BoChemian: Large Language Model Embeddings for Bayesian Optimization of Chemical Reactions

Bojana Ranković BOJANA.RANKOVIC@EPFL.CH
Laboratory of Artificial Chemical Intelligence (LIAC),
National Centre of Competence in Research (NCCR) Catalysis
Institut des Sciences et Ingénierie Chimiques Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

Philippe Schwaller PHILIPPE.SCHWALLER@EPFL.CH
Laboratory of Artificial Chemical Intelligence (LIAC),
National Centre of Competence in Research (NCCR) Catalysis
Institut des Sciences et Ingénierie Chimiques Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

Abstract

This paper explores the integration of Large Language Models (LLM) with Bayesian Optimization (BO) in the domain of chemical reaction optimization with the showcase study on Buchwald-Hartwig reactions. By leveraging LLMs, we can transform textual chemical procedures into an informative feature space suitable for BO. Our findings show that even out-of-the-box open-source LLMs can map chemical reactions for optimization tasks, highlighting their latent specialized knowledge. The results motivate the consideration of further model specialization through adaptive fine-tuning within the BO framework for on-the-fly optimization. This work serves as a foundational step toward a unified computational framework that synergizes textual chemical descriptions with machine-driven optimization, aiming for more efficient and accessible chemical research.
The code is available at: https://github.com/schwallergroup/bochemian

Keywords: Bayesian Optimization, Gaussian processes, Large Language Models, Chemical Reaction Optimization

1. Introduction

Navigating the landscape of chemical reaction optimization is an inherently complex task, as this space is characterized by numerous variables and parameters that influence each other in ways both subtle and profound (Taylor et al., 2023). The application of machine learning (ML) to chemistry has made strides (Coley et al., 2020; Jorner et al., 2021; Schwaller et al., 2022), yet chemical reaction optimization is a field often constrained by a scarcity of data. This limitation renders the application of many ML techniques less than optimal. The answer to these challenges has shifted the spotlight to Bayesian Optimization (BO) as an efficient strategy in low-data scenarios where data are sparse but the stakes are high (Shields et al., 2021; Schweidtmann et al., 2018; Eyke et al., 2020; Felton et al., 2021; Häse

et al., 2021; Pomberger et al., 2022; Müller et al., 2022; Torres et al., 2022; Hickman et al., 2022; Wigh et al., 2023; Guo et al., 2023).

While BO provides a powerful framework for guiding the search in the vast chemical space, its effectiveness is deeply entwined with the quality of the representations used for the reactions (Ranković et al., 2023). The field offers a plethora of ways to encode these, from one-hot encoding (Chuang and Keiser, 2018) and molecular fingerprints (Rogers and Hahn, 2010; Schneider et al., 2015; Capecchi et al., 2020; Probst et al., 2022), to quantum mechanical descriptors (Ahneman et al., 2018; Shields et al., 2021) and data-driven representations (Schwaller et al., 2021). Each comes with its set of trade-offs, be it computational overhead, interpretability, or the required expertise to create and employ them. In this intricate context, one medium stands out for its simplicity, flexibility and depth — natural language. Chemists have long documented the fine details of reactions in textual formats in research papers and supplementary materials, creating a rich collection of information vital for reproducibility and deeper understanding of the nature of chemical reactions (Vaucher et al., 2020; Guo et al., 2021; White, 2023).

Recent advancements in Large Language Models (LLMs) have garnered considerable attention, particularly their utility in a multitude of scientific endeavors (Wei et al., 2022; Bran et al., 2023; Boiko et al., 2023). These models, initially employed for text encoding and generation in data-rich problems, have evolved to solve nuanced challenges in data-scarce fields (Jablonka et al., 2023b). Their capabilities extend beyond mere text generation to potential reasoning and understanding, making them rational candidates for converting human-readable text into computationally actionable insights.

The synergy between the Bayesian optimization and large language models is a novel approach evaluated both in terms of facilitating the usage of BO in highly specialized chemistry domains (Jablonka et al., 2023a) or employing LLMs through in-context learning (Han et al., 2023) for direct catalyst optimization (Caldas Ramos et al., 2023). Moreover, Caldas Ramos et al. (2023) shows that Bayesian optimization coupled with language model embeddings provides promising results for synthesizing novel catalysts. These approaches motivate the exploration of this paradigm further.

Natural language plays a dual role in both chemistry, where it is used to describe chemical reactions, and in large language modeling, where models trained on general text gain expertise in specialized fields, including chemistry. If LLMs internally maintain a state capable of understanding chemical language and if chemists have been leveraging natural language for reaction descriptions in both daily and academic communications, can we harness the power of LLMs to transform the "chemist language" into meaningful and powerful representations for downstream tasks including Bayesian optimization for chemical reactions?

This study takes up the task of answering this question by rigorously evaluating various LLM representations in the context of Bayesian optimization with Gaussian process surrogate models (MacKay et al., 1998; Quinonero-Candela and Rasmussen, 2005) on chemical reaction optimization data. Specifically, we leverage LLMs to transform chemical procedures into a feature space amenable for BO and analyze the intersection of chemical language and optimization. By demonstrating the efficacy of this hybrid approach in identifying high-yield chemical reactions, we make a compelling case for the synergy between textual descriptions of chemical procedures and machine-driven optimization strategies.

2. Methods

The main focus of this study is to evaluate the approach of using different LLM embeddings in Bayesian optimization for chemical reactions. Traditionally, chemical reactions have been explored using quantum mechanic (QM) descriptors. Pioneering work in Bayesian optimization for chemistry (Shields et al., 2021) utilized these representations in the search space of Buchwald-Hartwig coupling reactions (Ahneman et al., 2018) showcasing their applicability to navigate the space towards rich high-yielding regions. These descriptors, although effective, come with computational overhead and require domain-specific expertise for their calculation and interpretation. Recently, simpler and more computationally efficient representations on this dataset have been explored in (Griffiths, 2023). However, there are no examples of reporting performance of LLM embedded features in the domain of Buchwald-Hartwig reactions.

This dataset provides a set of evaluated chemical compounds, namely three bases, four ligands, 22 additives, and 15 aryl halides (Ahneman et al., 2018; Sandfort et al., 2020), which constitute the search space for BO algorithm. Importantly, the dataset comprises reactions that yield five different products. We leverage this diversity by optimizing our model for each individual product, thereby partitioning the dataset accordingly. To enable the analysis, we started by generating a procedure template with placeholders for each of the chemical compounds used and populating them with associated molecular Simplified molecular-input line-entry system (SMILES) (Anderson et al., 1987; Weininger, 1988; Wang et al., 2019).

2.1 Data representation

These procedures are then propagated to various language models for text embedding. We used Massive Text Embedding Benchmark (MTEB) Leaderboard to differentiate five best performing classes of models, namely BGE (Xiao et al., 2023), GTE (Li et al., 2023), E5 (Wang et al., 2022), Instructor model (Su et al., 2022) and OpenAI's text embeddings (ADA) (Neelakantan et al., 2022) and within each class the models containing largest set of parameters. Only the closed-source ADA embeddings were previously used by (Caldas Ramos et al., 2023). For a comprehensive understanding of each model's architecture and training, refer to the Appendinx 4. To compare LLM embeddings to more chemistry specialized features, we also employ RXNFP and DRFP reaction representations. RXNFP (Schwaller et al., 2021) is a data-driven approach that directly maps reaction SMILES to continuous space by finetuning transformer models (Vaswani et al., 2017) on reaction type classification tasks. DRFP (Probst et al., 2022) - differential reaction fingerprint, evaluates and hashes the symmetric difference of the sets containing the circular molecular n-grams generated from 1) reactants and reagents and 2) reaction products, resulting in a chemically meaningful and computationally efficient binary reaction fingerprints. The DRFP was the best-performing chemistry-informed representation in work on additive optimization (Ranković et al., 2023).

2.2 Bayesian Optimization

Bayesian optimization is a powerful framework for optimizing black-box functions by guided sampling of the points in the search space through balancing the trade off between explo-

ration and exploitation. The central component of the BO technique is a probabilistic surrogate model that provides a representation of the underlying function, alongside the uncertainty into its predictions. The uncertainty estimates can be crucial in the decision-making process, indicating the model's confidence in the generated version of the objective function. To navigate through these uncertainties, Bayesian optimization involves an acquisition function that pinpoints the prospective areas of the design space where the true objective function could possibly contain optimal values.

We employ Bayesian optimization (BO) to guide the exploration of the search space towards these promising regions. Our BO setup consists of a Gaussian process surrogate model with a Matern kernel ($\nu = 2.5$) and expected improvement (EI) acquisition function. We select the 10 initial points using k-means clustering method (Morishita and Kaneko, 2022) to widely cover the exploration space for the initial stages of BO algorithm. In order to mimic the real-life scenario of an experimental chemist, we optimize the objective using Kriging believer batch strategy (Ginsbourger et al., 2010) with five suggestions per iteration. The optimization runs for 20 iterations and we repeat each configuration over 20 different seeds to ensure robust results.

3. Results & Discussion

Following the methodology outlined in the Method section, we employed various reaction representations to embed reactions from the Buchwald-Hartwig dataset. Figure 1 **(f)** offers a direct comparison of these representations in their effectiveness to reach the 99th quantile of the reaction search space. Notably, all examined representations outperform a random selection approach, underscoring their utility in systematically exploring the reaction landscape. Among the LLM-based representations we can observe slight differences between the representations using different language models. Instructor embeddings, that map the procedures to vectors through the specialized query `Represent the chemistry procedure:` are particularly interesting. These embeddings surpass other LLM-based approaches but also closely approximate the performance of more chemically explicit methods - DRFP. DRFP maps reactants and products from the reaction SMILES to a binary vector representing the interplay of diverse reaction elements. They explicitly describe chemical realm, yet perform only marginally better than LLM-based embeddings stemming from encoding reaction procedures. Interestingly, representations derived from simply embedding the reaction procedure using large language models outperform other data-driven methods coming from reaction classes fine-tuned transformer models, such as RXNFP. This outcome suggests the synthesis text-based LLM embeddings, are not just semantically rich but also remarkably informative in the context of chemical reactions. Nevertheless, a RXNFP transformer model fine-tuned on a more related task could yield improved embeddings.

To further analyze their performance in uncovering high-yielding reactions in the process of chemical optimization we kept track of the progress of reaching the optimal parameters for each reaction dataset. We present the BO trajectories in plots **(a-e)** of Figure 1. For the sake of brevity and clarity in visualization, we grouped different LLM-based representations under a single category. We can see that even though the maximum achievable yields vary across the five different reactions in this dataset, (BO paths consistently guide the search towards these maxima, reaffirming the informativeness of LLM-based representations.

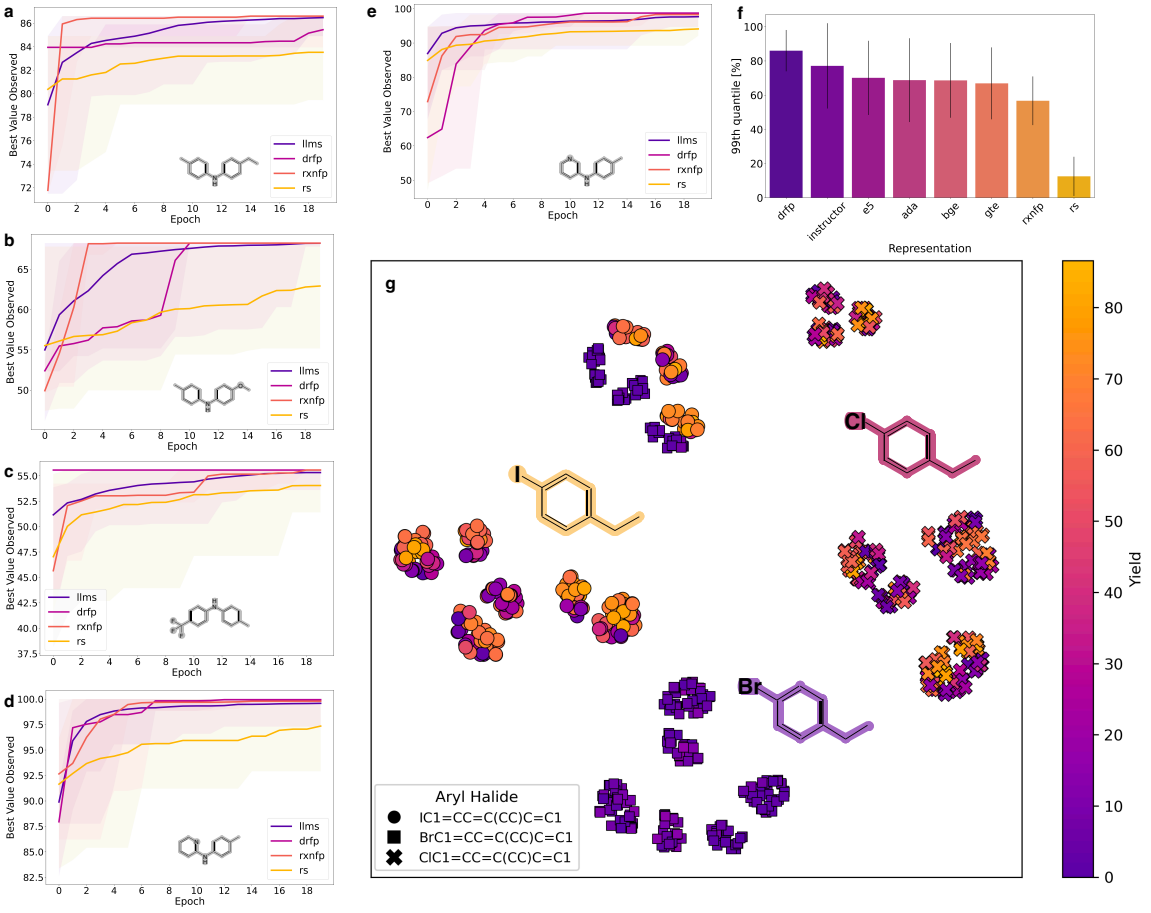

Figure 1: Visualization of Bayesian optimization (BO) paths and reaction representation performance. **(a-e)**, BO paths for different reaction representations over 20 iterations using the Kriging batch strategy, based on 20 repeated seed runs. Each line represents the trajectory of the maximum value found during optimization, while the shaded regions depict the standard deviation. Language modeling embeddings for reactions are averaged and presented as a single line for each reaction. **(f)** Performance of reaction representations, averaged across five reactions, showcasing the efficacy of each representation. Presented metric measures the percentage of the 99th quantile achieved during the optimization. **(g)** t-SNE visualization of the latent space derived from instructor embeddings. Data points are colored by yield, with shapes differentiating the reactant aryl halide used in the reaction.

Additionally, we validate their informativeness through a t-SNE latent space visualization. As shown in Figure 1 **(g)**, instructor embeddings are able to distinguish between the high-yield and low-yield reactions. The yield-based color coding displays evident clusters of reaction with similar outcomes. Diving deeper into the analysis of the latent space, we observe chemical significance of its particular organization. More specifically, we can see how different reactants (marked with different symbols in the plot) influence the yield. Beyond the yield differentiation, the arrangement of data points in the latent space offers additional insights. Specifically, the organization of the latent space pushes reactions with the same

aryl halide closer to each other. This layered insight reinforces the multidimensional content richness of LLM embeddings in capturing both semantic properties and chemical nuances.

4. Conclusion

In this work, we have extended the boundaries of LLM applications further into the realm of chemistry. We showcased that these models, even when deployed in an out-of-the-box manner, possess the capacity to map complex chemical reactions into a form suitable for Bayesian optimization. This outcome underscores the notion that LLMs, despite their generalist training, harbor a latent specialized knowledge that can be tapped into for various scientific directions.

The findings naturally lead to the question: If general-purpose embeddings are already so effective, what can be achieved through specialization through fine-tuning? Future work aims to incorporate these approaches within the Bayesian optimization framework. The objective is to evolve the embeddings during the optimization campaign and adapt to different chemical reactions on the fly.

Moreover, this work serves as a foundational step toward a unified computational framework designed to assist chemists across the complete research cycle—from the conceptualization of reaction procedures to their evaluation, vectorization, and subsequent optimization using BO. Additionally, the ability of LLMs to not only represent the chemical data well but also interact effectively with domain experts, sets the stage for a new era of machine-assisted chemical research that is both more efficient and accessible.

Acknowledgments

We would like to acknowledge support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

Appendix A.

Generating reaction procedures

The following reaction setup was used to generate reaction procedures. The template was populated with associated smiles for each compound for all data points. In comparison to the original reaction procedure extracted from the supplementary material in (Ahneman et al., 2018), this procedure keeps only the variable components of the procedure to ensure efficient mapping of the reaction search space. We found that including the general instructions that are same for all the data points damages the representations for BO purpose so we therefore restrict the information in procedures to variable components that describe the search space. Additional experimentation is needed to assess whether different encoding strategies could help alleviate this problem. For example, general (and mutual for all the data points in the search space) instructions could be mapped separately and added (in vector form) to the design space embeddings.

```
Reaction Setup

The following solutions were prepared in DMSO:

- Ligand: {ligand_smile}
- Aryl halide: {aryl_halide_smile}
- Additive: {additive_smile}
- Base: {base_smile}
```

In the future, more information could be added to this synthesis description, including reaction conditions and amounts. For the Buchwald-Hartwig dataset, the conditions and amounts were kept constant (Ahneman et al., 2018). The original procedure contains more fine-grained information.

Original procedure: Reaction Setup The following solutions were prepared in DMSO: catalyst (0.05 M), aryl halide (0.50 M), toluidine (0.50 M), additive (0.50 M), and base (0.75 M). These solutions were added to a 384-well source plate (80 µL per well). The Mosquito HTS liquid handling robot was used to dose each of these solutions (200 nL each) into a 1536-well plate. The plate was sealed and heated to 60 °C. After 16 h, the plate was opened and the Mosquito was used to add internal standard to each well (3 µL of 0.0025 M di-tert-butylbiphenyl solution in DMSO). At that point, aliquots were sampled into 384-well plates and analyzed by UPLC (Ahneman et al., 2018).

Large language models for embeddings

BGE embeddings

BGE embeddings (Xiao et al., 2023) are derived from the C-Pack suite, a comprehensive package aimed at advancing the field of general Chinese embeddings. It consists of three major components: a benchmarking suite (C-MTEB), a large text embedding dataset (C-MTP), and a family of embedding models (C-TEM), among which BGE is one. Their state-of-the-art performance on the MTEB benchmark with the subset of English models,

made them a valid choice for the BO framework. Topping the leader board scores and exceeding prior models by up to 10% at the time of release, attested to their effectiveness in various downstream tasks from classification to clustering. These results also suggested potential applicability in the domain of chemical reaction optimization.

The model behind these embeddings incorporates a diligent training procedure from pretraining, general purpose fine-tuning to task-specific fine-tuning. The models are pretrained on a massive corpus using a tailored algorithm designed to support the embedding task. The Wudao corpora (Yuan et al., 2021) served as the foundational dataset. The released English data, however, is reported to be twice the size of Chinese corpora. Following the pretraining procedure, the models underwent fine-tuning on C-MTP via contrastive learning, aimed at discriminating paired texts from negative samples. Further refinements were made using labeled data from C-MTP employing strategies such as instruction-based fine-tuning to help the model adapt to different tasks.

GTE embeddings

GTE (Li et al., 2023) is a versatile text embedding model trained with multi-stage contrastive learning. It capitalizes on advancements in unifying diverse NLP tasks and is trained over a variety of datasets. With a relatively modest parameter count of 110M, GTEbase not only outperforms OpenAI's black-box embedding API (Neelakantan et al., 2022) but also surpasses models with 10x larger parameters on key benchmarks.

The model employs a deep Transformer encoder, initialized with pretrained language models like BERT (Devlin et al., 2018). It uses a dual-encoder architecture with mean pooling. GTE is initially pretrained on approximately 800M text pairs from diverse sources, such as web pages, academic papers and code repositories. Fine-tuning is performed on smaller, annotated datasets, and incorporates both symmetric and asymmetric tasks. Additionaly, the authors applied data sampling and improved contrastive loss mechanisms for effective training.

E5 embeddings

E5 embeddings (Wang et al., 2022) emanate from a contrastive training approach with a large-scale dataset named CCPairs. These embeddings are designed for general-purpose tasks like retrieval, clustering, and classification. E5 outperforms models with 40x more parameters when fine-tuned on the MTEB benchmark. The model is pretrained on their own curated CCPairs constructed by combining various semistructured data sources such as CommunityQA, Common Crawl and Scientific paper. Additionally they use biencoder architecture with a pre-trained Transformer encoder for text embeddings. In-batch negatives are employed for contrastive loss. Further training is conducted on labeled data from diverse tasks. Moreover they apply hard negatives and knowledge distillation techniques and score embeddings using a cosine similarity scaled by a temperature parameter.

Instructor embeddings

INSTRUCTOR (Su et al., 2022) is a multi-task text embedding model trained with task-specific instructions. It is built on the GTR model family and designed to generate task-

and domain-specific embeddings without further training. INSTRUCTOR achieves state-of-the-art performance on 70 diverse datasets, improving the average score by 3.4% compared to previous best models.

It utilizes GTR models (Ni et al., 2021) as the backbone encoder. Embeddings are generated by concatenating input text and task instructions, followed by mean pooling. For the training objective, it uses a text-to-text problem formulation for various tasks, maximizing the similarity between positive pairs and minimizing it for negative pairs. Additionally, the authors apply bidirectional in-batch sampled loss. The models are trained on a curated MEDI dataset, which comprises 330 tasks annotated with human-written instructions.

Ada embeddings

Ada embeddings (Neelakantan et al., 2022) stem from contrastive pretraining on a large scale showcasing that this procedure can produce high-quality text and code embeddings. The model is trained unsupervised and achieves state-of-the-art results in various tasks. It is also the only closed-source model evaluated in this study.

The embeddings generate a relative improvement of 4% and 1.8% over the previous highest-scoring unsupervised and supervised models in linear-probe classification. In semantic search, they show a relative improvement of 23.4%, 14.7%, and 10.6% on MSMARCO, Natural Questions, and TriviaQA. The code embeddings improve by 20.8% on code search.

The training procedure involves a Transformer encoder to map text and code to vector representations. The authors employ special token delimiters for a more stable training together with contrastive learning with in-batch negatives. The model is initialized with other pretrained models to achieve optimal performance. The training is done on naturally occurring paired data for text and uses (text, code) pairs for code. It is noted, however, that large batch sizes are crucial for optimal performance which could hinder general applicability for groups without extensive computational resources.

Additional analysis of the latent space

Figure 2 shows the remaining components of the reaction, displayed using different shape markers in the scatter plot of t-SNE latent space. While it is difficult to uncover the structural properties of the latent space for different additives, the bases, similarly to aryl halide presented in the main part of the paper, show clear separation in the space, however still dispersed across. These properties are not as present for the latent organization of different ligands and additives. Regardless, the positioning of bases and reactants already provides relevant information in estimating how well the embeddings decipher important chemical information from the textual procedures.

Bayesian optimization framework

Gaussian process prior

The function $f(x)$ is modeled as a Gaussian Process (GP) with mean $\mu(x)$ and covariance $K(x, x')$.

$$f(x) \sim \mathcal{GP}(\mu(x), K(x, x'))$$

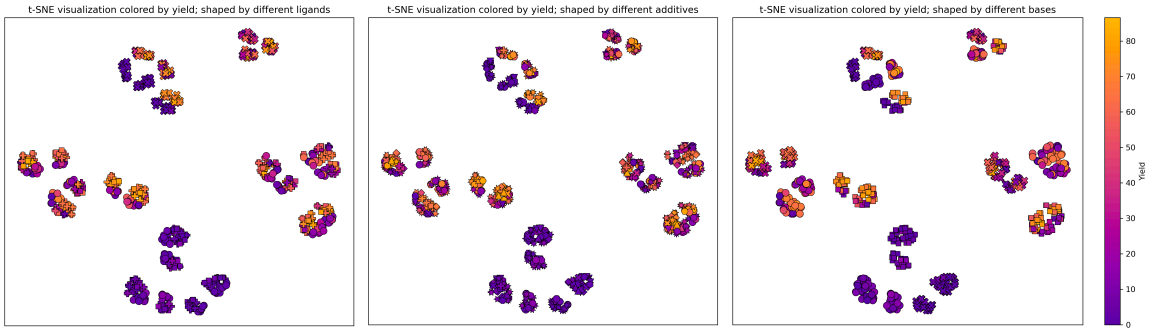

Figure 2: t-SNE visualized instructor embeddings, colored by yield. Each of the plots show different reaction components (ligand, additive, base), while the data points are represented with specific shape based on the unique values of the used reaction components.

Posterior update

After observing some data $D = \{(x_i, y_i)\}_{i=1}^{n}$, the posterior distribution of the function is updated:

$$f(x)|D \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$$

where $\mu_n(x)$ and $\sigma_n^2(x)$ can be computed using the kernel matrix K and the observed data.

Acquisition Function

The acquisition function $\alpha(x)$ guides the next sampling point.

$$x_{\text{next}} = \arg\max \alpha(x)$$

Popular acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB).

- **Expected Improvement (EI):**

$$\alpha_{\text{EI}}(x) = \mathbb{E}[f(x) - f(x_{\text{best}})|D]$$

- **Probability of Improvement (PI):**

$$\alpha_{\text{PI}}(x) = P(f(x) \geq f(x_{\text{best}}) + \epsilon)$$

- **Upper Confidence Bound (UCB):**

$$\alpha_{\text{UCB}}(x) = \mu_n(x) + \kappa\sigma_n(x)$$

Optimization loop

The BO loop iteratively updates the GP model and selects new points based on the acquisition function.

$$x_{\text{next}} = \arg\max \alpha(x|D)$$
$$D = D \cup \{(x_{\text{next}}, f(x_{\text{next}}))\}$$

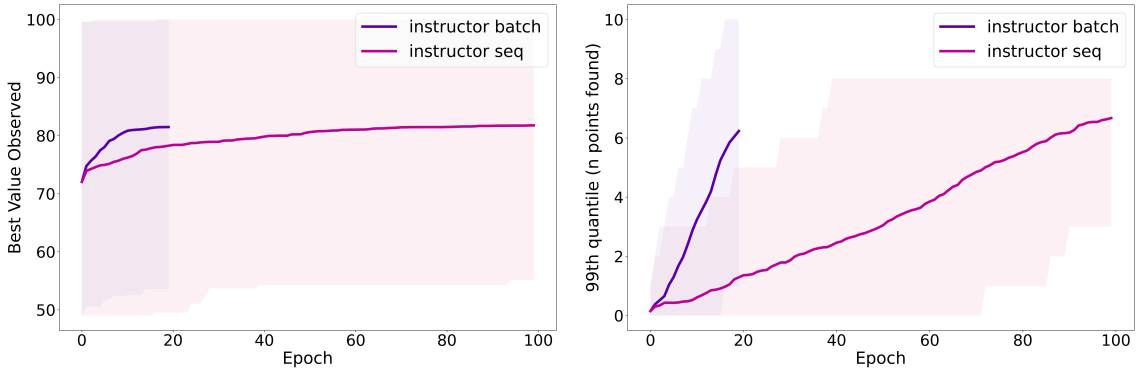

Figure 3: BO paths for instructor embeddings over five different reactions and 20 seed runs (averages across total of 100 different runs). Sequential optimization iterates for 100 steps, at each selecting only one point to evaluate. In contrast, with batch optimization we take a selection of five points at each step and run the whole optimization process shorter (20 iterations).

Batch strategy versus sequential optimization

In order to emulate the typical scenario evolving in a chemical lab, Bayesian optimization is often trained to propose batch of experiments instead of a single point. While this is a highly beneficial strategy for the laboratory, where parallel experiments could be run, therefore saving time and utilizing resources optimally, it can have negative impact on the Bayesian optimization surrogate models. Batching using Kriging believer strategy operates on predicted values as true evaluations and retrains the surrogate model with predictions until the batch is not filled. The model is eventually retrained with the addition of actual observation once the batch is evaluated. In our experiments, batching minimally degrades the optimization procedure observed by reaching similar outcomes with either strategy. On the contrary, batch optimization covers the more realistic chemical optimization scenarios and its facilitation is an important aspect when thinking of employing Bayesian optimization in the real world.

Averaged performance metrics across seed runs for different configurations

Repr.	Rxn	batch	Quant. 99 [%]	Quant. 95 [%]	Train/R^2	Val./R^2	Train/nlpd
ADA	1	1.0	0.78 ± 0.07	0.66 ± 0.02	0.97 ± 0.03	0.56 ± 0.04	2.45 ± 1.67
		5.0	0.78 ± 0.15	0.67 ± 0.05	0.98 ± 0.02	0.50 ± 0.08	1.24 ± 2.29
	2	1.0	0.99 ± 0.03	0.66 ± 0.03	0.84 ± 0.06	0.22 ± 0.53	3.52 ± 0.13
		5.0	0.87 ± 0.17	0.59 ± 0.10	0.85 ± 0.07	0.39 ± 0.12	3.24 ± 1.17
	3	1.0	0.72 ± 0.08	0.64 ± 0.05	0.83 ± 0.07	0.45 ± 0.05	3.15 ± 0.22
		5.0	0.74 ± 0.08	0.59 ± 0.05	0.92 ± 0.09	0.40 ± 0.08	1.05 ± 2.52
	4	1.0	0.56 ± 0.26	0.57 ± 0.09	0.89 ± 0.06	0.34 ± 0.11	2.88 ± 1.55

Continued on next page

Repr.	Rxn	batch	Quant. 99 [%]	Quant. 95 [%]	Train/R^2	Val./R^2	Train/nlpd
		5.0	0.47 ± 0.33	0.48 ± 0.12	0.96 ± 0.06	0.34 ± 0.10	0.17 ± 2.40
	5	1.0	0.47 ± 0.17	0.59 ± 0.06	0.93 ± 0.06	0.18 ± 0.09	3.40 ± 0.46
		5.0	0.58 ± 0.20	0.67 ± 0.08	0.96 ± 0.05	0.07 ± 0.13	2.65 ± 1.69
BGE	1	1.0	0.86 ± 0.22	0.58 ± 0.05	0.90 ± 0.10	0.22 ± 0.17	3.36 ± 0.32
		5.0	0.54 ± 0.16	0.47 ± 0.10	0.99 ± 0.02	0.23 ± 0.07	-0.99 ± 1.50
	2	1.0	1.00 ± 0.00	0.59 ± 0.02	0.87 ± 0.03	0.13 ± 0.10	3.35 ± 0.13
		5.0	0.96 ± 0.07	0.54 ± 0.06	1.00 ± 0.00	-0.41 ± 0.37	-1.58 ± 0.04
	3	1.0	0.47 ± 0.17	0.50 ± 0.07	0.86 ± 0.08	0.09 ± 0.23	2.90 ± 0.21
		5.0	0.55 ± 0.18	0.52 ± 0.06	0.87 ± 0.14	-0.04 ± 0.41	1.02 ± 2.60
	4	1.0	0.59 ± 0.07	0.52 ± 0.06	0.96 ± 0.05	0.29 ± 0.17	1.67 ± 2.18
		5.0	0.76 ± 0.20	0.62 ± 0.10	0.92 ± 0.05	0.32 ± 0.13	2.73 ± 1.20
	5	1.0	0.66 ± 0.06	0.70 ± 0.06	0.95 ± 0.05	-0.55 ± 0.44	1.41 ± 2.47
		5.0	0.62 ± 0.12	0.70 ± 0.14	0.95 ± 0.05	-0.03 ± 0.30	2.85 ± 1.44
DRFP	1	1.0	0.88 ± 0.05	0.71 ± 0.04	1.00 ± 0.00	0.44 ± 0.06	0.64 ± 1.87
		5.0	0.74 ± 0.09	0.70 ± 0.04	0.99 ± 0.06	0.18 ± 0.66	-1.20 ± 1.16
	2	1.0	1.00 ± 0.00	0.67 ± 0.03	0.99 ± 0.02	0.45 ± 0.09	0.56 ± 2.00
		5.0	0.99 ± 0.03	0.64 ± 0.03	0.98 ± 0.05	0.24 ± 0.53	1.80 ± 1.42
	3	1.0	0.88 ± 0.06	0.80 ± 0.03	0.91 ± 0.08	0.06 ± 1.00	2.53 ± 0.35
		5.0	0.86 ± 0.06	0.80 ± 0.02	0.98 ± 0.06	0.29 ± 0.63	-1.50 ± 1.81
	4	1.0	0.96 ± 0.09	0.82 ± 0.04	0.99 ± 0.03	0.33 ± 0.19	-0.96 ± 1.53
		5.0	0.86 ± 0.13	0.71 ± 0.06	0.98 ± 0.04	0.41 ± 0.20	-0.28 ± 1.95
	5	1.0	0.86 ± 0.08	0.74 ± 0.07	1.00 ± 0.00	0.03 ± 0.31	-1.26 ± 0.06
		5.0	0.84 ± 0.11	0.72 ± 0.06	1.00 ± 0.00	-0.08 ± 0.46	-1.25 ± 0.09
E5	1	1.0	0.66 ± 0.15	0.60 ± 0.07	0.93 ± 0.04	0.38 ± 0.29	3.19 ± 0.25
		5.0	0.68 ± 0.16	0.57 ± 0.08	0.95 ± 0.07	0.45 ± 0.11	2.15 ± 1.83
	2	1.0	0.93 ± 0.12	0.65 ± 0.06	0.79 ± 0.20	-0.37 ± 0.79	3.48 ± 0.41
		5.0	0.98 ± 0.06	0.65 ± 0.04	0.91 ± 0.16	0.06 ± 0.56	2.13 ± 1.98
	3	1.0	0.64 ± 0.11	0.60 ± 0.05	0.91 ± 0.04	0.21 ± 0.13	2.78 ± 0.28
		5.0	0.52 ± 0.12	0.55 ± 0.05	0.95 ± 0.08	0.19 ± 0.11	-0.08 ± 2.49
	4	1.0	0.63 ± 0.30	0.60 ± 0.09	0.85 ± 0.09	0.26 ± 0.17	3.42 ± 0.28
		5.0	0.66 ± 0.25	0.56 ± 0.07	0.90 ± 0.09	0.33 ± 0.12	2.78 ± 1.53
	5	1.0	0.64 ± 0.14	0.56 ± 0.09	0.98 ± 0.04	-0.44 ± 0.59	1.18 ± 2.20
		5.0	0.66 ± 0.12	0.56 ± 0.08	0.97 ± 0.03	-0.38 ± 0.43	1.80 ± 2.12
GTE	1	1.0	0.50 ± 0.11	0.53 ± 0.07	0.86 ± 0.19	0.31 ± 0.43	3.13 ± 1.41
		5.0	0.49 ± 0.17	0.52 ± 0.08	0.95 ± 0.05	0.39 ± 0.17	2.10 ± 2.06
	2	1.0	0.88 ± 0.25	0.58 ± 0.09	0.87 ± 0.12	-0.08 ± 0.77	3.12 ± 1.10
		5.0	0.90 ± 0.17	0.56 ± 0.08	0.92 ± 0.04	0.42 ± 0.08	2.80 ± 1.51
	3	1.0	0.65 ± 0.10	0.55 ± 0.11	0.84 ± 0.08	0.34 ± 0.23	3.07 ± 0.31
		5.0	0.63 ± 0.12	0.54 ± 0.11	0.85 ± 0.10	0.08 ± 0.53	2.15 ± 2.06
	4	1.0	0.72 ± 0.14	0.64 ± 0.04	0.89 ± 0.08	0.31 ± 0.20	3.40 ± 0.22

Continued on next page

Repr.	Rxn	batch	Quant. 99 [%]	Quant. 95 [%]	Train/R^2	Val./R^2	Train/nlpd
		5.0	0.71 ± 0.15	0.66 ± 0.07	0.90 ± 0.08	0.23 ± 0.22	2.62 ± 1.84
	5	1.0	0.61 ± 0.17	0.63 ± 0.06	0.97 ± 0.03	-0.48 ± 0.56	3.20 ± 0.32
		5.0	0.61 ± 0.19	0.62 ± 0.08	0.97 ± 0.03	-0.42 ± 0.56	2.72 ± 1.39
INS.	1	1.0	0.79 ± 0.10	0.76 ± 0.08	0.93 ± 0.04	0.21 ± 0.21	3.03 ± 1.08
		5.0	0.68 ± 0.21	0.59 ± 0.16	0.95 ± 0.03	0.17 ± 0.20	2.57 ± 1.73
	2	1.0	0.98 ± 0.06	0.60 ± 0.06	0.88 ± 0.06	0.30 ± 0.22	3.36 ± 0.26
		5.0	0.98 ± 0.17	0.59 ± 0.09	0.90 ± 0.07	0.21 ± 0.34	2.17 ± 2.25
	3	1.0	0.65 ± 0.12	0.69 ± 0.03	0.93 ± 0.07	-0.04 ± 0.51	0.92 ± 2.51
		5.0	0.64 ± 0.31	0.69 ± 0.15	0.89 ± 0.06	0.34 ± 0.08	2.45 ± 1.48
	4	1.0	0.89 ± 0.18	0.67 ± 0.10	0.92 ± 0.06	-0.30 ± 0.64	2.46 ± 1.78
		5.0	0.76 ± 0.17	0.66 ± 0.06	0.95 ± 0.05	0.36 ± 0.23	2.33 ± 1.65
	5	1.0	0.86 ± 0.17	0.73 ± 0.07	0.98 ± 0.03	-0.48 ± 0.40	0.61 ± 2.16
		5.0	0.81 ± 0.23	0.72 ± 0.07	0.99 ± 0.04	-0.19 ± 0.39	-0.50 ± 1.79
RS	1	1.0	0.17 ± 0.13	0.14 ± 0.05	/ ± /	/ ± /	/ ± /
		5.0	0.17 ± 0.13	0.14 ± 0.05	/ ± /	/ ± /	/ ± /
	2	5.0	0.09 ± 0.09	0.12 ± 0.04	/ ± /	/ ± /	/ ± /
	3	5.0	0.16 ± 0.11	0.14 ± 0.04	/ ± /	/ ± /	/ ± /
	4	5.0	0.09 ± 0.10	0.14 ± 0.06	/ ± /	/ ± /	/ ± /
	5	5.0	0.07 ± 0.09	0.12 ± 0.05	/ ± /	/ ± /	/ ± /
RXNFP	1	1.0	0.49 ± 0.04	0.45 ± 0.04	0.69 ± 0.06	0.49 ± 0.05	3.77 ± 0.10
		5.0	0.55 ± 0.08	0.43 ± 0.05	0.79 ± 0.09	0.52 ± 0.03	3.61 ± 0.15
	2	1.0	0.58 ± 0.13	0.45 ± 0.04	0.61 ± 0.07	0.45 ± 0.08	3.81 ± 0.08
		5.0	0.46 ± 0.12	0.41 ± 0.04	0.57 ± 0.07	0.38 ± 0.08	3.83 ± 0.09
	3	1.0	0.54 ± 0.06	0.41 ± 0.04	0.90 ± 0.05	0.32 ± 0.03	3.02 ± 0.12
		5.0	0.51 ± 0.14	0.42 ± 0.11	0.92 ± 0.05	0.25 ± 0.13	2.90 ± 0.22
	4	1.0	0.72 ± 0.13	0.58 ± 0.04	0.93 ± 0.05	0.23 ± 0.11	3.33 ± 0.35
		5.0	0.73 ± 0.12	0.58 ± 0.07	0.92 ± 0.04	0.19 ± 0.08	3.43 ± 0.25
	5	1.0	0.66 ± 0.12	0.67 ± 0.02	0.71 ± 0.08	0.26 ± 0.13	4.15 ± 0.15
		5.0	0.57 ± 0.09	0.75 ± 0.06	0.72 ± 0.10	0.27 ± 0.08	4.10 ± 0.24

Table 1: Summary of optimization performance metrics across different reaction representations and batch strategies: The table presents averaged metrics for Quantile 99% and 95% values, indicative of the highest-performing regions discovered during the optimization. Additionally, R^2 scores and negative log-probability density (NLPD) are reported. The training set is extended at each step of the optimization with evaluated data points. We report results at the end of the optimization routine with 100 new points selected during optimization, while the validation set comprises the remaining design space, slightly under 700 points. Higher R^2 and quantile percentages suggest better optimization efficacy and model generalization, while lower NLPD values indicate a more accurate and well-calibrated probabilistic model.

References

Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

Eric Anderson, Gilman D Veith, and David Weininger. *SMILES, a line notation and computerized interpreter for chemical structures*. US Environmental Protection Agency, Environmental Research Laboratory, 1987.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Mayk Caldas Ramos, Shane S Michtavy, Marc D Porosoff, and Andrew D White. Bayesian optimization of catalysts with in-context learning. *arXiv e-prints*, pages arXiv–2304, 2023.

Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 12(1): 1–15, 2020.

Kangway V Chuang and Michael J Keiser. Comment on "predicting reaction performance in c–n cross-coupling using machine learning". *Science*, 362(6416):eaat8603, 2018.

Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie International Edition*, 59(51): 22858–22893, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Natalie S Eyke, William H Green, and Klavs F Jensen. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, 5(10):1963–1972, 2020.

Kobi C Felton, Jan G Rittig, and Alexei A Lapkin. Summit: benchmarking machine learning methods for reaction optimisation. *Chemistry-Methods*, 1(2):116–122, 2021.

David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. Kriging is well-suited to parallelize optimization. In *Computational intelligence in expensive optimization problems*, pages 131–162. Springer, 2010.

Ryan-Rhys Griffiths. Applications of gaussian processes at extreme lengthscales: From molecules to black holes. *arXiv preprint arXiv:2303.14291*, 2023.

Jeff Guo, Bojana Ranković, and Philippe Schwaller. Bayesian optimization for chemical reactions. *CHIMIA*, 77(1/2):31, Feb. 2023. doi: 10.2533/chimia.2023.31. URL `https://www.chimia.ch/chimia/article/view/2023_31`.

Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. Automated chemical reaction extraction from scientific literature. *Journal of chemical information and modeling*, 62(9):2035–2045, 2021.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

Florian Häse, Matteo Aldeghi, Riley J Hickman, Loïc M Roch, Melodie Christensen, Elena Liles, Jason E Hein, and Alán Aspuru-Guzik. Olympus: a benchmarking framework for noisy optimization and experiment planning. *Machine Learning: Science and Technology*, 2(3):035021, 2021.

Riley Hickman, Jurgis Ruža, Loïc Roch, Hermann Tribukait, and Alberto García-Durán. Equipping data-driven experiment planning for self-driving laboratories with semantic memory: case studies of transfer learning in chemical reaction optimization. *ChemRxiv*, 2022.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2023a.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Is gpt-3 all you need for low-data discovery in chemistry? 2023b.

Kjell Jorner, Anna Tomberg, Christoph Bauer, Christian Sköld, and Per-Ola Norrby. Organic reactivity from mechanism to machine learning. *Nature Reviews Chemistry*, 5(4): 240–255, 2021.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.

Toshiharu Morishita and Hiromasa Kaneko. Initial sample selection in bayesian optimization for combinatorial optimization of chemical compounds. *ACS omega*, 8(2):2001–2009, 2022.

Pia Müller, Adam D Clayton, Jamie Manson, Samuel Riley, Oliver S May, Norman Govan, Stuart Notman, Steven V Ley, Thomas W Chamberlain, and Richard A Bourne. Automated multi-objective reaction optimisation: which algorithm should i use? *Reaction Chemistry & Engineering*, 7(4):987–993, 2022.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

Alexander Pomberger, AA Pedrina McCarthy, Ahmad Khan, Simon Sung, CJ Taylor, MJ Gaunt, Lucy Colwell, David Walz, and AA Lapkin. The effect of chemical representation on active machine learning towards closed-loop optimization. *Reaction Chemistry & Engineering*, 2022.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discovery*, 1(2): 91–97, 2022.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.

Bojana Ranković, Ryan-Rhys Griffiths, Henry B. Moss, and Philippe Schwaller. Bayesian optimisation for additive screening and yield improvements in chemical reactions – beyond one-hot encoding. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2022-nll2j-v3.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390, 2020.

Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of chemical information and modeling*, 55(1):39–53, 2015.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, 2021.

Philippe Schwaller, Alain C Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, page e1604, 2022.

Artur M Schweidtmann, Adam D Clayton, Nicholas Holmes, Eric Bradford, Richard A Bourne, and Alexei A Lapkin. Machine learning meets continuous flow chemistry: Automated optimization towards the pareto front of multiple objectives. *Chemical Engineering Journal*, 352:277–282, 2018.

Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

Connor J Taylor, Kobi C Felton, Daniel Wigh, Mohammed I Jeraal, Rachel Grainger, Gianni Chessari, Christopher N Johnson, and Alexei A Lapkin. Accelerated chemical reaction optimization using multi-task learning. *ACS Central Science*, 2023.

Jose Antonio Garrido Torres, Sii Hong Lau, Pranay Anchuri, Jason M Stevens, Jose E Tabora, Jun Li, Alina Borovika, Ryan P Adams, and Abigail G Doyle. A multi-objective active learning platform and web app for reaction optimization. *Journal of the American Chemical Society*, 144(43):19999–20007, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):3601, 2020.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436, 2019.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

Andrew D White. The future of chemistry is language. *Nature Reviews Chemistry*, pages 1–2, 2023.

Daniel S Wigh, Matthieu Tissot, Patrick Pasau, Jonathan M Goodman, and Alexei A Lapkin. Quantitative in silico prediction of the rate of protodeboronation by a mechanistic density functional theory-aided algorithm. *The Journal of Physical Chemistry A*, 127(11): 2628–2636, 2023.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68, 2021.