Are Large Language Models the Best Estimators of Chemical Reaction Yields?

Anastasia Orlova[1], **Andrei Dmitrenko**[1,2], **Aleksei Dmitrenko**[1], **Vladimir Vinogradov**[1]

[1]Center for AI in Chemistry, SCAMT institute, ITMO University, Saint Petersburg 191002, Russian Federation
[2]D ONE AG, Zurich, Switzerland
dmitrenko@scamt-itmo.com

Abstract

Chemical reaction yield, defined as the percentage of reactants turned into products, is the main criterion for selecting reaction conditions and evaluating success of a synthesis. Various machine learning (ML) models have been reported to predict reaction yields based on high-throughput experiment datasets. However, in the face of sparse and insufficient data typical for regular laboratory experiments, the performance and applicability of such models remain limited. More recently, the capabilities of large language models (LLMs) have been explored in predictive chemistry. Following up on this work, we investigate how LLMs perform in the generalized yield prediction task treated as a binary classification problem. In this regard, we engineer four different chemical reaction datasets to systematically evaluate performance of the top rated LLMs. We demonstrate that in the few-shot classification task LLMs outperform baseline approaches in F1-score up to 9% and show competitive performance in terms of accuracy. Moreover, we observe superiority of ML models trained on LLM embeddings with the best average accuracy of 0.70 versus 0.67 achieved with current state-of-the-art approaches on the USPTO data. In this context, we discuss the potential of LLM embeddings to become the new state-of-the-art chemical reaction representations. Additionally, we share our empirical results on practical aspects of the few-shot LLM classifiers, such as the optimal size of the training set, and discuss peculiarities and prospects of the proposed methods.

Introduction

Chemical reaction yield estimation plays a crucial role in chemical synthesis planning (Zhang et al. 2023; Noyori 2009). It allows assessing the amount of product that can be obtained with a particular chemical reaction, which is essential for optimizing reaction conditions and designing efficient synthesis routes (Schwaller et al. 2021; Meuwly 2021).

In the last decade, machine learning has made significant strides in prediction of reaction yields. The main idea behind the existing works is to formulate and solve a regression problem of predicting the yield values based on reaction representations and/or reaction parameters, such as temperature, solvent, reagents and others (Haywood et al. 2021; Sato, Miyao, and Funatsu 2022). Nevertheless, estimating reaction yields remains a rather challenging task (Voinarovska et al. 2023). The complexity of chemical experiments and the extensive number of factors involved both contribute to the difficulty of the regression problem (Chen et al. 2023). However, the primary reason why yield prediction still remains a challenge is the lack of reliable data available for training the machine learning models (Bustillo and Rodrigues 2023).

High-throughput experimentation (HTE) is often used to generate training data and power deep learning applications (Callaghan 2021; Eyke, Koscher, and Jensen 2021). Allowing large numbers of experiments to be run simultaneously, HTE provides high-quality reaction datasets containing detailed and organized information on the reaction conditions and the corresponding product yields (Biyani, Moriuchi, and Thompson 2021). However, HTE datasets have one significant drawback: they usually focus on a particular reaction type and offer information for only a small number of reaction templates applied to a limited set of reactant combinations (Schwaller et al. 2021). For this reason, models trained on such datasets usually perform very well but cannot generalize to other datasets (Schwaller et al. 2021; Fitzner et al. 2023).

In contrast, there are datasets of reactions covering a wide chemical space. Public ones include ORD (Kearnes et al. 2021) and ChEMU (Nguyen et al. 2020). Reaxys (Goodman 2009) and CAS (CAS 2008) belong to commercial databases. However, the most commonly used data is the US Patent and Trademark Office (USPTO) dataset (Lowe 2012). USPTO found many applications in the context of retrosynthesis (Lin et al. 2020; Karpov, Godin, and Tetko 2019), and since the topics of synthesis planning and prediction of reaction yields largely overlap, the same dataset was used for yield prediction as well (Schwaller et al. 2021; Probst, Schwaller, and Reymond 2022). Unlike many of the aforementioned datasets, USPTO is public, well-structured and contains a large number of organic reactions of various types. However, since the data originates from different studies, it also suffers from sparse and noisy chemical reaction information as well as biases in the reported yield values. Consequently, although models trained on such data are more universal and widely applicable, they achieve rather modest results and cannot be put into practice. This fact is confirmed by several studies where no satisfactory results were obtained regardless of the complexity of the machine learning techniques applied (Schwaller et al. 2021; Probst, Schwaller, and Reymond 2022; Yarish et al. 2023; Jiang et al. 2021).

Large language models (LLMs) have made a pivotal breakthrough in the development of AI in recent years. Naturally, this technology has started to extend its reach into various non-linguistic highly-specialized domains such as physics, biology or material sciences (Latif, Parasuraman, and Zhai 2024; Li et al. 2024; Jablonka et al. 2023). Recent studies reveal the potential for LLM-based approaches in predictive chemistry as well (Zhao et al. 2024; M. Bran et al. 2024; McNaughton et al. 2024). Therefore, for the first time in the field, we investigated how large language models perform in the generalized yield prediction task formulated as a binary classification problem. In certain scenarios, knowing whether a reaction is high-yielding or not without requiring precise yield values can be sufficient. In HTE, for instance, assessing whether a reaction is high-yielding or not provides quick feedback and allows to prioritize reactions for further optimization (Collins and Glorius 2015). In addition, in routine synthetic operations where the main objective is to produce sufficient quantities of a compound for further applications, knowing whether a reaction consistently yields high amounts of product is often more critical (Schwaller et al. 2021).

The main contributions of this paper can be summarized as follows:

1. We explore the predictive capabilities of six state-of-the-art market-leader LLMs in reaction yield prediction using few-shot approaches. For this, we design four reactions datasets and conduct a comparative analysis of LLM-based approaches against several baseline models. We demonstrate the competitive performance of LLMs with an average accuracy value of 0.61. Moreover, we show that few-shot classification surpasses baselines in F1-scores by a maximum of 9%.
2. We are the first to predict reaction yields based on two state-of-the-art LLM embeddings. We showcase an increase in accuracies and F1-scores up to 10% obtained with the models trained on LLM embeddings compared to the baseline approaches. Based on our empirical results, we hypothesize that LLM embeddings could become novel state-of-the-art representations of chemical entities.
3. In addition, we provide practical considerations on using LLMs and identify the optimal size of the training set required to achieve competitive performance in the few-shot setting. These findings can be especially helpful for chemists seeking to employ LLM-based reaction yield prediction for their own experimental data. The code and data used in this study are available at: https://github.com/ai-chem/LLMYieldPred.

Related Work

Yield Prediction for Specific Reaction Types

Cross-coupling reactions are the fundamental part of the pharmaceutical synthesis (Ruiz-Castillo and Buchwald 2016). Modern pharmaceutical laboratories usually screen cross-coupling reactions with the help of HTE setups, making this reaction class a popular basis for yield prediction tasks. Ahneman et al. (2018) were among the first to predict yields based on HTE cross-coupling reactions data. Having collected a dataset of 4.6k reactions, they trained a Random Forest model achieving a coefficient of determination (R^2) of 0.92. In the study of Fu et al. (2020) a feed-forward neural network showed impressive results in the prediction of Suzuki-Muyaura cross-coupling reactions' yields and scored with R^2=0.95 on the test set. Apart from classical machine learning approaches, complex deep learning architectures such as transformers (Schwaller et al. 2020, 2021) and graph neural networks (Kwon et al. 2022; Sato, Miyao, and Funatsu 2022; Zhao et al. 2021; Saebi et al. 2021) were proposed. Despite the fact that all the aforementioned models cope with yield prediction within a specific reaction type, they cannot generalize to other reaction classes. This limitation makes them hardly applicable in general practice of experimental laboratories. For this reason, we opt for using the USPTO and ORD datasets with many reaction types in our study.

Generalized Yield Prediction

Some considerable work has been done in the field of general-purpose yield prediction without binding to a specific reaction class. An early attempt to predict reaction outcomes for a multi-type reaction dataset was made by Skoraczyński et al. (2017). In this study ML models were trained to solve a simpler binary classification task. Still, their work reached a conclusion that existing ML-methods do not cope well with the task due to insufficient number of descriptors and the unsystematic way of reporting reactions in organic chemistry literature. Schwaller et al. (2021) were the first to predict reaction yields given text-based representations of reactions. Although the proposed Yield-BERT transformer attained outstanding results for HTE cross-coupling reactions, the model failed on the USPTO data with R^2 value of 0.195 due to the strong difference of reaction yields for close reactions. Lu and Zhang (2022) followed the idea of Schwaller et al. and introduced T5Chem model based on Text-to-Text Transfer Transformer borrowed from NLP (Raffel et al. 2020). The authors created a special multi-task dataset based on the USPTO data. Even though the number of reaction classes in this dataset was reduced, the suggested model only managed to achieve R^2 of 0.47 in yield prediction. Although this metric is higher than those of other authors, they could not be fairly compared as the problem was significantly narrowed. Another BERT-based model called Egret was suggested by Yin et al. (2024). It was reported to achieve R^2 of 0.128 on the USPTO dataset. At the same time, the study of Probst, Schwaller, and Reymond (2022) proved that classical ML models, such as XGBoost, outperform transformer-based architectures when trained on the novel differential reaction fingerprints (DRFP) as opposed to SMILES strings. However, the overall predictive performance of such method remains rather low with R^2 of 0.197. It can be observed that regression problem continues to be a difficult undertaking in yield prediction (also confirmed by our own regression experiments described in Appendix A.1), which motivated us to opt for classification task in our study.

LLMs in Reaction Yield Prediction

Some recent studies have investigated the potential of LLMs to predict chemical reaction yields based on SMILES representations of reactions. In the research of Guo et al. few-shot classification was carried out to evaluate several LLMs on two benchmarking HTE datasets. GPT-4 was reported to show the best results among other LLMs and achieved competitive performance to the baseline graph neural network. Additionally, authors raised the need for future research and improvement in the performance of LLMs on challenging chemistry datasets, which motivated us to consider the task of generalized yield prediction. Similar to Guo et al., we formulate this task as a classification problem and apply a few-shot approach. However, since the problem we are solving is more universal and practically oriented, we try alternative sampling techniques, experiment with data formats and complement our research by investigating the optimal size of the training subset.

The alternative approach for evaluating LLMs performance in predictive tasks is to extract embeddings from LLMs and train ML models on them. The success of this method has already been proved in some text classification and clustering tasks (Petukhova, Matos-Carvalho, and Fachada 2024; Keraghel, Morbieu, and Nadif 2024). To the best of our knowledge, however, we are the first to use LLMs embeddings for reaction yield prediction.

It is important to note we intentionally opted for general-purpose LLMs in this work. Beside their strong performance demonstrated across various chemical tasks (Dubrovsky et al. 2024; Liu et al. 2024a; Jablonka et al. 2023), such LLMs possess major advantages such as broad accessibility and ease of use, regular updates improving their performance over time, as well as state-of-the-art generalization capabilities. Models pretrained or fine-tuned on chemical data lack those benefits and do not always provide top results (Sadeghi et al. 2024; Liu et al. 2024b; Schwaller et al. 2021; Yu et al. 2024).

Preparation of Datasets

Public datasets containing a wide range of reaction classes are usually quite large and may contain several thousands of reactions in a test set. In this study, however, we designed four smaller datasets based on two public reactions databases. This choice is particularly suitable for few-shot classification tasks, where the goal is to assess the model's ability to generalize from a limited number of examples it is presented with. Moreover, this setup provides a clearer understanding of how well LLMs can perform in real-world scenarios where labeled data is often scarce.

Two datasets were obtained using the USPTO dataset (Schwaller et al. 2018) containing over 1M organic reactions in the format of canonicalized Simplified Molecular-Input Line-Entry System (SMILES). The dataset was additionally processed resulting in more than 526k entries. Another pair of datasets were derived from the ORD database (Kearnes et al. 2021) including only reactions that do not overlap with USPTO dataset and include reported yields as well as two reactants maximum.

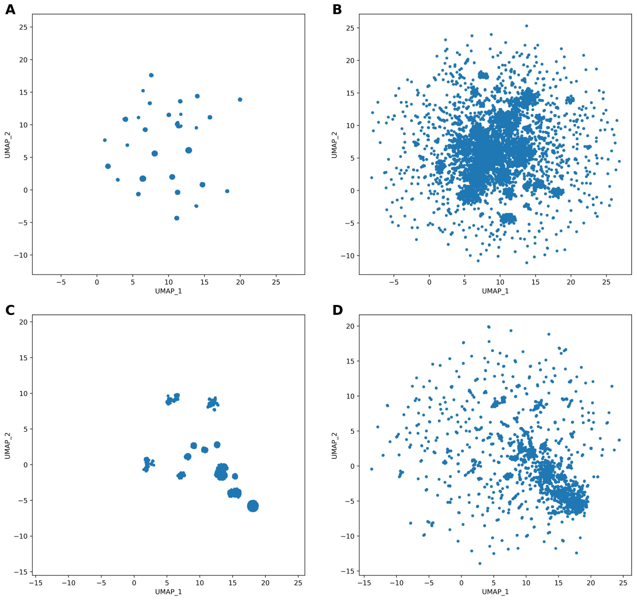


Figure 1: Visualization of the datasets: A) USPTO-C; B) USPTO-R; C) ORD-C; D) ORD-R

We used two techniques to sample reactions from the USPTO and ORD datasets. The first technique involved random sampling. We randomly selected 5300 reactions from ORD (ORD-R dataset) and 11300 reactions from USPTO (USPTO-R dataset). The second sampling technique was based on clustering. We employed differential reaction fingerprints (DRFP) with radius of 2 and length of 2048 bits as reaction representations using the `drfp` Python framework. DRFP is an NLP-inspired representation that has previously been shown well suitable for reaction clustering tasks (Probst, Schwaller, and Reymond 2022). We reduced fingerprints dimensionality to 2D space with UMAP algorithm (McInnes, Healy, and Melville 2018) and clustered the embeddings with HDBSCAN (McInnes et al. 2017). Reactions for the final datasets were sampled proportionally from each cluster, resulting in the ORD-C dataset containing 7313 reactions and USPTO-C dataset containing 9329 reactions. From each dataset a validation (300 reactions) and a test (100 reactions) subsets were sampled either randomly (for USPTO-R and ORD-R) or based on the proportion of the obtained clusters (for USPTO-C and ORD-C). Visualizations of the obtained datasets are represented in Figure 1. A thorough description of the datasets' preparation process is provided in Appendix A.2.

Similar to Guo et al., we categorize reaction yields into two distinct classes: ”Not high-yielding” for reactions yielding below 70%, and ”High-yielding” for reactions yielding 70% and above. This division not only minimizes the imbalance of classes, but also corresponds to the commonly accepted practices among chemists: reactions with yields below 70% are typically called “fair” or “poor”, while reactions with yields above 70% are considered “good”, “very good” and “excellent” (Vogel 1974).

Experiments

Few-shot Classification

Experimental Settings Few-shot classification is a task where the model is prompted to classify objects (or data points) into predefined categories, but with the constraint of having access to only a limited number of training samples (shots) (Mar and Liu 2022).

In our experiments, we selected three market leader providers of LLMs, namely, OpenAI, Anthropic and Mistral AI. We employed two models from each provider: a top performant one and another smaller one that is more affordable. More specifically, we conducted experiments with GPT-3.5 Turbo, GPT-4, Claude 3 Haiku, Claude 3 Opus, Mistral Small and Mistral Large.

The models were prompted with a few example reactions from the training set and asked to predict yield categories for reactions in the test set. We used two types of reaction representations: SMILES strings (as in the original dataset) and text descriptions. The latter were obtained by converting SMILES strings of molecules into their chemical names using the PubChem API. A few examples of such conversions are given in Appendix A.3.

For each type of reaction representation, we adapted prompts previously proposed by Guo et al. (2023). The prompt for the textual data format was as follows:

You are an expert chemist. Based on text descriptions of organic reactions, you predict their yields using your experienced reaction yield prediction knowledge. You can only predict whether the reaction is 'High-yielding' or 'Not high-yielding'. 'High-yielding' reaction means the yield rate of the reaction is above 70%. 'Not high-yielding' means the yield rate of the reaction is below 70%. You will be provided with several examples of reactions and the corresponding yield rates. Please answer with only 'High-yielding' or 'Not high-yielding', no other information can be provided.

For SMILES representations, some information about basic SMILES rules was included in the prompt (see Appendix A.4).

We experimented with the number of shots $k = \{2, 4, 6, 8, 10\}$ and two sampling strategies. The Tanimoto sampling strategy involves calculating Tanimoto similarity (Bajusz, Rácz, and Héberger 2015) between DRFPs and random selection of k reactions with Tanimoto similarity not less than 0.8. Such sampling technique selects examples from the chemical point of view. In the Random sampling strategy, we randomly selected at least one reaction from each yield category, which provides examples based on the general classification idea with no regard to the nature of reactions. For the SMILES data format, both Tanimoto and Random sampling strategies were applied, while for the textual data format we used only the Random sampling strategy.

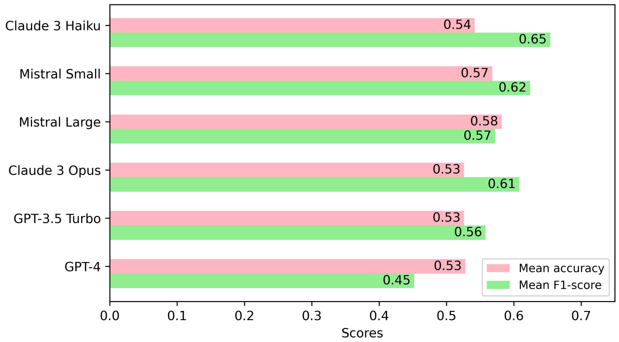


Figure 2: Comparison of different LLMs with the following prompting strategy: textual data format, Random sampler

USPTO-R dataset was used to select the best prompting strategy as it is the largest and the noisiest dataset among the others. We evaluated performance using standard classification metrics: accuracy and F1-score. Each experiment was conducted 5 times with different random states to calculate mean and standard deviation of classification metrics. The same algorithm was performed for all the further experiments, which are represented in the format of mean ± std in this paper.

Results The comparison of prompting strategies revealed that the textual data format with Random sampling technique outperformed other approaches by up to 12% in accuracy (Table 1). Although SMILES data format along with Random sampler demonstrated the highest F1-scores, the combination of the textual data format with Random sampler was selected as the optimal strategy due to the smaller gap between the two classification metrics. To our surprise, however, two smaller LLMs (i.e., Claude 3 Haiku and Mis-

Prompting strategy	SMILES Tanimoto sampler		SMILES Random sampler		Text descriptions Random sampler	
Metrics	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Claude 3 Haiku	0.55 ± 0.04	0.63 ± 0.06	**0.55 ± 0.00**	**0.69 ± 0.00**	0.55 ± 0.02	0.67 ± 0.02
Claude 3 Opus	**0.53 ± 0.02**	0.63 ± 0.11	0.51 ± 0.00	**0.68 ± 0.00**	0.53 ± 0.03	0.64 ± 0.07
Mistral Small	0.49 ± 0.03	0.64 ± 0.03	0.50 ± 0.00	**0.66 ± 0.00**	**0.61 ± 0.02**	0.63 ± 0.04
Mistral Large	0.53 ± 0.03	0.62 ± 0.04	0.54 ± 0.02	0.65 ± 0.02	**0.59 ± 0.02**	**0.66 ± 0.03**
GPT-3.5 Turbo	0.51 ± 0.01	0.52 ± 0.20	**0.51 ± 0.00**	**0.68 ± 0.00**	**0.51 ± 0.00**	**0.68 ± 0.00**
GPT-4	0.50 ± 0.02	0.58 ± 0.16	**0.51 ± 0.00**	**0.68 ± 0.00**	0.51 ± 0.01	0.67 ± 0.00

Table 1: Comparison of prompting strategies on the USPTO-R dataset. For each LLM, only results with k corresponding to the highest sum between accuracy and F1-score are included. Results for all k values are presented separately in Tables 4, 5, 6 of the Appendix A.5. For each LLM, the best accuracy and F1-score are highlighted in bold.

LLM	Data format	USPTO-C		USPTO-R		ORD-C		ORD-R	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Mistral 7B	SMILES	0.56 ± 0.01	0.50 ± 0.02	0.54 ± 0.04	0.59 ± 0.04	0.66 ± 0.04	0.71 ± 0.04	0.59 ± 0.02	0.52 ± 0.04
	Text	0.65 ± 0.01	0.61 ± 0.02	**0.62 ± 0.04**	**0.64 ± 0.05**	**0.66 ± 0.02**	**0.71 ± 0.02**	0.60 ± 0.01	0.55 ± 0.01
`text-embedding-3-large`	SMILES	0.66 ± 0.03	0.62 ± 0.04	0.58 ± 0.01	0.57 ± 0.01	0.61 ± 0.02	0.66 ± 0.01	0.67 ± 0.01	0.60 ± 0.02
	Text	**0.70 ± 0.02**	**0.68 ± 0.02**	0.59 ± 0.02	0.61 ± 0.01	0.59 ± 0.01	0.66 ± 0.00	**0.71 ± 0.01**	**0.67 ± 0.01**

Table 2: Results of classification based on LLM embeddings of reactions

tral Small) provided top performance in terms of average accuracy and F1-score (Figure 2). Due to their high predictive power and minimal costs (a detailed comparison of LLMs pricing is summarized in Table 7 of the Appendix A.5), Claude 3 Haiku ($k = 6$) and Mistral Small ($k = 6$) (see Figure 4 of the Appendix A.5) were selected for further evaluation of the few-shot approaches on all four datasets. The corresponding results are shown in Table 8 of the Appendix A.5.

Embeddings Classification

Experimental Settings In large language models, embeddings refer to the representation of words, phrases, or tokens as dense, low-dimensional vectors in a continuous vector space. These embeddings capture semantic and syntactic relationships between words and are learned during the training process of the language model. To obtain reaction embeddings, `text-embedding-3-large` from OpenAI with the embeddings size of 3072 and the public Mistral-7B with 4096 dimensions were utilized. The choice of the models was motivated by several factors. Firstly, we wanted to compare few-shot and embeddings results for the same providers (Anthropic does not provide any embedding models). Secondly, in order to fairly compare embeddings models between each other, we opted for models of different sizes and costs.

Similar to few-shot experiments, we used SMILES strings and text descriptions of reactions as input reaction representations. Embeddings classification was performed using the XGBoost algorithm (see Appendix A.7 for details).

Results The results are shown in Table 2. Models trained on the textual data representations demonstrated significantly better performance in comparison with SMILES embeddings, when trained on the USPTO-C, USPTO-R and ORD-R datasets. In the case of the ORD-C dataset, metrics obtained with text embeddings were identical to those of SMILES representations, but had lower standard deviations, which indicates improved consistency. The impressive average accuracy of 0.71 was achieved by `text-embedding-3-large` on the ORD-R dataset.

In an attempt to find an explanation for this outcome, we visualized the space of USPTO-R embeddings reduced to 3D with UMAP. We did the same for the USPTO-R DRFP representations (one of the most common and computationally efficient ways of representing reactions (Probst, Schwaller, and Reymond 2022)) and applied a colormap highlighting chemical reaction yield values (Figure 3a). Strikingly, we observed a clear gradient from low to high yielding reactions for the Mistral-7B embeddings, whereas DRFPs appeared as a random cloud of data points. This finding holds a great promise for the future of chemical reaction representations, already serving as the new state of the art for the generalized reaction yield prediction.

Comparison with Baseline Approaches

Fingerprints Fingerprints are one of the most popular and efficient representations of chemical substances. In particular, the XGBoost model trained on differential reaction fingerprints was shown to deliver top performance in generalized yield prediction task (Probst, Schwaller, and Reymond 2022), which motivated us to use this combination as one of our baseline models. A description of the models' optimization process is presented in Appendix A.8.

Transformers The transformer architecture was introduced in 2017 and was repeatedly reported to achieve state-of-the-art performance on various NLP tasks (Van Nguyen et al. 2021; Ramos-Pérez, Alonso-González, and Núñez-Velázquez 2021; Grail, Perez, and Gaussier 2021). Transformers have been evaluated in yield prediction task as well. The most interesting works on this topic include Yield-BERT and Egret, which were discussed in subsection . Both models were also trained on the USPTO data, so we used them as additional baselines. Since the models were originally created to solve the regression problem, we converted their outputs into yield categories and thus evaluated their classification performance. The weights of the pre-trained models were taken from the official GitHub repositories and the forward pass was performed once for each of our test sets to calculate the corresponding performance metrics.

Results The comparison of LLMs with baseline approaches is shown in Table 3. Notably, the two transformer-based state-of-the-art approaches showed rather modest performance, despite being trained on the USPTO data. Our own baseline model achieved comparable metrics and even outperformed the transformer-based approaches on the USPTO-C and ORD-R datasets. These empirical results suggest poor generalization power of the existing models in the reaction yield prediction task.

Conversely, the LLM-based approaches produced the best scores across all the datasets. Interestingly, we observed competitive performance of the LLM few-shot approach with accuracy of 0.61 and overall best F1-score of 0.68 achieved on the USPTO-R dataset. The few-shot performance on ORD-based datasets was slightly lower compared

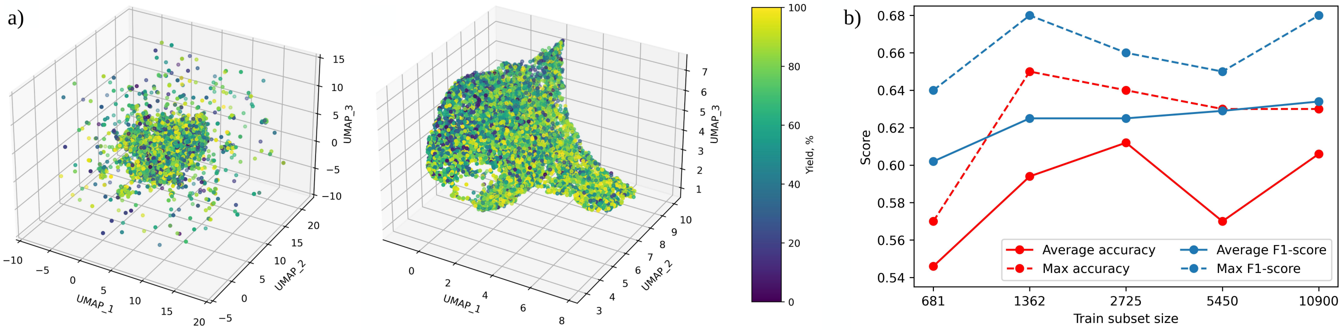


Figure 3: a) Visualization of UMAP-reduced reactions representations: DRFPs (on the left) and Mistral-7B embeddings derived from texts of reaction descriptions (on the right); b) Investigation of the optimal training subset size with Mistral Small model

	USPTO-C		USPTO-R		ORD-C		ORD-R	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Yield-BERT*	0.58	0.57	0.61	0.59	0.60	0.60	0.59	0.55
Egret*	0.59	0.55	0.51	0.42	0.62	0.61	0.54	0.51
XGB + DRFP**	0.67 ± 0.03	0.67 ± 0.03	0.51 ± 0.02	0.53 ± 0.02	0.58 ± 0.02	0.61 ± 0.02	0.69 ± 0.03	**0.68 ± 0.03**
LLM few-shot	0.62 ± 0.02[a]	0.65 ± 0.07[b]	0.61 ± 0.02[a]	**0.68 ± 0.03**[b]	0.51 ± 0.01[a]	0.65 ± 0.02[b]	0.53 ± 0.04[a]	0.57 ± 0.04[b]
XGB + LLM emb.	**0.70 ± 0.02**[d]	**0.68 ± 0.02**[d]	**0.62 ± 0.04**[c]	0.64 ± 0.05[c]	**0.66 ± 0.02**[c]	**0.71 ± 0.02**[c]	**0.71 ± 0.01**[d]	0.67 ± 0.01[d]

Table 3: Performance of LLMs and baseline approaches on ORD-C, ORD-R, USPTO-C, USPTO-R datasets: a) Mistral Small (Random sampler, $k = 6$); b) Claude 3 Haiku (Random sampler, $k = 6$); c) Mistral-7B embeddings derived from text descriptions; d) `text-embedding-3-large` derived from text descriptions. *reproduced baseline state-of-the-art models, **our own trained baseline models

to USPTO data. To make sure that the prompting strategy we have chosen is appropriate for all datasets, we conducted additional experiments with a wider range of k values and LLMs, that revealed no significant changes for ORD-R dataset (see Appendix A.6 for details).

The best accuracy was consistently achieved by the XGB trained on LLM embeddings. Strikingly, we obtained up to 12% increase in accuracy (for ORD-R) and F1-score (for USPTO-C and ORD-R) compared to the transformer-based approaches. This highlights the enriched information contents of the LLM embeddings compared to DRFP and, likely, molecular fingerprints in general.

Optimal Training Set Investigation

Experimental Settings While LLM embeddings provide the best results according to our experiments, the few-shot approach is especially interesting for applications due to its strong performance and the ease of use. Rather than training and optimizing machine learning models, one can just provide an LLM with examples from the training set and ask a question of interest. In this regard, knowing the minimum size of the training set required to achieve acceptable results is of great importance. Particularly in scenarios when the data has to be experimentally generated.

We used the USPTO-R dataset to evaluate the optimal training set size with Mistral Small model (textual data format, Random sampler, $k = 6$). Apart from the original USPTO-R dataset containing 10900 reactions, we randomly sampled from it to obtain four other datasets, each time reducing the number of training samples by half. This procedure resulted in datasets of 5450, 2725, 1362 and 681 training reactions. The test subset was never changed.

Results To compare Mistral Small performance on datasets with different training sizes, we calculated mean accuracy and F1-score as well as the maximum of these metrics across 5 random states. The results are shown in Figure 3b. Somewhat intuitive, we observed a positive correlation between the scores and the train subset sizes with the worst predictions corresponding to the smallest dataset of 681 reactions. Increasing the number of reactions in the training subset leads to improved performance with maximum accuracy achieved at 1362 samples and best average accuracy - at 2725 samples. Thus, the training size can be reduced by 8 times without considerable losses in accuracy. This discovery not only streamlines the preparation of training data but also significantly reduces the time required to conduct experiments.

Surprisingly, a drop in average accuracy and maximum F1-score was observed at 5450 reactions. The emergence of such performance anomalies serves as a reminder of the complexity of the LLM-based solutions. In this regard, we strongly recommend conducting practically oriented experiments for each individual application to ensure consistent and stable predictions.

Discussion

Are LLMs Embeddings the Novel State-of-the-Art Reaction Representation?

Comparison of models trained on reaction fingerprints with those trained on LLM embeddings revealed the superior performance of the latter. One plausible explanation for this phenomenon is the enhanced context of reaction representations reflected in LLM embeddings. Unlike reaction fingerprints, which primarily capture structural information based on atom connectivity, LLM embeddings grasp the nuanced semantics present within chemical reactions. By representing reactions as sentences, LLMs capture not only the molecular structures involved but also the relationships, transformations, and conditions dictating the yield of the reaction.

This astonishing result opens a new research horizon for the yield prediction and potentially many other tasks of predictive chemistry. Similarly to the 8 benchmarking tasks for the few-shot applications proposed by Guo et al. (Guo et al. 2023), it is essential to formulate a comprehensive set of benchmarks to evaluate the performance of LLM embeddings in different areas of chemistry.

Peculiarities of LLMs in Few-Shot Reaction Yield Classification

During our experiments with the few-shot classification approach, we noticed some curious aspects of LLM behavior, which we would like to discuss in this section.

First, we observed that standard deviations of the few-shot metrics are, in some cases, lower than those of XGB and DRFP combination (Table 3). It suggests that a proper configuration of the few-shot approach offers competitive advantages to traditional approaches not only in terms of accuracy, but also reflected in consistency and reliability of predictions.

Second, we took some notes on the performance of GPT models. Even with the best prompting strategy (Random sampler with the textual data format), GPT-4 and GPT-3.5 Turbo produced the highest standard deviations up to 56% of the mean value for F1-scores (see Figure 5 of the Appendix A.5). Strikingly, in some cases the standard deviations of F1-scores for GPT-3.5 Turbo exceeded the mean values. GPT models are known for their high complexity and capacity to capture intricate patterns in data. This complexity might lead to a wider range of behaviors when adapting to specialized few-shot tasks, resulting in higher variability in performance. Moreover, we noticed that in some cases GPT-4 provided the same response for all the reactions in the test subset resulting in almost zero F1-scores. Therefore, we would not recommend using GPT models for the generalized yield prediction task.

Finally, we came to a conclusion that in certain scenarios LLMs with a smaller number of parameters cope better with the task than generally more performant models of the same provider. Apparently, more advanced models in classical NLP tasks do not always excel in chemistry-related problems. The researchers should take this observation into account when designing their own LLM-based solutions.

Advantages and Limitations of the Proposed Methods

Using LLMs for reaction yield prediction has undoubted advantages over traditional ML approaches. Existing yield prediction models often rely on handcrafted features derived from molecular descriptors or reaction conditions, which take a lot of time and computational resources to prepare. In the few-shot approach, on the other hand, only the original training reactions are needed, possibly in the textual format. Another benefit of using LLMs is their ability to capture relationships between chemical substances in reaction descriptions. Such ability is often absent in fingerprints and graphs, which only encode the structures of molecules involved in a particular reaction. Finally, as opposed to traditional approaches, LLMs can be used effectively for both generalized and type-specific yield prediction, as they are originally trained on a diverse range of texts and have a higher generalization ability.

Despite the aforementioned benefits, LLM-based approaches have certain limitations. The first one is the lack of domain-specific expertise. LLMs are typically trained on a wide range of general-purpose public texts, which may not fully capture the nuanced knowledge required for such a specific field as chemistry. Additionally, LLMs may struggle with novel scientific achievements (for example, novel reactions) that appeared after the model had been trained. One other drawback of LLMs is the lack of interpretability. Decisions made by LLMs in the few-shot classification as well as interconnections encoded in LLM embeddings are difficult to comprehend. We encountered this limitation in our study when a mysterious drop in LLM performance with 5450 training reactions was observed. However, it is noteworthy that many of the traditional approaches have similar limitations, perhaps to a lower extent.

Conclusion and Future Work

In this study, we explored the efficacy of general-purpose LLMs in the domain of chemical reaction yield prediction without binding to a specific reaction class. We demonstrated the superior performance of LLM-based approaches over baseline models, investigated their robustness and practical applicability in scarce data scenarios. We also discussed the limitations of the presented approaches and the prospects of even broader use of LLMs in predictive chemistry.

While our study provides valuable insights into the efficacy of the few-shot approach and LLM embeddings in generalized yield prediction, several avenues for future research emerge from our findings. One promising direction is to explore the reasoning behind the LLM answers to grasp the aspects of key importance in the LLM decision making. Predicting reaction yields based on the complete descriptions of synthesis procedures represent another avenue for future research with potentially high impact.

Acknowledgements

The authors thank the Priority 2030 Federal Academic Leadership Program for infrastructure support.

References

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; and Doyle, A. G. 2018. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385): 186–190.

Bajusz, D.; Rácz, A.; and Héberger, K. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7: 1–13.

Biyani, S. A.; Moriuchi, Y. W.; and Thompson, D. H. 2021. Advancement in organic synthesis through high throughput experimentation. *Chemistry-Methods*, 1(7): 323–339.

Bustillo, L.; and Rodrigues, T. 2023. A focus on the use of real-world datasets for yield prediction. *Chemical Science*, 14(19): 4958–4960.

Callaghan, S. 2021. Toward machine learning-enhanced high-throughput experimentation for chemistry. *Patterns*, 2(3).

CAS. 2008. CAS SciFinder - Chemical Compound Database. https://scifinder-n.cas.org/.

Chen, K.; Chen, G.; Li, J.; Huang, Y.; Wang, E.; Hou, T.; and Heng, P.-A. 2023. MetaRF: attention-based random forest for reaction yield prediction with a few trails. *Journal of Cheminformatics*, 15(1): 43.

Collins, K. D.; and Glorius, F. 2015. Intermolecular reaction screening as a tool for reaction evaluation. *Accounts of chemical research*, 48(3): 619–627.

Dubrovsky, I.; Dmitrenko, A.; Dmitrenko, A.; Serov, N.; and Vinogradov, V. 2024. Unveiling the Potential of AI for Nanomaterial Morphology Prediction. *arXiv preprint arXiv:2406.02591*.

Eyke, N. S.; Koscher, B. A.; and Jensen, K. F. 2021. Toward machine learning-enhanced high-throughput experimentation. *Trends in Chemistry*, 3(2): 120–132.

Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; and Schindler, T. 2023. Machine Learning C–N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS omega*, 8(3): 3017–3025.

Fu, Z.; Li, X.; Wang, Z.; Li, Z.; Liu, X.; Wu, X.; Zhao, J.; Ding, X.; Wan, X.; Zhong, F.; et al. 2020. Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction. *Organic Chemistry Frontiers*, 7(16): 2269–2277.

Goodman, J. 2009. Computer software review: Reaxys.

Grail, Q.; Perez, J.; and Gaussier, E. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, 1792–1810.

Guo, T.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X.; et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36: 59662–59688.

Haywood, A. L.; Redshaw, J.; Hanson-Heine, M. W.; Taylor, A.; Brown, A.; Mason, A. M.; Gaertner, T.; and Hirst, J. D. 2021. Kernel methods for predicting yields of chemical reactions. *Journal of Chemical Information and Modeling*, 62(9): 2077–2092.

Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; et al. 2023. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5): 1233–1250.

Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B.-L.; and Xia, N. 2021. When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*, 9: 85071–85083.

Karpov, P.; Godin, G.; and Tetko, I. V. 2019. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*, 817–830. Springer.

Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; and Coley, C. W. 2021. The open reaction database. *Journal of the American Chemical Society*, 143(45): 18820–18826.

Keraghel, I.; Morbieu, S.; and Nadif, M. 2024. Beyond words: a comparative analysis of LLM embeddings for effective clustering. In *International Symposium on Intelligent Data Analysis*, 205–216. Springer.

Kwon, Y.; Lee, D.; Choi, Y.-S.; and Kang, S. 2022. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics*, 14: 1–10.

Latif, E.; Parasuraman, R.; and Zhai, X. 2024. PhysicsAssistant: An LLM-Powered Interactive Learning Robot for Physics Lab Investigations. *arXiv preprint arXiv:2403.18721*.

Li, T.; Shetty, S.; Kamath, A.; Jaiswal, A.; Jiang, X.; Ding, Y.; and Kim, Y. 2024. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digital Medicine*, 7(1): 40.

Lin, K.; Xu, Y.; Pei, J.; and Lai, L. 2020. Automatic retrosynthetic route planning using template-free models. *Chemical science*, 11(12): 3355–3364.

Liu, H.; Yin, H.; Luo, Z.; and Wang, X. 2024a. Integrating Chemistry Knowledge in Large Language Models via Prompt Engineering. *arXiv preprint arXiv:2404.14467*.

Liu, Y.; Ding, S.; Zhou, S.; Fan, W.; and Tan, Q. 2024b. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv preprint arXiv:2406.12950*.

Lowe, D. M. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, Pembroke College.

Lu, J.; and Zhang, Y. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 62(6): 1376–1387.

M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1–11.

Mar, J.; and Liu, J. 2022. Constrained Few-Shot Learning: Human-Like Low Sample Complexity Learning and Non-Episodic Text Classification. *arXiv preprint arXiv:2208.08089*.

McInnes, L.; Healy, J.; Astels, S.; et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11): 205.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McNaughton, A. D.; Ramalaxmi, G.; Kruel, A.; Knutson, C. R.; Varikoti, R. A.; and Kumar, N. 2024. CACTUS: Chemistry Agent Connecting Tool-Usage to Science. *arXiv preprint arXiv:2405.00972*.

Meuwly, M. 2021. Machine learning for chemical reactions. *Chemical Reviews*, 121(16): 10218–10239.

Nguyen, D. Q.; Zhai, Z.; Yoshikawa, H.; Fang, B.; Druckenbrodt, C.; Thorne, C.; Hoessel, R.; Akhondi, S. A.; Cohn, T.; Baldwin, T.; et al. 2020. ChEMU: named entity recognition and event extraction of chemical reactions from patents. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, 572–579. Springer.

Noyori, R. 2009. Synthesizing our future. *Nature chemistry*, 1(1): 5–6.

Petukhova, A.; Matos-Carvalho, J. P.; and Fachada, N. 2024. Text clustering with LLM embeddings. *arXiv preprint arXiv:2403.15112*.

Probst, D.; Schwaller, P.; and Reymond, J.-L. 2022. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery*, 1(2): 91–97.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramos-Pérez, E.; Alonso-González, P. J.; and Núñez-Velázquez, J. J. 2021. Multi-transformer: A new neural network-based architecture for forecasting S&P volatility. *Mathematics*, 9(15): 1794.

Ruiz-Castillo, P.; and Buchwald, S. L. 2016. Applications of palladium-catalyzed C–N cross-coupling reactions. *Chemical reviews*, 116(19): 12564–12649.

Sadeghi, S.; Bui, A.; Forooghi, A.; Lu, J.; and Ngom, A. 2024. Can large language models understand molecules? *BMC bioinformatics*, 25(1): 225.

Saebi, M.; Nan, B.; Herr, J.; Wahlers, J.; Wiest, O.; and Chawla, N. 2021. Graph neural networks for predicting chemical reaction performance. *ChemRxiv*.

Sato, A.; Miyao, T.; and Funatsu, K. 2022. Prediction of Reaction Yield for Buchwald-Hartwig Cross-coupling Reactions Using Deep Learning. *Molecular Informatics*, 41(2): 2100156.

Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; and Laino, T. 2018. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28): 6091–6098.

Schwaller, P.; Vaucher, A. C.; Laino, T.; and Reymond, J.-L. 2020. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. In *Proceedings of NeurIPS 2020 Machine Learning for Molecules Workshop*.

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

Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; and Gambin, A. 2017. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific reports*, 7(1): 3582.

Van Nguyen, M.; Lai, V. D.; Veyseh, A. P. B.; and Nguyen, T. H. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. *arXiv preprint arXiv:2101.03289*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vogel, I. 1974. *Practical organic chemistry*. Citeseer.

Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; and Tetko, I. V. 2023. When yield prediction does not yield prediction: an overview of the current challenges. *Journal of Chemical Information and Modeling*, 64(1): 42–56.

Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; and Gurbych, O. 2023. Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *Journal of Computational Chemistry*, 44(2): 76–92.

Yin, X.; Hsieh, C.-Y.; Wang, X.; Wu, Z.; Ye, Q.; Bao, H.; Deng, Y.; Chen, H.; Luo, P.; Liu, H.; et al. 2024. Enhancing Generic Reaction Yield Prediction through Reaction Condition-Based Contrastive Learning. *Research*, 7: 0292.

Yu, B.; Baker, F. N.; Chen, Z.; Ning, X.; and Sun, H. 2024. LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset. *arXiv preprint arXiv:2402.09391*.

Zhang, S.-Q.; Xu, L.-C.; Li, S.-W.; Oliveira, J. C.; Li, X.; Ackermann, L.; and Hong, X. 2023. Bridging chemical knowledge and machine learning for performance prediction of organic synthesis. *Chemistry–A European Journal*, 29(6): e202202834.

Zhao, Y.; Liu, X.; Lu, H.; Zhu, X.; Wang, T.; Luo, G.; Zheng, R.; and Luo, Y. 2021. An optimized deep convolutional neural network for yield prediction of Buchwald-Hartwig amination. *Chemical Physics*, 550: 111296.

Zhao, Z.; Ma, D.; Chen, L.; Sun, L.; Li, Z.; Xu, H.; Zhu, Z.; Zhu, S.; Fan, S.; Shen, G.; et al. 2024. ChemDFM: Dialogue Foundation Model for Chemistry. *arXiv preprint arXiv:2401.14818*.

A Appendix

A.1 Prediction of Reaction Yields as a Regression Problem

In order to demonstrate challenges in predicting the exact value of reaction yield we conducted experiments using three state-of-the-art approaches described in subsection 2.1 of the main part, namely Yield-BERT, Egret and XGBoost regressor trained on DRFPs (XGB + DRFP). Models were evaluated using 4 reactions datasets described in section 3.

In case of XGB and DRFP combination, the optimal fingerprints length was determined with default XGB parameters for each dataset (see Figure 1). Then grid-search over the following XGB parameters was ran: `n_estimators`, `max_depth`, `learning_rate`, `gamma`, `colsample_bytree`. The parameters grid is represented in Table 1. Best parameters set was searched using validation subsets of the datasets, while the overall models performance was evaluated on the test subsets. Best hyperparameters for each dataset are represented in Table 2. Each experiment with the best hyperparameters was conducted 5 times with different random states and mean and standard deviation of regression metrics were calculated. The results are represented in Table 3.

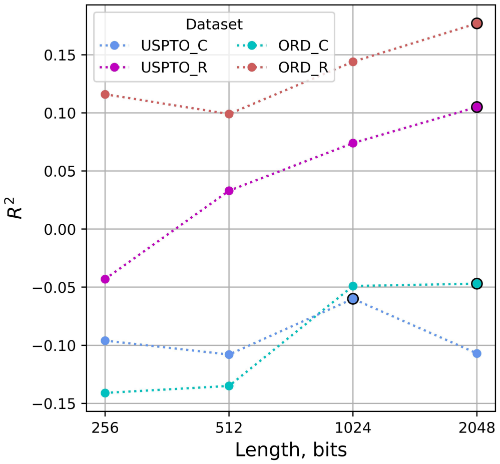


Figure 1: XGBoost regressor performance depending on DRFP length. DRFPs radius is consistent and equals 2. Best configuration for each dataset is circled.

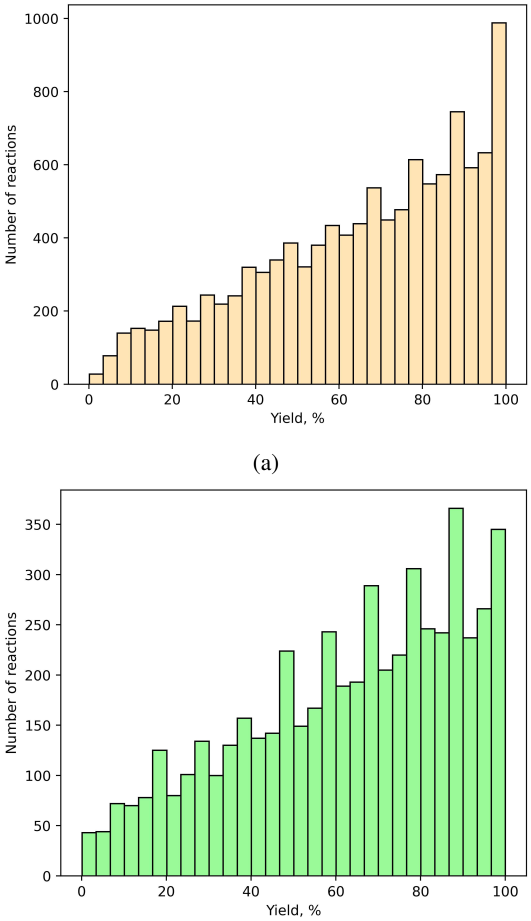
It can be observed, that the overall efficacy of the models is quite low. XGBoost models trained on DRFPs show top performance among other approaches, however such level of predictive ability remains unsatisfactory for practical tasks where precise yield values are required. Possible reasons for that could be strong dependence of yield values on reaction conditions (such as temperature, pressure, the choice of solvents, catalysts etc.) as well as the uneven distribution of yields in datasets with prevalence of higher values due to the tendency of scientists to publish only successful results of syntheses. The yields distributions in the USPTO-R and ORD-R datasets are shown in Figure 2.

parameter	values
`n_estimators`	100, 200, 300, 400, 500
`max_depth`	2, 4, 6, 8, 10
`learning_rate`	0.01, 0.05, 0.1
`gamma`	0.1, 0.5, 1, 1.5
`colsample_bytree`	0.5, 0.7, 0.9

Table 1: Parameters grid for grid-search

	USPTO-C	USPTO-R	ORD-C	ORD-R
`n_estimators`	300	500	200	500
`max_depth`	4	6	10	10
`learning_rate`	0.01	0.05	0.01	0.05
`gamma`	0.1	1.5	1.5	1.5
`colsample_bytree`	0.9	0.9	0.9	0.5

Table 2: Best hyperparameters of XGBoost regressors trained on DRFPs



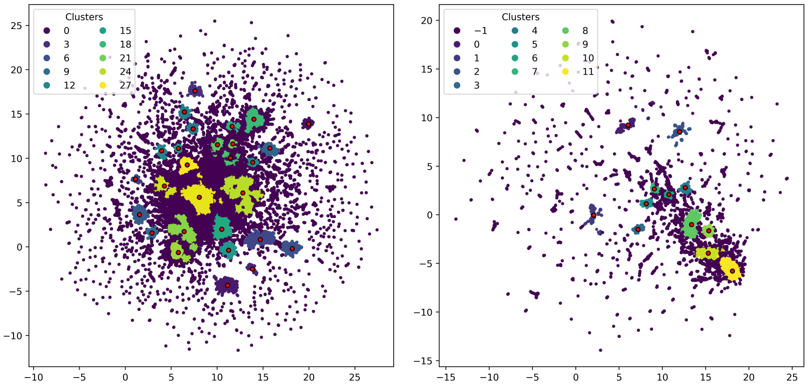
(a)

(b)

Figure 2: Distribution of yield values: a) USPTO-R; b) ORD-R

	USPTO-C		USPTO-R		ORD-C		ORD-R	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Yield-BERT	0.020	25.404	-0.030	25.913	-0.017	24.750	0.009	28.305
Egret	-0.070	26.542	-0.161	27.509	-0.054	25.195	-0.043	29.035
XGB + DRFP	**0.092 ± 0.002**	**24.454 ± 0.026**	**-0.029 ± 0.012**	**25.902 ± 0.156**	**0.084 ± 0.003**	**23.478 ± 0.043**	**0.109 ± 0.003**	**26.829 ± 0.049**

Table 3: The performance of state-of-the-art methods in prediction of reaction yields treated as a regression problem. We do not provide mean and standard deviation of metrics obtained with Yield-BERT and Egret since they were reproduced using pre-trained weights available on the official GitHub repositories of these models



(a)

(b)

Figure 3: Clustering of UMAP embeddings with HDBSCAN: a) USPTO; b) ORD

A.2 Preparation of Datasets

Chemical Reaction A chemical reaction includes several components: *i)* reactants - chemically interacting organic molecules consumed in the reaction process to make products; *ii)* reagents - substances added to cause a reaction that are not necessarily consumed and not always present in a reaction. Reagents usually include solvents, salts, catalysts etc.; *iii)* products - substances produced as a result of a reaction.

USPTO Data Processing The USPTO dataset was additionally processed to remove erroneous and ambiguous yields. All miscalculated and misprinted values as well as erroneous yields (containing no measurement units, negative, zero or exceeding 100% values) were removed. If a yield was reported two times for the same reaction or reported as a range, the larger value was preserved. As a result, approximately 50% of incomplete or missing values were filtered out. Reactions containing more than two reactants were also removed, since they made up only 6.7% of the dataset. The resulting dataset contained more than 526k reactions represented as SMILES strings.

ORD Data Processing The original ORD database contained around 2.3M reactions. We filtered out reactions where no yields and no reaction SMILES were reported. We also dropped reactions with multiple yields and canonicalized reaction SMILES using RDKit package. Then, we excluded reactions that overlap with the USPTO dataset and the resulting ORD dataset contained around 21k reactions.

USPTO and ORD Clustering Reactions from ORD dataset represented as DRFPs were compressed to 2 components using UMAP algorithm with `sokalmichener` metric for binary data. HDBSCAN with parameters `min_cluster_size=225`, `max_cluster_size=10000` was used for clustering UMAP embeddings resulting in 12 clusters. The same operations were performed on the USPTO dataset `kulsinski` metric for UMAP and `min_cluster_size=2000`, `max_cluster_size=60000` for HDBSCAN resulting in 28 clusters. The ”-1” cluster was not taken into account when USPTO-C and ORD-C datasets were engineered. Visualization of clustering results for USPTO and ORD is shown in Figure 3.

A.3 Conversion of Reaction SMILES into Text Descriptions

Example 1:

SMILES: NN.O=C(Cl)c1c(F)ccc(F)c1F>CCO.CO.ClCCl.O>NNC(=O)c1c(F)ccc(F)c1F

Text description: Hydrazine and 2,3,6-trifluorobenzoyl chloride react together in the presence of ethanol, methanol, dichloromethane, oxidane to produce 2,3,6-trifluorobenzohydrazide.

Example 2:

SMILES: C#CCBr.COCCc1nc2cnc3ccccc3c2n1CCO>> C#CCOCCn1c(CCOC)nc2cnc3ccccc3c21

Text description: 3-bromoprop-1-yne and 2-[2-(2-methoxyethyl)imidazo[4,5-c]quinolin-1-yl]ethanol react together to produce 2-(2-methoxyethyl)-1-(2-prop-2-ynoxyethyl)imidazo[4,5-c]quinoline.

A.4 Instruction Prompt for SMILES Data Format

In the study of Guo et al. (2023) it was shown, that it is difficult for LLMs to generate accurate responses when SMILES strings appear in prompt. In an attempt to improve this situation, we added some information about reaction SMILES into the instruction prompt. In particular, we explained the order and meanings of the reaction SMILES components. The prompt is as follows:

You are an expert chemist. Your task is to predict reaction yields based on SMILES representations of organic reactions. Reaction SMILES consist of potentially three parts (reactants, agents, and products) each separated by an arrow symbol '>'. Reactants are listed before the arrow symbol. If a reaction includes agents, such as catalysts or solvents, they can be included after the reactants. Products are listed after the second arrow symbol, representing the resulting substances of the reaction. You can only predict whether the reaction is 'High-yielding' or 'Not high-yielding'. 'High-yielding' reaction means the yield rate of the reaction is above 70%. 'Not high-yielding' means the yield rate of the reaction is below 70%. You will be provided with several examples of reactions and corresponding yield rates. Please answer with only 'High-yielding' or 'Not high-yielding', no other information can be provided.

A.5 Few-Shot Classification Results

Results for few-shot classification experiments with different prompting strategies are shown in Tables 4, 5 and 6.

Comparison of LLMs pricing is represented in Table 7.

Performance analysis of GPT-4 and GPT-3.5 Turbo in few-shot classification is shown in Figure 5.

The choice of the optimal number of shots for Mistral Small and Cluade 3 Haiku can be observed in Figure 4. Results for Mistral Small and Claude 3 Haiku with the best prompting strategy across all datasets are represented in Table 8.

	$k = 2$		$k = 4$		$k = 6$		$k = 8$		$k = 10$	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Claude 3 Haiku	0.52 ± 0.02	0.51 ± 0.29	0.55 ± 0.03	0.47 ± 0.28	0.52 ± 0.01	0.63 ± 0.07	**0.55 ± 0.04**	**0.63 ± 0.06**	0.54 ± 0.04	0.62 ± 0.08
Claude 3 Opus	0.51 ± 0.02	0.54 ± 0.30	0.50 ± 0.01	0.47 ± 0.26	0.51 ± 0.02	0.56 ± 0.16	**0.53 ± 0.02**	**0.63 ± 0.11**	0.50 ± 0.02	0.48 ± 0.30
Mistral Small	**0.49 ± 0.03**	0.64 ± 0.03	0.49 ± 0.06	0.48 ± 0.13	0.44 ± 0.03	0.39 ± 0.14	0.43 ± 0.02	0.37 ± 0.18	0.44 ± 0.03	0.36 ± 0.19
Mistral Large	0.51 ± 0.02	0.12 ± 0.06	0.51 ± 0.03	0.29 ± 0.07	0.54 ± 0.05	0.58 ± 0.09	0.53 ± 0.03	0.6 ± 0.05	**0.53 ± 0.03**	**0.62 ± 0.04**
GPT-3.5 Turbo	0.49 ± 0.04	0.48 ± 0.30	0.49 ± 0.02	0.28 ± 0.36	0.49 ± 0.02	0.39 ± 0.36	**0.51 ± 0.01**	**0.52 ± 0.20**	0.50 ± 0.02	0.45 ± 0.32
GPT-4	0.51 ± 0.01	0.46 ± 0.31	0.50 ± 0.01	0.30 ± 0.35	0.50 ± 0.01	0.40 ± 0.37	**0.50 ± 0.02**	**0.58 ± 0.16**	0.50 ± 0.01	0.50 ± 0.29

Table 4: The performance of LLMs with the following prompting strategy: SMILES data format, Tanimoto sampler

	$k = 2$		$k = 4$		$k = 6$		$k = 8$		$k = 10$	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Claude 3 Haiku	**0.55 ± 0.00**	**0.69 ± 0.00**	0.49 ± 0.01	0.14 ± 0.07	0.55 ± 0.03	0.58 ± 0.09	0.56 ± 0.02	0.59 ± 0.09	0.54 ± 0.04	0.61 ± 0.05
Claude 3 Opus	**0.51 ± 0.00**	**0.68 ± 0.00**	0.50 ± 0.02	0.17 ± 0.09	0.53 ± 0.01	0.59 ± 0.19	0.52 ± 0.01	0.52 ± 0.22	0.51 ± 0.0	0.51 ± 0.23
Mistral Small	**0.50 ± 0.00**	**0.66 ± 0.00**	0.48 ± 0.03	0.35 ± 0.06	0.48 ± 0.04	0.51 ± 0.13	0.46 ± 0.03	0.48 ± 0.11	0.47 ± 0.04	0.49 ± 0.13
Mistral Large	0.48 ± 0.00	0.07 ± 0.00	0.51 ± 0.02	0.25 ± 0.08	0.54 ± 0.03	0.57 ± 0.04	0.53 ± 0.04	0.62 ± 0.07	**0.54 ± 0.02**	**0.65 ± 0.02**
GPT-3.5 Turbo	**0.51 ± 0.00**	**0.68 ± 0.00**	0.48 ± 0.00	0.00 ± 0.00	0.51 ± 0.02	0.50 ± 0.28	0.50 ± 0.03	0.41 ± 0.36	0.49 ± 0.02	0.43 ± 0.33
GPT-4	**0.51 ± 0.00**	**0.68 ± 0.00**	0.49 ± 0.00	0.00 ± 0.00	0.52 ± 0.02	0.53 ± 0.29	0.50 ± 0.01	0.41 ± 0.37	0.50 ± 0.01	0.41 ± 0.36

Table 5: The performance of LLMs with the following prompting strategy: SMILES data format, Random sampler

	$k = 2$		$k = 4$		$k = 6$		$k = 8$		$k = 10$	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Claude 3 Haiku	0.52 ± 0.00	0.68 ± 0.00	0.57 ± 0.02	0.59 ± 0.04	**0.55 ± 0.02**	**0.67 ± 0.02**	0.54 ± 0.03	0.68 ± 0.01	0.53 ± 0.01	0.65 ± 0.03
Claude 3 Opus	0.48 ± 0.00	0.63 ± 0.00	0.57 ± 0.04	0.57 ± 0.04	**0.53 ± 0.03**	**0.64 ± 0.07**	0.53 ± 0.02	0.61 ± 0.09	0.52 ± 0.04	0.59 ± 0.11
Mistral Small	0.51 ± 0.00	0.66 ± 0.00	0.59 ± 0.03	0.60 ± 0.04	**0.61 ± 0.02**	**0.63 ± 0.04**	0.56 ± 0.03	0.62 ± 0.02	0.57 ± 0.03	0.61 ± 0.05
Mistral Large	0.58 ± 0.00	0.61 ± 0.00	0.56 ± 0.03	0.40 ± 0.11	0.61 ± 0.02	0.57 ± 0.05	**0.59 ± 0.02**	**0.66 ± 0.03**	0.57 ± 0.03	0.62 ± 0.04
GPT-3.5 Turbo	**0.51 ± 0.00**	**0.68 ± 0.00**	0.54 ± 0.03	0.39 ± 0.07	0.53 ± 0.02	0.61 ± 0.13	0.53 ± 0.04	0.55 ± 0.21	0.52 ± 0.04	0.56 ± 0.16
GPT-4	**0.51 ± 0.01**	**0.67 ± 0.00**	0.51 ± 0.01	0.07 ± 0.04	0.55 ± 0.01	0.53 ± 0.2	0.55 ± 0.02	0.51 ± 0.25	0.52 ± 0.03	0.48 ± 0.27

Table 6: The performance of LLMs with the following prompting strategy: text data format, Random sampler

	Claude 3 Haiku	Claude 3 Opus	Mistral Small	Mistral Large	GPT-3.5 Turbo	GPT-4
Price per 1M input tokens, USD	0.25	15	1	4	0.5	30
Price per 1M output tokens, USD	1.25	75	3	12	1.5	60

Table 7: Comparison of LLMs pricing for input and output tokens

	USPTO-C		USPTO-R		ORD-C		ORD-R	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Claude 3 Haiku ($k = 6$)	0.58 ± 0.06	**0.65 ± 0.07**	0.56 ± 0.01	**0.68 ± 0.03**	**0.51 ± 0.01**	**0.65 ± 0.02**	0.50 ± 0.03	**0.57 ± 0.04**
Mistral Small ($k = 6$)	**0.62 ± 0.02**	0.64 ± 0.05	**0.61 ± 0.02**	0.63 ± 0.04	**0.51 ± 0.01**	0.53 ± 0.06	**0.53 ± 0.04**	0.46 ± 0.03

Table 8: Results for the best models and corresponding prompting strategies on all datasets

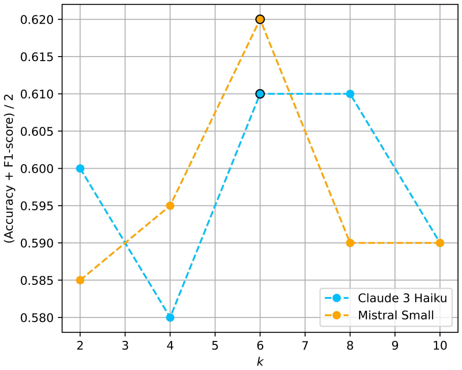


Figure 4: The average between accuracy and F1-score for each k for Mistral Small and Claude 3 Haiku. The optimal configuration for each LLM is circled.

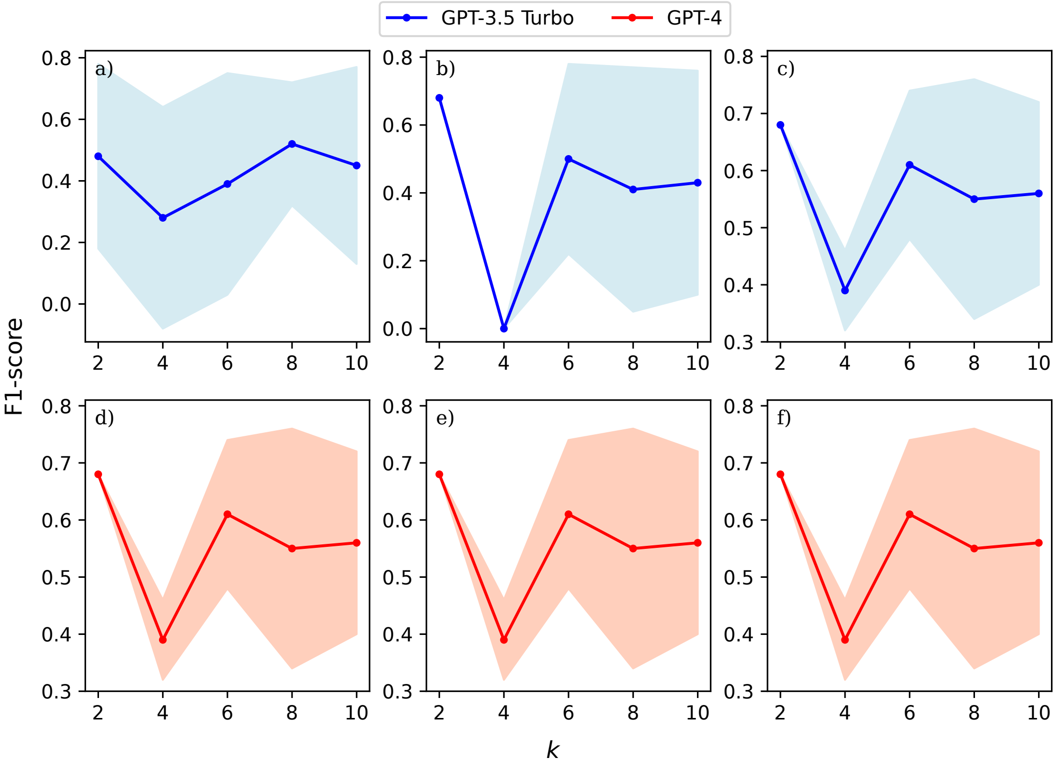
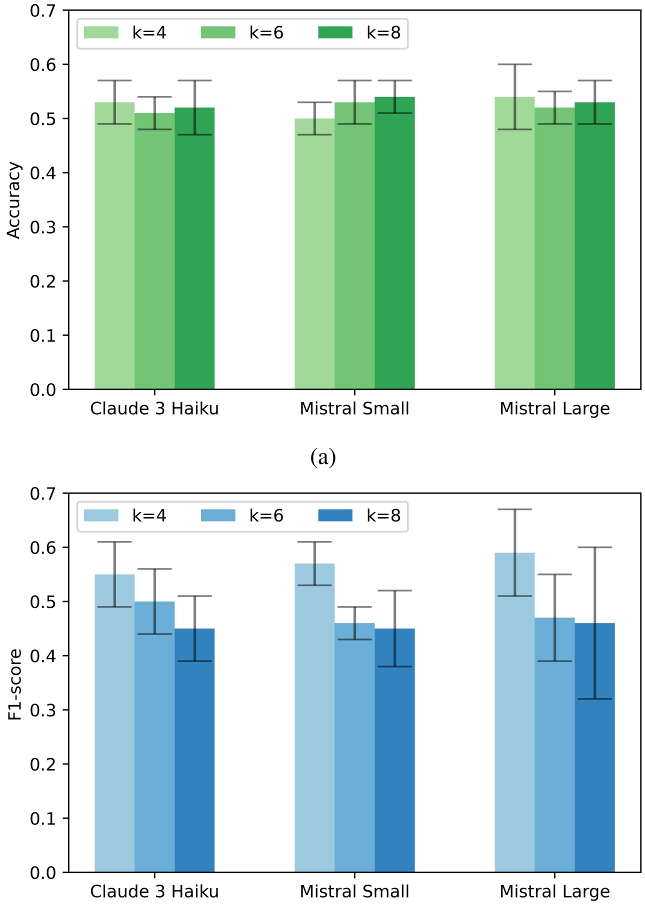


Figure 5: F1-scores and their standard deviations for GPT-3.5 Turbo and GPT-4: a,d) SMILES data format, Tanimoto sampler; b,e) SMILES data format, Random sampler; c,f) textual data format, Random sampler

A.6 Verification of the Prompting Strategy on ORD-Based Datasets

Through experiments on the USPTO-R dataset Mistral Small ($k = 6$) and Claude 3 Haiku ($k = 6$) were selected as the best models for few-shot classification task. To ensure that the choice of LLMs and corresponding k values was appropriate for ORD-based datasets as well, we additionally explored the performance of Mistral Small and Claude 3 Haiku with $k = \{4, 8\}$ and Mistral Large with $k = \{4, 6, 8\}$. Results are represented in Figure 6.

It can be seen, that changes in the k values do not significantly influence LLMs performance. Moreover, standard deviations of metrics within $k = \{4, 8\}$ are higher that within $k = 6$ in some cases. It indicates, that the prompting strategy was selected correctly and does not contribute to the difference in few-shot performance on USPTO- and ORD-based datasets.



(a)

(b)

Figure 6: Performance of Claude 3 Haiku, Mistral Small, and Mistral Large on the ORD-R dataset: a) Average accuracies; b) Average F1-scores

A.7 Embeddings Classification

Grid-search over the parameters grid as in Table 1 was ran. Best parameters set was searched using validation subsets of the datasets, while the overall models performance was evaluated on the test subsets. Hyperparameters leading to the models overfitting were excluded from grid-search results. Best hyperparameters for each dataset are represented in Tables 9, 10, 11 and 12.

	USPTO-C	USPTO-R	ORD-C	ORD-R
n_estimators	400	100	100	100
max_depth	2	8	4	4
learning_rate	0.1	0.05	0.1	0.1
gamma	0.1	1.5	0.5	1
colsample_bytree	0.5	0.5	0.7	0.9

Table 9: Best hyperparameters of XGBoost classifiers trained on Mistral 7B embeddings derived from reaction SMILES

	USPTO-C	USPTO-R	ORD-C	ORD-R
n_estimators	500	100	300	500
max_depth	6	6	4	2
learning_rate	0.01	0.05	0.05	0.1
gamma	0.5	0.1	0.5	0.5
colsample_bytree	0.9	0.9	0.7	0.7

Table 10: Best hyperparameters of XGBoost classifiers trained on Mistral 7B embeddings derived from text descriptions of reactions

	USPTO-C	USPTO-R	ORD-C	ORD-R
n_estimators	100	200	100	200
max_depth	6	4	4	2
learning_rate	0.05	0.1	0.1	0.05
gamma	0.5	1	1.5	0.1
colsample_bytree	0.9	0.9	0.5	0.9

Table 11: Best hyperparameters of XGBoost classifiers trained on `text-embedding-3-large` embeddings derived from reaction SMILES

	USPTO-C	USPTO-R	ORD-C	ORD-R
n_estimators	100	500	200	100
max_depth	10	8	10	8
learning_rate	0.1	0.05	0.05	0.1
gamma	0.5	1	0.1	0.5
colsample_bytree	0.7	0.7	0.9	0.9

Table 12: Best hyperparameters of XGBoost classifiers trained on `text-embedding-3-large` embeddings derived from text descriptions of reactions

A.8 Comparison with Baseline Models

For XGB classification models trained on DRFPs, the optimal fingerprints length was determined with default XGB parameters for each dataset. The results are shown in Figure 7.

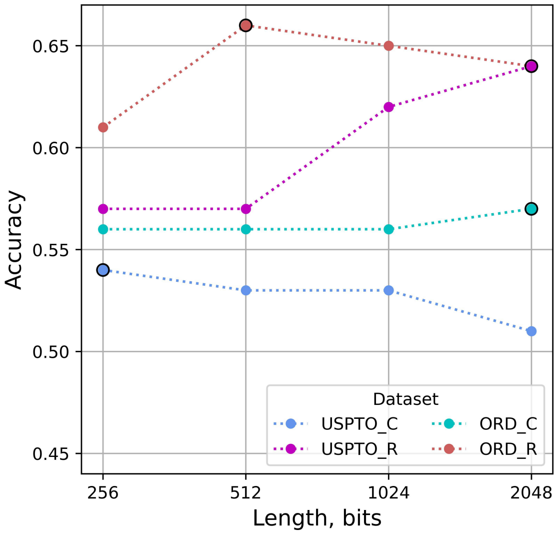


Figure 7: XGBoost performance depending on DRFPs length. DRFPs radius is consistent and equals 2. Best configuration for each dataset is circled.

After that, grid-search over the parameters grid as in Table 1 was ran. Best parameters set was searched using validation subsets of the datasets, while the overall models performance was evaluated on the test subsets. Best hyperparameters for each dataset are represented in Table 13.

	USPTO-C	USPTO-R	ORD-C	ORD-R
n_estimators	100	500	200	100
max_depth	10	8	10	8
learning_rate	0.1	0.05	0.05	0.1
gamma	0.5	1	0.1	0.5
colsample_bytree	0.7	0.7	0.9	0.9

Table 13: Best hyperparameters of XGBoost classifiers trained on DRFPs

Each experiment with the best hyperparameters was conducted 5 times with different random states and mean and standard deviation of regression metrics were calculated.

A.9 Technical Details

CPU	AMD EPYC 7763 64-Core Processor
GPU	NVIDIA RTX A6000
RAM	512 GB
Operating system	Linux
Python	3.10

Table 14: Computing infrastructure used for the experiments

A.10 Societal Impacts of the Study

The use of LLMs in reaction yield prediction presents a range of societal impacts. On the positive side, LLMs have the potential to revolutionize various industries by significantly accelerating the process of drug discovery and development. This can lead to faster production of better medications, ultimately improving public health and saving lives. Moreover, LLMs enable optimization of chemical processes, which can reduce waste and thus minimize environmental harm, contributing to more sustainable industrial practices. The cost savings achieved through accurate prediction of reaction yields can make high-quality products more affordable, benefiting consumers and enhancing economic accessibility. Additionally, the democratization of scientific knowledge facilitated by LLMs allows researchers worldwide, especially in resource-constrained settings, to access advanced predictive tools, fostering innovation and collaboration across borders.

However, these advancements come with potential negative societal impacts as well. The adoption of LLMs may lead to job displacement, particularly for professionals involved in traditional data analysis roles. In addition, ethical concerns arise, such as the risk of data privacy breaches, inherent biases in model predictions, and the misuse of technology for harmful purposes. There is a danger of overreliance on these models, which could undermine critical thinking and problem-solving skills in scientific research. Security risks are another critical concern, as malicious actors could exploit vulnerabilities in LLMs to disrupt chemical processes or introduce errors intentionally. Thus, it is crucial to address the accompanying challenges through ethical guidelines and robust security measures.

References

Guo, T.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X.; et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36: 59662–59688.