OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Consistent Estimation of Gibbs Energy Using Component Contributions

Elad Noor[1☯], Hulda S. Haraldsdóttir[2☯], Ron Milo[1]*, Ronan M. T. Fleming[2,3]*

1 Plant Sciences Department, Weizmann Institute of Science, Rehovot, Israel, **2** Center for Systems Biology, University of Iceland, Reykjavik, Iceland, **3** Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

Standard Gibbs energies of reactions are increasingly being used in metabolic modeling for applying thermodynamic constraints on reaction rates, metabolite concentrations and kinetic parameters. The increasing scope and diversity of metabolic models has led scientists to look for genome-scale solutions that can estimate the standard Gibbs energy of all the reactions in metabolism. Group contribution methods greatly increase coverage, albeit at the price of decreased precision. We present here a way to combine the estimations of group contribution with the more accurate reactant contributions by decomposing each reaction into two parts and applying one of the methods on each of them. This method gives priority to the reactant contributions over group contributions while guaranteeing that all estimations will be consistent, i.e. will not violate the first law of thermodynamics. We show that there is a significant increase in the accuracy of our estimations compared to standard group contribution. Specifically, our cross-validation results show an 80% reduction in the median absolute residual for reactions that can be derived by reactant contributions only. We provide the full framework and source code for deriving estimates of standard reaction Gibbs energy, as well as confidence intervals, and believe this will facilitate the wide use of thermodynamic data for a better understanding of metabolism.

Citation: Noor E, Haraldsdóttir HS, Milo R, Fleming RMT (2013) Consistent Estimation of Gibbs Energy Using Component Contributions. PLoS Comput Biol 9(7): e1003098. doi:10.1371/journal.pcbi.1003098

Editor: Daniel A. Beard, Medical College of Wisconsin, United States of America

Received December 14, 2012; **Accepted** April 30, 2013; **Published** July 11, 2013

Funding: EN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship (http://www.azrielifoundation.org/). RM is supported by the European Research Council, (http://erc.europa.eu/, [260392 - SYMPAC]) and is the incumbent of the Anna and Maurice Boukstein Career Development Chair in Perpetuity. RMTF and HSH were supported by the U.S. Department of Energy (Office of Advanced Scientific Computing Research, http://science.energy.gov/ascr/, and Office of Biological and Environmental Research, http://science.energy.gov/ber/) as part of the Scientific Discovery Through Advanced Computing program, grant DE-FG02-09ER25917 and the Icelandic Research Fund, grant No. 100406022. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ron.milo@weizmann.ac.il (RM); ronan.mt.fleming@gmail.com (RMTF)

☯ These authors contributed equally to this work.

This is a *PLOS Computational Biology* Methods Article.

Introduction

A living system, like any other physical system, obeys the laws of thermodynamics. In the context of metabolism, the laws of thermodynamics have been successfully applied in several modeling schemes to improve accuracy in predictions and eliminate infeasible functional states. For instance, several methodologies that reflect the constraints imposed by the second law of thermodynamics have been developed [1–3] and were shown to remove thermodynamically infeasible loops and improve overall predictions. Alternatively, thermodynamic data have been integrated directly into genome-wide models and analysis methods [4–10]. Unfortunately, this integration has been hindered by the fact that thermodynamic parameters for most reactions are effectively missing (sometimes due to scattered accessibility or non-standard annotations).

The nearly ubiquitous method for experimentally obtaining thermodynamic parameters for biochemical reactions, specifically their standard transformed Gibbs energies $\Delta_r G'^{\circ}$, is directly measuring the apparent equilibrium constant K' and then applying the formula $\Delta_r G'^{\circ} = -RT\ln(K')$, where R is the gas constant and T is the temperature. Typically, the substrates of the reaction are added to a buffered medium together with an enzyme that specifically catalyzes the reaction. After the concentrations reach a steady-state, the reaction quotient Q is calculated by dividing the product concentrations by the substrate concentrations. It is recommended to do the same measurement in the opposite direction as well (starting with what were earlier the products). If the experiment is successful, and the steady-state reached is an equilibrium state then both values for Q (measured in both directions) will be equal to K' and thus to each other. Notably, due to the nature of this method, only reactions with $\Delta_r G'^{\circ}$ close to the equilibrium value of zero can be directly measured since current technology allows measuring metabolite concentrations only within a range of a few orders of magnitude. Although this method involves purifying substantial amounts of the enzyme, it has been applied to many of the enzyme-catalyzed reactions studied throughout the last century and the results were published in hundreds or even thousands of papers. A comprehensive collection of measured K' (among other thermodynamic parameters), for more than 400 reactions, has been published by the National Institute of Standards and Technology (NIST) in the Thermodynamics of Enzyme-Catalyzed Reactions Database (TECRDB [11]). However, even this wide collection covers less than 10% of biochemical

Author Summary

The metabolism of living organisms is a complex system with a large number of parameters and interactions. Nevertheless, it is governed by a strict set of rules that make it somewhat predictable and amenable to modeling. The laws of thermodynamics play a pivotal role by determining reaction feasibility and by governing the kinetics of enzymes. Here we introduce estimations for the standard Gibbs energy of reactions, with the best combination of accuracy and coverage to date. The estimations are derived using a new method which we denote *component contribution*. This method integrates multiple sources of information into a consistent framework that obeys the laws of thermodynamics, and provides a significant improvement in accuracy compared to previous genome-wide estimations of standard Gibbs energies. We apply and test our method on reconstructions of *E. coli* and human metabolism and, in addition, do our best to facilitate the use of these estimations in future models by providing open-source software that performs the integration in a streamlined process.

reactions in standard metabolic reconstructions, such as the *E. coli* model iAF1260 [5].

In 1957 [12], K. Burton recognized that these apparent equilibrium constants can be used (together with chemically derived standard Gibbs energies for some simple compounds) to calculate equilibrium constants of reactions with no known K' values. This method is based on the notion that by knowing the $\Delta_r G'^{\circ}$ of two different reactions, one can calculate the $\Delta_r G'^{\circ}$ of the combined reaction by summing the two known standard transformed Gibbs energies – as dictated by the first law of thermodynamics. For example, although the reaction of ATP hydrolysis ($ATP+H_2O \rightleftharpoons ADP+P_i$) might be too far from equilibrium to be measured directly, one can more easily measure the K' of the reactions of glucose kinase ($ATP+glucose \rightleftharpoons ADP+glucose-6P$; $\Delta_r G'^{\circ} \approx -25$ kJ/mol) and of glucose-6P phosphatase ($glucose-6P+H_2O \rightleftharpoons glucose+P_i$; $\Delta_r G'^{\circ} \approx -12$ kJ/mol), which are both closer to equilibrium. The standard transformed Gibbs energy for the total reaction (i.e. ATP hydrolysis) would thus be $\Delta_r G'^{\circ} \approx -37$ kJ/mol.

In order to facilitate these K' calculations, Burton published a table of about 100 inferred standard Gibbs energies of formation ($\Delta_f G^{\circ}$) which are defined as the standard Gibbs energy $\Delta_r G^{\circ}$ of the *formation reaction*, i.e. the reaction of forming a compound out of pure elements in their standard forms (e.g. $\frac{1}{2}O_2+H_2 \rightleftharpoons H_2O$). Some of these values were taken from chemical thermodynamic tables, and the rest were derived by Burton using the arithmetic approach of combining reactions. For instance, if all species except one in an enzyme-catalyzed reaction have known $\Delta_f G^{\circ}$, and the reaction's $\Delta_r G^{\circ}$ can be obtained experimentally, then the last remaining $\Delta_f G^{\circ}$ can be calculated and added to the table. After compiling such a table, the $\Delta_r G^{\circ}$ of any reaction involving species that appear in the table can be easily calculated by summing the formation energies of all the products and subtracting those of the substrates. Throughout this paper we will refer to this method of calculating $\Delta_r G^{\circ}$ as the Reactant Contribution (RC) method, since it is based on the contribution of each reactant to $\Delta_r G^{\circ}$ (i.e. its standard Gibbs energy of formation).

In the 50 years following Burton's work, several such tables of formation Gibbs energies have been published. Some of the most noteworthy are the table by R. Thauer [13] and the larger collection by R. Alberty [14,15]. Using these values, one can determine Gibbs energies for more reactions at a wider range of physiological conditions (pH, ionic strength) than the set of reactions measured and stored in TECRDB. However, even this advanced method covers less than 10% of reactions in the *E. coli* model. This gap has prompted scientists to develop methods that can estimate the missing thermodynamic parameters for genome-wide models.

Quite coincidentally, a year after Burton published his thermodynamic tables, S. Benson and J. Buss [16] published their work on additivity rules for the estimation of molecular properties. Benson and Buss called the law of additivity of atomic properties a *zero-order* approximation, the additivity of bond properties a *first-order* approximation, and the additivity of group properties a *second-order* approximation. Groups were defined as pairs of atoms or structural elements with distances of 3–5 Å. The contribution of each group to the total was determined by linear regression. Using the second-order approximation, $\Delta_r G^{\circ}$ is calculated as the sum of the standard Gibbs energy contributions of groups that are produced in the reaction, minus the contributions of groups that are consumed. This method is commonly called the Group Contribution (GC) method. Burton's method of calculating the Gibbs energy of formation for compounds (which we denote RC) can be thought of as a ∞-*order* approximation, where the entire molecule is taken as the basic additivity unit for estimating the $\Delta_r G^{\circ}$ (of course, this is not actually an approximation).

Group contribution methods were relatively successful in estimating the thermodynamic parameters of ideal gases [16–19], and later extended to liquid and solid phases [20]. Only in 1988 [21] was it brought to the world of biochemical reactions in aqueous solutions and has since become the most widely used technique for estimating the Gibbs energy of reactions [22–24]. GC methods can cover the majority of relevant biochemical reactions ($\approx$90% and $\approx$70% of the reactions in *E. coli* and human cell metabolic models respectively) [5,10,24]. The downside of GC lies in its accuracy, since it relies on a simplifying assumption that the contributions of groups are additive. Evidently, the average estimation error attributed to GC is about 9–10 kJ/mol [23]. In a recent study, we showed that an improvement of $\approx$14% can be achieved by considering different pseudoisomers that exist simultaneously for many of the compounds [24] (see Section S1 in Text S1 for details). Even with this improvement, error in GC estimates is still significantly higher than in RC estimates (Figure 1).

In this paper, we aim to unify GC and RC into a more general framework we call the Component Contribution method. We demonstrate that component contribution combines the accuracy of RC with the coverage of GC in a fully consistent manner. A plot comparing the component contribution method to other known methods is given in Figure 1.

Unifying reactant and group contribution methods

The extensive use of formation Gibbs energies for calculating $\Delta_r G^{\circ}$ might create the impression that combining these two frameworks (RC and GC) is a trivial task. Traditionally, the formation energy of all pure elements in their standard forms is set to zero by definition. All other compounds' formation energies are calculated in relation to these reference points. This is a sound definition which creates a consistent framework for deriving the $\Delta_r G^{\circ}$ of any reaction which is chemically balanced. However, the difficulty of calculating the formation energy for some complex but useful co-factors has been side-stepped by creating a somewhat looser definition of formation Gibbs energy, where several non-elemental compounds are defined as reference points as well (with a standard formation energy of zero). For some rare reactions, this

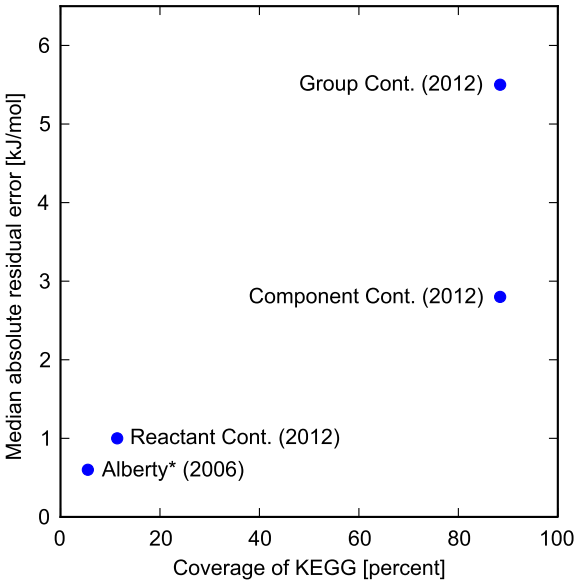

Figure 1. The development of Gibbs energy estimation frameworks. The coverage is calculated as the percent of the relevant reactions in the KEGG database (i.e. reactions that have full chemical descriptions and are chemically balanced). The median residual (in absolute values) is calculated using leave-one-out cross-validation over the set of reactions that are within the scope of each method. Note that the reason component contribution has a higher median absolute residual than RC is only due to its higher coverage of reactions (for reactions covered by RC, the component contribution method gives the exact same predictions). *The residual value for Alberty's method is not based on cross-validation since it is a result of manual curation of multiple data sources – a process that we cannot readily repeat.
doi:10.1371/journal.pcbi.1003098.g001

definition can create a conflict that will result in a very large mistake in the $\Delta_r G^\circ$.

For instance, Alberty's formation energy table [15] lists 16 compounds as having $\Delta_f G^\circ = 0$. Among these, only 5 are elements (H_2, I_2, N_2, O_2, and S) and 11 are co-factors (CoA^{-1}, $NAD(ox)^{-1}$, $FAD(ox)^{-2}$, $FMN(ox)^{-2}$ and seven other redox carriers). In most reactions which use these co-factors as substrates, the "zeros" will cancel out since one of the products will match it with a formation energy which is defined according to the same reference point (e.g. $FAD(ox)^{-2}$ will be matched with $FAD(red)^{-2}$ whose formation energy is -38.88 kJ/mol in Albery's table). Nevertheless, there are a handful of reactions where this matching doesn't occur. In the reaction $FAD(ox)^{-2} + H_2O \rightleftharpoons FMN(ox)^{-2} + AMP^{-1} + H^+$ (catalyzed by *FAD nucleotidohydrolase*, EC 3.6.1.18), there is a violation of conservation laws for FAD and FMN (both have $\Delta_f G^\circ = 0$ in Alberty's table). Therefore, using the table naïvely for this reaction would yield a wrong value, namely $\Delta_r G^\circ = -880$ kJ/mol. Combining formation energies derived using GC with ones from RC greatly increases the number of reactions where different reference-points are mixed together, and mistakes such as the one described above become much more common.

One way to deal with the problem of reference-point conflicts, is to use either RC or GC exclusively for every reaction $\Delta_r G^\circ$ being estimated. Specifically, RC will by applied to all reactions that can be completely covered by it. Only if one or more reactants are missing from the formation energy table, we would use the less precise GC method for the entire reaction. Unfortunately, combining the frameworks in such a way can easily lead to violations of the first law of thermodynamics. This stems from the fact that inconsistent use of formation energies across several reactions, coming from inaccuracies in the estimation method that do not cancel out, can create situations where futile cycles will have a non-zero change in Gibbs energy. An example for such a futile cycle is given in Figure 2. Applying this method for large-scale metabolic reconstructions will most likely lead to non-physical solutions.

Reference-point conflicts and first-law violations can both be avoided, by adjusting baseline formation energies of compounds with non-elemental reference points to match group contribution estimates. This approach was taken in [8] and [10]. The formation energies of $FAD(ox)^{-2}$ and all other reference points in Alberty's table were set equal to their group contribution estimates. All formation energies that were determined relative to each reference point were then adjusted according to Alberty's table to maintain the same relative formation energies. The main disadvantage of this approach is that the set of reference points is fixed and limited to a few common cofactors. The coverage of reactant contributions could be increased by also defining less common metabolites as reference points, but listing them all in a static table would be impractical and inefficient.

The component contribution method, which is described in detail in the following sections of this paper, manages to combine the estimates of RC and GC while avoiding any reference-point conflicts or first-law violations. In the component contribution framework, the maximal set of reference points given a set of measured reactions is automatically determined. We maintain the notion of prioritizing RC over GC, but rather than applying only one method exclusively per reaction, we split every reaction into two independent reactions. One of these sub-reactions can be evaluated using RC, while the other cannot – and thus its $\Delta_r G^\circ$ is calculated using GC. We use linear orthogonal projections in order to split each of the reactions, ensuring that all estimated

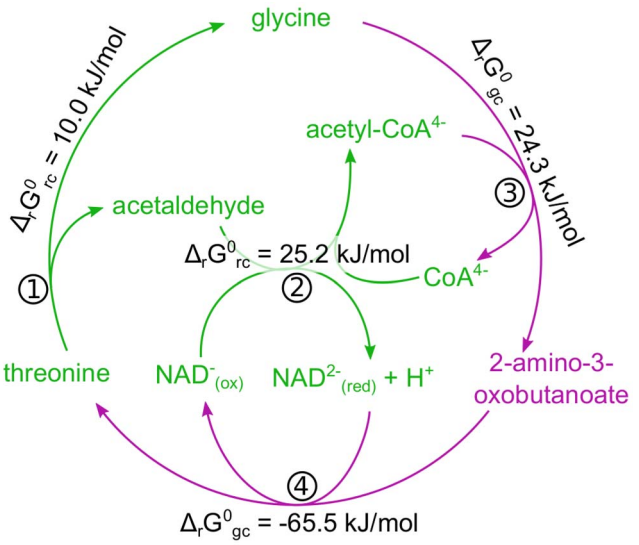

Figure 2. An example of a futile cycle where Gibbs energies are derived using RC and GC. The combined stoichiometry of (1) threonine aldolase, (2) acetaldehyde dehydrogenase (acetylating), (3) glycine C-acetyltransferase, and (4) threonine:NAD oxidoreductase creates a futile cycle where all the inputs and outputs are balanced. Using RC we are able to derive the $\Delta_r G^\circ$ of reactions 1 and 2 (green), but since 2-amino-3-oxobutanoate does not appear in formation energy tables – we must use GC for reactions 3 and 4 (magenta). The sum of all $\Delta_r G^\circ$ in this case is -6.0 kJ/mol which is a violation of the first law of thermodynamics.
doi:10.1371/journal.pcbi.1003098.g002

$\Delta_r G^\circ$ values are self-consistent. The choice of orthogonal projections is somewhat arbitrary, and is based on the assumption that it is beneficial to minimize the euclidean distance to the projected point that can be estimated using RC. This framework also enables us to calculate confidence intervals for standard reaction Gibbs energies in a mathematically sound way.

Results

The component contribution method

The component contribution method integrates reactant contributions and group contributions in a single, unified framework using a layered linear regression technique. This technique enables maximum usage of the more accurate reactant contributions, and fills in missing information using group contributions in a fully consistent manner. The inputs to the component contribution method are the stoichiometric matrix of measured reactions, denoted $S \in \mathbb{R}^{m \times n}$, and a list of measurements of their standard Gibbs energies $\Delta_r G^\circ_{obs} \in \mathbb{R}^n$ (see Table S2 in Text S1 for a list of mathematical symbols). In our case, all data is taken from TECRDB [11] and tables of compound formation energies [13,15]. As a pre-processing step which is used to linearize the problem, we apply an inverse Legendre transform to the observed equilibrium constants in TECRDB and the formation energies, if necessary (same as in [24], see Section S1 in Text S1). To provide context for the mathematical formulation of the component contribution method, we precede it with general formulations of the reactant and group contribution methods, and discuss the limitations of each. The reactant and group contribution methods are both based on linear regression. The difference between the two methods lies in the regression models used in each.

Reactant contribution method. The regression model used in the reactant contribution method is based on the first law of thermodynamics (conservation of energy). The first law dictates that the overall standard Gibbs energy of a reaction that takes place in more than one step, is the sum of the standard Gibbs energies of all the intermediate steps at the same conditions [25]. Consequently, if $\Delta_f G^\circ \in \mathbb{R}^m$ is the vector of standard Gibbs energies of formation for compounds in S, then the standard Gibbs energies of reactions in S are given by the equation

$$\Delta_r G^\circ = S^{\mathrm{T}} \cdot \Delta_f G^\circ. \tag{1}$$

From Eq. 1 it is apparent that $\Delta_r G^\circ$ is in the range of S^{T}, which we denote by $\mathcal{R}(S^{\mathrm{T}})$. In practice, this may not be readily true for $\Delta_r G^\circ_{obs}$ from TECRDB, since its values are empirically derived and thus subject to measurement noise. Also, the exact ionic strength is not known for most measurements and the extended Debye-Hückel theory of electrolyte solutions [26] (which the inverse Legendre transform is based on [27]) is itself an approximation. The linear regression model used in the reactant contribution method for $\Delta_r G^\circ_{obs}$ therefore takes the form

$$\Delta_r G^\circ_{obs} = S^{\mathrm{T}} \cdot \Delta_f G^\circ + \varepsilon_{rc}, \tag{2}$$

where ε_{rc} encompasses the error from the aforementioned sources.

Least-squares linear regression on the system in Eq. 2 gives the reactant contribution estimate of the standard Gibbs energies of formation for compounds in S

$$\Delta_f G^\circ_{rc} = \left(S^{\mathrm{T}}\right)^+ \cdot \Delta_r G^\circ_{obs}. \tag{3}$$

The Moore–Penrose pseudoinverse $\left(S^{\mathrm{T}}\right)^+$ is used because S^{T} is typically column rank deficient. Reactant contribution fitted standard Gibbs energies for reactions in S are,

$$\Delta_r G^\circ_{rc} = S^{\mathrm{T}} \cdot \Delta_f G^\circ_{rc} = S^{\mathrm{T}} \left(S^{\mathrm{T}}\right)^+ \cdot \Delta_r G^\circ_{obs} \tag{4}$$

i.e., they are the orthogonal projection of $\Delta_r G^\circ_{obs}$ onto $\mathcal{R}(S^{\mathrm{T}})$. $\Delta_r G^\circ_{rc}$ is therefore the closest point to $\Delta_r G^\circ_{obs}$ that is consistent with the first law of thermodynamics. The residual of the fit

$$e_{rc} = \Delta_r G^\circ_{obs} - \Delta_r G^\circ_{rc}, \tag{5}$$

gives an estimate of the error term ε_{rc} in Eq. 2. We stress that there is a conceptual distinction between the residual (e_{rc}) and the statistical error (ε_{rc}). e_{rc} is dependent on the specific sample of equilibrium constants we use in the training set, while ε_{rc} is a random variable that can only be approximated. We use the term *error* for the deviation of an observed or estimated Gibbs energy (known values), from the (unknown) true Gibbs energy. The term *residual* is used for the deviation of an observed Gibbs energy from an estimate. We note that e_{rc} is in the null space of S, denoted $\mathcal{N}(S)$, since the null space is the orthogonal complement of $\mathcal{R}(S^{\mathrm{T}})$, according to the fundamental theorem of linear algebra.

The standard Gibbs energy $\Delta_r G^\circ_x \in \mathbb{R}$ of an unmeasured reaction with stoichiometric vector $x \in \mathbb{R}^m$ can be estimated with the reactant contribution method as

$$\Delta_r G^\circ_{rc,x} = x^{\mathrm{T}} \cdot \Delta_f G^\circ_{rc} = x^{\mathrm{T}} \left(S^{\mathrm{T}}\right)^+ \cdot \Delta_r G^\circ_{obs}. \tag{6}$$

This result is consistent with the first law of thermodynamics in the following sense. In general, the first law implies that the standard Gibbs energy of a linear combination of reactions, is the same combination applied to the respective standard reaction Gibbs energies. Mathematically, if $x = Sw$ then $\Delta_r G^\circ_{rc,x} = w^{\mathrm{T}} \cdot \Delta_r G^\circ_{obs}$. The former equation gives $w = S^+ x$ which is precisely the result in Eq. 6. Having compliance with the first law as the only assumption explains the high accuracy of the reactant contribution method.

The reactant contribution method can be used to evaluate standard Gibbs energies for x in the range of S, i.e. reactions that are linear combinations of reactions in S (and thus have at least one solution for $x = Sw$). Any component of x that is not in $\mathcal{R}(S)$ cannot be evaluated. Since S is rank deficient, its range represents only a fraction of the entire space of reactions and thus most reactions are under-determined by this method. For instance, the CMP phosphohydrolase reaction ($CMP + H_2O \rightleftharpoons cytidine + P_i$) exists in the *E. coli* model but is not listed as a measured reaction in TECRDB. Although CMP and cytidine both appear in other measured reactions ($CMP + ATP \rightleftharpoons CDP + ADP$ and $cytidine + H_2O \rightleftharpoons uridine + NH_4$), it is impossible to use a combination of reactions in TECRDB to find the $\Delta_r G^\circ$ of the CMP phosphohydrolase reaction.

Group contribution method. Increased reaction coverage can be achieved using the group contribution method, where each compound in S is decomposed into a predefined set of structural subgroups. Each decomposition is represented by a row of the group incidence matrix $\mathcal{G} \in \mathbb{R}^{m \times g}$, and $\Delta_r G^\circ$ is assumed to be a linear combination of the standard Gibbs energy contributions $\Delta_g G^\circ$ of the groups in $\mathcal{G}$. The linear regression model for the group contribution method is

$$\Delta_r G^\circ_{obs} = S^{\mathrm{T}} \mathcal{G} \cdot \Delta_g G^\circ + \varepsilon_{gc}. \tag{7}$$

$S^{\mathrm{T}}\mathcal{G}\in\mathbb{R}^{n\times g}$ describes the group decompositions of reactions in S i.e., the stoichiometry of groups that are consumed or produced in the reactions. An estimate of $\Delta_g G^\circ$ is obtained using linear regression on the system in Eq. 7 i.e.,

$$\Delta_g G^\circ_{gc} = \left(S^{\mathrm{T}}\mathcal{G}\right)^{+}\cdot\Delta_r G^\circ_{obs}, \tag{8}$$

and like in reactant contribution we define $\Delta_r G^\circ_{gc} = S^{\mathrm{T}}\mathcal{G}\left(S^{\mathrm{T}}\mathcal{G}\right)^{+}\cdot\Delta_r G^\circ_{obs}$ and $e_{gc} = \Delta_r G^\circ_{obs} - \Delta_r G^\circ_{gc}$. The group contribution estimate of the standard reaction Gibbs energy for x can then be derived as

$$\Delta_r G^\circ_{gc,x} = x^{\mathrm{T}}\mathcal{G}\cdot\Delta_g G^\circ_{gc} = x^{\mathrm{T}}\mathcal{G}\left(S^{\mathrm{T}}\mathcal{G}\right)^{+}\cdot\Delta_r G^\circ_{obs}. \tag{9}$$

The reaction coverage of the group contribution method is much greater than that of the reactant contribution method in Eq. 6, because the column rank deficiency of $S^{\mathrm{T}}\mathcal{G}$ is much smaller than that of S^{T}. However, this greater coverage comes at a price, since the assumption of group additivity underlying the model in Eq. 7 is not always accurate. We estimated the root-mean-square error resulting from this assumption as 6.8 kJ/mol for all reactions in S (see Section S4 in Text S1 for details). The reaction coverage of group contribution methods is still limited to $\mathcal{G}^{\mathrm{T}}x\in\mathcal{R}\left(\mathcal{G}^{\mathrm{T}}S\right)$, i.e. reactions with group decompositions that are linear combinations of the group decompositions of measured reactions.

Mathematical formulation of the component contribution method. The reactant contribution method covers any vector in the range of S. The component contribution method takes advantage of the fact that any reaction vector x in $\mathbb{R}^m$ can be decomposed into a component x_R in the range of S, and an orthogonal component x_N in the null space of S^{T}. Let $P_{\mathcal{R}(S)}$, $P_{\mathcal{N}(S^{\mathrm{T}})}\in\mathbb{R}^{m\times m}$ be the orthogonal projection matrices onto the range of S and the null space of S^{T}, respectively. Then $x_R = P_{\mathcal{R}(S)}\cdot x$ and $x_N = P_{\mathcal{N}(S^{\mathrm{T}})}\cdot x$ (so $x = x_R + x_N$ and $x_R \perp x_N$). The component contribution method applies the more reliable reactant contribution method to evaluate x_R, and only applies the less reliable group contribution method to x_N (see Figure 3). The standard reaction Gibbs energy estimate for x is obtained by summing up the contributions from the two components (see Equations 6 and 9) i.e.,

$$\begin{aligned}\Delta_r G^\circ_{cc,x} &= x_R^{\mathrm{T}}\cdot\Delta_f G^\circ_{rc} + x_N^{\mathrm{T}}\cdot\mathcal{G}\cdot\Delta_g G^\circ_{gc} = \\ &= x^{\mathrm{T}}\left(P_{\mathcal{R}(S)}\left(S^{\mathrm{T}}\right)^{+} + P_{\mathcal{N}\left(S^{\mathrm{T}}\right)}\mathcal{G}\left(S^{\mathrm{T}}\mathcal{G}\right)^{+}\right)\cdot\Delta_r G^\circ_{obs}\end{aligned} \tag{10}$$

(see the full derivation in Section S3 in Text S1). We note that using the two orthogonal projections is only one option for separating x to two components and applying RC and GC on each one respectively. Other pairs of linear projections could be applied as long as they fulfill the requirement that they sum up to the identity matrix, and that the range of the first one is $\subseteq\mathcal{R}(S)$. Here we chose $P_{\mathcal{R}(S)}$ and $P_{\mathcal{N}\left(S^{\mathrm{T}}\right)}$ because they minimize the norm of the second component, and we assume there is benefit to it.

The component x_N in the null space of S^{T} can only be evaluated if $\mathcal{G}^{\mathrm{T}}x_N$ is in the range of $\mathcal{G}^{\mathrm{T}}S$, i.e. the space covered by group contributions. We thus require that $x_N = x_{NR}$ where x_{NR} is the component of x_N in $\mathcal{R}\left(\mathcal{G}^{\mathrm{T}}S\right)$. If x_N has a nonzero component $x_{NN}\equiv x_N - x_{NR}\in\mathcal{N}\left(S^{\mathrm{T}}\mathcal{G}\right)$ then the overall reaction x cannot be evaluated using component contributions. In practice we assign an infinite uncertainty to reactions where $x_{NN}\neq 0$ as discussed in section *Calculation of confidence intervals*. The two orthogonal components of $x_N = x_{NR} + x_{NN}$ are determined by orthogonal projections onto the subspaces of $\mathcal{G}^{\mathrm{T}}S$, in the same way that $x = x_R + x_N$ was decomposed by projections onto the subspaces of S. Component contribution is thus based on two layers of orthogonal decompositions; a first layer where x is decomposed into x_R and x_N, and a second layer where x_N is decomposed into x_{NR} and x_{NN} (Figure 3).

A common example where $x_{NN}\neq 0$ occurs where x_N is a reaction that includes the formation of an uncommon group. If this group does not appear (or is always conserved) in all of the reactions in the training set then the contribution of that group is unknown. Since $\mathcal{G}^{\mathrm{T}}x_N$ has a non-zero value corresponding to that group, x_N cannot be in the range of $\mathcal{G}^{\mathrm{T}}S$.

Validation results

In order to evaluate the improvement in estimations derived using component contribution compared to an implementation of group contribution [24], we ran a cross-validation analysis (see section *Leave-one-out cross-validation* for details). The results of this analysis are shown in Figure 4, where we compare the distributions of the absolute residuals (the difference between each method's estimated $\Delta_r G^\circ$ and the observed $\Delta_r G^\circ$ according to TECRDB). For each estimation, the value of $\Delta_r G^\circ$ for that reaction (or any other measurement of the same reaction) was not used for training the group contribution and component contribution methods.

Our results show a significant improvement for component contribution compared to group contribution when focusing on reactions in the range of S. The median of all residuals (absolute value) was reduced from 4.6 to 1.0 kJ/mol (p-value $< 10^{-36}$) for this set of reactions. For reactions that were not in $\mathcal{R}(S)$, there was no significant difference (p-value = 0.45) in the median absolute residual between the two methods. The error in group contribution estimates that is due to the assumption of group additivity does not depend on the extent to which group contribution is used (see Section S4.2 in Text S1). Because component contribution uses group contribution to some extent for all reactions that are not in $\mathcal{R}(S)$, the error in component contribution estimates for those reactions is not significantly lower than the error in group contribution estimates. Note that it is still very important to use component contribution for these reactions (and not GC) for the sake of having consistent estimations across whole metabolic models (see section *Unifying reactant and group contribution methods* in the Introduction).

In each iteration of the cross-validation, one reaction was excluded from the training set. To further validate the component contribution method, we used the results of each iteration to predict independent observations of the reaction that was excluded. All available observations of that reaction were then compared against the prediction intervals for its standard Gibbs energy (see section *Calculation of prediction intervals* in the Methods). Overall, we found that 73% of observations fell within their respective 68% prediction intervals, 89% fell within their 90% prediction intervals, 92% fell within their 95% prediction intervals, and 97% within their 99% intervals. Prediction intervals obtained with the component contribution method were on average 36% smaller than those obtained with group contribution. Taken together, these results show that the component contribution method yields estimates with reliable confidence intervals, as well as increased accuracy and reduced uncertainty compared to group contribution.

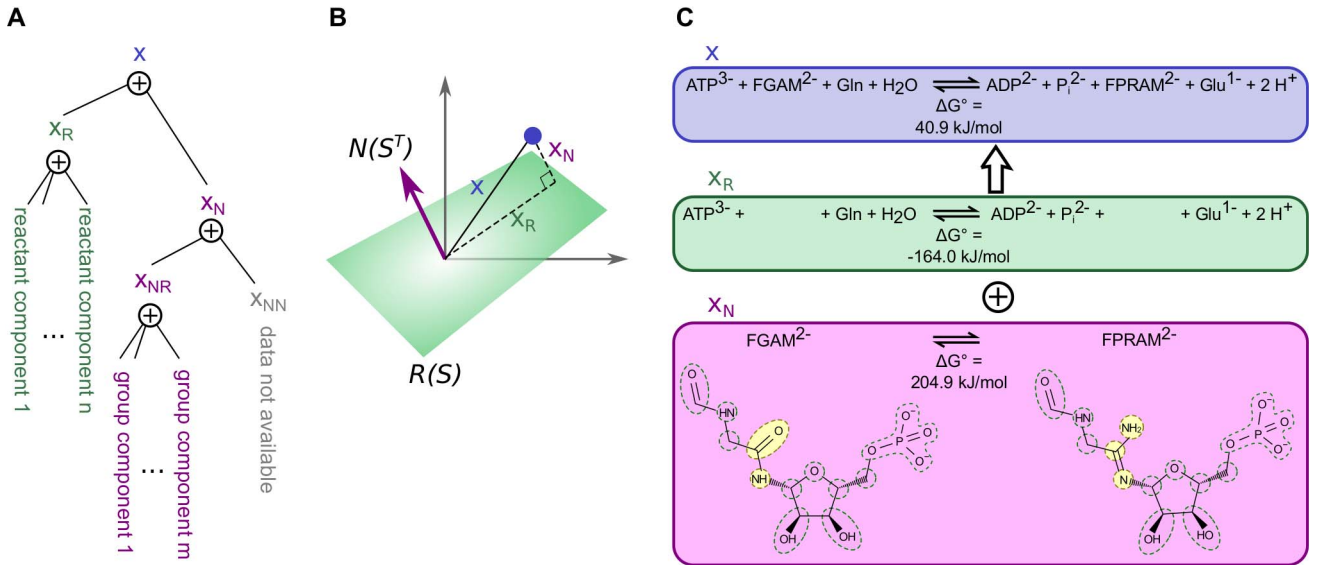

Figure 3. A diagram illustrating how the component contribution method projects the stoichiometric vector onto the different spaces. (A) The reaction vector x is decomposed into the two components x_R and x_N, where the reactant contribution and group contribution methods are used for the relevant components. Later, x_N is decomposed into x_{NR} and x_{NN}. The same projection is shown graphically in (B) where the green plane represents the range of S and the normal to that plane represents the null space of S^{T}. (C) An example for a reaction which decomposes into two non-zero components. In this case, the component x_{NN} is equal to 0, which means that the reaction is covered by the component contribution method.
doi:10.1371/journal.pcbi.1003098.g003

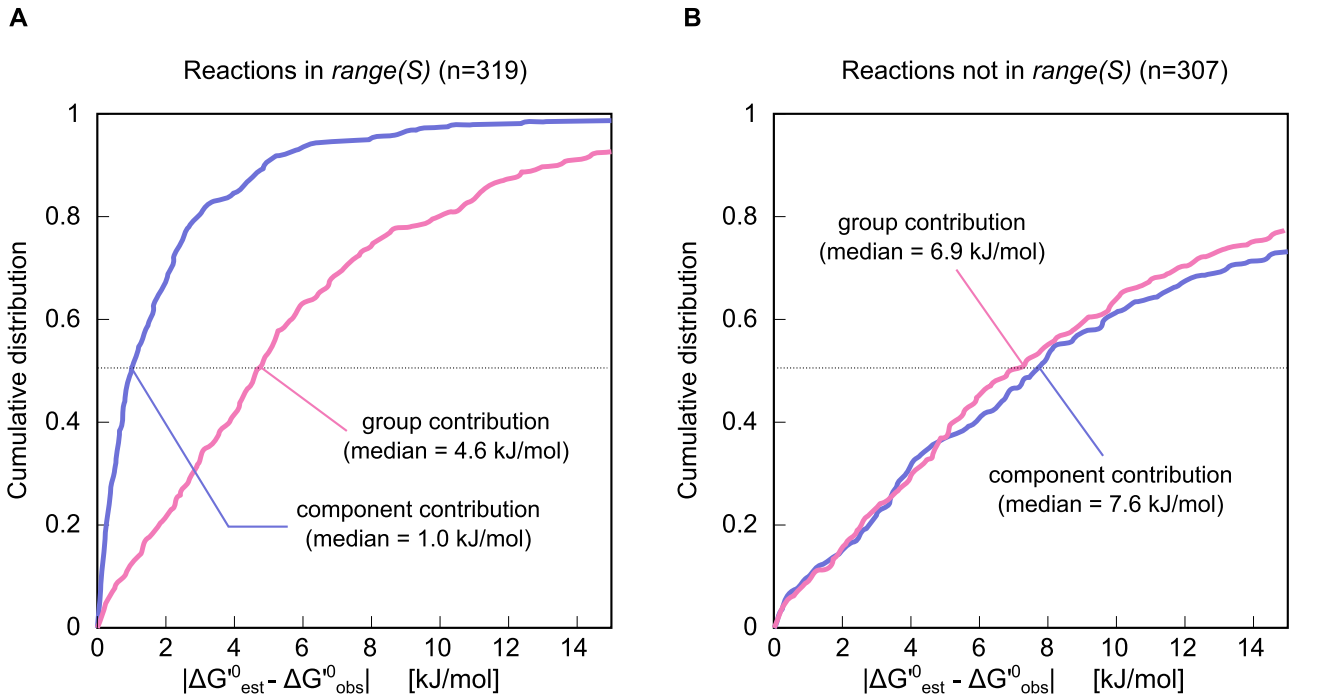

Figure 4. Cumulative distributions for the cross-validation results. The CDF of the absolute-value residuals for both group contribution ($||e_{rc}||$, pink) and component contribution ($||e_{cc}||$, purple). The reactions were separated to ones which are (A) linearly-dependent on the set of all other reactions (s_j is in the range of $S_{(j)}$, the stoichiometric matrix of all reactions except s_j), and (B) to ones which are linearly-independent (and thus component contribution uses group decompositions for at least part of the reaction). We found an 80% reduction in the median for the former set and no significant change for the latter (p-value = 0.45).
doi:10.1371/journal.pcbi.1003098.g004

Application to genome-scale metabolic reconstructions

A major application of the component contribution method is estimation of standard Gibbs energies for reactions in genome-scale reconstructions. Such large reaction networks require consistent and reliable estimates with high coverage. If estimates are not consistent, the risk of reference point violations increases with network size. As discussed in section *Adjustment to in vivo conditions*, metabolic models generally require estimates of standard transformed Gibbs energies, $\Delta_r G'^{\circ}_{est}$, at *in vivo* conditions. To meet this requirement, we have integrated the component contribution method into a new version (2.0) of von Bertalanffy [28] (see section *Implementation and availability of code*).

Here, we apply von Bertalanffy 2.0 to two reconstruction; the *E. coli* reconstruction iAF1260 [5] and the human reconstruction Recon 1 [29]. Standard transformed reaction Gibbs energies had previously been estimated for both reconstructions, with older versions of von Bertalanffy [8,10]. Those older versions relied on a combination of experimentally derived standard formation energies from [15], and estimated standard formation energies obtained with the group contribution method presented in [23]. We compare estimates obtained with the new version of von Bertalanffy, to both experimental data in TECRDB, and estimates obtained with the older versions.

$\Delta_r G'^{\circ}_{est}$ were obtained for 90% (1878/2078) of internal reactions in iAF1260, and 72% (2416/3362) of internal reactions in Recon 1. External reactions i.e., exchange, demand and sink reactions are not mass or charge balanced and therefore have no defined Gibbs energies. To validate our estimates we compared them to available experimental data. Measurements of apparent equilibrium constants (K') were available in TECRDB for 163 of the evaluated iAF1260 reactions, and 186 Recon 1 reactions. Multiple measurements, made at different experimental conditions, were often available for a single reaction. To enable comparison, the data in TECRDB was first normalized to standard conditions by applying an inverse Legendre transform as described in Section S1 in Text S1. The resulting standard reaction Gibbs energies ($\Delta_r G^{\circ}_{obs}$) were then adjusted to the conditions in Tables 1 and 2 with von Bertalanffy, to obtain standard transformed reaction Gibbs energies, $\Delta_r G'^{\circ}_{obs}$. Comparison of $\Delta_r G'^{\circ}_{est}$ to $\Delta_r G'^{\circ}_{obs}$ gave a root mean square error (RMSE) of 2.7 kJ/mol for iAF1260, and 3.1 kJ/mol for Recon 1.

von Bertalanffy 2.0 relies on component contribution estimated standard reaction Gibbs energies, whereas older versions relied on a combination of experimental data and group contribution estimates. Table 3 compares standard transformed Gibbs energy estimates, for iAF1260 and Recon 1, between versions. Use of component contribution resulted in both higher coverage and lower RMSE than was achieved with the previously available data.

Table 1. pH and electrical potential in each compartment of the *E. coli* reconstruction iAF1260.

Compartment	pH	Electrical potential (mV)
Cytosol	7.70	0
Periplasm	7.70	90
Extracellular fluid	7.70	90

Electrical potential in each compartment is relative to electrical potential in the cytosol. Temperature was set to 310.15 K (37°C), and ionic strength was assumed to be 0.25 M [14] in all compartments. Taken from [8].
doi:10.1371/journal.pcbi.1003098.t001

Table 2. pH and electrical potential in each compartment of the human reconstruction Recon 1.

Compartment	pH	Electrical potential (mV)
Cytosol	7.20	0
Extracellular fluid	7.40	30
Golgi apparatus	6.35	0
Lysosomes	5.50	19
Mitochondria	8.00	−155
Nucleus	7.20	0
Endoplasmic reticulum	7.20	0
Peroxisomes	7.00	12

Electrical potential in each compartment is relative to electrical potential in the cytosol. Temperature was set to 310.15 K (37°C), and ionic strength was assumed to be 0.15 M [14] in all compartments. Taken from [10].
doi:10.1371/journal.pcbi.1003098.t002

The greater coverage was due to reactions where groups or compounds that were not covered by component contributions canceled out, because they appeared unchanged on both sides of the reactions. Such reactions are easily identified and evaluated within the component contribution framework.

Another improvement achieved with the component contribution method was the lower standard error, s_r, of standard reaction Gibbs energy estimates compared with previously available methods (Table 3). This is an important improvement as standard error has previously been shown to affect predictions made based on reaction Gibbs energy estimates [6,8,10]. The reduction in s_r was obtained by accounting for covariances in parameter estimates (see section *Calculation of confidence intervals*). As we showed in section *Validation results*, the lower standard errors of component contribution estimates yielded reliable prediction intervals for observed standard reaction Gibbs energies. They can therefore be expected to also yield reliable confidence intervals for true standard reaction Gibbs energies.

The lower RMSE achieved with component contribution stems primarily from two factors. The first is the normalization of the training data by the inverse Legendre transform, which in [24] was shown to lead to significant improvements in group contribution estimates of Gibbs energies. The second factor is the greater number of reactions that are fully evaluated with reactant contribution (Eq. 6). Close to 10% of all evaluated reactions in both iAF1260 and Recon 1, were fully evaluated using only reactant contribution (Figure 5). Although this category represents a minority of all reactions, it includes the majority of reactions in central carbon metabolism. The greater accuracy in Gibbs energy estimates for reactions in central carbon metabolism is expected to have a disproportionally large effect, as these reactions are involved in most metabolic activities. To support this claim, we predicted 312 flux distributions for iAF1260 and 97 flux distributions for Recon 1 (see Section S6 in Text S1 for details). We found that the tenth of reactions that were fully evaluated with reactant contributions carried approximately half of the total flux in iAF1260 and a third of the total flux in Recon 1 (Figure 5).

Discussion

The component contribution method presented in this paper merges two established methods for calculating standard Gibbs energies of reactions while maintaining each of their advantages;

Table 3. Comparison of standard transformed reaction Gibbs energy estimates based on component contributions, to estimates based on previously available data.

	iAF1260		Recon 1	
	Fleming et al. [8]	**Current study**	**Haraldsdóttir et al. [10]**	**Current study**
Coverage	85%	90%	63%	72%
RMSE (kJ/mol)	9.9	2.7	11.6	3.1
Mean s_r (kJ/mol)	20.3	2.3	3.4	2.2

doi:10.1371/journal.pcbi.1003098.t003

accuracy in the case of reactant contribution (RC) and the wide coverage of group contribution (GC). By representing every reaction as a sum of two complementary component reactions, one in the subspace that is completely covered by RC and the other in the complementary space, we maximize the usage of information that can be obtained with the more accurate RC method. Overall, we find that there is a 50% reduction in the median absolute residual compared to standard GC methods, while providing the same wide coverage and ensuring that there are no reference-point inconsistencies that otherwise lead to large errors. Furthermore, since our method is based on least-squares linear regression, we use standard practices for calculating confidence intervals for standard Gibbs energies (see section *Calculation of confidence intervals*), and for weighing the measured standard Gibbs energies used as training data (see Section S1.2 in Text S1).

Since the empirical data used in our method is measured in various conditions (temperature, pH, ionic strength, metal ion concentrations, etc.) – it is important to "standardize" the input data before applying any linear regression model [24]. In this work, we used an inverse Legendre transform to normalize the pH and ionic strength, but ignore the temperature effect and the metal ion concentrations (see Section S1.1 in Text S1). In addition, the proton dissociation constants were obtained from a third party software estimator (by Marvin, see Methods) and have a mean absolute error of about 0.9 pH units [30]. Notably, a commendable effort for creating a database of thermodynamic quantities [31] has been published recently, where the data was standardized using more reliable parameters and considering more effects. This database currently only covers reactions from glycolysis, the tricarboxylic acid cycle, and the pentose phosphate pathway. Therefore, we chose to use the more extensive TECRDB database and perform the inverse Legendre transform ourselves, effectively increasing the coverage while compromising on the accuracy of the data. Since the changes brought forward in the component contribution method are independent of the source of input data, we believe that it will benefit from any future improvements in these databases.

The precision of the component contribution method is limited by the accuracy of the measured reaction equilibrium constants used in the regression model. In cases of isolated reactions, where the empirical data cannot be corroborated by overlapping measurements, large errors will be directly propagated to our estimate of those reactions' standard Gibbs energies. As the number of measurements underlying an estimate is reflected in its standard error, however, confidence intervals for such reactions will be large. It is therefore recommended to use confidence intervals, and not point estimates, for simulations and predictions based on standard Gibbs energy estimates. In the future, it might be worthwhile to integrate several promising computational

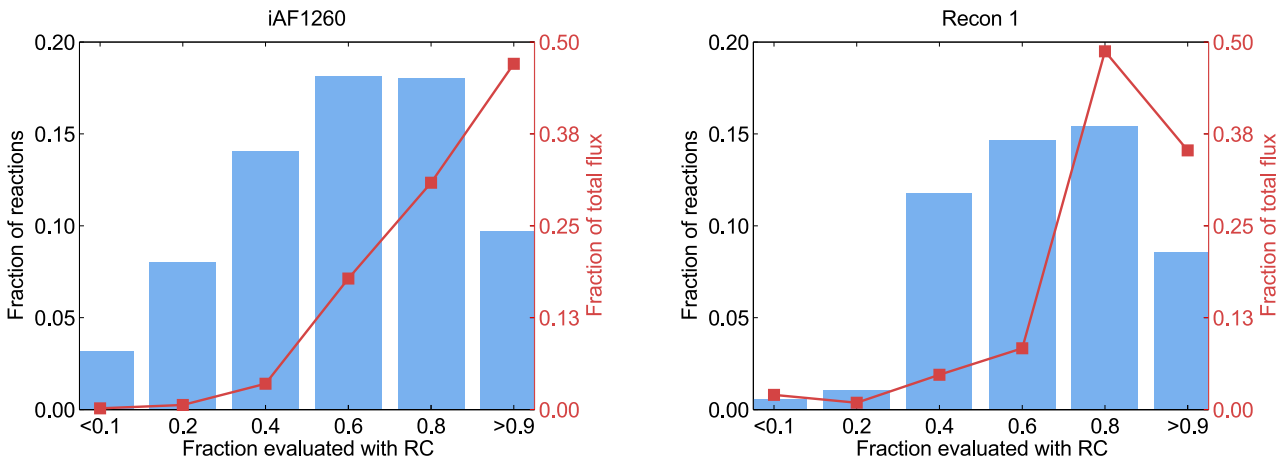

Figure 5. Distribution of the fractions of reaction vectors (black) in iAF1260 (*E. coli*) and Recon 1 (human), that were in the range of S, and were thus evaluated with reactant contribution (RC). For a reaction x, this fraction was calculated as $\|x_R\|^2/\|x\|^2$. Passive and facilitated diffusion reactions, where the reactants undergo no chemical changes, are not included in the figure. 9.4% of all evaluated reactions in iAF1260 were fully evaluated using only reactant contributions. These reactions carried approximately half of the total flux (red) in 312 predicted flux distributions. The 8.3% of evaluated reactions in Recon 1 that were fully evaluated with reactant contributions, carried close to a third of the total flux in 97 predicted flux distributions.
doi:10.1371/journal.pcbi.1003098.g005

prediction approaches [32] which are not based on RC and GC, such as molecular mechanics methods [33], density functional methods [34], and post Hartree-Fock approaches [35,36]. Although the computational cost of these methods can be substantial depending on the theoretical method and the solvation models [37] used, they have the advantage of being based on computable chemical and physical principles, implying that a 100% coverage of all biochemical reactions is achievable (though not yet practical). Currently, the accuracy of these methods for reactions in solution is limited. Nevertheless, they might already be useful for estimating $\Delta_r G^\circ$ of reactions that are not covered by component contributions, or for validating the sparse measurements. Alternatively, a method that infers $\Delta_r G^\circ$ from reaction similarities named IGERS [38] manages to be much more accurate than GC when predicting the standard Gibbs energy of reactions which are very similar to a reaction with a measured $\Delta_r G^\circ$. Adding IGERS as another layer between RC and GC using the ideas presented in this paper might contribute to the overall accuracy of our estimations. Finally, the laws of additivity suggested by [16] include single atom (zero-order) and single bond (first-order) contributions, which would be too crude to use for approximating Gibbs energies directly, but might be useful as two extra layers in a method like component contribution and help cover a wider fraction of the reaction space.

The use of thermodynamic parameters in modeling living systems has been hindered by the fact that it is mostly inaccessible or requires a high level of expertise to use correctly, especially in genome-scale models. In order to alleviate this limitation, we created a framework that facilitates the integration of standard reaction Gibbs energies into existing models and also embedded our code into the openCOBRA toolbox. The entire framework (including the source code and training data) is freely available. We envisage a collaborative community effort that will result in a simple and streamlined process where these important thermodynamic data are widely used and where future improvements in estimation methods will seamlessly propagate to modelers.

Methods

Calculation of confidence intervals

The component contribution estimated standard Gibbs energy $\Delta_r G'^\circ_{cc,x}$ in Eq. 10, is a point estimate of the true standard Gibbs energy $\Delta_r G'^\circ_x$ for reaction vector x. To quantify the uncertainty in this estimate, we need to calculate confidence intervals for $\Delta_r G'^\circ_x$. An important advantage of integrating the reactant and group contribution methods in a single, unified framework is that it greatly simplifies calculation of confidence intervals. We present the key equations in this section. A summary of the statistical theory underlying these equations [39] is given in Section S7 in Text S1.

The covariance matrix V_{rc} for the reactant contribution estimates ($\Delta_f G'^\circ_{rc}$ in Eq. 3) is calculated as

$$\begin{aligned} V_{rc} &= s_{rc}^2 \cdot \left(SS^T\right)^+ \\ &= \frac{\|e_{rc}\|^2}{n - \mathrm{rank}(S)} \cdot \left(SS^T\right)^+, \end{aligned} \tag{11}$$

where the matrix $(SS^{\mathrm{T}})^+$ is scaled by the estimated variance s_{rc}^2 of the error term ε_{rc} in Eq. 2. Our estimate of the variance was $s_{rc}^2 = 17.8$ (kJ/mol)^2. The covariance matrix V_{gc} for the group contribution estimates ($\Delta_g G'^\circ_{gc}$) is likewise obtained as

$$\begin{aligned} V_{gc} &= s_{gc}^2 \cdot \left(\mathcal{G}^T S S^T \mathcal{G}\right)^+ \\ &= \frac{\|e_{gc}\|^2}{n - \mathrm{rank}(S^T\mathcal{G})} \cdot \left(\mathcal{G}^T S S^T \mathcal{G}\right)^+, \end{aligned} \tag{12}$$

where the estimated variance of ε_{gc} from Eq. 7 was $s_{gc}^2 = 62.0$ (kJ/mol)^2.

For a reaction x, the *standard error* of $\Delta_r G'^\circ_{cc,x}$ is given by

$$\begin{aligned} {s_{cc,x}}^2 &= x_R^T \cdot V_{rc} \cdot x_R + x_N^T \cdot \mathcal{G} V_{gc} \mathcal{G}^T \cdot x_N \\ &= x^T \cdot \left(P_{\mathcal{R}(S)} V_{rc} P_{\mathcal{R}(S)} + P_{\mathcal{N}(S^T)} \mathcal{G} V_{gc} \mathcal{G}^T P_{\mathcal{N}(S^T)}\right) \cdot x. \end{aligned} \tag{13}$$

The confidence interval for $\Delta_r G'^\circ_x$, at a specified confidence level $\gamma \in [0\%, 100\%]$, is given by

$$\Delta_r G'^\circ_{cc,x} \pm z_\gamma s_{cc,x}, \tag{14}$$

where z_γ is the value of the standard normal distribution at a cumulative probability of $(100\% + \gamma)/2$. The 95% confidence interval for $\Delta_r G'^\circ_x$ is therefore $\Delta_r G'^\circ_{cc,x} \pm 1.96 \times s_{cc,x}$.

In calculating $s_{cc,x}$, we employ the covariance matrices for estimated parameters $\Delta_f G'^\circ_{rc}$ and $\Delta_g G'^\circ_{gc}$. In contrast, Jankowski et al. used only the diagonal of the covariance matrix for $\Delta_g G'^\circ_{gc}$ in their implementation of the group contribution method [23]. The main advantage of using covariance matrices is that it leads to more appropriate confidence intervals for $\Delta_r G'^\circ_x$, that can be much smaller. Knowledge about the relative Gibbs energy of two groups or compounds, increases with the number of measurements for reactions where those groups or compounds occur together. This knowledge should be reflected in smaller confidence intervals for reactions where the groups or compounds co-occur. Covariance matrices provide a means for propagating this knowledge. If only the diagonal of the covariance matrix is used, this knowledge is lost and confidence intervals will often be unnecessarily large.

The covariance matrices can likewise be used to propagate lack of knowledge to $s_{cc,x}$. If $\mathcal{G}^{\mathrm{T}} x$ is not in $\mathcal{R}\left(\mathcal{G}^{\mathrm{T}} S\right)$ then the reaction x is not covered by the group contribution method or by the component contribution method. Then $\Delta_r G'^\circ_{cc,x}$ obtained with Eq. 10 will not be a valid estimate of $\Delta_r G'^\circ_x$, and should have a large (infinite) standard error. This can be achieved by adding a term to Eq. 13;

$$\begin{aligned} s_{cc,x}^2 = x^{\mathrm{T}} \cdot \big(&P_{\mathcal{R}(S)} V_{rc} P_{\mathcal{R}(S)} \\ &+ P_{\mathcal{N}(S^{\mathrm{T}})} \mathcal{G} V_{gc} \mathcal{G}^{\mathrm{T}} P_{\mathcal{N}(S^{\mathrm{T}})} \\ &+ \mathcal{G} V_\infty \mathcal{G}^{\mathrm{T}}\big) \cdot x \end{aligned} \tag{15}$$

where $V_\infty = P_{\mathcal{N}(S^{\mathrm{T}}\mathcal{G})} \cdot \infty$, and $P_{\mathcal{N}(S^{\mathrm{T}}\mathcal{G})} \in \mathbb{R}^{g \times g}$ is a projection matrix onto the null-space of $S^{\mathrm{T}}\mathcal{G}$. Eq. 15 will give $s_{cc,x} = \infty$ for all reactions that cannot be evaluated with component contributions because $x^{\mathrm{T}}\mathcal{G}$ has a nonzero component in the null-space of $S^{\mathrm{T}}\mathcal{G}$. In practice, we use a very large value instead of ∞ (e.g. 10^{10} kJ/mol) which will dominate any reasonable Gibbs energy in case $x^{\mathrm{T}}\mathcal{G}$ is not orthogonal to this null-space.

Leave-one-out cross-validation

Both group contribution and component contribution are parametric methods that use a set of training data in order to evaluate a long list of parameters. In order to validate these models, we need to use more empirical data which has not been used in the training phase. Since data regarding reaction Gibbs energies is scarce, we apply the leave-one-out method in order to maximize the amount of data left for training in each cross-validation iteration. As a measure for the quality of the standard Gibbs energy estimations from each method we use the median absolute residual of the cross-validation results compared to the observations.

Our entire training set consists of 4146 distinct reaction measurements. However, since many of them are experimental replicates – measurements of the same chemical reaction in different conditions or by different researchers – we can only use each distinct reaction once. We thus take the median $\Delta_r G'^{\circ}_{obs}$ over all replicates (after applying the inverse Legendre transform) as the value to be used for training or cross-validation. We choose the median rather than the mean to avoid sensitivity to outliers. After this process of unifying observations, we are left with 694 unique reaction observations. Note that the repetitions do play a role in determining the standard error in standard Gibbs energy estimates (see section *Calculation of confidence intervals*). Finally, the vector of $\Delta_r G'^{\circ}_{obs}$ values for the unique reactions is projected onto the range of S^{T} since we assume that the actual values comply with the first law of thermodynamics (see section *Reactant contribution method*) and that any deviation is caused by experimental error.

Calculation of prediction intervals

The γ prediction interval for a reaction x, with estimated standard Gibbs energy $\Delta_r G'^{\circ}_{cc,x}$, is calculated as

$$\Delta_r G'^{\circ}_{cc,x} \pm z_{\gamma}\sqrt{s^2_{cc}+s^2_{cc,x}}, \tag{16}$$

where z_{γ} was defined in Eq. 14, and $s^2_{cc,x}$, the standard error of $\Delta_r G'^{\circ}_{cc,x}$, was defined in Eq. 15. s^2_{cc} is calculated as

$$s^2_{cc} = \frac{\|x_R\|^2}{\|x\|^2}\cdot s^2_{rc} + \frac{\|x_N\|^2}{\|x\|^2}\cdot s^2_{gc} \tag{17}$$

i.e., it is a weighted mean of the estimated variances for reactant and group contribution, where the weights are the fractions of x that are in $\mathcal{R}(S)$ and $\mathcal{N}(S)$, respectively. A summary of the statistical theory underlying calculation of prediction intervals [39] is given in Section S7 in Text S1.

Adjustment to *in vivo* conditions

For an input reaction x, the component contribution method outputs an estimate of the reaction's standard *chemical* Gibbs energy $\Delta_r G'^{\circ}_x$. In a chemical reaction each compound is represented in a specific protonation state. This is in contrast to biochemical reactions, where each compound is represented as a pseudoisomer group of one *or more* species in different protonation states. To thermodynamically constrain models of living organisms we require Gibbs energies of biochemical reactions at *in vivo* conditions, known as standard *transformed* reaction Gibbs energies $\Delta_r G'^{\circ}$.

We estimated $\Delta_r G'^{\circ}$ with version 2.0 of von Bertalanffy [8,10,28]; a Matlab implementation of biochemical thermodynamics theory as presented in [14]. A comprehensive summary of the relevant theory is given in [10]. In addition to component contribution estimates of standard Gibbs energies, required inputs to von Bertalanffy are a stoichiometric matrix S_{recon} for a metabolic reconstruction of an organism, pK_a values for compounds in S_{recon}, and literature data on temperature, pH, ionic strength (I) and electrical potential (ϕ) in each cell compartment in the reconstruction.

We estimated $\Delta_r G'^{\circ}$ for reactions in two multi-compartmental, genome scale metabolic reconstructions; an *E. coli* reconstruction iAF1260 [5], and a human reconstruction Recon 1 [29]. The environmental parameters pH, I and ϕ were taken from [8] for *E. coli* (Table 1), and from [10] for human (Table 2). pK_a values were estimated with Calculator Plugins, Marvin 5.10.1, 2012, ChemAxon (http://www.chemaxon.com).

Implementation and availability of code

The component contribution method has been implemented in both Matlab and Python. The Matlab implementation is tailored towards application to genome-scale metabolic reconstructions. It is fully compatible with the COBRA toolbox [40] and is freely available as part of the openCOBRA project on Sourceforge (http://sourceforge.net/projects/opencobra/). The component contribution method has been integrated into version 2.0 of von Bertalanffy to provide an easy-to-use tool to estimate transformed Gibbs energies at *in vivo* conditions. The Python implementation is a stand-alone package that can be used by researchers with suitable programming skills. The Python package includes a simple front-end called eQuilibrator (http://equilibrator.weizmann.ac.il/), which is a freely available online service. The Python code for component contribution is licensed under the open source MIT License and available on GitHub (https://github.com/eladnoor/component-contribution). Our code depends on the open source chemistry toolbox called Open Babel [41].

Supporting Information

Text S1 Supporting text with sections on 1) the inverse Legendre transform of the training data, 2) group decomposition, 3) the full mathematical derivation of the component contribution method, 4) estimation of error in the group model, 5) reaction type statistics, 6) prediction of flux distributions, 7) the theory underlying calculation of confidence and prediction intervals, and 8) mathematical symbols used throughout the manuscript.
(PDF)

Acknowledgments

We thank Arren Bar-Even, Wolfram Liebermeister, Naama Tepper, Tomer Shlomi, Bastian Niebel, Steinn Gudmundsson, Adrian Jinich, Dmitrij Rappoport, and William R. Cannon for helpful discussions.

Author Contributions

Conceived and designed the experiments: EN HSH RM RMTF. Performed the experiments: EN HSH. Analyzed the data: EN HSH. Contributed reagents/materials/analysis tools: EN HSH RM RMTF. Wrote the paper: EN HSH RM RMTF.

References

1. Beard DA, Babson E, Curtis E, Qian H (2004) Thermodynamic constraints for biochemical networks. journal of theoretical biology 228: 327–33.
2. Schellenberger J, Lewis NE, Palsson BO (2011) Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophysical journal 100: 544–53.
3. Fleming RMT, Maes CM, Saunders Ma, Ye Y, Palsson BO (2012) A variational principle for computing nonequilibrium uxes and potentials in genome-scale biochemical networks. Journal of theoretical biology 292: 71–7.
4. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V (2006) Genome-scale thermodynamic analysis of Escherichia coli metabolism. Biophysical journal 90: 1453–61.
5. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular systems biology 3: 121.
6. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophysical journal 92: 1792–805.
7. Zamboni N, Kummel A, Heinemann M (2008) fanNET:g a tool for network-embedded thermodynamic analysis of quantitative metabolome data. BMC Bioinf 9: 199.
8. Fleming RMT, Thiele I, Nasheuer HP (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. Biophysical chemistry 145: 47–56.
9. Fleming RMT, Thiele I, Provan G, Nasheuer HP (2010) Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. Journal of Theoretical Biology 264: 683–692.
10. Haraldsdóttir HS, Thiele I, Fleming RMT (2012) Quantitative assignment of reaction directionality in a multicompartmental human metabolic reconstruction. Biophysical journal 102: 1703–11.
11. Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions–a database for quantitative biochemistry. Bioinformatics (Oxford, England) 20: 2874–7.
12. Krebs HA, Kornberg HL, Burton K (1957) Energy transformation in living matter. Berlin, Germany: Springer.
13. Thauer RK, Jungermann K, Decker K (1977) Energy conservation in chemotrophic anaerobic bacteria. Bacteriological reviews 41: 809.
14. Alberty RA (2003) Thermodynamics of Biochemical Reactions. Hoboken N.J.: John Wiley & Sons, 0–2 pp.
15. Alberty RA (2006) Biochemical Thermodynamics: Applications of Mathematica (Methods of Biochemical Analysis). Wiley-Interscience, 480 pp.
16. Benson SW, Buss JH (1958) Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. The Journal of Chemical Physics 29: 546.
17. Benson SW (1967) Thermochemical Kinetics. New York, NY: John Wiley & Sons, Inc., 320 pp.
18. Benson SW, Cruickshank FR, Golden DM, Haugen GR, O'Neal HE, et al. (1969) Additivity rules for the estimation of thermochemical properties. Chemical Reviews 69: 279–324.
19. Ritter ER, Bozzelli JW (1991) THERM: Thermodynamic property estimation for gas phase radicals and molecules. International Journal of Chemical Kinetics 23: 767–778.
20. Domalski ES, Hearing ED (1988) Estimation of the Thermodynamic Properties of Hydrocarbons at 298.15 K. Journal of Physical and Chemical Reference Data 17: 1637.
21. Mavrovouniotis ML, Bayol P, Lam TkM, Stephanopoulos G, Stephanopoulos G (1988) A group contribution method for the estimation of equilibrium constants for biochemical reactions. Biotechnology Techniques 2: 23–28.
22. Mavrovouniotis ML (1991) Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. Biotechnology and Bioengineering 38: 803–804.
23. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophysical journal 95: 1487–99.
24. Noor E, Bar-Even A, Flamholz A, Lubling Y, Davidi D, et al. (2012) An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. Bioinformatics (Oxford, England) 28: 2037–2044.
25. Berry SR, Rice SA, Ross J (2000) Thermochemistry and its applications. Oxford: Oxford University Press, 2nd edition, 388–419 pp.
26. Dill KA, Bromberg S (2003) Molecular driving forces: Statistical thermodynamics in Chemistry and Biology. London: Garland Science.
27. Alberty RA (2002) Inverse Legendre Transform in Biochemical Thermodynamics: Illustrated with the Last Five Reactions of Glycolysis. The Journal of Physical Chemistry B 106: 6594–6599.
28. Fleming RMT, Thiele I (2011) von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. Bioinformatics (Oxford, England) 27: 142–3.
29. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences of the United States of America 104: 1777–82.
30. Lee AC, Crippen GM (2009) Predicting pKa. Journal of chemical information and modeling 49: 2013–33.
31. Li X, Wu F, Qi F, Beard DA (2011) A database of thermodynamic properties of the reactions of glycolysis, the tricarboxylic acid cycle, and the pentose phosphate pathway. Database : the journal of biological databases and curation 2011: bar005.
32. Irikura KK, Frurip DJ (1998) Computational Thermochemistry. In: ACS Symposium Series, Washington, DC: American Chemical Society, volume 677 of ACS Symposium Series, chapter 1. pp. 2–18.
33. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. Journal of the American Chemical Society 118: 11225–11236.
34. Goerigk L, Grimme S (2010) A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions - Assessment of Common and Reparameterized (meta -)GGA Density Functionals. Journal of Chemical Theory and Computation 6: 107–126.
35. Peterson Ka, Feller D, Dixon Da (2012) Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges. Theoretical Chemistry Accounts 131: 1079.
36. Bylaska EJ, Glaesemann KR, Felmy AR, Vasiliu M, Dixon DA, et al. (2010) Free energies for degradation reactions of 1,2,3-trichloropropane from ab initio electronic structure theory. The journal of physical chemistry A 114: 12269–82.
37. Marenich AV, Ding W, Cramer CJ, Truhlar DG (2012) Resolution of a Challenge for Solvation Modeling: Calculation of Dicarboxylic Acid Dissociation Constants Using Mixed DiscreteContinuum Solvation Models. The Journal of Physical Chemistry Letters 3: 1437–1442.
38. Rother K, Hoffmann S, Bulik S, Hoppe A, Gasteiger J, et al. (2010) IGERS: inferring Gibbs energy changes of biochemical reactions from reaction similarities. Biophysical journal 98: 2478–86.
39. Kutner MH, Nachtsheim CJ, Neter J, Li W (2004) Multiple Regression I. In: Applied Linear Statistical Models, McGraw-Hill/Irwin, chapter 6. 5 edition, pp. 214–255.
40. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature protocols 6: 1290–307.
41. O'Boyle NM, Banck M, James Ca, Morley C, Vandermeersch T, et al. (2011) Open Babel: An open chemical toolbox. Journal of cheminformatics 3: 33.