Molecular Field Analysis Using Computational-Screening Data in Asymmetric N-Heterocyclic Carbene-Copper Catalysis toward Data-driven in silico Catalyst Optimization

Masakiyo Mukai†, Kazunori Nagao†, Shigeru Yamaguchi‡,* and Hirohisa Ohmiya†,§,*

†Division of Pharmaceutical Sciences, Graduate School of Medical Sciences, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan.
‡RIKEN Center for Sustainable Resource Science, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.
§JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.

Supporting Information Placeholder

ABSTRACT: A molecular-field-based regression analysis using computational screening data for N-heterocyclic carbene (NHC)-Cu-catalyzed asymmetric carbonyl additions of a silylboronate to aldehydes is reported. A computational screening was performed to collect enantioselectivity data (ΔΔG‡: energy differences between the transitions states leading to each enantiomer) via transition-state (TS) calculations using density functional theory (DFT) methods. A molecular field analysis (MFA) was carried out using the obtained calculated ΔΔG‡ values and TS structures (30 samples in total). Important structural information for enantioselectivity extracted by the MFA was visualized on the TS structures, which provided insight into an asymmetric induction mechanism. Based on the obtained information, chiral NHC ligands were designed, which showed improved enantioselectivity in these carbonyl additions.

Data-driven machine-learning-based optimization of molecular catalysis is an emerging and promising research area.1 Regressi

ional analysis, i.e., quantitative structure-property relationships (QSPR) modeling of catalytic reactions typically correlates molecular descriptors with experimental catalytic activity values. High quality-experimental data suitable for QSPR modeling are, however, not always available. In some cases, experimental data include non-negligible noise derived from various factors, such as side reactions and experimental errors. The use of such data in regression analysis/QSPR can reduce its predictive performance, and in some cases, the information extracted by the analysis may not provide reliable insights into the reactions. The use of computational-screening data obtained through ab-initio calculations avoids this potential shortcoming, as the calculated activity data do not include the aforementioned noise. In addition, computational-screening approaches do not require experiments to collect data, thereby enabling data analysis of reactions that involve expensive and synthetically difficult catalysts. Despite these attractive features, reports of data-driven molecular design in molecular catalysis using the QSRR framework with the calculated catalytic-activity values as the target variables are scarce,2 although virtual screening using transition-state force fields developed via the quantum-guided molecular-mechanics method has been investigated in asymmetric catalysis.3,4,5 As transition-state (TS) DFT calculations with appropriate calculation conditions have become a reliable method to design asymmetric catalytic reactions,4,5 the use of the calculated selectivity values as target variables for QSRR or quantitative structure-selectivity relationships (QSSR) can be expected to further accelerate the development of asymmetric catalytic reactions. Recently, a QSSR study of the asymmetric propargylation of aryl aldehydes has been reported,6 in which more than 600 training samples with literature-reported calculated activation-energy values were used. However, the calculation cost of DFT-based TS modeling is high. Therefore, it would be highly desirable to develop a regression-based method that enables molecular design to improve enantioselectivity by analyzing a relatively small set of DFT-based computational-screening data. Here, we report the data-driven design of chiral catalysts for N-heterocyclic carbene (NHC)-Cu-catalyzed asymmetric carbonyl additions of a silylboronate to aldehydes through molecular-field-based regression analysis using the calculated ΔΔG‡ values and transition-state structures with a total of 30 training samples (Figure 1a).
Structures. We hypothesized that the reaction proceeded via the

four-centered reaction mechanism that is frequently observed in
related Cu(I)-catalyzed reactions.\textsuperscript{8,9} As an example, the TS
structure derived from ligand L\textsubscript{3} and substrate S\textsubscript{1} (L\textsubscript{3}S\textsubscript{1}) is
shown in Figure 2. The calculated \( \Delta \Delta G^2 \) values exhibited a correlation
with the experimental \( \Delta \Delta G^2 \) values (see Figure S9). For
data analysis, we employed a 3D-QSSR approach,\textsuperscript{10} which we refer
to here as molecular field analysis (MFA). MFA is a regression analysis between reaction outcomes and molecular
fields calculated from three-dimensional molecular structures.

In our previous study, we found that MFA of the intermediates
in the enantio-determining step of asymmetric catalytic reac-
tions enabled the extraction and visualization of structural in-
formation that led to the design of molecules with improved en-
antioselectivity (Figure 1b).\textsuperscript{10,11} Thus, we were interested in
whether MFA between the calculated \( \Delta \Delta G^2 \) values and the TS
structures would be able to extract similar information (Figure
1a). The seminal report of the use of MFA in asymmetric catal-
ysis by the Kozlowski group employed TS structures; however,
the authors did not use the extracted and visualized information
for molecular design, and their target variables were the exper-
imental \( \Delta \Delta G^2 \) values.\textsuperscript{12} Our specific goal in this study was to
evaluate our approach through the data-driven design of the cur-
rent optimum NHC ligand L\textsubscript{6} in the targeted carbonyl addi-
tions\textsuperscript{7} starting from a set of 18 training samples (Table 1).

Table 1. Enantioselectivity of the training samples (%ee).

<table>
<thead>
<tr>
<th>Substrate (Ar\textsuperscript{1})</th>
<th>L\textsubscript{1}</th>
<th>L\textsubscript{2}</th>
<th>L\textsubscript{3}</th>
<th>L\textsubscript{4}</th>
<th>L\textsubscript{5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>21 (0.34)</td>
<td>31 (0.58)</td>
<td>73 (1.77)</td>
<td>87 (1.83)</td>
<td>82 (2.25)</td>
</tr>
<tr>
<td>S2</td>
<td>18 (0.59)</td>
<td>41 (0.65)</td>
<td>17 (1.67)</td>
<td>85 (2.29)</td>
<td>85 (2.32)</td>
</tr>
<tr>
<td>S3</td>
<td>15 (0.60)</td>
<td>38 (0.71)</td>
<td>63 (1.33)</td>
<td>80 (2.31)</td>
<td>78 (2.81)</td>
</tr>
<tr>
<td>S4</td>
<td>11 (0.04)</td>
<td>18 (0.70)</td>
<td>66 (1.82)</td>
<td>78 (2.43)</td>
<td>74 (2.41)</td>
</tr>
<tr>
<td>S5</td>
<td>19 (0.16)</td>
<td>27 (0.30)</td>
<td>69 (0.85)</td>
<td>82 (2.18)</td>
<td>81 (2.76)</td>
</tr>
<tr>
<td>S6</td>
<td>17 (0.38)</td>
<td>31 (-0.25)</td>
<td>63 (1.28)</td>
<td>88 (2.17)</td>
<td>78 (2.30)</td>
</tr>
</tbody>
</table>

Calculated \( \Delta \Delta G^2 \) values (kcal/mol) are shown in parentheses. The
18 initial training samples are highlighted in grey.
As molecular fields, we employed steric indicator fields composed of indicator variables (with values of 0 or 1), as shown in Figure 1. The indicator fields were calculated for each unit cell of a grid space. A unit cell that includes the van der Waals radius of any atom was assigned a value of 1, otherwise, the cell was given a value of 0 (Figures 1 and S2). The indicator fields were calculated from the TS structures of a minor pathway (the R pathway), as we aimed to obtain guidelines for the design of chiral NHC ligands that would destabilize the minor pathway. By correlating the calculated ΔG‡ values and the indicator fields using LASSO or Elastic Net regression according to the reported procedure\textsuperscript{15} employing the R package glmnet\textsuperscript{16}, we generated a regression model (for more details regarding the MFA, see the SI, page S4). Thus, we successfully improved the enantioselectivity through data-driven catalyst design using computational-screening data.

**Figure 2.** Transition-state structures of L\textsubscript{3S1}; calculations were performed at the \textit{ab}97XD/SDD(Cu) & 6-311+G**-SMD(toluene)//B3LYP-D3/LANL2DZ(Cu) & 6-31G* level of theory.

\[ \Delta G^\ddagger = 1.8 \text{ kcal/mol} \]

73\% ee

\[ \Delta G^\ddagger = 0.3 \text{ kcal/mol} \]

21\% ee

\[ \Delta G^\ddagger = 1.8 \text{ kcal/mol} \]

87\% ee

**Figure 3.** Important structural information visualized in the TS structures. The information extracted by the MFA using 18 samples is shown, together with the TS structures (a) L\textsubscript{3S1}, (b) L\textsubscript{1S1}, and (c) L\textsubscript{4S1}. Calculated ΔG‡ and experimentally determined enantiomeric excess values are also shown. The silyl group was removed for the MFA (for more details, see Figure S2), and thus, we have omitted the silyl group in Figures 3 and 4.

the blue point described above (Figures 3c and S4). The calculated ΔG‡ values for L\textsubscript{4S1} (1.8 kcal/mol) and L\textsubscript{5S1} (2.2 kcal/mol) are higher than that of the design template L\textsubscript{1S1} (0.3 kcal/mol). We experimentally confirmed that the enantioselectivity of the reactions L\textsubscript{4S1} (87\% ee) and L\textsubscript{5S1} (82\% ee) are improved relative to the selectivity of L\textsubscript{3S1} (73\% ee), which exhibited the highest value among the 18 training samples (for more details of the molecular design workflow, see the SI, page S4). Thus, we successfully improved the enantioselectivity through data-driven catalyst design using computational-screening data.
We then performed the MFA again, this time including the data obtained from the designed ligands (30 samples, Table 1). Figure 4a shows the structural information in the TS structures calculated from the designed ligands. Based on this information, we designed L6 and L7, which bear Me or iPr groups at the 3- and 5-positions of the aromatic ring. The structures L6 and L7 overlapped with the structural information used for the design (Figures 4b, 4c and Figure S5). The calculated ΔΔG‡ values are higher than those of the design template (L4S1: 1.8 kcal/mol; L6S1: 3.3 kcal/mol; L7S1: 2.1 kcal/mol). We then performed the reactions L6S1 and L7S1. To our delight, further improvement of the enantioselectivity was observed (L6S1: 90% ee; L7S1: 96% ee; L4S6: 88% ee, which is the highest value among the 30 training samples). Although DFT calculations are useful to design chiral catalysts, DFT-based computational predictions vary vastly depending on the calculation conditions and often fail as replacements for experiments. In this case, we employed the B97XD functional for the single-point calculations; we also confirmed that MFA using ΔΔG‡ values obtained via single-point calculations using other representative functionals for the TS calculations in asymmetric catalysis led to the same molecular design as described above (Figures S10–S12). Although we employed 30 training samples to secure the quality and reliability of the regression models, practically, we can in this case successfully design appropriate molecules using a smaller sample size (for details, see Figure S13).

In summary, a molecular field analysis (MFA) using computational-screening data and the corresponding transition-state (TS) structures with a relatively small dataset (30 samples) enabled the design of chiral NHC ligands with improved enantioselectivity for the Cu-catalyzed carbonyl addition of a silylboronate. The design of ligands L6 and L7 was achieved through iterative MFA-based molecular design. Further optimization of asymmetric catalysis using this strategy is currently in progress. As the structural information comprises Cartesian coordinates that computers can understand, the combination of our MFA-based strategy and a structure generator could lead to the future development of efficient computational chiral catalyst optimization programs/software, as has been seen in the field of the computational design of drug-like molecules.

**ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge on the ACS Publications website.

Experimental details, characterization data for all new compounds, details of calculations and data analysis (PDF)

Input data for the data analysis (Zip)

Output data obtained from the data analysis (Zip)

Coordinates of all the TS structures (XYZ)

**AUTHOR INFORMATION**

**Corresponding Author**

**Shigeru Yamaguchi** – RIKEN Center for Sustainable Resource Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan; orcid.org/0000-0003-2304-8448; Email: shigeru.yamaguchi.hw@riken.jp

**Hirohisa Ohmiya** – Division of Pharmaceutical Sciences, Graduate School of Medical Sciences, Kanazawa University, Kanazawa 920-1192, Japan; orcid.org/0000-0002-1374-1137; Email: ohmiya@p.kanazawa-u.ac.jp

**Authors**

**Masakiyo Mukai** – Division of Pharmaceutical Sciences, Graduate School of Medical Sciences, Kanazawa University, Kanazawa 920-1192, Japan
Kazunori Nagao – Division of Pharmaceutical Sciences, Graduate School of Medical Sciences, Kanazawa University, Kanazawa 920-1192, Japan; orcid.org/0000-0003-3141-5279

Notes
The authors declare no competing interests.

ACKNOWLEDGMENT
This work was supported by JSPS KAKENHI grants JP17H06449 (H. O.) and 20H04831 (S. Y.) (Hybrid Catalysis). The DFT calculations were performed on the RIKEN HOKUSAI supercomputer system.

REFERENCES


11. Although careful selection of training samples is required for predictive purposes as the Denmark group has demonstrated (cf. ref. 1d), our methodology aims at a mechanistic interpretation. The molecular design is performed using a combination of the researcher’s intuition and obtained insights (not predicted values), thus allowing a rough sample selection.
based on practical considerations such as availability as described in this manuscript.


