Bioinformatics, 35(15), 2019, 2634–2643
doi: 10.1093/bioinformatics/bty1035
Advance Access Publication Date: 24 December 2018
Original Paper

OXFORD

Systems biology

Systematic selection of chemical fingerprint features improves the Gibbs energy prediction of biochemical reactions

Meshari Alazmi[1,†], Hiroyuki Kuwahara [1,*,†], Othman Soufan[2], Lizhong Ding[3] and Xin Gao[1,*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal 23955-6900, Saudi Arabia, [2]Institute of Parasitology, McGill University, Montreal, Quebec, Canada and [3]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on May 28, 2018; revised on September 26, 2018; editorial decision on December 16, 2018; accepted on December 19, 2018

Abstract

Motivation: Accurate and wide-ranging prediction of thermodynamic parameters for biochemical reactions can facilitate deeper insights into the workings and the design of metabolic systems.
Results: Here, we introduce a machine learning method with chemical fingerprint-based features for the prediction of the Gibbs free energy of biochemical reactions. From a large pool of 2D fingerprint-based features, this method systematically selects a small number of relevant ones and uses them to construct a regularized linear model. Since a manual selection of 2D structure-based features can be a tedious and time-consuming task, requiring expert knowledge about the structure-activity relationship of chemical compounds, the systematic feature selection step in our method offers a convenient means to identify relevant 2D fingerprint-based features. By comparing our method with state-of-the-art linear regression-based methods for the standard Gibbs free energy prediction, we demonstrated that its prediction accuracy and prediction coverage are most favorable. Our results show direct evidence that a number of 2D fingerprints collectively provide useful information about the Gibbs free energy of biochemical reactions and that our systematic feature selection procedure provides a convenient way to identify them.
Availability and implementation: Our software is freely available for download at http://sfb.kaust.edu.sa/Pages/Software.aspx.
Contact: hiro.kuwahara@kaust.edu.sa or xin.gao@kaust.edu.sa.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Thermodynamic data provide useful information to constrain the functional repertoire of metabolic networks from their structures (Ataman and Hatzimanikatis, 2015; Beard *et al.*, 2004; Held and Sadowski, 2016; Toure and Dussap, 2016). With advances in the characterization of the metabolome, thus, increasingly important becomes the role of thermodynamics in functional analysis of the endogenous metabolism of organisms (Feist *et al.*, 2007; Großkopf and Soyer, 2016; Henry *et al.*, 2006; Kümmel *et al.*, 2006) and metabolic engineering for natural product biosynthesis (Carbonell *et al.*, 2014; Kuwahara *et al.*, 2016; Lee *et al.*, 2012; Nielsen, 1998; Yim *et al.*, 2011). Unfortunately, however, experimental thermodynamic data for metabolic reactions have thus far been limited to only a small fraction of known biochemical reactions (Flamholz *et al.*, 2012;

Goldberg *et al.*, 2004), making *in silico* prediction of biochemical thermodynamic parameters not only necessary but also essential to a deeper understanding of the workings of metabolic systems.

The Gibbs free energy prediction problem can be treated as a regression problem with linear constraints imposed by the first law of thermodynamics, and a number of computational methods have been proposed to tackle this thermodynamics constrained regression task (e.g. Jankowski *et al.*, 2008; Jinich *et al.*, 2014; Mavrovouniotis *et al.*, 1988; Noor *et al.*, 2013; Rother *et al.*, 2010). Most commonly applied ones are variants of the group contribution method (Jankowski *et al.*, 2008; Mavrovouniotis *et al.*, 1988; Noor *et al.*, 2012, 2013). The group contribution method is a linear regression method which uses predefined 2D substructures as features and estimates the Gibbs free energy of formation of a compound by the sum of the weight of substructure fragments into which this compound can be decomposed. Because these substructure fragments can be combined to compose molecules that are not seen in the training set, the group contribution method has a potential to cover a wide range of biochemical reactions. The main challenge of the group contribution method is to identify useful 2D substructure fragments that can be used as its features. The manual selection of such substructure fragments is a complex, tedious and time-consuming task requiring expert knowledge on the structure-activity-relationship (SAR) of metabolites. This is because the selection of substructure features needs to satisfy two objectives: (i) to compose all compounds in the training by the substructure fragments and (ii) to have useful substructure fragments for the biochemical thermodynamic prediction. Furthermore, since the selection of these features depends strongly on the compounds present in the training set, its decision also needs to take into account the characterization of the 2D structure of unseen compounds so as to avoid severely limited prediction coverage.

In this paper, we introduce a new linear regression-based method, called the fingerprint contribution (FC) method, that we developed for the prediction of Gibbs free energy of biochemical reactions. The FC method is a two-step method which represents chemical compounds by features based on 2D fingerprints and molecular descriptors. In the first step, from a large pool of 2D fingerprint-based features, it systematically selects a smaller set of relevant ones that is expected to exhibit low generalization error, and in the second step, it uses the selected features in a regularized regression method to construct the final linear model. This new method overcomes usability limitations found in the group contribution method. While substructure fragment-based features used in the group contribution method can only represent chemical compounds that can be composed by some substructure fragments in the feature set, 2D fingerprint-based features used in our new method can represent any chemical compounds with concrete 2D structures. Thus, the FC model can cover a much wider range of chemical reactions than the group contribution variants. Indeed, the FC model is able to predict Gibbs free energy of virtually any biochemical reactions in which structurally characterized compounds participate. Furthermore, unlike the group contribution method, in which substructure fragment-based features are manually selected in a tedious process to ensure that they cover the composition of all chemical compounds in the training set, the FC method uses 2D fingerprinting methods to generate potential features and systematically select relevant ones in a much more convenient and efficient way.

Here, to analyze the value of the FC method, we compared its performance with that of state-of-the-art linear regression-based methods for the Gibbs free energy prediction. Our results demonstrated that the FC method outperformed the other method in terms of the prediction accuracy and the prediction coverage. These suggest that a number of 2D fingerprints provide useful information about the biochemical thermodynamics and that our systematic feature selection procedure can identify relevant fingerprints to improve the prediction of the Gibbs free energy of biochemical reactions.

2 Materials and methods

2.1 Significance test of the range-based partition for accuracy

With a null hypothesis that the observed difference in the prediction error between these linear dependency-based subgroups can be obtained by chance, we performed random permutation test, in which we randomly partitioned the Noor *et al.*-based dataset 1 million times into two subgroups based on the size of the in-range-reactions and the out-of-range reactions. Let n_{OR} and ϵ_{OR} be the size and the mean absolute error (MAE) of the out-of-range reactions, respectively. Then, we performed random permutation test, in which we randomly sampled reactions of size n_{OR} from the Noor *et al.*-based dataset 1 million times and measured the MAE for each sample as the test statistic. With this, we computed the *P*-value as the probability that the test statistic is higher than or equal to ϵ_{OR} in this sampling distribution. Clearly, this *P*-value is also the same as the probability that the MAE of the unchosen reactions is lower than or equal to the MAE of the in-range reactions.

2.2 Prediction error estimation for the KEGG dataset

The weighted average approach we used for the estimation of prediction error for the KEGG reactions is as follows:

$$\hat{\epsilon}_{KEGG}(m) = \alpha_{KEGG}(m)\epsilon_{OR}(m) + (1 - \alpha_{KEGG}(m))\epsilon_{IR}(m), \qquad (6)$$

where $\hat{\epsilon}_{KEGG}(m)$ is a predicted MAE for model m on the KEGG dataset, $\alpha_{KEGG}(m)$ is the fraction of the out-of-range reactions in the KEGG dataset with respect to the design matrix for the construction of model m, $\epsilon_{OR}(m)$ is the MAE of the out-of-range reactions from m in the leave-one-out cross validation (LOOCV), and $\epsilon_{IR}(m)$ is the MAE of the in-range reactions from m in the LOOCV. To evaluate the prediction accuracy of the FC model, the, group contribution (GC) model and the reactant contribution (RC) model with this measure, thus, the linear dependency of each reaction in the KEGG dataset was analyzed by examining whether its feature vector is a linear combination of the row vectors of the design matrix (Noor *et al.*, 2012). In the case of component contribution (CC) model, since it is a hybrid model of the RC model and the GC model, a reaction in the KEGG dataset was determined to be an in-range reaction if it is an in-range reaction in either the RC model or the GC model and an out-of-range reaction otherwise (Noor *et al.*, 2013).

3 Results

3.1 Fingerprint contribution model

3.1.1 Energy conservation constraint

The reaction Gibbs energy prediction problem can be treated as a regression problem with a linear constraint imposing the principle of energy conservation. The energy conservation constraint can be seen via the expression of the standard reaction Gibbs free energy of a chemically balanced reaction based on the standard Gibbs free energy of formation of the participating compounds. Let $\mathbf{C} = \{c_1, \ldots, c_n\}$ be a set of n chemical compounds and $\Delta_f \boldsymbol{G}^0 = (\Delta_f G_1^0, \ldots, \Delta_f G_n^0)^T$ be an n-dimensional vector whose ith element, $\Delta_f G_i^0$, represents the standard Gibbs free energy of formation of compound c_i. Then, the first law of thermodynamics implies that the

standard reaction Gibbs free energy of chemical reaction R_j, $\Delta_r G_j^0$, be expressed by the following expression:

$$\Delta_r G_j^0 = (\Delta_f \boldsymbol{G}^0)^T \boldsymbol{\nu_j}, \tag{1}$$

where $\boldsymbol{\nu_j} = (\nu_{j1}, \ldots, \nu_{jn})^T$ is an n-dimensional vector whose ith element, ν_{ji}, represents the stoichiometric coefficient of compound c_i in reaction R_j. This shows that the standard reaction Gibbs free energy is a linear transformation of the stoichiometric vector, and the energy conservation constraint can be described by the additive property of linear transformation. To illustrate this point, suppose there is a set of the following three reactions:

$$R_1: \ c_1 \rightleftharpoons c_2, \ \ R_2: \ c_2 \rightleftharpoons c_3, \ \ R_3: \ c_3 \rightleftharpoons c_1.$$

Then, since these three reactions create a futile cycle, the energy conservation constraint implies that $\Delta_r G_1^0 + \Delta_r G_2^0 + \Delta_r G_3^0 = 0$. Because the violation of the first law of thermodynamics most likely leads to physically irrelevant solutions, similar to state-of-the-art methods such as those based on the group contribution method (Jankowski *et al.*, 2008; Mavrovouniotis *et al.*, 1988; Noor *et al.*, 2012, 2013), we treat the energy conservation constraint as a hard constraint in the reaction Gibbs energy prediction problem.

3.1.2 Chemical fingerprint-based linear model

Fingerprint-contribution (FC) model is a linear model with chemical fingerprint and molecular descriptor-based features for the prediction of the standard Gibbs free energy of biochemical reactions. By representing the chemical transformation of each reaction R_j by a numerical vector of 2D fingerprint-based features, $\mathbf{x_j} \in \mathbb{R}^p$, the FC model forms the following linear function:

$$f(\mathbf{x_j}) = \mathbf{w}^T \mathbf{D} \mathbf{x_j}, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^p$ is a vector of the feature weights and $\mathbf{D}$ is a P-by-P diagonal matrix to normalize the input feature vector. This linear model can be derived from Equation 1 by assuming that each $\Delta_f G_i^0$ be represented by the weighted sum of chemical fingerprint-based features. That is, with a function $h: \mathbf{C} \mapsto \mathbb{R}^p$ which maps compound $c_i \in \mathbf{C}$ to a P-dimensional numerical vector of chemical fingerprint-based features, we express $\Delta_f \boldsymbol{G^0}$ by $\mathbf{FDw}$ where $\mathbf{F} = (h(c_1), \ldots, h(c_n))^T$ is an n-by-P compound-feature matrix. From this, the standard Gibbs free energy of reaction R_j can be characterized in terms of the chemical fingerprint-based features as

$$\Delta_r G_j^0 = (\mathbf{FDw})^T \boldsymbol{\nu_j} = \mathbf{w}^T \mathbf{D} \mathbf{x_j}, \tag{3}$$

where $\mathbf{x_j} = \mathbf{F}^T \boldsymbol{\nu_j}$. This shows that the P-dimensional feature vector in the FC model is a linear transformation of the n-dimensional stoichiometric vector via the compound-feature matrix. Thus, the FC model satisfies the energy conservation constraint.

3.2 Overview of the FC method

We assume to be given a training set $\{(R_1, y_1), \ldots, (R_m, y_m)\}$ where each y_j is the observed standard reaction Gibbs free energy of reaction R_j. Given this training set, we first select a subset of relevant fingerprint-based features, and then we apply a regularized linear regression using only the selected features based on the following equation

$$\mathbf{y} = \mathbf{DXw} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{y} = (y_1, \ldots, y_m)^T$ is an m-dimensional vector for the observed standard reaction Gibbs free energies, $\mathbf{X} = \mathbf{SF}$ is the m-by-P design matrix derived from the product of the m-by-n stoichiometric matrix, $\mathbf{S} = (\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_m})^T$, and an n-by-P compound-feature matrix, $\mathbf{F}$, and $\boldsymbol{\varepsilon}$ is an m-dimensional vector of uncorrelated random variables with zero mean and finite variance. In this regression, since $\mathbf{X}$ is known and $\mathbf{D}$ is derived from $\mathbf{X}$, our objective is to learn the weight $\mathbf{w}$.

We first select a subset of relevant fingerprint-based features, and then we apply a regularized linear regression using only the selected features to learn the weights $\mathbf{w}$. Let $h_0: \mathbf{C} \mapsto \mathbb{R}^{p_0}$ be a function which maps each compound to a p_0-dimensional vector that contains the original set of chemical fingerprint-based features. With feature selection, we wish to filter out many features and select a smaller subset of relevant ones from the original features (i.e. $p \ll p_0$). By using h_0, we can represent each compound by the initial chemical fingerprint-based features and generate an n-by-p_0 compound-feature matrix $\mathbf{F}_0$, which, in turn, allows us to construct the m-by-p_0 initial design matrix $\mathbf{X}_0 = \mathbf{SF}_0$ where $\mathbf{S} = (\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_m})^T$ is the stoichiometric matrix.

We first remove each feature that gives a zero-column in $\mathbf{X}_0$. To further filter the features, we analyze the multicollinearity of the remaining features and remove those features that can be safely represented by some other features. By defining highly correlated features to be those whose columns have pairwise correlation values greater than threshold ρ, we remove features so that none of the pairs in the remaining features has a correlation value larger than ρ. Note that, in this unsupervised filtering step, we only screen for strong positive correlations, and we do not consider those features that have strong negative correlations for filtering.

Next, we construct a linear regression model based on the remaining features using lasso, which is a regularized least square regression that penalizes the sum of the absolute value of the feature weights (Tibshirani, 1996). By regularizing the feature weights by the ℓ-1 penalty, lasso tends to obtain a sparse solution (i.e. solution with many zero weights), allowing us to identify irrelevant features. However, because of this ℓ-1 penalty, lasso cannot guarantee a unique optimal solution, and the presence of collinearity in the features in such cases can lead to inconsistent feature selection (Leng *et al.*, 2006; Rajaratnam *et al.*, 2016; Zou, 2006). Thus, to alleviate the chance of inconsistent feature selection, we deliberately apply the aforementioned collinearity-based filtering as a preprocessing of this lasso-based feature selection. In this lasso-based filtering, we first optimize λ_{lasso}, the tuning parameter that controls the amount of regularization. To this end, we perform grid search with leave-one-out cross validation (LOOCV) based on the mean absolute error (MAE) criterion on the training set. After identifying the optimal value of λ_{lasso}, we focus on the LOOCV results of the regression model trained with this hyperparameter choice and measure the weight of the features for each left-out sample. By choosing threshold value θ which represents the cutoff for the number of zero weights, we filter out features whose weights are assigned zero values at least θ times in the LOOCV samples. Note that, while this grid search is able to fine-tune hyperparameters, it cannot guarantee to find the globally optimal hyperparameter combination.

Through this feature selection step, we obtain a function h that maps each compound to a P-dimensional vector of 2D fingerprint-based features. Given the training set and this feature representation of each compound, we seek to learn the value of $\mathbf{w}$ to estimate $\mathbf{y}$ by $\mathbf{DXw}$ where $\mathbf{X}$ is the m-by-P design matrix which is expressed by $\mathbf{X} = \mathbf{SF}$ and $\mathbf{D}$ is the P-by-P diagonal matrix that is used to normalize each column vector of $\mathbf{X}$ by its infinity norm.

Since the size of available thermodynamic quantities is typically small, we often have $p \gg m$ (Flamholz *et al.*, 2012; Goldberg *et al.*, 2004; Noor *et al.*, 2013). In addition, many biochemical reactions can often be represented by linear combinations of other

reactions (i.e. many rows of **S** are often linearly dependent) (Gunawardena, 2003; Kuwahara *et al.*, 2017; Lee *et al.*, 2000; Orth *et al.*, 2010). Even when we have $P < m$ with the filtering of features, the rank of **X** often ends up being smaller than P. In such cases, thus, the ordinary least-square regression becomes ill-posed and results in an infinite number of optimum solutions for **w**. To construct an FC model under such circumstances, thus, we use ridge regression, which is a regularized least-square regression that, by penalizing the amount of the squared weights, obtains a unique global optimum solution as follows:

$$\hat{\mathbf{w}} = (\mathbf{Z}^T\mathbf{Z} + \lambda_{ridge} I)^{-1}\mathbf{Z}^T\mathbf{y}, \tag{5}$$

where $\lambda_{ridge} > 0$ is a tunable parameter that controls the amount of shrinkage and $\mathbf{Z} = \mathbf{DX}$ is the normalized design matrix. By reducing a squared Euclidean norm of the weights, ridge regression can reduce variance, which helps reduce the generalization error (i.e. alleviate the overfitting problem). Because of the inclusion of the penalty term in the objective function, however, one consequence of this is that the resulting linear model introduces a bias. Thus, to find a regularization parameter value for a good compromise which is expected to achieve a low bias with an acceptable variance, we use a cross validation.

3.3 Learning of an FC model

To learn an FC model, we used a dataset that contains experimental measurements of standard reaction Gibbs energies for 697 unique reactions. This dataset was derived from the thermodynamic dataset curated by Noor *et al.* (2013) (see Supplementary Section S1). Since it has 681 chemical compounds, the dataset contains more reactions than the compounds. However, since a number of reactions in this dataset are linearly dependent, the rank of the stoichiometric matrix is 523, making the identification of the unique solution for the standard Gibbs energy of formation of the compounds via the ordinary least square regression impossible.

To represent each compound in this Noor *et al.*-based dataset, we generated 2D fingerprint-based numerical features by gathering 881 binary features from the Pubchem fingerprint scheme, 307 binary features from Open Babel fingerprint (FP4), 166 binary features from MACCS Keys and 190 molecular descriptors implemented in RDKit. In addition to these features, since the 2D structure of 41 chemical compounds in the dataset was not concretely specified—mainly due to the presence of the R group in their structures—we created additional features to accommodate these unknown-structure compounds. In total, thus, we represented each compound in the dataset by using 1585 numerical features (see Supplementary Section S2).

By performing our systematic feature selection procedure on the initial 1585 features, we were able to remove a substantial number of features and retained a small fraction of relevant ones. We first removed 687 features with zero-columns in the initial design matrix. Among these unused features, 639 were not used at all to represent the compounds in the training set. The other 48 features correspond to non-zero column vectors that are in the null space of the stoichiometric matrix. These 48 features were, thus, determined to be canceled out because they were conserved between the reactants and the products of each reaction. Next, by defining the correlation of 0.99 as the threshold value for a high degree of collinearity (i.e. $\rho = 0.99$), the collinearity-based filtering removed 222 features, making the number of remaining features 676.

With these remaining features, we applied the lasso-based feature selection. Grid search to minimize the validation error found $\lambda_{lasso} = 0.1$ to be the optimal value of the regularization parameter (see Supplementary Table S1). By examining the distribution of the sign of the feature weights from the LOOCV samples, we observed highly consistent patterns (Fig. 1). Among the 676 features, 266 consistently had zero weight for all 697 left-out samples, while 414 and 423 had zero weight for $\geq 90\%$ and $\geq 50\%$ of the samples, indicating that irrelevant features were highly consistent in our LOOCV results.

Of the 410 non-zero-weight features, there were 121 features that contributed to the spontaneity of reactions in at least one LOOCV sample. Out of these 121 features, 115 had negative weights in more than 90% of the samples, of which 38 consistently had negative weights in all samples. For example, among those features which had negative weights for all 697 samples, the topological polar surface area feature (Open Babel FP4 90) had the average weight of −233.15, while a feature to test a specific atom neighbor pattern based on hydrogen, carbon and oxygen atoms (Pubchem fingerprint 339) had the average weight of −6.60. Because a reaction with negative Gibbs free energy favors the forward direction, fingerprint features with negative weights in a given compound contribute to attracting the flow of the reaction to produce that compound.

We also found highly consistent patterns in the features with positive weights. Among the 135 features that had positive weights in at least one sample, 122 had positive weights in more than 90% of the samples, of which 47 consistently had positive weights in all LOOCV samples. For example, the Pubchem substructure feature to test the presence of atom pair O–O (Pubchem fingerprint 309) had positive weights for all 697 reactions with the average weight of 1746.58, while another Pubchem feature which checks the presence of simple substructure pattern based on nitrogen and carbon atoms (Pubchem 516) also had positive weight consistently with the average weight of 10.92.

We searched for adequate values of the zero-count threshold θ and the ridge regularization parameter λ_{ridge} through the minimization of the MAE from the LOOCV as the objective in search space with 14 different values for each parameter (Fig. 2). Among these hyperparameter combinations, $\theta = 14$ and $\lambda_{ridge} = 0.0001$ resulted in the lowest LOOCV error. To further optimize the hyperparameter combination, we adjusted the value of λ_{ridge} from 0.0001 with a fine increment, while keeping the value of θ as 14. This fine-tuning allowed us to identify $\theta = 14$ and $\lambda_{ridge} = 0.0006$ as the optimal hyperparameter combination.

In total, with the feature selection procedure, we were able to reduce the number of features from 1585 to 223. With these selected features, the design matrix became skinny and had a dimension of 697 by 223. However, the rank of the design matrix is 218, which is less than the full rank. Thus, we used the aforementioned regularized linear regression method to construct an FC model.

3.4 Effects of feature selection

To understand how our systematic feature selection procedure affects the accuracy performance, we first performed LOOCV for an FC model with original 898 non-zero features for various values of λ_{ridge} (Supplementary Table S2). That is, by comparing the LOOCV results with and without applying the systematic feature selection procedure for the best performing λ_{ridge}, we set out to analyze the effects of the feature selection using various accuracy criteria. We found that the FC model with the final features outperformed the one with the original non-zero features (Table 1). In particular, the results show that the feature selection enabled a 25% improvement in the MAE (from 21.24 to 16.02 kJ/mol). Moreover, the results from other accuracy measures such as Pearson's

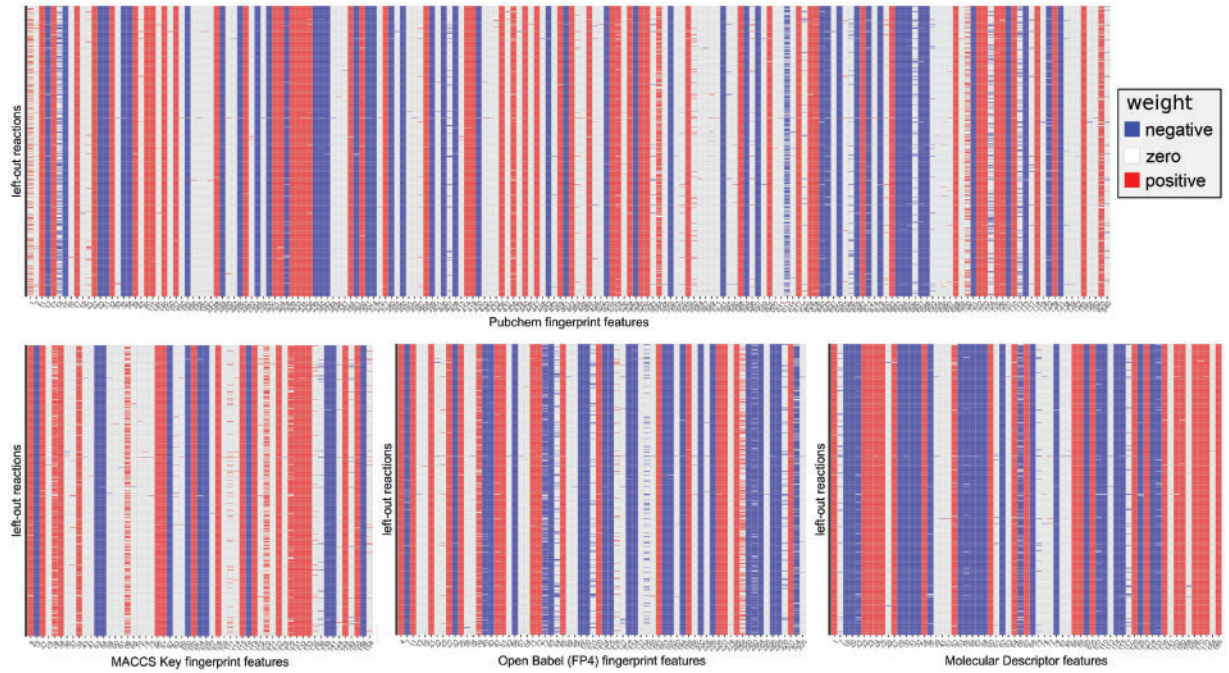

Fig. 1. Heatmaps showing the classification of the weight of selected chemical fingerprint and molecular descriptor features based on the leave-one-out cross-validation analysis of the lasso model

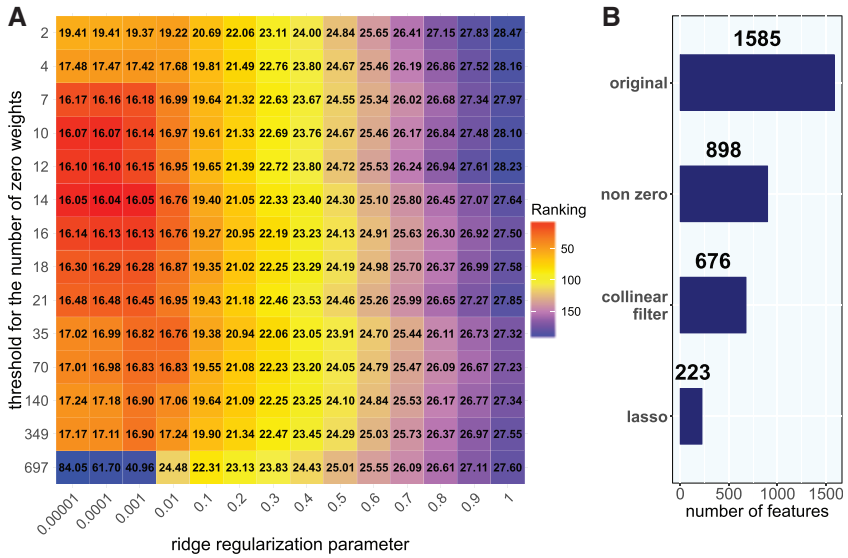

Fig. 2. Hyperparameter based on the leave-one-out cross-validation (LOOCV) results of the ridge regression. (**A**) Grid search of hyperparameters. Each cell shows the mean absolute error (MAE) from the LOOCV results for a specified combination of the zero count threshold θ and the ridge regularization parameter λ_{ridge}. The color of each cell indicates the ranking of its prediction performance based on the mean absolute error. **(B)** The number of selected features for each feature filtering step. With $\theta = 14$, the number of the selected features turns out to be 223. The unit of the LOOCV MAE is kJ/mol

correlation, Spearman's rank correlation and the root mean squared error consistently demonstrated the performance gain achieved by the feature selection. Next, we performed LOOCV for FC models with various subsets of the original features and analyzed the MAEs (Supplementary Table S3). We found that the validation accuracy of these FC models varied significantly depending on combinations of 2D fingerprint features. We also found that the accuracy of the FC model with the final features was higher compared to these models. Our results indicate that a combination of 2D fingerprints strongly affects the prediction accuracy our systematic feature selection procedure is able to determine a small subset of relevant ones from a large pool of 2D fingerprint-based features to increase the prediction accuracy.

3.5 Performance comparison via cross validation

To evaluate the value of the FC method, we compared its prediction performance with that of state-of-the-art methods on the same dataset and performed the same LOOCV. In this comparison, we used a least-square regression method based on Equation 1 that constructs a linear model with representatives of pseudoisomers as its features

Table 1. The effects of feature selection on the prediction performance

FC model	# features	MAE[a]	Pearson[b]	Spearman[c]	RMSE[d]
None-zero features	898	21.24	0.993	0.90	51.06
Final features	223	16.02	0.994	0.95	49.46

Note: Various prediction performance measures from the leave-one-out cross validation (LOOCV) in kJ/mol are computed and are compared between an FC model with the initial none-zero features and an FC model with the final selected features.

[a]The mean absolute error from LOOCV.

[b]The Pearson correlation coefficient.

[c]The Spearman rank correlation coefficient.

[d]The root mean square error.

(Noor *et al.*, 2012) which we refer to as the reactant contribution (RC) method. We also used two versions of the group contribution-based methods: one is a version developed by Noor *et al.* (2012) which, similar to the RC method, has the pseudoisomer-based pre-processing for the compounds, which we call the GC method, and the other is a more recent addition of a GC variant, called the component contribution (CC) method, which is a hybrid of the RC method and the GC method (Noor *et al.*, 2013). Based on the observation that the prediction accuracy of the RC method is higher than the GC method for certain biochemical reactions, the main idea of the CC method is to use the RC method when it is expected to perform well and use the GC method for the other cases (Noor *et al.*, 2013). These methods are all publicly accessible at https://github.com/eladnoor/component-contribution.

From the LOOCV results, we compared predicted values of $\Delta_r G^0$ to the corresponding observed ones and examined the distribution of their absolute error (Fig. 3). We found that the FC model achieved the strongest positive correlation ($r = 0.99$) and the smallest MAE ($\mu = 16.02$ kJ/mol) among the four models (Fig. 3A). The CC and GC models performed similarly in terms of the quality of the linear correlation with the observed data, both resulting in correlation coefficient of $r = 0.95$ (Fig. 3B and C). However, their data showed that, while many of the estimates appeared to be in close agreement with the corresponding observed ones, there was a small subset of the estimates that had noticeably large deviations from the observed values. While these deviations had small effects on the correlation, they might have contributed substantially to their inferior performance in terms of the MAE ($\mu = 32.29$ kJ/mol for the CC model and $\mu = 33.17$ kJ/mol for the GC model). This issue was further pronounced in the RC model, resulting in much lower correlation ($r = 0.66$) and substantially higher prediction error ($\mu = 217.9$ kJ/mol). These LOOCV results, thus, indicate that the predictive performance of the FC method is superior to the other methods, mainly because its prediction error was less sensitive to changes in the training set compared with the other three.

3.6 The effects of linear dependency on the prediction accuracy

Although its overall LOOCV results were poor, the RC method was reported to perform well for the prediction of the Gibbs free energy of certain reactions (Noor *et al.*, 2013). Specifically, these reactions are those whose stoichiometric vectors are in the row space of the stoichiometric matrix **S** for the training set. That is, whether or not the stoichiometric vector of a given reaction is a linear combination

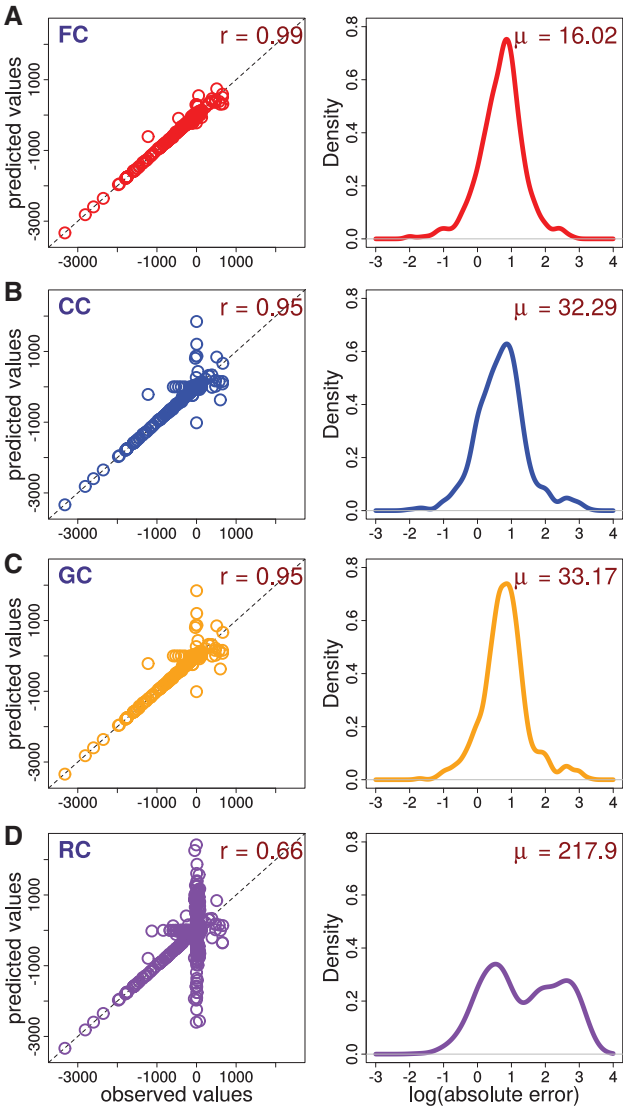

Fig. 3. Comparison of the results from the leave-one-out cross validation (LOOCV). The left pane shows scatter plots in which the observed values for the standard Gibbs energy (x-axis) and the predicted values (y-axis) are compared. The Pearson correlation coefficient (r) between the observed data and the predicted data is shown for each model. The right pane displays the distribution of the absolute error computed for each pair of observed and predicted values. The x-axis uses a base-10 log scale. The mean prediction error (μ) is shown for each model. (**A**) fingerprint-contribution (FC) model, (**B**) component-contribution (CC) model, (**C**) group-contribution (GC) model and (**D**) reaction-contribution (RC) model. The unit of ΔG^0 is kJ/mol

of those for the reactions in the training set was demonstrated to be an important factor for the prediction accuracy of the RC method.

Building on this observation, we analyzed the extent to which the prediction error is influenced by the linear dependency of reaction features in the validation set with respect to the reaction features in the design matrix. To this end, we classified reactions into two groups: in-range reactions and out-of-range reactions. An in-range reaction is a reaction whose feature representation is linearly dependent on those of the reactions in the training set, while an out-of-range reaction is a reaction whose feature representation is linearly independent of those of the reactions in the training set. In other words, reaction R_k is an in-range reaction if $\mathbf{x_k}$ is in the column space of $\mathbf{X}^T$, the transpose of the design matrix and an out-of-range reaction otherwise.

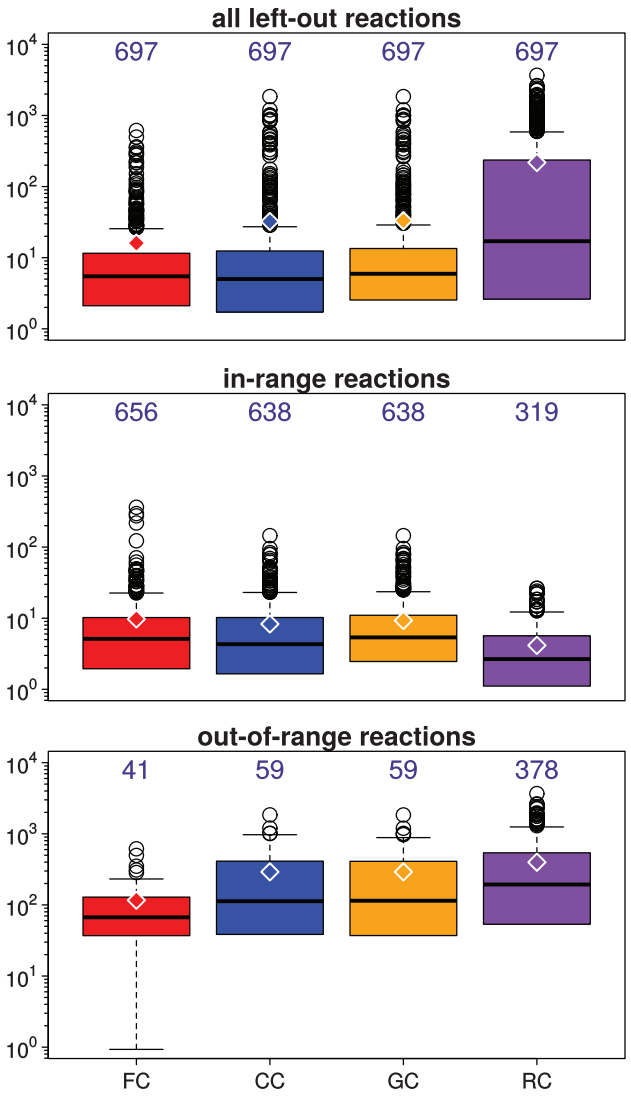

Fig. 4. Boxplots showing how the distribution of the absolute errors is partitioned for in-range reaction samples and out-of-range reaction samples in each model. In each boxplot, the number shown in the upper side of the plot indicates the sample size, while the diamond-shaped point represents the mean absolute error (MAE). Open circle points represent outliers which are defined to be those samples that are outside of the range between the lower quartile minus 1.5 times the interquartile distance (IQD) and the upper quartile plus 1.5 times the IQD. The unit of absolute error is kJ/mol

By measuring the distribution of the absolute error for the two groups, we found that all of the models had large discrepancies in the prediction accuracy between the two groups in the LOOCV results (Fig. 4). We consistently observed substantially higher prediction error in the out-of-range reactions. Specifically, the MAE of the out-of-range reactions was 14.77, 35.16, 31.37 and 95.95 times as high as that of the in-range reactions for FC, CC, GC and RC, respectively. Our results confirm the previous study (Noor *et al.*, 2013) in that, while the RC model produced the highest LOOCV error ($\mu = 217.86$ kJ/mol), it had the lowest prediction error for the in-range reactions ($\mu = 4.15$ kJ/mol). This indicates that the RC model was highly overfit towards the prediction of the standard Gibbs free energy of the in-range reactions since the weights of the chemical compounds participating in those reactions were uniquely determined from the training set. Conversely, we found that the FC method can contain the deviations between the two groups the most. That is, while the MAE of the in-range reactions for the FC model was on par with those for the CC model and the GC model, the FC model was able to produce the lowest overall LOOCV error because its proportion of the out-of-range reactions was the smallest ($n = 41$) and its MAE for the out-of-range reactions was the lowest ($\mu = 116.36$ kJ/mol) among the four models. These suggest that the FC model was generalized the most among the four models to deal with unseen biochemical reactions.

3.7 Performance analysis using the KEGG dataset

To further examine the generalization property, we analyzed the performance of the four methods on the KEGG REACTION dataset which contains a wide range of biochemical reactions. By inspecting 10 668 reactions, we decided to use 7929 that were deemed to be valid for this analysis (see Supplementary Section S3). However, since the KEGG dataset does not contain experimentally observed thermodynamic data, we first needed to determine how to evaluate the prediction performance in order to use the KEGG dataset for analysis of the prediction accuracy.

Since we found that the prediction error of in-range reactions were much lower than that of out-of-range reactions (Fig. 4), we considered an evaluation approach that estimates the prediction accuracy based on the fractions of in-range reactions and out-of-range reactions in a testing set. To understand whether this approach is sound, however, we first analyzed the statistical significance of the relation between this linear dependency-based grouping and the LOOCV prediction error (see Section 2). Table 2 shows the results from our analysis on various models, which consistently indicates that it is highly unlikely to find a partition of the LOOCV prediction errors into two subgroups by chance to produce the MAE as extreme as the one observed in the in-range reactions and the out-of-range reactions ($P < 10^{-6}$). This shows statistically significant evidence that in-range reactions are expected to have low prediction error, while out-of-range reactions are expected to have high prediction error.

With these results in hand, we proceeded to estimate the prediction error on the KEGG dataset by a weighted average approach based on the partition of the in-range and the out-of-range reactions (see Section 2). To this end, we first generated the distribution of the in-range reactions, the out-of-range reactions and reactions outside of the prediction coverage (i.e. 'not-covered' reactions) for each model (Fig. 5A). Of 7929 valid KEGG reactions, we found that all of the reactions are in-range reactions in the FC model. Both the GC model and the CC model had the same distribution, and they had 5950 in-range reactions, 1005 out-of-range reactions and 974 not-covered reactions. Among the four models, the RC model had the highest proportion of not-covered reactions and the lowest proportion of the in-range reactions. It had only 803 in-range reactions, while it had 200 out-of-range reactions and 6926 not-covered reactions. Because of this substantially limited prediction coverage, we excluded the RC model from the performance analysis.

By using these distributions, we computed the accuracy estimate of the three models (Fig. 5B). Since all of the valid KEGG reactions are in-range reactions, the MAE estimate for the FC model was exactly the same as the MAE of the in-range reactions from the LOOCV results, which is 9.75 kJ/mol. On the other hand, since about 15% are out-of-range reactions in both the CC model and the GC model among the 6595 covered reactions, we estimated the MAE for the CC model and the GC model to be 49.26 and 50.05 kJ/mol, respectively. Thus, our performance analysis indicates that the FC model would outperform the other three models for the

Table 2. Relation between the magnitude of the mean absolute error (MAE) in kJ/mol and the linear dependency-based subgrouping in the leave-one-out cross-validation results

	All samples	In-range		Out-of-range		Permutation test	
Model	MAE	Size	MAE	Size	MAE	Samples	*P*-value[a]
FC	16.02	656	9.75	41	116.36	10^6	$<10^{-6}$
FC-50[b]	16.90	649	8.97	48	124.16	10^6	$<10^{-6}$
FC-orig[c]	21.24	494	9.16	203	50.64	10^6	$<10^{-6}$
CC	32.29	638	8.30	59	291.80	10^6	$<10^{-6}$
GC	33.17	638	9.29	59	291.39	10^6	$<10^{-6}$
RC	217.86	319	4.15	378	398.21	10^6	$<10^{-6}$

Note: Statistical significance was measured by computing the *P*-value of MAE based on the partitions for the in-range reaction subset and the out-of-range reaction subset for various models.

[a]Each *P*-value was computed as the probability that the MAE of a randomly selected in-range-reaction-size reaction set is lower than or equal to the observed MAE for the in-range reactions in the sampling distribution.

[b]FC model generated based on the lasso-based feature selection with the zero-count threshold being 349 (i.e. 50% of 697).

[c]FC model based on the original non-zero features.

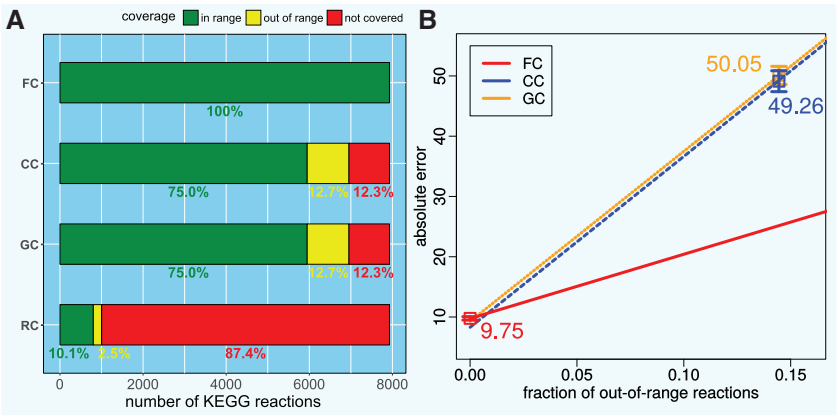

Fig. 5. Estimation of the prediction accuracy for the KEGG reactions by each model trained by the Noor *et al.*-based dataset. (**A**) A bar graph showing the proportion of valid KEGG reactions that are partitioned into the three coverage-based groups: 'in range,' 'out of range' and 'not covered.' In the KEGG dataset, there are 7929 valid reactions, each of which is chemically balanced and reacts chemical compounds whose 2D structures are specified. Here, 'in range' means reactions whose feature vectors are linear combinations of the feature vectors in the training set, 'out of range' means a group of reactions whose feature vectors are not linear combinations of the training set, and 'not covered' means a subset of 'out of range' reactions that cannot be represented by the features in a given model. (**B**) Estimation of prediction accuracy based on the weighted average of the prediction errors for in-range reaction group and out-of-range reaction group from the leave-one-out cross-validation (LOOCV) results for the three models with a higher reaction coverage. Square points indicate the sample mean of absolute error of the three models for the KEGG reaction set based on this weighted average approach with 100 reaction sets, while error bars represent their sample standard deviation. For each model, the reaction set is sampled from the LOOCV results to have the same proportion of in-range and out-of-range reactions as the KEGG reactions. The unit of the absolute error is kJ/mol

prediction of the standard Gibbs free energy for the KEGG reactions, largely because of its proportion for the in-range reactions. Furthermore, this analysis demonstrated that the prediction coverage of systematically selected 2D fingerprint-based features used in the FC method can be substantially higher than those features used in the other three methods.

4 Discussion

In summary, we have developed a statistical method called fingerprint-contribution (FC) which, by systematically selecting relevant 2D fingerprint-based features, constructs a regularized linear model for the prediction of the Gibbs free energy of biochemical reactions. By representing each chemical compound by features based on 2D fingerprints and molecular descriptors, the FC method can predict the Gibbs free energy of reaction in a manner that is consistent with the first law of thermodynamics, and its prediction can cover virtually any biochemical reactions in which compounds with concrete 2D structures participate. At the same time, the systematic feature filtering procedure allows for a convenient way to select a small set of relevant 2D fingerprint-based features to improve the quality of prediction accuracy.

With the ability to represent the 2D structure of each molecule in a high-dimensional feature space, 2D fingerprints have been widely used as a means to quantify the similarity of molecules (Willett *et al.*, 1998). In the SAR analysis, such structural similarity coefficients have been successfully applied, for example, to ligand-based virtual screening to reduce the search space for the experimental evaluation for the identification of novel hits (Cereto-Massagué *et al.*, 2015; Eckert and Bajorath, 2007; Lavecchia, 2015; Ripphausen *et al.*, 2011; Willett, 2006). The idea of 2D fingerprint-based similarity has also been applied to the prediction of $\Delta_r G^0$ before. Indeed, IGERS measures reaction similarity via Tanimoto coefficient and infers $\Delta_r G^0$ of a reaction by that of the most similar one

from a set of predefined reference reactions based on manually picked chemical attributes (Rother *et al.*, 2010). However, since this method does not consider ΔG^0 at the compound level, its prediction may lead to the violation of the first law of thermodynamics, which can result in modeling of metabolic systems with severely inconsistent thermodynamic parameters. Furthermore, since this reaction similarity-based approach can only predict $\Delta_r G^0$ of reactions that are sufficiently similar to those in the reference set, its prediction coverage can be greatly limited (Noor *et al.*, 2012; Rother *et al.*, 2010).

Here, we have demonstrated the value of the FC method over the state-of-the-art methods in terms of the prediction accuracy and coverage. By classifying reactions into the in-range reactions and the out-of-range reactions, we found that the superior accuracy achieved by the FC method in the LOOCV results was due to the reduction in the prediction error and the size for the out-of-range reactions. Because the FC method had the smallest accuracy difference between the in-range reactions and the out-of-range reactions, our results suggest that the FC model has the best generalization quality to deal with the Gibbs energy prediction of unseen biochemical reactions. Indeed, our results from the performance analysis on the KEGG reactions supported this and showed further evidence that the FC method performs well on a wide range of biochemical reactions in terms of prediction accuracy and coverage. Since all of the prediction methods examined here were linear regression-based methods, our study also points to the value of 2D fingerprint-based features on the prediction of reaction Gibbs free energies. In addition, we have demonstrated that the systematic feature filtering procedure improved the prediction accuracy of an FC model by selecting a small number of relevant features. Taken together, this study suggests the effectiveness of the use of 2D fingerprints and molecular descriptors on the biochemical thermodynamic prediction and highlights that a systematic filtering procedure allows for a convenient way to select most relevant ones which provide useful information to quantify the Gibbs free energy.

Our future work includes the development of a fingerprinting method to generate more suitable features and a nonlinear modeling approach to achieve higher prediction accuracy for the Gibbs free energy prediction problem. To that end, we have already performed several computational experiments. For example, we have analyzed the performance of FC models with a different type of chemical fingerprints, which were generated via a neural network (see Supplementary Section S5). While we found that these existing neural network features did not perform as well as expected, we saw a potential to such learning method to customize chemical fingerprint features for the Gibbs energy prediction problem. In addition, we have studied the possibility of using neural network models for this prediction problem (see Supplementary Section S6). Although our initial neural network model was not very useful as it violated the energy conservation principle, its validation accuracy was found to be reasonable, which was higher than that of the GC and CC models. Thus, to develop a high-performing nonlinear modeling approach to the Gibbs energy prediction problem, we plan to study how neural network models can be customized to meet the energy conservation constraints and to increase the prediction accuracy.

Funding

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Awards No. FCC/1/1976-23, FCC/1/1976-26, URF/1/2602-01, URF/1/3007-01, and URF/1/3450-01.

Conflict of Interest: none declared.

References

Ataman,M. and Hatzimanikatis,V. (2015) Heading in the right direction: thermodynamics-based network analysis and pathway engineering. *Curr. Opin. Biotechnol.*, **36**, 176–182.

Beard,D.A. *et al.* (2004) Thermodynamic constraints for biochemical networks. *J. Theor. Biol.*, **228**, 327–333.

Carbonell,P. *et al.* (2014) XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.*, **42**, W389–W394.

Cereto-Massagué,A. *et al.* (2015) Molecular fingerprint similarity search in virtual screening. *Methods*, **71**, 58–63.

Eckert,H. and Bajorath,J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, **12**, 225–233.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 1–18.

Flamholz,A. *et al.* (2012) eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res.*, **40**, D770–D775.

Goldberg,R.N. *et al.* (2004) Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics*, **20**, 2874–2877.

Großkopf,T. and Soyer,O.S. (2016) Microbial diversity arising from thermodynamic constraints. *ISME J.*, **10**, 2725–2733.

Gunawardena,J. (2003) Chemical reaction network theory for *in-silico* biologists. http://vcp.med.harvard.edu/papers/crnt.pdf.

Held,C. and Sadowski,G. (2016) Thermodynamics of bioreactions. *Annu. Rev. Chem. Biomol. Eng.*, **7**, 395–414.

Henry,C.S. *et al.* (2006) Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys. J.*, **90**, 1453–1461.

Jankowski,M.D. *et al.* (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.*, **95**, 1487–1499.

Jinich,A. *et al.* (2014) Quantum chemical approach to estimating the thermodynamics of metabolic reactions. *Sci. Rep.*, **4**, 7022.

Kümmel,A. *et al.* (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.*, **2**, 2006.0034.

Kuwahara,H. *et al.* (2016) MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. *Nucleic Acids Res.*, **44**, W217–W225.

Kuwahara,H. *et al.* (2017) ACRE: absolute concentration robustness exploration in module-based combinatorial networks. *Synth. Biol.*, **2**, ysx001.

Lavecchia,A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today*, **20**, 318–331.

Lee,J.W. *et al.* (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.*, **8**, 536.

Lee,S. *et al.* (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng.*, **24**, 711–716.

Leng,C. *et al.* (2006) A note on the lasso and related procedures in model selection. *Stat. Sin.*, **16**, 1273–1284.

Mavrovouniotis,M.L. *et al.* (1988) A group contribution method for the estimation of equilibrium constants for biochemical reactions. *Biotechnol. Tech.*, **2**, 23–28.

Nielsen,J. (1998) Metabolic engineering: techniques for analysis of targets for genetic manipulations. *Biotechnol. Bioeng.*, **58**, 125–132.

Noor,E. *et al.* (2012) An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics*, **28**, 2037–2044.

Noor,E. *et al.* (2013) Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.*, **9**, 1003098.

Orth,J.D. *et al.* (2010) What is flux balance analysis? *Nat. Biotechnol.*, **28**, 245.

Rajaratnam,B. *et al.* (2016) Lasso regression: estimation and shrinkage via the limit of Gibbs sampling. *J. R. Stat. Soc. Ser. B*, **78**, 153–174.

Ripphausen,P. *et al.* (2011) State-of-the-art in ligand-based virtual screening. *Drug Discov. Today*, **16**, 372–376.

Rother,K. *et al*. (2010) IGERS: inferring Gibbs energy changes of biochemical reactions from reaction similarities. *Biophys. J*., **98**, 2478–2486.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Toure,O. and Dussap,C.-G. (2016) Determination of Gibbs energies of formation in aqueous solution using chemical engineering tools. *Bioresour Technol*., **213**, 359–368.

Willett,P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.

Willett,P. *et al*. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci*., **38**, 983–996.

Yim,H. *et al*. (2011) Metabolic engineering of *Escherichia coli* for direct production of 1, 4-butanediol. *Nat. Chem. Biol*., **7**, 445.

Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc*., **101**, 1418–1429.