Supporting Information for

Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts

Simone Gallarati,a,+ Raimon Fabregat,a,+ Rubén Laplaza,a,b Sinjini Bhattacharjee,a,c Matthew D. Wodricha and Clemence Corminboeufa,b,d,*

aLaboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

bNational Center for Competence in Research – Catalysis (NCCR-Catalysis), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

cIndian Institute of Science Education and Research, Dr Homi Bhabha Rd, Ward No. 8, NCL Colony, Pashan, Pune, Maharashtra 411008, India

dNational Center for Computational Design and Discovery of Novel Materials (MARVEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

*Email: clemence.corminboeuf@epfl.ch
+These authors contributed equally to this work.

Contents

1. Ligand Configurations for Boltzmann Weighting 1
2. Learning Curves 2
3. Feature Importances 2
4. Hyperparameters 3
5. Predicted e.e. Values 4
6. Out-of-sample Predictions with Retrained Model 5
7. DFT Optimised XYZ Structures and Energies 6
8. Out-Of-Sample Machine Learning Predicted Activation Energies 6
1. Ligand Configurations for Boltzmann Weighting

**Figure S1.** 10 distinct ligand arrangements leading to the (R)- or (S)-propargyl alcohol for $C_2$-symmetric bidentate Lewis-based catalysed propargylation reactions. Nu = alkyl nucleophile. For each ligand configuration BP1–5, the alkyl nucleophile can add to either face of benzaldehyde, yielding 10 possible diastereomeric TSs (R)- or (S)-.
2. Learning Curves

**Figure S2.** Learning curves for the different molecular representations used. *a*) Curves correspond to the SLATM representations of 3 and 2 (dashed and solid blue, respectively), 3 – 2 (orange), 3 – 2 with 500 features selected using Mutual Information importances (red), and 3 – 2 with 500 features selected using $r^2$ linear regression coefficients (green). *b*) Curves correspond to the learning curves of 3 – 2 using different standard atomistic ML representations: Coulomb Matrix (blue), Bag of Bonds (orange), and SLATM (green).

3. Feature Importances

**Figure S3.** Feature importances of the SLATM$_{DIFF}$ representations of the dataset, computed using: (blue) the variance, (orange) the $r^2$ linear regression coefficient, and (green) the Mutual Information.
4. Hyperparameters

Figure S4. a-c) Average hyperparameter fitting curves for the 100 train/test splits. The error bars are calculated with the standard deviation in the 100 splits. d) Importances of features sorted by the average feature importance in the 100 train/test splits. The error bars are computed using the standard deviation in the feature importance for the 100 splits.

Table S1. Optimised hyperparameters, obtained through grid-search optimisation, of the ML model for each of the representations discussed in the main text. $\sigma$ controls kernel width and $\lambda$ is the ridge parameter for regularization.

<table>
<thead>
<tr>
<th></th>
<th>$\sigma$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SLATM_2$</td>
<td>180</td>
<td>$1 \times 10^{-5}$</td>
</tr>
<tr>
<td>$SLATM_{DIFF}$</td>
<td>1.5</td>
<td>$1 \times 10^{-6}$</td>
</tr>
<tr>
<td>$SLATM_{DIFF+}$</td>
<td>1.5</td>
<td>$1 \times 10^{-6}$</td>
</tr>
</tbody>
</table>
5. Predicted \textit{e.e.} Values

\textbf{Figure S5.} ML-predicted vs. reference DFT \textit{e.e.} values for the 76 catalysts using each of the three different approaches discussed in the main text: SLATM$_2$ (blue), SLATM$_{\text{DIFF}}$ (orange) and SLATM$_{\text{DIFF+}}$ (green). Most of the points are hidden by the overlaps at the 100/100 region. Data corresponds to \textbf{Figure 3} of the main text and details on their generation are given in the machine learning section.
6. Out-of-sample Predictions with Retrained Model

Figure S6. ML-predicted vs. reference DFT $E_a$ values of out-of-sample catalysts 7j and 7k. The ML model was re-trained on all of the 754 data points, without splitting them into the 90/10 train/test sets, using the same hyperparameters previously obtained in the cross-validation training. The features of SLATM$_{\text{DIFF+}}$ were also selected using the full dataset (754 points), but they did not vary from those selected in the previous cross-validation splits.
7. DFT Optimised XYZ Structures and Energies

The structures of the 1508 catalytic cycle intermediates, optimised at the PCM_{DCM/B97-D/TZV(2p,2d)} level, are provided in the folders DFTgeomInt2 and DFTgeomInt3. The absolute energies (in atomic units) of intermediates 2, 3, and of the enantiodetermining TSs are provided in DFTEnergies.csv. The ML-predicted relative $E_a$ values for each species, in kcal mol$^{-1}$, using the three representations discussed in the main text, are provided in ActivationEnergiesPredictions.csv.

Note that all our data (optimised structures, energies, ML predictions) can be found in the Materials Cloud.

8. Out-Of-Sample Machine Learning Predicted Activation Energies

The ML-predicted and DFT-computed activation energies of the out-of-sample catalysts 7j and 7k with the SLATM_{DIFF+} representation are given in the OOSPredictions.csv file, while the geometries of catalytic cycle intermediates 2 and 3 and of the enantiodetermining transitions states are given in the folders DFTgeomOOSInt2, DFTgeomOOSInt3 and DFTgeomOOSTS.