Biophysical Journal
Article

Biophysical Society

Temperature-Dependent Estimation of Gibbs Energies Using an Updated Group-Contribution Method

Bin Du,[1] Zhen Zhang,[1] Sharon Grubner,[1] James T. Yurkovich,[1] Bernhard O. Palsson,[1] and Daniel C. Zielinski[1,*]
[1]Department of Bioengineering, University of California San Diego, La Jolla, California

ABSTRACT Reaction-equilibrium constants determine the metabolite concentrations necessary to drive flux through metabolic pathways. Group-contribution methods offer a way to estimate reaction-equilibrium constants at wide coverage across the metabolic network. Here, we present an updated group-contribution method with 1) additional curated thermodynamic data used in fitting and 2) capabilities to calculate equilibrium constants as a function of temperature. We first collected and curated aqueous thermodynamic data, including reaction-equilibrium constants, enthalpies of reaction, Gibbs free energies of formation, enthalpies of formation, entropy changes of formation of compounds, and proton- and metal-ion-binding constants. Next, we formulated the calculation of equilibrium constants as a function of temperature and calculated the standard entropy change of formation ($\Delta_f S^\circ$) using a model based on molecular properties. The median absolute error in estimating $\Delta_f S^\circ$ was 0.013 kJ/K/mol. We also estimated magnesium binding constants for 618 compounds using a linear regression model validated against measured data. We demonstrate the improved performance of the current method (8.17 kJ/mol in median absolute residual) over the current state-of-the-art method (11.47 kJ/mol) in estimating the 185 new reactions added in this work. The efforts here fill in gaps for thermodynamic calculations under various conditions, specifically different temperatures and metal-ion concentrations. These, to our knowledge, new capabilities empower the study of thermodynamic driving forces underlying the metabolic function of organisms living under diverse conditions.

INTRODUCTION

The First and Second Laws of Thermodynamics connect reaction-flux directions, metabolite concentrations, and reaction-equilibrium constants. An increasing number of systems biology methods have begun to take advantage of the intimate connection between thermodynamics and metabolism to obtain insights into the function of metabolic networks. These methods have been used in a number of applications, including the calculation of thermodynamically feasible optimal states (1,2), the identification of thermodynamic bottlenecks in metabolism (3,4), and the constraint of kinetic constants via Haldane relationships (5).

To perform thermodynamic analyses on metabolic networks, it is necessary to have values for the equilibrium constants of reactions carrying flux in the network. Experimentally, the equilibrium constant of a reaction is determined by calculating the mass action ratio (the ratio of product to substrate concentrations), also called the reaction quotient, when the reaction is at equilibrium. A collection of experimentally measured equilibrium constants for over 600 reactions has been published in the National Institute of Standards and Technology (NIST) Thermodynamics of Enzyme-Catalyzed Reactions database (TECRdb) (6). However, the equilibrium constants of the majority of known metabolic reactions are still unmeasured, making computational estimation necessary. The most commonly used approach for estimating thermodynamic constants in aqueous solutions is the group-contribution method (7,8). This method is based on the simplifying assumption that the Gibbs energy of formation ($\Delta_f G^\circ$) of a compound is based on the sum of the contributions of its composing functional groups, which are independent of each other. The contribution of each group can be estimated through linear regression, using existing data on $\Delta_f G^\circ$ and the Gibbs energies of reactions ($\Delta_r G^\circ$).

Recent iterations of group-contribution methods for reactions in aqueous solutions have incorporated pH corrections into estimations of equilibrium constants (9) and improved accuracy by taking advantage of fully defined reaction-stoichiometric loops forming First Law energy-conservation

Submitted September 22, 2017, and accepted for publication April 16, 2018.
*Correspondence: dczielin@ucsd.edu
Editor: Daniel Beard.
https://doi.org/10.1016/j.bpj.2018.04.030

relationships within the training data (10). These methods also have begun to take advantage of computational chemistry software to estimate the pK_a-values of compounds as part of thermodynamic parameter estimation. However, a number of issues remain for thermodynamic estimation of reaction-equilibrium constants in metabolic networks, including 1) significant estimation errors in many cases, which may be attributed to a number of factors, including missing or erroneous reaction conditions; and 2) the lack of an established method to handle correction of thermodynamic data with respect to temperature changes across conditions. Additionally, existing group-contribution methods have not taken into account the substantial metal-ion binding of many metabolites at physiological ion concentrations, although established theory exists to correct reaction-equilibrium constants for metal-ion binding when ion-dissociation constants are available (11).

The geochemistry field has developed a sophisticated theory to handle thermodynamic variables as a function of temperature for a wide variety of compounds in aqueous solutions (12–18). The parameters used to calculate thermodynamic transformation across temperature are specific for different compounds. However, the available literature only covers less than half of the compounds in the NIST TECRdb. Therefore, the estimation of a large number of compound-specific parameters is required. It is possible to use a group-contribution approach by incorporating these parameters into the formulation of $\Delta_r G'^\circ$ and fitting them against experimental data at different temperatures. However, because of the lack of data in necessary depth and resolution, the parameter estimation procedure on the fully parameterized thermodynamic model can suffer from significant error due to overfitting of parameters. Therefore, a simplified approach with fewer parameters to transform $\Delta_r G'^\circ$ across temperature is desirable.

In this study, we extend the capabilities of computational estimation of reaction-equilibrium constants for metabolic networks. We first curate the NIST TECRdb of reaction-equilibrium constants to obtain missing reaction conditions and correct any other errors. We further incorporate additional thermodynamic data, including $\Delta_f G^\circ$ and data related to proton and metal-ion binding, from a number of other sources (11,19–23). The equilibrium constants and $\Delta_f G^\circ$-values are commonly used as the training data for group contribution. The proton- and metal-ion-binding data are required to transform $\Delta_r G'^\circ$ across different pH and metal-ion concentrations. To enable the calculation of equilibrium constants as a function of temperature, we adapt the thermodynamic theory from the geochemistry literature (12–18) given certain simplifying assumptions. The thermodynamic parameters required for such calculation, $\Delta_f S^\circ$ of aqueous species, are estimated through a regression model using various molecular descriptors. Next, to fill gaps in the magnesium-binding correction of equilibrium constants, we estimate magnesium-binding constants for 618 compounds using molecular descriptors and magnesium-binding groups defined based on known magnesium-binding compounds. Finally, we incorporate these new data and functionalities into the most recently published group-contribution framework, termed the component contribution (10), to obtain a new group-contribution estimator for reaction-equilibrium constants with expanded capabilities.

MATERIALS AND METHODS

Workflow for estimation of equilibrium constants

We first introduce the workflow for estimation of equilibrium constants illustrated in Fig. 1 *A*. The following sections expand upon the workflow in greater detail. We collected and curated 4298 equilibrium constants (K') for 617 unique reactions measured under different conditions (temperature, pH, ionic strength, metal-ion concentrations) as the training data set for the current group-contribution method (Fig. 1 *B*). We also collected $\Delta_f G^\circ$-values from multiple sources as the training data (11,19,20). We collected and curated stability constants of metal-ion complexes from the International Union of Pure and Applied Chemistry (IUPAC) Stability Constants Database (SC-database) and $\Delta_f S^\circ$ from various literature sources and online databases (11,19,20) (Fig. 1 *C*). To complete the necessary thermodynamic transformations to reference conditions, we estimated different thermodynamic properties for compounds for which data were not available. We estimated pK_a-values using ChemAxon (Budapest, Hungary) (http://www.chemaxon.com). We used regression models to estimate magnesium binding constants (pK_{Mg}) and $\Delta_f S^\circ$ based on collected data.

First, we transformed all measurements to the same reference conditions at 298.15 K, pH 7, 0 M ionic strength, and no metal concentration. We applied a Legendre transform to account for the different ion-binding states of each compound as in the previous component-contribution method (10). The transformation of the Gibbs free energy of reaction across pH and ionic strength is also based on the previous method. However, we used the Davies equation rather than the extended Debye-Hückel equation to calculate activity coefficients of electrolyte solutions, as the Davies equation was used in the previous work for thermodynamic transformations across temperature (12–15). The transformation of Gibbs free energy of reaction across different metal concentrations is based on the formulation described by Alberty (11,24). The transformation of Gibbs free energy of reaction across temperature is based on adapted thermodynamic theory from the geochemistry literature (12–15) with simplifying assumptions. The relevant equations and theory above can be found in the Supporting Materials and Methods.

Using $\Delta_r G^\circ$ and $\Delta_f G^\circ$ data at reference conditions, we applied the component contribution method by Noor et al. (10) and obtained estimates of $\Delta_r G^\circ$ and $\Delta_f G^\circ$ at reference conditions. Using these values, as well as the estimated $\Delta_r S^\circ$ to transform $\Delta_r G'^\circ$ across temperature (more details in Results) and other thermodynamic transformations applied in the previous work (10), we are able to calculate the equilibrium constant of a given reaction at defined temperature, pH, and ionic strength.

Curation of the IUPAC SC-database

The IUPAC SC-database contains ion-binding data, i.e., dissociation/binding/stability constants, under various conditions from primary literature. Additionally, the database contains several different annotations for the binding of protons and metal ions to specific aqueous species. When the ligand is a proton, the related dissociation constant is a pK_a constant, whereas when the ligand is a metal ion such as magnesium, the dissociation constant is a pK_{Mg} (modified to the specific ion) constant. For each compound of interest, we categorized the available binding data specific to each ion-bound state. We then corrected binding data to 0 M ionic strength

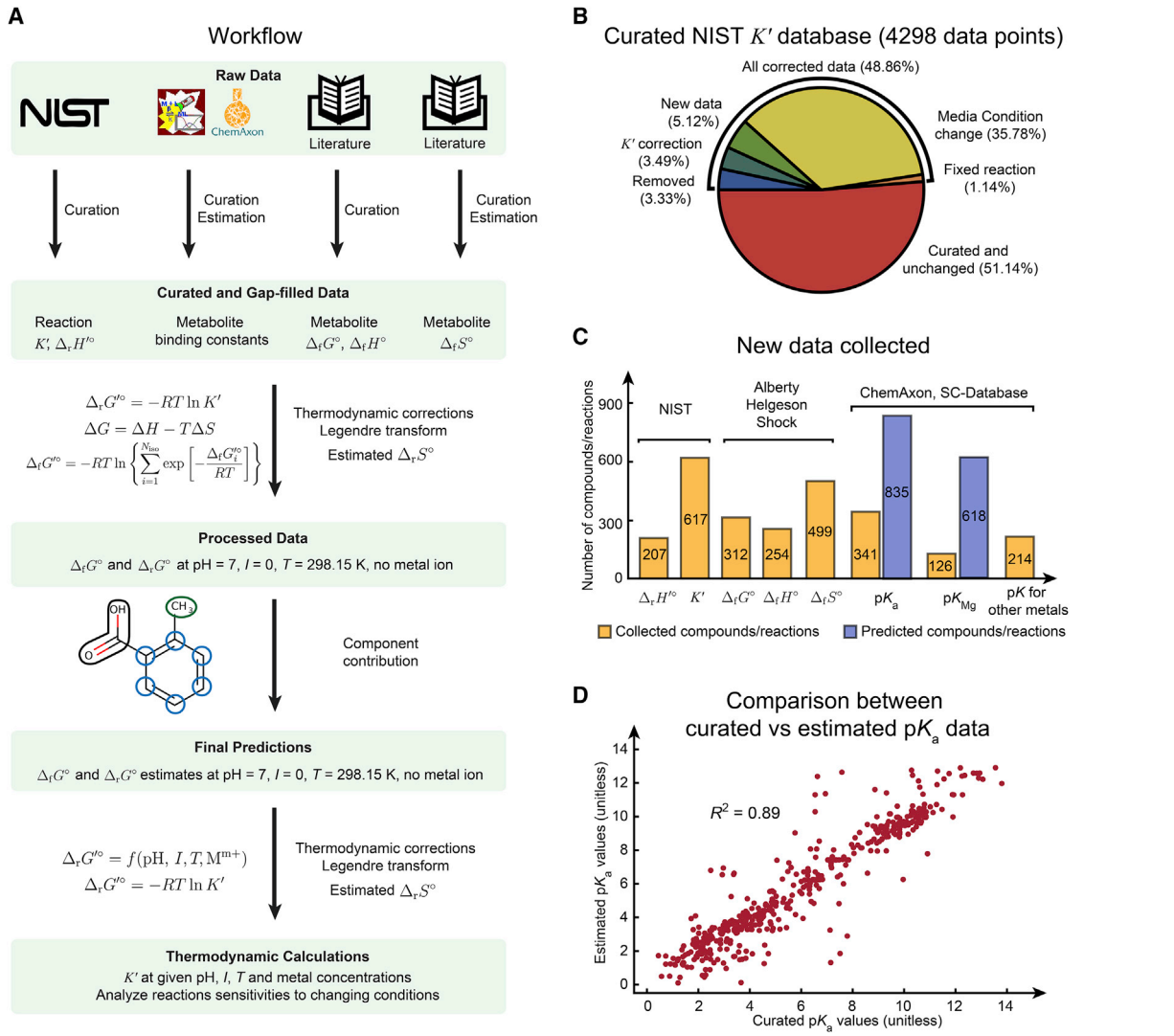

FIGURE 1 Estimation of reaction-equilibrium constants. (*A*) The workflow of data curation and parameter fitting for equilibrium constant estimation is given. (*B*) The results of curation of equilibrium constants from the NIST TECRdb are given. (*C*) New thermodynamic properties generated, either collected from sources shown or computationally estimated, are given. (*D*) A comparison between curated pK_a data from the IUPAC SC-database as well as literature with computationally estimated pK_a-values from ChemAxon is shown. To see this figure in color, go online.

using the Davies equation (25). For each ion-binding reaction, we calculated the median of all available binding data as the value utilized in the fitting (Table S1, tabs 4, 5, and 7).

Features and data used in regression models to estimate pK_{Mg} and $\Delta_f S^\circ$

For estimation of pK_{Mg}, we included a total of 140 data points (Table S1, tab 5) and 128 molecular descriptors as features for regression models. The molecular descriptors included magnesium binding groups identified from existing pK_{Mg} data (Table S1, tab 8), the charge of the compound excluding any magnesium binding groups, sums of partial charge and numbers of different types of atoms, and several additional molecular descriptors from ChemAxon and RDkit. For estimation of $\Delta_f S^\circ$, we included 762 data points (Table S1, tab 3) and 195 features including group decompositions, sums of partial charge and numbers of different types of atoms, and molecular descriptors from ChemAxon and RDkit. The molecular descriptors of compound were estimated with Calculator Plugins, Marvin 16.11.21, 2016, ChemAxon (http://www.chemaxon.com) and RDKit: open-source cheminformatics (http://www.rdkit.org). The full list of molecular descriptors used can be found in Table S1, tab 15.

Comparison of regression methods using nested 10-fold cross-validation

We tested six different regression methods to estimate pK_{Mg} and $\Delta_f S^\circ$. These methods are ridge regression, lasso regression, elastic net regularization, random forests, extra trees, and gradient boosting. We applied nested 10-fold cross-validation to compare the performance of these regression methods. The specific implementation of nested 10-fold cross-validation involves generating an outer loop and inner loop of cross-validation. The outer loop separates the whole dataset into 10 folds, with one fold for testing and the rest for training in each iteration. The training data in each iteration is further separated into 10 folds, and cross-validation is performed in the inner loop to select the optimal model hyperparameters through grid search (Table S1, tab 16). We repeated the nested 10-fold cross-validation on each

regression method five different times by splitting the data into different subdivisions.

We then assessed model performance through the median absolute residual of testing errors calculated from the outer loop, for a total of 50 folds (10 folds × 5 repetitions). The testing errors calculated here also reflect how well the model generalizes on unseen data and are thus used as a metric to evaluate model performance. We also evaluated model stability by calculating the relative standard deviation (RSD, SD/mean) of hyperparameters selected by the inner loop for a total of 50 folds (10 folds × 5 repetitions). We evaluated both testing error and the RSD of hyperparameters when selecting the final regression model to use. For every fitting procedure, we applied standardization on both the training and testing set using the mean and SD of features calculated from the training set.

The regression models, including linear models, tree-based methods, and gradient boosting, were implemented using the Python package scikit-learn 0.19.1 (26).

Lasso regression for estimation of pK_{Mg} and $\Delta_f S^\circ$

Based on the evaluation of different regression methods through nested 10-fold cross-validation (more details in Results), we used lasso regression as the model to estimate pK_{Mg} and $\Delta_f S^\circ$. Specifically, the objective function to minimize is

$$\min_w \frac{1}{2n_{\text{samples}}} \|y - Xw\|^2 + \alpha \|w\|_1, \quad (1)$$

where y is the vector of data with length n_{samples}, X is the matrix with features in the row corresponding to each data point, w is the vector of coefficients of the model, and α is a constant that tunes the degree of the l_1 penalty.

We repeated 10-fold cross-validation 100 times on pK_{Mg} and $\Delta_f S^\circ$ data sets, respectively, to find the optimal α-values that lead to the lowest testing errors. We then constructed a lasso-regression-based estimator for each pK_{Mg} and $\Delta_f S^\circ$ dataset using the selected α-value and applying standardization on the dataset.

Comparison of previous and current group-contribution method

We compared how the previous (10) and the current group-contribution methods perform at different temperatures. Because the previous group-contribution method does not involve an explicit term to correct for $\Delta_r G'^\circ$ at different temperatures, we were only able to substitute different temperatures in thermodynamic transformations and Legendre transform (Eqs. S8 and S9, the RT term) as the temperature transformation on $\Delta_r G'^\circ$. On the other hand, the current method includes an explicit term ($\Delta_r S^\circ$) besides the RT term to calculate $\Delta_r G'^\circ$ at different temperatures. Using the two methods, we calculated $\Delta_r G'^\circ$-values of all the TECRdb data measured at different temperatures and the absolute residual of the estimated $\Delta_r G'^\circ$-values against experimental data.

We then performed 10-fold cross-validation on the 432 reactions that overlapped between the previous and the current group-contribution method. Specifically, we first transformed experimentally measured $\Delta_r G'^\circ$ data to the reference state $\Delta_r G^\circ$ (298.15 K, pH 7, 0 M ionic strength), with different sequential modifications on this procedure (based on the previous method). These modifications include updated media conditions, the Davies equation to correct for the effect of ionic strength, new compound groups, temperature correction, and metal correction. For each set of $\Delta_r G^\circ$-values obtained, we calculated the median $\Delta_r G^\circ$ of all data points in each unique reaction, and performed 10-fold cross-validation on those 432 $\Delta_r G^\circ$-values. We repeated this procedure 100 times by splitting the data into different subdivisions. We then calculated the median absolute residual of 100 repetitions for each reaction.

Additionally, we also compared how well the two methods perform on the 185 new reactions collected in this work. The first method is based on the previous work by Noor et al. (10), whereas the second method in the current work is similar to the first but has several modifications, including updated media conditions, the Davies equation, new compound groups, and the temperature correction. We fit the group-contribution model using both methods with $\Delta_r G^\circ$-values of the original 432 overlapping reactions as training data and calculated the absolute residual in predicting $\Delta_r G^\circ$ for the 185 new reactions as the testing set.

Calculation of standard entropy change of formation

The standard entropy change of formation ($\Delta_f S^\circ$) of the compound is not directly available. Given the type of data available, it can be calculated either from $\Delta_f G^\circ$ and the standard enthalpy of formation ($\Delta_f H^\circ$) of the compound

$$\Delta_f S^\circ = \left(\Delta_f H^\circ - \Delta_f G^\circ\right)/T \quad (2)$$

or from the standard molar entropy (S°) of the compound

$$\Delta_f S^\circ = S^\circ - \sum_{i=1}^{N_e} n_e S_e^\circ, \quad (3)$$

where S_e° is the standard molar entropy of the element N_e composing the compound and n_e is the number of atoms for the element N_e.

Implementation and availability of source code

The updated group-contribution method has been implemented in Python 2.7.6. The source code is available on GitHub (https://github.com/bdu91/group-contribution), together with detailed instructions on how to install it and examples using the package.

RESULTS

Collection and curation of thermodynamic data

The workflow for estimating reaction-equilibrium constants under given pH, temperature, ionic strength, and metal-ion concentrations is demonstrated in Fig. 1 *A* (Materials and Methods). To obtain the necessary data for this estimation, we curated a number of databases and primary literature sources. First of all, from the NIST TECRdb (https://randr.nist.gov/enzyme) (6), we obtained measured equilibrium constants (K') and enthalpies of reactions ($\Delta_r H'^\circ$) for 617 and 207 unique reactions, respectively. Noticing a number of gaps in experimental conditions and other minor issues, we curated a total of 4298 measured K' data from the NIST TECRdb. This curation effort resulted in 48.9% corrected data entries, including updated experimental media conditions (35.78%), addition of new data (5.12%), correction of K'-values (3.49%), removal of problematic data (3.33%) (examples in Table S1, tab 13), and correction of reaction formulae (1.14%) (Fig. 1 *B*).

Next, we collected data on standard Gibbs free energies of formation ($\Delta_f G^\circ$), standard enthalpies of formation

($\Delta_f H^\circ$), and standard entropy of formation changes ($\Delta_f S^\circ$) for 312, 254, and 499 unique compounds, respectively (Fig. 1 *C*). $\Delta_f S^\circ$ data are usually not directly measured but instead are calculated from either $\Delta_f G^\circ$ and $\Delta_f H^\circ$ data or standard molar entropy (S°) of the compound (Materials and Methods). The above data are from multiple sources: *Thermodynamics of Biochemical Reactions* by Alberty (11), the SUPCRT92 database (19), and the Organic Compounds Hydration Properties Database (20).

Lastly, we collected and curated pK_a data for 341 compounds, magnesium binding constants for 126 compounds, and other metal-type binding constants for 214 compounds (including cobalt, iron, zinc, sodium, potassium, manganese, calcium, and lithium) from the IUPAC SC-database and primary literature (21–23) (Fig. 1 *C*). We also predicted pK_a data for 835 compounds using ChemAxon (http://www.chemaxon.com) (Fig. 1 *C*). We compared the collected pK_a data and the predicted values from ChemAxon for the same compounds (Fig. 1 *D*). We found that the differences between the collected and predicted pK_a-values can be as large as 5.84 (unitless), with a median of 0.42 (unitless). This error is a large enough difference to substantially alter the major protonation states for metabolites containing groups with pK_a-values around physiological pH. We examined the specific cause of the largest discrepancies and found that they are due to issues such as assignment of the pK_a-value to the wrong charged form by ChemAxon (e.g., 4-oxo-L-proline) or error in calculating pK_a-values related to particular molecular moieties, such as nitrogenous bases and nitrogen atoms on unsaturated rings (e.g., 2′-deoxyguanosine 5′-monophosphate, xanthine-8-carboxylate, deaminocozymase). We thus used measured pK_a data when available. All collected and curated data can be found in Table S1, tabs 1–7.

Thermodynamic parameters for transformation of $\Delta_r G'^\circ$ across temperature

We then sought to develop the capability to calculate standard transformed Gibbs energy of reaction ($\Delta_r G'^\circ$) as a function of temperature. Specifically, we adapted theory from the geochemistry literature under constant enthalpy and entropy assumptions (12–15), as well as the assumption that the contribution of heat capacity to change in Gibbs energy over temperature is negligible compared to the contribution of entropy (derivation in Supporting Materials and Methods). Thus, we obtained a simple linear formulation of $\Delta_r G'^\circ$ at a given temperature T using the standard entropy change of reaction $\Delta_r S^\circ$ at a reference T_r (298.15 K) (derivation in Supporting Materials and Methods):

$$\Delta_r G'^\circ_T = \Delta_r G'^\circ_{T_r} - (T - T_r)\Delta_r S^\circ_{T_r}. \quad (4)$$

As $\Delta_r S^\circ_{T_r}$ (we use $\Delta_r S^\circ$ in later references because T_r is the only condition of interest, and the same for $\Delta_f S^\circ$) of reactions can be calculated from the $\Delta_f S^\circ$ of the compounds involved, we sought to construct a regression model to estimate $\Delta_f S^\circ$-values. Besides collecting 669 $\Delta_f S^\circ$-values for 499 compounds at different protonation states (Table S1, tab 3) as training data, we also collected $\Delta_r S^\circ$-values from multiple sources. These $\Delta_r S^\circ$-values are effectively linear combinations of $\Delta_f S^\circ$-values and can also be used as training data for $\Delta_f S^\circ$ estimation. From the NIST TECRdb, we selected reactions with K' data measured under at least four different temperatures. We then calculated the $\Delta_r S^\circ$ of each reaction using the $\Delta_r G'^\circ$ of the reaction at different temperatures based on Eq. 4, obtaining 51 $\Delta_r S^\circ$-values. Next, we picked reactions in the NIST TECRdb with both $\Delta_r G^\circ$ and $\Delta_r H^\circ$ data available and calculated their $\Delta_r S^\circ$-values, obtaining 41 additional data points. Together, we obtained a total of 762 data points for $\Delta_f S^\circ$ estimation.

Estimation of standard entropy change of formation $\Delta_f S^\circ$

We found that simple molecular descriptors, notably the number of atoms in the compound and the compound charge, were highly useful as predictors for $\Delta_f S^\circ$. Specifically, we found $\Delta_f S^\circ$ data to be highly correlated simply with the total numbers of atoms in the compound, with an R^2 of 0.89 (Fig. 2 *A*). The $\Delta_f S^\circ$ data as a function of atom number are separated into two main clusters, one of which contains aqueous species with large atom numbers and large absolute $\Delta_f S^\circ$-values (oxidized nicotinamide adenine dinucleotide, reduced nicotinamide adenine dinucleotide, oxidized nicotinamide adenine dinucleotide phosphate, reduced nicotinamide adenine dinucleotide phosphate). The other cluster contains a wide variety of aqueous species, with a few categories labeled in Fig. 2 *A*. We noticed clear separations among aqueous species with −5, −4, −3, and −2 charge, but less so for those with −1, 0, and +1 charge (Fig. 2 *A*). We found the trend between $\Delta_f S^\circ$ and number of atoms exists even more strongly among compounds within the same homologous series, in which the compound structures differ only by the number of CH_2 units in the main carbon chain. Specifically, $\Delta_f S^\circ$-value decreases by ~0.11 kJ/K/mol with every additional CH_2 unit. This trend was observed in a number of homologous series including alkanes, alkenes, alkynes, aldehydes, single carboxylic acids, amines, amides, and thiols. However, the change in $\Delta_f S^\circ$ with respect to the number of atoms across different homologous series is inconsistent, thus requiring additional molecular descriptors.

As an additional descriptor, we found that partial charge of atoms can help distinguish $\Delta_f S^\circ$ from different homologous series. For example, the carbon atoms in glycerol (alcohol containing multiple hydroxyl groups) have larger partial charges than those in methanol (alcohol containing a single group). The prediction of glycerol $\Delta_f S^\circ$ from methanol $\Delta_f S^\circ$ based on their difference in atom numbers yielded

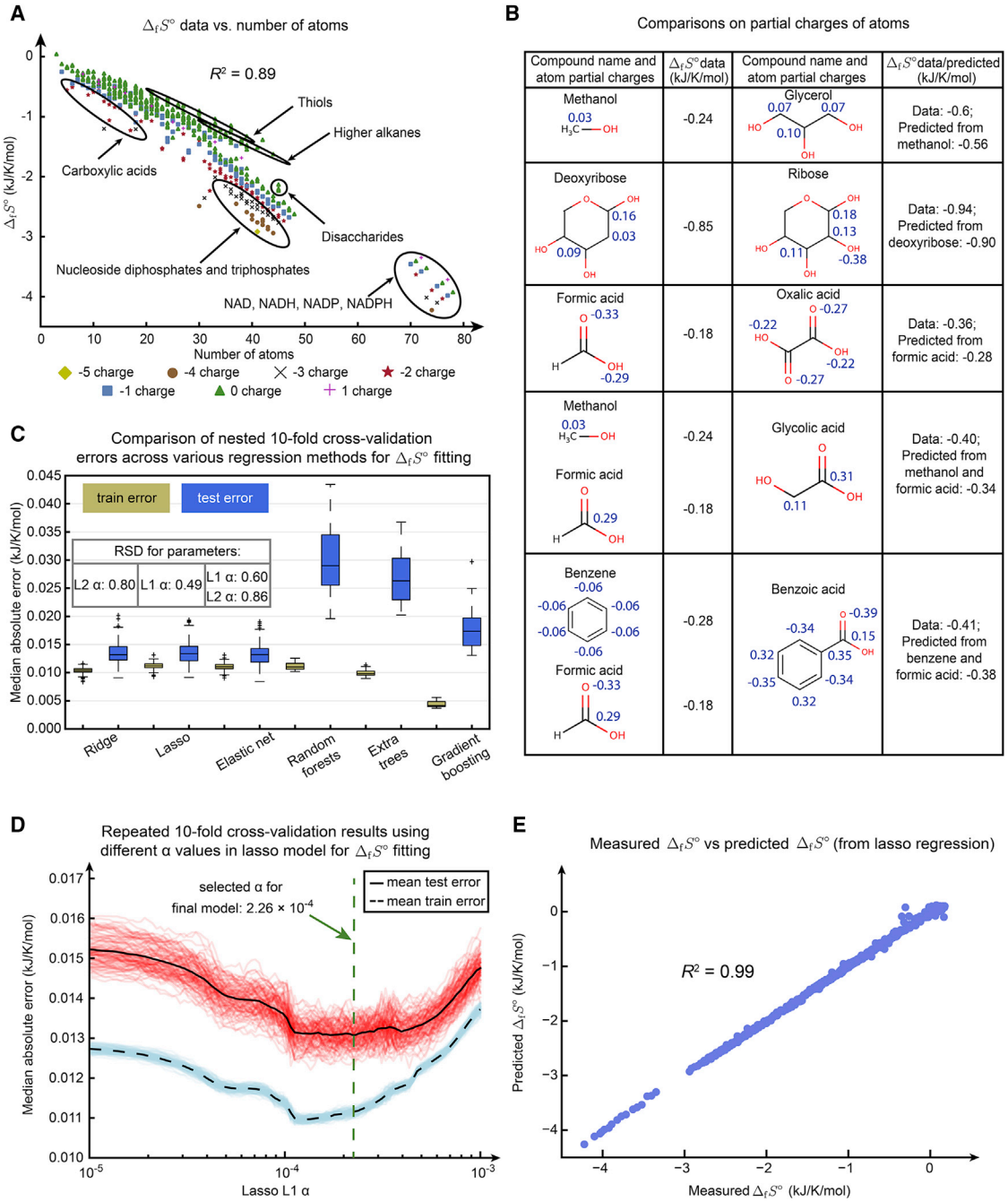

FIGURE 2 Estimation of standard entropy change of formation $\Delta_f S°$. (*A*) $\Delta_f S°$ versus number of atoms is shown. The compounds with different charges are encoded with different colored symbols. We also labeled compounds belonging to the same category or containing the same functional groups. (*B*) Comparisons of partial charges of atoms between compounds are shown. Each row contains a pair of compounds and their $\Delta_f S°$ data. In each pair, the $\Delta_f S°$ of the latter compound can be predicted from that of the former compound(s) based on the difference in atom number. The partial charges of atoms that are different within each pair are marked in blue. (*C*) Training and testing errors of nested 10-fold cross-validation (repeated five different times) on $\Delta_f S°$ data using six different regression methods are shown. We also used relative standard deviation (RSD, SD/mean to assess the relative variability of linear regression model parameters selected by the inner loops of nested cross-validation. (*D*) A selection of parameters in the lasso regression using 10-fold cross-validation on all $\Delta_f S°$ data is shown. We repeated 10-fold cross-validation 100 times and calculated training (*blue*) and testing (*red*) errors at α from 10^{-5} to 10^{-3}. The mean training and testing errors are shown in dashed and solid black lines. The selected α at the lowest mean testing error is 2.26×10^{-4} (unitless). (*E*) A comparison of 762 measured $\Delta_f S°$ training data versus predicted $\Delta_f S°$-values from the final lasso-regression model. To see this figure in color, go online.

a smaller absolute $\Delta_f S°$-value than the actual glycerol data (Fig. 2 *B*) (calculation in Table S1, tab 9). The correlation of larger partial charges of carbon atoms with larger absolute $\Delta_f S°$ is also observed in other pairs in Fig. 2 *B* (deoxyribose versus ribose, methanol + formic acid versus glycolic acid, benzene + formic acid versus benzoic acid). Besides carbon

atoms, we also found differences in partial charges of oxygen atoms to be associated with $\Delta_f S^\circ$ differences, as shown between formic acid and oxalic acid (Fig. 2 *B*). After these observations, we included the sums of absolute partial charge of each type of atom as molecular descriptors for the regression model.

In addition to partial charge, we also considered a number of other molecular descriptors from ChemAxon and RDkit (Materials and Methods). We obtained a total of 195 features and 762 $\Delta_f S^\circ$ data for regression models. We performed nested 10-fold cross-validation to compare between multiple regression models (Fig. 2 *C*). We selected lasso regression as the final model to use, because it has significantly smaller testing errors compared to more complex methods and the least variation in parameters selected from cross-validation compared to other linear regression methods (Fig. 2 *C*). Using parameters selected from cross-validation on the entire $\Delta_f S^\circ$ dataset (Fig. 2 *D*), we constructed a lasso-regression model and predicted 672 $\Delta_f S^\circ$-values (Table S1, tab 17). We obtained 121 predictive variables from the final lasso model, including 1) the number of carbon, hydrogen, and oxygen atoms; 2) the partial charge of hydrogen and oxygen atoms; 3) the formal charge of the compound; 4) the presence of phosphate groups; and 5) the solvent-accessible surface area. The median absolute residual of the lasso-regression model for $\Delta_f S^\circ$ estimation is 0.013 kJ/K/mol (Fig. 2 *C*). Because $\Delta_r S^\circ$-values are linear combinations of $\Delta_f S^\circ$-values, we used the final lasso-regression model to estimate the $\Delta_r S^\circ$-values for all 617 reactions in the TECRdb (Table S1, tab 10).

Evaluation of temperature-dependent estimation of $\Delta_r G'^\circ$

We next evaluated the performance of our method in estimating $\Delta_r G'^\circ$ at different temperatures. We calculated $\Delta_r G'^\circ$-values of all the K' data measured at different temperatures in the TECRdb using the current method with estimated $\Delta_f S^\circ$-values and the previous group-contribution method (10). We calculated the absolute residuals of $\Delta_r G'^\circ$ estimation and compared the two methods across temperature. We found that our method resulted in smaller residuals than the previous method in all temperature ranges (Fig. 3 *A*). This result is also confirmed in different reactions for which we identified series of K' data measured at different temperatures. In all those cases, our estimated $\Delta_r G'^\circ$ across temperature agreed well with the experimental data, in contrast to the estimations by the previous method (Fig. 3, *B–D*). Additionally, we found the temperature-dependent estimation of $\Delta_r G'^\circ$ to be quite robust in the temperature range of available data in the TECRdb (0–90°C), which covers the living conditions of most organisms. Examining reactions for which $\Delta_r G^\circ$-values are predicted to be sensitive to change in temperature (large $\Delta_r S^\circ/\Delta_r G^\circ$ ratio), a number of interesting cases in central metabolism were identified, including malate dehydrogenase, amino acid transaminase, and transketolase (Table S1, tab 14).

Estimation of unknown magnesium binding constants

In addition to its dependence on temperature, the standard transformed Gibbs free energy of the compound ($\Delta_f G'^\circ$) can also depend on pH and the concentrations of metal ions because of the presence of different protonation states and various metal bound species. Specifically, $\Delta_f G'^\circ$ can be calculated based on the standard transformed Gibbs energies of its different ion bound states ($\Delta_f G'^\circ_1$, $\Delta_f G'^\circ_2$, etc.) through Legendre transform (11).

$$\Delta_f G'^\circ = -RT\ln\left\{\sum_{i=1}^{N_{iso}} \exp\left[-\frac{\Delta_f G'^\circ_i}{RT}\right]\right\}. \quad (5)$$

The equation can be rewritten as

$$\Delta_f G'^\circ = \Delta_f G'^\circ_1 - RT\ln\left\{1 + \exp\left(\frac{\Delta_f G'^\circ_1 - \Delta_f G'^\circ_2}{RT}\right) + \exp\left(\frac{\Delta_f G'^\circ_1 - \Delta_f G'^\circ_3}{RT}\right) + \ldots\right\}, \quad (6)$$

where $\Delta_f G'^\circ_1$ is the Gibbs energy of a particular ion-bound state (typically with the least hydrogens and metal ions bound). The Gibbs energy of a specific ion-bound state ($\Delta_f G'^\circ_i$) can then be written in terms of $\Delta_f G'^\circ_1$ and the binding polynomial P_i,

$$\Delta_f G'^\circ_i = \Delta_f G'^\circ_1 - RT\ln P_i, \quad (7)$$

where P_i is expressed in terms of the proton concentration and metal-ion concentration as well as the binding constants of successive proton- and metal-ion-binding steps to obtain the i^{th} ion-bound state (11) (derivation in Supporting Materials and Methods). Therefore, metal-binding constants are important parameters that affect $\Delta_f G'^\circ$ and subsequently reaction-equilibrium constants.

We focused on magnesium binding, because the magnesium ion is well-known to bind to various metabolites and its binding to ATP and other phosphate-containing compounds has been characterized experimentally (27,28). However, magnesium-binding data is still lacking for a large number of compounds that contain similar structural groups to those known to bind magnesium, suggesting that many more compounds may have substantial magnesium binding than have been measured.

Based on the structures of compounds with known magnesium binding, we determined 31 magnesium-binding groups (Table S1, tab 8), most of which are phosphate and carboxyl groups. We were unable to determine the specific binding groups for certain categories of compounds that

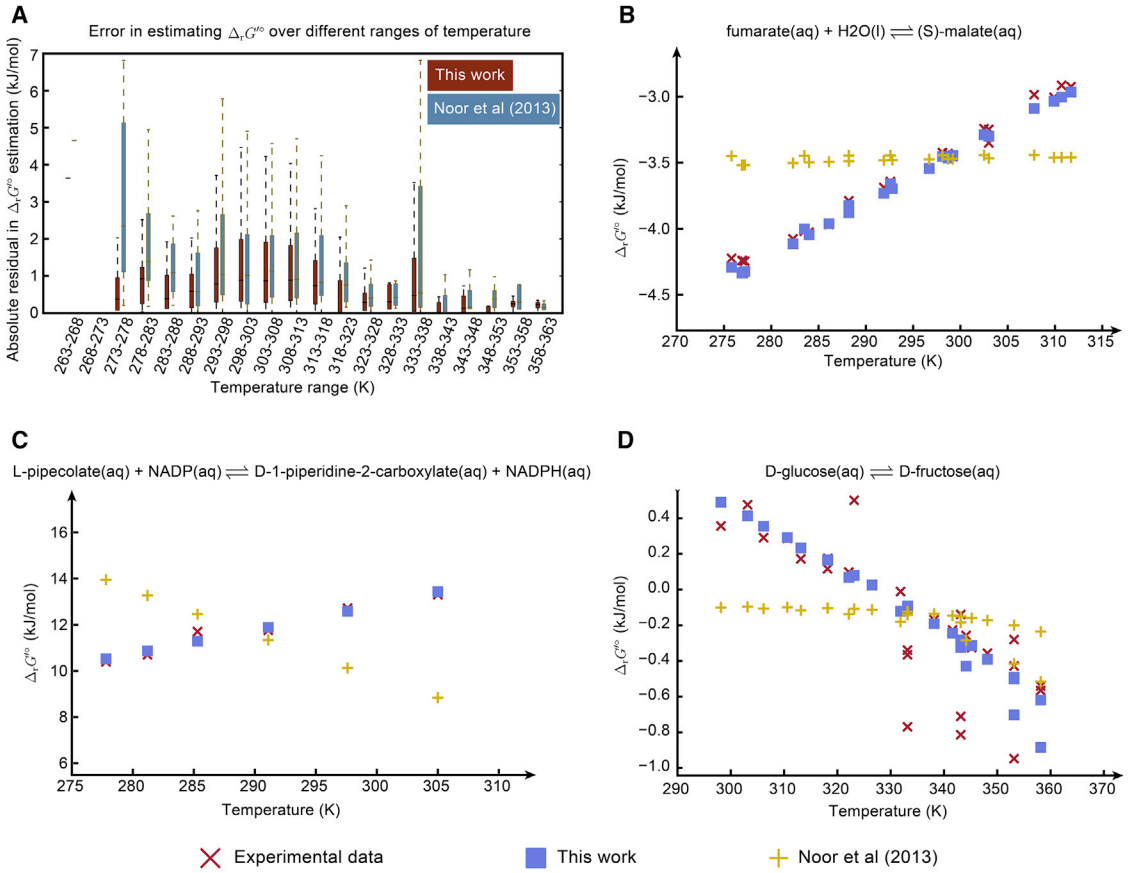

FIGURE 3 Evaluation of temperature-dependent estimation of $\Delta_r G'^\circ$. (*A*) A comparison of absolute residuals on estimating $\Delta_r G'^\circ$ at different temperatures between the previous group-contribution method (10) and the current method is shown. For all the TECRdb data measured at different temperatures, we estimated the $\Delta_r G'^\circ$-values using the previous method and the current method and calculated the absolute residual against experimental data. For clarity in comparison, we divided the entire temperature range into windows with 5 K difference. (*B*) Estimated $\Delta_r G'^\circ$-values for the fumarate hydratase reaction at different temperatures using the previous method and the current method are given. (*C*) Estimated $\Delta_r G'^\circ$-values for the 1-piperidine-2-carboxylate reductase reaction at different temperatures using the previous method and the current method are given. (*D*) Estimated $\Delta_r G'^\circ$-values for the xylose isomerase reaction at different temperatures using the previous method and the current method are given. To see this figure in color, go online.

were measured to complex with magnesium, including nucleobases, ribonucleosides, and purine derivatives. To try to capture metabolite properties responsible for Mg binding in these cases, we added molecular properties (Materials and Methods) as additional descriptors. Together, we used 128 features and 140 measured magnesium-binding constants to construct several candidate regression models for the prediction of magnesium-dissociation constants. We performed nested 10-fold cross-validation to compare between multiple regression models (Fig. S2, *A* and *B*). We selected lasso regression as the best predictor because of its superior generalizability compared to more complex methods (Fig. S2 *A*) and stable model parameters across cross-validation replicates compared to other linear methods (Fig. S2 *B*). Using 140 measured magnesium-binding constants as training data, we constructed a lasso-regression model with parameters tuned through cross-validation (Fig. 4 *A*) and predicted 1707 magnesium-binding constants for aqueous species from 618 compounds (Table S1, tab 5). We obtained 35 predictive variables from the final lasso model, including the formal charge, the solvent-accessible surface area, the presence of various phosphate groups for magnesium binding, the partial charge of nitrogen atoms, the compound charge excluding its magnesium-binding groups, and the dipole moment of the molecule. We found 34 of the 618 compounds are predicted to bind to magnesium at physiological concentrations (2–3 mM) (29). The median absolute residual of the lasso-regression model for magnesium-binding-constant estimation is 0.39 (unitless), as calculated by the nested 10-fold cross-validation (Fig. S2 *A*).

Estimation of standard Gibbs free energy of reaction

Utilizing the curated and estimated datasets mentioned above, as well as the estimation of $\Delta_r S^\circ$, we adapted the most recent group-contribution-based method, termed

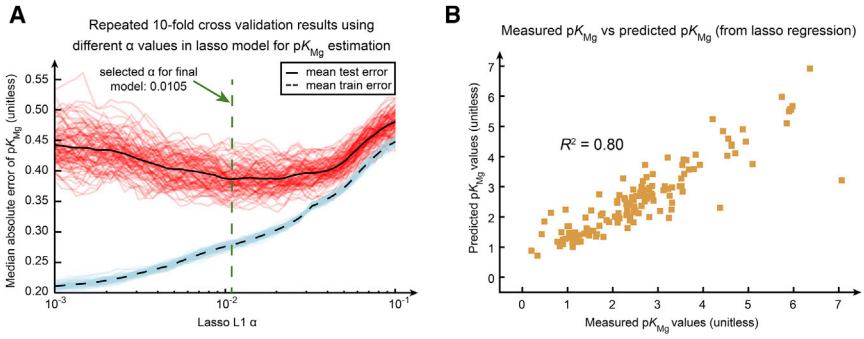

FIGURE 4 Estimation of compound magnesium binding constants (pK_{Mg}). (*A*) A selection of parameters in the lasso regression using 10-fold cross-validation on all pK_{Mg} data is given. We repeated 10-fold cross-validation 100 times and calculated training (*blue*) and testing (*red*) errors at α from 10^{-3} to 10^{-1}. The mean training and testing errors are shown in dashed and solid black lines. The selected α at the lowest mean testing error is 0.0105 (unitless). (*B*) A comparison of 140 measured pK_{Mg} training data versus predicted pK_{Mg}-values from the final lasso-regression model is shown. To see this figure in color, go online.

component contribution (10), to calculate reaction-equilibrium constants for a set of 617 unique reactions in the NIST TECRdb. Besides the addition of transformation of $\Delta_r G'^\circ$ across temperature, we also included 17 novel group definitions to account for compounds with new functional groups not covered by the previous component-contribution method. The novel group definitions can be found in Table S1, tab 11. Additionally, we used the Davies equation (25) rather than the extended Debye-Hückel equation (used in the previous component-contribution method (10)) to correct for the effect of ionic strength, as the Davies equation was used in the previous work on temperature-dependent thermodynamic calculations (12–15). We also showed that the Davies equation was slightly more effective in correcting data at high ionic strength compared to the extended Debye-Hückel equation (Fig. S5). On top of the new functionalities, we also added additional $\Delta_r G^\circ$-values for 185 reactions and $\Delta_f G^\circ$-values for 178 compounds over the dataset used in the previous method.

We compared the accuracy of the updated component-contribution method with the previous work using repeated 10-fold cross-validation (Materials and Methods) for a set of 432 overlapping reactions (10). We applied the modifications mentioned above sequentially on the framework to examine how each new functionality affects the estimation error globally (Fig. S4 *A*). We first noted that the updated media conditions increased the median absolute residual of $\Delta_r G^\circ$ estimation (6.21 kJ/mol), which we found to be due to the addition of data at high ionic strength (>0.5 M, beyond the working range of the Davies equation). Removal of those data resulted in similar errors as in the previous work (5.95 kJ/mol). We found a modest decrease in median absolute residual with the additional group definitions (5.82 kJ/mol) and capability to transform Gibbs energy of reaction across temperature (5.71 kJ/mol) (Fig. S4 *A*). Surprisingly, we observed a considerable increase in error (6.47 kJ/mol) after applying the correction on magnesium concentration globally (Fig. S4 *A*). We investigated this issue in detail and found that problems related to inconsistency in measured K' data (involving magnesium binding) and report of total magnesium concentration can be major sources of error, even though the correction works with well curated data (Supporting Materials and Methods). Therefore, we proceed by omitting the global correction on magnesium concentration from our procedure.

Additionally, we compared our method to the most recent method by predicting $\Delta_r G^\circ$ for 185 new reactions collected in this work, using the 432 overlapping previous reactions as training data. We found the median absolute residual from the current method (8.17 kJ/mol) is notably smaller than that from the previous work (11.47 kJ/mol) (Fig. S4 *B*).

To summarize, we included the Davies equation, new group definitions, and temperature transformation capabilities but not the magnesium correction in our final group-contribution framework. We used the equilibrium constants from the TECRdb and the collected $\Delta_f G^\circ$-values as the training data (Table S1, tabs 1and 3). Additionally, we used the collected pK_a data from the SC-database when possible and estimated the rest using ChemAxon (Table S1, tab 4). Overall, our method led to improved performance compared to the most recent group-contribution method while adding the capability to correct equilibrium constants with respect to temperature and substantially expanding the scope of predictions and thermodynamic datasets used in estimation.

DISCUSSION

In this work, we expanded the scope of thermodynamic calculations to more compounds and reactions with both curated and estimated data and also extended the group-contribution methods for estimating reaction-equilibrium constants to account for variations in temperature. We first collected and curated thermodynamic data including K', $\Delta_r H'^\circ$, $\Delta_f G^\circ$, $\Delta_f H^\circ$, $\Delta_f S^\circ$, and various ion-binding constants from a number of databases. We then applied an existing thermodynamic theory with simplifying assumptions to enable the calculation of Gibbs free energy of reaction across temperature and estimated the necessary parameters ($\Delta_f S^\circ$) using a linear regression model. We also estimated magnesium-binding constants for 618 compounds using molecular descriptors and magnesium-binding groups based on existing binding data. With new capabilities and new data, we utilized an updated group-contribution method to

calculate equilibrium constants with improved accuracy over previous work.

The curation of the NIST TECRdb revealed that fully specified media conditions, which influence the ionic strength and metal-ion-concentration corrections, were often lacking. Surprisingly, curating the literature and filling in media conditions did not improve the resulting fit on the estimation of equilibrium constants, with one possible cause that we added data at high ionic strengths that exceed the intended range of the Davies and Debye-Hückel models for chemical activity. Another possible source of error could be related to the relatively simple model used to account for the effect of ionic strength on activity coefficients of aqueous electrolytes. The Davies equation fails to account for specific interactions between various ions present in solution and is unable to calculate activity coefficients at temperatures other than 298.15 K. Equations with a more comprehensive handling of these thermodynamic theories are established (12–15,30,31) but require substantially more data than is currently available for the vast majority of compounds.

Utilizing reasonable assumptions of constant enthalpy and entropy over the range of biological interest, we formulated a simplified approach to calculate temperature transformation of Gibbs energy of reaction and reduced the number of parameters needed for estimation drastically. With the incorporation of temperature transformation capabilities, we obtained similar errors in estimating $\Delta_r G^\circ$ compared to the previous method (10) (Fig. S4 *A*). Such similar errors seem to be largely due to the fact that most of the data were measured not far from 298.15 K (83.5% of the data were measured under 295.15–313.15 K), resulting in a minor change in correction of K' to the reference conditions. However, we do predict large changes in the Gibbs energy of many reactions at high temperatures (approaching 373 K), which thus may be significant for high-interest thermophilic organisms such as those living in hot springs and hydrothermal vents.

The compound-specific parameter required for temperature transformation in our simplified model is $\Delta_f S^\circ$, which is missing for a large number of compounds in the TECRdb. Using a regression model, we predicted $\Delta_f S^\circ$ of a comprehensive collection of compounds with high accuracy by identifying key chemical properties such as number of atoms and partial charge. The linear correlation of other thermodynamic properties (e.g., standard molar entropy, standard partial molal volume, $\Delta_f G^\circ$) with number of atoms has been demonstrated in previous work (32–35), but only for compounds in the same homologous series. We found the partial charge of atoms to be useful to distinguish $\Delta_f S^\circ$ from different homologous series, possibly because of the fact that the partial charge of atoms of the aqueous species influences its interaction with surrounding water molecules. The regression model was unable to clearly differentiate $\Delta_f S^\circ$ of compounds within certain categories, however, such as monosaccharides and disaccharides. For example, the differences in $\Delta_f S^\circ$ for fructose, mannose, and sorbose are around 10 to 20 J/K/mol, whereas the model only predicts up to 5 J/K/mol difference because of their similar chemical properties. Such error is not evident when evaluating the accuracy of $\Delta_f S^\circ$ estimation, as $\Delta_f S^\circ$ of monosaccharides are around 1000 J/K/mol. However, when calculating $\Delta_r S^\circ$ of the isomerization reaction between monosaccharides, we found that the errors of $\Delta_f S^\circ$ prediction, though small compared to $\Delta_f S^\circ$-values, are significant compared to the calculated $\Delta_r S^\circ$-values. We observed this issue to be prevalent for a number of reactions in the NIST TECRdb. Thus, identification of new molecular properties or additional features describing group interactions to more accurately differentiate these complex carbohydrates can be a productive next step to improve $\Delta_f S^\circ$ estimation. Additionally, the error in $\Delta_f S^\circ$ estimation can be incorporated into the calculation of confidence intervals developed by the previous method (10), offering the capability to assess the error in estimating $\Delta_r G'^\circ$ at different temperatures.

We demonstrated that magnesium-binding groups (specifically the phosphate groups) that could be identified from known magnesium-binding compounds are useful features to estimate magnesium-binding constants with good accuracy. However, we found a number of compounds that complex with magnesium do not contain the binding groups we defined. These compounds include nucleobases, ribonucleosides, deoxyribonucleosides, purine derivatives, and small chemicals such as ammonia, thiocyanate, and urea. Currently, we use molecular properties to describe their binding to magnesium. Such an issue in identifying the chemical moiety responsible for magnesium binding can still make it difficult to extend our predictions to new compounds with similar structures as the compounds described above. The approach of estimating magnesium-binding constants can also be applied to other metals. However, we did not perform such predictions here because of the scarcity of binding data available for other metals.

We found the overall error in estimating $\Delta_r G^\circ$ increases with the incorporation of magnesium correction using curated and predicted magnesium binding data (Fig. S4 *A*). We identified inconsistency in K' data (with magnesium binding involved) to be one primary source of error. Another source of error can be due to the uncertainty in estimation of magnesium-binding constants and missing binding data for other metals. Additionally, most measurements only reported total metal-ion concentrations, whereas the metal-correction formulation uses free-metal-ion concentrations. Therefore, additional effort is necessary to calculate free-metal-ion concentrations from measured data. Because of the lack of binding data and uncertainty in estimated data, an iterative approach might be taken in which free-metal-ion concentrations calculated using the current binding data are applied to optimize the binding data, which

are then fed into the calculation of free-metal-ion concentrations.

The current work expands opportunities toward an understanding of thermodynamic factors underlying metabolic network and function in biological systems. This area has generated a number of exciting results, such as the discovery that amino acid biosynthesis, which is endergonic at surface conditions, is exergonic under the conditions of life in hydrothermal vents (36). Another recent effort proposed proteomic constraints because of thermodynamic bottlenecks as a critical factor underlying metabolic pathway choice (4). As methods for estimating the thermodynamic properties of metabolic networks continue to improve, these efforts are likely to be increasingly fruitful in uncovering the physical constraints driving the function and evolution of metabolic networks.

CONCLUSION

The work here provides an updated group-contribution method with an expanded set of thermodynamic data and extended capabilities to calculate equilibrium constants as a function of temperature. We collected and curated thermodynamic data for compounds and reactions from a number of databases and primary literature sources. We established a simple yet well-justified framework, which includes formulations derived from existing theory and the necessary parameters $(\Delta_f S^\circ)$, to calculate equilibrium constants as a function of temperature. We also used molecular properties and magnesium binding groups defined from existing data to estimate magnesium-binding constants for 618 compounds through a linear regression model. Taken together, this work fills a gap in previous group-contribution methods to calculate equilibrium constants to temperature conditions and better correct for magnesium-ion binding. These efforts should facilitate the growing number of applications to apply thermodynamic principles to better understand cell metabolism.

SUPPORTING MATERIAL

Supporting Materials and Methods, six figures, and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30524-1.

AUTHOR CONTRIBUTIONS

B.D. and D.C.Z. conceived and designed the study. B.D., Z.Z., S.G., J.T.Y., and D.C.Z. collected the data. B.D. and D.C.Z. performed the analysis. B.D., Z.Z., S.G., J.T.Y., B.O.P., and D.C.Z. wrote the manuscript. B.D. and J.T.Y. wrote the Supporting Materials and Methods. All authors read and approved the final content.

ACKNOWLEDGMENTS

We would like to thank Nikolaus Sonnenschein for valuable discussions. We would also like to thank the reviewers for their thoughtful comments.

This work was supported by the Novo Nordisk Foundation Grant Number NNF10CC1016517.

SUPPORTING CITATIONS

References (37–53) appear in the Supporting Material.

REFERENCES

1. Henry, C. S., L. J. Broadbelt, and V. Hatzimanikatis. 2007. Thermodynamics-based metabolic flux analysis. *Biophys. J.* 92:1792–1805.
2. Hamilton, J. J., V. Dwivedi, and J. L. Reed. 2013. Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. *Biophys. J.* 105:512–522.
3. Kümmel, A., S. Panke, and M. Heinemann. 2006. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* 2:2006.0034.
4. Noor, E., A. Bar-Even, …, R. Milo. 2014. Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Comput. Biol.* 10:e1003483.
5. Beard, D. A., K. C. Vinnakota, and F. Wu. 2008. Detailed enzyme kinetics in terms of biochemical species: study of citrate synthase. *PLoS One.* 3:e1825.
6. Goldberg, R. N., Y. B. Tewari, and T. N. Bhat. 2004. Thermodynamics of enzyme-catalyzed reactions–a database for quantitative biochemistry. *Bioinformatics.* 20:2874–2877.
7. Mavrovouniotis, M. L. 1990. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* 36:1070–1082.
8. Jankowski, M. D., C. S. Henry, …, V. Hatzimanikatis. 2008. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* 95:1487–1499.
9. Noor, E., A. Bar-Even, …, R. Milo. 2012. An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics.* 28:2037–2044.
10. Noor, E., H. S. Haraldsdóttir, …, R. M. Fleming. 2013. Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* 9:e1003098.
11. Alberty, R. A. 2003. Thermodynamics of Biochemical Reactions. John Wiley & Sons, Hoboken, NJ.
12. Helgeson, H. C., and D. H. Kirkham. 1974. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures; I, Summary of the thermodynamic/electrostatic properties of the solvent. *Am. J. Sci.* 274:1089–1198.
13. Helgeson, H. C., and D. H. Kirkham. 1974. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures; II, Debye-Huckel parameters for activity coefficients and relative partial molal properties. *Am. J. Sci.* 274:1199–1261.
14. Helgeson, H. C., and D. H. Kirkham. 1976. Theoretical prediction of the thermodynamic properties of aqueous electrolytes at high pressures and temperatures. III. Equation of state for aqueous species at infinite dilution. *Am. J. Sci.* 276:97–240.
15. Helgeson, H. C., D. H. Kirkham, and G. C. Flowers. 1981. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes by high pressures and temperatures; IV, calculation of activity coefficients, osmotic coefficients, and apparent molal and standard and relative partial molal properties to 600 degrees C and 5kb. *Am. J. Sci.* 281:1249–1516.
16. Shock, E. L., and H. C. Helgeson. 1988. Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: correlation algorithms for ionic species and equation of state predictions to 5 kb and 1000°C. *Geochim. Cosmochim. Acta.* 52:2009–2036.
17. Plyasunov, A. V., and E. L. Shock. 2001. Correlation strategy for determining the parameters of the revised Helgeson-Kirkham-Flowers

model for aqueous nonelectrolytes. *Geochim. Cosmochim. Acta.* 65:3879–3900.

18. Plyasunov, A. V., J. P. O. Connell, …, E. L. Shock. 2001. Semiempirical equation of state for the infinite dilution thermodynamic functions of hydration of nonelectrolytes over wide ranges of temperature and pressure. *Fluid Phase Equilib.* 183:133–142.
19. Johnson, J. W., E. H. Oelkers, and H. C. Helgeson. 1992. SUPCRT92: a software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C. *Comput. Geosci.* 18:899–947.
20. Plyasunova, N. V., A. V. Plyasunov, and E. L. Shock. 2004. Database of thermodynamic properties for aqueous organic compounds. *Int. J. Thermophys.* 25:351–360.
21. Pettit, L. D., and K. J. Powell. 2006. The IUPAC stability constants database. *Chemistry International – Newsmagazine for IUPAC.* 28:14–15.
22. Kortüm, G., and K. Andrussow. 1961. Dissociation Constants of Organic Acids in Aqueous Solution. Butterworths, London, UK.
23. Perrin, D. D. 1965. Dissociation Constants of Organic Bases in Aqueous Solution. Butterworths, London, UK.
24. Alberty, R. A. 1968. Effect of pH and metal ion concentration on the equilibrium hydrolysis of adenosine triphosphate to adenosine diphosphate. *J. Biol. Chem.* 243:1337–1343.
25. Davies, C. W. 1938. 397. 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.* 0:2093–2098.
26. Pedregosa, F., G. Varoquaux, …, É. Duchesnay. 2011. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12:2825–2830.
27. Goldberg, R. N., and Y. B. Tewari. 1991. Thermodynamics of the disproportionation of adenosine 5′-diphosphate to adenosine 5′-triphosphate and adenosine 5′-monophosphate. I. Equilibrium model. *Biophys. Chem.* 40:241–261.
28. Larson, J. W., Y. B. Tewari, and R. N. Goldberg. 1993. Thermochemistry of the reactions between adenosine, adenosine 5′-monophosphate, inosine, and inosine 5′-monophosphate; the conversion of d-histidine to (urocanic acid+ammonia). *J. Chem. Thermodyn.* 25:73–90.
29. Cayley, S., B. A. Lewis, …, M. T. Record, Jr. 1991. Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo. *J. Mol. Biol.* 222:281–300.
30. Pitzer, K. S., and J. J. Kim. 1974. Thermodynamics of electrolytes. IV. Activity and osmotic coefficients for mixed electrolytes. *J. Am. Chem. Soc.* 96:5701–5707.
31. Elizalde, M. P., and J. L. Aparicio. 1995. Current theories in the calculation of activity coefficients-II. Specific interaction theories applied to some equilibria studies in solution chemistry. *Talanta.* 42:395–400.
32. Helgeson, H. C. 1992. Calculation of the thermodynamic properties and relative stabilities of aqueous acetic and chloroacetic acids, acetate and chloroacetates, and acetyl and chloroacetyl chlorides at high and low temperatures and pressures. *Appl. Geochem.* 7:291–308.
33. Shock, E. L. 1995. Organic acids in hydrothermal solutions: standard molal thermodynamic properties of carboxylic acids and estimates of dissociation constants at high temperatures and pressures. *Am. J. Sci.* 295:496–580.
34. Schulte, M. D., and K. L. Rogers. 2004. Thiols in hydrothermal solution: standard partial molal properties and their role in the organic geochemistry of hydrothermal environments. *Geochim. Cosmochim. Acta.* 68:1087–1097.
35. Schulte, M. D., and E. L. Shock. 1993. Aldehydes in hydrothermal solution: standard partial molal thermodynamic properties and relative stabilities at high temperatures and pressures. *Geochim. Cosmochim. Acta.* 57:3835–3846.
36. Amend, J. P., and E. L. Shock. 1998. Energetics of amino acid synthesis in hydrothermal ecosystems. *Science.* 281:1659–1662.
37. Ono, K., K. Yanagida, …, K. Soda. 2006. Alanine racemase of alfalfa seedlings (Medicago sativa L.): first evidence for the presence of an amino acid racemase in plants. *Phytochemistry.* 67:856–860.
38. Woolf, B. 1929. Some enzymes in B. coli communis which act on fumaric acid. *Biochem. J.* 23:472–482.
39. Quastel, J. H., and B. Woolf. 1926. The equilibrium between l-aspartic acid, fumaric acid and ammonia in presence of resting bacteria. *Biochem. J.* 20:545–555.
40. Siekevitz, P., and V. R. Potter. 1953. The adenylate kinase of rat liver mitochondria. *J. Biol. Chem.* 200:187–196.
41. Nishizuka, Y., M. Takeshita, …, O. Hayaishi. 1959. beta-Alanine-alpha-alanine transaminase of Pseudomonas. *Biochim. Biophys. Acta.* 33:591–593.
42. Nixon, P. F., and R. L. Blakley. 1968. Dihydrofolate reductase of Streptococcus faicium. II. Purification and some properties of two dihydrofolate reductases from the amethopterin-resistant mutant Streptococcus faecium var. Durans strain A. *J. Biol. Chem.* 243:4722–4731.
43. Blasi, F., F. Fragomele, and I. Covelli. 1969. Thyroidal phenylpyruvate tautomerase. Isolation and characterization. *J. Biol. Chem.* 244:4864–4870.
44. Haagensen, P., L. G. Karlsen, …, J. Villadsen. 1983. The kinetics of penicillin-V deacylation on an immobilized enzyme. *Biotechnol. Bioeng.* 25:1873–1895.
45. Hassan Ansari, N. C. P., and L. Stevens. 1985. Effects of high concentrations of proteins on the equilibrium and kinetic properties of four enzymes. *Biochem. Soc. Trans.* 13:362.
46. Huber, R. E., and K. L. Hurlburt. 1986. Reversion reactions of β-galactosidase (Escherichia coli). *Arch. Biochem. Biophys.* 246:411–418.
47. Johansson, E., L. Hedbys, …, S. Svensson. 1989. Studies of the reversed α-mannosidase reaction in high concentrations of mannose. *Enzyme Microb. Technol.* 11:347–352.
48. Hori, N., M. Watanabe, and Y. Mikami. 1991. The effects of organic solvent on the ribosyl transfer reaction by thermostable purine nucleoside phosphorylase and pyrimidine nucleoside phosphorylase from bacillus stearothermophilus JTS 859. *Biocatalysis.* 4:297–304.
49. Manchester, J., G. Walkup, …, Z. You. 2010. Evaluation of pKa estimation methods on 211 druglike compounds. *J. Chem. Inf. Model.* 50:565–571.
50. Settimo, L., K. Bellman, and R. M. Knegtel. 2014. Comparison of the accuracy of experimental and predicted pKa values of basic and acidic compounds. *Pharm. Res.* 31:1082–1095.
51. Newville, M., T. Stensitzki, …, A. Nelson. 2016. Lmfit: Non-Linear Least-Square Minimization and Curve-Fitting for Python. Astrophysics Source Code Library https://lmfit.github.io/lmfit-py/.
52. Levenberg, K. 1944. A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* 2:164–168.
53. Marquardt, D. 1963. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* 11:431–441.

Biophysical Journal, Volume *114*

Supplemental Information

Temperature-Dependent Estimation of Gibbs Energies Using an Updated Group-Contribution Method

Bin Du, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, and Daniel C. Zielinski

Temperature-dependent estimation of Gibbs energies using an updated group contribution method
Supporting Material

Bin Du, Zhen Zhang, Sharon Grubner, James T. Yurkovich, Bernhard O. Palsson, Daniel C. Zielinski

Department of Bioengineering, University of California, San Diego, La Jolla, CA, United States, 92093-0412

Transformation of standard Gibbs free energy of formation ($\Delta_f G^\circ$) of aqueous species across temperature

Considering constant pressure, the standard Gibbs free energy of formation of an aqueous species at a given temperature T and the reference temperature T_r (298.15 K) can be written as $\Delta_f G^\circ_T$ and $\Delta_f G^\circ_{T_r}$, respectively. Based on the Second law of thermodynamics,

$$\Delta_f G^\circ_T = \Delta_f H^\circ_T - T\Delta_f S^\circ_T \quad [1]$$

$$\Delta_f G^\circ_{T_r} = \Delta_f H^\circ_{T_r} - T_r\Delta_f S^\circ_{T_r}. \quad [2]$$

Subtracting Equation 1 by Equation 2, we get

$$\begin{aligned}\Delta_f G^\circ_T = \Delta_f G^\circ_{T_r} + (\Delta_f H^\circ_T - \Delta_f H^\circ_{T_r}) - \\ T(\Delta_f S^\circ_T - \Delta_f S^\circ_{T_r}) - (T - T_r)\Delta_f S^\circ_{T_r}\end{aligned} \quad [3]$$

Using the definition of enthalpy and entropy in terms of heat capacity at constant pressure (1, 2), Equation 3 is expressed as

$$\begin{aligned}\Delta_f G^\circ_T = \Delta_f G^\circ_{T_r} + \int_{T_r}^{T} C_{P_r} dT - \\ T\int_{T_r}^{T} C_{P_r} d\ln T - (T - T_r)\Delta_f S^\circ_{T_r}\end{aligned} \quad [4]$$

where C_{P_r} is the heat capacity of the aqueous species at T_r. It is worth mentioning that the formulation above is slightly different from that in the geochemistry literature (1, 2), where we replaced $S^\circ_{T_r}$ with $\Delta_f S^\circ_{T_r}$.

Based on Shock et al. (2), the heat capacity of an aqueous species is a function of temperature and depends on three parameters c_1, c_2, and ω, which are different for different aqueous species. We found that heat capacities of aqueous species at different temperatures generally vary in a small range from their values at T_r. Specifically, we examined a total of 399 compounds with data available (3) and found that their C_P values at different temperatures vary maximally around 16% from their C_{P_r} values, for the temperature range we are working with (283.5 K to 360.5 K). The temperature range is based on temperatures of measured data in TECRdb (4). Additionally, the maximum variation in C_P across temperature is smaller than that across different compounds, as shown in Figure S1A. Thus, given the assumption that heat capacity is a constant with respect to temperature, we take integrals in Equation 4 and get

$$\begin{aligned}\Delta_f G^\circ_T = \Delta_f G^\circ_{T_r} + C_{P_r}(T - T_r) - \\ TC_{P_r}\ln\left(\frac{T}{T_r}\right) - (T - T_r)\Delta_f S^\circ_{T_r}.\end{aligned} \quad [5]$$

Combining the terms involving C_{P_r}, we have

$$\begin{aligned}\Delta_f G^\circ_T = \Delta_f G^\circ_{T_r} + \left[T - T_r - T\ln\left(\frac{T}{T_r}\right)\right] C_{P_r} - \\ (T - T_r)\Delta_f S^\circ_{T_r}\end{aligned} \quad [6]$$

From Equation 6, we have the term involving C_{P_r} and the term involving $\Delta_f S^\circ_{T_r}$ together affecting the change of standard Gibbs free energy of formation across temperature.

We define the coefficient in front of C_{P_r} to be a ($a = T - T_r - T\ln\left(\frac{T}{T_r}\right)$) and the coefficient in front of $\Delta_f S^\circ_{T_r}$ to be b ($b = T - T_r$). Comparing the magnitude of a and b as a function of temperature, we found that a is much smaller than b (Figure S1B). The value of a/b is at most 0.025 for the most frequent temperatures of TECRdb measured data (295.5 K to 313.5 K), and at most 0.1 in the overall temperature range of interest.

Given that C_{P_r} and $\Delta_f S^\circ_{T_r}$ of the same aqueous species are generally on the same order of magnitude (Figure S1C) and C_{P_r} coefficient is much smaller than $\Delta_f S^\circ_{T_r}$ coefficient, it is reasonable to neglect the term involving C_{P_r} in Equation 6. Thus, we have

$$\Delta_f G^\circ_T = \Delta_f G^\circ_{T_r} - (T - T_r)\Delta_f S^\circ_{T_r} \quad [7]$$

to transform the standard Gibbs free energy of formation of an aqueous species across temperature.

Equilibrium constant as a function of pH, temperature, ionic strength and metal ion concentration

In aqueous solutions, each compound exists as several different pseudoisomer forms distributed according to the Boltzmann distribution. The pseudoisomer forms refer to the different protonation and ion bound states of the same compound (5, 6). For example, the pseudoisomer forms of orthophosphate include but are not limited to PO_4^{3-} and $MgPO_4^-$. Adapted from Alberty (6) and the formulation in the last section, the standard transformed Gibbs free energy of formation of pseudoisomer i ($\Delta_f G'^\circ_i$) of a given compound under certain pH, temperature (T), ionic strength (I) and metal ion concentration (pM) is expressed as

$$\Delta_f G_i'^\circ = \Delta_f G_i^\circ(I=0,T_r) - (T-T_r)\Delta_f S_i^\circ + N_H(i)RT\ln(10)\text{pH} - N_M(i)(\Delta_f G_M^\circ(T) - RT\ln(10)\text{pM}) - RT\alpha(z_i^2 - N_H(i))\left(\frac{\sqrt{I}}{1+\sqrt{I}} - 0.3I\right) \quad [8]$$

where $\Delta_f S_i^\circ$ is the standard entropy change of formation of pseudoisomer i at 298.15 K, z_i, $N_H(i)$, and $N_M(i)$ are the charge, number of hydrogen atoms and number of metal ions M bound to pseudoisomer i (due to availability of metal binding data, we only handle pseudoisomer form bound with at most one type of metal ion), $\Delta_f G_M^\circ(T)$ is the standard Gibbs free energy of formation of aqueous ionic metal species M at T (can be calculated using equations and data from Shock et al. (2)), pM (pM $= -\log_{10}[M^{m+}]$) is the potential of ionic metal species M with concentration [M] and charge +m in aqueous solutions, and α is the Debye-Hückel Constant and is temperature dependent (6). The correction on ionic strength is based on Davies equation, which is an empirical extension of Debye–Hückel theory and can be used to calculate activity coefficients of electrolytes at relatively high ion concentrations (7).

The standard transformed Gibbs free energy of formation of the compound ($\Delta_f G_j'^\circ$) can be calculated based on the energies of its pseudoisomer forms using Legendre transform (6):

$$\Delta_f G_j'^\circ = -RT\ln\left\{\sum_{i=1}^{N_{iso}}\exp\left[-\frac{\Delta_f G_i'^\circ}{RT}\right]\right\}. \quad [9]$$

Additionally, the equilibrium mole fraction m_i of the ith pseudoisomer in the pseudoisomer group is given by

$$m_i = \exp\left\{\frac{\Delta_f G_j'^\circ - \Delta_f G_i'^\circ}{RT}\right\} \quad [10]$$

The standard transformed Gibbs free energy of reaction ($\Delta_r G'^\circ$) can thus be calculated based on the energies of its participating compounds ($\Delta_f G_j'^\circ$) and their corresponding stoichiometries (r_j) in the reaction

$$\Delta_r G'^\circ = \sum_{j=1}^{N} r_j \Delta_f G_j'^\circ \quad [11]$$

Thus, we are able to calculate thermodynamics of the reaction as a function of pH, temperature, ionic strength and metal ion concentrations.

Under a specified condition, we can identify the dominant pseudoisomer form for a compound (the form with the largest concentration). Such dominant form also has a dominant contribution to the Gibbs free energy of formation of the compound, according to Equation 10 ($m_i = 1$ when $\Delta_f G_j'^\circ = \Delta_f G_i'^\circ$). Therefore, the transformation of $\Delta_r G'^\circ$ across temperature can be calculated as $\Delta_r S_{T_r}^\circ = \sum_{j=1}^{N} r_j \Delta_f S_j^\circ$, where $\Delta_f S_j^\circ$ of the compound can be approximated to that of its dominant pseudoisomer form. We thus have

$$\Delta_r G_T'^\circ = \Delta_r G_{T_r}'^\circ - (T-T_r)\Delta_r S_{T_r}^\circ \quad [12]$$

The reaction equilibrium constant can thus be calculated through the equation

$$\Delta_r G'^\circ = -RT\ln K'. \quad [13]$$

The above procedures can also be used to transform the measured equilibrium constants to $\Delta_r G^\circ$ at the reference state (298.15 K, pH 7, 0M ionic strength, no metal ion), by applying corrections on pH, ionic strength and metal ion concentrations as in Equation 8 and correction on temperature as in Equation 12. Then, we can use corrected $\Delta_r G^\circ$ data to estimate $\Delta_r G^\circ$ and $\Delta_f G^\circ$ for new reactions and compounds, based on the latest group contribution method, termed component contribution (8).

Example of binding constant and binding polynomial formulation

We introduce the concept of binding constant and describe its relationship with the binding polynomial. Binding constant describes the equilibrium of binding and unbinding reaction between a receptor (compound) and a ligand (proton, metal ion). Here, we specifically refer the binding constant to be the equilibrium constant of the unbinding step. For example, a reactant is composed of three ion bound states: A (with least hydrogens and metal ions bound), HA (A bound with H^+), MgHA (A bound with H^+ and Mg^{2+}). There are two binding steps between A and MgHA: HA $\rightleftharpoons$ A + H^+ and MgHA $\rightleftharpoons$ HA + Mg^{2+}. The respective binding constants are

$$K_1 = \frac{[\text{A}][\text{H}^+]}{[\text{HA}]} \quad [14]$$

$$K_2 = \frac{[\text{HA}][\text{Mg}^{2+}]}{[\text{MgHA}]} \quad [15]$$

For practical purposes, it is more convenient to express the logarithmic form of the constants, where $\text{p}K_1 = -\log_{10}K_1$ and $\text{p}K_2 = -\log_{10}K_2$. Based on the type of ligand, $\text{p}K_1$ is known as the acid dissociation constant ($\text{p}K_a$) and $\text{p}K_2$ is the stability constant for magnesium binding ($\text{p}K_{Mg}$). The logarithmic form of the binding constant is what we used for estimation in regression models and calculation in the group contribution framework.

Binding polynomial gives the partition of a reactant between various aqueous species that make it up. Binding polynomial is the measure of the difference in Gibbs energy between one ion bound state and another. For convenience of calculation, we usually write the binding polynomial of an ion bound state with respect to the one with the least hydrogens and metal ions bound. Thus, the binding polynomial P of MgHA is defined as (6):

$$P = \frac{[\text{A}] + [\text{HA}] + [\text{MgHA}]}{[\text{A}]} \quad [16]$$

Substituting Equations 14 and 15 into 16, we get

$$P = 1 + \frac{[\text{H}^+]}{K_1} + \frac{[\text{H}^+][\text{Mg}^{2+}]}{K_1 K_2} \quad [17]$$

The energy difference between A and MgHA is $-RT\ln P$, which can be used in Equation 9 of the main text and calculate $\Delta_f G'^\circ$ of the reactant. Therefore, binding polynomial can be expressed in terms of proton and metal ion concentrations, as well as the binding constants of different binding steps. This example can be extended to any other ion bound states with defined number of hydrogens and metal ions bound.

Case studies on correcting K' data measured at different magnesium concentrations

Through several case studies, we examined how well the magnesium binding constants correct K' data measured at different magnesium concentrations to the same reference conditions. Specifically, we transformed the $\Delta_r G'^{\circ}$ data calculated from K' ($\Delta_r G'^{\circ} = -RT \ln K'$) to the reference state $\Delta_r G^{\circ}$ (298.15 K, pH 7, 0 M ionic strength, no metal ion) using Legendre transforms (6). The resulting reference state $\Delta_r G^{\circ}$ values should be within a small range, which would indicate that the correction to Gibbs energy for magnesium binding is accurate. Taking data from the reaction catalyzed by adenylate kinase, one of the best characterized reactions, as an example, we found a substantial decrease in $\Delta_r G^{\circ}$ variation with respect to magnesium concentration after applying corrections to account for magnesium binding (Figure S2C). We observed similar trend in arginine kinase (Figure S2D) and creatine kinase reactions (Figure S2E), when accounting for the binding of ATP and ADP to magnesium. We found more cases where applying correction to account for magnesium binding reduced the variation in $\Delta_r G^{\circ}$ significantly (Figure S2F-S2H).

However, in some cases, the differences in $\Delta_r G^{\circ}$ remained substantial (Figure S2I-S2K). One example is the dataset from the hexokinase reaction, where we also applied the correction on the binding of ATP and ADP to magnesium. To address such inconsistency in magnesium correction, which we hypothesized to be errors in measured binding constants, we attempted to adjust the binding constants through optimization to maximize the correction on K' data at different magnesium concentrations. Specifically, we optimized the binding constants of ATP, ADP, AMP and glucose 6-phosphate together using a Levenberg-Marquardt algorithm that minimizes the squared distance from the average inferred $\Delta_r G^{\circ}$ values of hexokinase and adenylate kinase reactions (9, 10). However, we found that while the optimized binding constants resulted in a smaller variation in $\Delta_r G^{\circ}$ for data in hexokinase reaction (Figure S3A), the variation in $\Delta_r G^{\circ}$ for adenylate kinase reaction increases significantly using those optimized values (Figure S3B). We observed similar trend for arginine kinase reaction when optimizing its data together with data from the hexokinase reaction. The optimized binding constants resulted in much greater variation in $\Delta_r G^{\circ}$ for arginine kinase reaction data compared to $\Delta_r G^{\circ}$ values calculated from the original binding data (Figure S3C). We also observed such inconsistency in magnesium correction for data from different sources of the same reaction, where the optimized binding constants of aconitase reaction failed to reduce the variation in $\Delta_r G^{\circ}$ across all three datasets (Figure S3D-S3F). Additionally, we noted that the dataset where magnesium correction did not help (Figure S3F) had reported total magnesium concentrations rather than the more direct free concentrations. However, after applying the free magnesium concentrations calculated from the total substrate and magnesium concentrations reported, we found that the variation in $\Delta_r G^{\circ}$ remained large.

To summarize, we found that magnesium binding correction works well in cases where high quality K' data and magnesium binding constants are available. However, issues such as inconsistency in measured data (involving magnesium binding) and report of total magnesium concentration exist, which can be problematic when applying the correction on magnesium binding. These issues help explain why the fit is worse when applying the magnesium correction globally (main text), even though the correction works with well curated data (e.g. adenylate kinase reaction). Therefore, we proceed by omitting the global magnesium binding correction from our procedure.

Optimization of ion binding constants using the Levenberg-Marquardt algorithm

For selected reactions with compound magnesium binding constants to optimize, we first collected K' data measured at different magnesium concentrations. We then formulated equations that correct the standard transformed Gibbs energy of reaction ($\Delta_r G'^{\circ}$) (calculated from K') to $\Delta_r G^{\circ}$, where magnesium binding constants are variables in the equations. We allowed $\pm$ 0.5 (unitless) variation for each ion binding constant from its original value, consistent with reported error in these parameters (11, 12). We optimized the binding constants using an iterative Levenberg-Marquardt algorithm with decreasing step sizes for the gradient approximation parameter. At each iteration, we input the optimized values from the previous iteration into the transformation equations and calculated the squared distance from the average inferred $\Delta_r G^{\circ}$ values. The termination criterion for the optimization was a fractional difference in the sum of squares between two consecutive iterations below 0.00001 (unitless). The Levenberg-Marquardt algorithm was performed using python package lmfit 0.9.2 (13).

Considering p number of equations of the same reaction and q number of $\mathrm{p}K_{\mathrm{Mg}}$ values to optimize, the optimization program is as follows

$$\min_{\mathrm{p}K_{\mathrm{Mg_1}},\ldots,\mathrm{p}K_{\mathrm{Mg_q}}} \sum_{i=1}^{p} (\Delta_r G_i^{\circ}(\mathrm{p}K_{\mathrm{Mg_1}},\ldots,\mathrm{p}K_{\mathrm{Mg_q}}) - \Delta_r G_{avg}^{\circ})^2 \quad [18]$$

$$\mathrm{p}K_{\mathrm{Mg_{j,data}}} - 0.5 \le \mathrm{p}K_{\mathrm{Mg_j}} \le \mathrm{p}K_{\mathrm{Mg_{j,data}}} + 0.5 \ (j = 1, 2, \ldots, q) \quad [19]$$

where $\Delta_r G_i^{\circ}(\mathrm{p}K_{\mathrm{Mg_1}},\ldots,\mathrm{p}K_{\mathrm{Mg_q}})$ is the standard transformed Gibbs energy of reaction as a function of $\mathrm{p}K_{\mathrm{Mg_1}},\ldots,\mathrm{p}K_{\mathrm{Mg_q}}$ and $\Delta_r G_{avg}^{\circ} = \frac{\sum_{i=1}^{p} \Delta_r G_i^{\circ}}{p}$. If there are multiple reactions used together to optimize the set of $\mathrm{p}K_{\mathrm{Mg}}$ values, we apply the same procedures except that the residual of each equation is with respect to the $\Delta_r G_{avg}^{\circ}$ of its corresponding reaction. The optimized $\mathrm{p}K_{\mathrm{Mg}}$ values are then applied on those reaction data to check the consistency of optimized values in reducing variation of $\Delta_r G^{\circ}$ values in different reactions.

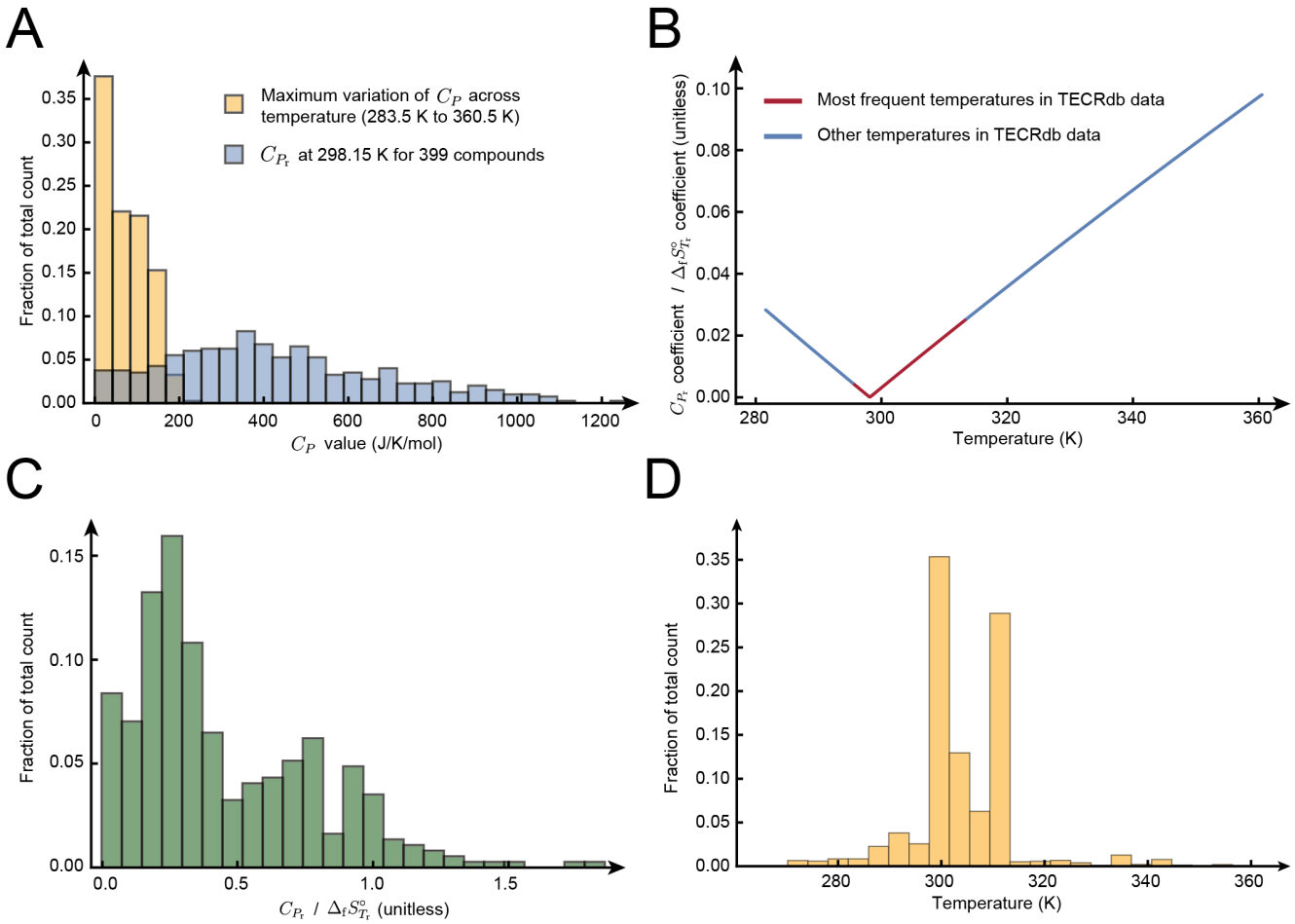

Fig. S1. (A) Order of magnitude comparison between maximum variation of C_P across temperature vs. C_{P_r} at 298.15 K for 399 compounds (3). (B) The ratio of C_{P_r} coefficient to $\Delta_f S^\circ_{T_r}$ coefficient as a function of temperature. The temperature range is based on the distribution of temperatures for measured equilibrium constants in TECRdb. (C) The distribution of $C_{P_r}/\Delta_f S^\circ_{T_r}$ of the same aqueous species from a total 370 aqueous species collected. (D) Distribution of temperature for all measured equilibrium constants in TECRdb.

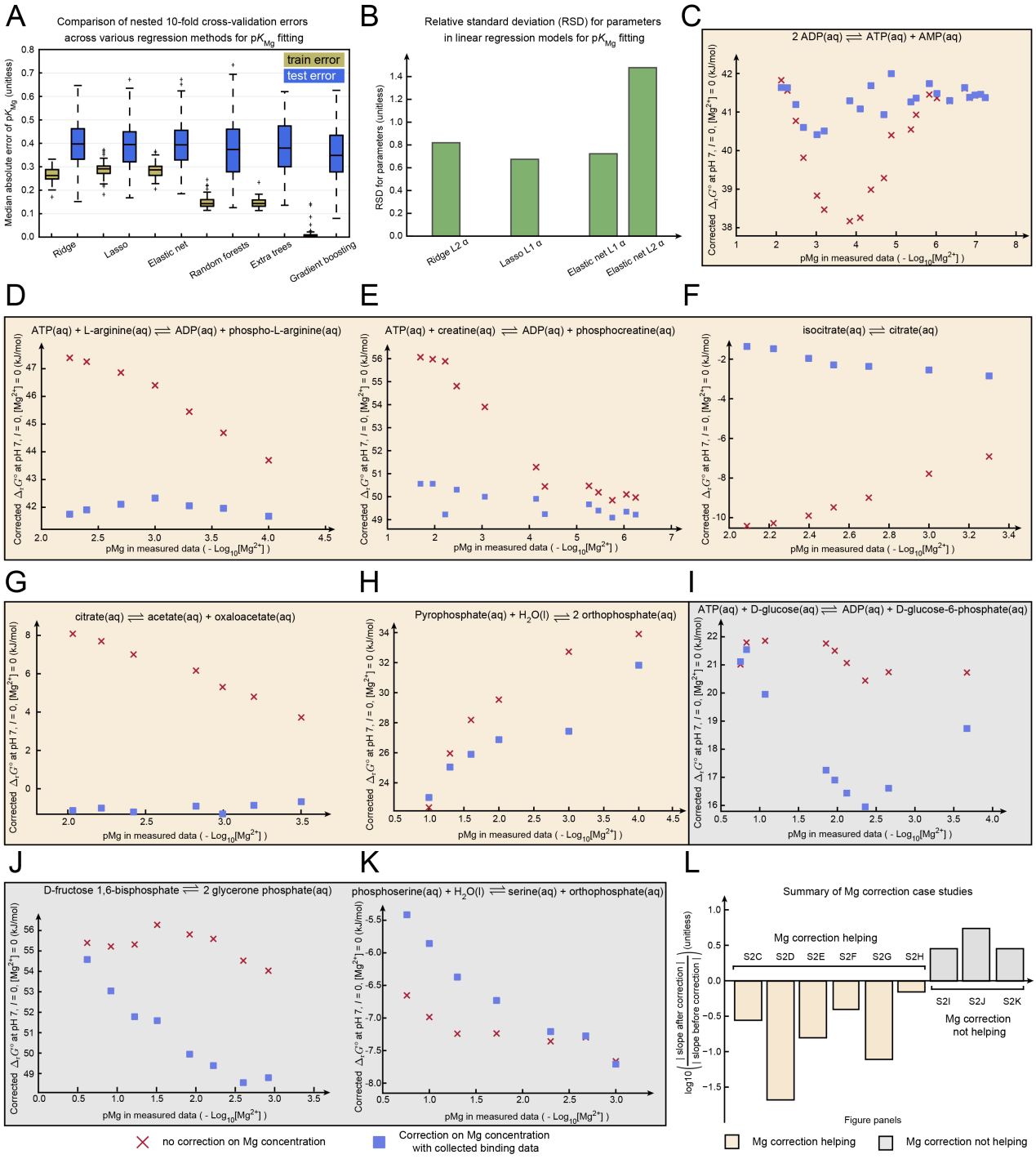

Fig. S2. (A) Training and testing errors of nested 10-fold cross validation on magnesium (Mg) binding data using the ridge regression, lasso regression, elastic net regularization, random forests, extra trees and gradient boosting. We repeated cross-validation 5 times by splitting all Mg binding data into different subdivisions. We included a total of 140 Mg binding data points and 128 features including metal binding groups, the partial charge, and molecular properties from ChemAxon and RDKit. (B) Relative standard deviation (RSD) for parameters in linear regression models used for Mg binding fitting. We calculated the mean and standard deviation of parameters selected by the inner loops of nested cross validation (repeated 5 different times). We used RSD (standard deviation/mean) to assess the relative variability of the parameters and model stability in Mg binding fitting. (C-K) Case studies on corrected $\Delta_r G^\circ$ values of different reactions calculated from equilibrium constants (K') measured at different Mg concentrations. We applied different corrections to transform $\Delta_r G'^\circ$ (calculated from K') to $\Delta_r G^\circ$ values: no correction on Mg concentrations (red Xs) and correction on Mg concentrations using collected binding constants (blue squares). The reaction whose $\Delta_r G'^\circ$ data are used to optimize the binding constants can be found in Table S12. Panels with yellow background are cases where applying Mg binding constants reduces the variation in $\Delta_r G^\circ$, while those with gray background are cases that did not help. (J) Summary of Mg correction case studies shown in different Figure panels (x axis label). We calculated the log10 value of the ratio between slope of $\Delta_r G^\circ$ values after correction and that of $\Delta_r G^\circ$ values before correction with respect to Mg concentrations. A negative log10 ratio corresponds to the case where Mg correction helps reduce the variation in $\Delta_r G^\circ$ values. pK_{Mg}: magnesium binding constant.

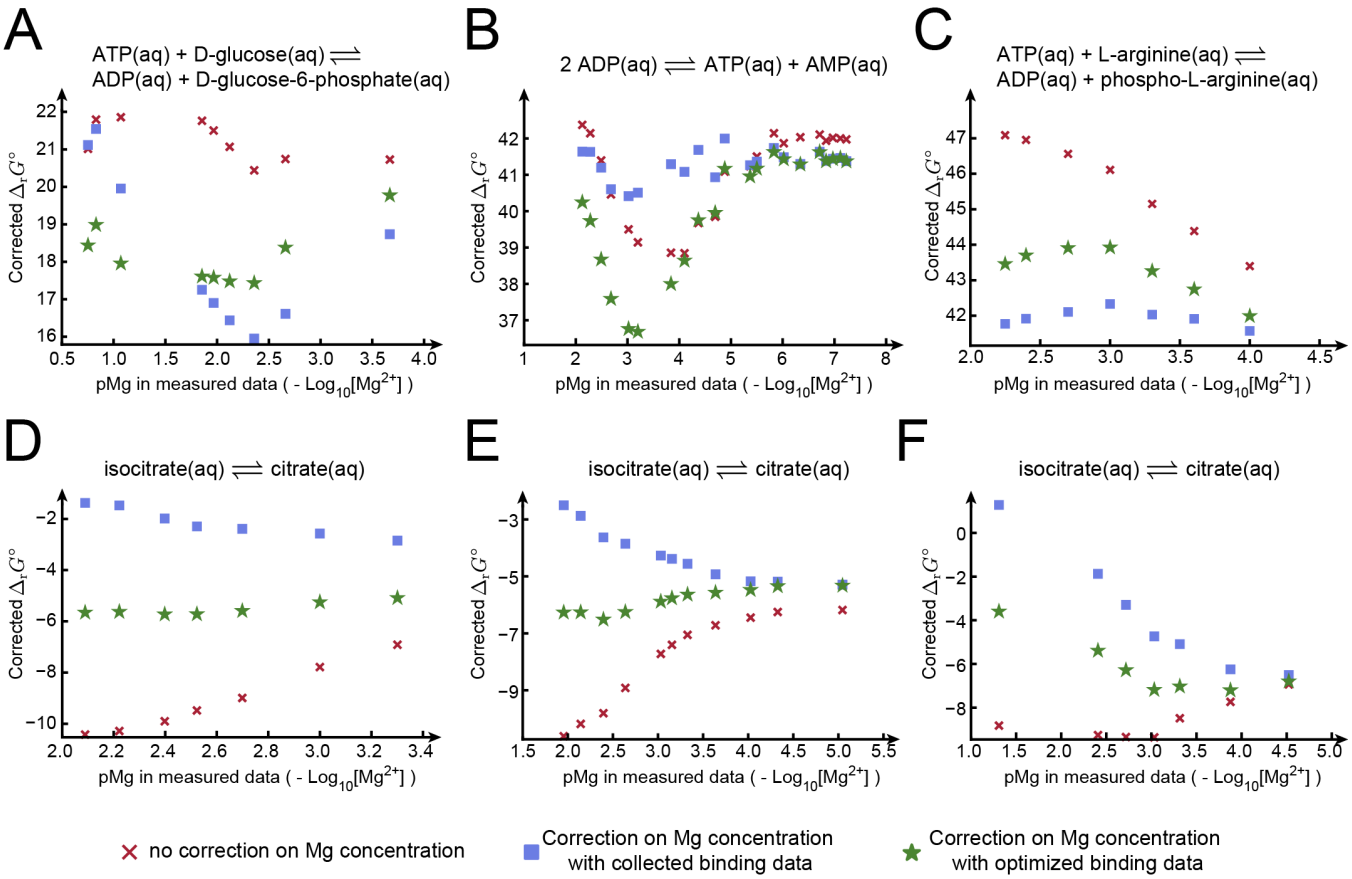

Fig. S3. (A) Corrected $\Delta_r G^\circ$ values of hexokinase reaction calculated from equilibrium constants (K') measured at different Mg concentrations. We applied different corrections to transform $\Delta_r G'^\circ$ (calculated from K') to $\Delta_r G^\circ$ values: no correction on Mg concentrations (red Xs), correction on Mg concentrations using collected binding constants (blue squares), correction on Mg concentrations using optimized binding constants (green stars). Ideally, the difference between standard $\Delta_r G^\circ$ values after correction is 0. The optimized binding constants are obtained by minimizing the least-squares errors on $\Delta_r G^\circ$ values of hexokinase and adenylate kinase reactions. (B) Corrected $\Delta_r G^\circ$ values of adenylate kinase reaction calculated from equilibrium constants (K') measured at different Mg concentrations. The labels of $\Delta_r G^\circ$ values are the same as panel A, so do the optimized binding constants used. (C) Corrected $\Delta_r G^\circ$ values of arginine kinase reaction calculated from equilibrium constants (K') measured at different Mg concentrations. The labels of $\Delta_r G^\circ$ values are the same as panel A, so do the optimized binding constants used. (D-F) Case studies on corrected $\Delta_r G^\circ$ values of aconitase reaction calculated from K' measured at different Mg concentrations. The different panels represent data from different literature sources. We found that using $\mathrm{p}K_{\mathrm{Mg}}$ data or optimized $\mathrm{p}K_{\mathrm{Mg}}$ values helped reduce the variation in standard $\Delta_r G^\circ$ values (green stars and blue squares) for panel D and E. However, in panel F, the variation in $\Delta_r G^\circ$ values is still considerably large, whether using $\mathrm{p}K_{\mathrm{Mg}}$ data or optimized values to correct on Mg concentration.

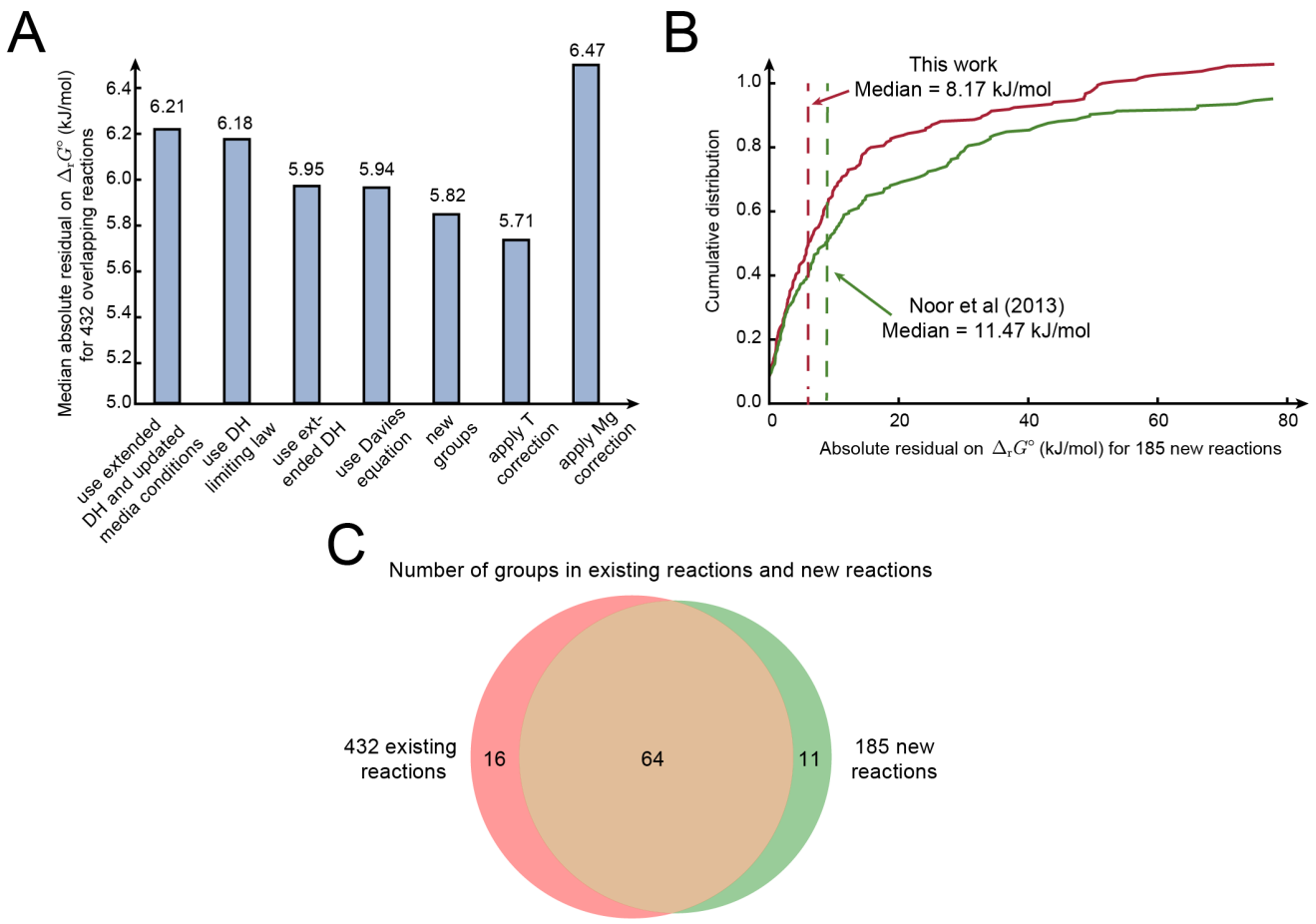

Fig. S4. (A) Comparison of absolute residuals on estimating $\Delta_r G^\circ$ for 432 reactions with various modifications on data and methods. We applied modifications on data or method one at a time sequentially and evaluated the change of median absolute residual from 10-fold cross-validation repeated 100 times. First, we used the extended Debye-Hückel (DH) equation (as used in Noor et al (8)) and updated media conditions, obtaining a median absolute residual of 6.21 kJ/mol. Removing the data with high ionic strength (> 0.5 M), we obtained an error of 5.95 kJ/mol. Next using the Davies equation, the error is 5.94 kJ/mol. We next included new compound groups in current work (5.82 kJ/mol), next with temperature correction in current work (5.71 kJ/mol) and finally with metal correction in current work (6.47 kJ/mol). We also compared the different equations to correct for the effect of ionic strength side by side (updated media condition data with ionic strength $\leq$ 0.5M) and showed that using DH limiting law results in higher error than the other two. The median absolute residual is 6.18 kJ/mol using DH limiting law, while 5.95 kJ/mol using extended DH equation and 5.94 kJ/mol using the Davies equation. (B) Comparison of absolute residuals on estimating 185 new reactions in the current method. We calculated $\Delta_r G^\circ$ for 185 new reactions by constructing the group contribution model using $\Delta_r G^\circ$ values of 432 overlapping reactions from the previous method and the current method. We then calculated the absolute residual between estimated $\Delta_r G^\circ$ and $\Delta_r G^\circ$ data for those 185 reactions. (C) Comparison of group coverage between 432 reactions in the previous group contribution method and 185 new reactions added in the current method. DH: Debye-Hückel.

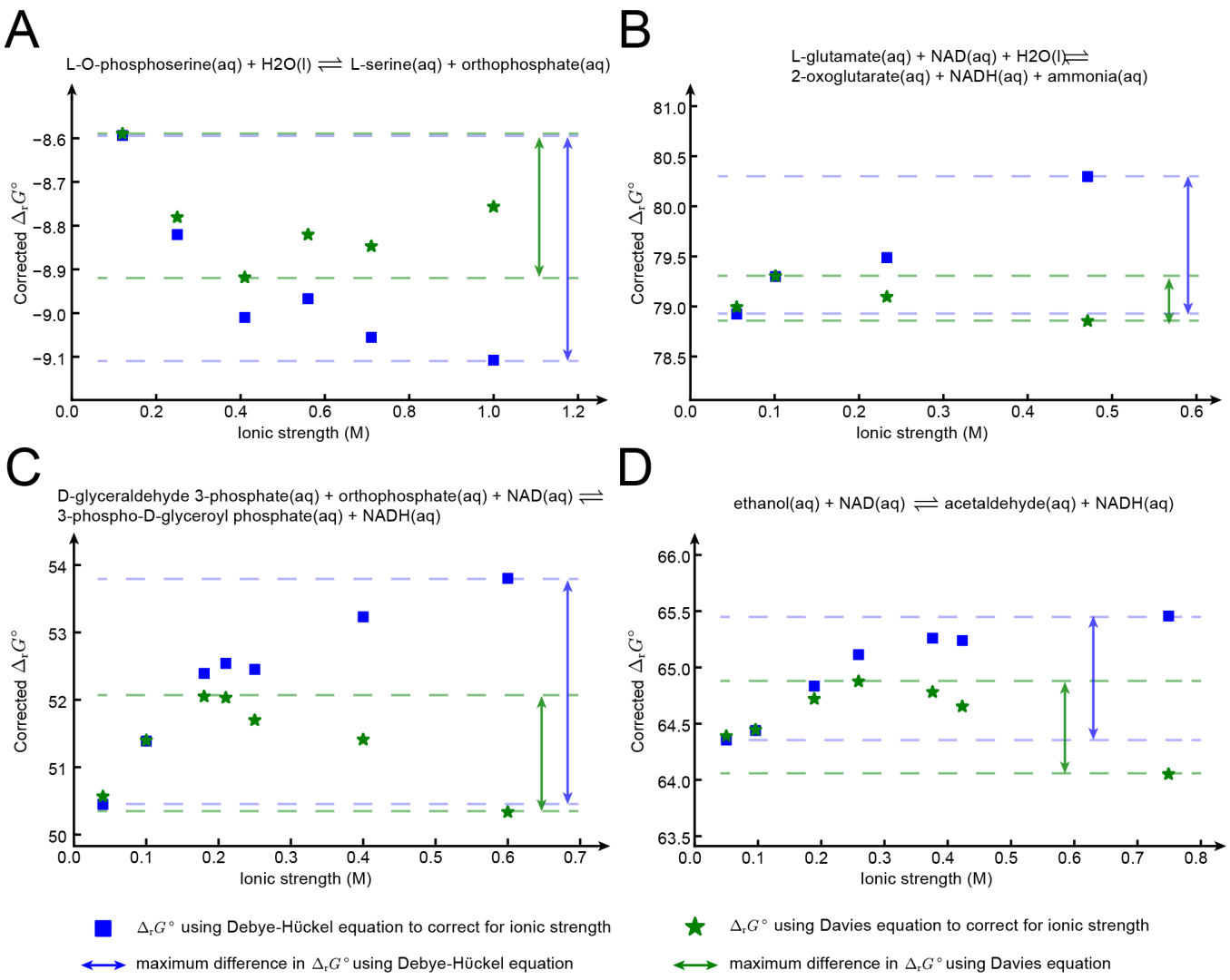

Fig. S5. (A-D) Case studies on corrected standard $\Delta_r G^\circ$ values of reactions calculated from equilibrium constants (K') measured at different ionic strength. We calculated $\Delta_r G^\circ$ at standard state using the extended Debye-Hückel equation (blue squares) and the Davies equation (green stars) to correct for varying ionic strength. We also showed the maximum differences in corrected standard $\Delta_r G^\circ$ values using different equations to correct for ionic strength. Ideally, the difference between standard $\Delta_r G^\circ$ values after correction is 0. We found that generally applying the Davies equation to correct for ionic strength results in smaller variations in $\Delta_r G^\circ$ values compared to using the extended Debye-Hückel equation.

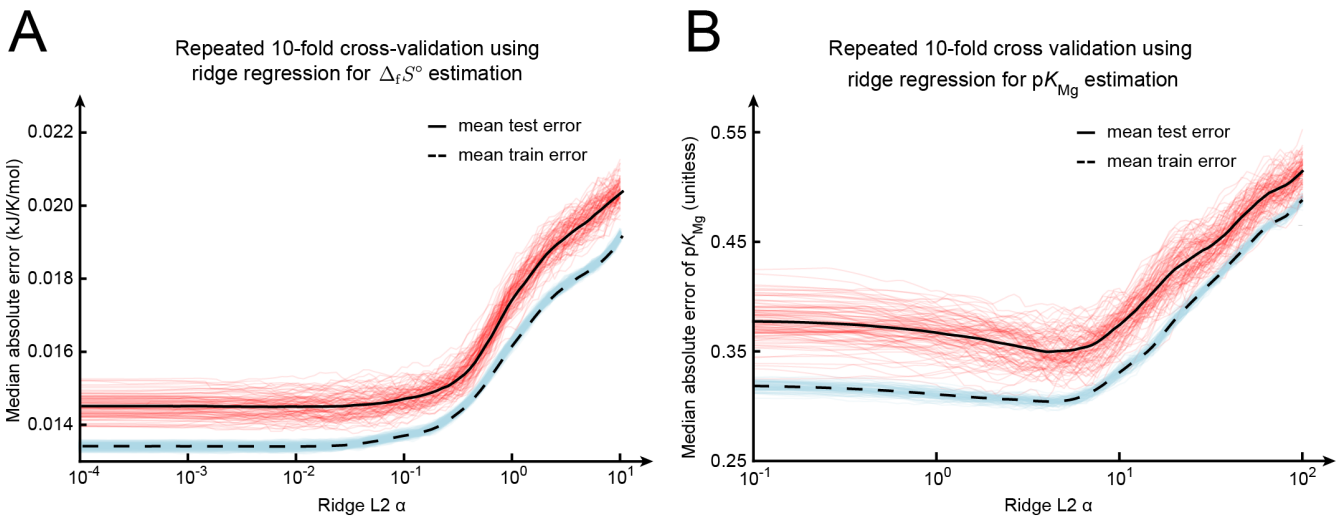

Fig. S6. (A-B) Repeated 10-fold cross-validation using ridge regression for estimation of $\Delta_f S^\circ$ and $\mathrm{p}K_{\mathrm{Mg}}$. We selected the variables with the largest absolute coefficients from the final $\Delta_f S^\circ$ and $\mathrm{p}K_{\mathrm{Mg}}$ lasso regression models (Figure 2D and 4A). We used those variables as features for the ridge regression model and performed repeated 10-fold cross-validation (100 times) on different L2 α values. We found that the resulting lowest errors are similar to those in the final lasso regression models. For $\Delta_f S^\circ$ lasso regression model, we selected variables with nonzero coefficients greater than 0.01, thus 55 out of 121 variables. For $\mathrm{p}K_{\mathrm{Mg}}$ lasso regression model, we selected variables with nonzero coefficients greater than 0.1, thus 18 out of 35 variables.

Supporting References

1. Helgeson HC, Kirkham DH (1976) Theoretical prediction of thermodynamic properties of aqueous electrolytes at high pressures and temperatures. III. equation of state for aqueous species at infinite dilution. *Am. J. Sci.* 276(2).
2. Shock EL, Helgeson HC (1988) Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: Correlation algorithms for ionic species and equation of state predictions to 5 kb and 1000°c. *Geochim. Cosmochim. Acta* 52(8):2009–2036.
3. Johnson JW, Oelkers EH, Helgeson HC (1992) SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°c. *Comput. Geosci.* 18(7):899–947.
4. Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics* 20(16):2874–2877.
5. Noor E, et al. (2012) An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics* 28(15):2037–2044.
6. Alberty RA (2003) *Thermodynamics of biochemical reactions*. (Massachusetts Institute of Technology, Cambridge, MA).
7. Davies CW (1938) 397. 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.* pp. 2093–2098.
8. Noor E, Haraldsdóttir HS, Milo R, Fleming RMT (2013) Consistent estimation of gibbs energy using component contributions. *PLoS Comput. Biol.* 9(7):e1003098.
9. Levenberg K (1944) A METHOD FOR THE SOLUTION OF CERTAIN NON-LINEAR PROBLEMS IN LEAST SQUARES. *Quart. Appl. Math.* 2(2):164–168.
10. Marquardt D (1963) An algorithm for Least-Squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11(2):431–441.
11. Manchester J, Walkup G, Rivin O, You Z (2010) Evaluation of pka estimation methods on 211 druglike compounds. *J. Chem. Inf. Model.* 50(4):565–571.
12. Settimo L, Bellman K, Knegtel RMA (2014) Comparison of the accuracy of experimental and predicted pka values of basic and acidic compounds. *Pharm. Res.* 31(4):1082–1095.
13. Newville M, et al. (2016) Lmfit: Non-Linear Least-Square minimization and Curve-Fitting for python. *Astrophysics Source Code Library*.