PREDICTING SOLVENTS WITH THE HELP OF ARTIFICIAL INTELLIGENCE

Oliver Schilter[*], **Carlo Baldassari, Teodoro Laino**
IBM Research Europe,
8803 Rüschlikon, Switzerland
&
National Center for Competence in Research-Catalysis
(NCCR-Catalysis),
8093 Zürich, Switzerland

Philippe Schwaller
EPFL Lausanne
Laboratory of Artificial Chemical Intelligence (LIAC)
1015 Lausanne, Switzerland

ABSTRACT

The right solvent is a crucial factor in achieving environmentally friendly, selective, and highly converted chemical reactions. Artificial intelligence-based tools, often lack the ability to reliably predict reaction conditions such as the appropriate solvent. Here, we present a comprehensive investigation into the efficacy of data-driven machine-learning models for solvent prediction for a broad spectrum of single-solvent organic reactions. Remarkably, our models achieve a Top-3 accuracy of 86.88%, showcasing outstanding performance in predicting solvents from underrepresented classes. An uncertainty analysis revealed that the models' misclassifications could be explained by the fact that the reaction can be run in multiple solvents. In the experimental validation, 8 out of 11 reactions succeeded with the predicted solvent. Our work addresses a key challenge in organic synthesis and demonstrates the practical application of machine learning models in predicting reaction solvents for more efficient and sustainable chemical synthesis.

***K*eywords** Solvent prediction · Reaction Conditions · BERT · DRFP

1 Introduction

Chemistry plays a critical role in our daily lives, from the fabrics we wear and the medicines we take to the materials that surround us and the technologies we use. When it comes to chemical reactions, the choice of solvents can profoundly impact the reaction outcome, such as yield, purity, selectivity, and overall sustainability. A good summary of commonly used greenness criteria for solvent in pharmaceutical companies [2, 10, 16] or academia [6, 22] can be found in the "green solvent selection guide" by [5].

The optimal solvent for a given reaction depends on various factors, such as the reaction mechanism, the type of reactants, and the reaction conditions. However choosing the best solvent for a given reaction is a challenging task, as there are numerous solvents available, each with its own set of properties and interactions with the reactants and products. Solvents with low boiling points, generally preferable for reducing the complexity of workouts, evaporate easily, contributing to air pollution and increasing risks to the workers' health and the environment. While experimental solvent screening demands significant time and resources, atomistic modeling like quantum chemistry calculations offer the ability to predict solvent properties and interactions with reactants and products [23, 14]. However, quantum chemical calculations can be computationally intensive thus limiting the extensiveness of the sampling required to properly account entropy factors.

Machine learning (ML) has shown great potential in predicting solvents properties. Vermeire and Green [23] used a graph-based directed-message passing neural network (NN) to predict the solvation-free energies and demonstrate their model capability on an experimental dataset. Boobier et al. [4] trained a variety of machine learning models such as random forest (RF), NN, and support vector machines (SVM) using a small set of molecular descriptors describing the

[*] *Corresponding Author*: **oli@zurich.ibm.com**

dissolution process. These models demonstrated exceptional accuracy in predicting the solubility of organic solvents in water.

Sanchez-Lengeling et al. [18] employed Gaussian processes, a Bayesian machine learning approach, to derive the Hansen solubility parameters. These parameters are essential for predicting solubility between solute and solvent. The authors effectively utilized their model to predict solubility interactions among 193 solvents and 31 polymers or 8,000 organic compounds. This demonstrated the model's capability to learn the Hansen parameters from a collection of quantum descriptors and molecular fingerprints.

In the realm of predicting reaction conditions, the prediction of the missing solvent has also been explored, with particular case studies conducted on various reaction classes. For instance, Walker et al. [24] reported the successful implementation of neural networks trained to predict solvents for Friedel-Crafts, Aldol addition, Claisen condensation, Diels-Alder, and Wittig reactions. Shim et al. [21] investigated the use of transfer learning for predicting reaction conditions, including solvent prediction, in palladium-catalyzed coupling reactions. Gao et al. [9] used Reaxys [1] to extract information regarding reactants, solvents, and catalysts involved in a reaction and trained a neural network to predict in sequence catalysts and subsequently solvents and reactants. They achieved a top-3 accuracy of 75.8% for predicting the major solvent.

Recently, Beker et al. [3] challenged the paradigm, that ML can be used to find optimal reaction conditions, by demonstrating that a database of carefully curated literature data may be insufficient to create accurate and meaningful ML models. The authors illustrate this point by examining the prediction of optimum reaction conditions for Suzuki-Miyaura coupling with heterocyclic building blocks. Even when restricting the search space to solvents and bases, the proposed ML models failed to make predictions better than naive assignments based on the sheer frequency of certain reaction conditions.

Transformed-based neural networks, including BERT models, have emerged as leading performers in classifying text-based inputs into distinct categories [8]. In the field of chemistry, BERT models have demonstrated success in tasks such as yield prediction [20] and reaction classification [19]. Recently, researchers have applied transformed-based neural networks to predict solvation-free energy. SolvBERT, introduced by Yu et al. [25], represents the latest advancement in this area, achieving state-of-the-art performance by leveraging the SMILES of a solute and the solvent to predict either experimental solvation-free energy or solubility.

Despite recent advancements in machine learning techniques to predict reaction solvents, the accuracy and reliability of such models are still limited. To address this issue, we have developed a BERT-based classifier and simpler machine learning classifiers using DRFP [17] to reliably predict the missing solvent in a reaction. Our models show improved accuracy in predicted solvents compared to existing reaction condition models [9]. Furthermore, our approach outperforms the models proposed by Beker et al. [3] and their popularity baseline in predicting solvent classes. One notable feature of our BERT-based models is the analysis of prediction uncertainty using Monte Carlo dropout, which revealed that most of the uncertainty is attributed to the use of multiple solvents in the training data for certain reaction classes. We report an experimental validation campaign that showcases the models' ability to correctly predict the solvent in 8 out of 11 cases.

2 Methods

2.1 Data-set analysis and prepossessing

We used two main datasets, namely the open-source USPTO dataset from Lowe [12] and the Pistachio dataset from NextMove (version 2022Q4) [13], both of which were curated by extracting information from reaction procedures found in patents. To ensure consistency and quality of the data, we used the RXN reaction preprocessing pipeline [2] to remove atom mapping, eliminate duplicate reactions, add reagents to the reactants, and canonicalize each molecule in the reaction string. After preprocessing, the Pistachio dataset contained 3'996'348 reactions, while the USPTO dataset contained 1'435'481 reactions.

To determine the solvents used in each reaction, we compiled a list of 227 solvents based on two solvent sets found in the literature [18, 3]. This concatenated list included the SMILES string of each solvent, which was then canonicalized for matching with the SMILES string of each molecule in a reaction. Each reaction in the Pistachio dataset was processed to detect possible solvents, and subsequently group them into one of three categories: reactions containing no solvents, reactions containing exactly one solvent, and reactions containing more than one solvent. It is worth noting that since duplicate molecules were removed during preprocessing, the number of solvents in a given reaction corresponds to the unique number of solvents used. Reactions without a solvent were excluded from this study, given the uncertainty of

[2]Available at `https://github.com/rxn4chemistry/rxn-reaction-preprocessing`

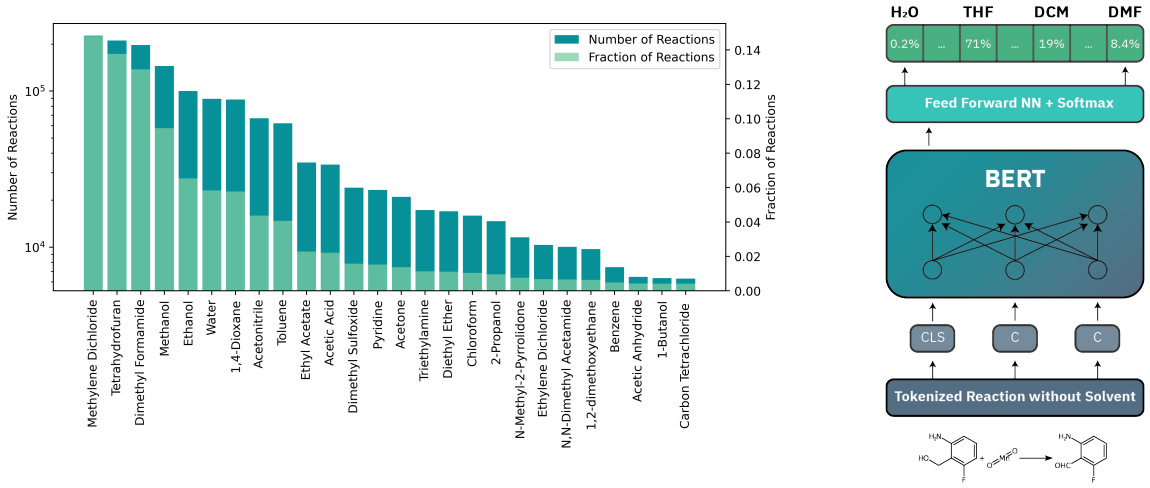


Figure 1: (left) Distribution of the 26 most used Solvents in the Pistachio dataset. (right) The BERT architecture used in our study employs tokenized reactions SMILES as input for a classifier neural network, which is jointly trained. The [CLS]-embedding is utilized for generating input to the classifier head, which learns to predict the probabilities of each solvent class.

being run without any solvent or the solvent not being extracted correctly from the reaction procedure paragraph. We analyzed a subset of the dataset consisting of reactions involving multiple solvents. Nevertheless, during inspection of the corresponding paragraphs, we noticed instances where one of the listed solvents was not used as a reaction solvent but rather as a solvent for workup steps, such as product extraction or water in crystal structures of reagents. To mitigate this limitation, we decided to focus primarily on reactions that contained only one solvent for the main part of our study. Table 1 provides a summary of the number of reactions in each subset.

Among the 3'996'348 reactions in the Pistachio dataset, 1'530'590 reactions contained exactly one solvent. These reactions were analyzed to determine the distribution of the 227 solvents. As shown in Figure 1, the 26 most common solvents account for 95.2% of the reactions with a single solvent, while the top 50 solvents cover 98.4% of all the one-solvent cases. Here, we aim to predict the 26 most commonly used solvent classes, with additional experiments conducted using the top 50 most common solvents.

We excluded reactions that did not involve any of the 26 specified solvents, as this 4.8% of the data might have unique solubility characteristics. Creating a separate category to encompass all such cases would not be ideal, as the model could fail to capture the underlying trends due to the varying solubility properties.

The extracted solvent was removed from the reactions SMILES to form the input for the machine learning models. After this prepossessing the data was split into 90% training data, 5% validation data, and 5% test data.

Table 1: Number of Reactions containing zero, one, or two solvents

Number of Reactions	Pistachio	USPTO
Unique Reactions	3'996'348	1'435'481
Reactions without a solvent	879'506	270'655
Reactions with 1 solvent	**1'530'590**	**561'817**
Reactions with 2 solvents	144'754	421'726
Reactions with more than 1 solvent	1'441'279	603'009
Reactions with more than 2 solvents	1'296'525	181'283
Reactions Train	1'410'064	522'223
Reactions Validation	78'337	29'012
Reactions Test	78'337	29'013

2.2 Machine Learning Architectures

In our study, we employed various machine learning models to predict the solvent of a reaction using representations with the absence of the solvent information. Specifically, we used BERT models, which were first pre-trained on a mask language modeling (MLM) task and then fine-tuned on the task of solvent classification. To achieve this, we utilized a pre-trained MLM model from Schwaller et al. [19] that has prior domain knowledge and was trained on a version of the Pistachio dataset [3].

Subsequently, we fine-tuned the selected BERT model by attaching a classifier head that identifies solvent class. We generated the input for the classifier head using the embeddings of the [CLS] token, which was already included during the MLM pretraining specifically for classification tasks. During the fine-tuning process, we optimized the model using the cross-entropy loss function and the AdamW optimizer[11]. We also utilized a learning rate scheduler to optimize the learning rate during training.

In addition to the BERT model, we also trained simpler machine learning classifiers using differential reaction fingerprints (DRFP) as input. Specifically, we trained k-nearest neighbors (KNN) clustering, XGBoost and Random Forest (RF) classifiers. Additional details on the training procedure and hyperparameters are provided in the supplemental information 5.3.

2.3 Molecular Representation

For BERT models, the reactions were represented as SMILES strings and tokenized into tokens corresponding to the same procedure and vocabulary used in the training of the MLM task, while the solvent was one hot encoded into n-classes. The one-hot encoding leads to a vector length of n-classes where each position corresponds to one of the classes. If a solvent is present the corresponding position in the vector is set to 1, while the non-present solvent positions are represented by 0.

For the models using the fingerprints as an input (Random Forest, XGBoost, K-NN Clustering) the DRFP of the reaction without solvent was constructed according to Probst et al. [17], additionally the embedding of the pooled last hidden state layer (so-called reaction fingerprint [19]) of the fine-tuned BERT model was tested as input for traditional machine learning approaches.

2.4 Uncertainty Analysis

To assess the uncertainty of our BERT models' predictions, we performed a Monte Carlo dropout-based analysis. This involved reactivating the dropout layers of the neural network during inference, resulting in different predicted probabilities for each solvent class when the same reaction is sampled multiple times. We used this method to predict the solvent for each reaction in the test set 25 times and averaged the probabilities for each class over these 25 samples.

To investigate cases where the model was uncertain about its prediction, we performed a reaction subclass analysis. Each reaction in the pistachio dataset has a reaction class associated with it (e.g., "reductive amination"). We went iterated over the entire dataset and listed the number of solvent occurrences for each reaction class, disregarding solvents with an occurrence below 5% for a given reaction class, due to their relatively rare nature.

For each sample in the test set, we analyzed the average predicted solvent probabilities of the twenty-five predictions. If a solvent had an average probability above 20%, we investigated whether it was a valid choice for the given reaction class. If the solvent was found among the frequently occurring solvents for the reaction class in question, we assigned it to the valid suggestion category. If not, we assigned it to the not valid suggestion category. For example, if a reaction belonging to the reaction class "reductive amination" had an average solvent probability of 40% for methanol and 30% for ethanol, we checked whether methanol and ethanol are commonly used solvents for reductive amination in our dataset. If they were, we assigned them as valid predictions, even if the model was not certain about them, since it is likely that the reaction can be run in multiple solvents.

2.5 Experimental validation

A machine learning model is only as useful as their real-world implication, which in our case is the model's ability to reliably predict the solvent in which the reaction runs and the desired product is formed. As chemical reactions are influenced by various parameters, including temperature, concentration, and atmosphere, it is not sufficient to test a reaction in a predicted solvent and assume that if the desired product is not found it is solely due to the incorrect solvent

[3] Models available under `https://github.com/rxn4chemistry/rxnfp/tree/master/rxnfp/models/transformers/bert_pretrained`

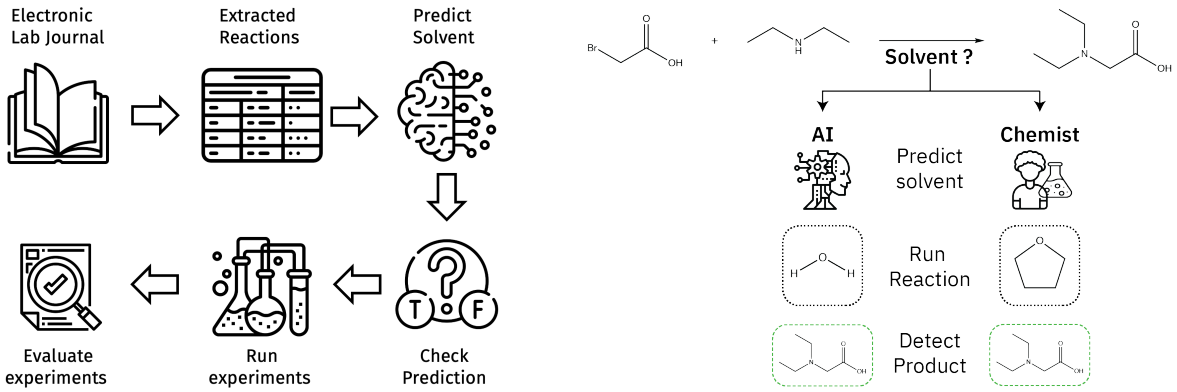


Figure 2: Experimental workflow: Successful reactions were extracted from the electronic lab notebook. From these reactions, solvent information was removed and then subjected to solvent prediction using the BERT model. Subsequently, a subset of reactions with incorrectly predicted solvents was selected for experimental validation using the model's recommended solvent and the original solvent choice of the Chemist. The presence of the desired product was confirmed by analyzing the mass spectrum. Reactions yielding detectable product peaks in the mass spectrum were labeled as successful.

prediction. Instead, we leveraged our electronic lab notebooks (ELN) to identify reactions that were successfully run in a solvent chosen by an expert chemist. We then used both our DRFP-based Random Forest and BERT models (trained on Pistachio 26 Solvents) to predict the solvent for these reactions, after removing the solvent from the reactions SMILES. Since the BERT model achieved higher accuracy on the ELN data, we used it as the predictor for this campaign (8 of 26 reactions = 30.76% vs. 6 of 26 reactions = 23.08%). Next, we compared the predicted solvent with the originally used solvent for each reaction. When the predicted solvent differed from the original solvent, we identified them as potential candidates for experimental validation. To validate these predictions, we ran the reaction again, once in the original solvent as a control and once in the predicted solvent, while keeping all other conditions, such as concentration, temperature, and reaction time, the same (see Figure 2). We selected 12 reactions to be screened, based on spanning a diverse set of reaction classes and differences in the predicted and experimental solvents.

3 Results & Discussion

3.1 Solvent Prediction

The models' predictive performance was evaluated by analyzing their Top-n scores to identify the most accurate predictions, and their ability to detect deviations from statistical trends was measured using F1, recall, and precision scores as defined in Section 5.1 of the Supplementary Information. The results of the evaluation are presented in Table 3.1, which shows the performance of the models and input representations on the two datasets' test sets. While the models demonstrated similar predictive ability overall, the Random Forest model using DRFP as a molecular representation achieved the highest Top-1 accuracy in both the USPTO and Pistachio datasets. Interestingly, the BERT-based models exhibited more balanced learning for the underrepresented solvents in the Pistachio dataset, resulting in improved F1 scores compared to the DRFP-based models. To benchmark our performance against existing models, we compared our results with the models developed by Gao et al. [9] capable of predicting reaction solvents, temperature, and catalyst using a reaxys-derived dataset. To account for the difference in solvents between their dataset and ours, we assumed that any solvent predicted by their model, not included in our 26 solvents, was correct. Additionally, for multiple solvent predictions, we evaluated each solvent individually and treated the prediction as correct if at least one of the predicted solvents matched the true solvent. We can see that our model is significantly better in predicting the solvent for a single solvent reaction. It should be noted that the model proposed by Gao et al. [9] achieved a higher accuracy on the solvent prediction task on Reaxys data, which is most likely caused by the fact that no-solvent reactions were permitted in their dataset.

3.2 Suzuki case study

Beker et al. [3] tried to utilize machine learning to predict ideal reaction conditions for the Suzuki cross-coupling. They divided this task into different sub-tasks of which one of which is the reaction solvent prediction. Instead of

Table 2: Predictive performance of the machine learning models classifying 26 solvents.

	Model	Input	Top 1	Top 3	Top 10	F1 macro	F1 micro	Precision	Recall
Pistachio	BERT	SMILES	66.20	85.96	**97.09**	**56.87**	**66.20**	62.06	**62.06**
	Random Forest	DRFP	**68.92**	**86.88**	96.14	52.98	62.42	**83.62**	49.79
	XGBoost	DRFP	60.70	82.62	**96.19**	45.55	54.80	78.37	42.13
	K-NN	DRFP	61.30	80.26	87.84	53.28	62.03	73.17	53.83
	Askcos [9]	Morgan FP	43.69	63.51	73.78	45.66	32.64	30.11	43.41
USPTO	BERT	SMILES	64.29	82.50	95.10	57.17	64.29	64.29	60.64
	Random Forest	DRFP	**65.39**	**85.10**	**95.92**	**59.26**	**65.35**	**65.35**	**63.35**
	XGBoost	DRFP	59.97	82.00	96.66	52.77	59.97	59.97	59.97
	K-NN	DRFP	54.99	75.14	83.87	49.39	55.63	68.80	55.63
	Askcos [9]	Morgan FP	49.21	68.85	77.80	39.41	51.06	51.17	35.79

Table 3: Predictive ability of various models for Suzuki cross-coupling solvent prediction task: A comparison of the feed-forward neural network, DRFP-based random forest, Bert models and popularity baseline predicting 6 and 13 solvent classes.

	Model	Input	6 Solvent classes			13 Solvent classes		
			Top 1	Top 2	Top3	Top 1	Top 2	Top 3
Reaxys[3]	Popularity baseline		29.8	57.4	75.5	29.7	41.4	52.6
	Feed-forward NN	Morgan FP	51.7	69.4	81.2	43.3	57.4	67.0
USPTO-Suzuki-RXN	Popularity baseline		37.4	63.0	76.8	37.2	62.5	76.7
	Feed-forward NN	Morgan FP	52.2	70.2	83.1	37.7	60.5	73.8
USPTO-Suzuki-RXN	Random Forest	DRFP	60.2	78.9	88.1	60.7	78.9	**87.5**
USPTO-Full	Random Forest	DRFP	**60.7**	**80.2**	**89.7**	60.8	**79.6**	87.1
USPTO-Suzuki-RXN	BERT	SMILES	50.8	68.8	82.6	53.4	70.4	77.7
USPTO-Full	BERT	SMILES	46.3	71.9	86.4	47.8	69.0	81.2
USPTO-Full finetuned	BERT	SMILES	55.3	73.1	84.4	58.8	70.8	79.2
USPTO-Suzuki-RXN	BERT	8xAug SMILES	56.7	73.9	83.0	60.4	76.1	80.8
USPTO-Full finetuned	BERT	8xAug SMILES	57.3	75.3	85.0	**62.8**	72.5	84.2

directly predicting the solvent they split the solvents into 6 classes ({alcohols, water/polar solvents, water/alcohols, water/amides, water, amides}, {water/aromatics, alcohols/aromatics, water/alcohols/aromatics}, {aromatics}, {ethers}, {water/ethers}, {other}) or 13 classes (water/ethers, ethers, water/alcohols/aromatics, water/amides, alcohols/aromatics, aromatics, amides, water/aromatics, low boiling polar aprotic solvents/ water, water/alcohols, water, alcohols, and other).

The performance of Beker et al. [3]'s best-performing model, a feed-forward neural network using the reactants and products Morgan fingerprints as molecular representation, was compared against a simple baseline called the "popularity baseline" on Suzuki reactions extracted from Reaxys. The "popularity baseline" predicts the most commonly occurring solvent in the training data as the prediction for all test cases. However, the neural network failed to achieve significantly higher accuracy than this baseline. To further evaluate their method, the authors extracted an additional 5,434 Suzuki reactions from USPTO [12] for reaction conditions testing. Our study benchmarked their best-performing Neural Network against our models for the solvent prediction task using these 5'434 USPTO Suzuki reactions for predicting the defined 6 and 13 solvent classes. We compared different varieties of our models: A DRFP-based Random Forest and the BERT architecture only using the 5'434 reactions with an 80:20 train:test split (USPTO-Suzuki-RXN), and the Random Forest and BERT architecture using the full scope of USPTO data (USPTO-Full). In the latter, we first removed all 5'434 Suzuki reactions from the full USPTO dataset, then after splitting it into training, validation, and test set, we added 80% of the 5'434 Suzuki reactions back into the training data and used this to train the BERT and Random Forest models to learn to predict the solvent classes. We evaluated directly the remaining 20% of the 5'434 Suzuki reactions (USPTO-Full) or did an additional finetuning of the BERT models on 80% of the Suzuki reactions (USPTO-Full fine-tuned) before evaluating on the 20%. We additionally implemented an 8-times SMILES augmentation of the Suzuki reactions for the fine-tuning step based on the yield prediction task improvement of Schwaller et al. [20].

The results, summarized in Table 3.2, indicate that the DRFP-based Random Forest model generally outperforms the BERT models. Furthermore, utilizing the USPTO-Full dataset helps enhance the predictive performance. In the case of the Bert models, finetuning improves the Top-1 accuracy, as well as that SMILES augmentation increases the accuracy and generalization ability of the models. Overall, our approach, particularly for the 13 solvent models, outperforms previously proposed models by Beker et al. [3]. We would also like to emphasize that predicting the correct solvent for the Suzuki cross-coupling is a particularly challenging task. The difficulty arises from the fact that the reaction class remains the same for all reactions, and the presence of a reaction in a solvent mixture does not imply that it cannot be conducted in a different solvent mixture.

3.3 Uncertainty Analysis

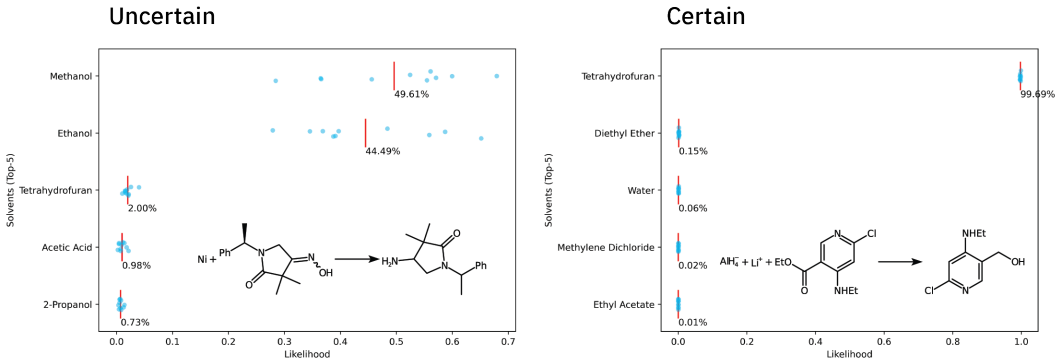


Figure 3: (left) Even though the model is uncertain about the prediction the two most probable solvents, ethanol and methanol, are commonly used solvents for the reaction class. (right) A reaction where the model is certain about is prediction since the solvent class with the highest probability doesn't change over the 25 samples.

The in section 2.4 introduced uncertainty analysis via Monte Carlo dropout allows us to understand how convinced the BERT-based model is in its prediction. Across the entire test set, the average probability of the highest three predicted classes was found to be $76.48 \pm 24.56\%$, $10.90 \pm 11.73\%$, and $4.27 \pm 5.52\%$, which indicates that the model is reasonably certain in it is prediction. To gain more insight into the uncertain predictions via further performed reaction subclass analysis (see section 2.4). The core concept of this is that reactions in the same reaction class, e.g. reductive amination, can be run in multiple solvents, therefore it is reasonable to assume that model is uncertain about its prediction. The reaction subclass analysis revealed that 79.79% of the predicted solvents with an average probability over 20% could be found as commonly used solvents for the given reaction class. On the left side of Figure 14 this behavior is exemplified, the solvents with an over 20% prediction (methanol and ethanol) both are commonly used solvents for the given reaction class.

Therefore, we conclude that when the dropout layers are activated during inference and the model changes its prediction, it is more likely because the reaction in question can be run in multiple solvents rather than the model's inability to predict the correct solvent with high certainty.

3.4 Experimental validation

We selected 12 reactions based on a variety of factors including the solvent type, chemical transformation, and availability of the required chemicals for experimentally validating our model's solvent prediction. These twelve reactions were previously run successfully in solvents chosen by an expert chemist but have a different predicted solvent (see Section 2.5). The selected 12 reactions and their experimental procedures can be found in the SI 5.5. The reactions were classified as either successful or not based on the presence of the expected product in the mass spectrum. Details of the reactions and their solvents and success can be found in the table provided 4. Out of the 12 reactions, 11 were successfully run under their original conditions. Among these working reactions, we observed that the product was obtained in 8 cases when the reactions were run in the predicted solvent. This results in a success rate of 72.72% in the run reactions, which if we assume that the correctly predicted reactions would have been run rate would actually be an 81.11% success rate on the overall ELN data (see equation 4). This indicates that our model can reliably predict the missing solvent of a broad spectrum of organic reactions and demonstrates the usefulness of such a model.

Table 4: Experimental validation of 12 reactions in two different solvents. If the product was found is marked by (S) or if the reaction failed by a (F).

	Reaction Type	Chemist		Bert Model	
1	N-methylation	Toluene	(S)	DMF	(S)
2	Steglich reaction	DMF	(S)	DCM	(S)
3	N-acylation to amide	THF	(S)	DMF	(S)
4	Williamson ether synthesis	DMF	(S)	THF	(F)
5	Schiff base formation	MeOH	(S)	Toluene	(F)
6	Williamson ether synthesis	Acetone	(F)	DMF	(F)
7	Thiourea cyclization	THF	(S)	Pyridine	(F)
8	N-acylation	THF	(S)	DCM	(S)
9	N-substitution	THF	(S)	Water	(S)
10	Schotten-Baumann reaction	DCM	(S)	THF	(S)
11	N-acylation to amide	THF	(S)	DCM	(S)
12	Steglich reaction	DCM	(S)	DMF	(S)

4 Conclusion

In this study, we successfully demonstrated the ability of our models to predict solvents with high reliability using both patent-derived data and experimental validation. Our evaluation of the models' performance on two datasets, Pistachio and USPTO, revealed that the Random Forest model using DRFP as a molecular representation achieved the highest Top-1 accuracy for solvent prediction in both datasets.

Furthermore, we compared the performance of our models with existing models developed by Gao et al. [9] and found that our models outperformed them in predicting the solvent for single-solvent reactions when the solvent information was missing. These results highlight the potential of machine learning-based methods in providing accurate and efficient solutions for predicting solvent identities in chemical reactions.

The experimental validation campaign showed that our models' predictions were in good agreement with the experimental results, confirming the models' reliability and accuracy. Our findings have important implications for the development of more sustainable and efficient chemical reactions, as an accurate prediction of solvent identities can greatly aid in the design and optimization of reaction conditions. Overall, this study demonstrates the potential of machine learning in advancing the field of chemical synthesis.

Acknowledgments

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

5 Supporting information

5.1 Classification Results

Dataset	Model	Input	Top 1	Top 3	Top 10	F1 macro	F1 micro	Precision	Recall
USPTO	BERT	SMILES	62.79	82.55	94.85	51.19	62.79	68.57	44.75
USPTO	RF	drfp	64.50	84.10	94.84	43.39	58.72	70.17	47.60
USPTO	xgboost	drfp	59.28	81.26	94.61	46.20	53.59	67.13	45.66
USPTO	knn	drfp	53.97	73.57	79.05	44.29	54.62	56.07	40.54
USPTO	RF	bertfp	63.03	81.63	92.34	49.99	63.01	61.79	48.49
USPTO	xgboost	bertfp	62.61	80.49	93.91	52.80	63.12	60.34	48.38
USPTO	knn	bertfp	62.54	72.82	75.75	52.75	63.40	57.24	49.61

5.2 Classification Metrics

Accuracy is a commonly used metric in classification, defined as the ratio of true positive and true negative predictions to the total number of predictions, as shown below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Here, TP represents true positives, FP represents false positives, TN represents true negatives, and FN represents false negatives. However, accuracy and top-k accuracy can be limited in their ability to evaluate unbalanced datasets in a multi-class classification problem. A more appropriate metric for assessing the predictive performance of reactions involving less common solvents is the F1 score, which is the harmonic mean of precision and recall, averaged in a macro or micro fashion. In macro averaging, the F1 score is determined for each class and then averaged with equal weight, while in micro averaging, the overall true positives, false positives, true negatives, and false negatives across all classes are used.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{2}$$

5.3 Machine learning architecture

5.3.1 Random Forest

We used the scikit-learn implementation [15] of the random forest classifier. We set the number of trees to 250 and the class weight to balanced to train more favorably on underrepresented classes. We kept the remaining hyperparameters at their default values. In our study, we aimed to assist chemists in selecting a suitable solvent. Therefore, instead of directly predicting the solvent, we chose the solvent with the highest probability as the predicted class. This approach differs from the implementation of [7, 15, 26], where a cutoff of 0.5 is used to determine if a solvent was predicted or not, enabling multi-label classification. However, this approach can lead to cases where the model is uncertain between two solvents, resulting in the failure to predict a solvent if all probabilities are below 0.5. Our approach is selecting the solvent with the highest probability and set the remaining solvents as not-predicted.

5.3.2 XGBoost

For XGBoost, we used the implementation from Chen and Guestrin [7]. We set the number of trees to 250 and the class weight to balanced to train more effectively on underrepresented classes. The remaining hyperparameters were left at their default configuration.

5.3.3 K-NN Clustering

To implement K-NN clustering, we used the scikit-learn implementation [15]. We set the number of neighbors to 5 and kept the remaining hyperparameters at their default configuration.

5.4 Calculation of success rate on ELN data

The success rate of the experimental campaign is twofold

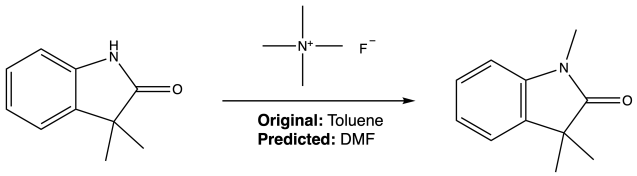
$$R_{experiments} = \frac{n_{success}}{n_{experiments}} \tag{3}$$

$$\frac{n_{correct\ pred\ RXN} + (n_{uncorrect\ pred\ RXN} * R_{experiments})}{n_{total\ RXN\ ELN}} = \frac{8 + (18 * \frac{8}{11}))}{26} = 81.11\% \quad (4)$$

5.5 Experimental Validation

Each reaction was run twice, once in the original solvent **Original** and once in the predicted solvent **Predicted**.

5.5.1 Reaction 1: N-Methylation of 3,3-dimethyl-1H-indol-2-one to 1,3,3-trimethylindolin-2-one



3,3-dimethyl-1H-indol-2-one (10mg, 0.062mmol, 1eq) and tetramethylammonium fluoride tetrahydrate (25.6 mg, 0.155mmol, 2.5 eq) added to a vial and dissolved with **Original** 1ml toluene or **Predicted** 1ml DMF. The mixture was stirred at 100C overnight. The formation of the product 1,3,3-trimethylindolin-2-one was determined in the HPLC/MS spectrum for **Original** and **Predicted**. MS (ESI): m/z 176.2 [M+H] calculated, found 176.1 m/z. Area peak: **Original**: 345'922'608 **Predicted** 121'640'946

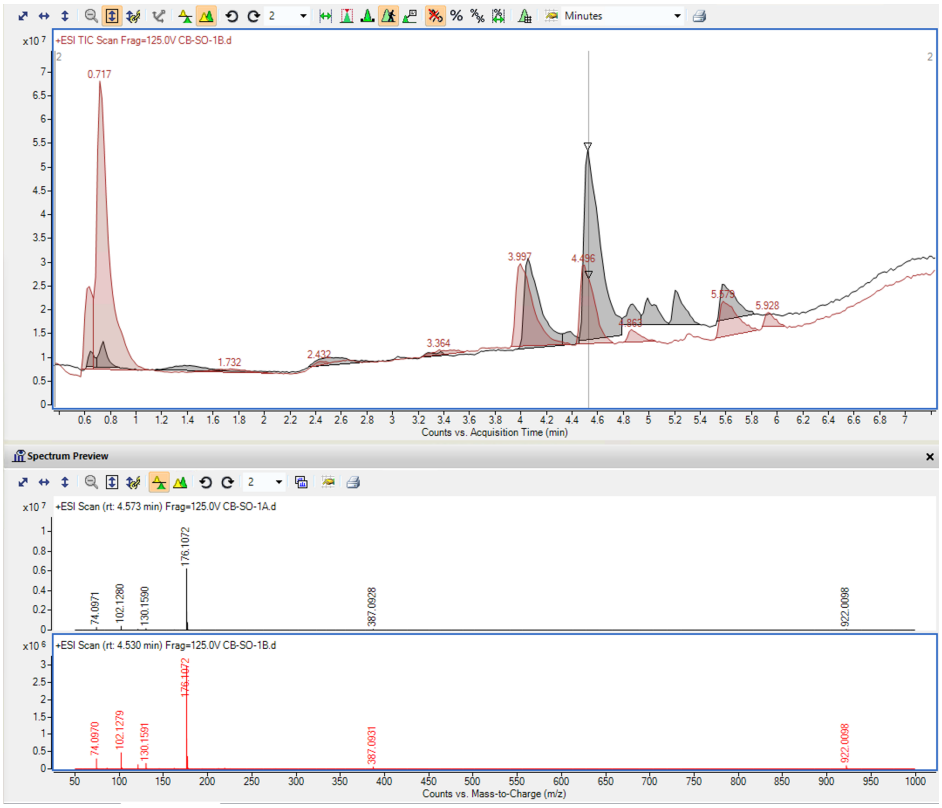
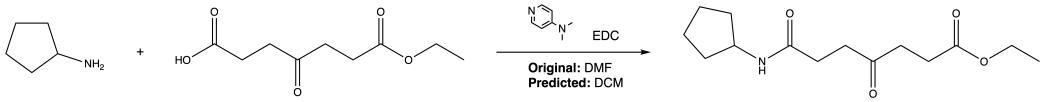


Figure 4: MS spectra of Reaction 1: **Original** in black and on top, **Predicted** in red and bottom

5.5.2 Reaction 2: Amide formation of ethyl 7-(cyclopentylamino)-4,7-dioxoheptanoate



7-ethoxy-4,7-dioxoheptanoic acid (15mg, 0.074mmol, 1eq) and cyclopentylamine (12.63mg, 0.184mmol, 2eq) were added to vial along EDC hydrochloride (17.07 mg, 0.089mmol, 1.2eq), N,N-dimethylpyridin-4-amine (0.9mg 0.007mmol, 0.1eq). A mixture was formed by adding **Original** 1ml of DMF respectively **Predicted** 1ml of DCM. The mixture was stirred at room temperature overnight to form ethyl 7-(cyclopentylamino)-4,7-dioxoheptanoate determined via HPLC/Ms spectrum for **Original** and **Predicted**. MS (ESI): m/z 270.12[M+H] calculated, found 270.17 m/z.

Area peak: **Original**: 3'649'001'684 **Predicted** 4'282'004'193

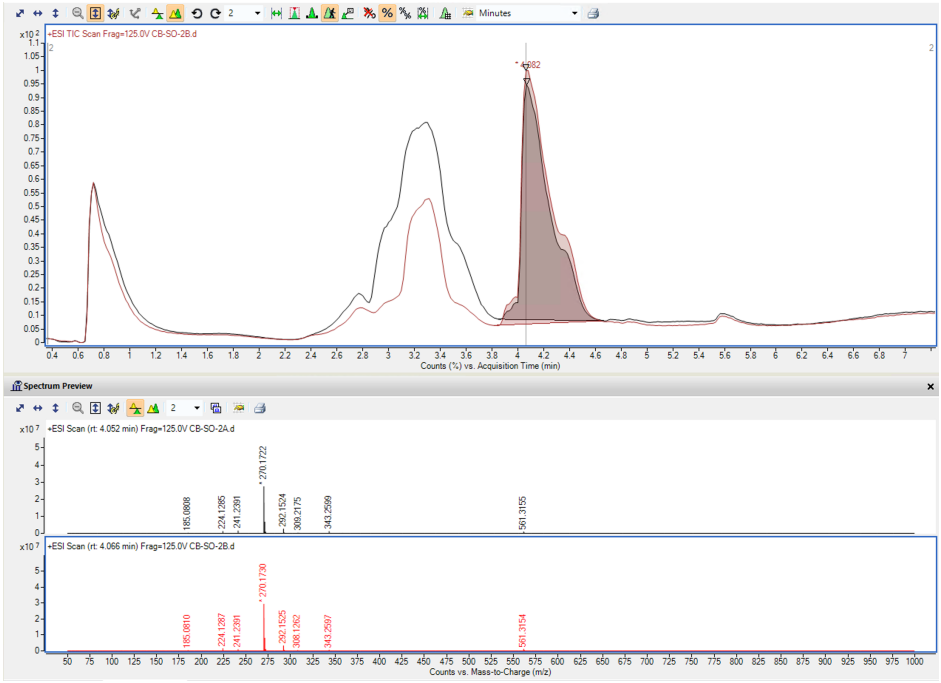
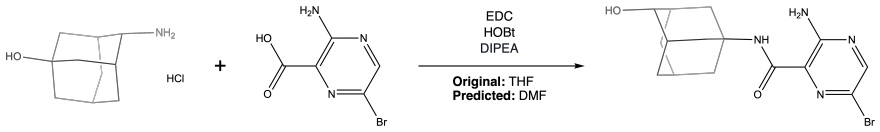


Figure 5: MS spectra of Reaction 2: **Original** in black and on top, **Predicted** in red and bottom

5.5.3 Reaction 3: Formation of 3-amino-6-bromo-N-(4-hydroxy-1-adamantyl)pyrazine-2-carboxamide



4-aminoadamantan-1-ol:hydrochloride (10mg, 0.046mmol, 1eq) and 3-amino-6-bromo-pyrazine-2-carboxylic acid (10.37mg, 0.046mmol, 1eq) were added alongside EDC HCl (10.73mg 0.059mmol, 1.2eq) and 1-hydroxybenzotriazole:hydrate (8.83mg 0.059mmol, 1.2eq) and Diisopropylethylamine (21.10mg, 0.0284ml, 1.642mmol, 3.5eq) and dissolved in **Original** 1ml THF or **Predicted** 1ml of DMF. The mixture was stirred over night at room temperature. The product 3-amino-6-bromo-N-(4-hydroxy-1-adamantyl)pyrazine-2-carboxamide was detected for **Original** and **Predicted** via HPLC/MS (ESI): m/z 367.07[M+H] calculated, found 367.07 m/z.

Area peak: **Original**: 526’414’862 **Predicted** 383’741’438

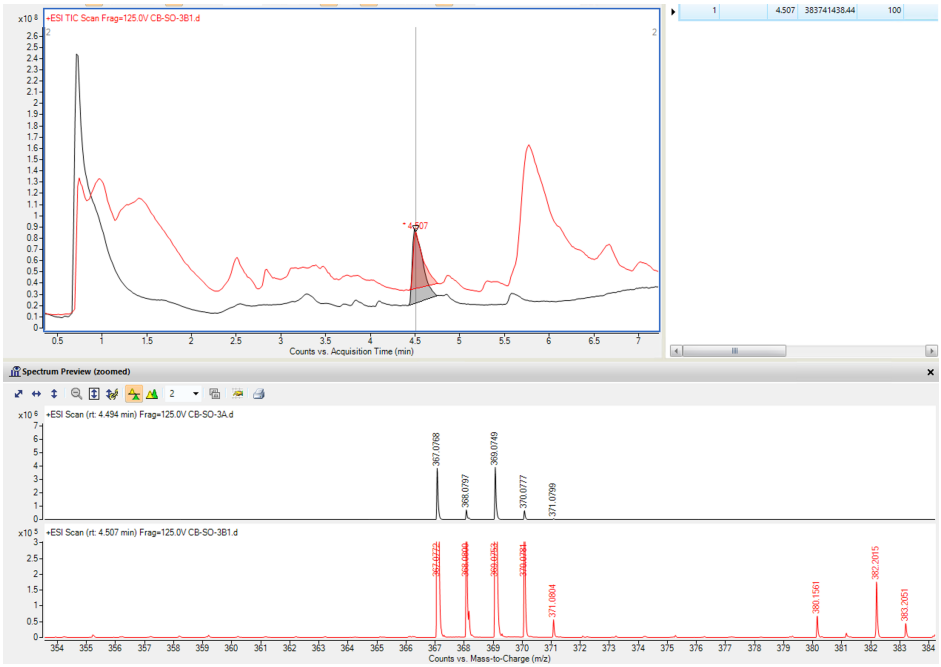
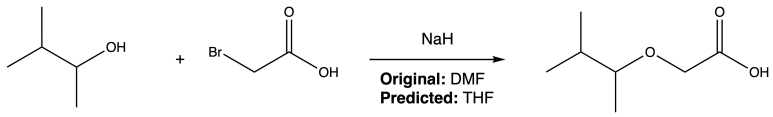


Figure 6: MS spectra of Reaction 3: **Original** in black and on top, **Predicted** in red and bottom

5.5.4 Reaction 4: Williamsom ether synthesis of 2-(3-methylbutan-2-yloxy)acetic acid



Bromoacetic Acid (15mg, 0.108mmol, 1eq) and cyclopentylamine (9.51mg, 0.0116ml, 0.108mmol, 1eq) were added to a vial. These starting materials were dissolved in **Original** 1ml DMF or **Predicted** 1ml THF. Sodium hydroyide (8.64mg, 0.216mmol, 2eq) was added and stirred overnight at room temperature. The product 2-(3-methylbutan-2-yloxy)acetic acid was only found in reaction **Original** via HPLC/MS (ESI): m/z 147.102 [M+H] calculated, found 147.113 m/z.

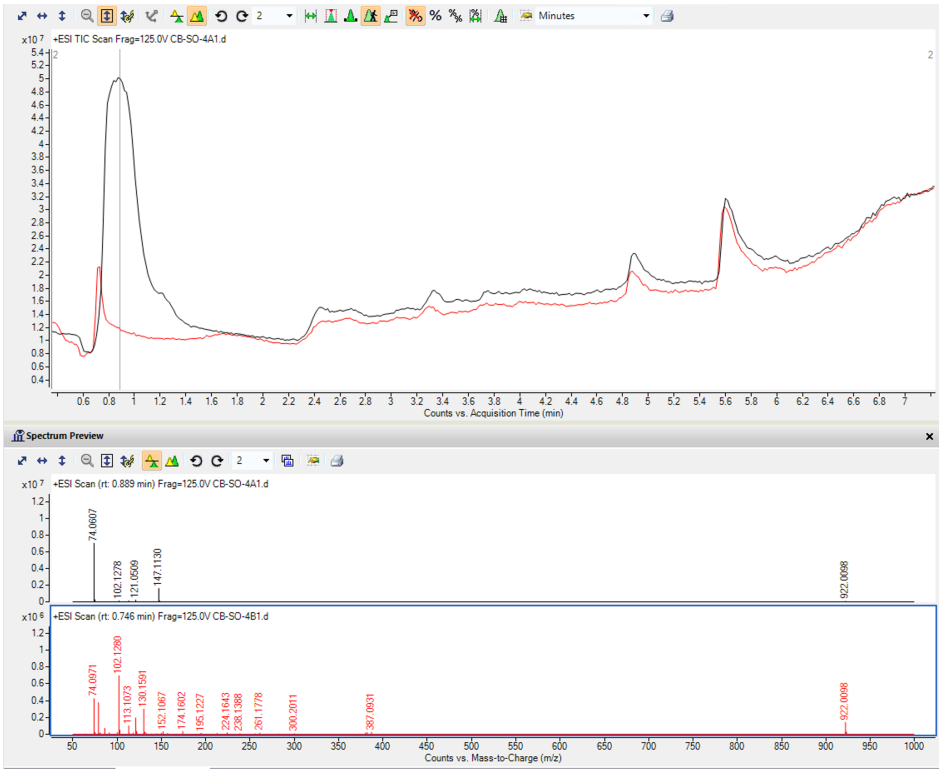


Figure 7: MS spectra of Reaction 4: **Original** in black and on top, **Predicted** in red and bottom

5.5.5 Reaction 5: Schiff base formation of (E)-N-(5-methylthiazol-2-yl)pentan-1-imine

5-methyl-2-thiazolamine (10mg, 0.0858mmol, 1eq) and valeraldehyde (7.6mg 0.0094ml, 0.0858mmol, 1eq) were added to a vial. The reagents were dissolved in **Original** 1ml methanol or **Predicted** 1ml toluene. The mixture was stirred at room temperature overnight. The product (E)-N-(5-methylthiazol-2-yl)pentan-1-imine was mainly found in solvent **Original** via HPLC/MS (ESI): m/z 183.08[M+H] calculated, found 183.10m/z. While in **Predicted** trace amount of the product was determined the reaction was classified as not successful.

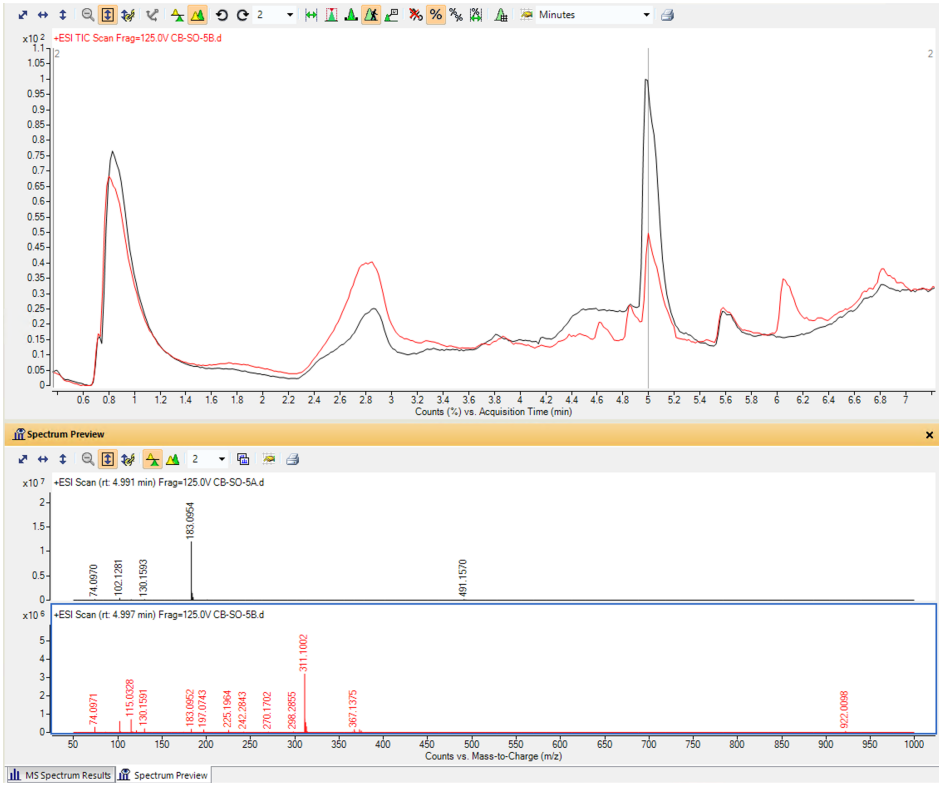
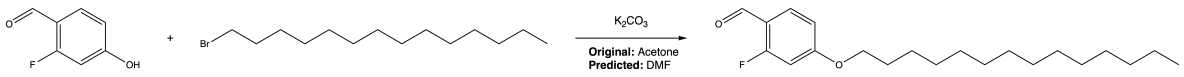


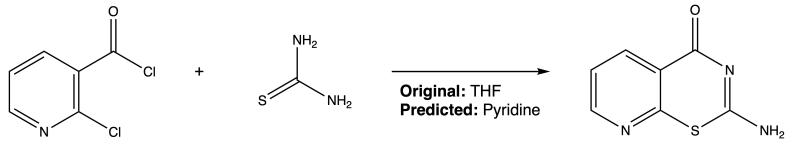
Figure 8: MS spectra of Reaction 5: **Original** in black and on top, **Predicted** in red and bottom

5.5.6 Reaction 6: Williamson ether formation to build 2-fluoro-4-(tetradecyloxy)benzaldehyde



2-fluoro-4-hydroxy-benzaldehyde (15mg, 0.107mmol, 1eq) and 1-bromotetradecane (44.53mg, 0.048ml, 0.161mmol, 1.5eq) were added to a vial alongside Potassium carbonate (25.59mg, 0.214mmol, 2eq). To the vial, **Original** 1ml of acetone or **Predicted** 1ml of DMF was added and the resulting mixture was stirred at 56 C overnight. No product 2-fluoro-4-(tetradecyloxy)benzaldehyde was detected via HPLC/MS(ESI) analysis.

5.5.7 Reaction 7: Formation of 2-aminopyrido[3,2-e][1,3]thiazin-4-one



2-chloropyridine-3-carbonyl chloride (15mg, 0.0124ml, 0.085mmol, 1eq) and isothiourea (6.5mg, 0.085mmol, 1eq) are added to a vial and were dissolved with **Original** 1ml THF or **Predicted** 1ml pyridine. The mixture was stirred overnight at room temperature. The product only form 2-aminopyrido[3,2-e][1,3]thiazin-4-one only found in solvent **Original** via HPLC/MS (ESI): m/z 180.19[M+H] calculated, found m/z.

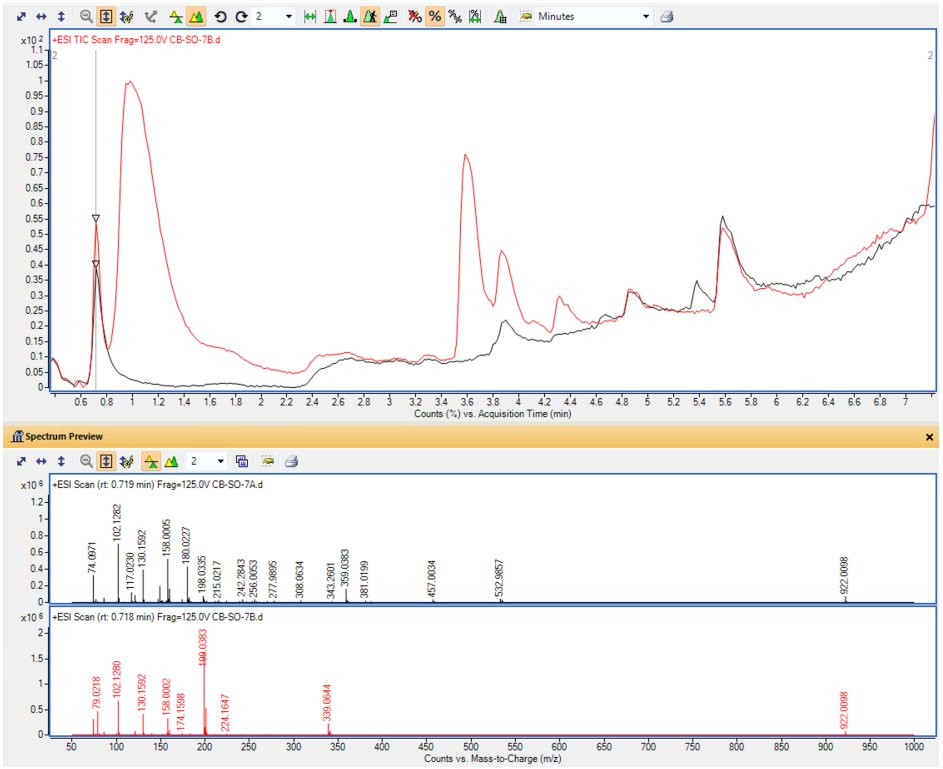
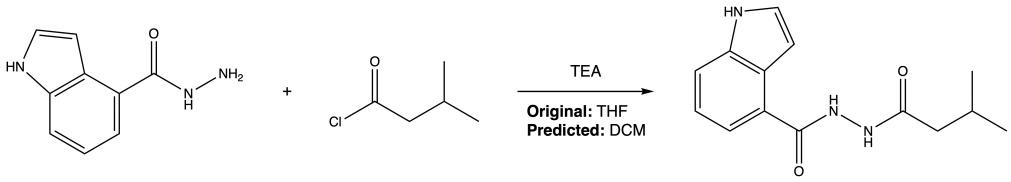


Figure 9: MS spectra of Reaction 7: **Original** in black and on top, **Predicted** in red and bottom

5.5.8 Reaction 8: Formation of N'-(3-methylbutanoyl)-1H-indole-4-carbohydrazide



1H-indole-4-carbohydrazide (15mg, 0.086mmol, 1eq) and 3-methylbutyryl chloride (12.3mg, 0.0125ml, 0.103mmol, 1.2eq) added to a vial alongside triethylamine (18.5mg, 0.025ml, 0.183mmol, 2.135eq) and dissolved in **Original** 1ml of THF or **Predicted** 1ml of DCM. The mixture was stirred at room temperature overnight. The product N'-(3-methylbutanoyl)-1H-indole-4-carbohydrazide was detected via HPLC/MS (ESI): m/z 260.13[M+H] calculated, found m/z.

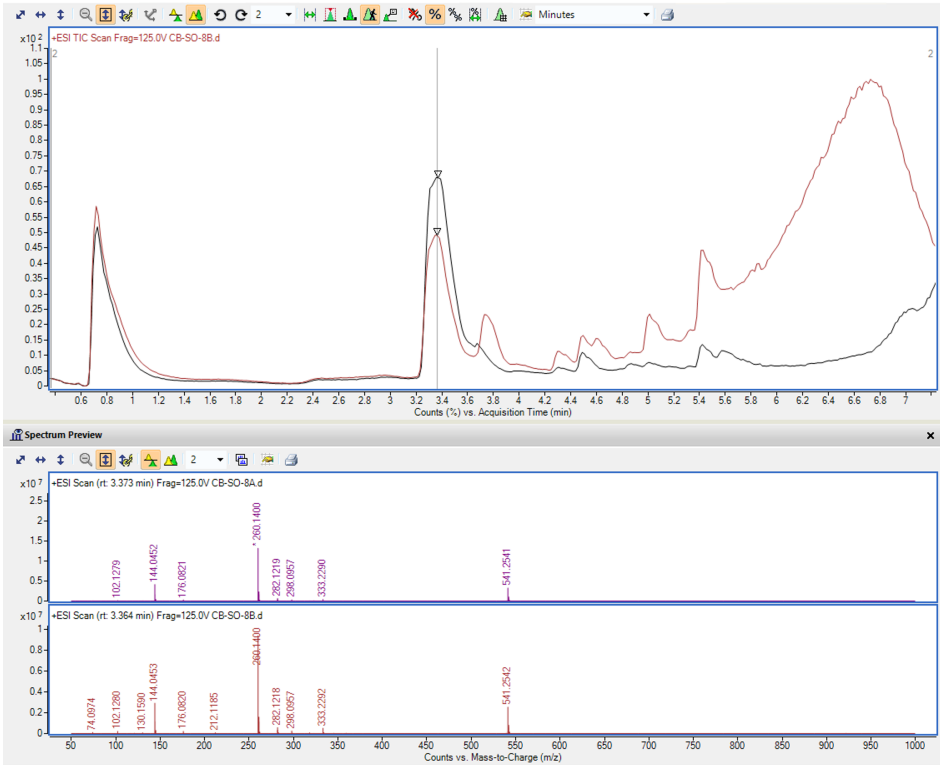
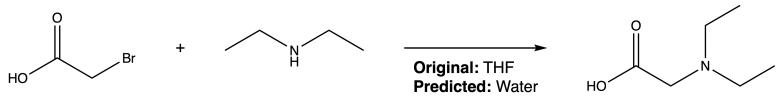


Figure 10: MS spectra of Reaction 8: **Original** in black and on top, **Predicted** in red and bottom

5.5.9 Reaction 9: Tertiary amine formation of 2-(diethylamino)acetic acid



2-bromoacetic acid (10mg, 0.072mmol, 1eq) and N-ethylethanamine (10.53mg, 0.015ml, 0.144mmol, 2eq) were added to a vial. A mixture was formed by adding **Original** 1ml of THF or **Predicted** 1ml of water. The reaction mixture was stirred overnight at room temperature. The product 2-(diethylamino)acetic acid was detected via HPLC/MS (ESI): m/z 132.09 [M+H] calculated, 132.10 found m/z.

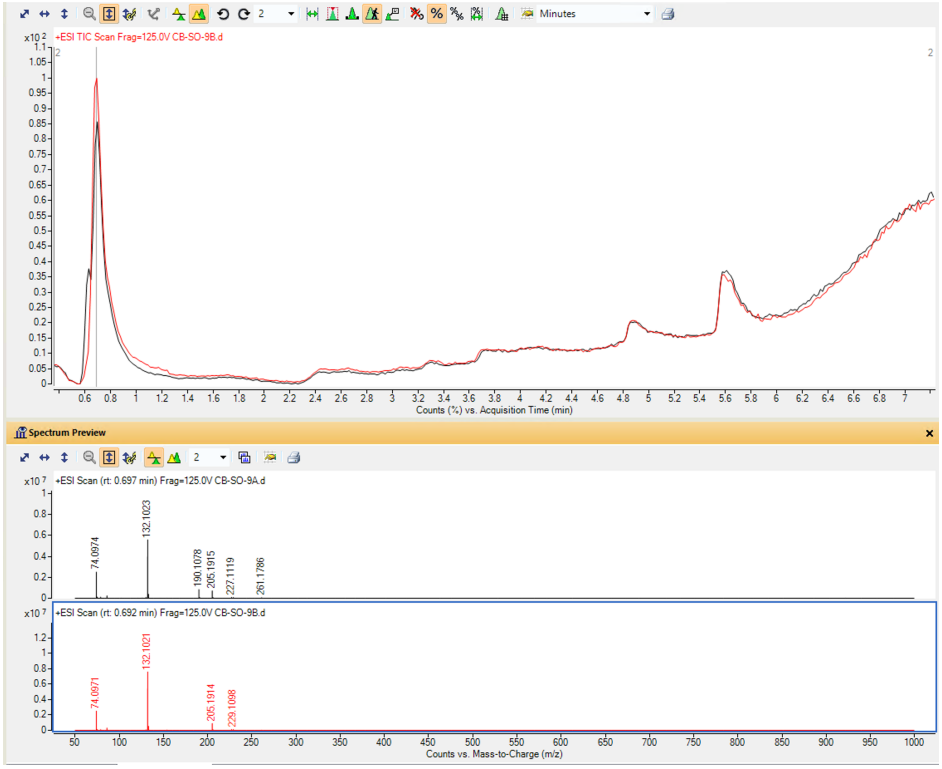
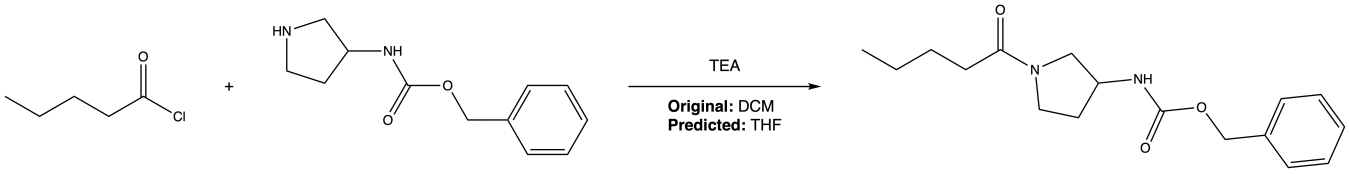


Figure 11: MS spectra of Reaction 9: **Original** in black and on top, **Predicted** in red and bottom

5.5.10 Reaction 10: Schotten-Baumann reaction to form benzyl (1-pentanoylpyrrolidin-3-yl)carbamate



Benzyl N-pyrrolidin-3-ylcarbamate (20mg, 0.091mmol, 1eq) and valeryl chloride (10.9mg 0.011ml 0.091mmol, 1eq) were added alongside triethylamine (11.6mg, 0.016ml) to a vial and dissolved by adding **Original** 1ml DCM or **Predicted** 1ml THF. The mixture was stirred at room temperature overnight. The formed benzyl (1-pentanoylpyrrolidin-3-yl)carbamate was detected in **Original** and **Predicted** via HPLC/MS (ESI): m/z 305.18 [M+H] calculated, 305.18 found m/z.

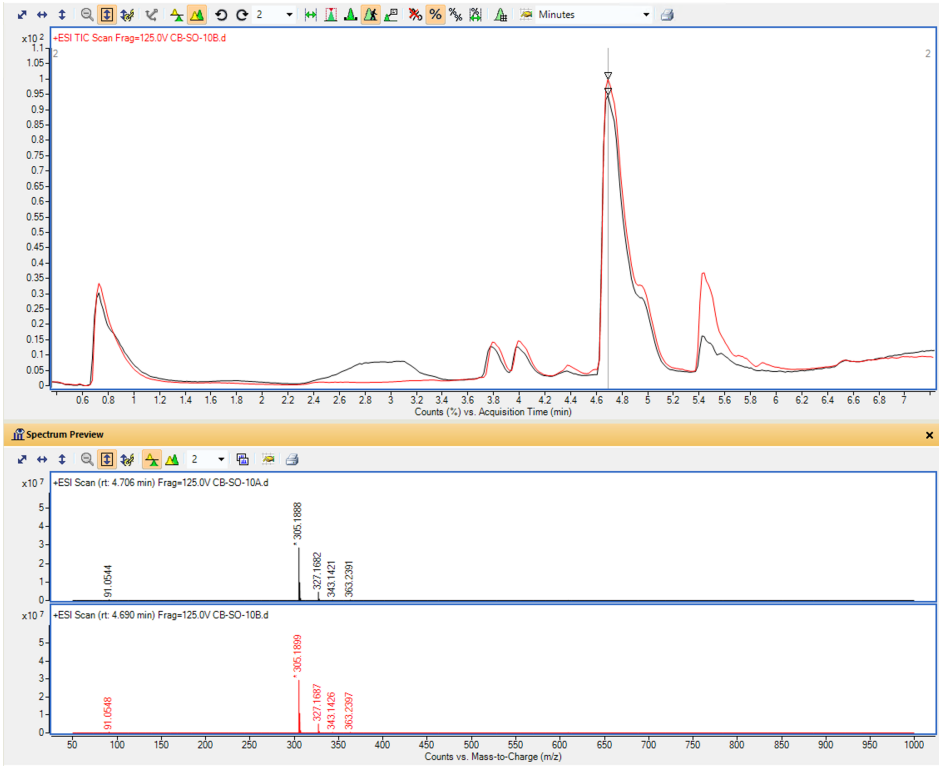
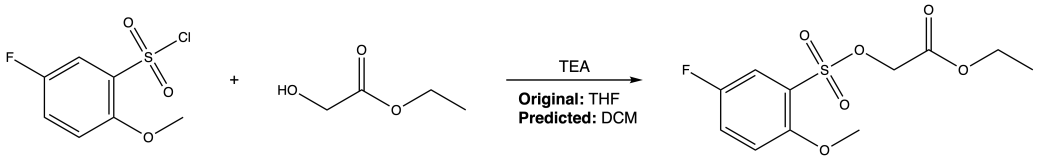


Figure 12: MS spectra of Reaction 10: **Original** in black and on top, **Predicted** in red and bottom

5.5.11 Reaction 11: Sulfonic ester formation of ethyl 2-(((5-fluoro-2-methoxyphenyl)sulfonyl)oxy)acetate



5-fluoro-2-methoxybenzenesulfonyl chloride (15mg, 0.066mmol, 1eq) and ethyl 2-hydroxyacetate (9.1mg, 0.086mmol, 1.2eq) were added to a vial alongside triethylamine (10.1mg, 0.014ml, 0.1mmol, 1.5eq) and dissolved in **Original** 1ml of THF or **Predicted** 1ml of DCM. The mixture was stirred overnight at room temperature. The product ethyl 2-(((5-fluoro-2-methoxyphenyl)sulfonyl)oxy)acetate formed and was detected in **Original** and **Predicted** via HPLC/MS (ESI): m/z 310.07 [M+NH4] calculated, 310.18 found m/z.

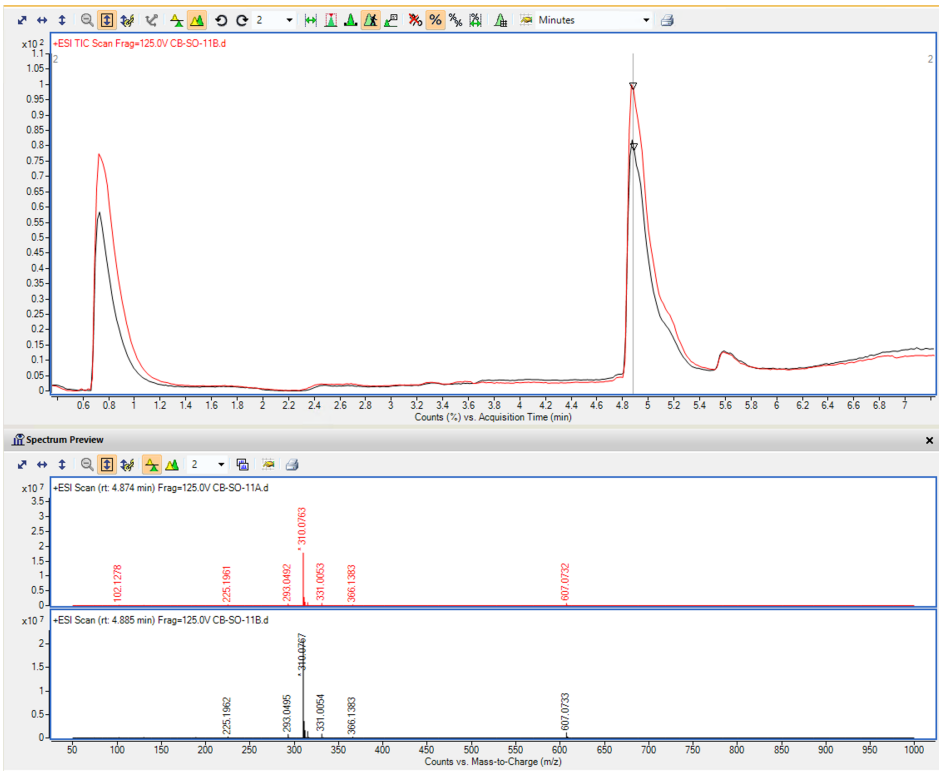
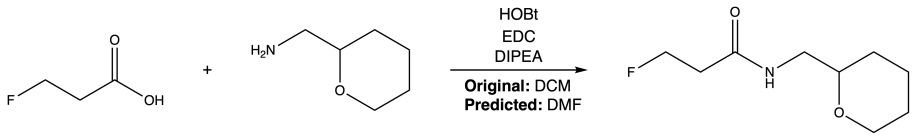


Figure 13: MS spectra of Reaction 10: **Original** in black and on top, **Predicted** in red and bottom

5.5.12 Reaction 12: Steglich reaction to form 3-fluoro-N-((tetrahydro-2H-pyran-2-yl)methyl)propanamide



3-fluoropropanoic acid (10mg, 0.104mmol, 1eq) and tetrahydropyran-2-ylmethylamine (12.0mg, 0.104mmol, 1eq) were added to a vial alongside EDC (23.9mg, 0.121mmol, 1.2eq) DIPEA(33.7mg, 0.045ml, 0.261mmol, 2.5eq) and HOBt (19.8mg, 0.121mmol, 1.2eq). To the vial, either **Original** 1ml of DCM or **Predicted** 1ml of DMF were added. The mixture was stirred at room temperature overnight. 3-fluoro-N-((tetrahydro-2H-pyran-2-yl)methyl)propanamide was detected in **Original** and **Predicted** via HPLC/MS (ESI): m/z 192.12 [M+H] calculated, found m/z.

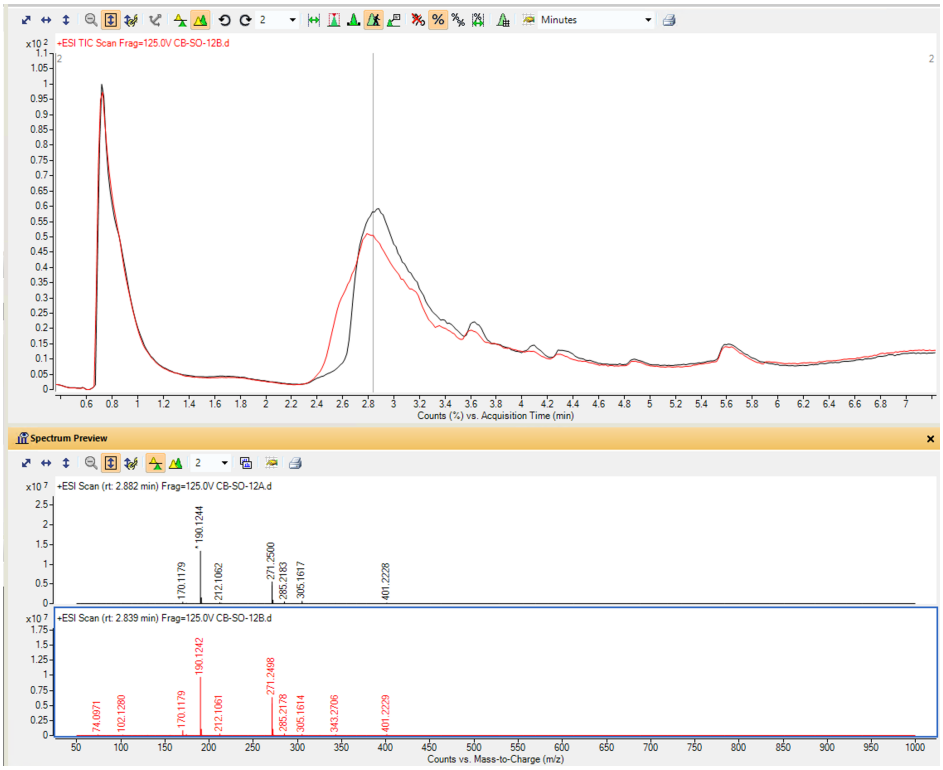


Figure 14: MS spectra of Reaction 10: **Original** in black and on top, **Predicted** in red and bottom

References

[1] Reaxys. https://www.reaxys.com, Accessed on: 20 March 2023.

[2] K. Alfonsi, J. Colberg, P. J. Dunn, T. Fevig, S. Jennings, T. A. Johnson, H. P. Kleine, C. Knight, M. A. Nagy, D. A. Perry, et al. Green chemistry tools to influence a medicinal chemistry and research chemistry based organisation. *Green Chemistry*, 10(1):31–36, 2008.

[3] W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, and B. A. Grzybowski. Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic suzuki–miyaura coupling. *Journal of the American Chemical Society*, 144(11):4819–4827, 2022.

[4] S. Boobier, D. R. Hose, A. J. Blacker, and B. N. Nguyen. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature communications*, 11(1):5753, 2020.

[5] F. P. Byrne, S. Jin, G. Paggiola, T. H. Petchey, J. H. Clark, T. J. Farmer, A. J. Hunt, C. Robert McElroy, and J. Sherwood. Tools and techniques for solvent selection: green solvent selection guides. *Sustainable Chemical Processes*, 4:1–24, 2016.

[6] C. Capello, U. Fischer, and K. Hungerbühler. What is a green solvent? a comprehensive framework for the environmental assessment of solvents. *Green Chemistry*, 9(9):927–934, 2007.

[7] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476, 2018.

[10] R. K. Henderson, C. Jiménez-González, D. J. Constable, S. R. Alston, G. G. Inglis, G. Fisher, J. Sherwood, S. P. Binks, and A. D. Curzons. Expanding gsk's solvent selection guide–embedding sustainability into solvent selection starting at medicinal chemistry. *Green Chemistry*, 13(4):854–862, 2011.

[11] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[12] D. M. Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[13] J. Mayfield, D. Lowe, and R. Sayle. Pistachio, 2018.

[14] D. L. Mobley and J. P. Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] D. Prat, O. Pardigon, H.-W. Flemming, S. Letestu, V. Ducandas, P. Isnard, E. Guntrum, T. Senac, S. Ruisseau, P. Cruciani, et al. Sanofi's solvent selection guide: A step toward more sustainable processes. *Organic Process Research & Development*, 17(12):1517–1525, 2013.

[17] D. Probst, P. Schwaller, and J.-L. Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discovery*, 2022.

[18] B. Sanchez-Lengeling, L. M. Roch, J. D. Perea, S. Langner, C. J. Brabec, and A. Aspuru-Guzik. A bayesian approach to predict solubility parameters. *Advanced Theory and Simulations*, 2(1):1800069, 2019.

[19] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, and J.-L. Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, 2021.

[20] P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1):015016, 2021.

[21] E. Shim, J. A. Kammeraad, Z. Xu, A. Tewari, T. Cernak, and P. M. Zimmerman. Predicting reaction conditions from limited data through active transfer learning. *Chemical Science*, 13(22):6655–6668, 2022.

[22] C. S. Slater and M. Savelski. A method to characterize the greenness of solvents used in pharmaceutical manufacture. *Journal of Environmental Science and Health, Part A*, 42(11):1595–1605, 2007.

[23] F. H. Vermeire and W. H. Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021.

[24] E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari, and P. M. Zimmerman. Learning to predict reaction conditions: relationships between solvent, molecular structure, and catalyst. *Journal of chemical information and modeling*, 59(9):3645–3654, 2019.

[25] J. Yu, C. Zhang, Y. Cheng, Y.-F. Yang, Y.-B. She, F. Liu, W. Su, and A. Su. Solvbert for solvation free energy and solubility prediction: A demonstration of an nlp model for predicting the properties of molecular complexes. *Digital Discovery*, 2023.

[26] H. Zhang, S. Si, and C.-J. Hsieh. Gpu-acceleration for large-scale tree boosting. *arXiv preprint arXiv:1706.08359*, 2017.