

Photocatalyst Prediction

Design and Experimental Validation of a Photocatalyst Recommender Based on a Large Language Model

Francis Millward⁺, Michał Kulczykowski⁺, Jay Badland-Shaw, Sara Szymkuc, Rajan Suraksha, Aniket Kumar Srivastawa, Violaine Manet, Máire Griffin, Megan Bryden, Thomas Comerford, Lea Hämmerling, Aminata Mariko, Bartosz A. Grzybowski,* and Eli Zysman-Colman*

Abstract: Utilizing an extensive library of literature on photocatalytic transformations, we disclose the development of a machine learning (ML) model for the recommendation of photocatalysts most suitable for reactions of interest. The model is trained on > 36 000 such literature examples and uses an architecture inspired by the Bidirectional Encoder Representations from Transformer (BERT) large language model. Under cross-validation, it can suggest the “correct” photocatalysts with ~90% accuracy. When experimentally tested on five out-of-box reactions, this algorithm consistently suggested photocatalysts that gave yields competitive to those chosen by human researchers and frequently suggested alternative photocatalysts that are potentially more appealing than the originally selected photocatalyst. Altogether, this platform serves as a valuable tool for researchers undertaking reaction optimization programs. The model is free to use at <https://photocatalysts.grzybowski.org/pl/predict/>.

Introduction

Machine Learning (ML) algorithms are at the heart of the ongoing AI revolution. They are impacting numerous areas of chemical research, from protein design^[1] to synthetic chemistry^[2–7] to catalysis. In the latter context, there have been notable recent studies using ML to design homogeneous catalysts offering improved performance. For instance,

Sigman and coworkers pioneered data-driven workflows and regression models to design catalysts offering improved enantiomeric outcomes^[8–10] and, in some *tour de force* demonstrations, industrially relevant scalability.^[11] ML models to find new ligands have been developed by the Denmark,^[12,13] Ackermann,^[14] and List and Varnek^[15] groups and have been based on both 2D and 3D featurization. Schoenebeck and coworkers extended these efforts to dinuclear catalysts,^[16] whereas Hong and coworkers demonstrated systems that seek metal replacements.^[17]

In parallel to these studies seeking unprecedented catalysts, there is also a desire for ML recommender systems that would suggest the most suitable *known* catalyst for a particular transformation. In a recent study from one of our groups, we described a proof-of-principle model of this kind that used a standard multilayer perceptron (MLP) architecture to propose Mg-based catalysts with an accuracy of around 80%.^[18] This study was accompanied by experiments that validated most—but not all—predictions.

Herein, we disclose the development and experimental validation of an efficient recommender of photocatalysts for visible light-driven organic synthesis reactions. Photocatalysis is now a burgeoning area of research, with rapidly expanding applications in numerous sectors, such as the pharmaceutical industry.^[19–21] ML has seen some uptake in the photocatalysis community; e.g., Glorius et al. developed an ML platform for guiding substrate discovery in energy transfer catalysis,^[22] while Noto, Saito, and coauthors have reported the prediction of organic photocatalyst performance in both [2+2] cycloadditions and the nickel co-catalyzed metallaphotoredox synthesis of phenols.^[23,24]

Still, when it comes to selecting the most suitable photocatalyst for a given reaction during optimization campaigns, thermodynamic parameters such as redox potentials and


[*] F. Millward⁺, J. Badland-Shaw, R. Suraksha, A. K. Srivastawa, V. Manet, M. Griffin, Dr. M. Bryden, Dr. T. Comerford, Dr. L. Hämmerling, A. Mariko, Prof. Dr. E. Zysman-Colman
Organic Semiconductor Centre, EaStCHEM School of Chemistry, University of St Andrews, St Andrews KY16 9ST, UK
E-mail: eli.zysman-colman@st-andrews.ac.uk


Dr. M. Kulczykowski⁺, Dr. S. Szymkuc
Allchemy, Inc., Highland, Indiana 46322, USA

Dr. M. Kulczykowski⁺
Institute of Organic Chemistry, Polish Academy of Science, ul. Kasprzaka 44/52, Warsaw 02–224, Poland

Prof. Dr. B. A. Grzybowski
Center for Algorithmic and Robotized Synthesis (CARS) of Korea's Institute for Basic Science (IBS) and Department of Chemistry, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, South Korea
E-mail: nanogrzybowski@gmail.com

[+] Both authors contributed equally to this work.

 Additional supporting information can be found online in the Supporting Information section

 © 2025 The Author(s). Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

excited-state energies are often not predictive of reaction outcomes, as photocatalysts tend to degrade,^[25] and extensive photocatalyst screening programs are necessary. Furthermore, the motivations for the choice of specific photocatalysts surveyed are rarely discussed, prompting questions about whether the identified conditions are premeditated or opportunistic (according to photocatalyst availability and/or popularity/bias in prior literature^[26]), and whether better photocatalysts for a given reaction of interest can be found. To this end, we disclose here the development of a photocatalyst recommender that suggests photocatalysts likely to catalyze a given reaction.

We first tested an ML recommender akin to our previously disclosed platform^[18]—similar to the said reference, it achieves an accuracy of around 80%. We then sought to improve this metric using a transformer architecture previously used in large language models. By first pre-training this transformer on a large collection of reactions (not necessarily photocatalytic), we taught it to understand the general syntax of molecular representations, which, in turn, proved beneficial in training the ultimate model for photocatalyst prediction. This architecture offers significantly improved accuracy (~90%). This model was then used to suggest alternative photocatalysts for several literature reactions that were described in the literature only after the development of our model and were therefore not in its training/test set. The results of these studies are encouraging in that all photocatalysts suggested by the recommender gave appreciable yields, and in four out of five cases, the model identified at least one photocatalyst that performs competitively with the “optimal” one identified in the original publication or with photocatalysts chosen by a human researcher.

Results and Discussion

Data Curation

We began by collecting literature data for model training. Having identified 31 archetypal photocatalysts (see Supplementary Information and subset in Figure 2), we queried the Reaxys database for reactions using these photocatalysts. Following filtering of the data (see Supporting Information for details), a dataset of 36097 unique reactions remained. It should be noted that the data set is heavily imbalanced, reflecting the prevailing historical bias in the choice of photocatalysts, where some were associated with >7000 reactions, while others with a mere 20. The photocatalysts were one-hot encoded (that is, given unique labels), whereas the reactions were represented in two ways: i) by the concatenated SMILES of the reaction substrates and ii) by the reaction cores (that is, atoms that change their bonding patterns and flanking atoms to within a bond-distance radius of 3) that were extracted after atom-mapping the reactions using our MAPPET algorithm.^[27,28] Altogether, 4854 distinctive reaction cores/types were identified. It should be noted that the model did not incorporate information about the presence or choice of co-catalyst, base, solvent, reaction time, excitation wavelength, photon flux, and other reaction parameters. In

Ref. [18], we showed that automatic extraction of the relevant information from Reaxys is problematic due to errors and differences in nomenclature. In the said reference, we were able to curate some of this data manually by consulting ~1500 source publications—in the present case, however, such manual curation for >36 000 entries would be prohibitive. In a broader context, we^[26,29,30] and others^[31–34] have previously shown that prediction of reaction conditions and/or yields is extremely difficult due to the historical biases in the data and, for yields, scarcity of negative examples.

With these preliminaries and limitations, the overall learning strategy was to use both the substrate and reaction-core SMILES. As we previously showed,^[18] this approach is not redundant, as both the reaction type and the specifics of the reaction (e.g., functional groups present in the substrates outside of the core) independently contribute to the prediction outcome. We tested two types of models, an MLP and a Bidirectional Encoder Representations from Transformer (BERT)-based language model.

The MLP Model

(Figure 1a) used two inputs: the substrates SMILES represented as a standard Morgan fingerprint of size 64 and, for the reaction cores, a Differential Reaction Fingerprint (DRFP)^[35] with a vector size of 64 and radius 3. Each input was processed through a distinct hidden layer, the outputs of which were subsequently concatenated. This concatenated layer was followed by two additional hidden layers and a final output layer, which yielded a 31-dimensional vector of logits corresponding to the numbers/labels of one-hot-encoded photocatalysts present in the dataset. This vector was transformed into the model's final rankings via the *softmax* function. The model's hyperparameters, including the activation function, dimensions of the hidden layers, kernel initializers, and the method for merging the inputs at the hidden layer, were optimized using the Optuna library.

To address dataset imbalance, the model was trained using a weighted cross-entropy loss. Model validation was carried out using a 5-fold cross-validation and leveraging three performance metrics to ensure appropriate training: accuracy, F1 score with weighted averaging, and F1 score with macro averaging. The accuracy and F1 scores were, respectively, $79.0\% \pm 1.6\%$ and $79.1\% \pm 1.6\%$, while F1 with macro averaging was $\sim 71.5\% \pm 1.5\%$. Performance was further evaluated using the top-*k* metric (that assesses the model's ability to predict the ground truth photocatalyst within its top *k* predictions). The correct photocatalyst was chosen by the model as its top prediction (top-1) in 79% of cases, within the top-4 predictions in ~95% of cases, and within the top-10, in 99% of cases.

BERT-Based Model

(Figure 1b). The aforementioned performance of the MLP model was on par with a conceptually similar architecture we recently developed to predict one-hot-encoded Mg-based

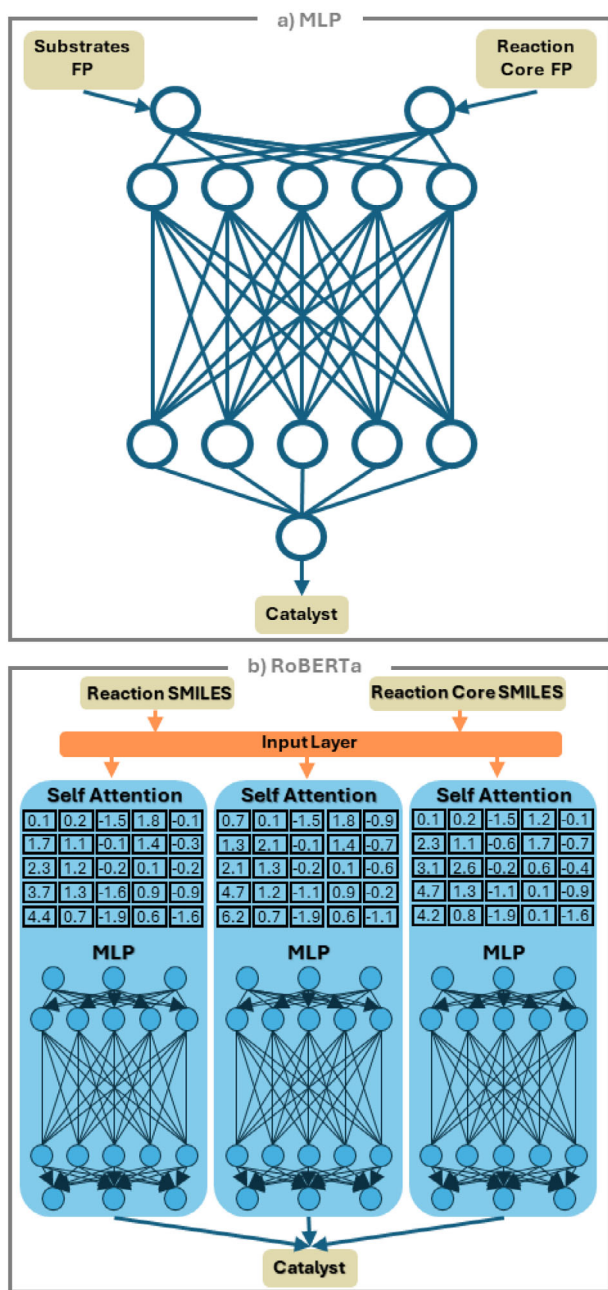


Figure 1. The architectures of a) the multilayered perceptron, MLP, model and b) the language model based on the RoBERTa extension of BERT. The essence of the BERT (Bidirectional Encoder Representations from Transformer) approach is a self-attention mechanism, which, in the case of SMILES strings, enables the model to capture complex relationships between molecular substructures across long distances by trying to put attention on how tokens relate to one another. The bidirectional understanding is particularly valuable for chemical reactions, where reactivity often depends on “interactions” between functional groups that may be separated by multiple atoms. Pre-training on a large, unlabeled dataset helps the model to understand principles of chemistry, which increases performance on supervised fine-tuning tasks where data is scarce and allows for better generalization to unseen data during inference.

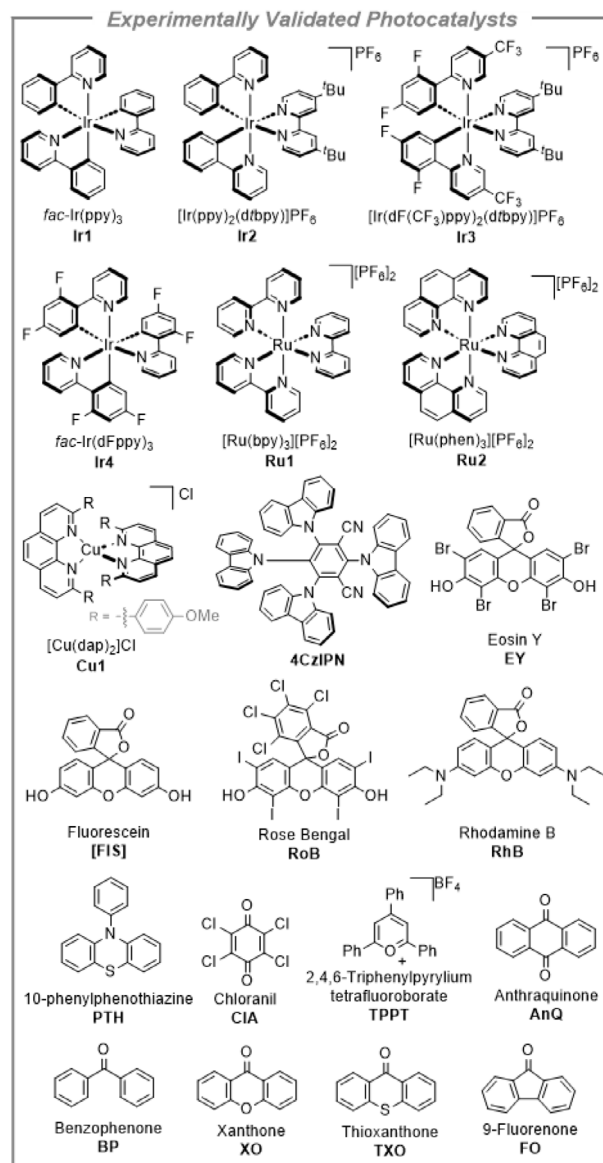


Figure 2. A subset of 31 photocatalysts used to train the ML model (for the full set, please consult the Supporting Information). For any given reaction, the model always ranks all 31 catalysts. Here, for each of the reactions committed to experimental validation, the top-5 suggestions were taken. Across five reactions tested, 19 catalysts shown here were among the top-5 choices; the 20th catalyst, Ir3, was never suggested among the top-5 but was used as a control chosen by human researchers.

catalysts most suitable for a given reaction.^[18] However, we sought to improve these metrics and therefore considered another approach specifically designed to learn language representations, here the syntax of SMILES strings. This model uses RoBERTa,^[36] a variant of Google’s BERT^[37] that is fine-tuned for optimal hyperparameters. BERT, an advanced ML model for natural language processing tasks, introduces a transformer-based architecture that leverages bidirectional training. This innovative approach contrasts with traditional unidirectional models by considering contextual influences from both the “left” and the “right” directions

within a given text. RoBERTa, an enhancement developed by Facebook AI, refines BERT's performance further. It uses architecture illustrated in Figure 1b and was pre-trained on 1715395 reactions taken mostly from the Open Reaction Database (<https://open-reaction-database.org/>) and standardized using the MolVS library. Analogously to the photocatalyst dataset, reaction cores were extracted and combined with reaction SMILES, forming a two-sentence input to the model. The ByteTokenizer, trained on this dataset, tokenized the input SMILES into distinctive classes from the trained tokenizer dictionary. Each token was assigned an integer value corresponding to that token text value in the vocabulary. The input vector dimension was set to 1024 elements, with 15% of all inputs not representing special tokens statically masked. The model was then trained over two epochs to predict these missing parts and reconstruct the ground-truth SMILES. This pre-training was important for the model to learn the "grammar" of correct SMILES describing reactions and pinpointing elements of the reactants that undergo reaction, thus gaining "intuition" about the reaction process (see also caption to Figure 1b).

Subsequently, the model underwent fine-tuning to perform the actual task of interest—that is, to assign each of the reactions from the original 36097 set to one of the 31 photocatalysts. This fine-tuning phase spanned 10 epochs. Under five-fold cross validation, this pre-trained model performed significantly better than MLP: both accuracy and F1 were $90.0\% \pm 0.4\%$ while F1 with macro averaging was $\sim 84\% \pm 0.6\%$. The top-k statistics also improved, as the ground-truth catalyst was the top-1 prediction in 89.99% of cases, within the top-2 predictions in 95.7% of cases, and within the top-7 in 99% of cases.

WebApp Predictor

To facilitate wider use of the model, we developed a web application freely available at <https://photocatalysts.grzybowski-group.pl/predict/>. The WebApp is straightforward to use. The SMILES code for a given reaction is copied into the search box, and the model reduces the scheme to a simplified reaction group SMILES. Only the starting substrates and product, separated by a reaction arrow, are required; considerations of co-catalysts and other additives are not included in the model's design (see Supporting Information). The top-5 photocatalysts are then suggested, ranked by the model's probability/confidence rating that the photocatalyst will successfully catalyze the reaction. It is important to note that the confidence rating is an internal metric of the model and, as demonstrated below, does not correlate with experimental yields, though every top-5 recommendation made by the recommender was effective in catalyzing the reactions discussed in the remaining part of the paper.

Experimental Validation

We proceeded to validate the model. In doing so, we note that the model will suggest photocatalysts for *any and all* reactions,

even those that cannot be performed photocatalytically—this is a well-known limitation of models in which catalysts are one-hot-encoded. Accordingly, we focus on literature-precedented reactions that were previously shown to be suitable for photocatalysis but were not included in the training-set data. The model was trained on reactions published up until the end of 2021; therefore, reactions for the experimental screen used publications from 2022 or later. We then assessed whether the ML model would suggest photocatalysts that are i) competitive (that is, equivalent or higher yielding) with the optimal photocatalyst identified in the original publication and ii) better than photocatalysts that were not among the model's top-five suggestions but were selected by human researchers. The human researcher photocatalyst selections included at least two of the following photocatalysts: **4CzIPN**, $[\text{Ir}(\text{dF}(\text{CF}_3)\text{ppy})_2\text{dtbbpy}](\text{PF}_6)$ (**Ir3**), *fac*- $\text{Ir}(\text{ppy})_3$ (**Ir1**), or $[\text{Ru}(\text{bpy})_3](\text{PF}_6)_2$ (**Ru1**). These photocatalysts are widely employed within screening campaigns and are all commercially available, and were thus included in the trials to benchmark the performance of the ML model. This ML model only suggests the ionic form of charged photocatalysts and does not consider the counterion; in situations where these were suggested, the most commonly used counterions for the photocatalysts were employed. SMILES for the validation reactions are provided in the Supporting Information as example inputs for using the WebApp.

ATRA Reaction

We began by exploring the atom transfer radical addition (ATRA) reaction of phenylsulfonyl chloride to potassium allyltrifluoroborate, Figure 3.^[38] The original paper evaluated **4CzIPN**, **Ir3**, $[\text{Ir}(\text{ppy})_2\text{dtbbpy}](\text{PF}_6)$ (**Ir2**), Eosin Y (**EY**), and **Ir1**, identifying the latter as the optimal photocatalyst in their setup.^[38] Our algorithm ranked **Ir1** among the top five hits for this reaction; however, it was more confident in the use of both **Ru1** and **EY**, while **Ir2** and $[\text{Cu}(\text{dap})\text{Cl}_2$ (**Cu1**) were also suggested. All five photocatalysts suggested by the model promoted the reaction, affording the desired product in 48%–72% yield (Figure 3). Most notably, the top-ranked **Ru1**, which was not screened in the original publication, gave the highest yield of the photocatalysts tested. The use of **EY**, a more sustainable organic photocatalyst, ranked by the ML model as the second most likely to work, was also productive using our photoreactors, achieving comparable yields to the more expensive **Ir1**. Interestingly, the authors reported that **EY** did not work as a photocatalyst, which is in contrast to our experiments and highlights the impact of the reaction setup (e.g., light source, photon flux, etc.) in impacting reaction yields. We also assessed **Ir3** and **4CzIPN**, which are human researcher choices, as they are two of the most popularly screened photocatalysts for transformations requiring balanced redox potentials and are frequently included in photocatalyst library screens. Both performed adequately (each affording yields of 46%), though they performed worse than all of the ML-predicted catalysts. This first reaction clearly showcases the potential utility of the disclosed ML model: the suggested

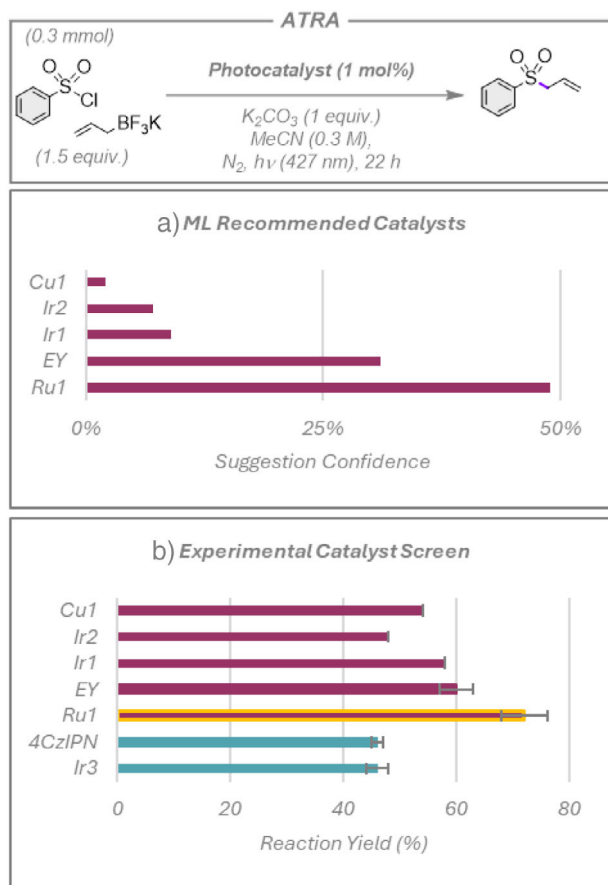


Figure 3. ATRA reaction with phenylsulfonamide. a) Top five photocatalysts suggested by the model and ranked by confidence rating. b) Experimental yields. Red bars show the photocatalysts suggested by the model, with the yellow outline highlighting a previously untested yet productive photocatalyst, **Ru1**, which the ML model identified. Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative 1H NMR experiment repeats, using 1,3,5-trimethoxybenzene as an internal standard, with the error bars indicating the standard deviation.

photocatalysts all worked, and the ML model identified **Ru1**, which outperformed all other photocatalysts screened, and was not considered in the photocatalyst optimization program in the initial publication.^[38]

We next probed how the choice of substrate could influence the ML model's recommendation. The same five photocatalysts were suggested for *p*-toluenesulfonyl chloride, or *p*-trifluoromethylphenylsulfonamide, Figure 4; however, there was a significant difference in the order and confidence ratings outputted by the ML for each of these substrates. Again, **Ru1** afforded the highest experimental yields of 70% and 58%, further evidencing the value of utilizing this ML model for photocatalyst selection.

Phosphorylation

We next explored the photocatalyzed phosphorylation of tertiary amines, using triethylamine as a model substrate,

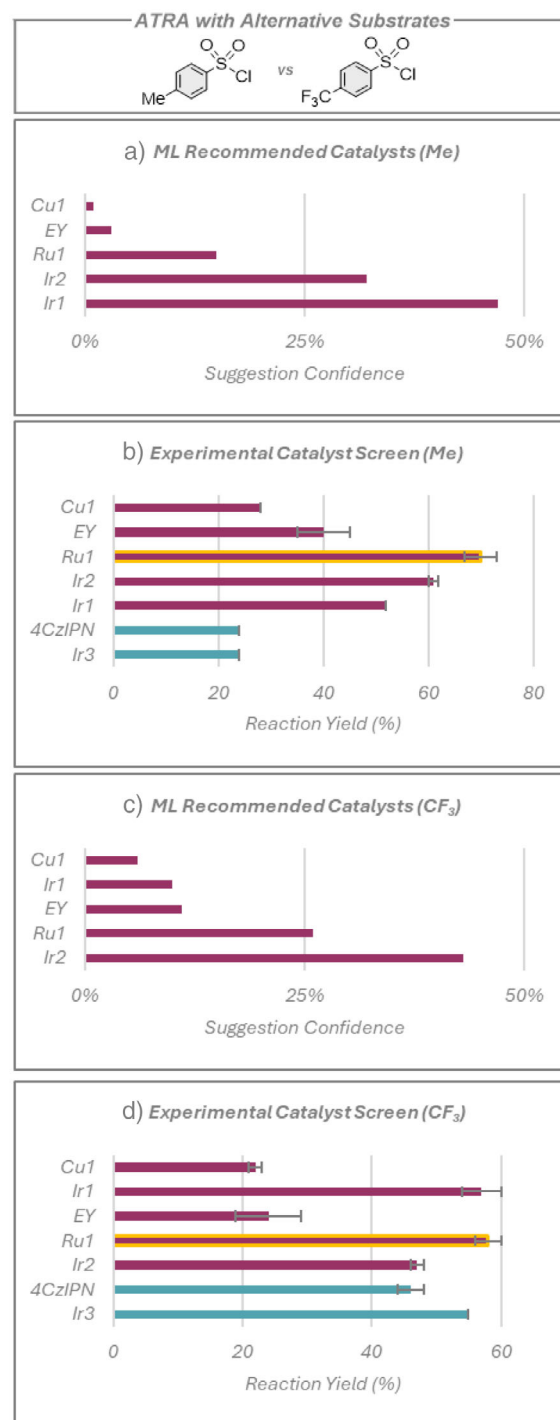


Figure 4. ATRA reaction using alternative substrates. a) Top five RoBERTa-suggested photocatalysts ranked by confidence rating with *p*-toluenesulfonyl chloride. b) Experimental yields with *p*-toluenesulfonyl chloride. a) Top five photocatalysts suggested by the ML model and ranked by confidence rating with *p*-trifluoromethylphenylsulfonamide. b) Experimental yields with *p*-trifluoromethylphenylsulfonamide. Red bars show the photocatalysts suggested by the model, with the yellow outline highlighting a previously untested yet highly productive photocatalyst, **Ru1**, which the ML model identified. Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative 1H NMR experiment repeats, using 1,3,5-trimethoxybenzene as an internal standard, with the error bars indicating the standard deviation.

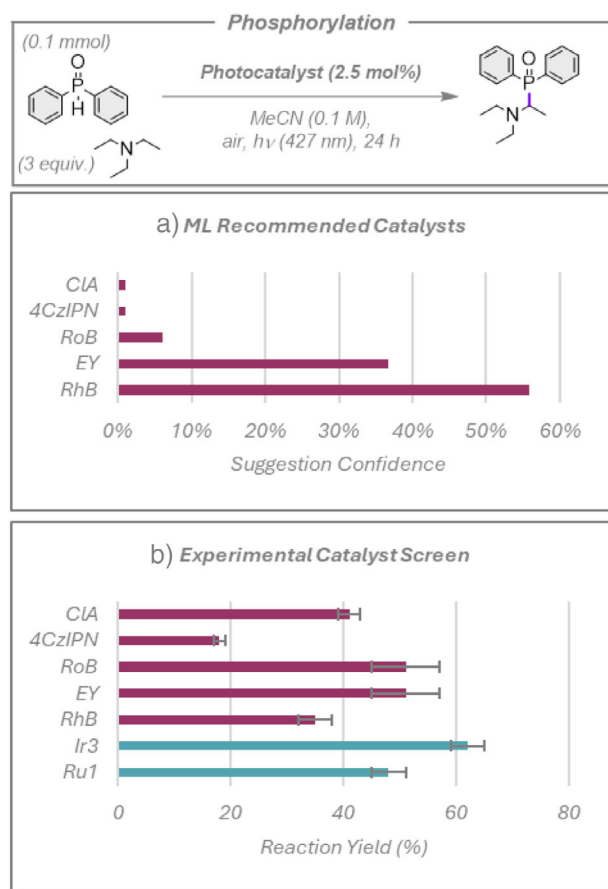


Figure 5. Photocatalyzed phosphorylation of triethylamine. a) Top five photocatalysts suggested by the model and ranked by confidence rating. b) Experimental yields. Red bars show the catalysts suggested by the model. Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative ^1H NMR experiment repeats, using 1,3,5-trimethoxybenzene as an internal standard, with the error bars indicating the standard deviation.

Figure 5.^[39] The authors used a bespoke photocatalyst, $[\text{Ru}(\text{CF}_3\text{-bpy})_2(\text{OMe-bpy})](\text{PF}_6)_2$, which was not accessible to us nor included in the library of photocatalysts available to the ML model; thus, this photocatalyst was not tested. The algorithm heavily favored organic photocatalysts in its recommendation for this phosphorylation reaction, providing high confidence ratings to **EY** and Rhodamine B (**RhB**) and lower confidence suggestions to chloranil (**CIA**), **4CzIPN**, and Rose Bengal (**RoB**). All five photocatalysts promoted the reaction in 35%–51% (Figure 5), which is to be expected, as the key photocatalytic step in this reaction is the oxidation of the tertiary amine ($E_{\text{ox}}(\text{triethylamine}) = 0.83$ V versus SCE in MeCN).^[40] We also assessed **Ir3** and **Ru1** as the human researcher choices, which can both also oxidize triethylamine, with the former giving the highest yield of 62% of the seven photocatalysts tested. While the ML model consistently suggested viable photocatalysts for this reaction, the current version did not identify the photocatalyst that produced the highest yield.

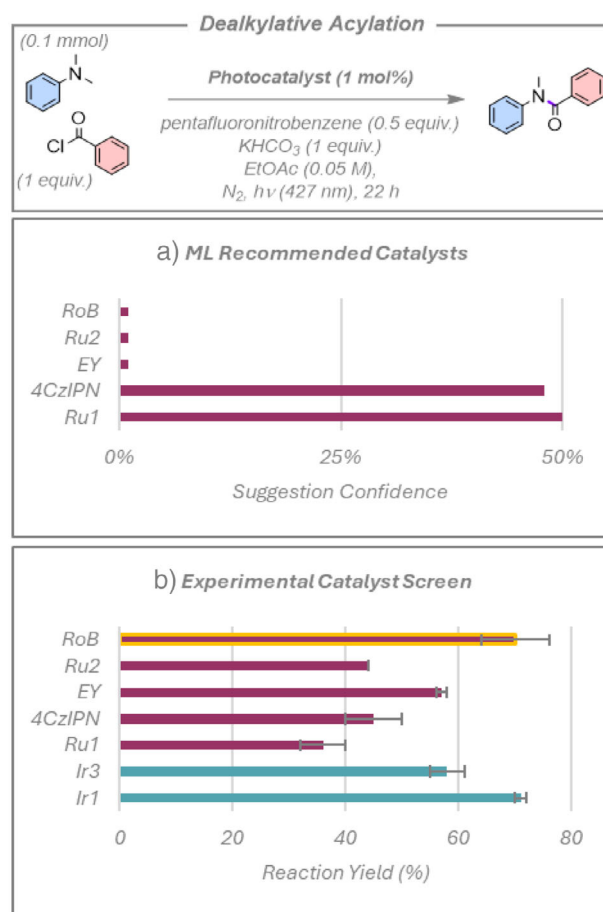


Figure 6. Photocatalyzed dealkylative acylation. a) Top five photocatalysts suggested by the model and ranked by confidence rating. b) Experimental yields. Red bars show the photocatalysts suggested by the model, with the yellow outline highlighting a previously untested yet productive photocatalyst, **RoB**, which the ML model identified. Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative ^1H NMR experiment repeats, using 1,4-bis(trimethylsilyl)benzene as an internal standard, with the error bars indicating the standard deviation.

Dealkylative Acylation

We next explored the dealkylative amide formation from alkyl amines, Figure 6.^[41] The proposed reaction mechanism involves the in situ formation of a secondary amine, driven by pentafluoronitrobenzene acting as a hydrogen atom transfer (HAT) agent, which is then trapped by benzoyl chloride. Once again, the authors used a bespoke photocatalyst, Cz-NI-Ph, that was not available to us or included in the ML model and was therefore not tested. As shown in Figure 6, the algorithm gave a high confidence rating for **4CzIPN** and **Ru1**, while also suggesting **RoB**, $[\text{Ru}(\text{phen})_3](\text{PF}_6)_2$ (**Ru2**), and **EY**, albeit with much lower confidence ratings. All five proposed photocatalysts are viable, with **RoB** affording a high yield of 70%. We also tested **Ir1** and **Ir3** as the human researcher choices, both of which also performed well; in fact, **Ir1** gave a comparable yield to **RoB** of 71%. However, it is important to note that the principal photochemistry in this reaction

is being performed by pentafluoronitrobenzene, acting as a strong electron acceptor and proposed HAT agent that is compatible with all of the photocatalysts, of which the ML has no knowledge or mechanistic insight. From the perspective of reaction development, the current ML model therefore has limited value, as there is no included thermodynamic discriminator in terms of the optoelectronic properties of the photocatalysts, nor an understanding of the mechanistic importance of co-additives that are needed for realizing the reaction. However, the ML model did identify **RoB** as a competitive photocatalyst, one that is typically not included in photocatalyst screening in the literature, and it performed exceptionally well in this reaction, on par with **Ir1**.

Aldehyde to Nitrile Conversion

The final electron transfer reaction explored was the synthesis of benzonitrile from benzaldehyde using TEMPO as an organic co-catalyst, Figure 7.^[42] For this reaction, the original publication identified **4CzIPN** as the optimal photocatalyst during their photocatalyst screen. The top two photocatalysts recommended by the ML model (see Figure 7), **Ru1** and 2,4,6-triphenylpyrylium tetrafluoroborate (**TPPT**), performed well (48% and 69%, respectively), but returned somewhat lower yields than **4CzIPN** (84%). The model's suggestions of anthraquinone (**AnQ**) and 10-phenylphenothiazine (**PTH**) also gave comparably high yields of 65 and 73%, respectively. However, the model's suggestion of fluorescein (**FIS**), which the authors notably did not screen during the reaction development, gave comparable yields to **4CzIPN** (85%). This is a remarkable outcome for the model, as **FIS** is significantly cheaper than **4CzIPN**. However, we should again note that the reaction is driven primarily by the TEMPO co-catalyst, of which the ML has no knowledge and does not factor into its decision-making algorithm. While this means the model could not have been used to develop the initial reaction protocol during the reaction discovery phase, it clearly has demonstrated value during the optimization phase to identify superior photocatalysts for this transformation.

Energy Transfer Example

Finally, we examined the performance of the ML model in a photoinduced energy transfer reaction, the [1,3]-sigmatropic shift of (S)-verbenone to chrysanthenone, Figure 8.^[43] For this transformation, the ML model suggested predominantly organic diarylketone-based photosensitisers, including benzophenone (BP), xanthone (XO), thioxanthone (TXO), and 9-fluorenone (FO), as well as suggesting *fac*-Ir(dFppy)₃ (Ir4). Due to the onset of the absorbance spectra of the diarylketone photocatalysts, a 390 nm light source was used to drive the reaction (at this excitation wavelength, an inefficient, 19% yield reaction is observed in the absence of any catalysts). Once again, **4CzIPN** and **Ir3** were chosen as the human researcher choices to test alongside the model's suggestions, and both produced moderate yields of 49% and 56%, respectively. All of the model's suggested photocatalysts

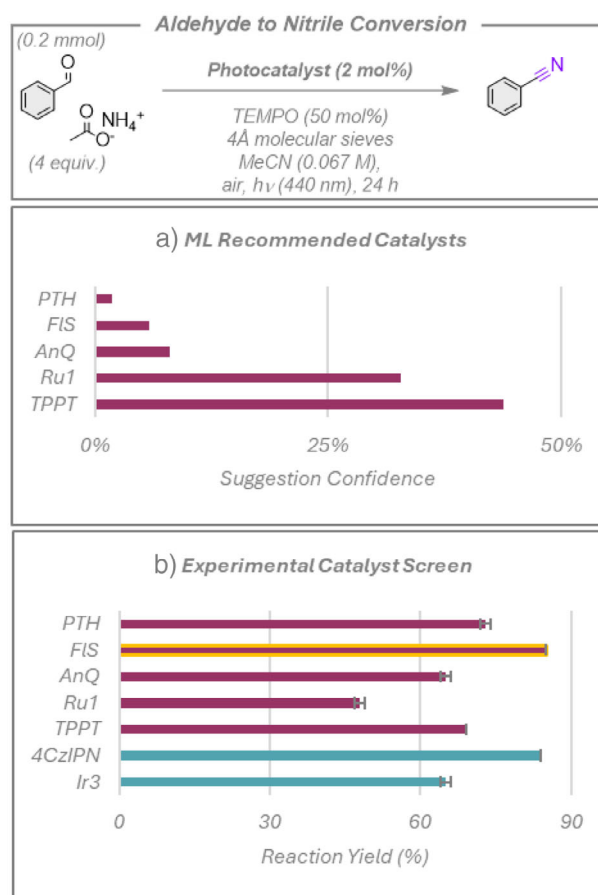


Figure 7. Photocatalyzed aldehyde to nitrile conversion. a) Top five photocatalysts suggested by the model and ranked by the confidence rating. b) Experimental yields. Red bars show the photocatalysts suggested by the model, with the yellow outline highlighting a previously untested yet highly productive photocatalyst, **FIS**, which the ML model identified. Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative ¹H NMR experiment repeats, using 1,3,5-trimethoxybenzene as an internal standard, with the error bars indicating standard deviations.

outperformed these “go-to” examples, with the diarylketone-based ones achieving 65%–74% yield, while **Ir4** afforded an excellent 91% yield. Diarylketone-based photocatalysts are known to undergo deactivation via pinacol homo-coupling,^[44] which may explain why they performed less efficiently than **Ir4** at the 2 mol% catalyst loading used for this reaction. In a broader context, because of the high triplet energy of the starting material ($E_T = 3.0$ eV),^[43] it could be expected that photocatalysts possessing similarly high E_T should perform more efficiently in the reaction than those whose E_T was much lower in energy. Therefore, it is not surprising that **4CzIPN** and **Ir3** ($E_T = 2.53$ and 2.68 eV, respectively) were less efficient photocatalysts than those suggested by the ML model ($E_T = 2.75$, 3.00, 3.22, and 2.84 eV for **Ir4**, **BP**, **XO**, and **TXO**, respectively).^[45] What is surprising, however, is that the model—despite having no prior knowledge of the E_T of either the photocatalysts or substrates—exclusively suggested photocatalysts that were more likely to facilitate

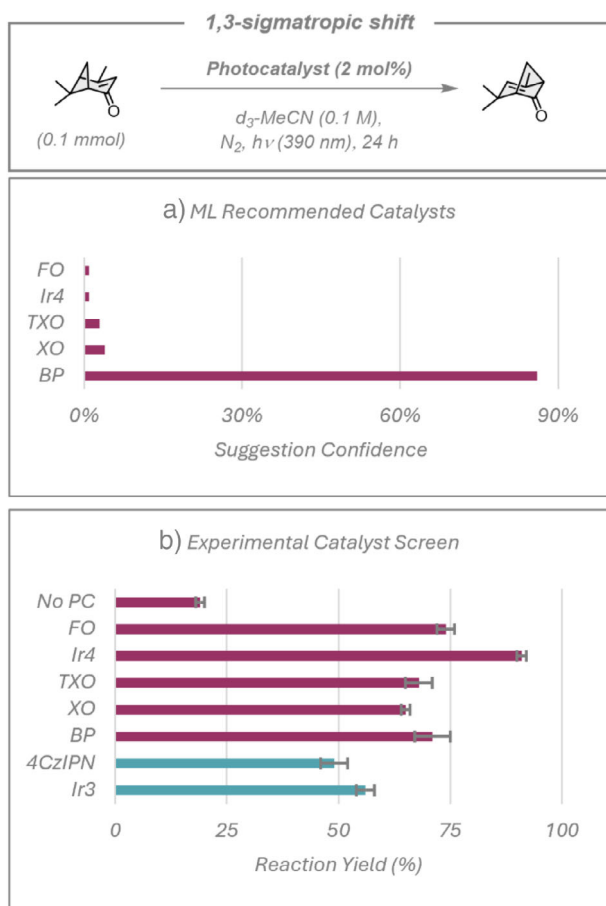


Figure 8. [1,3]-sigmatropic shift of (S)-verbenone to chrysanthenone. a) Top five photocatalysts suggested by the model and ranked by confidence rating. b) Experimental yields. Red bars show the catalysts suggested by the model (and a “No PC” control without photocatalyst). Blue bars represent additional photocatalysts chosen by human researchers. Yields presented are the average of two separate quantitative 1H NMR experiment repeats, using mesitylene as an internal standard, with the error bars indicating the standard deviation.

favorable Dexter energy transfer to the substrate. While an experienced photochemist likely would have predicted this result, a researcher who has more limited experience in this field may have unnecessarily devoted resources to testing catalysts such as 4CzIPN and Ir3 as part of a photocatalyst screening campaign.

Conclusions

The conclusions derived from these experiments are nuanced. In all of the reactions screened, every photocatalyst suggested by the ML model was productive, although rankings within each set of suggested photocatalysts do not correlate with experimental yields and cannot be used to infer which photocatalyst is optimal. ML-suggested photocatalysts afforded yields that were consistently better than or on par with the photocatalysts chosen for a particular reaction by human researchers who selected only popular photocatalysts such as 4CzIPN or Ir3 in four out of the five reactions

tested. We have also compared the outputs of the ML model to three other complex reaction systems from the recent literature, where the ML model selected similar photocatalysts to those chosen by expert photochemists during their screening programs (see Supporting Information for this discussion). From a practical point of view, one of the main advantages of the recommender is that it is not biased toward particular photocatalysts that human researchers consider “worth screening.” Here, the algorithm frequently suggested less popular photocatalysts that performed the reactions comparably well to the more widely used and more expensive “go-to” photocatalysts (see Figure S1). For example, in the ATRA reaction, the original publication did not screen **Ru1**, which was suggested by the recommender, and this was, in fact, the best-performing photocatalyst. In the dealkylative acylation, the ML model suggested **RoB** (albeit with a low confidence rating), and this turned out to be one of the best-performing photocatalysts; notably, this is typically no longer a commonly screened photocatalyst. Similarly, in the aldehyde to nitrile conversion reaction, **FIS** performed as well as the much more popular 4CzIPN.

While the ML model shows considerable promise as a tool to rapidly suggest photocatalysts, in its current form, it is not suited toward reaction discovery, where the selection of co-catalysts, additives, solvents, and other reaction parameters is often more important than the choice of photocatalyst in optimizing the reaction yield, and this is a current limitation of the model; furthermore, there is no correlation between the confidence rating from the model and reaction yields. Overall, we suggest that in its current form, the recommender acts as an essential “sense check” tool to suggest the most promising photocatalysts to include in a photocatalyst screen, and we suggest that all photocatalysts recommended by the ML model be tested. Future versions of this ML tool will include thermodynamic and kinetic parameters of both photocatalysts and substrates, as well as excitation wavelengths, which will lead to stronger correlations between the confidence rating of the recommendation and the product yield. The model would also benefit from interrogating a more expansive photocatalyst library beyond the original 31 selected and an expanded library of reactions for each photocatalyst. Nonetheless, the ML model presented here may already act as a valuable tool for photocatalysis researchers.

Supporting Information

Supporting Information provides additional theoretical and experimental details. All model code and data is now made available at: <https://github.com/GlockPL/Photocatalyst-Recommender-based-on-a-Large-Language-Model>.

The research data supporting this publication can be accessed at <https://doi.org/10.17630/21981110-a42c-44cd-91aa-b5c52a60a7b7>.

Author Contributions

M.K. and S.S. curated the data and developed the recommender models. F.M. co-wrote the manuscript and directed

the experimental validation efforts. J. B.-S. and R.S. each conducted screening for two of the reactions. A.K.S. conducted screening for one reaction. V.M. scouted potential reactions. M.G. contributed to experimental data analysis, initial data set sorting, and designed the Table of Contents. M.B., T.C., L.H., A.M. contributed to initial data set sorting. B.A.G. and E.Z.-C. conceived and supervised the project, and co-wrote the manuscript.

Acknowledgements

M.K. and S.S. were supported by Allchemy, Inc. Analysis of results and writing of the paper by B.A.G. was supported by the Institute for Basic Science, Korea (project code IBS-R020-D1). The UK authors thank the Leverhulme Trust for support (RPG-2023–110), the Engineering and Physical Sciences Research Council for funding (EP/W007517/1, EP/W015137/1, EP/Z535291/1, and EP/W52259/1), the European Commission (PhotoReAct ITN: 956324), Syngenta and Johnson Matthey for support. F.M. M.A.B., and T.C. thank the EaSI-CAT CDT at the University of St Andrews for support in the form of a studentship.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Pure at <https://doi.org/10.17630/21981110-a42c-44cd-91aa-b5c52a60a7b7>, reference number 321347017.

Keywords: Large language models • Machine learning • Photocatalysis

- [1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, et al., *Nature* **2021**, *596*, 583–589, <https://doi.org/10.1038/s41586-021-03819-2>.
- [2] F. Strieth-Kalthoff, S. Szymkuć, K. Molga, A. Aspuru-Guzik, F. Glorius, B. A. Grzybowski, *J. Am. Chem. Soc.* **2024**, *146*, 11005–11017.
- [3] B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83–88, <https://doi.org/10.1038/s41586-020-2855-y>.
- [4] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281–1289, <https://doi.org/10.1021/acs.accounts.8b00087>.
- [5] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Touthkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem.* **2018**, *4*, 522–532, <https://doi.org/10.1016/j.chempr.2018.02.002>.
- [6] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610, <https://doi.org/10.1038/nature25978>.
- [7] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.
- [8] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348, <https://doi.org/10.1038/s41586-019-1384-z>.
- [9] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217, <https://doi.org/10.1021/jacs.1c09718>.
- [10] M. E. Akana, S. Tcyrulnikov, B. D. Akana-Schneider, G. P. Reyes, S. Monfette, M. S. Sigman, E. C. Hansen, D. J. Weix, *J. Am. Chem. Soc.* **2024**, *146*, 3043–3051, <https://doi.org/10.1021/jacs.3c09554>.
- [11] J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Püntener, K. A. Mack, M. S. Sigman, *J. Am. Chem. Soc.* **2023**, *145*, 110–121, <https://doi.org/10.1021/jacs.2c08513>.
- [12] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631, <https://doi.org/10.1126/science.aau5631>.
- [13] N. I. Rinehart, R. K. Saunthwal, J. Wellauer, A. F. Zahrt, L. Schlemper, A. S. Shved, R. Bigler, S. Fantasia, S. E. Denmark, *Science* **2023**, *381*, 965–972, <https://doi.org/10.1126/science.adg2114>.
- [14] X. Hou, S. Li, J. Frey, X. Hong, L. Ackermann, *Chem.* **2024**, *10*, 2283–2294, <https://doi.org/10.1016/j.chempr.2024.03.027>.
- [15] N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, B. List, *Angew. Chem. Int. Ed.* **2023**, *62*, e202218659.
- [16] J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen, F. Schoenebeck, *Science* **2021**, *374*, 1134–1140, <https://doi.org/10.1126/science.abj0999>.
- [17] X.-Y. Xu, L.-G. Liu, L.-C. Xu, S.-Q. Zhang, X. Hong, *J. Am. Chem. Soc.* **2025**, *147*, 15318–15328, <https://doi.org/10.1021/jacs.5c00838>.
- [18] P. Baczewska, M. Kulczykowski, B. Zambroń, J. Jaszczewska-Adamczak, Z. Pakulski, R. Roszak, B. A. Grzybowski, J. Mlynarski, *Angew. Chem. Int. Ed.* **2024**, *63*, e202318487, <https://doi.org/10.1002/anie.202318487>.
- [19] L. Candish, K. D. Collins, G. C. Cook, J. J. Douglas, A. Gómez-Suárez, A. Jolit, S. Keess, *Chem. Rev.* **2022**, *122*, 2907–2980, <https://doi.org/10.1021/acs.chemrev.1c00416>.
- [20] A. Y. Chan, I. B. Perry, N. B. Bissonnette, B. F. Buksh, G. A. Edwards, L. I. Frye, O. L. Garry, M. N. Lavagnino, B. X. Li, Y. Liang, E. Mao, A. Millet, J. V. Oakley, N. L. Reed, H. A. Sakai, C. P. Seath, D. W. C. Macmillan, *Chem. Rev.* **2022**, *122*, 1485–1542, <https://doi.org/10.1021/acs.chemrev.1c00383>.
- [21] D. F. Fernández, M. González-Esguevillas, S. Keess, F. Schäfer, J. Mohr, A. Shavnya, T. Knauber, D. C. Blakemore, D. W. C. MacMillan, *Org. Lett.* **2024**, *26*, 2702–2707, <https://doi.org/10.1021/acs.orglett.3c00994>.
- [22] L. Schlosser, D. Rana, P. Pflüger, F. Katzenburg, F. Glorius, *J. Am. Chem. Soc.* **2024**, *146*, 13266–13275, <https://doi.org/10.1021/jacs.4c01352>.
- [23] N. Noto, R. Kunisada, T. Rohlf, M. Hayashi, R. Kojima, O. García Mancheño, T. Yanai, S. Saito, *Nat. Commun.* **2025**, *16*, 3388, <https://doi.org/10.1038/s41467-025-58687-5>.
- [24] N. Noto, A. Yada, T. Yanai, S. Saito, *Angew. Chem. Int. Ed.* **2023**, *62*, e202219107, <https://doi.org/10.1002/anie.202219107>.
- [25] M. A. Bryden, F. Millward, O. S. Lee, L. Cork, M. C. Gather, A. Steffen, E. Zysman-Colman, *Chem. Sci.* **2024**, *15*, 3741–3757, <https://doi.org/10.1039/D3SC06499A>.

- [26] W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *J. Am. Chem. Soc.* **2022**, *144*, 4819–4827, <https://doi.org/10.1021/jacs.1c12005>.
- [27] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, B. A. Grzybowski, *Nat. Commun.* **2019**, *10*, 1434, <https://doi.org/10.1038/s41467-019-09440-2>.
- [28] S. Szymkuć, T. Badowski, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2021**, *60*, 26226–26232.
- [29] S. Szymkuć, A. Wołos, R. Roszak, B. A. Grzybowski, *Nat. Commun.* **2024**, *15*, 10286.
- [30] G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 3582.
- [31] F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, F. Glorius, *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647, <https://doi.org/10.1002/anie.202204647>.
- [32] M. Fitzner, G. Wuitschik, R. Koller, J.-M. Adam, T. Schindler, *ACS Omega* **2023**, *8*, 3017–3025.
- [33] Z. Liu, Y. S. Moroz, O. Isayev, *Chem. Sci.* **2023**, *14*, 10835–10846, <https://doi.org/10.1039/D3SC03902A>.
- [34] M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle, N. V. Chawla, O. Wiest, *Chem. Sci.* **2023**, *14*, 4997–5005, <https://doi.org/10.1039/D2SC06041H>.
- [35] D. Probst, P. Schwaller, J.-L. Reymond, *Digital Discovery* **2022**, *1*, 91–97, <https://doi.org/10.1039/D1DD00006C>.
- [36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Yoshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanow, **2019**, arXiv:1907.11692.
- [37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *Proc. Conf. N. Am. Chapt. Assoc. Comput. Ling.: Human Lang. Tech.* **2019**, *1*, 4171–4186.
- [38] S.-P. Liu, Y.-H. He, Z. Guan, *J. Org. Chem.* **2023**, *88*, 11161–11172, <https://doi.org/10.1021/acs.joc.3c01124>.
- [39] Z. Mahmood, J. He, S. Cai, Z. Yuan, H. Liang, Q. Chen, Y. Huo, B. König, S. Ji, *Chem. - Eur. J.* **2023**, *29*, e202202677.
- [40] D. Nicewicz, H. Roth, N. Romero, *Synlett* **2015**, *27*, 714–723, <https://doi.org/10.1055/s-0035-1561297>.
- [41] C. Liu, H.-N. Chen, T.-F. Xiao, X.-Q. Hu, P.-F. Xu, G.-Q. Xu, *Chem. Commun.* **2023**, *59*, 2003–2006, <https://doi.org/10.1039/D2CC05842A>.
- [42] X. He, Y.-W. Zheng, B. Chen, K. Feng, C.-H. Tung, L.-Z. Wu, *Sci. China Chem.* **2023**, *66*, 2852–2857, <https://doi.org/10.1007/s11426-023-1748-4>.
- [43] L. Schmid, F. Glaser, R. Schaer, O. S. Wenger, *J. Am. Chem. Soc.* **2022**, *144*, 963–976, <https://doi.org/10.1021/jacs.1c11667>.
- [44] B. Abadie, D. Jardel, G. Pozzi, P. Toullec, J. M. Vincent, *Chem. - Eur. J.* **2019**, *25*, 16120–16127, <https://doi.org/10.1002/chem.201904111>.
- [45] S. Dutta, J. E. Erchinger, F. Strieth-Kalthoff, R. Kleinmans, F. Glorius, *Chem. Soc. Rev.* **2024**, *53*, 1068–1089, <https://doi.org/10.1039/D3CS00190C>.

Manuscript received: July 03, 2025

Revised manuscript received: October 30, 2025

Manuscript accepted: November 27, 2025

Version of record online: ■ ■ ■ ■ ■