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.

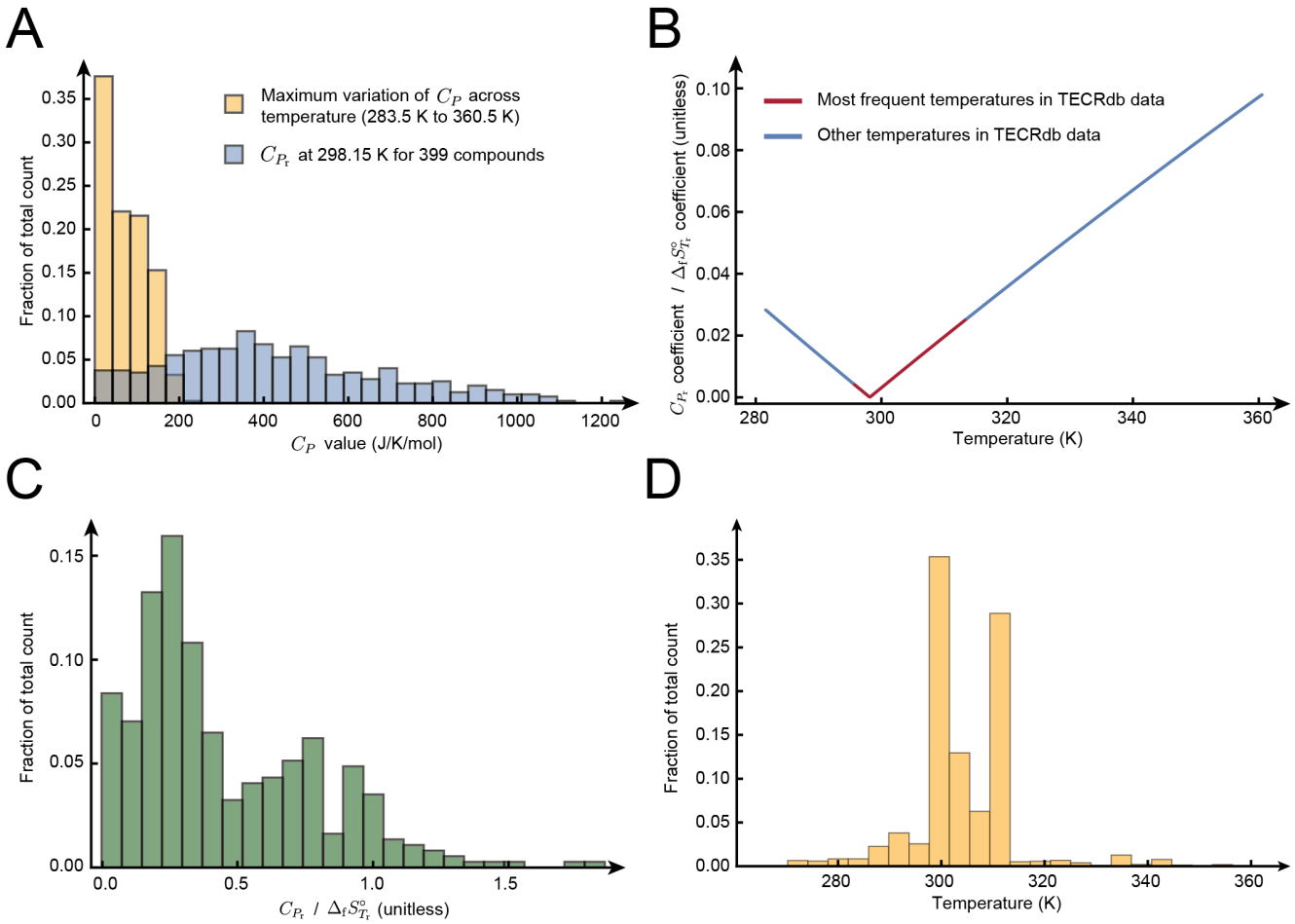

Fig. S1. (A) Order of magnitude comparison between maximum variation of C_P across temperature vs. C_{P_r} at 298.15 K for 399 compounds (3). (B) The ratio of C_{P_r} coefficient to $\Delta_f S^\circ_{T_r}$ coefficient as a function of temperature. The temperature range is based on the distribution of 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.

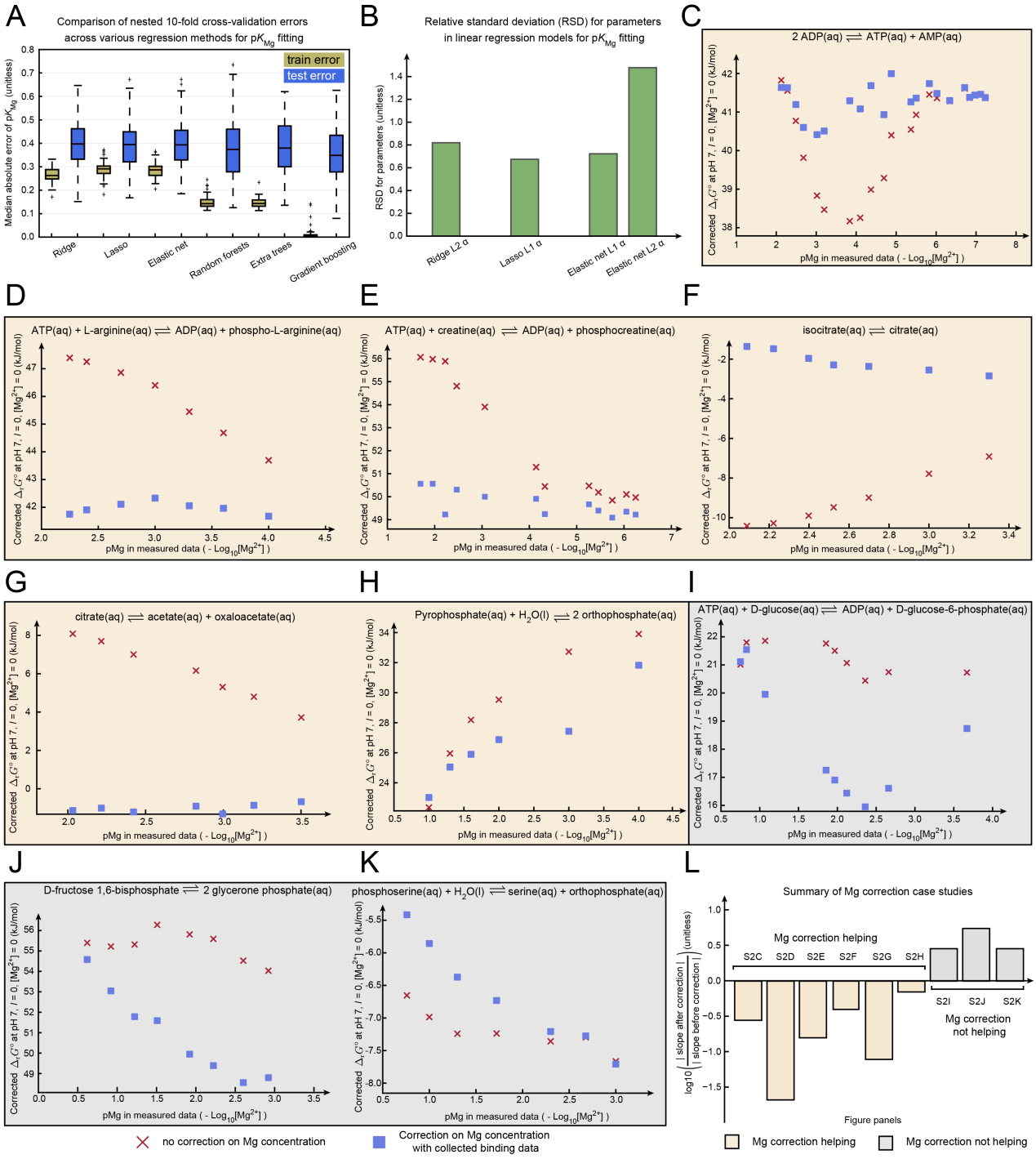

Fig. S2. (A) Training and testing errors of nested 10-fold cross validation on magnesium (Mg) binding data using the ridge regression, lasso regression, elastic net regularization, random forests, extra trees and gradient boosting. We repeated cross-validation 5 times by splitting all Mg binding data 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.

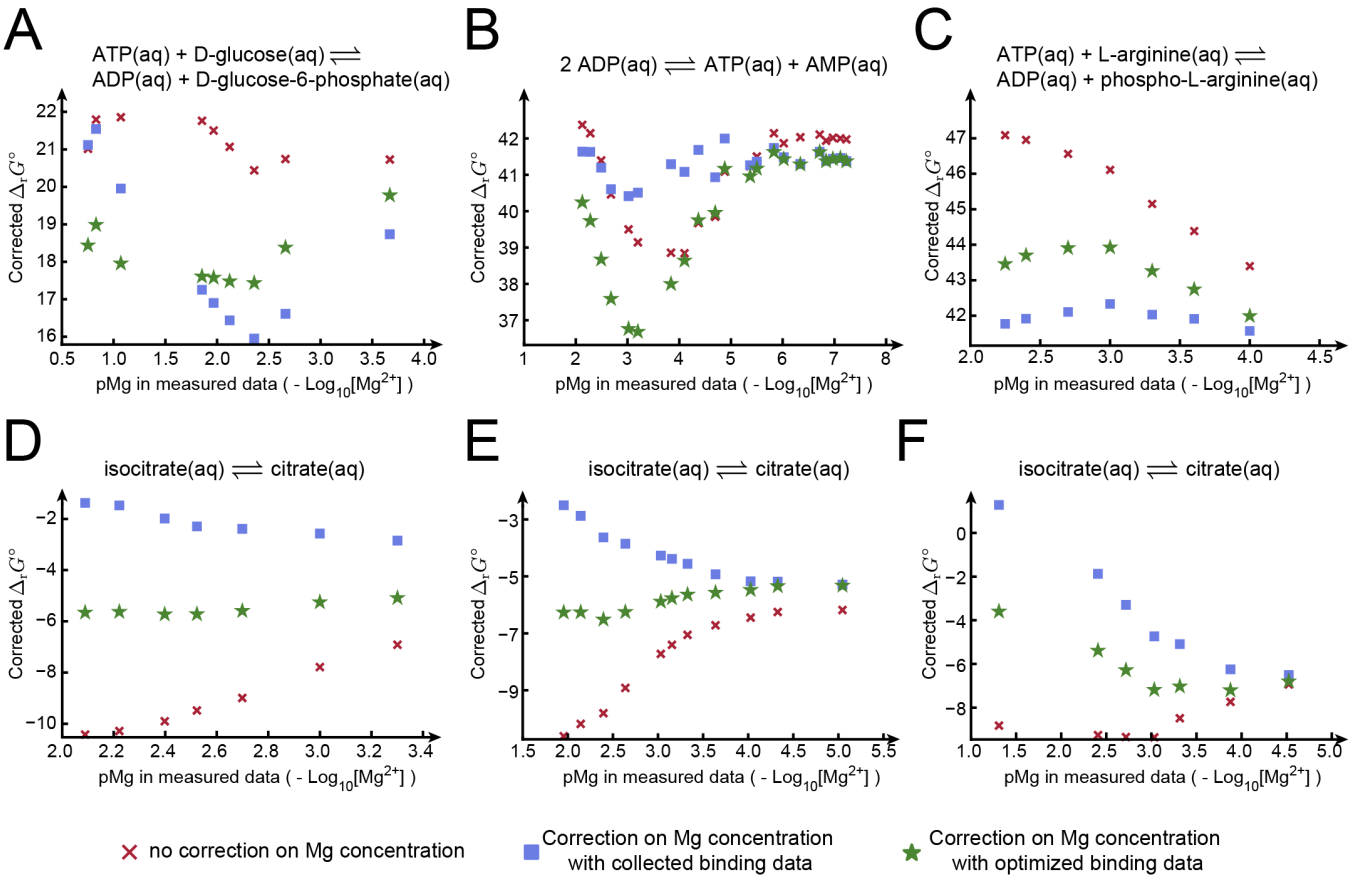

Fig. S3. (A) Corrected $\Delta_r G^\circ$ values of hexokinase reaction calculated from equilibrium constants (K') measured at different Mg concentrations. We applied different corrections to transform $\Delta_r G'^\circ$ (calculated from K') to $\Delta_r G^\circ$ values: no correction on Mg concentrations (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.

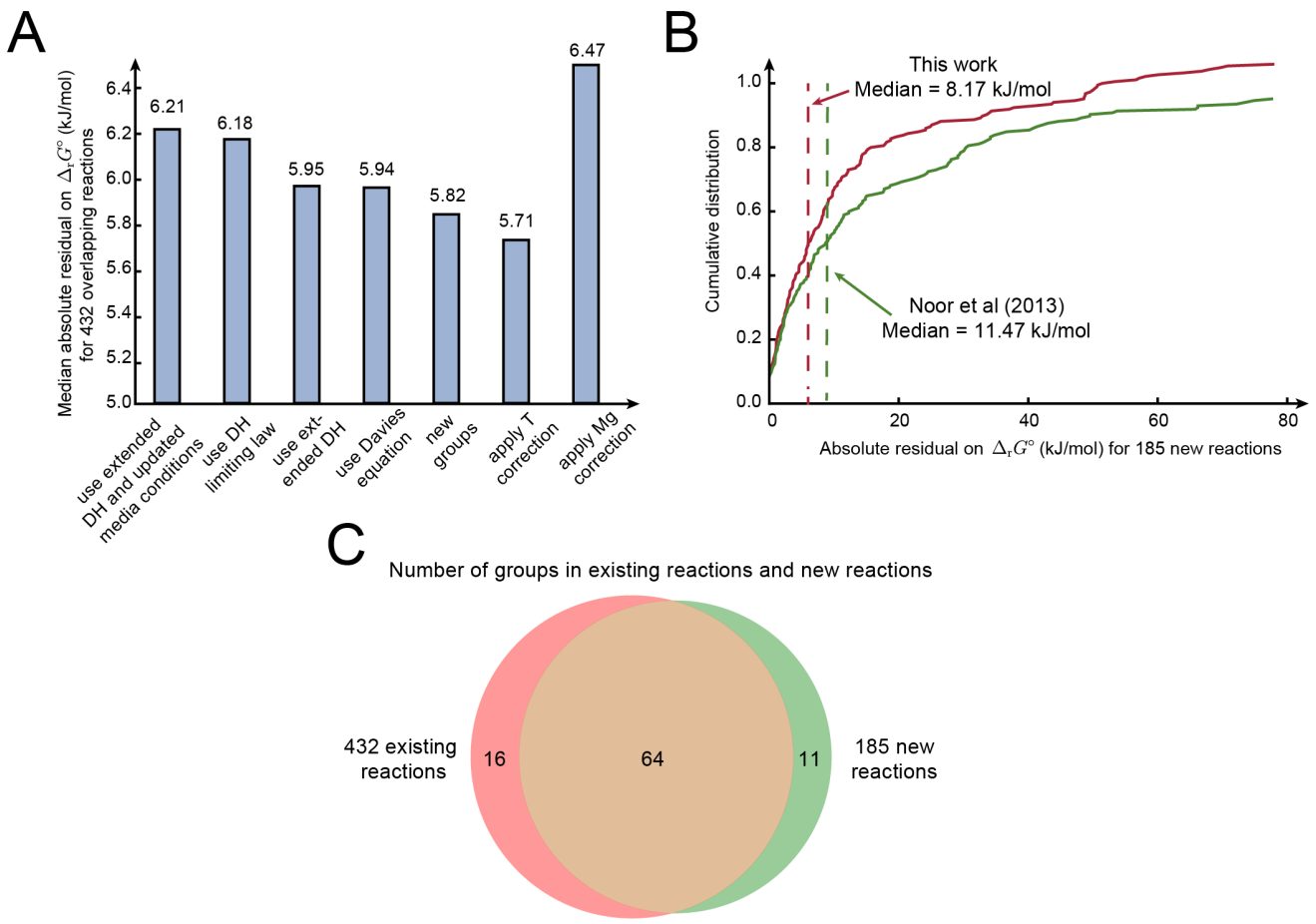

Fig. S4. (A) Comparison of absolute residuals on estimating $\Delta_r G^\circ$ for 432 reactions with various modifications on data and methods. We applied modifications on data or method one at a time sequentially and evaluated the change of median absolute residual from 10-fold cross-validation repeated 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.

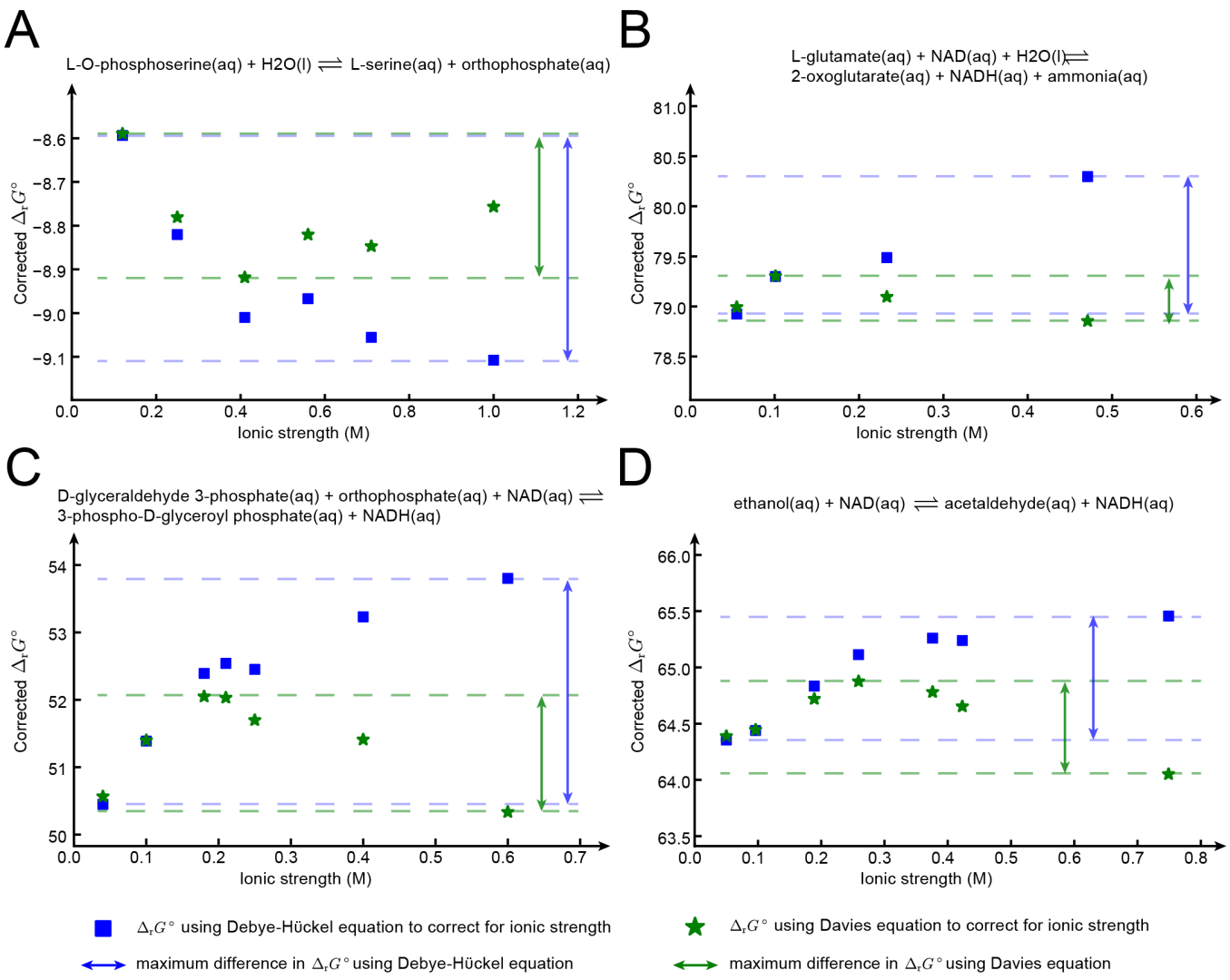

Fig. S5. (A-D) Case studies on corrected standard $\Delta_r G^\circ$ values of reactions calculated from equilibrium constants (K') measured at different ionic strength. We calculated $\Delta_r G^\circ$ at standard state using the extended Debye-Hückel equation (blue squares) and the Davies equation (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.

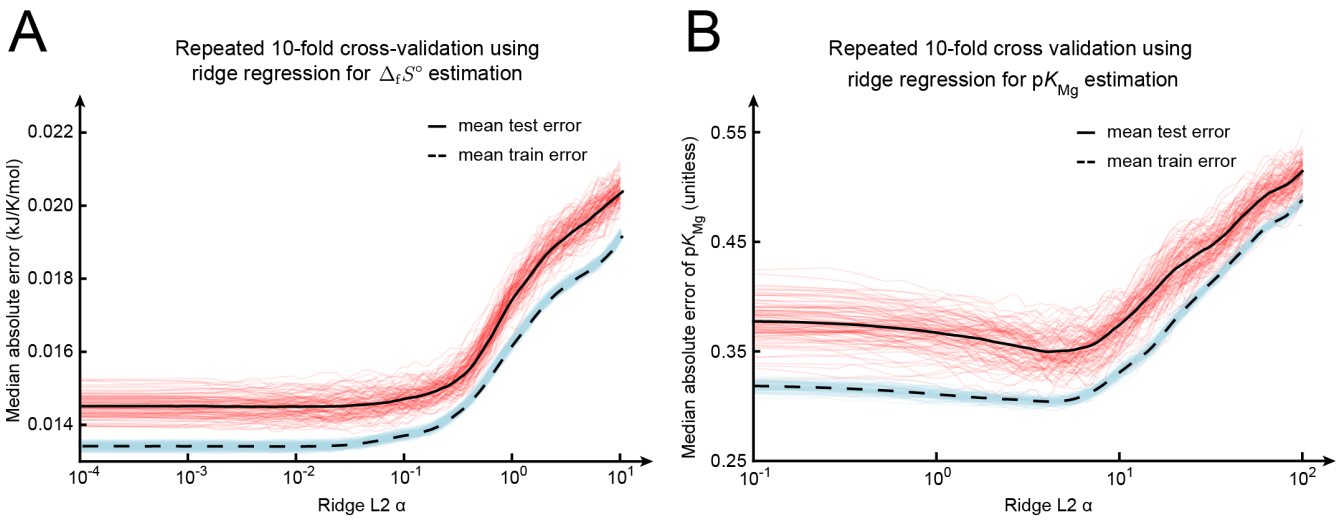

Fig. S6. (A-B) Repeated 10-fold cross-validation using ridge regression for estimation of $\Delta_f S^\circ$ and $\mathrm{p}K_{\mathrm{Mg}}$. We selected the variables with the largest absolute coefficients from the final $\Delta_f S^\circ$ and $\mathrm{p}K_{\mathrm{Mg}}$ lasso regression models (Figure 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*.