# Supplementary text and tables

## Comparison with the group contribution method of Mavrovouniotis

Although most of the *interaction factors* and *structural groups* used in the new group contribution method are based on the *interaction factors* and *structural groups* used in the group contribution method of Mavrovouniotis (1, 2), significant changes were made in the development of this new group contribution method. 20 new *structural groups* (all noted in Table 1) were introduced into the method to allow for the calculation of $\Delta G^{'\circ}_{est}$ for a wider variety of compounds and reactions. Additionally, some of the groups used in the Mavrovouniotis group contribution method were combined into more general *structural groups*. In the Mavrovouniotis method, primary, secondary, and tertiary hydroxyl groups and a hydroxyl group attached to a benzene ring were all treated as separate *structural groups*. After fitting with the four hydroxyl groups delineated, $\Delta_{gr} G^{'\circ}_{i}$ values for the hydroxyl groups were found to be similar, with a maximum difference of 4 kcal/mol and average difference of 2.28 kcal/mol between the values (Table SI). Primary, secondary, and tertiary phosphates were also treated as separate groups in the Mavrovouniotis method, and the $\Delta_{gr} G^{'\circ}_{i}$ values initially calculated for these three phosphate groups were also very similar with a maximum difference of 2 kcal/mol and an average difference of 1.2 kcal/mol. In order to simplify the group contribution method, all hydroxyl groups and all phosphate groups were combined into the single *structural groups* -OH and -O-$PO_3^{2-}$, respectively. Although these changes did result in a 3.3% increase in the $SE_{MLR}$ for the fitting from 1.84 to 1.90 kcal/mol, this increase is acceptable given the corresponding 5.8% reduction in the number of parameters involved in the group contribution model. Note that in both cases, the number of datapoints in the training set involving the generic -OH group or the generic -O-$PO_3^{2-}$ group is not equal to the total number of datapoints involving each of the specific -OH or -O-$PO_3^{2-}$ groups (Table SI). One reason for this apparent discrepancy is that some datapoints contain more than one of the specific -OH or -O-$PO_3^{2-}$ groups. For example, glucose contains a primary and a secondary hydroxyl group which means that glucose is counted among both the 683 datapoints containing primary -OH and the 902 datapoints containing secondary -OH. Another reason for the discrepancy is that some reactions involve the destruction of a primary -OH or -O-$PO_3^{2-}$ along with the simultaneous creation of a secondary -OH or -O-$PO_3^{2-}$, but the generic -OH and -O-$PO_3^{2-}$ cancel out in these reactions.

Changes were also made to the *interaction factors* used in the group contribution method. Three new types of *interaction factors* were introduced in this new group contribution method. These new *interaction factors* are discussed in detail in the methods section of the manuscript. Additionally, some of the *interaction factors* which were included in the group contribution method of Mavrovouniotis were not included in the final implementation of the new group contribution method: (i) the *NAD(P)H factor*, (ii) the *CoA factor*, and (iii) the *origin factor*. As implemented by Mavrovouniotis, the *NAD(P)H factor* was added to $\Delta_r G^{'\circ}_{est}$ of all reactions involving NAD(P) and NAD(P)H as cofactors in order to improve the agreement between $\Delta_r G^{'\circ}_{est}$ and $\Delta_r G^{'\circ}_{obs}$ for reactions. Similarly, the *CoA factor* was added to all reactions involving the addition or removal of

CoA. Finally, the *origin factor* was added to $\Delta_f G^{'\circ}_{est}$ of every compound that was decomposed into structural subgroups. These interaction factors were not included in the new group contribution method because all three had high T-tests and an insignificant effect on the $SE_{MLR}$ of the fitting (Table SII).

All of these changes introduced in this new group contribution method have not only expanded the applicability of the method to calculate $\Delta G^{'\circ}_{est}$ for a wider range of compounds and reactions, but also improved the accuracy of the method. For the compounds and reactions in the training set for which $\Delta G^{'\circ}_{est}$ can be calculated using the Mavrovouniotis group contribution method, the standard deviation of the residuals is 3.92 kcal/mol, compared to a standard deviation of 1.98 kcal/mol when the new group contribution method is used to calculate $\Delta G^{'\circ}_{est}$ for the same reactions and compounds.

## F-test calculation details

The validity of the linear model proposed in the group contribution hypothesis was assessed using a statistical F-test. An F-test indicates whether or not the variability in the $\Delta G^{'\circ}_{obs}$ values that make up the training set that is captured by the group contribution model is statistically significant compared to the variability not captured by the model (the variances between $\Delta G^{'\circ}_{obs}$ and $\Delta G^{'\circ}_{est}$) (3). The *F* value for the model is calculated as follows (3):

$$F = \frac{\left(\mathbf{\Delta G^{'\circ}_{est}}\right)^{'}\left(\mathbf{\Delta G^{'\circ}_{est}}\right)/N_{gr}}{\left(SE_{MLR}\right)^{2}} \qquad (1)$$

The location of this *F* value in the F-cumulative distribution function is then determined,, and if the location of the F value corresponds to a probability value greater than 90%, the linear model is accepted (3).

## T-test calculation details

While the F-test is used to test if the entire group contribution model is statistically significant compared to the uncertainty in the model, the t-test is used to determine if each $\Delta_{gr} G^{'\circ}_{i}$ in the group contribution model is statistically significant compared to the uncertainty in $\Delta_{gr} G^{'\circ}_{i}$, $SE_{gr,i}$ (3). The t-value for each parameter is calculated as follows (3):

$$t_i = \frac{\Delta_{gr} G^{'\circ}_{i}}{SE_{gr,i}} \qquad (2)$$

The $\Delta_{gr} G^{'\circ}_{i}$ of a group is considered to be statistically significant compared to the $SE_{gr,i}$ for the group if the location of the $t_i$ value in the student t-cumulative distribution corresponds to a probability value less than 5% (3). Only *interaction factors* with t-tests over 5% were removed from the group contribution method. While t-tests were performed on the *structural groups* as well, *structural groups* with high t-tests were not removed from the model because they were required for the complete decomposition of the molecular structures involved in the training set.

## Table S1. Generalized alcohol and phosphate structural subgroups

| Description of molecular substructure | $\Delta_{gr}G'^{\circ}$ | $SE_{gr}$ | Frequency |
|---|---|---|---|
| | kcal/mol | kcal/mol | |
| **Alcohol groups** | | | |
| -OH (generalized) † | -41.5 | 0.126 | 1117 |
| -OH (primary) | -40.8 | 0.125 | 683 |
| -OH (secondary) | -43.2 | 0.165 | 902 |
| -OH (tertiary) | -44.8 | 0.510 | 314 |
| -OH (attached to a benzene ring) | -41.5 | 0.562 | 28 |
| **Phosphate groups** | | | |
| $-O-PO_3^{2-}$ (generalized) † | -254 | 0.159 | 380 |
| $-O-PO_3^{2-}$ (primary) | -253 | 0.155 | 483 |
| $-O-PO_3^{2-}$ (secondary) | -254 | 0.229 | 207 |
| $-O-PO_3^{2-}$ (tertiary) | -252 | 1.83 | 2 |

†These are the groups actually included in the final version of the method

## Table S2. Mavrovouniotis interaction factors that were removed

| Interaction factor | $\Delta_{gr}G'^{\circ}$ | $SE_{gr}$ | t-test | Frequency | $SE_{MLR}$ |
|---|---|---|---|---|---|
| | kcal/mol | kcal/mol | | | with/without |
| Origin term | 0.352 | 0.850 | 0.68 | 660 | 1.90/1.90 |
| NADH | 0.605 | 0.427 | 0.16 | 690 | 1.90/1.90 |
| CoA | 1.64 | 0.777 | 0.03 | 165 | 1.90/1.90 |

1. Mavrovouniotis, M. L. 1990. Group contributions for estimating standard Gibbs energies of formation of biochemical-compounds in aqueous-solution. Biotechnology and Bioengineering 36:1070-1082.
2. Mavrovouniotis, M. L. 1991. Estimation of standard Gibbs energy changes of biotransformations. Journal of Biological Chemistry 266:14440-14445.
3. Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied linear statistical models : regression, analysis of variance, and experimental designs. Irwin, Homewood, IL.