive Index to Complex Stability, Can. J. Chem. 51 (1973) 2512-2524.

[56] G. Hefter, Simple Electrostatic Correlations of Fluoride Complexes in Aqueous Solution, Coord. Chem. Rev. 12 (1974) 221-239.

[57] R. D. Hancock, F. Marsicano, The Chelate Effect: A Simple Quantitative Approach, J. C. S. Dalton Trans. (1976) 1096-1098.

[58] W. R. Harris, Structure-Reactivity Relation for the Complexation of Ni, Cd, Zn, and Fe, J. Coord. Chem. 13 (1983) 17-27.

[59] F. Eblinger, H.-J. Schneider, Stabilities of Hydrogen-Bonded Supramolecular Complexes with Various Numbers of Single Bonds: Attempts to Quantify a Dogma in Host-Guest Chemistry, Angew. Chem. Int. Ed. 37 (1998) 826-829.

[61] M. Mammen, E. I. Shakhnovich, G. M. Whitesides, Using a Convenient, Quantitative Model for Torsional Entropy to Establish Qualitative Trends for Molecular Processes that Restrict Conformational Freedom, J. Org. Chem. 63 (1998) 3168-3175.

[62] K. N. Houk, A. G. Leach, S. P. Kim, X. Zhang, Binding Affinities of Host-Guest, Protein-Ligand, and Protein-Transition-State Complexes, Angew. Chem. Int. Ed. 42 (2003) 4872-4897.

[63] F. Deanda, K. M. Smith, J. Liu, R. S. Pearlman, GSSI, a General Model for Solute-Solvent Interactions. 1. Description of the Model, Mol. Pharm. 1 (2004) 23-39.

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

[65] CHEMPROP code is available at no cost from https://github.com/chemprop/

[66] (a) *Pcmodel*, Version 9.3, Serena Software, Bloomington, IN. (b) An upgraded version of this code, *Pcmodel*, Version 10.0, is available at no cost from http://www.serenasoft.com/license.html

[67] B .P. Hay, The Application of Molecular Mechanics in Coordination Chemistry, Coord. Chem. Rev. 126 (1993) 177-236.

[68] B. P. Hay, L. Yang, N. L. Allinger, J.-H. Lii, An Extended MM3(96) Force field for Complexes of the Group 1A and 2A Cations with Ligands Bearing Conjugated Ether Donor Groups, J. Mol. Struct. (THEOCHEM) 428 (1998) 203-219.

[69] B. P. Hay, O. Clement, G. Sandrone, D. A. Dixon, A MM3(96) Force Field for Metal Amide Complexes, Inorg. Chem. 37 (1998) 5887-5894.

[70] B. P. Hay, E. J. Werner, K. N. Raymond, Estimating the Number of Bound Waters in Gd(III) Complexes Revisited. Improved Methods for the Prediction of q-Values, Bioconjugate Chem. 15 (2004) 1496-1502.

[71] B. P. Hay, J. Uddin, T. K. Firman, Eight-Coordinate Stereochemistries of U(IV) Catecholate and Aquo Complexes, Polyhedron 23 (2004) 145-154.

[72] (a) GNU Compiler Collection, Copyright© Free Software Foundation, Inc., 2020. (b) https://gcc.gnu.org/

[73] J. E. Huheey, Inorganic Chemistry, Principles of Structure and Reactivity, 2nd edition, Harper and Row, New York (1978).

[74] C. W. Davies, "The Extent of Dissociation of Salts in Water. Part VIII. An Equation for the Mean Ionic Activity Coefficient of an Electrolyte in Water, and a Revision of the Dissociation Constants of Some Sulphates." J. Chem. Soc. (1938) 2093-2098.

[75] D. Dyrssen, M. Wedborg, Equilibrium Calculations of the Speciation of Elements in Seawater, 5 (1974) 181–195.

[76] MINEQL+, Chemical equilibrium and speciation modeling software, https://www.mineql.com

[77] M. Quiros, S. Grazulis, S. Gridzijauskaite, A. Merkys, A. Vaitkus, Using SMILES strings for the Description of Chemical Connectivity in the Crystallography Open Database, J. Cheminform. 10 (2018) 17 pages.

[78] (a) A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer, Description of Several Chemical Structure File Formats used by Computer Programs Developed at Molecular Design Limited, J. Chem. Inf. Model. 32 (1992) 244-255. (b) C. A. James, OpenSMILES Specification, http://opensmiles.org.html

[79] Y. Marcus, Thermodynamics of Solvation of Ions, Part 5. – Gibbs Free Energy of Hydration at 298.15 K, J. Chem. Soc. Faraday Trans. 87 (1991) 2995-2999.

[80] R. D. Hancock, F. Marsicano, Parametric Correlation of Formation Constants in Aqueous Solution. 1. Ligands with Small Donor Atoms, Inorg. Chem 17 (1978) 560-564.

[81] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, J. J. Collins, A Deep Learning Approach to Antibiotic Discovery, Cell 180 (2020) 688-702.

[82] RDKit documentation is available at https://www.rdkit.org/docs/

Appendix A. Descriptors Generated by RDKit Chem.Fragment Module

This information is from http://www.rdkit.org/docs/source/rdkit.Chem.Fragments.html

rdkit.Chem.Fragments.fr_Al_COO(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aliphatic carboxylic acids

rdkit.Chem.Fragments.fr_Al_OH(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aliphatic hydroxyl groups

rdkit.Chem.Fragments.fr_Al_OH_noTert(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aliphatic hydroxyl groups excluding tert-OH

rdkit.Chem.Fragments.fr_ArN(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of N functional groups attached to aromatics

rdkit.Chem.Fragments.fr_Ar_COO(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of Aromatic carboxylic acide

rdkit.Chem.Fragments.fr_Ar_N(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aromatic nitrogens

rdkit.Chem.Fragments.fr_Ar_NH(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aromatic amines

rdkit.Chem.Fragments.fr_Ar_OH(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aromatic hydroxyl groups

rdkit.Chem.Fragments.fr_COO(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of carboxylic acids

rdkit.Chem.Fragments.fr_COO2(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of carboxylic acids

rdkit.Chem.Fragments.fr_C_O(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of carbonyl O

rdkit.Chem.Fragments.fr_C_O_noCOO(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of carbonyl O, excluding COOH

rdkit.Chem.Fragments.fr_C_S(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thiocarbonyl

rdkit.Chem.Fragments.fr_HOCCN(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of C(OH)CCN-Ctert-alkyl or C(OH)CCNcyclic

rdkit.Chem.Fragments.fr_Imine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of Imines

rdkit.Chem.Fragments.fr_NH0(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of Tertiary amines

rdkit.Chem.Fragments.fr_NH1(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of Secondary amines

rdkit.Chem.Fragments.fr_NH2(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of Primary amines

rdkit.Chem.Fragments.fr_N_O(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of hydroxylamine groups

rdkit.Chem.Fragments.fr_Ndealkylation1(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of XCCNR groups

rdkit.Chem.Fragments.fr_Ndealkylation2(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of tert-alicyclic amines (no heteroatoms, not quinine-like bridged N)

rdkit.Chem.Fragments.fr_Nhpyrrole(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of H-pyrrole nitrogens

rdkit.Chem.Fragments.fr_SH(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thiol groups

rdkit.Chem.Fragments.fr_aldehyde(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aldehydes

rdkit.Chem.Fragments.fr_alkyl_carbamate(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of alkyl carbamates (subject to hydrolysis)

rdkit.Chem.Fragments.fr_alkyl_halide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of alkyl halides

rdkit.Chem.Fragments.fr_allylic_oxid(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of allylic oxidation sites excluding steroid dienone

rdkit.Chem.Fragments.fr_amide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of amides

rdkit.Chem.Fragments.fr_amidine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of amidine groups

rdkit.Chem.Fragments.fr_aniline(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of anilines

rdkit.Chem.Fragments.fr_aryl_methyl(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of aryl methyl sites for hydroxylation

rdkit.Chem.Fragments.fr_azide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of azide groups

rdkit.Chem.Fragments.fr_azo(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of azo groups

rdkit.Chem.Fragments.fr_barbitur(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of barbiturate groups

rdkit.Chem.Fragments.fr_benzene(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of benzene rings

rdkit.Chem.Fragments.fr_benzodiazepine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of benzodiazepines with no additional fused rings

rdkit.Chem.Fragments.fr_bicyclic(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Bicyclic

rdkit.Chem.Fragments.fr_diazo(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of diazo groups

rdkit.Chem.Fragments.fr_dihydropyridine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of dihydropyridines

rdkit.Chem.Fragments.fr_epoxide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of epoxide rings

rdkit.Chem.Fragments.fr_ester(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of esters

rdkit.Chem.Fragments.fr_ether(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of ether oxygens (including phenoxy)

rdkit.Chem.Fragments.fr_furan(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of furan rings

rdkit.Chem.Fragments.fr_guanido(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of guanidine groups

rdkit.Chem.Fragments.fr_halogen(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of halogens

rdkit.Chem.Fragments.fr_hdrzine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of hydrazine groups

rdkit.Chem.Fragments.fr_hdrzone(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of hydrazone groups

rdkit.Chem.Fragments.fr_imidazole(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of imidazole rings

rdkit.Chem.Fragments.fr_imide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of imide groups

rdkit.Chem.Fragments.fr_isocyan(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of isocyanates

rdkit.Chem.Fragments.fr_isothiocyan(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of isothiocyanates

rdkit.Chem.Fragments.fr_ketone(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of ketones

rdkit.Chem.Fragments.fr_ketone_Topliss(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha

rdkit.Chem.Fragments.fr_lactam(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of beta lactams

rdkit.Chem.Fragments.fr_lactone(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of cyclic esters (lactones)

rdkit.Chem.Fragments.fr_methoxy(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of methoxy groups -OCH3

rdkit.Chem.Fragments.fr_morpholine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of morpholine rings

rdkit.Chem.Fragments.fr_nitrile(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of nitriles

rdkit.Chem.Fragments.fr_nitro(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of nitro groups

rdkit.Chem.Fragments.fr_nitro_arom(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of nitro benzene ring substituents

rdkit.Chem.Fragments.fr_nitro_arom_nonortho(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of non-ortho nitro benzene ring substituents

rdkit.Chem.Fragments.fr_nitroso(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of nitroso groups, excluding NO2

rdkit.Chem.Fragments.fr_oxazole(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of oxazole rings

rdkit.Chem.Fragments.fr_oxime(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of oxime groups

rdkit.Chem.Fragments.fr_para_hydroxylation(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of para-hydroxylation sites

rdkit.Chem.Fragments.fr_phenol(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of phenols

rdkit.Chem.Fragments.fr_phenol_noOrthoHbond(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of phenolic OH excluding ortho intramolecular Hbond substituents rdkit.Chem.Fragments.fr_phos_acid(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of phosphoric acid groups

rdkit.Chem.Fragments.fr_phos_ester(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of phosphoric ester groups

rdkit.Chem.Fragments.fr_piperdine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of piperdine rings

rdkit.Chem.Fragments.fr_piperzine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of piperzine rings

rdkit.Chem.Fragments.fr_priamide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of primary amides

rdkit.Chem.Fragments.fr_prisulfonamd(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of primary sulfonamides

rdkit.Chem.Fragments.fr_pyridine(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of pyridine rings

rdkit.Chem.Fragments.fr_quatN(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of quarternary nitrogens

rdkit.Chem.Fragments.fr_sulfide(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thioether

rdkit.Chem.Fragments.fr_sulfonamd(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of sulfonamides

rdkit.Chem.Fragments.fr_sulfone(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of sulfone groups

rdkit.Chem.Fragments.fr_term_acetylene(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of terminal acetylenes

rdkit.Chem.Fragments.fr_tetrazole(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of tetrazole rings

rdkit.Chem.Fragments.fr_thiazole(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thiazole rings

rdkit.Chem.Fragments.fr_thiocyan(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thiocyanates

rdkit.Chem.Fragments.fr_thiophene(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of thiophene rings

rdkit.Chem.Fragments.fr_unbrch_alkane(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)

rdkit.Chem.Fragments.fr_urea(*mol*, *countUnique=True*, *pattern=<rdkit.Chem.rdchem.Mol object>*)

Number of urea groups