

Roboticized AI-assisted microfluidic photocatalytic synthesis and screening up to 10,000 reactions per day

Corresponding Author: Professor Qun Fang

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)

In this work, the authors have developed a robotic system using microfluidic liquid-core waveguide and AI-assisted techniques to perform ultra-high-throughput chemical synthesis and screening of photocatalytic reactions, achieving up to 10,000 reactions per day and significantly reducing reaction times from hours to seconds. This is a nice work with that can be insightful for the development of autonomous experimental workflow.

I'd be happy to recommend this for publication in Nature Communications, once these few things are addressed:

1. The methods section contains some information on how the descriptors were generated. However, there is no information on model hyper-parameters used in Figures 3A2, Figures 4F&G, and how they were selected. These include the learning rates, max-depths, regularization terms, etc. Please elaborate.

2. t-SNE is generally not considered reasonable for feature selection and is rather used for visualization and understanding the structure of high-dimensional data. How was 15 decided as the number of features? Please elaborate. Model performance is quite high in the current setting by typically RFE methods are used for choosing the most promising uncorrelated features.

3. To allow reproducibility of the work, please provide the source code on github.

Reviewer #2

(Remarks to the Author)

The manuscript by Jia-Min Lu, Hui-Feng Wang, Qi-Hang Guo et al. presents a roboticized platform for high-throughput experimentation and screening in organic chemistry.

In general, the manuscript presents a highly innovative, state-of-the-art system that addresses one of the main problems in organic chemistry and catalysis research, namely the impossibility of covering the vast space of organic chemistry. Additionally, such vast data-collection efforts are a requisite for the advancement of machine learning in chemistry.

To this end, the authors use advanced microfluidics (liquid-core waveguide techniques) combined with photocatalysis to shorten reaction times dramatically from hours to seconds. This is then coupled with UV-vis spectroscopy and integrated in the iChemFoundry platform for screening and handling of conditions and reagents. Then, they push the system by moving to non-steady state measurements, which would shorten the time spent per reaction even more. To assist them in processing the non-steady state data, they use standard machine learning (ML) regressors.

These latter efforts, unlike the rest of the manuscript, are not convincing for me. Thus, I'd have some major comments that I would want to see addressed before publication:

- The specific task at hand for the ML is never clearly stated. In page 12-13 it should be clearly said what are the inputs and the outputs of the models, what exactly is the label that is then used to compute the RMSE, etc. It is claimed that the goal is to "process the large numbers of the non-steady-state peak signals and predict the corresponding steady-state absorbance

data". But how are those peak signals fed to the model?

- It is overall hard for me to believe that a model can be trained on steady state data to predict from non-steady state data unless both the steady state and non steady state from the same conditions can be associated with the same label. Is this the case?

- Tests that predict steady state from steady state, such as the bottom of page 12, are not representative at all -- how would those be? There is no data from the non-steady peak signal for those. So it has nothing to do, unless the goal is to predict the absorbance directly from the conditions (how are the chemicals digitalized?) and then this has nothing to do with non-steady state.

- Additionally, the ML regression tasks are not properly cross validated, and therefore it is very hard to trust the reported metrics. A thorough cross-validation study should be performed clearly, at least in the SI.

- On that vein, note that tests with a "random" 10% of the data used for training are meaningless without either thorough 10-fold cross-validation or several random splits (which then lead to a st. deviation in the metrics).

- The same questions apply to the section "AI-assisted cross species prediction". Not clear, not cross validated.

- Figure 3 is unreadable due to small font sizes and tiny panels.

- Figure 4 is extremely hard to understand, should be cut into pieces probably and overhauled completely.

In general, I suggest the authors overhaul the ML parts of the manuscript to make them clear, accompanied by concrete plots that actually check the main hypothesis, which according to the test is predicting steady-state abs. from non-steady state signal + conditions, with proper cross-validation.

Otherwise, the authors can remove the ML part, which at the moment does not feel like an integrated part of the paper, and just carry out detailed data analysis (as hinted in Figure 4). The experimental platform and the generated data is good and valuable on its own.

In my opinion, an interesting way to combine their platform with ML would be to use the 12,000 datapoints to predict yield, as it is implied in "AI-assisted cross species prediction" (but explaining clearly how everything is represented in the ML model inputs) then use an architecture with uncertainty, then use this in a bayesian opt. setting for quick optimization towards new products, for instance. At the moment, much of the latter sections of the paper feels either blurry and unclear or plainly disconnected (i.e. we can use these 12,000 data points for ML, but thats it).

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed all my comments in great details, and I am happy with the changes and recommend this work for publication in Nature Communications.

Reviewer #2

(Remarks to the Author)

The authors have devoted significant work to overhaul the ML/AI parts of the manuscript following my suggestions, clarifying what was done, how and why. I appreciate their effort and congratulate them for this work.

Nonetheless, I still think the AI/ML parts only serve an exemplary purpose and do not really add much to the story, since they are not used proactively but rather retrospectively.

In any case, I would ask the authors to give the new text blocks an additional read to smooth the writing, which is a bit weird at some points.

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 comments of the Reviewers

Responses to the comments of Reviewer #1

Comment 1: The methods section contains some information on how the descriptors were generated. However, there is no information on model hyper-parameters used in Figures 3A2, Figures 4F&G, and how they were selected. These include the learning rates, max-depths, regularization terms, etc. Please elaborate.

Response:

We appreciate the Reviewer for the comment.

We are sorry for not clarifying how the model hyperparameters were generated. In the revised version, we have added detailed descriptions in both the **AI-assisted absorbance prediction** and **AI-assisted cross-species prediction** sections.

We briefly mentioned the use of grid search to optimize the hyperparameters as “The training set was subjected to k-fold cross-validation, and the hyperparameter combinations of the models were optimized by grid search to improve the model performance.” on page 36, line 665-666. The related description about the **AI-assisted absorbance prediction** (Fig. 4 in the revised version) has been added in the revised version as “In the AI-assisted absorbance prediction, the performance of each AI model usually related to several to even a dozen of hyperparameters, from which 1-6 hyperparameters with large impacts on model performance were empirically selected and optimized using the Grid Search method (i.e., the GridSearchCV in sklearn.model_selection). Based on the default values of each hyperparameter, the single training time consumed by the model, the characteristics of the dataset (i.e., amount of data and number of features), and prior experience, the selection ranges of these hyperparameters to be searched (i.e., the param_grid parameter in GridSearchCV) were generated in advance, and the AI models with different hyperparameter values were trained, evaluated, and optimized with 5-fold cross-validation. The optimal hyperparameter values were obtained from GridSearchCV's best_params_, and the optimal model was obtained from GridSearchCV's best_estimator_. For example, for

the Random Forest regression model, the selected hyperparameters and their values included: param_grid = {'n_estimators': [250, 350, 500], 'max_depth':[10, 30, 50], 'max_features': [4, 6]}. The optimal hyperparameters and their values included: {'max_depth': 30, 'max_features': 6, 'n_estimators': 350}. Considering that each AI model involved different hyperparameters, each of the 10 AI models was analyzed specifically to obtain the optimal hyperparameter values, respectively (Fig. S7). The linear regression model was not optimized for hyperparameters as sklearn does not have hyperparameters set for it.” in Supplementary Information on page S24-S25.

In addition, the software with user interface is also available on Github with a link as <https://github.com/LJM-1997/NCOMMS-24-15301-T> and a password for the datasets as NCOMMS-24-15301-T, and the Reviewers can obtain information related to model hyperparameter screening from the source code.

Regression models	Selected hyperparameters and values	Optimal hyperparameter values by grid search
PLSR	{'n_components': [1, 2, 5, 9]}	{'n_components': 9}
Linear	No model parameters need to screen	-
Adaboosting	{'n_estimators': [50, 100, 200], 'learning_rate':[0.1, 0.5, 1]}	{'learning_rate': 0.1, 'n_estimators': 100}
ExtraTree	{'max_depth': [3, 10, 20], 'min_samples_leaf':[5, 10, 20], 'min_samples_split':[10, 30, 50]}	{'max_depth': 20, 'min_samples_leaf': 5, 'min_samples_split': 10}
SVR	{'C': [0.1, 1, 5, 10]}	{'C': 1}
KNN	{'weights':['distance'], 'n_neighbors': [5, 10, 30], 'p':[1, 2, 4]}	{'n_neighbors': 5, 'p': 1, 'weights': 'distance'}
MLP	{'solver':['lbfgs'], 'activation':['relu'], 'learning_rate_init':[0.001], 'hidden_layer_sizes':[(50,50), (50,20), (50,50,20)], 'max_iter': [100, 200], 'alpha':[0.001, 0.01]}	{'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (50, 50, 20), 'learning_rate_init': 0.001, 'max_iter': 200, 'solver': 'lbfgs'}
RandomForest	{'n_estimators': [250, 350, 500], 'max_depth':[10, 30, 50], 'max_features': [4, 6]}	{'max_depth': 30, 'max_features': 6, 'n_estimators': 350}
CascadeForest	{'max_layers': [20, 30], 'n_estimators':[2, 10], 'n_trees':[100, 150]}	{'max_layers': 20, 'n_estimators': 2, 'n_trees': 150}
XGB	{'n_estimators': [200, 300, 500]}	{'n_estimators': 500}

Figure S7. Selected hyperparameters and their values of the 10 AI models, with the optimal hyperparameter values.

The related description about the AI-assisted cross-species prediction (Fig. 6 in the revised version) has been added as “In the AI-assisted cross-species prediction, we used the XGBoost's default hyperparameters because they could offer good performance.” in Supplementary Information on page S25.

Comment 2: t-SNE is generally not considered reasonable for feature selection and is rather used for visualization and understanding the structure of high-dimensional data. How was 15 decided as the number of features? Please elaborate. Model performance

is quite high in the current setting by typically RFE methods are used for choosing the most promising uncorrelated features.

Response:

We appreciate the Reviewer for the comment.

Our feature includes two parts, Mordred descriptors of the substrate and the photocatalyst species described with the SMILES string and reaction conditions (including the 6 numerical variables, i.e., laser light intensity, flow rate, substrate concentration, photocatalyst ratio, photocatalyst concentration, and absorbance wavelength), both of which are of great importance for the photocatalytic [2+2] cycloaddition. However, the extensive number of Mordred descriptors (more than 1800 descriptors per molecule) is significantly higher compared to reaction conditions (represented by 6 numerical variables), which may hinder the model's predictive performance. Given the specificity of our problem, where the number of Mordred descriptors far exceeds that of other numerical variables, we didn't perform feature selection. Instead, we opted for dimensionality reduction while striving to preserve the high-dimensional structural relationships as much as possible, using the reduced molecular representation to represent the molecules.

Given the large number of Mordred descriptors, the TSNE's ability to manage and reduce high-dimensional data efficiently is preferable. It allows us to reduce the dimension without losing significant information about the molecular relationships. Therefore, we applied TSNE and observed remarkable predictive performance. The exact dimensionality is not critical as long as it is reasonably close to the dimensions of other numerical variables. Initially, as we described in Methods, the number 15 was chosen randomly during our experiments. In the revised version, we also tried to reduce it to 2 dimensions and observed that this setting also yielded excellent predictive performance. Additionally, reducing to 2 dimensions could leverage TSNE's strengths, which included visualizing different molecules and observing their distribution, distances, and similarities. We have uploaded the code related to dimensionality reduction and the results after reduction to GitHub, hoping to facilitate future research

in this area.

In responding to the comment of the Reviewer, we have further clarified this as “With the SMILES string of each molecule as part of the inputs of the models (Extended Data Figure 5), Mordred could generate both 2D and 3D descriptors, encompassing a list of more than 1800 descriptors. However, such a dimension of the Mordred descriptors was very high compared with other numerical variables (i.e., laser light intensity, flow rate, substrate concentration, photocatalyst ratio, photocatalyst concentration, and absorbance wavelength), which may hamper the predictive capabilities of the model. Given the specificity of our problem, where the number of Mordred descriptors far exceeded those of other numerical variables, we did not perform feature selection. Instead, we reduced the dimension of Mordred descriptor of each substrate and photocatalyst species from more than 1800 to 2 with T-distributed stochastic neighbor embedding (TSNE) algorithm, which is a widely used unsupervised dimension reduction technique owing to its advantage in capturing local data characteristics and revealing subtle data structures. Given the large number of Mordred descriptors, the TSNE algorithm is preferable due to its ability to manage high-dimensional data. The reduced Mordred descriptors were then concatenated with other 6 numerical variables to construct the final reaction fingerprints.

We used the XGB algorithm for cross-substrate and cross-photocatalyst prediction, which is a highly efficient and flexible machine learning algorithm based on the gradient boosting framework. It is renowned for its outstanding performance and high efficiency, which are optimized through parallel processing and tree-pruning. Additionally, the XGB's regularization method could prevent overfitting, thereby improving the model's generalizability.” in Method on page 42, line 742-760.

We appreciate the reviewer's insights and hope this will be helpful for clarifying our approach.

***Comment 3:** To allow reproducibility of the work, please provide the source code on github.*

Response:

We appreciate the Reviewer for the reminding.

In the AI-assisted prediction, the source code and datasets are available on Github with a link as <https://github.com/LJM-1997/NCOMMS-24-15301-T> and a password for the datasets as **NCOMMS-24-15301-T**.

Responses to the comments of Reviewer #2

Comment 1: The specific task at hand for the ML is never clearly stated. In page 12-13 it should be clearly said what are the inputs and the outputs of the models, what exactly is the label that is then used to compute the RMSE, etc. It is claimed that the goal is to "process the large numbers of the non-steady-state peak signals and predict the corresponding steady-state absorbance data". But how are those peak signals fed to the model?

It is overall hard for me to believe that a model can be trained on steady state data to predict from non-steady state data unless both the steady state and non steady state from the same conditions can be associated with the same label. Is this the case?

Tests that predict steady state from steady state, such as the bottom of page 12, are not representative at all -- how would those be? There is no data from the non-steady peak signal for those. So it has nothing to do, unless the goal is to predict the absorbance directly from the conditions (how are the chemicals digitalized?) and then this has nothing to do with non-steady state.

Response:

We appreciate the Reviewer for the comment.

We are very apologetic that our insufficiently clear and detailed description has caused the Reviewers' understanding to be difficult.

In fact, there are two applications of AI models in this work, **AI-assisted absorbance prediction** and **AI-assisted cross-species prediction**. In the response to the present comment (*Comment 1*) of **Reviewer II**, we provide further explanation and descriptions about **AI-assisted absorbance prediction**, and provide the explanation and descriptions about **AI-assisted cross-species prediction** in the response to *Comment 3*.

We are sorry that the task as well as the inputs and outputs of the models in **AI-assisted absorbance prediction** were not stated clearly enough.

When ultrafast photocatalytic synthesis was achieved, we attempted to accomplish fast on-line characterization as well as high-throughput screening. Although the UV-

Vis absorption spectroscopy method is a fast on-line characterization method, during the experiments we observed that the screening throughput was limited to 2,600 reaction conditions per day due to the relatively long waiting time required when switching between different reaction conditions. Specifically, when the synthesis and characterization of the previous condition experiment were completed in the flow system, a new reaction solution needed to be introduced into the reactor and detection flow-cell channels.

Due to the convection and molecular diffusion effects existed in the flow system, the previous reacted solution and the newly-introduced unreacted solution would mix with each other at their junction region, and the absorbance signals detected by the UV-Vis detector exhibited dynamically-changing format (Fig. 3A1) during the solution switching process. The conventional method is to wait for the newly-introduced reaction solution to flush all of the previous solution out of the flow system to enable the detector to obtain a stable plateau-type steady-state absorbance signal for data reading. However, such a method requires much longer waiting time for obtaining the steady-state signal, such as 27 s, which far exceeded the time (<4 s) for photocatalytic synthesis and characterization for a reaction solution, severely limiting the improvement of the screening throughput for different reaction conditions (Fig. 3A2). In fact, this is one of the major limiting bottlenecks in the application of current flow chemistry systems to high-throughput screening.

In the present work, we proposed the strategy of non-steady-state experimental mode instead of the steady-state mode to significantly improve the screening throughput by using the laser pulse irradiation method to rapidly switch the experimental laser intensity conditions, producing a series of non-steady-state continuous peak-shaped signals as shown in Figure 3A3. Under the non-steady-state mode, the waiting time for reaction solution switching was shortened to 6 s and the average time for each experimental cycle was shortened to 8.5 s.

However, these non-steady-state peak signals included the combined absorbance information of the reactants and products from the previous and the newly-introduced

reaction solutions, which were influenced by multiple factors related to the convection and molecular diffusion effects, such as the reaction solution flow rate, the inner diameters and lengths of the reactor and the detection flow-cell channels, and the reactants and products molecular weights. Therefore, it was a great challenge to acquire the corresponding steady-state absorbance data from the non-steady-state peak signals for evaluating the reaction progress.

In the AI prediction models, the non-steady-state absorbance data (i.e., 40 absorbance data points recorded for each non-steady-state signal peak, Fig. 3A4) as well as the corresponding 8 reaction variables (i.e., flow rate, light intensity, wavelength, substrate concentration, photocatalyst ratio, photocatalyst concentration, substrate species, and photocatalyst species) were set as **the inputs of the models** (Fig. 4A).

The corresponding steady-state absorbance data obtained experimentally using the steady-state mode with the same conditions (i.e., real steady-state absorbance data) were set as **the targets of the models**, and the predicted steady-state absorbance data were set as **the outputs of the models**.

Since the aim of the AI prediction models was to predict the steady-state absorbance data from the non-steady-state data under the same conditions, there was no need to input the chemical structures of the reactants. The substrate and photocatalyst species were expressed in terms of relative molecular weight, which is directly related to the molecular diffusion effect. A total of 10 different AI regression models were evaluated, and the model with the best performance was selected based on the R^2 and RMSE values of the test set. The RMSE values were calculated based on the targets (i.e., real steady-state absorbance data) and the outputs (i.e., predicted steady-state absorbance data) of the models.

In responding to the comment of the Reviewer, in the revised version, we have further clarified this as “During the high-throughput screening experiments for different reaction conditions, when the synthesis and characterization of the previous condition experiment were completed in the flow system, a new reaction solution needed to be introduced into the reactor and detection flow-cell channels. Due to the convection and

molecular diffusion effects existed in the flow system, the previous reacted solution and the newly-introduced unreacted solution would mix with each other at their junction region, and the absorbance signals detected by the UV-Vis detector exhibited dynamically-changing format (Fig. 3A1) during the solution switching process. The conventional method is to wait for the newly-introduced reaction solution to flush all of the previous solution out of the flow system to enable the detector to obtain a stable plateau-type steady-state absorbance signal for data reading. However, such a steady-state experimental mode requires much longer waiting time for obtaining the steady-state signals. In the above photocatalytic screening experiment under the steady-state mode, the lasers kept irradiating the microreactor channel and the system spent most of the time (27 s of the 32 s of one experimental cycle time) in switching the different experimental conditions and waiting for a steady-state detection signal to be obtained (such as the laser light intensity experiment as shown in Figure 3A2). Such a waiting time far exceeded the time (<4 s) for photocatalytic synthesis and characterization for a reaction solution, severely limiting the screening throughput for different reaction conditions. In fact, this is one of the major limiting bottlenecks in the application of current flow chemistry systems to high-throughput screening.” on page 10, line 148-166;

“To increase the efficiency of time utilization and screening throughput, we proposed the strategy of non-steady-state experimental mode instead of the steady-state mode by using the laser pulse irradiation method to rapidly switch the experimental laser intensity conditions, producing a series of non-steady-state continuous peak-shaped signals as shown in Figure 3A3. Under the non-steady-state mode, the waiting time for reaction solution switching was shortened to 6 s and the average time for each experimental cycle was shortened to 8.5 s, achieving an ultra-high throughput up to 10,000 reaction conditions per day (Fig. 3A3, 3A4, 3B).

However, these non-steady-state peak signals included the combined absorbance information of the reactants and products from the previous and the newly-introduced reaction solutions, which were influenced by multiple factors related to the convection

and molecular diffusion effects, such as the reaction solution flow rate, the inner diameters and lengths of the reactor and the detection flow-cell channels, and the reactants and products molecular weights.” on page 12-13, line 184-195;

“We developed the AI-assisted absorbance prediction method by using the AI method to analyze the influencing factors related to the convection and molecular diffusion effects and decoupling the non-steady-state data of the adjacent reaction solutions mixed with each other, to predict the corresponding steady-state absorbance data of the respective reaction solutions.” on page 13, line 198-202;

“We attempted to use 10 regression models based on the principles of linear models, decision tree, neural networks and integrated learning, to process the large numbers of the non-steady-state absorbance data and to predict the corresponding steady-state absorbance data under the same reaction condition. In these models, the non-steady-state absorbance data (i.e., 40 absorbance data points recorded for each non-steady-state signal peak, Fig. 3A4) as well as the all corresponding 8 variables (i.e., flow rate, light intensity, wavelength, substrate concentration, photocatalyst ratio, photocatalyst concentration, substrate species, and photocatalyst species) of the present flow photocatalytic system were set as the inputs of the models (Fig. 4A). The corresponding steady-state absorbance data obtained experimentally using the steady-state mode with the same conditions (i.e., real steady-state absorbance data) were set as the targets of the models, and the predicted steady-state absorbance data were set as the outputs of the models. The substrate and photocatalyst species were input to the models in the form of relative molecular weights instead of chemical structures, since they are directly related to the molecular diffusion effect. On the basis of the massive amounts of the output and target data of the 12,000 reaction conditions (Fig. 4B1), we evaluated the performance of the 10 regression models based on the R^2 and RMSE values of the test set. The RMSE values were calculated based on the targets (i.e., real steady-state absorbance data) and the outputs (i.e., predicted steady-state absorbance data) of the models.” on page 15-16, line 230-246.

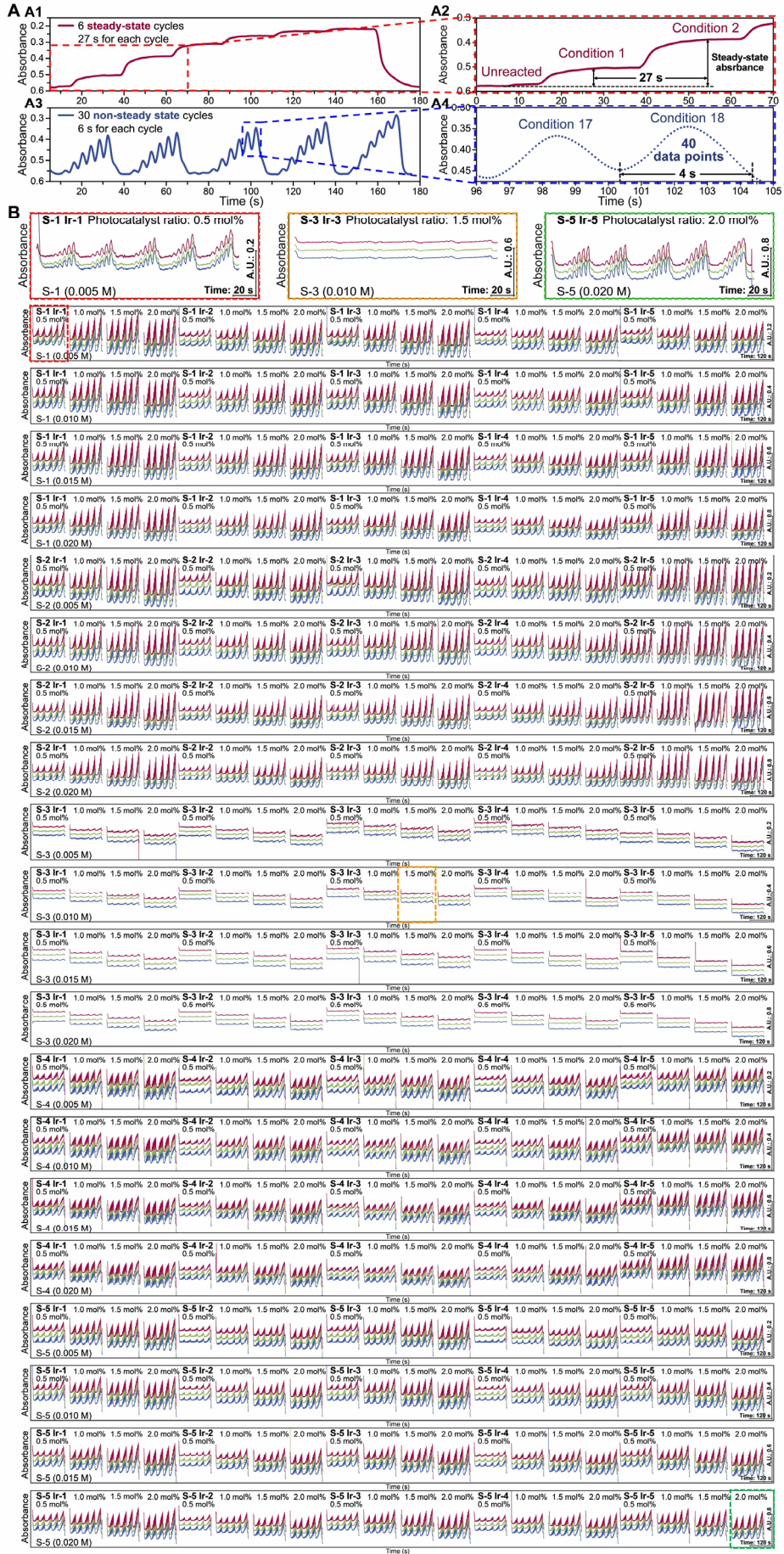


Fig. 3. Recordings of the typical and total absorbance signals obtained in the large-scale screening experiment. (A) Typical recordings of the steady-state and non-steady-state absorbance signals obtained in the screening for S-5, with a photocatalyst ratio of 2 mol% and a S-5 concentration of 0.01 M. (A1) Typical absorbance recordings of 6 different reaction conditions under the steady-state screening mode. (A2) Enlarged view of the absorbance recordings for the first two conditions of the 6 reaction conditions in (A1). It took an average time of 27 s in each condition cycle to obtain the steady-state absorbance signal, which was calculated from the difference between the reacted steady-state plateau absorbance and the unreacted blank absorbance. (A3) Typical absorbance recordings of 30 different reaction conditions under the non-steady-state screening mode. (A4) Enlarged view of the absorbance recordings for the 17th and 18th conditions of the 30 reaction conditions in (A3). Each condition cycle took an average time of ca. 4s, with a non-steady-state signal peak containing 40 absorbance data points, obtained using the laser pulse irradiation method. (B) Recordings of the non-steady-state peak signals obtained in the screening experiment of the total 12,000 reaction conditions, which was replicated three times to test the repeatability.

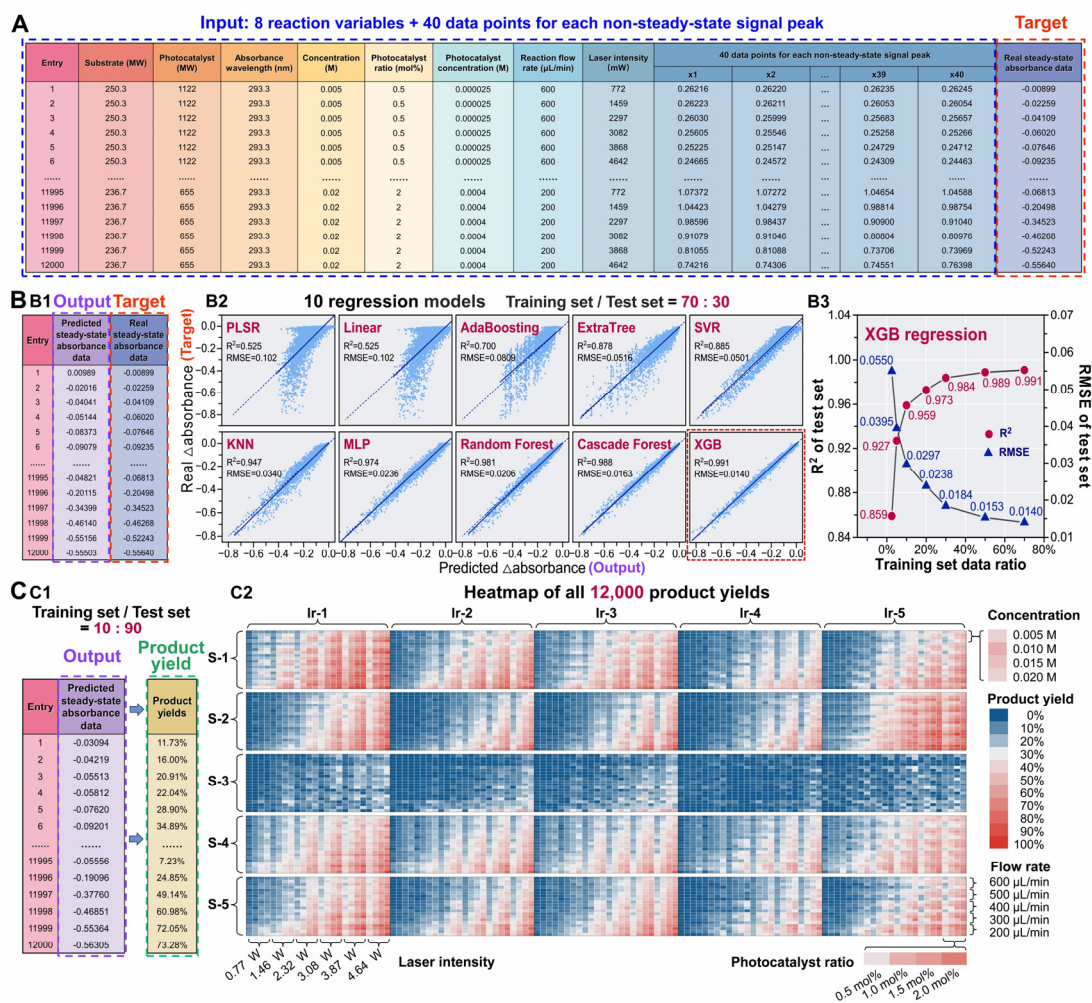


Fig. 4. AI-assisted prediction of steady-state absorbance data from non-steady-state absorbance data for condition screening of photocatalytic [2+2] cycloaddition. (A) Composition of the large dataset used in the AI-assisted steady-state absorbance prediction, including the 12,000 absorbance data with 48 eigenvalues (8 reaction variables and 40 data points for each non-steady-state signal peak) and 1 target (steady-state absorbance data). (B) AI-assisted prediction models for predicting the steady-state absorbance data from the non-steady-state absorbance data. (B1) Predicted and real steady-state absorbance as the outputs and targets of the models, respectively. (B2) 10 different regression models for the prediction of the steady-state absorbance data from the non-steady-state absorbance data, including partial least squares regression (PLSR), linear regression, adaptive boosting (AdaBoosting) regression, extremely randomized trees (ExtraTree) regression, support vector regression (SVR), k-nearest neighbor (KNN) regression, multi-layer perceptron (MLP) regression, random forest regression,

cascade forest regression, and extreme gradient boosting (XGB) regression models. The dashed line is the $y=x$ line, and the solid line is a linear fit curve between the predicted and the true steady-state absorbance values. The two metrics, R^2 and RMSE values, were calculated using the corresponding functions within sklearn.metrics. (B3) Performance of the test set of the XGB regression model. The prediction accuracy of the test set gradually decreases as the proportion of randomly selected training set data decreases from 70% to 2.5%. (C) Screening results of the total 12,000 reaction conditions obtained using the AI-assisted steady-state absorbance prediction method. (C1) Predicted steady-state absorbance data output from the XGB model with training set : test set = 10 : 90 were converted to product yields. (C2) Heatmap showing the screening results of a total of 12,000 reaction conditions, including the orthogonal combination of 5 substrate species, 5 photocatalyst species, 4 concentrations, 4 photocatalyst ratios, 5 flow rates, and 6 laser light intensities under the non-steady-state mode.

Comment 2: Additionally, the ML regression tasks are not properly cross validated, and therefore it is very hard to trust the reported metrics. A thorough cross-validation study should be performed clearly, at least in the SI.

On that vein, note that tests with a "random" 10% of the data used for training are meaningless without either thorough 10-fold cross-validation or several random splits (which then lead to a st. deviation in the metrics).

Response:

We appreciate the Reviewer for the comment.

As suggested by the Reviewer, in the task of AI-assisted absorbance prediction, i.e., using 10% of the non-steady-state absorbance data as the training set for the XGB model to predict the 12,000 steady-state absorbance data prediction, we have added a 10-fold cross-validation study. The results of the 10-fold cross-validation study of the AI-assisted absorbance prediction are shown in Figure S8 in the revised Supplementary Information, and the methods and results were also added, described as “a 10-fold

cross-validation study was performed and the results are shown in Supplementary Information (Fig. S8).” on page 17, line 266-267, and “In the task of predicting 12,000 steady-state absorbance data using 10% of the non-steady-state absorbance data as the training set of the XGB model, a 10-fold cross-validation study was performed on the entire dataset by dividing the dataset equally into 10 parts (i.e., 0-10%, 10-20%, ..., 90-100%). Each time, one part of them was taken as the training set and the remaining 9 parts as the test set, resulting in a total of 10 kinds of training/test sets. To ensure the randomness in the extracting of the training sets, the data in the dataset were disrupted in order before performing the 10-fold cross-validation. The optimized XGB regression model (i.e., $n_estimators$ hyperparameter taking the value of 500 filtered by grid search) was trained and evaluated using the 10 training/test sets, and the model performance metrics including the MAE, RMSE, and R^2 values and their means, standard deviations, and coefficients of variation (CV) were calculated (Fig. S8). The results showed that the model performance metrics obtained from the dataset as shown in Figure 4A had no significant difference, and the coefficients of variation were 4.50% for MAE, 5.44% for RSME and 0.49% for R^2 , respectively, indicating the good generalization ability of the model.” in Supplementary Information on page S25.

Entry	Train Index	Test Index	Metrics of the test set		
			MAE	RMSE	R^2
1	[0 1 2 ... 1197 1198 1199]	[1200 1201 1202 ... 11997 11998 11999]	0.0182	0.0289	0.961
2	[1200 1201 1202 ... 2397 2398 2399]	[0 1 2 ... 1198 1199 2400... 11997 11998 11999]	0.0188	0.0296	0.959
3	[2400 2401 2402 ... 3597 3598 3599]	[0 1 2 ... 2398 2399 3600... 11997 11998 11999]	0.0188	0.0304	0.956
4	[3600 3601 3602 ... 4797 4798 4799]	[0 1 2 ... 3598 3599 4800... 11997 11998 11999]	0.0179	0.0292	0.960
5	[4800 4801 4802 ... 5997 5998 5999]	[0 1 2 ... 4798 4799 6000... 11997 11998 11999]	0.0191	0.0309	0.955
6	[6000 6001 6002 ... 7197 7198 7199]	[0 1 2 ... 5998 5999 7200... 11997 11998 11999]	0.0180	0.0286	0.962
7	[7200 7201 7202 ... 8397 8398 8399]	[0 1 2 ... 7198 7199 8400... 11997 11998 11999]	0.0177	0.0277	0.964
8	[8400 8401 8402 ... 9597 9598 9599]	[0 1 2 ... 8398 8399 9600... 11997 11998 11999]	0.0203	0.0327	0.950
9	[9600 9601 9602 ... 10797 10798 10799]	[0 1 2 ... 9598 9599 10800... 11997 11998 11999]	0.0199	0.0325	0.951
10	[10800 10801 10802 ... 11997 11998 11999]	[0 1 2 ... 10797 10798 10799]	0.0184	0.0299	0.958
Mean			0.0187	0.0300	0.958
Standard deviation			0.000842	0.00164	0.00469
Coefficient of variation (CV)			4.50%	5.44%	0.49%

Fig. S8. The 10-fold cross-validation study performed on the entire dataset in the AI-assisted absorbance prediction. The model performance metrics include the MAE, RMSE, and R^2 values and their means, standard deviations, and coefficients of variation

(CV).

Comment 3: The same questions apply to the section "AI-assisted cross species prediction". Not clear, not cross validated.

Response:

We appreciate the Reviewer for the comment.

We are very sorry that the task as well as the inputs and outputs of the models in “**AI-assisted cross-species prediction**” were not stated clearly enough.

In the section of “**AI-assisted cross-species prediction**” on page 20-21, the model was **tasked with** cross-species prediction of the product yields. The reaction conditions (i.e., the same 8 reaction variables as in the absorbance prediction section) were set as the inputs of the models, the product yields obtained experimentally were set as the targets of the models, and the predicted product yields were set as the outputs of the models. However, unlike the AI-assisted absorbance prediction under the same condition, the cross-species prediction of the product yields required detailed information of the chemical structures of the substrate and photocatalyst species. Therefore, SMILES strings were used to describe, digitize and input the chemical structure of the substrate and photocatalyst compounds to the models.

In responding to the comment of the Reviewer, we have further clarified this as “To further utilize the above 12,000 data and preliminarily explore the potential possibility of applying AI technique to intelligent chemical synthesis screening, we used the XGB algorithm to perform AI-assisted prediction of product yields cross-substrate and cross-photocatalyst. For cross-species prediction of product yields, the inputs of the models were the reaction conditions (i.e., the 8 reaction variables), the targets were the product yields obtained experimentally, and the outputs were the predicted product yields (Fig. 6A). Differing from the AI-assisted absorbance prediction, the cross-species prediction required detailed chemical structure information of the substrate and photocatalyst species, which were described, digitized, and input to the models using the SMILES strings to generate Modred descriptors

(Extended Data Figure 5). The Mordred descriptor dimensions of each substrate and photocatalyst species were reduced to 2 in order to match the dimensions of the other variables and facilitate visualization in subsequent study, as described in Methods. The inputs consisted of 10 dimensions, with 2 representing the substrate species, 2 representing the photocatalyst species, and the remaining 6 representing the other variables (Fig. 6A).” on page 21, line 344-357, and “With the SMILES string of each molecule as part of the inputs of the models (Extended Data Figure 5), Mordred could generate both 2D and 3D descriptors, encompassing a list of more than 1800 descriptors. However, such a dimension of the Mordred descriptors was very high compared with other numerical variables (i.e., laser light intensity, flow rate, substrate concentration, photocatalyst ratio, photocatalyst concentration, and absorbance wavelength), which may hamper the predictive capabilities of the model. Given the specificity of our problem, where the number of Mordred descriptors far exceeded those of other numerical variables, we did not perform feature selection. Instead, we reduced the dimension of Mordred descriptor of each substrate and photocatalyst species from more than 1800 to 2 with T-distributed stochastic neighbor embedding (TSNE) algorithm, which is a widely used unsupervised dimension reduction technique owing to its advantage in capturing local data characteristics and revealing subtle data structures. Given the large number of Mordred descriptors, the TSNE algorithm is preferable due to its ability to manage high-dimensional data. The reduced Mordred descriptors were then concatenated with other 6 numerical variables to construct the final reaction fingerprints.” on page 42, line 742-755.

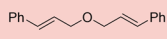
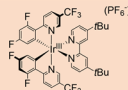
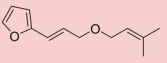
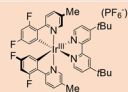
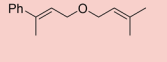
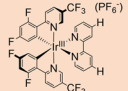
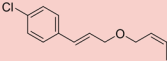
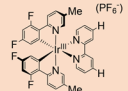
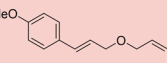
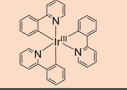
Settings in AI-assisted **cross-species** prediction

A Input: 10 dimensions (8 variables) Target

Entry	Photocatalyst		Substrate		Absorbance wavelength (nm)	Substrate concentration (M)	Photocatalyst ratio (mol%)	Photocatalyst concentration (M)	Reaction flow rate (μL/min)	Laser intensity (mW)	Product yield (%)
	Ir_0	Ir_1	s_0	s_1							
1	0	0.90758	0	0.026131	293.3	0.005	0.5	0.000025	600	772	11.7
2	0	0.90758	0	0.026131	293.3	0.005	0.5	0.000025	600	1459	16.0
5955	0.37390	0.42933	0.19517	1	285.4	0.01	1.5	0.00015	400	2297	7.2
5956	0.37390	0.42933	0.19517	1	285.4	0.01	1.5	0.00015	400	3082	12.2
11999	1	0	0.25776	0.56653	293.3	0.02	2	0.0004	200	3868	72.0
12000	1	0	0.25776	0.56653	293.3	0.02	2	0.0004	200	4642	73.3

Fig. 6. (A) Settings used for AI-assisted cross-species prediction, including inputs and

targets of the models.

Substrate species	Substrate (SMILES)	Photocatalyst species	Photocatalyst (SMILES)
	<chem>C1(/C=C/C/COC/C=C/C2=CC=CC=C2)=CC=CC=C1</chem>		<chem>FC(F)(C1=CN=C(C=C1)C2=C(F)C=C(C=C2)F.F.[P-](F)(F)(F)F.CC(C)(C3=CC(C4=CC(C(C)C)C)C)C=CC=N4)=NC=C3)C.FC5=CC(F)=C([C-]=C5)C6=NC=C(C=C6)C(F)(F)F.[Ir+3]</chem>
	<chem>C/C(C)=C/COC/C=C/C1=C=C=C1</chem>		<chem>FC1=CC(F)=C(C2=[N]3C=C(C(C)C=C2)C([Ir]435([N]6=C(C(C5=CC(F)=C7)=C7F)C=CC(C)=C6)[N]8=CC=C(C(C)C)C)C=C8C9=[N]4C=CC(C(C)C)=C9)=C1.F[P-](F)(F)(F)F</chem>
	<chem>C/C(C)=C/COC/C=C(C)C1=CC=CC=C1</chem>		<chem>FC1=CC(F)=C(C2=[N]3C=C(C(C)C)C=C2)C([Ir]435([N]6=C(C(C5=CC(F)=C7)=C7F)C=CC(C)=C6)[N]8=CC=C([H])C=C8C9=[N]4C=CC([H])=C9)=C1.F[P-](F)(F)(F)F</chem>
	<chem>C/C(C)=C/COC/C=C/C1=C=C(C(Cl))C=C1</chem>		<chem>FC1=CC(F)=C(C2=[N]3C=C(C(C)C=C2)C([Ir]435([N]6=C(C(C5=CC(F)=C7)=C7F)C=CC(C)=C6)[N]8=CC=C([H])C=C8C9=[N]4C=CC([H])=C9)=C1.F[P-](F)(F)(F)F</chem>
	<chem>C/C(C)=C/COC/C=C/C1=C=C(C(OC))C=C1</chem>		<chem>C12=CC=CC=C1C3=[N]([Ir]245(C6=C(C7=CC=CC=[N]75)C=C(C=C6)C8=CC=CC=C8C9=[N]4C=CC=C9)C=CC=C3</chem>

Extended Data Figure 5. SMILES strings for all the substrate and photocatalyst species.

In the section “AI-assisted cross-species prediction”, we have also supplemented the **cross-validation studies** based on the Reviewers' valuable suggestions, and added the experiments with cross-photocatalyst and cross-substrate prediction for all possible cases.

In the Supplementary Information, we have further clarified the **cross-validation studies** for **AI-assisted cross-species prediction** as “In the cross-species prediction, the cross-validation studies were performed for both cross-photocatalyst and cross-substrate prediction. In the cross-photocatalyst prediction, the data of 4 photocatalyst species were used as the training set to predict the yields of the other 1 photocatalyst species, and the data of 3 photocatalyst species were used as the training set to predict the yields of the other 2 photocatalyst species, with box plots showing the product yield results (Fig. 5G, S9A).

In the cross-substrate prediction, the data of 3 substrate species were used as the training set to predict the yields of the other 1 substrate species, and the data of 2 substrate species were used as the training set to predict the yields of the other 2

substrate species, with box plots showing the product yield results (Fig. 5G, 2S9B). For instance, we used the data of S-1, S-2, and S-4 as the training set to predict the yields of S-5, achieving MAE=0.0698 and RMSE=0.0878 (Fig. S9B1). The distinct effects of the 5 photocatalyst species on S-5 were accurately predicted with Ir-1 as the optimal photocatalyst. With a smaller training set of S-1 and S-2, the prediction for S-4 and S-5 were also completed with MAE=0.0772 and RMSE=0.0999 (Fig. S9B2).

The performance parameters of the test set for all possible cases were obtained in the cross-validation studies, and their small errors demonstrated the stability of the method (Fig. 5H, Fig. S10).” in Supplementary Information on page S26.

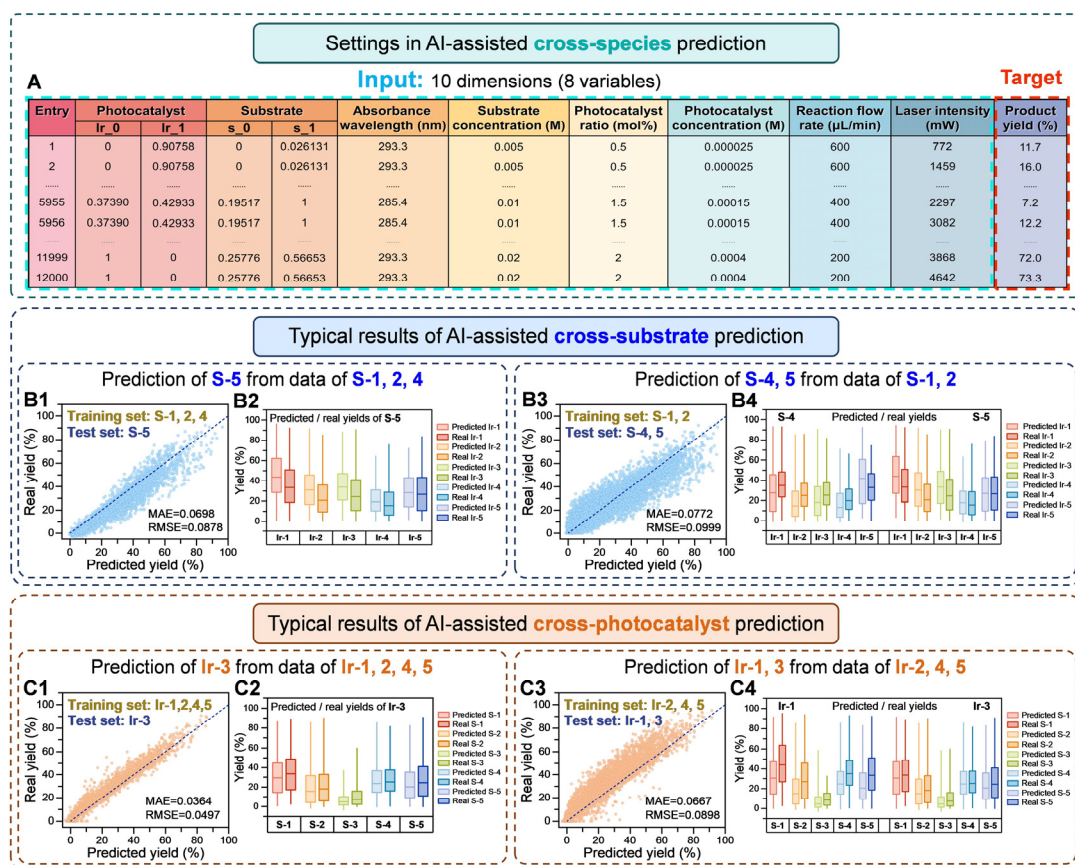


Fig. 6. AI-assisted cross-species prediction. (A) Settings used for AI-assisted cross-species prediction, including inputs and targets of the models. (B) Typical results of the AI-assisted cross-substrate prediction, including prediction of S-5 from data of S-1, S-2, and S-4 (B1), and prediction of S-4, S-5 from data of S-1 and S-2 (B3) with box plots showing product yields results (B2, B4), respectively. (C) Typical results of the AI-

assisted cross-photocatalyst prediction, including prediction of Ir-3 from data of Ir-1, Ir-2, Ir-4 and Ir-5 (C1), and prediction of Ir-1, Ir-3 from data of Ir-2, Ir-4 and Ir-5 (C3) with box plots showing product yields results (C2, C4), respectively. Model performance metrics for the cross-validation studies, including the MAE and RMSE values are shown in (B1), (B3), (C1), and (C3). The entire results of the cross-species prediction are shown in Figure S9.

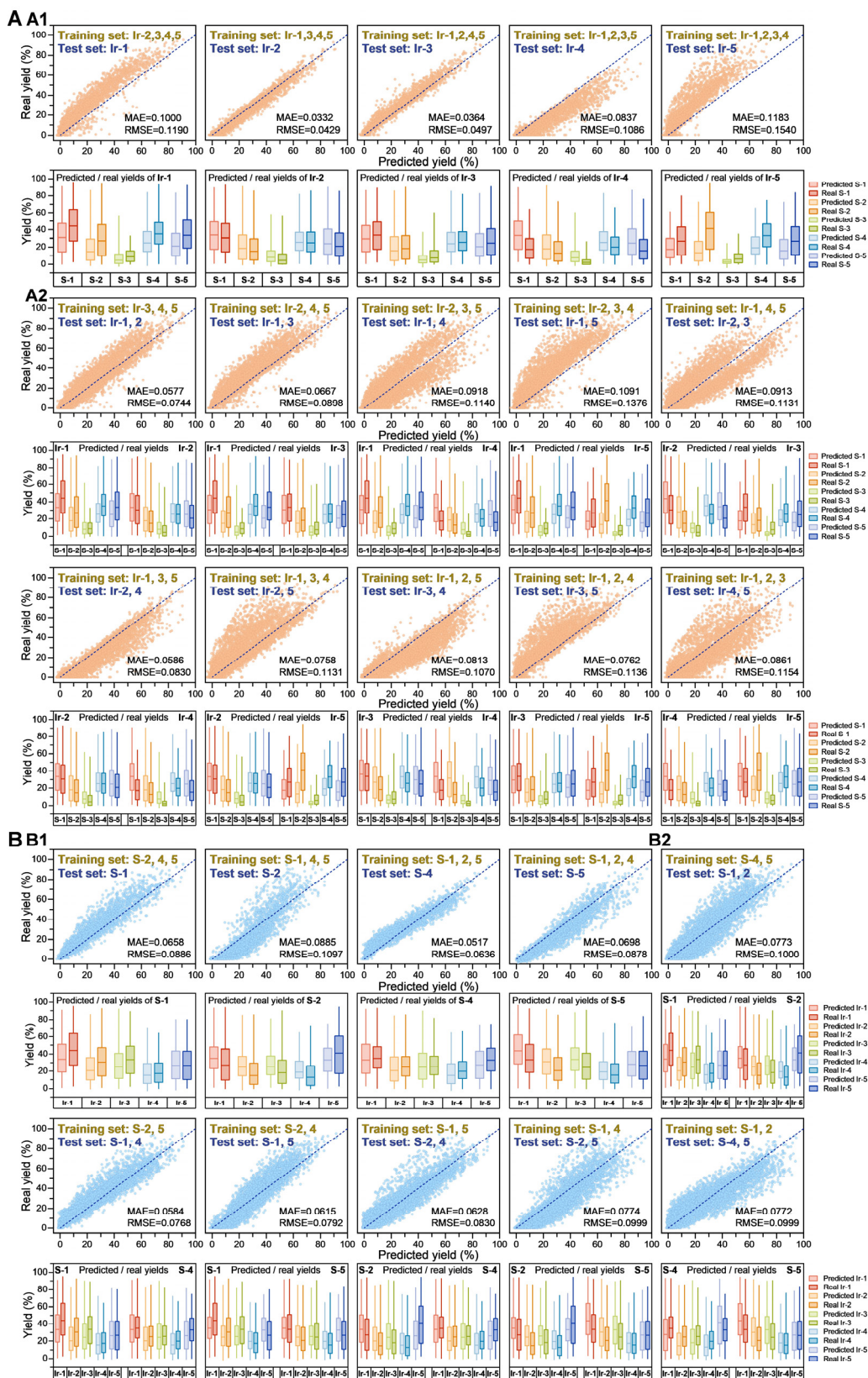


Fig. S9. Results of cross-validation studies for the cross-photocatalyst prediction (A) and the cross-substrate prediction (B). (A1) The cross-photocatalyst prediction using

data from 4 photocatalyst species as a training set to predict the yields of the other 1 photocatalyst species. (A2) The cross-photocatalyst prediction using data from 3 photocatalyst species as a training set to predict the yields of the other 2 photocatalyst species. (B1) The cross-substrate prediction using data from 3 substrate species as a training set to predict the yields of the other 1 substrate species. (B2) The cross-substrate prediction using data from 2 substrate species as a training set to predict the yields of the other 2 substrate species.

A Cross-validation of cross-photocatalyst prediction					B Cross-validation of cross-substrate prediction				
Entry	Train set	Test set	Metrics of the test set		Entry	Train set	Test set	Metrics of the test set	
			MAE	RMSE				MAE	RMSE
1	Ir-2,3,4,5	Ir-1	0.1000	0.1190	1	S-2,4,5	S-1	0.0658	0.0886
2	Ir-1,3,4,5	Ir-2	0.0332	0.0429	2	S-1,4,5	S-2	0.0885	0.1097
3	Ir-1,2,4,5	Ir-3	0.0364	0.0497	3	S-1,2,5	S-4	0.0517	0.0636
4	Ir-1,2,3,5	Ir-4	0.0837	0.1086	4	S-1,2,4	S-5	0.0698	0.0878
5	Ir-1,2,3,4	Ir-5	0.1183	0.1540	Mean			0.0690	0.0874
Standard deviation			0.0341	0.0424	Standard deviation			0.0131	0.0164
1	Ir-3,4,5	Ir-1,2	0.0577	0.0774	1	S-4,5	S-1,2	0.0773	0.1000
2	Ir-2,4,5	Ir-1,3	0.0667	0.0898	2	S-2,5	S-1,4	0.0584	0.0768
3	Ir-2,3,5	Ir-1,4	0.0918	0.1140	3	S-2,4	S-1,5	0.0615	0.0792
4	Ir-2,3,4	Ir-1,5	0.1091	0.1376	4	S-1,5	S-2,4	0.0628	0.0830
5	Ir-1,4,5	Ir-2,3	0.0913	0.1131	5	S-1,4	S-2,5	0.0774	0.0999
6	Ir-1,3,5	Ir-2,4	0.0586	0.0830	6	S-1,2	S-4,5	0.0772	0.0999
7	Ir-1,3,4	Ir-2,5	0.0758	0.1131	Mean			0.0691	0.0898
8	Ir-1,2,5	Ir-3,4	0.0813	0.1070	Standard deviation			0.00831	0.0103
9	Ir-1,2,4	Ir-3,5	0.0762	0.1136	Standard deviation			0.00831	0.0103
10	Ir-1,2,3	Ir-4,5	0.0861	0.1154	Mean			0.0691	0.0898
Standard deviation			0.0152	0.0171	Standard deviation			0.00831	0.0103

Fig. S10. Model performance metrics include the MAE and RMSE values, with their means and standard deviations, obtained in the cross-validation studies for the cross-photocatalyst prediction (A) and cross-substrate prediction (B).

Comment 4: Figure 3 is unreadable due to small font sizes and tiny panels.

Response:

We appreciate the Reviewer for the reminding about Figure 3.

We apologize that the images sizes of Figure 3 were too small for reading. We have revised Figure 3 and split it into two figures, as the new **Figure 3** and **Figure 4** in

the revised version.

A large amount of data is contained in **Figure 3B** in the revised version, including the recording of all non-steady-state absorbance peak signals, so the size of each plot is largely limited by the page size. In order to maximize the size of each plot, we have modified the figure format from a horizontal version to a vertical version, making the size area of each plot increase to 3 times its original size and the font size is 2 sizes larger. To make it easier for readers to understand, three typical data plots under different conditions are used as examples, enlarged and placed on the top of Figure 3B.

In order to make the ML model section clear, some modifications have been made in the new **Figure 3** and **Figure 4** in the revised version. First, we have added Figure 3A2, showing a typical steady-state absorbance recording and a typical non-steady-state absorbance recording including 40 absorbance data points. Second, in order to illustrate the inputs, targets, and outputs of the ML models, the figure about the large dataset, which was originally placed in Extended Data Figure 4 in Methods, has been moved to Figure 4A with the inputs and targets labeled, respectively. The model outputs and targets are illustrated (Fig. 4B1), and the R^2 and RMSE values for each ML model were calculated according to these data (Fig. 4B2). Based on the relationship between the GC product yields and the steady-state absorbance data, the outputs of the optimal model (i.e., predicted steady-state absorbance data) were transformed into product yields and presented as a heatmap (Fig. 4C).

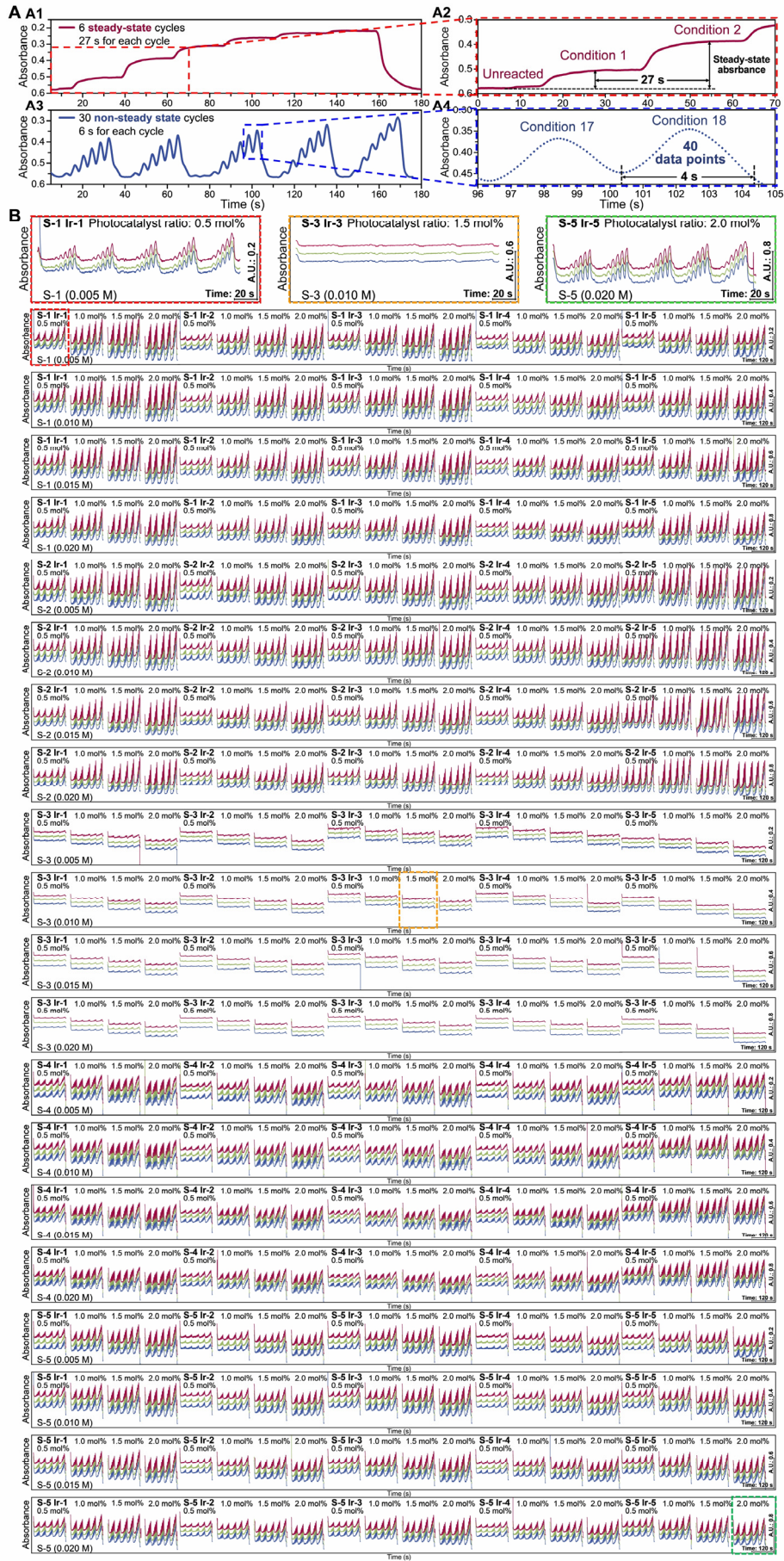


Fig. 3. Recordings of the typical and total absorbance signals obtained in the large-scale screening experiment. (A) Typical recordings of the steady-state and non-steady-state absorbance signals obtained in the screening for S-5, with a photocatalyst ratio of 2 mol% and a S-5 concentration of 0.01 M. (A1) Typical absorbance recordings of 6 different reaction conditions under the steady-state screening mode. (A2) Enlarged view of the absorbance recordings for the first two conditions of the 6 reaction conditions in (A1). It took an average time of 27 s in each condition cycle to obtain the steady-state absorbance signal, which was calculated from the difference between the reacted steady-state plateau absorbance and the unreacted blank absorbance. (A3) Typical absorbance recordings of 30 different reaction conditions under the non-steady-state screening mode. (A4) Enlarged view of the absorbance recordings for the 17th and 18th conditions of the 30 reaction conditions in (A3). Each condition cycle took an average time of ca. 4s, with a non-steady-state signal peak containing 40 absorbance data points, obtained using the laser pulse irradiation method. (B) Recordings of the non-steady-state peak signals obtained in the screening experiment of the total 12,000 reaction conditions, which was replicated three times to test the repeatability.

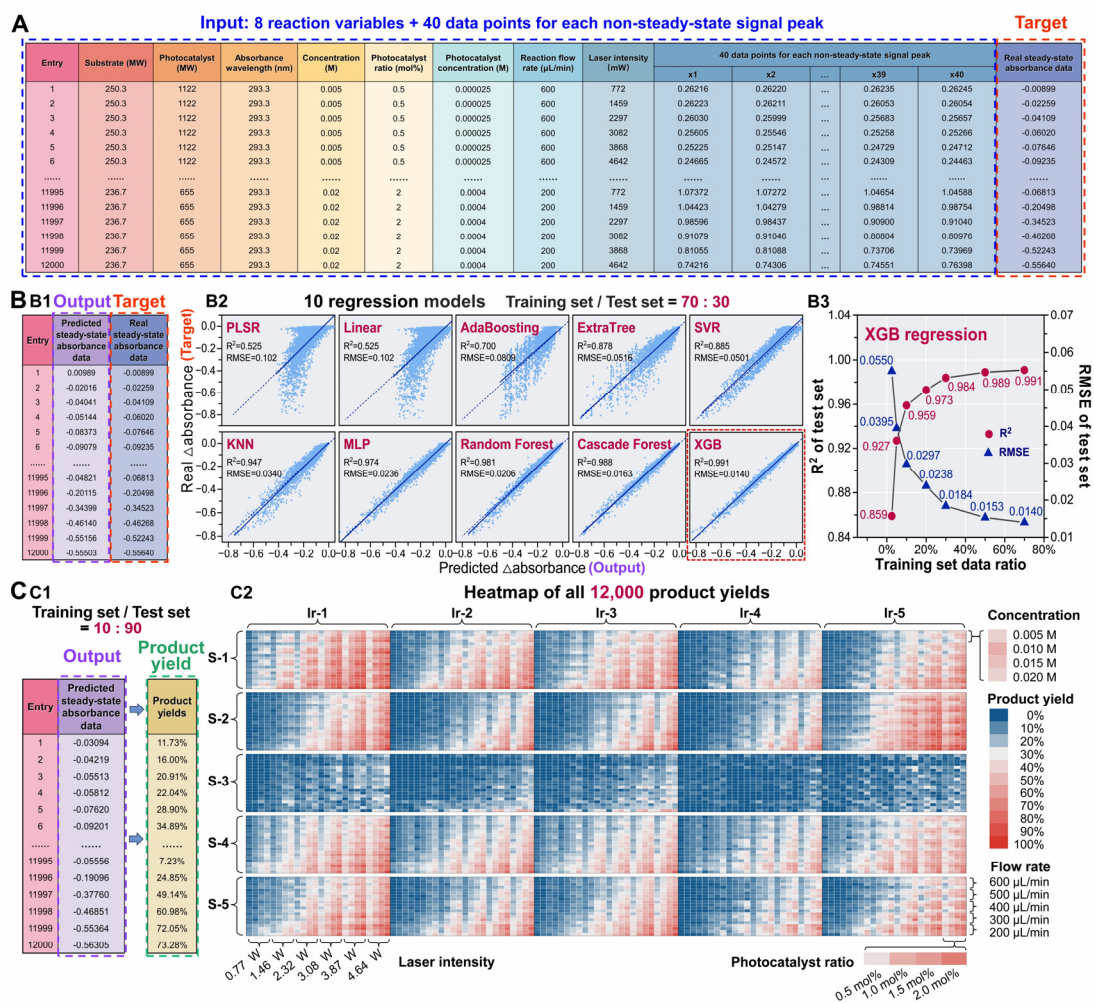
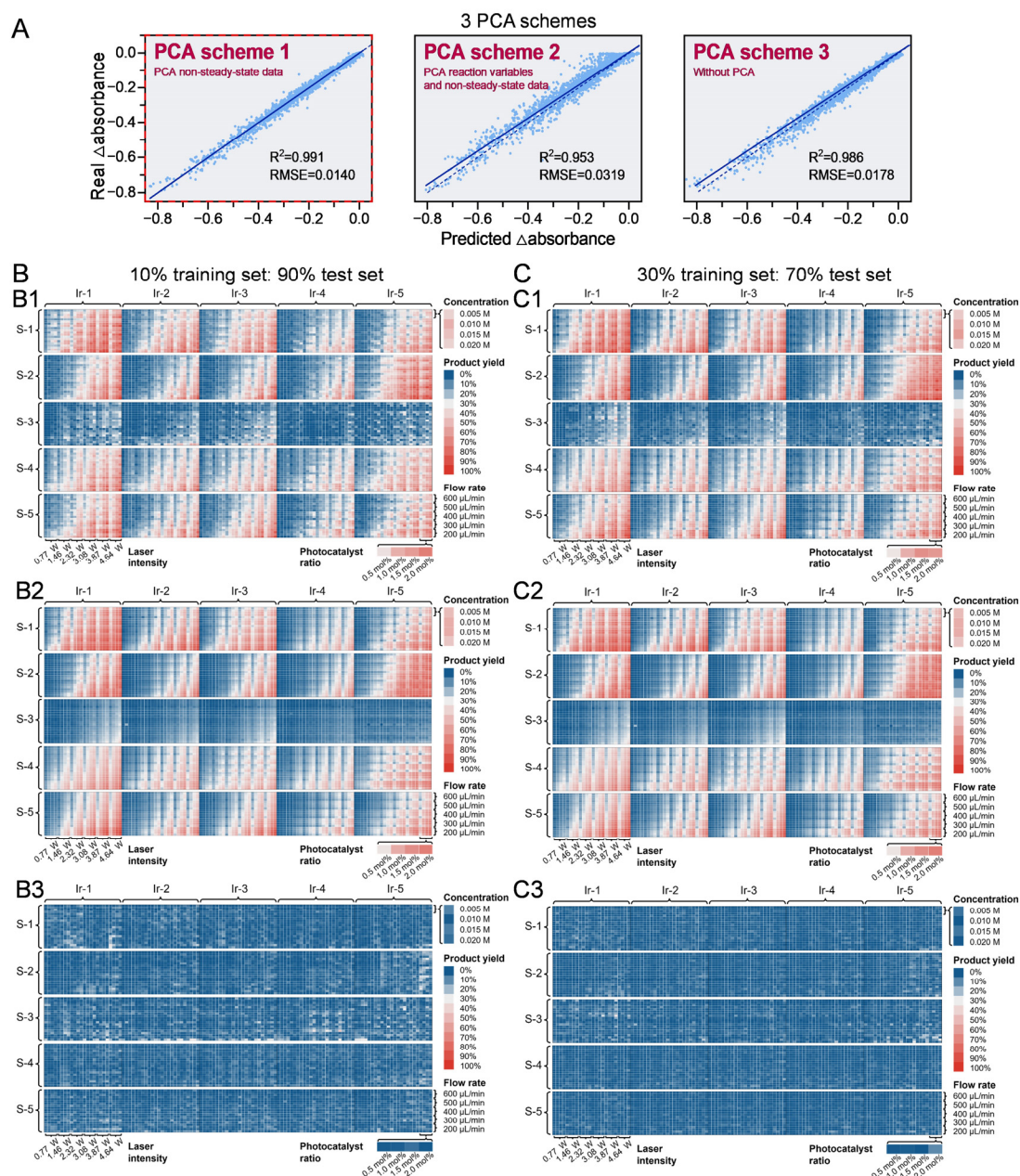


Fig. 4. AI-assisted prediction of steady-state absorbance data from non-steady-state absorbance data for condition screening of photocatalytic [2+2] cycloaddition. (A) Composition of the large dataset used in the AI-assisted steady-state absorbance prediction, including the 12,000 absorbance data with 48 eigenvalues (8 reaction variables and 40 data points for each non-steady-state signal peak) and 1 target (steady-state absorbance data). (B) AI-assisted prediction models for predicting the steady-state absorbance data from the non-steady-state absorbance data. (B1) Predicted and real steady-state absorbance as the outputs and targets of the models, respectively. (B2) 10 different regression models for the prediction of the steady-state absorbance data from the non-steady-state absorbance data, including partial least squares regression (PLSR), linear regression, adaptive boosting (AdaBoosting) regression, extremely randomized trees (ExtraTree) regression, support vector regression (SVR), k-nearest neighbor (KNN) regression, multi-layer perceptron (MLP) regression, random forest regression,

cascade forest regression, and extreme gradient boosting (XGB) regression models. The dashed line is the $y=x$ line, and the solid line is a linear fit curve between the predicted and the true steady-state absorbance values. The two metrics, R^2 and RMSE values, were calculated using the corresponding functions within `sklearn.metrics`. (B3) Performance of the test set of the XGB regression model. The prediction accuracy of the test set gradually decreases as the proportion of randomly selected training set data decreases from 70% to 2.5%. (C) Screening results of the total 12,000 reaction conditions obtained using the AI-assisted steady-state absorbance prediction method. (C1) Predicted steady-state absorbance data output from the XGB model with training set : test set = 10 : 90 were converted to product yields. (C2) Heatmap showing the screening results of a total of 12,000 reaction conditions, including the orthogonal combination of 5 substrate species, 5 photocatalyst species, 4 concentrations, 4 photocatalyst ratios, 5 flow rates, and 6 laser light intensities under the non-steady-state mode.



Extended Data Figure 4. AI-assisted prediction of the absorbance data and heatmap visualization of the product yields. (A) Three PCA schemes for downscaling the original eigenvalues using the XGB model, with the dataset divided into training and test sets with a proportion of 70% and 30%. (B) The heatmaps of the product yields converted from the non-steady-state absorbance data (B1), the steady-state absorbance data (B2), and the absolute errors between them (B3), with the dataset divided into training and test sets with a proportion of 10% and 90%. (C) The heatmaps of the product yields converted from the non-steady-state absorbance data (C1), the steady-state absorbance data (C2), and the absolute errors between them (C3), with the

dataset divided into training and test sets with a proportion of 30% and 70%. The product yields are displayed in color gradient with blue indicating low product yields, and red indicating high yields.

Comment 5: Figure 4 is extremely hard to understand, should be cut into pieces probably and overhauled completely.

Response:

We appreciate the Reviewer for the comment about Figure 4.

We apologize that the previous Figure 4 is difficult to understand. In order to help understanding, we have added some related descriptions and optimized the figure design by dividing it into two figures (new Figure 5 and Figure 6) with several pieces cut in each figure. Since Figure 3 has split into 2 figures as we responded in *Comment 4*, the original Figure 4 is now **Figure 5** and **Figure 6** in the revised version.

Specifically, Figure 5A contains product yields of all 12,000 conditions in the form of multidimensional bubble plot, showing the effects of all discrete and continuous variables on the product yields. Box plots (Fig. 5B and 5C) show the effect of photocatalyst species and concentration on product yields of all 12,000 conditions, respectively. The results of the increased concentration experiments are shown in Figures 5D and 5E, including the yields and *d.r.* values obtained both in batch and flow synthesis mode.

The dataset for AI-assisted cross-species prediction is illustrated in Figure 6A, along with input and output information. As part of the cross-species prediction, some results of the AI-assisted cross-substrate and cross-photocatalyst prediction are demonstrated in Figures 6B, 6C, with the remaining cross-species prediction results shown in the Supplementary Information (Fig. S9, Fig S10).

In the revised version, the description about the AI-assisted cross-species prediction section in **Figure 6** has been modified as “For cross-species prediction of product yields, the inputs of the models were the reaction conditions (i.e., the 8 reaction variables), the targets were the product yields obtained experimentally, and the outputs

were the predicted product yields (Fig. 6A). Differing from the AI-assisted absorbance prediction, the cross-species prediction required detailed chemical structure information of the substrate and photocatalyst species, which were described, digitized, and input to the models using the SMILES strings to generate Mordred descriptors (Extended Data Figure 5). The Mordred descriptor dimensions of each substrate and photocatalyst species were reduced to 2 in order to match the dimensions of the other variables and facilitate visualization in subsequent study, as described in Methods. The inputs consisted of 10 dimensions, with 2 representing the substrate species, 2 representing the photocatalyst species, and the remaining 6 representing the other variables (Fig. 6A).

Both the results of the cross-species prediction with different training set ratios and the cross-validation studies are described in the Supplementary Information (Fig. S9B, Fig. S10). As a typical result of the cross-substrate prediction, we used the data of S-1, S-2, and S-4 as the training set to predict the yields of S-5, achieving MAE=0.0698 and RMSE=0.0878 (Fig. 6B1). The distinct effects of the 5 photocatalyst species on S-5 were accurately predicted and Ir-1 showed to be the optimal photocatalyst, which is consistent with the experimental results (Fig. 6B2). With a smaller training set of S-1 and S-2, the prediction for S-4 and S-5 achieved with MAE=0.0772 and RMSE=0.0999 (Fig. 6B3, 6B4). For the cross-photocatalyst prediction, the data of Ir-1, Ir-2, Ir-4 and Ir-5 could be used to predict the yields of Ir-3, with MAE=0.0364 and RMSE=0.0497, which presented similar results to the real product yields (Fig. 4C1, 4C2). When the training set was reduced to include three photocatalysts of Ir-2, Ir-4 and Ir-5, pretty good prediction for Ir-1 and Ir-3 could still be obtained with MAE=0.0667 and RMSE=0.0898 (Fig. 4G3, 4G4). ” on page 23, line 370-381.

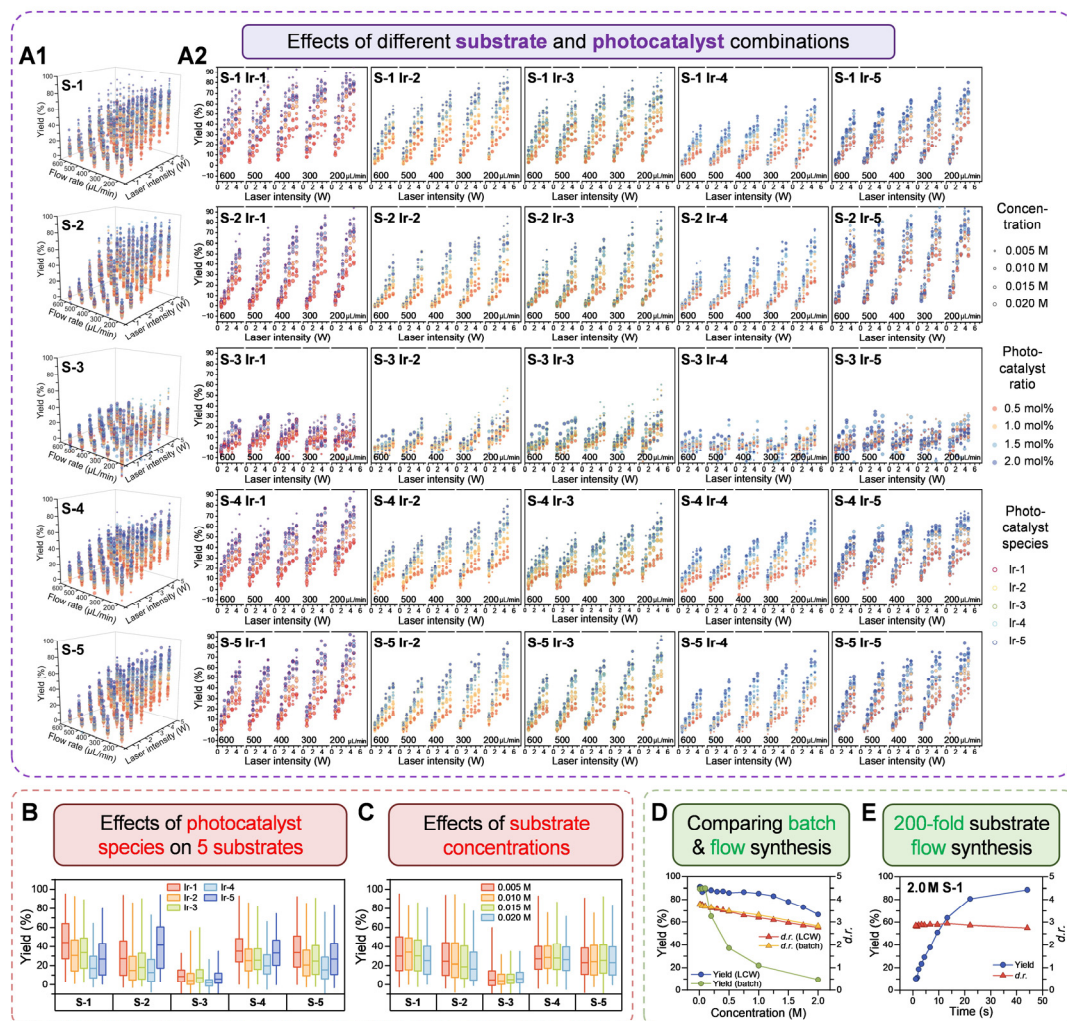


Fig. 5. Screening results of 12,000 reaction conditions. (A) Multidimensional bubble plots of the 12,000 data, showing the effects of different variables on the product yield, including the species, concentrations, and ratios of the substrates and photocatalysts, laser light intensity, and flow rate of the reaction solutions. Each 6-dimensional bubble plot in (A1) contains 2400 product yield data of 1 substrate specie and 5 photocatalyst species. Each 5-dimensional bubble plots in (A2) contains 480 product yield data of 1 photocatalyst and 1 substrate specie, corresponding to a data square in the heatmap shown in Figure 4C2. In each 5-dimensional bubble plot, the colors of the contour lines of the bubbles represent different photocatalyst species, the colors filled in the bubbles represent different photocatalyst ratios, and the sizes of the bubbles represent different substrate concentrations. (B) Box plot showing the effect of the photocatalyst species on the product yields of the 5 substrates. The three horizontal lines of each box from

top to bottom represent the first, median, and third quartiles of the product yield data, respectively. (C) Box plot showing the effect of the concentrations of the 5 substrates on the product yields. The three horizontal lines of each box from top to bottom represent the first, median, and third quartiles of the product yield data, respectively. (D) Comparisons of the variations of the product yield and *d.r.* with the increase of the S-1 concentration using the batch and present flow methods. (E) Variations of the product yield and *d.r.* with the increase of the residence time in the LCW microreactor using S-1 with a high concentration of 2.0 M, which is 200 fold of that in conventional batch systems.

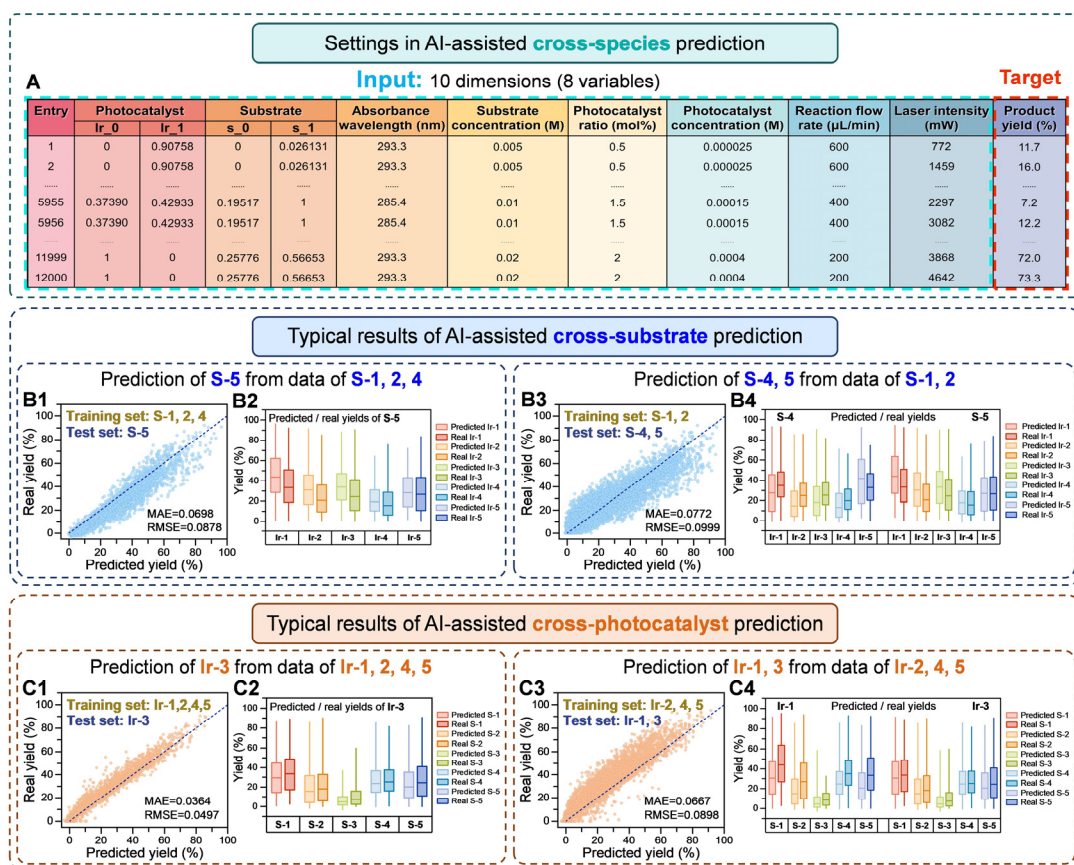


Fig. 6. AI-assisted cross-species prediction. (A) Settings used for AI-assisted cross-species prediction, including inputs and targets of the models. (B) AI-assisted cross-substrate prediction, including prediction of S-5 from data of S-1, S-2, and S-4 (B1), and prediction of S-4, S-5 from data of S-1 and S-2 (B3), with box plots showing product yields results (B2, B4), respectively. (C) AI-assisted cross-photocatalyst

prediction, including prediction of Ir-3 from data of Ir-1, Ir-2, Ir-4 and Ir-5 (C1), and prediction of Ir-1, Ir-3 from data of Ir-2, Ir-4 and Ir-5 (C3), with box plots showing product yields results (C2, C4), respectively. The model performance metrics for the cross-validation studies include the MAE and RMSE values. The entire results of the cross-species prediction are shown in Figure S9.

Comment 6: In general, I suggest the authors overhaul the ML parts of the manuscript to make them clear, accompanied by concrete plots that actually check the main hypothesis, which according to the test is predicting steady-state abs. from non-steady state signal + conditions, with proper cross-validation.

Otherwise, the authors can remove the ML part, which at the moment does not feel like an integrated part of the paper, and just carry out detailed data analysis (as hinted in Figure 4). The experimental platform and the generated data is good and valuable on its own.

Response:

We appreciate the Reviewer for the suggestions and for recognizing the capability of our experimental platform and the quality of our data. We apologize for the imprecise statement about the ML part, both in the text and in the figures, which led to some misunderstanding. In responding to the comments of the Reviewer, we have overhauled the ML parts to make it clearer and added descriptions of the relationship between ML experiments and robotic system.

As we responded to the *Comment 1* of the Reviewer, there are two applications of AI models in this work, **AI-assisted absorbance prediction** and **AI-assisted cross-species prediction**. In fact, the application of the **AI-assisted absorbance prediction** is very helpful for this robotic system to break through the limitations of the screening throughput reported so far. With the aid of the ML absorbance prediction models, it is possible to analyze the complex influences on the convection and molecular diffusion effects in flow synthesis and predict steady-state absorbance data accurately from the non-steady-state absorbance data and the reaction variables, thus significantly

improving the system's screening throughput for 2,600 to 10,000 conditions per day. Based on the Reviewers' valuable suggestions, we have modified the ML parts to make it clearer as we responded in *Comment 1*, and added the cross-validation studies to complete the ML section as we responded in *Comment 2* and *Comment 3*.

As suggested by the Reviewer, we have added the related description about the importance of the ML models to improve the screening throughput in the revised version, described as “These results showed that with the use of the non-steady-state mode and the AI-assisted absorbance prediction method, the long-standing challenge limiting the improvement of screening throughput of flow chemical screening systems caused by inefficient and time-consuming condition switching could be solved. Correspondingly, the screening throughput for the photocatalytic [2+2] cycloaddition reaction conditions increased from 2,600 to 10,000 conditions per day, which is the highest level reported in the field of organic synthesis so far.” on page 16, line 252-257.

In addition, in the form of subheadings, we have added two sub-headings in the part "**AI-assisted ultra-high-throughput photocatalytic synthesis and screening**" in the revised version, as “*Steady-state and non-steady-state experimental mode*” (on page 10, line 143) and “*AI-assisted absorbance prediction*” (on page 13, line 180), to facilitate readers' understanding.

As for the section of **AI-assisted cross-species prediction**, since we obtained large-scale of 12,000 high-quality data, for further utilizing these data, we attempted to perform AI-assisted prediction of product yields cross-substrate and cross-photocatalyst to preliminarily explore the potential possibility of applying AI technique to intelligent chemical synthesis screening. As shown in Figure 6, Figure S9 and Figure S10, although the attempts were preliminary, the results were encouraging, with some good results obtained in both cross-substrate and cross-photocatalyst predictions. We hope that these results may provide some inspiration for the study of intelligent chemical synthesis screening.

In responding to the comment of the Reviewer, we have added the related description about the aim of the AI-assisted cross-species prediction as “To further

utilize the above 12,000 data and preliminarily explore the potential possibility of applying AI technique to intelligent chemical synthesis screening, we used the XGB algorithm to perform AI-assisted prediction of product yields cross-substrate and cross-photocatalyst.” on page 21, line 344-346. In the revised version, according to the Reviewers' valuable suggestions, we have specifically used a new figure (Figure 6) showing the parameters and results of the cross-species predictions to illustrate this section more clearly. We have also modified the ML parts to make it clearer, and added the cross-validation studies to complete the ML section as we responded in *Comment 2* and *Comment 3*.

Comment 7: In my opinion, an interesting way to combine their platform with ML would be to use the 12,000 datapoints to predict yield, as it is implied in "AI-assisted cross species prediction" (but explaining clearly how everything is represented in the ML model inputs) then use an architecture with uncertainty, then use this in a bayesian opt. setting for quick optimization towards new products, for instance. At the moment, much of the latter sections of the paper feels either blurry and unclear or plainly disconnected (i.e. we can use these 12,000 data points for ML, but thats it).

Response:

We are very grateful to the Reviewer for the valuable insights and indicating a very promising direction for our future research.

In this work, we mainly focused on building a fully automated high-throughput platform enabling automated chemical photocatalytic synthesis, characterization and screening with the ability to generate large-scale data. As a preliminary extension of this main work, we used the acquired data to initially explore the potential possibility of using AI techniques for cross-species prediction. In the revised version, we have clarified the aim of performing the cross-species prediction as “To further utilize the above 12,000 data and preliminarily explore the potential possibility of applying AI technique to intelligent chemical synthesis screening, we used the XGB algorithm to perform AI-assisted prediction of product yields cross-substrate and cross-

photocatalyst.” on page 21, line 344-346.

We also realize that there is still a lot of work to do in further exploring the large amount of data and utilizing the cross-species prediction method. The inspiring suggestions of the Reviewer pointed us in a meaningful research direction. In the future work, we will continue to explore this data deeper and conduct more and widespread attempts, to explore more possibilities of intelligent synthesis, such as using an architecture with uncertainty and Bayesian optimization setting for quick optimization of new products, as pointed by the Reviewer.

In response to the Reviewer’s comment, we have added a description about this in the conclusion in the revised version, as “As we initially demonstrated in AI-assisted cross-species prediction, such a large amount of data from the same experimental system could provide a solid data base for AI applications. In the future, it would be meaningful to make full use of the 12,000 data and further incorporate AI techniques, such as Bayesian optimization method for rapid optimization of new products.” on page 24, line 394-398.

Responses to the comments of the Reviewers

Responses to the comments of Reviewer #2

Comment 1: The authors have devoted significant work to overhaul the ML/AI parts of the manuscript following my suggestions, clarifying what was done, how and why. I appreciate their effort and congratulate them for this work.

Nonetheless, I still think the AI/ML parts only serve an exemplary purpose and do not really add much to the story, since they are not used proactively but rather retrospectively.

In any case, I would ask the authors to give the new text blocks an additional read to smooth the writing, which is a bit weird at some points.

Response:

We sincerely appreciate Reviewer #2 for the recognition to our revisions to the manuscript and the comment.

According to the Reviewer's comment, we have carefully read through both the **AI-assisted absorbance prediction** and **AI-assisted cross-species prediction** sections and modified some descriptions to further improve the smoothness of the writing in the revised version as “Due to the convection and molecular diffusion effects existed in the flow system, the previous reacted solution and the newly-introduced unreacted solution would mix with each other at their junction region, and the absorbance signals detected by the UV-Vis detector exhibited a dynamically-changing format (Fig. 3A1) during the switching process **of different solutions**. The conventional method is to wait for the newly-introduced **unreacted** reaction solution to flush all of the previous **reacted** solution out of the flow system to enable the detector to obtain a stable plateau-type steady-state absorbance signal for data reading.” on page 7, line 135-142;

“To increase the efficiency of time utilization and screening throughput, we proposed the strategy of non-steady-state experimental mode instead of the steady-state mode by using the laser pulse irradiation method to **turn the irradiation laser on and off for achieving the rapid switching between the reacted and unreacted solutions**, producing a series of non-steady-state continuous peak-shaped signals as shown in

Figure 3A3.” on page 8, line 152-156;

“For achieving this complex and challenging task, we developed the AI-assisted absorbance prediction method by using the AI method to analyze the influencing factors related to the convection and molecular diffusion effects and decoupling the non-steady-state data of the adjacent reaction solutions mixed with each other, to predict the corresponding steady-state absorbance data of the respective reaction solutions.

In order to obtain accurate prediction results, we attempted to use 10 regression models based on the principles of linear models, decision tree, neural networks and integrated learning, to process the large numbers of the non-steady-state absorbance data and to predict the corresponding steady-state absorbance data under the same reaction condition, from which we searched for the best-performing model.” on page 8-9, line 167-176;

“We tried to use less data to predict more data (such as 2.5% of the data as the training set and 97.5% of the data as the test set) to further test the predictive performance of the XGB regression model. Pretty good result was still obtained where 300 conditions data were used to predict the remaining 11,700 conditions with a RMSE of 0.0550 and R^2 of 0.859 (Fig. 4B3).” on page 9-10, line 192-196;

“Correspondingly, the screening throughput for the photocatalytic [2+2] cycloaddition reaction conditions increased from 2,600 to 10,000 conditions per day using the non-steady-state experimental mode, which is the highest level reported in the field of organic synthesis so far.” on page 10, line 100-203;

“We further increased the concentration of substrate S-1 using 50 $\mu\text{L}/\text{min}$ flow rate (i.e., 13.2 s residence time) in the present system, the results showed that the product yield decreased slightly from 91% to 85% when the substrate S-1 concentration increased from 0.01 M to 1.0 M, and further reduced to 67% when the substrate S-1 concentration reached its solubility limit of 2.0 M (Fig. 5D).” on page 12, line 254-259;

“The 12,000 experimental data were divided into training and test sets with different ratios for AI-assisted cross-species prediction.” on page 13, line 284-285.

For details of the revisions, please see the revised manuscript with the marked revisions.