Supporting Information for:

Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates

Yunsie Chung† and William H. Green*,†

†Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, U.S.A

E-mail: whgreen@mit.edu
Figure S1: Distributions of the pre-training data set for the reactions in all solvents (top), in water (middle), and in n-hexane (bottom). The absolute mean (abs. mean), mean and standard deviation (std.) are in kcal/mol. Water and n-hexane are chosen as sample solvents for data visualization.
Details on the fine-tuning data set

(a) Bimolecular hydrogen abstraction

(b) Unimolecular hydrogen migration

(c) Radical addition to a multiple bond

Figure S2: Templates of the reaction types found in the fine-tuning set.

A subset of the fine-tuning data set was selected for the model training. Around 25 solvents were sampled per each reaction with following procedure:

- Water, acetic acid, and acetonitrile are always included.
- Randomly sample 17 solvents from the following list of 44 more common solvents: carbon tetrachloride, benzene, tert-Butanol, hexane, anisole, T-amyl alcohol, ethyl acetate, propionitrile, chlorobenzene, methylformate, dioxane, tetrahydrofuran, tetrachloroethylene, bromobenzene, benzonitrile, cyclohexane, diethyl ether, methanol, 2-propanol, dichloromethane, chloroform, sulfolane, DMF, ethanol, 2,2,2-trifluoroethanol, acetone, toluene, isopropyl ether, veratrole, nitrobenzene, nitromethane, triethylene glycol, N-benzylacetamide, formamide, ethylene glycol, 2-methoxyethanol, benzyl alcohol, 2-phenoxyethanol, N-methylformamide, dimethyl sulfoxide, glycerol, isoctane, 3-methyl-3-pentanol, and 1-butanol.
- Sample 5 solvents from the remaining solvents.
Figure S3: Distributions of the fine-tuning data set for the reactions in all solvents (top), in water (middle), and in n-hexane (bottom). The absolute mean (abs. mean), mean and standard deviation (std.) are in kcal/mol. Water and n-hexane are chosen as sample solvents for data visualization.
## S3 Reactions in the experimental set

Table S1: List of reactions found in the experimental set.

<table>
<thead>
<tr>
<th>Index</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="beta-scission" alt="reaction" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="beta-scission" alt="reaction" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="H-abstraction" alt="reaction" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="H-abstraction" alt="reaction" /></td>
</tr>
<tr>
<td>5</td>
<td><img src="H-abstraction" alt="reaction" /></td>
</tr>
<tr>
<td>6</td>
<td><img src="H-abstraction" alt="reaction" /></td>
</tr>
<tr>
<td>7</td>
<td><img src="H-abstraction" alt="reaction" /></td>
</tr>
<tr>
<td>Page</td>
<td>Chemical Reaction</td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
</tr>
<tr>
<td>8</td>
<td>$\text{Diels–Alder}$</td>
</tr>
<tr>
<td>9</td>
<td>$\text{Diels–Alder}$</td>
</tr>
<tr>
<td>10</td>
<td>$\text{Diels–Alder}$</td>
</tr>
<tr>
<td>11</td>
<td>$\text{[2+2] Cycloaddition}$</td>
</tr>
<tr>
<td>12</td>
<td>$\text{Nucleophilic addition to carbon-carbon triple bond}$</td>
</tr>
<tr>
<td>13</td>
<td>$\text{Epoxidation of cyclohexene with peroxycarboxylic acid}$</td>
</tr>
<tr>
<td>14</td>
<td>$\text{Thermal decomposition of 3-sulfolene}$</td>
</tr>
</tbody>
</table>
The rates constants of Reaction 6 are reported relative to the rate constants of Reaction 1. The first step was used for the ML predictions for these reactions as the first step was previously identified as a rate limiting step.
S4 Details on the ML model

Table S2: Default atom and bond features used for the model.

<table>
<thead>
<tr>
<th>Atom features</th>
<th>Type</th>
<th>Bond features</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic number</td>
<td>One-hot</td>
<td>Bond type</td>
<td>One-hot</td>
</tr>
<tr>
<td>Number of bonds</td>
<td>One-hot</td>
<td>Conjugation</td>
<td>One-hot</td>
</tr>
<tr>
<td>Formal charge</td>
<td>One-hot</td>
<td>In ring</td>
<td>One-hot</td>
</tr>
<tr>
<td>Ring size</td>
<td>One-hot</td>
<td>Stereo</td>
<td>One-hot</td>
</tr>
<tr>
<td>Connected hydrogen atoms</td>
<td>One-hot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hybridization</td>
<td>One-hot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aromaticity</td>
<td>One-hot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molar mass</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S3: Optimized hyperparameters.

<table>
<thead>
<tr>
<th>Type</th>
<th>Optimized value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-MPNN hidden size</td>
<td>Reaction: 900, Solvent: 200</td>
</tr>
<tr>
<td>D-MPNN depth</td>
<td>Reaction: 4, Solvent: 2</td>
</tr>
<tr>
<td>D-MPNN bias</td>
<td>No</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Normalized</td>
</tr>
<tr>
<td>FNN hidden size</td>
<td>400</td>
</tr>
<tr>
<td>FNN depth</td>
<td>6</td>
</tr>
<tr>
<td>FNN bias</td>
<td>Yes</td>
</tr>
<tr>
<td>Activation function</td>
<td>ReLu</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.0</td>
</tr>
<tr>
<td>Batch size</td>
<td>30</td>
</tr>
<tr>
<td>Warm-up epochs</td>
<td>4</td>
</tr>
<tr>
<td>Max epochs for pre-training</td>
<td>80</td>
</tr>
<tr>
<td>Max epochs for fine-tuning</td>
<td>10</td>
</tr>
<tr>
<td>Learning rates (initial, max, final)</td>
<td>1.8e-4, 2.2e-4, 5.2e-5</td>
</tr>
</tbody>
</table>
S5 Results on the pre-training set

Figure S4: Parity plots and histograms of the $\Delta \Delta G_{\text{solv}}^\dagger$ errors on the 5-fold pre-training set reaction split. (a) The pre-trained model with no additional feature. (b) The pre-trained model with the RP-solv feature. (c) The fine-tuned model with no additional feature. (d) The fine-tuned model with the RP-solv feature. The MAE and RMSE are in kcal/mol. The numbers of reactions, solvents, total data points found in the test set are provided.
Figure S5: Parity plots and histograms of the $\Delta \Delta H^\dagger_{solv}$ errors on the 5-fold pre-training set reaction split. (a) The pre-trained model with no additional feature. (b) The pre-trained model with the RP-solv feature. (c) The fine-tuned model with no additional feature. (d) The fine-tuned model with the RP-solv feature. The MAE and RMSE are in kcal/mol. The numbers of reactions, solvents, total data points found in the test set are provided.
Figure S6: Parity plots and histograms of the test errors on the 5-fold pre-training set solvent split. The results correspond to the pre-trained model with the RP-solv feature. (a) \( \Delta \Delta G_{\text{solv}}^{\ddagger} \) results. (b) \( \Delta \Delta H_{\text{solv}}^{\ddagger} \) results. The MAE and RMSE are in kcal/mol. The numbers of reactions, solvents, total data points found in the test set are provided.
Table S4: Top 20 reactions with the highest RMSE on the pre-training test set with the reaction split. The test errors are computed with the fine-tuned model with no additional feature. The RMSE, mean $\Delta \Delta G^\ddagger_{\text{solv}}$ of the reaction across different solvents, and gas phase barrier height ($E_a, \text{gas}$) are shown in kcal/mol for both forward and reverse directions. The gas phase barrier heights are obtained from the published works by Grambow et al. \[1,2,3\]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Reaction</th>
<th>Forward: RMSE, Mean $\Delta \Delta G^\ddagger_{\text{solv}}$, $E_a, \text{gas}$</th>
<th>Reverse: RMSE, Mean $\Delta \Delta G^\ddagger_{\text{solv}}$, $E_a, \text{gas}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Reaction 1" /></td>
<td>RMSE = 10.39, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -15.13$, $E_a, \text{gas} = 106.82$</td>
<td>RMSE = 8.46, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -11.48$, $E_a, \text{gas} = 11.27$</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2" alt="Reaction 2" /></td>
<td>RMSE = 8.07, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -7.29$, $E_a, \text{gas} = 118.41$</td>
<td>RMSE = 8.28, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -7.61$, $E_a, \text{gas} = 109.81$</td>
</tr>
<tr>
<td>3</td>
<td><img src="image3" alt="Reaction 3" /></td>
<td>RMSE = 8.04, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -9.73$, $E_a, \text{gas} = 61.96$</td>
<td>RMSE = 5.26, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -3.68$, $E_a, \text{gas} = 37.76$</td>
</tr>
<tr>
<td>4</td>
<td><img src="image4" alt="Reaction 4" /></td>
<td>RMSE = 5.39, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -4.70$, $E_a, \text{gas} = 73.75$</td>
<td>RMSE = 7.78, Mean $\Delta \Delta G^\ddagger_{\text{solv}} = -7.34$, $E_a, \text{gas} = 64.63$</td>
</tr>
</tbody>
</table>
5
Forward: RMSE = 6.96, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -7.34$ $E_{a, \text{gas}} = 100.61$
Reverse: RMSE = 7.62, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -9.66$ $E_{a, \text{gas}} = 69.47$

6
Forward: RMSE = 6.18, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -7.94$ $E_{a, \text{gas}} = 91.40$
Reverse: RMSE = 6.67, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -9.90$ $E_{a, \text{gas}} = 90.10$

7
Forward: RMSE = 4.54, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -5.39$ $E_{a, \text{gas}} = 64.93$
Reverse: RMSE = 6.51, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -8.10$ $E_{a, \text{gas}} = 68.22$

8
Forward: RMSE = 6.51, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -7.66$ $E_{a, \text{gas}} = 82.50$
Reverse: RMSE = 0.73, Mean $\Delta\Delta G_{\text{solv}}^\dagger = 3.63$ $E_{a, \text{gas}} = 48.45$
9

Forward: RMSE = 4.92, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -7.85$ $E_a, \text{gas} = 67.94$
Reverse: RMSE = 6.49, Mean $\Delta\Delta G_{\text{solv}}^\dagger = 8.41$ $E_a, \text{gas} = 47.89$

10

Forward: RMSE = 5.88, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -3.05$ $E_a, \text{gas} = 82.90$
Reverse: RMSE = 4.58, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -2.10$ $E_a, \text{gas} = 58.42$

11

Forward: RMSE = 5.82, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -11.18$ $E_a, \text{gas} = 56.95$
Reverse: RMSE = 4.03, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -7.00$ $E_a, \text{gas} = 35.19$

12

Forward: RMSE = 5.81, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -6.83$ $E_a, \text{gas} = 95.34$
Reverse: RMSE = 5.33, Mean $\Delta\Delta G_{\text{solv}}^\dagger = -6.12$ $E_a, \text{gas} = 119.94$
13
Forward: RMSE = 3.39, Mean $\Delta \Delta G_{solv}^{\ddagger} = -2.42$  $E_{a,\text{gas}} = 63.52$
Reverse: RMSE = 5.78, Mean $\Delta \Delta G_{solv}^{\ddagger} = -0.22$  $E_{a,\text{gas}} = 2.80$

14
Forward: RMSE = 5.68, Mean $\Delta \Delta G_{solv}^{\ddagger} = -6.69$  $E_{a,\text{gas}} = 87.57$
Reverse: RMSE = 2.40, Mean $\Delta \Delta G_{solv}^{\ddagger} = -3.05$  $E_{a,\text{gas}} = 73.83$

15
Forward: RMSE = 5.61, Mean $\Delta \Delta G_{solv}^{\ddagger} = -6.09$  $E_{a,\text{gas}} = 77.90$
Reverse: RMSE = 4.87, Mean $\Delta \Delta G_{solv}^{\ddagger} = -5.28$  $E_{a,\text{gas}} = 30.34$

16
Forward: RMSE = 5.43, Mean $\Delta \Delta G_{solv}^{\ddagger} = -4.63$  $E_{a,\text{gas}} = 105.63$
Reverse: RMSE = 4.57, Mean $\Delta \Delta G_{solv}^{\ddagger} = -3.52$  $E_{a,\text{gas}} = 8.89$
17

Forward: $\text{RMSE} = 5.29$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -4.12$ $E_a, \text{gas} = 76.20$
Reverse: $\text{RMSE} = 0.35$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -0.82$ $E_a, \text{gas} = 5.24$

18

Forward: $\text{RMSE} = 5.24$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -11.53$ $E_a, \text{gas} = 93.04$
Reverse: $\text{RMSE} = 0.31$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = 0.28$ $E_a, \text{gas} = 22.60$

19

Forward: $\text{RMSE} = 3.74$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -6.78$ $E_a, \text{gas} = 58.09$
Reverse: $\text{RMSE} = 5.23$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -7.67$ $E_a, \text{gas} = 83.93$

20

Forward: $\text{RMSE} = 5.15$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -6.35$ $E_a, \text{gas} = 86.67$
Reverse: $\text{RMSE} = 4.36$, Mean $\Delta \Delta G_{\text{solv}}^\ddagger = -3.85$ $E_a, \text{gas} = 18.28$
Table S5: Examples of reactions found in the 10 most frequent reaction types

1. $+\text{C-H} \cdot \text{C-H} \cdot \text{C-C}$ (647 reactions)
   SMARTS: [C:1][C:2][H:3] $\rightarrow$ [C:2][C:1][H:3]

   ![Reaction 1](image1)

2. $+\text{O-H} \cdot \text{C} \equiv \text{C} \cdot \text{O} \cdot \text{C-H}$ (224 reactions)
   SMARTS: [C:1]==[C:2].[O:3][H:4]$\rightarrow$[C:1][([C:2][O:3])[H:4]]

   ![Reaction 2](image2)

3. $+\text{C} \equiv \text{C} \cdot \text{H} \cdot \text{C} \cdot \text{H} \cdot \text{C}$ (183 reactions)
   SMARTS: [C:1]==[C:2].[H:3][H:4]$\rightarrow$[C:1][([H:3])[C:2][H:4]]

   ![Reaction 3](image3)
4. $+\text{C-O,-C-O,-C-C}$ (133 reactions)
SMARTS: $[^{\text{C:1}}][^{\text{C:2}}][^{\text{O:3}}] \rightarrow [^{\text{C:2}}][^{\text{C:1}}][^{\text{O:3}}]$

5. $+\text{C-C\#N,-C-N\#C}$ (110 reactions)
SMARTS: $[^{\text{C:1}}][^{\text{N+:2}}] \rightarrow [^{\text{C:3}}][^{\text{N:2}}]$}

6. $+\text{C-C,-C-C,-C-C}$ (107 reactions)
SMARTS: $[^{\text{C:1}}][^{\text{C:2}}][^{\text{C:3}}] \rightarrow [^{\text{C:2}}][^{\text{C:1}}][^{\text{C:3}}]$
7. +O-H-C-O,-C-H (107 reactions)
SMARTS: [C:1].[O:2][H:3]»[C:1][H:3][O:2]

8. +C=C,+C-H,-C-C,-C-H (94 reactions)
SMARTS: [C:1]==[C:2].[C:3][H:4]»[C:3][C:1][C:2][H:4]

Top $\Delta\Delta G_{\text{solv}}^\dagger$ error reactions:

$\text{MAE} = 2.2 \text{ kcal/mol}$

$\text{MAE} = 1.8 \text{ kcal/mol}$

$\text{MAE} = 1.7 \text{ kcal/mol}$

$\text{MAE} = 1.6 \text{ kcal/mol}$
9. +C=O, +C-H, -C-O, -C-C, -C-H (87 reactions)
SMARTS: [C::1]=[O::2].[C::3][H:4]»[C::3][C::1][(H:4)](O:2)

SMARTS: [C::1]=[C::2][O:3][H:4]»[C::1][(C::2)=([O:3])][H:4]
References

