Supplementary Material for
Systematic selection of chemical fingerprint features improves the Gibbs energy prediction of biochemical reactions
Meshari Alazmi\textsuperscript{a}, Hiroyuki Kuwahara\textsuperscript{a,}\textsuperscript{*}, Othman Soufan, Lizhong Ding, and Xin Gao\textsuperscript{*}
King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Thuwal, 23955, Saudi Arabia
\textsuperscript{a}These authors contributed equally to this work.
\textsuperscript{*}Corresponding authors (email: hiro.kuwahara@kaust.edu.sa and xin.gao@kaust.edu.sa)
Contents

S1 Preprocessing of the Noor et al.-based dataset 3

S2 Mapping of compounds to their 2D fingerprints and molecular descriptors 3

S3 Preprocessing of the KEGG REACTION dataset 4

S4 Model performance using only neural fingerprint 4

S5 The contribution of the neural fingerprint 5
  S5.1 Lasso performance after removing collinear features 5
  S5.2 Ridge performance on the selected features 5

S6 Neural network performance 5

List of Tables 7
S1 Preprocessing of the Noor et al.-based dataset

The dataset we used contains experimental measurements of standard reaction Gibbs energies, $\Delta_r G^0$, for 697 unique reactions. To derive this dataset, we followed Noor et al. [1] and utilized their datasets (https://github.com/eladnoor/component-contribution) which consist of 4544 reactions from TECRDB [2], 13 redox reactions, and formation reactions of 224 compounds [3, 4]. The composition of the formation reactions of compounds was set based on the natural state of 13 elements: H, C, N, O, F, P, S, Mg, Cl, Mn, Fe, Br, and I.

The thermodynamic data were first mapped to the corresponding standard transformed Gibbs energies based on their specific biological conditions. For example, apparent equilibrium constant, $K'$, was mapped to the corresponding standard transformed Gibbs energy of reaction, $\Delta_r G'^0$, using the relation

$$K' = -RT \ln \left( \Delta_r G'^0 \right).$$  \hfill (S1)

To convert $\Delta_r G'^0$ into the corresponding $\Delta_r G^0$, we applied a script based on the Legendre transformation implemented by Noor et al. [5]. In this step, thermodynamics of pseudoisomers was taken into consideration, and each group of isomers was replaced by a representative “chemical compound.”

Many reactions from TECRDB had duplicates. For each duplicate reaction, we computed the median of $\Delta_r G^0$ and used it as the observed value. With this, we generated a dataset of $\Delta_r G^0$ for 697 unique reactions. In these 697 reactions, there were 681 unique chemical compounds that participate. We manually inspected these 681 compounds to see whether or not each of them has a concrete 2D structure by first identifying the InChI format of its structure. Those with either InChI or SMILES representations are further examined to see if their structures are concrete or if they are generic. With these inspections, 640 were determined to have concrete 2D structures, while the other 41 compounds were found not to possess concrete structures.

S2 Mapping of compounds to their 2D fingerprints and molecular descriptors

In this study, we used the PubChem fingerprint scheme, the Open Babel FP4 scheme, MACCS keys, and RDKit molecular descriptors to represent compounds with known 2D structures. This mapping was implemented in Python. To map a given compound to Pubchem fingerprints, we first checked if this compound had a Pubchem ID. If it does, then we used the pubchempy package to generate the PubChem fingerprints. If not, then we obtained the SMILES representation of its 2D structure and used it in the PubChem fingerprint API in CDK (https://cdk.github.io/) for this task. To generate the Open Babel
FP4 representation, we used the pybel package. For MACCS keys and RDKit molecular descriptors, we used the corresponding APIs in RDKit (see http://www.rdkit.org/docs/).

S3 Preprocessing of the KEGG REACTION dataset

The version of the KEGG REACTION dataset that we used consists of 10668 biochemical reactions. Of these KEGG reactions, 438 were either polymerization reactions or reactions that contain Glycans, and we excluded them for our analysis. In the remaining 10230 reactions, generic reactions with variable stoichiometric coefficients were changed to corresponding concrete reactions by replacing variable “n” by 2. Then, we inspected these reactions for the participation of compounds with unknown or generic 2D structures and for chemical imbalance. We found that there were 2047 biochemical reactions which used compounds with unknown or generic 2D structures. Among these 2047 reactions, however, 95 were found to be represented by the use of the 41 compound-specific features from the Noor et al.-based dataset. Thus, we excluded 1952 reactions in the KEGG for the participation of compounds with unknown or generic 2D structures.

Next, we inspected the remaining KEGG reactions for chemical imbalance. Before doing this, however, we attempted to balance electronic charges in biochemical reactions as many as possible. To this end, we followed the approach taken by Equilibrator [3]. For example, to correct the charge imbalance of a half-reaction, NAD and NADH were added to that reaction. With this, we found that 349 reactions were not chemically balanced.

Through the use of these inspections, thus, we concluded that 2301 out of the 10230 reactions were deemed to be invalid for our purpose. As a result, we used 7929 biochemical reactions from the KEGG REACTION dataset for the performance analysis.

S4 Model performance using only neural fingerprint

We extracted the neural fingerprint [7] of the compounds using the publicly available tool. The neural fingerprint schema has a fixed length of 50 features which can be used to represent such a compound. Our goal is to check if the neural fingerprint could contribute for a higher prediction accuracy of the standard Gibbs free energy. We performed LOOCV experiments using only the neural fingerprint. The first row in table S4 shows the results of ridge regression for the best $\lambda_{ridge}$ value without feature selection.

The other experiment is that we removed the features that are highly collinear ($\rho = 0.99$). This feature selection procedure removed 45 features out of 50 features from the neural fingerprint. The total number of the remaining features from the neural fingerprint is only five features. Then, we performed LOOCV using lasso with different $\lambda_{lasso}$ values. The second row in table S4 shows the best $\lambda_{lasso}$ results after removing the collinear features.
The last experiment, we analyzed the weights (coefficients) of the best lasso results by checking the total number of zero weights of each feature across all the LOOCV models. Lasso weighted three neural fingerprint features to have zero coefficients in all the LOOCV models. So, the remaining two features have been used to perform the final ridge model. The third row in table S4 shows the ridge regression results of the LOOCV for the best $\lambda_{ridge}$ using the final selected features.

S5 The contribution of the neural fingerprint

S5.1 Lasso performance after removing collinear features

We concatenated the previously used features with the neural fingerprint. Then, we removed the features that are highly collinear ($\rho = 0.99$) as we have done previously. This also removed 45 features out of 50 features from the neural fingerprint. Then, we performed LOOCV using lasso with different $\lambda_{lasso}$ values. Table S5 shows the best $\lambda_{lasso}$ performance after removing the collinear features when including or excluding neural fingerprint.

S5.2 Ridge performance on the selected features

We analyzed the weights of the best LOOCV lasso results (when $\lambda_{lasso} = 0.1$) by checking the total number of zero weights of a feature across the LOOCV models. The five features from the neural fingerprint have zero weights in all the models except one feature has non-zero weight in one LOOCV model. We performed ridge models using different values of the threshold $\theta$ which is a threshold for the number of zero weights of each feature across all lasso LOOCV models. All the features from the neural fingerprint did not go through the above mentioned feature selection procedure. That is why we observe similar performance in table S6.

S6 Neural network performance

In this section, we performed 5-fold cross validation (CV) to compare the performance of the state-of-the-art methods and FC method when using the previous approach (named here as ridge) and neural network (NN) method. The assumption of using NN is that non-linear models could have better performance over the linear models in our problem if we violate the first law of thermodynamics. For example, by having a reaction going from $A \rightarrow B \rightarrow C \rightarrow A$, the sum of the $\Delta G$ of these three reactions must be zero. We fine-tuned NN parameters such as the hidden layers, number of nodes in each layer, and the learning rate. We performed experiments using one hidden layer and two hidden layers, the number of nodes in each hidden layer are as follow: 1, 2, 5, 10, 20, 50, 100,
200, 300, and 400 nodes, and the learning rate is as follow: 0.0001, 0.001, 0.01, 0.1 and 1. When we used NN, we used all the features except the zero features are removed. We performed the same splits for the 5 folds. Table S7 shows the performance comparison between the NN model (the best results after fine-tuning) and the other methods. The NN model performs worse than ridge regression because NN model violates the first law of thermodynamics.
Tables

Table S1. Validation results from various $\lambda_{\text{lasso}}$ values for the lasso-based feature filtering. The row in bold face represents the optimal value.

| $\lambda_{\text{lasso}}$ | MAE     |
|--------------------------|---------|
| 0.001                    | 56.1153 |
| 0.01                     | 34.6777 |
| **0.1**                  | **20.5926** |
| 0.2                      | 21.4090 |
| 0.3                      | 22.0536 |
| 0.4                      | 22.2569 |
| 0.5                      | 22.2711 |
| 0.6                      | 22.5928 |
| 0.7                      | 22.8364 |
| 0.8                      | 23.2867 |
| 0.9                      | 23.2537 |
| 1                        | 23.5653 |
| 2                        | 25.5668 |
| 5                        | 30.0020 |
| 10                       | 35.5453 |
| 100                      | 71.9806 |
Table S2. LOOCV performance of FC models with the original non-zero features generated from different $\lambda_{\text{ridge}}$. The bold values indicate optimal.

| $\lambda_{\text{ridge}}$ | MAE       | Pearson$^a$ | Spearman$^b$ | RMSE$^c$ |
|---------------------------|-----------|-------------|--------------|----------|
| 1e-05                     | 116.4890  | 0.8461      | 0.6810       | 284.6930 |
| 0.0001                    | 76.6582   | 0.9314      | 0.7341       | 175.8954 |
| 0.001                     | 51.0992   | 0.9675      | 0.8103       | 117.2637 |
| 0.01                      | 31.8098   | 0.9876      | 0.8536       | 71.1071  |
| 0.1                       | 21.5397   | 0.9935      | 0.9014       | 51.1954  |
| 0.2                       | **21.2428** | **0.9936** | **0.9040** | **51.0575** |
| 0.3                       | 21.3383   | 0.9935      | **0.9057**   | 51.4272  |
| 0.4                       | 21.5696   | 0.9934      | 0.9055       | 51.8775  |
| 0.5                       | 21.7999   | 0.9932      | 0.9050       | 52.3403  |
| 0.6                       | 22.0299   | 0.9931      | 0.9042       | 52.7956  |
| 0.7                       | 22.2643   | 0.9930      | 0.9040       | 53.2357  |
| 0.8                       | 22.4855   | 0.9929      | 0.9033       | 53.6580  |
| 0.9                       | 22.7028   | 0.9928      | 0.9031       | 54.0617  |
| 1                         | 22.9177   | 0.9927      | 0.9024       | 54.4463  |
| 2                         | 24.9257   | 0.9919      | 0.8941       | 57.4856  |
| 5                         | 29.1796   | 0.9903      | 0.8739       | 62.6409  |
| 10                        | 34.0759   | 0.9887      | 0.8525       | 67.9050  |
| 100                       | 68.2748   | 0.9703      | 0.7152       | 116.1095 |

$^a$Pearson's correlation coefficient.

$^b$Spearman's rank correlation.

$^c$Root-mean-square error.
Table S3. Validation accuracy of various combinations of 2D fingerprint schemes. The reported MAE for each combination is from the LOOCV results of the optimal $\lambda_{ridge}$ via grid search. Each combination contains the 41 compound-specific features.

| Combination          | $\lambda_{ridge}$ | # of features | MAE    |
|----------------------|---------------------|---------------|--------|
| Pubchem$^a$          | 0.1                 | 922           | 39.1266|
| FP4$^b$              | 1                   | 348           | 147.3989|
| MACCS$^c$            | 0.01                | 207           | 36.1262|
| MD$^d$               | 0.1                 | 231           | 21.8788|
| Pubchem+FP4          | 0.2                 | 1229          | 38.3740|
| Pubchem+MACCS        | 0.3                 | 1088          | 38.0389|
| Pubchem+MD           | 0.2                 | 1112          | 21.2650|
| FP4+MACCS            | 0.01                | 514           | 32.1653|
| FP4+MD               | 0.1                 | 538           | 19.3918|
| MACCS+MD             | 0.1                 | 397           | 20.5514|
| Pubchem+FP4+MACCS    | 0.3                 | 1395          | 37.1850|
| Pubchem+FP4+MD       | 0.2                 | 1419          | 21.3405|
| Pubchem+MACCS+MD     | 0.2                 | 1278          | 21.7030|
| FP4+MACCSS+MD        | 0.1                 | 704           | 18.2714|
| Pubchem+FP4+MACCSS+MD| 0.2                | 1585          | 21.2428|

$^a$Pubchem fingerprint scheme.
$^b$Open Babel FP4 scheme.
$^c$MACCS keys.
$^d$2D molecular descriptors from RDKit.

Table S4. LOOCV performance using only the neural fingerprint

| Method             | Best $\lambda$ | MAE      | Pearson | Spearman | RMSE    |
|--------------------|----------------|----------|---------|----------|---------|
| Ridge$_{wo\_fs}^a$| 10             | 183.2040 | 0.4158  | 0.2554   | 413.9191|
| Lasso$_{rm\_co}^b$| 100            | 194.6206 | 0.3650  | 0.2472   | 435.0790|
| Ridge$_{w\_fs}^c$ | 10             | 183.2040 | 0.4158  | 0.2554   | 413.9191|

$^a$Ridge regression without feature selection.
$^b$Lasso after removing the highly collinear features.
$^c$Ridge regression on the final selected features.
Table S5. LOOCV performance using lasso after removing the collinear features when including or excluding neural fingerprint

| Features        | Best λ | MAE    | Pearson | Spearman | RMSE   |
|-----------------|--------|--------|---------|----------|--------|
| NF_included a   | 0.1    | **20.9204** | 0.9853  | 0.9229   | 77.4673|
| NF_excluded b   | 0.1    | **20.5926** | 0.9855  | 0.9262   | 76.8857|

a Neural fingerprint (NF) features are included.
b Neural fingerprint (NF) features are excluded.

Table S6. Performance comparison between the final selected features from the previous lasso models when including or excluding neural fingerprint features

| Features        | Best θ | Best λ | MAE    | Pearson | Spearman | RMSE   |
|-----------------|--------|--------|--------|---------|----------|--------|
| NF_included a   | 12     | 0.0001 | **16.0820** | 0.9938  | 0.9447   | 50.3123|
| NF_excluded b   | 14     | 0.0001 | **16.0358** | 0.9939  | 0.9447   | 49.6609|

a Neural fingerprint (NF) features are included in the lasso models.
b Neural fingerprint (NF) features are excluded from the lasso models.

The difference in the performance is not significant because none of the NF features were selected for the final features.

Table S7. 5-fold CV performance comparison for different methods.

| Method   | MAE    | Pearson | Spearman | RMSE   |
|----------|--------|---------|----------|--------|
| FC_NN a  | 28.7278| 0.9855  | 0.8653   | 64.9511|
| FC_ridge | 18.0750| 0.9917  | 0.9293   | 48.3777|
| RC       | 243.0060| 0.6150  | 0.5389   | 510.4016|
| GC       | 33.6459| 0.9561  | 0.8924   | 127.7596|
| CC       | 33.0885| 0.9559  | 0.8924   | 128.0255|

a This is the best performance achieved where the NN model has two hidden layers, the first hidden layer has 400 nodes and the second hidden layer has 300 nodes and the learning rate is 0.001.

References

1. Noor E, Haraldsdóttir HS, Milo R, Fleming RM (2013) Consistent estimation of gibbs energy using component contributions. PLoS Comput Biol 9: 1003098.
2. Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. Bioinformatics 20: 2874–2877.

3. Thauer RK, Jungermann K, Decker K (1977) Energy conservation in chemotrophic anaerobic bacteria. Bacteriol Rev 41: 100–180.

4. Alberty RA (2006) Biochemical thermodynamics: applications of Mathematica, volume 48. John Wiley & Sons.

5. 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 28: 2037–2044.

6. Flamholz A, Noor E, Bar-Even A, Milo R (2012) eQuilibrator—the biochemical thermodynamics calculator. Nucleic Acids Res 40: D770–D775.

7. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, et al. (2015) Convolutional networks on graphs for learning molecular fingerprints. In: Advances in Neural Information Processing Systems. pp. 2224–2232.