

Transfer learning across different photocatalytic organic reactions

Corresponding Author: Dr Naoki Noto

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The paper explores how knowledge can be transferred between different types of photocatalytic organic reactions using domain-adaptation-based transfer learning (TL) approach. By integrating experimental and computational protocols, this method has significantly improved the prediction of photocatalytic activity compared to conventional machine learning (ML) methods. Notably, small training data set led to the establishment of satisfying predictive performance, and the method also demonstrated potentials in identifying effective OPSs for alkene photoisomerization. However, some issues concerning effectiveness of methodology and reliability of conclusion should be addressed before consideration of publication.

1. Figure 1 exhibits DA method for concerning photocatalytic reactions, which is the key innovation of this paper. The simple figure in the present version gives poor insights into how exactly the methodology works. I recommend the author add a new Figure 1b to show the detailed procedure of DA (containing how the training data is derived, how the training is transferred, and other computational details).

2. In Figure 4c, Avg R2 and Max R2 for DFT_FE showed a slower increasing trend compared to DFT descriptor, especially for Max R2. Std R2 also hold this trend. If the training data is sufficient, will the deviation between these two methods becomes quite small? Can the author provide some clue for the effectiveness of method when the reaction possessing sufficient training data?

3. In application part of Figure 5, total 10 entries were selected as experimentally accessible small training dataset (OPS5, OPS9, OPS20, OPS23, OPS27, OPS31, OPS34, OPS44, OPS59, and OPS83). However, it is easy to discover that the performance of these ten samples ranging from 0% to 92%, which is a quite satisfying range for ML. It would be helpful for the author to provide some data demonstrating the results predicted from poor-performance entries (hypothesizing that OPS9 and OPS 44 is replaced by some moderate-performance samples). Under these circumstances, will the model also be able to predict high-performance catalyst?

4. In the experimental design section, the authors should further discuss the criteria used for selecting the training datasets. Although the importance of high similarity and data diversity is acknowledged, there remain challenges in selecting the most effective source-domain data. Authors could increase the importance of their research by adding more details about how they chose the most representative datasets, which provides more guidance for following researchers.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In recent years, the application of data-driven methodologies in organic synthesis has attracted considerable attention, particularly for their ability to comprehensively and accurately delineate structure-performance relationships. Such approaches hold significant promise for understanding and expanding the applicability of synthetic reactions, thereby enhancing efficiency and innovation in chemical research. In this manuscript, the authors explore the utility of domain-adaptation-based transfer learning to predict the catalytic behavior of organic photosensitizers across different photocatalytic reactions. By leveraging knowledge from cross-coupling reactions and applying it to [2+2] cycloaddition reactions, the study demonstrates substantial improvements in predictive performance. While the results presented are indeed intriguing and suggest potential advancements in ML-empowered synthetic chemistry, I maintain a cautious stance regarding the

robustness of the conclusions drawn. Consequently, I recommend that the manuscript undergo major revisions to address the following critical concerns before it can be considered for publication in Nature Communications.

Major Issues

1. My primary concern pertains to the statistical validation underpinning the significance of the domain adaptation approach utilized in this study. Transfer learning, by its very nature, relies on the robustness and reliability of the source model before any adaptation to the target domain. The observed substantial improvement in predictive performance appears unexpectedly high when compared to existing transfer learning studies. To substantiate the efficacy of this approach, the authors must provide rigorous evidence that isolates the impact of domain adaptation from other potential confounding factors, such as the introduction of additional data points.
2. As Figure 3b suggests a high correlation in yields, it might simply be due to increased data volume. Therefore, it is crucial that the authors directly apply a self-trained model from the source domain to evaluate its predictive performance on cycloaddition reactions under various combinations of algorithms and descriptors, as well as mix all data to examine the effects on cycloaddition reaction predictions. Only if the domain adaptation method further improves upon these combinations can it be considered effective.
3. With only limited data points in the target domain, it is challenging to exclude bias. Drawing a parallel, Doyle's previous publication on Buchwald-Hartwig coupling reactions presented a perfect/challenging task for transfer learning (10.1126/science.aar5169). In this paper, they found that the transfer learning between different halides are challenging, which could serve as a good reference point.
4. Last but not least, determining what constitutes a source or target domain in chemistry poses a significant challenge. It is crucial to establish clear criteria for these designations. What examples can be considered as the source domain? What qualifies as the target domain? In the context of this manuscript, are the domains designated as source simply because the authors already possess this data? For instance, if different coupling reactions were considered as the target domain, would the transfer learning effects still prove effective? This question underlines the need for a rigorous definition and justification of the selection criteria for source and target domains in the application of transfer learning in chemical research.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The manuscript presents an innovative approach to predicting suitable catalysts for new reactions using transfer learning (TL) and domain adaptation (DA) techniques. The authors claim to have successfully transferred knowledge from photocatalytic cross-coupling reactions to [2+2] cycloaddition reactions, achieving satisfactory predictive performance with only ten training data points.

These are some issues should be consider when publish this manuscript:

1. The authors claim to achieve satisfactory predictive performance using only ten training data points. This extremely small dataset raises significant concerns about the model's ability to generalize: It is challenging to train a robust machine learning algorithm with such a limited dataset, even with transfer learning. The impressive results reported may be indicative of overfitting rather than true predictive power. The authors should provide additional experiments to demonstrate that their algorithm is not overfitting and has good generalization capabilities. This could include: cross-validation results, performance on a separate, larger test set, and learning curves showing performance as the training set size increases.
2. The manuscript mentions using TrAdaBoostR2 and claims superior performance compared to other supervised DA methods. However, more details are needed: The authors should provide a clear explanation of how knowledge is transferred from one domain to another. The relationship between features in the source and target domains should be explicitly described. A comparative analysis with other DA methods, including quantitative results, would strengthen the claims of superiority.
3. As a reviewer with expertise in machine learning but limited knowledge in organic chemistry, I found some sections of the manuscript challenging to evaluate fully. It would be beneficial to have a domain expert in organic chemistry review the chemical aspects of this work to ensure the validity and significance of the findings in that context.

In addition, It would be helpful to include a discussion on the limitations of this approach and potential areas for future work.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have fully addressed my questions, the paper can be recommended for publication. One small issue, the comment 12 raised by reviewer 3 is an important issue, ML model training with only ten data points is problematic. The authors seem to already address this issue by supplementing Figure S5, however, in the corresponding discussions in revised manuscript, I do not see any corresponding discussion or mentioning of the performance comparison among different number of training data points, rather than the focused discussion on the generation performance comparison between different models.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have addressed my previous concerns. I support its publication as is.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Responses to the reviewers' comments (Noto et al, NCOMMS-24-59374-T)

Reviewer 1:

Comment 1: The paper explores how knowledge can be transferred between different types of photocatalytic organic reactions using domain-adaptation-based transfer learning (TL) approach. By integrating experimental and computational protocols, this method has significantly improved the prediction of photocatalytic activity compared to conventional machine learning (ML) methods. Notably, small training data set led to the establishment of satisfying predictive performance, and the method also demonstrated potentials in identifying effective OPSs for alkene photoisomerization. However, some issues concerning effectiveness of methodology and reliability of conclusion should be addressed before consideration of publication.

Response 1: We appreciate the favorable comments and valuable suggestions.

Comment 2: 1. Figure 1 exhibits DA method for concerning photocatalytic reactions, which is the key innovation of this paper. The simple figure in the present version gives poor insights into how exactly the methodology works. I recommend the author add a new Figure 1b to show the detailed procedure of DA (containing how the training data is derived, how the training is transferred, and other computational details).

Response 2: We revised Figure 1, where a new Figure 1b was fabricated for further explanation. Figure 1b provides information on the source and target datasets used for DA, along with the fundamental mechanism and purpose of knowledge transfer through DA.

Comment 3: 2. In Figure 4c, Avg R² and Max R² for DFT_FE showed a slower increasing trend compared to DFT descriptor, especially for Max R². Std R² also hold this trend. If the training data is sufficient, will the deviation between these two methods becomes quite small? Can the author provide some clue for the effectiveness of method when the reaction possessing sufficient training data?

Response 3: We tested the use of 60, 70, and 80 training data points for both the TrAB/DFT_FE model and the RF/DFT model. As a result, the Max R² value of the RF/DFT model increased with a larger training dataset but still did not reach that of the TrAB/DFT_FE model even when using 80 training data points due to the latter's exceptionally high prediction accuracy (TrAB/DFT_FE: Max R² = 0.95 vs. RF/DFT: Max R² = 0.81). Additionally, for the RF/DFT model, the Avg R² was not improved (70 training data: Avg R² = 0.37, 80 training data: Avg R² = 0.28), and the Std R² showed a significant increase (70 training data: Std R² = 0.19, 80 training data: Std R² = 1.05) when using 80 training data points, indicating large variability in the results. These findings suggest that while the RF/DFT model showed some improvement with increased training data, its predictive performance remained inferior to that achieved by the TrAB model (70 training data: Avg R² = 0.78, Std R² = 0.10, 80 training data: Avg R² = 0.78, Std R² = 0.14). Since we considered these results informative, they were added to the Supporting Information (Tables S10 and S18, and Figure S4).

Comment 4: 3. In application part of Figure 5, total 10 entries were selected as experimentally accessible small training dataset (OPS5, OPS9, OPS20, OPS23, OPS27, OPS31, OPS34, OPS44, OPS59, and OPS83). However, it is easy to discover that the performance of these ten samples ranging from 0% to 92%, which is a quite satisfying range for ML. It would be helpful for the author to provide some data demonstrating the results predicted from poor-performance entries (hypothesizing that OPS9 and OPS 44 is replaced by some moderate-performance samples). Under these circumstances, will the model also be able to predict high-performance catalyst?

Response 4: Since it is not easy to establish a criterion for selecting two moderate-performance OPSs

to replace high-performance OPSs, we simply removed **OPS9** and **OPS44** from the training dataset. When using eight training data points excluding the high-performance OPSs, the model still successfully identified the same OPSs (**OPS1**, **OPS7**, **OPS10**, **OPS11**, **OPS12**, and **OPS67**) as top performers, while the errors between predicted and experimental yields increased (**S1**: MAE = 14.8, **S6**: MAE = 9.6). We mentioned these findings in the revised manuscript (pages 17–18) and added the details to the Supporting Information (Figures S14 and S15).

Comment 5: 4. In the experimental design section, the authors should further discuss the criteria used for selecting the training datasets. Although the importance of high similarity and data diversity is acknowledged, there remain challenges in selecting the most effective source-domain data. Authors could increase the importance of their research by adding more details about how they chose the most representative datasets, which provides more guidance for following researchers.

Response 5: We added the necessary explanation to our revised manuscript to describe how source-domain datasets, including **S1**, were constructed when the target was **CA** (page 12). Source domains were constructed based on Pearson correlation coefficients representing trends in photocatalytic activity. We found that effective source-domain datasets generally showed higher correlation coefficients with the target reaction (**CA**) than ineffective datasets (Table 2). Further explanation of the relationship between the correlation coefficient and the effectiveness as a source domain is also included in the revised manuscript (page 13). For clarity, we primarily used source domains constructed using the correlation coefficients of the entire dataset, but this metric-based approach also proved effective when source domains were constructed using only the correlation coefficients among OPSs in the training data (Table S21; top 3: Avg R^2 = 0.64, bottom 3: Avg R^2 = 0.57). In addition, when predicting effective OPSs for alkene photoisomerization using 10 training data points, source-domain datasets were selected based on the correlation coefficients among the training data, as described on page 16 (see also Figures S8, S9, and S16). Overall, photoreaction datasets with catalytic activity trends similar to the target tend to be more useful for constructing the source domain. In the revised manuscript, we discuss that this similarity can be estimated to some extent using correlation coefficients.

Furthermore, to summarize the insights into source-domain selection gained throughout this study, we added a new section titled “Insights on the selection of source-domain datasets” to our revised manuscript (pages 22–24). This section emphasizes the importance of constructing source-domain datasets guided by organic-chemistry insights, because experimentally collecting data on photoreactions with trends in catalytic activity similar to the target is crucial, as mentioned earlier. Additionally, we mention the effectiveness of Pearson correlation coefficients in creating source domains, not only for **CA** but also for photocatalytic cross-coupling reactions, such as **CO_a**, **CO_b**, **CO_c**, **CO_d**, **CO_e**, **CS** and **CN** (see also Table S27). Meanwhile, the difference in correlation coefficients was not always sufficient to fully distinguish between useful and useless datasets. Additionally, in several cases, the DA competence of a source domain comprising all photoreactions other than the target was comparable to, or even outperformed, that of a source domain consisting solely of selected photoreactions biased toward trends in photocatalytic activity similar to the target (Tables 2 and S27). Therefore, we consider it challenging to draw decisive conclusions on the optimal method for selecting source-domain datasets solely from the results of this study. Meanwhile, considering the ability of TraAdaBoostR2 (TrAB) to mitigate the influence of uninformative data in the source domain, a practical, though not necessarily optimal, approach to addressing the issue of source-domain selection is to construct a source domain comprising more diverse photoreactions

with various substrate–product examples. Nonetheless, future studies are required to determine whether such a sample-weighting mechanism remains effective when the source domain includes reactions with catalytic activity trends significantly different from those of target EnT reactions, such as photoredox reactions.

We believe that the description added to the revised manuscript could provide valuable information for researchers across a wide range of fields. Once again, we appreciate the valuable suggestions that have helped enhance the quality of our work.

Reviewer 2:

Comment 6: In recent years, the application of data-driven methodologies in organic synthesis has attracted considerable attention, particularly for their ability to comprehensively and accurately delineate structure-performance relationships. Such approaches hold significant promise for understanding and expanding the applicability of synthetic reactions, thereby enhancing efficiency and innovation in chemical research. In this manuscript, the authors explore the utility of domain-adaptation-based transfer learning to predict the catalytic behavior of organic photosensitizers across different photocatalytic reactions. By leveraging knowledge from cross-coupling reactions and applying it to [2+2] cycloaddition reactions, the study demonstrates substantial improvements in predictive performance. While the results presented are indeed intriguing and suggest potential advancements in ML-empowered synthetic chemistry, I maintain a cautious stance regarding the robustness of the conclusions drawn. Consequently, I recommend that the manuscript undergo major revisions to address the following critical concerns before it can be considered for publication in Nature Communications.

Response 6: We appreciate the favorable comments and valuable suggestions.

Comment 7: 1. My primary concern pertains to the statistical validation underpinning the significance of the domain adaptation approach utilized in this study. Transfer learning, by its very nature, relies on the robustness and reliability of the source model before any adaptation to the target domain. The observed substantial improvement in predictive performance appears unexpectedly high when compared to existing transfer learning studies. To substantiate the efficacy of this approach, the authors must provide rigorous evidence that isolates the impact of domain adaptation from other potential confounding factors, such as the introduction of additional data points.

Response 7: We originally evaluated the predictive performance of each ML method using 100 different patterns of training and test data splits to validate statistical robustness. To further enhance statistical validation, we added paired t-test results to the revised Supporting Information, confirming statistically significant differences ($p < 0.05$) for all cases mainly discussed in the manuscript (Table S48). Furthermore, to address concerns of this reviewer as thoroughly as possible, we added results obtained from a comparison of prediction accuracy among domain adaptation and other methods, such as simply increasing the data volume, as well as investigations into the limitations of domain adaptation. For further details, please refer to Responses 8, 9 and 10, shown below. Overall, we appreciate these professional concerns, which gave us the opportunity to improve our research.

Comment 8: 2. As Figure 3b suggests a high correlation in yields, it might simply be due to increased data volume. Therefore, it is crucial that the authors directly apply a self-trained model from the source domain to evaluate its predictive performance on cycloaddition reactions under various

combinations of algorithms and descriptors, as well as mix all data to examine the effects on cycloaddition reaction predictions. Only if the domain adaptation method further improves upon these combinations can it be considered effective.

Response 8: We tested two additional methods for predicting the catalytic activity of OPSs. First, we employed a training dataset that includes all cross-coupling reactions (a total of 700 data points). Second, we utilized training datasets, in which the aforementioned 700 data points were combined with training datasets (50 data points) in the target domain (a total of 750 data points). These methods were referred to as Method B and Method C in the manuscript, respectively (see Figure 6a). For **CA**, Lasso, SVM, RF, and XGB models combined with various descriptor sets, such as Morgan Fingerprint (MF), RDKit descriptors (RDKit), MACCSKeys (MK), Mordred, their PCA-reduced versions, and a DFT-based descriptor set (denoted as DFT), were used. Consequently, none of the combinations used in Methods B and C outperformed TrAB (Method B: Avg $R^2 \leq 0.41$, Method C: Avg $R^2 \leq 0.52$, TrAB: Avg $R^2 = 0.65$), when TrAB was based on the DFT descriptors (DFT) and the source domain including the same 700 data points. Additionally, we tested Methods B and C for predicting photocatalytic activity in other cross-coupling reactions, as described in Figure 6a and Response 10. For these investigations, the XGB/DFT model was employed. While there were several results comparable to TrAB, TrAB consistently outperformed the other methods in all cases. TrAdaBoostR2 (TrAB) increases the weights of samples with large errors in the target domain and decreases those with large errors in the source domain within mixed datasets from the source and target domains. Given this mechanism, Method C is conceptually similar to TL used in this work, and the primary distinction of TrAB from Method C lies in its ability to perform this sample weighting. However, these findings highlight that the sample weighting mechanism of TrAB is the key to stably better performance in predicting photocatalytic activity.

Comment 9: 3. With only limited data points in the target domain, it is challenging to exclude bias. Drawing a parallel, Doyle's previous publication on Buchwald-Hartwig coupling reactions presented a perfect/challenging task for transfer learning ([10.1126/science.aar5169](https://doi.org/10.1126/science.aar5169)). In this paper, they found that the transfer learning between different halides are challenging, which could serve as a good reference point.

Response 9: We added the section titled “Investigations into applicability and limitations” to the manuscript (pages 19–22), which now includes the proposed paper and other relevant literature as references. In addition to evaluating the applicability of this approach to various photocatalytic cross-coupling reactions, as mentioned in Response 10, we also assessed the predictive performance for OPSs not included in the source domain. We randomly excluded 30 OPSs from the source domain and compared the predictive performance for these removed OPSs with those included in the source domain (Figure 6b). Consequently, the prediction accuracy for OPSs excluded from the source domain was much worse than for those included in the source domain (Figure 6b; included OPSs: Avg $R^2 = 0.76$, excluded OPSs: Avg $R^2 = 0.28$). By describing these results, we aim to ensure a fair discussion that underlines not only the successes but also the limitations of our approach, which need to be addressed in the future work. Meanwhile, we believe that the issues with the predictive performance of ML mentioned earlier can be addressed through efforts to construct an appropriate database informed by the expertise of organic chemists and are not inherently unsolvable.

Comment 10: 4. Last but not least, determining what constitutes a source or target domain in chemistry poses a significant challenge. It is crucial to establish clear criteria for these designations.

What examples can be considered as the source domain? What qualifies as the target domain? In the context of this manuscript, are the domains designated as source simply because the authors already possess this data? For instance, if different coupling reactions were considered as the target domain, would the transfer learning effects still prove effective? This question underlines the need for a rigorous definition and justification of the selection criteria for source and target domains in the application of transfer learning in chemical research.

Response 10: To address this reviewer's question regarding the applicability of our approach to other photoreactions, we evaluated the predictive performance for photocatalytic activity in **CO_a**, **CO_b**, **CO_c**, **CO_d**, **CO_e**, **CS** and **CN**, and the results obtained from these investigations were included in the newly added "Investigations into applicability and limitations" section (Figure 6a). The predictive performance was generally high for C–O bond-forming reactions (Avg $R^2 = 0.72\text{--}0.85$), likely because the source domain contains a sufficient number of reactions with trends in photocatalytic activity highly similar to the target domain. In contrast, while the prediction accuracy for **CS** was lower than that for the other cases (Avg $R^2 = 0.44$), TrAB still outperformed other methods tested. Overall, although alternative methods demonstrated comparable performance in some instances, the predictive performance by our approach using TrAB was proved consistently better and more stable across all cases evaluated in this study.

Furthermore, we added another new section titled "Insights on the selection of source-domain datasets" to the manuscript (pages 22–24). Both the [2+2] cycloaddition (**CA**) and alkene photoisomerization reactions, which were targeted in this work, are considered to proceed via energy transfer (EnT) processes. Accordingly, data on photocatalytic cross-coupling reactions, which also seemed to mainly involve EnT processes, were collected in this work as source-domain datasets. This database construction was informed by the organic-chemistry insight in that if OPSs play similar roles across photoreactions, data on their catalytic activity could be useful for TL. In addition to the description in the "Insights on the selection of source-domain datasets" section, additional arguments regarding this data-selection process were included on pages 8–9. Although one key factor contributing to the successful knowledge transfer in our case lies in this database construction guided by organic-chemistry insights, we also demonstrated that a simple metric, i.e., the Pearson correlation coefficient, could be used to distinguish between effective and ineffective datasets for the source-domain construction to some extent. This is derived from the observation that effective source-domain datasets generally exhibited higher correlation coefficients to the target reaction than ineffective datasets (Tables 2 and S27). Notably, this approach remains applicable even with limited knowledge of the reactions. However, the difference in correlation coefficients was not always sufficient to fully distinguish between useful and useless datasets. Furthermore, we observed the DA competence of a source domain comprising all photoreactions other than the target was comparable to, or even outperformed, that of a source domain consisting solely of selected photoreactions biased toward trends in photocatalytic activity similar to the target. Therefore, the overly selective use of datasets as the source domain, based on correlation coefficients, is not always effective. Meanwhile, considering the ability of TrAB to mitigate the influence of uninformative data in the source domain, a viable, though not necessarily optimal, approach to addressing the issue of source-domain selection is to construct a source domain comprising more diverse photoreactions with various substrate–product examples. Nonetheless, future studies are required to evaluate whether such a sample-weighting mechanism remains effective when the source domain includes reactions with catalytic activity trends significantly different from those of the target EnT reaction, such as photoredox reactions. These discussions are summarized in the corresponding sections of our revised

manuscript.

Reviewer 3:

Comment 11: The manuscript presents an innovative approach to predicting suitable catalysts for new reactions using transfer learning (TL) and domain adaptation (DA) techniques. The authors claim to have successfully transferred knowledge from photocatalytic cross-coupling reactions to [2+2] cycloaddition reactions, achieving satisfactory predictive performance with only ten training data points.

Response 11: We appreciate the favorable comments and valuable suggestions.

Comment 12: 1. The authors claim to achieve satisfactory predictive performance using only ten training data points. This extremely small dataset raises significant concerns about the model's ability to generalize: It is challenging to train a robust machine learning algorithm with such a limited dataset, even with transfer learning. The impressive results reported may be indicative of overfitting rather than true predictive power. The authors should provide additional experiments to demonstrate that their algorithm is not overfitting and has good generalization capabilities. This could include: cross-validation results, performance on a separate, larger test set, and learning curves showing performance as the training set size increases.

Response 12: In the corresponding investigations, we observed that the R^2 score improved from 0.72 to 0.83 by performing feature engineering on the original descriptor set (DFT_FE) to create a new descriptor set (DFT_FE2), using 10 OPSs as the training dataset. However, we also acknowledge that this result might partially reflect overfitting to this specific situation. To address this concern, we carried out ML experiments by testing 100 different patterns of training-test splits and varying the number of training data points (50, 40, 30, 20, and 10) for both the TrAB/DFT_FE and TrAB/DFT_FE2 models, and learning curves were constructed (Figure S5). The results showed that the TrAB/DFT_FE model outperformed the TrAB/DFT_FE2 model in terms of generalization performance. In contrast, unlike the TrAB/DFT_FE2 model, the TrAB/DFT_FE model failed to identify effective OPSs in high probability, when using 10 training data points for the prediction of high-performance OPSs for the alkene photoisomerization reaction. These results indicate that while the TrAB/DFT_FE2 model is inferior to the TrAB/DFT_FE model in terms of generalization performance, it is more useful for a practical catalyst exploration in specific organic reactions. Therefore, we believe that the TrAB/DFT_FE2 model is promising for practical applications of ML for catalyst discovery, but we also recognize the importance of fairly discussing the advantages and limitations of the ML model, as pointed out by this reviewer. To address this, we added a discussion to the revised manuscript (page 18) on the differences in predictive performance between the TrAB/DFT_FE and TrAB/DFT_FE2 models when using small training datasets, as well as the trade-off relationship between generalization performance and specific task performance. As mentioned earlier, we included learning curves to illustrate and compare the generalization performance of these models in the Supporting Information (Figure S5).

Comment 13: 2. The manuscript mentions using TrAdaBoostR2 and claims superior performance compared to other supervised DA methods. However, more details are needed: The authors should provide a clear explanation of how knowledge is transferred from one domain to another. The relationship between features in the source and target domains should be explicitly described. A

comparative analysis with other DA methods, including quantitative results, would strengthen the claims of superiority.

Response 13: We added a description of the mechanism of TrAdaBoostR2 (TrAB), and a comparison with other DA methods to the revised manuscript (pages 10–12). This description also includes details about the data structures in the source and target domains, the general differences in the mechanism of representative DA methods, and possible reasons behind observed results. TrAB combines source- and target-domain datasets and adjusts sample weights by decreasing the weights of instances in the source domain with large errors while increasing the weights of instances in the target domain with large errors, enabling efficient knowledge transfer. TrAB is one of the instance-based DA methods, and in the manuscript, we first compared it with feature-based DA methods such as FA and CORAL. The results showed that TrAB outperformed these feature-based methods (TrAB: Avg $R^2 = 0.65$, FA: Avg $R^2 = 0.55$, CORAL: Avg $R^2 = 0.08$). While instance-based DA, including TrAB, aims to address differences in data distribution through sample weighting, feature-based DA focuses on creating a new shared feature space, meaning that instance-based DA can place greater emphasis on individual data points. In cases of highly homogeneous domain shifts, such as in our study where the input and output structures are nearly identical between the source and target domains, it is reasonable that instance-based DA methods result in higher predictive performance. Thus, we additionally tested BW as another instance-based DA method. While BW demonstrated comparable performance to TrAB (BW: Avg $R^2 = 0.63$), the ability of TrAB to iteratively adjust sample weights seemed to allow it to adapt more flexibly to the target task, leading to slightly better performance.

Comment 14: 3. As a reviewer with expertise in machine learning but limited knowledge in organic chemistry, I found some sections of the manuscript challenging to evaluate fully. It would be beneficial to have a domain expert in organic chemistry review the chemical aspects of this work to ensure the validity and significance of the findings in that context.

Response 14: Thank you very much for your constructive suggestions to improve our ML section.

Comment 15: In addition, it would be helpful to include a discussion on the limitations of this approach and potential areas for future work.

Response 15: We added the section titled “Investigations into applicability and limitations” to the manuscript (pages 19–22). Specifically, we carried out following two investigations. First, we evaluated the applicability of the TrAB approach to photocatalytic cross-coupling reactions, i.e., **CO_a**, **CO_b**, **CO_c**, **CO_d**, **CO_e**, **CS** and **CN** (Figure 6a). The predictive performance was generally high for C–O bond-forming reactions (Avg $R^2 = 0.72$ – 0.85), likely because the source domain contains a sufficient number of reactions with highly similar trends in photocatalytic activity to the target domain. In contrast, while the prediction accuracy for **CS** was lower than that for the other cases (Avg $R^2 = 0.44$), TrAB still outperformed other methods tested. Overall, although alternative methods demonstrated comparable performance in some instances, our approach using TrAB-based TL models consistently exhibited better and more stable predictive performance across all cases evaluated in this study. Second, we evaluated the predictive performance for OPSs not included in the source domain, by removing 30 OPSs from the source domain (Figure 6b). Consequently, the prediction accuracy for OPSs excluded from the source domain was much worse than for those included in the source domain (included OPSs: Avg $R^2 = 0.76$, excluded OPSs: Avg $R^2 = 0.28$). This result suggests that achieving high prediction accuracy across a wide range of targets

depends not only on advancements in ML methods but also on constructing an appropriate database informed by the expertise of organic chemists. This proposal is also now included in this section.

Responses to the reviewers' comments (Noto et al, NCOMMS-24-59374A)

Reviewer 1

Comment 1: The authors have fully addressed my questions, the paper can be recommended for publication.

Response 1: Thank you for your favorable comment and careful review of the manuscript and the discussion.

Comment 2: One small issue, the comment 12 raised by reviewer 3 is an important issue, ML model training with only ten data points is problematic. The authors seem to already address this issue by supplementing Figure S5, however, in the corresponding discussions in revised manuscript, I do not see any corresponding discussion or mentioning of the performance comparison among different number of training data points, rather than the focused discussion on the generation performance comparison between different models.

Response 2: We have added a discussion on the problem of TL when using a small training dataset (page 16), with a particular focus on the influence of the number of training data points. As shown in Figure S5 and Table S22, although the TrAB/DFT_FE2 model trained on ten OPSs exhibited a Max R^2 value that was not significantly different from cases using a larger training dataset (50–20 data points: 0.87–0.82, 10 data points: 0.83), its Avg R^2 value declined substantially (50 data points: 0.71–0.62, 10 data points: 0.54). These results indicate that while TL can help mitigate performance degradation to some extent, ensuring generalization ability remains challenging with such a limited dataset. Nevertheless, developing ML methods trained on an experimentally accessible small dataset that can effectively propose potent catalysts is crucial in organic chemistry, even if this approach is not ideal from a data science perspective. We believe that this addition better justifies our subsequent attempt regarding DA in the alkene photoisomerization, and we appreciate your valuable suggestion.