Scaling High-Throughput Experimentation Unlocks Robust Reaction-Outcome Prediction

Michał Sadowski Molecule.one
Lukasz Sztukiewicz Molecule.one
Maria Wyrzykowska Molecule.one
Tadija Radusinović Molecule.one
Piotr Byrski Molecule.one
Paweł Włodarczyk-Pruszyński Molecule.one
Bartosz Matysiak Molecule.one
Jan Kulczycki Molecule.one
Filip Ulatowski Molecule.one
Ruard van Workum Molecule.one
Pawel Dabrowski-Tumanski Molecule.one
Paulina Wach Molecule.one
Filip Chmielewski Molecule.one
Jan Rzymkowski Molecule.one
Mateusz Bruno-Kamiński Molecule.one
Jan Busz Molecule.one
Artur Chołuj Molecule.one
Mateja Duda Molecule.one
Tomasz Dybowski Molecule.one
Marco Farinone Molecule.one
Tomasz Jeliński Molecule.one
Alicja Karczewska Molecule.one
Paweł Kowalczyk Molecule.one
Marek Pietrzak Molecule.one
Łukasz Szczupak Molecule.one
Aleksander Szkółka Molecule.one
Grzegorz Wojciechowski Molecule.one
Stanislaw Kamil Jastrzebski Molecule.one

Abstract

Organic chemistry underpins small-molecule drug discovery, yet—unlike structural biology—it lacks large, unbiased datasets for training broadly generalizable models. We report the largest microliter-scale high-throughput experimentation (HTE) campaign to date: $200{,}000$ reactions spanning three workhorse classes (Amide Coupling, Suzuki Coupling, Buchwald–Hartwig Coupling) involving $30{,}000$ products—over $4\times$ larger than the largest publicly disclosed dataset to date. This scale and diversity enable reaction-outcome predictors that generalize to unseen substrates. We introduce UniReact, a molecule-attention Transformer built on pretrained molecular encoders. Across the three reaction classes, our models achieve PR-AUC 2–3$\times$ over random and ROC-AUC in the 70–86% range. We further establish scaling laws for reaction-outcome prediction spanning three orders of magnitude of HTE data, and for one class up to $100{,}000$ reactions—*to our knowledge*, the broadest HTE scaling study to date. In a human study on Suzuki coupling prioritization, our models outperform PhD-level chemists (precision 87.1% at 50% recall vs. 60.8%). Finally, we show the first, to our best knowledge, demonstration of zero-shot transfer to an external HTE dataset. Taken together, these results support scaled HTE as a viable path to broadly applicable predictors of chemical reactivity that surpass human intuition and ultimately help discover novel chemistry.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

1 Introduction

The long-term ambition of synthetic chemistry is universal synthesis—the ability to make any physically realizable molecule. Unlocking broader chemical space requires two advances: discovering new reactions and developing robust models for synthesis planning and reaction-outcome prediction. Today, however, drug discovery remains constrained to compounds that are easy to synthesize [Blakemore et al., 2018].

The discovery of the Nobel Prize-winning Suzuki coupling in the 1980s reshaped medicinal chemistry. Drug hunters were able to form carbon–carbon bonds between sp^2 carbons. This invention plausibly contributed to the proliferation of small-molecule drugs rich in such bonds after the 1980s [Leeson et al., 2021].

Despite our mastery of organic chemistry, humans have relatively limited accuracy in predicting the outcomes of chemical reactions. This is evidenced by the high failure rate of human-executed experiments, reaching up to 40% [Buitrago Santanilla et al., 2015, Raghavan et al., 2024].[1] In many cases, this is due to issues beyond intrinsic reactivity, such as substrate instability, workup effects on the product, poor solubility, or unforeseen side reactions. Chemists routinely troubleshoot such situations [Frontier, 2025].

Limited predictive power is a pressing issue. Automation remains limited, and many small molecules in early-stage drug discovery are synthesized in countries with lower labor costs. This manifest in the fact that organic synthesis accounts for roughly 40% of the cost of discovering a drug and is a significant contributor to long delays [Paul et al., 2010].

High-throughput experimentation (HTE) is a natural way to generate large, relatively unbiased reaction datasets, unlocking both the discovery of novel chemistry and the training of robust predictive models. In other fields, major AI advances have closely followed the availability of large-scale datasets—for example, the Protein Data Bank enabled breakthroughs in structure prediction such as AlphaFold [PDB, 2022, 2025, Jumper et al., 2021]; Internet-scale corpora unlocked few-shot language models [Brown et al., 2020]; and massive labeled image collections made possible Transformer-based vision systems [Dosovitskiy et al., 2021]. Chemistry lacks an equivalent resource.

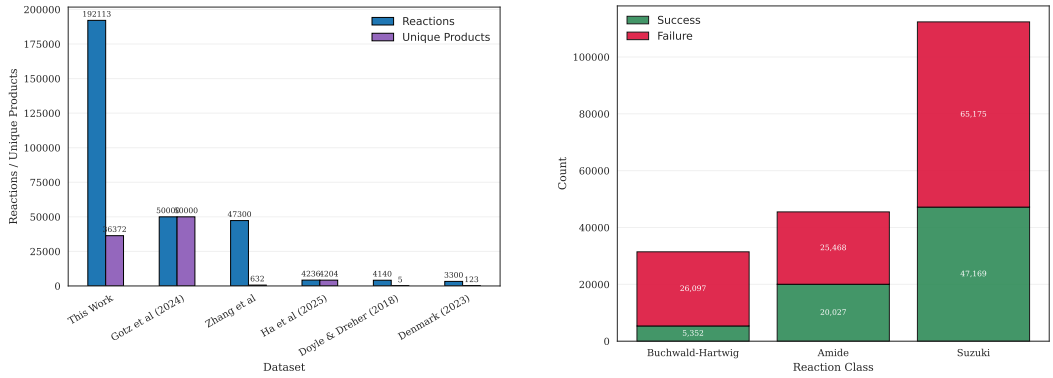
Most chemical HTE campaigns to date have focused on a narrow *product scope*, optimizing yields for a small number of products by varying conditions (e.g., temperature, time) [Shevlin, 2017, Mennen et al., 2019, Krska et al., 2017]. Such *targeted* designs provide limited data for learning about the broader chemical space, which is vast—the number of drug-like molecules exceeds 10^{60}. See also Figure 1.

We report the largest microliter-scale reaction campaign to date, spanning three key medicinal chemistry reaction classes: Amide Coupling, Suzuki Coupling, and Buchwald–Hartwig Coupling. These classes account for approximately 56.8% of reported large-scale syntheses [Brown and Boström, 2016]. The dataset comprises $200,000$ microliter-scale reactions, $> 1,000$ unique substrates, and $30,000$ unique products.

This chemical diversity and scale enable training of robust, superhuman models that predict outcomes on unseen substrates. Our main contributions are:

1. We show robust generalization to unseen building blocks on the largest HTE dataset across three workhorse reaction classes. We achieve PR-AUC 2–3$\times$ over a random baseline and ROC-AUC in the 70–86% range. We also present the first scaling laws for reaction-outcome prediction across three orders of magnitude of microliter-scale HTE data, with smooth power-law trends.
2. We show that our models outperform PhD-level synthetic chemists on Suzuki coupling prioritization (precision 87.1% at 50% recall vs. 60.8% for humans).
3. We introduce UniReact: a molecular-attention Transformer that surpasses a strong graph-based model on the largest subset of our dataset and exhibits complementary inductive biases that improve performance when ensembled.

[1] This may stem from limited feedback: unlike domains like chess, chemists perform only thousands of reactions in a lifetime, with most learning happening *offline* from textbooks, papers, and colleagues. These sources rarely report failed experiments, and the experiments performed are highly biased toward successes and chemist intuition.



(a) Comparison of the number of reactions the number of unique products with selected recent HTE datasets.

(b) The number of successful (green) and failed (red) reactions in each of the three reaction classes.

Figure 1: Summary of our microliter-scale HTE dataset.

4. We show the first, to our best knowledge, demonstration of zero-shot generalization to an external HTE dataset. We show a model trained on our subset of Amide Couplings achieves 70% ROC-AUC on the dataset from [Zhang et al., 2025].

2 Related work

Published reaction databases (textbooks, papers, patents, etc.) are heavily biased toward successful outcomes and seldom report negative or low-yield reactions. For example, Angello et al. [2022] attempted to mine the literature for general Suzuki–Miyaura conditions and explicitly noted that their ML approach "failed" in part because of "a lack of published (or otherwise accessibly archived) negative results." Saebi et al. [2023] likewise emphasize that the "lack of publicly available, large, and unbiased datasets" is a key roadblock for ML in chemistry. Efforts like the Open Reaction Database (ORD) [Kearnes et al., 2021] aim to standardize and share reaction data, but existing large collections (CAS, Reaxys, USPTO, commercial patent databases, etc.) often contain the same literature-derived chemistry. Indeed, King-Smith et al. [2024] note that datasets such as CAS, Reaxys, USPTO and even the ORD have "a high level of overlap" with published reactions, rendering their internal "reactomes" largely indistinguishable from the literature's. These observations underscore the need for new experimental data sources (especially including failed experiments) to train robust predictive models.

High-throughput experimentation (HTE) campaigns provide one such source. Chemical high-throughput experimentation (HTE) has evolved from early plate-based condition screens to a routine tool in pharmaceutical process and medicinal chemistry [Shevlin, 2017, Mennen et al., 2019, Krska et al., 2017]. Historically, most HTE campaigns have focused on optimizing reaction conditions, often leveraging Bayesian Optimization to sequentially prioritize experiments [Shields et al., 2021]. This focus is driven by the practical need to maximize yield in synthesis, from small to large scale. As a result, HTE has become a standard technique in commercial laboratories.

In contrast, relatively few studies have conducted HTE campaigns that target broad regions of chemical space. For instance, a recent study sought to identify a set of conditions effective across a wide range of products [Angello et al., 2022], but tested fewer than 100 unique products—limiting the generalizability of any models trained on this data. Another effort [King-Smith et al., 2024] reported $39,000$ chemical reactions spanning various reaction classes, yet included only 290 unique products. The largest HTE study reported to date is a $50,000$-reaction screen of the Ugi (3-component) reaction [Götz et al., 2025], which involved 171 building blocks, but was performed under a single set of conditions. Key datasets are summarized in Figure 1. [Zhang et al., 2025]

Machine learning for reaction outcomes has made rapid progress, but is indeed largely limited by data. Early work (e.g., Ahneman et al. [2018a]) showed that models like random forests or simple neural nets could predict yields for narrowly defined coupling reactions. More recently, message-passing graph neural networks have become popular (e.g., the Chemprop framework [Yang et al.,

2019]) for chemical property prediction, including reaction success. However, these models typically assume relatively small, domain-specific datasets and often fail to generalize beyond their training chemistries.

HTE coupled with automation and AI is also enabling autonomous discovery. Mahjour et al. [2024] proposed new multicomponent reactions via an automated workflow and confirmed two by robotic parallel experiments. Angello et al. [2022] used a closed-loop robotic system guided by ML to identify general Suzuki conditions. In contrast, purely theoretical approaches (e.g., quantum calculations) can illuminate mechanisms and catalyst design but are too resource-intensive for broad screening [Hayashi et al., 2023]. Overall, unbiased, high-throughput experimental data will be essential for training ML models that surpass human intuition.

3 Methods

3.1 High-throughput Experimentation

We begin by outlining our high-throughput experimentation (HTE) program.

While fields like structural biology and AI have advanced rapidly thanks to large, high-quality datasets such as the Protein Data Bank PDB [2022, 2025] and the Internet, chemistry still lacks a comparable resource. We see HTE as the key to building such a dataset, but it must be specifically designed to support broad, generalizable applications.

Our goal is to develop models with a broad, generalizable understanding of chemistry that can be readily fine-tuned for diverse downstream applications. To this end, our approach differs from much of the prior literature in several key ways:

1. We prioritize a wide diversity of substrates, while keeping the number of screened condition sets small (4–10);
2. We aim for $10\times$ to $100\times$ larger number of unique products than most previous studies;
3. We aim for semi-quantitative yield estimation using proprietary analytical software.

We conduct reactions at the microliter scale and millimolar concentrations. This scale offers a practical compromise: it is small enough for high-throughput, yet large enough to ensure reliable, high-quality data. Reagents are prepared as stock solutions in DMSO, with solubility checked by hand. Reactions are set up on 96-well plates using Opentrons pipetting robots. After reformatting, quenching, and workup, we analyze products by LC/MS, using autosampling from 384-well plates.

At the core of our workflow is proprietary software for processing analytical chemistry data. Unlike previous approaches, we use spectra curated by analytical chemists to train our software. This enables more accurate peak assignment and integration, producing a robust yield estimator. For semi-quantitative yield assessment, we calibrate the method on a held-out set of product standards.

The entire process is orchestrated by software. A centralized metadata store acts as the source of truth for both models and chemists. Analytical results are processed automatically and saved to cloud storage. These automation steps are crucial—they minimize human error and keep the operation running quickly.

3.2 UniReact: a model for scaled HTE

Models with minimal inductive bias, such as the Vision Transformer [Dosovitskiy et al., 2021], often outperform specialized architectures on large datasets. However, HTE datasets have historically favored models with stronger domain-specific assumptions [Ahneman et al., 2018b, Shields et al., 2021, Saebi et al., 2023].

Motivated by the scale of our dataset, we introduce UniReact: it embeds substrates and products using pretrained UniMolV2 [Zhou et al., 2024], processes each molecule with the Relative Molecule Attention Transformer (RMAT) [Maziarka et al., 2024], and aggregates per-compound representations into a reaction embedding. Figure 2 summarizes the architecture.

Let N_i denote the number of atoms in the ith molecule. Each UniMolV2 layer l operates on an atomic representation $\mathbf{x}^l \in \mathbb{R}^{N_i \times d_a}$ and a pair representation $\mathbf{p}^l \in \mathbb{R}^{N_i \times N_i \times d_p}$. Following [Zhou

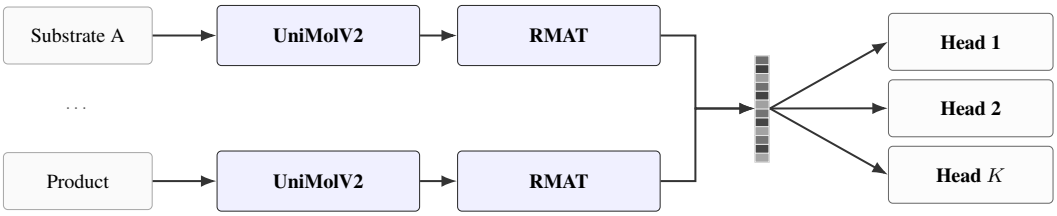


Figure 2: **UniReact architecture.** Each molecule is processed by pretrained UniMolV2 encoders to produce atom and pair embeddings. Relative Molecule Attention Transformer (RMAT) operates per molecule to yield a per-compound embedding; per-compound embeddings are averaged to obtain the reaction embedding $\mathbf{h}_{\text{react}}$, which feeds K prediction heads.

et al., 2024], we compute the initial $\mathbf{x}^0$ and $\mathbf{p}^0$ using RDKit and graph features and the molecular conformation.

Updating the pair representation scales computationally as $O(N_i^2 d_p)$. To alleviate this computational cost, we only keep the first k layers of the encoder and process the molecular representation further with a computationally cheaper Relative Molecule Attention Transformer (RMAT).

RMAT extends standard Transformer self-attention by incorporating molecular graph information directly into the attention computation. For a molecule with atom feature matrix $\mathbf{X}$ and pairwise features $\mathbf{p}^l$, the attention mechanism is further biased as follows:

$$\mathcal{A}(\mathbf{X}, \mathbf{p}^l) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + f_{\text{dist}}(\mathbf{D}) + f_{\text{adj}}(\mathbf{A}) + f_{\text{pair}}(\mathbf{p}^l)\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$ are the usual query, key, and value projections; $\mathbf{D}$ encodes pairwise atomic distances, $\mathbf{A}$ encodes adjacency (bond) information, and f_{dist}, f_{adj}, and f_{pair} are learned or fixed functions mapping these features to attention biases. The term $f_{\text{pair}}(\mathbf{p}^l)$ specifically injects the pairwise representations from UniMolV2 as an additional bias to the attention logits.

Finally, the representation of the first token (atom) is averaged across input molecules. The final classification is made by a separate MLP head for each set of conditions.

By default, we use 2D conformers to represent molecules to skip the costly 3D embedding procedure, which improves throughput without degrading performance empirically. We hypothesize that the pretrained UniMolV2 has the capability to quickly learn to update the conformation.

4 Experiments

Our primary objective is to develop models that can assist in planning syntheses during early-stage drug discovery. This goal shapes several key choices in our evaluation strategy.

We evaluate models on reactions where both substrates and the product are held out from training. This scenario closely mirrors real-world synthesis, where chemists typically encounter novel products and substrates due to the vastness of chemical space.

Because our focus is on early discovery, we consider the practical context: new structures are experimentally validated using only small amounts of product (typically milligrams). Therefore, achieving even modest yields is sufficient and cost-effective. In our experiments, we set a 5% yield threshold and frame the task as a classification problem.

For evaluation, we use the Precision Area Under the Curve (PR-AUC) as our primary metric. PR-AUC quantifies the average precision when reactions are prioritized by the model. For reference, a random baseline achieves a PR-AUC equal to the proportion of positive reactions in the dataset.

Our central claim is that scaling high-throughput experimentation enables the development of robust models for reaction outcome prediction. This principle underpins the design of our experiments.

Method	Suzuki		Amide		Buchwald-Hartwig	
	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
Random	13% ± 1%	50%	43% ± 4%	50%	19% ± 5%	50%
Chemprop	48.8% ± 5%	84.9% ± 3%	69.8% ± 8%	75.3% ± 2%	35% ± 14%	66% ± 5%
UniReact	52.8% ± 2%	**86.2%** ± 2%	69.5% ± 13%	74.5% ± 14%	34% ± 12%	66% ± 7%
Ensemble	**53.8%** ± 10%	86.0% ± 2%	**70.8%** ± 8%	**75.9%** ± 2%	**36%** ± 14%	**68%** ± 7%

Table 1: Performance comparison of methods on three reaction datasets, with error bars indicating standard deviation across 3 runs. Random baseline shows expected performance for random predictions. Best results for each column are in bold. Ensemble combines Chemprop and UniReact models with all hyperparameter configurations.

4.1 Robust generalization to unseen substrates

We compare UniReact to Chemprop, a widely used graph-based model [Heid et al., 2023]. For each of the three reaction classes, we evaluate models for predicting reaction outcomes for both novel substrates and products.

Unless noted otherwise, we use the UniMolV2-84M pretrained encoder. RMAT is configured with 2 layers and 8 attention heads (dropout 0.1), with hidden size tied to the UniMol embedding.

For UniReact, we train with learning rate in $\{1.2 \times 10^{-5}, 2.5 \times 10^{-5}\}$ and the number of compound encoder layers in $\{2, 3, 4\}$. For Chemprop, we train with hidden sizes in $\{250, 500\}$, depths in $\{2, 3\}$, and learning rates in $\{2 \times 10^{-4}, 5 \times 10^{-3}\}$, testing 12 hyperparameter combinations for Chemprop and 6 for UniReact. To evaluate generalization to unseen substrates, we exclude 40 boronic acids and 40 halides from the training set; the validation split remains random. We evaluate an ensemble of models trained with different hyperparameters, which we observe to achieve better performance on the out-of-distribution test set than tuning hyperparameters based on the validation set. All models use early stopping based on the ROC-AUC metric on the validation set.

Table 1 summarizes the results for all three reaction classes.

Both models demonstrate strong generalization to unseen building blocks, achieving PR-AUC scores that are 2–4$\times$ higher than the random baseline.

On the largest dataset (Suzuki coupling, $N \approx 100{,}000$ reactions), UniReact achieves a PR-AUC of $52.8\% \pm 2\%$ and ROC-AUC of $86.2\% \pm 2\%$, outperforming Chemprop (PR-AUC $48.8\% \pm 5\%$, ROC-AUC $84.9\% \pm 3\%$). This result supports our hypothesis that more expressive models excel as dataset size increases. To our knowledge, this is the first demonstration of a Transformer-based model surpassing a graph-based model on a high-throughput reaction dataset with unseen substrates.

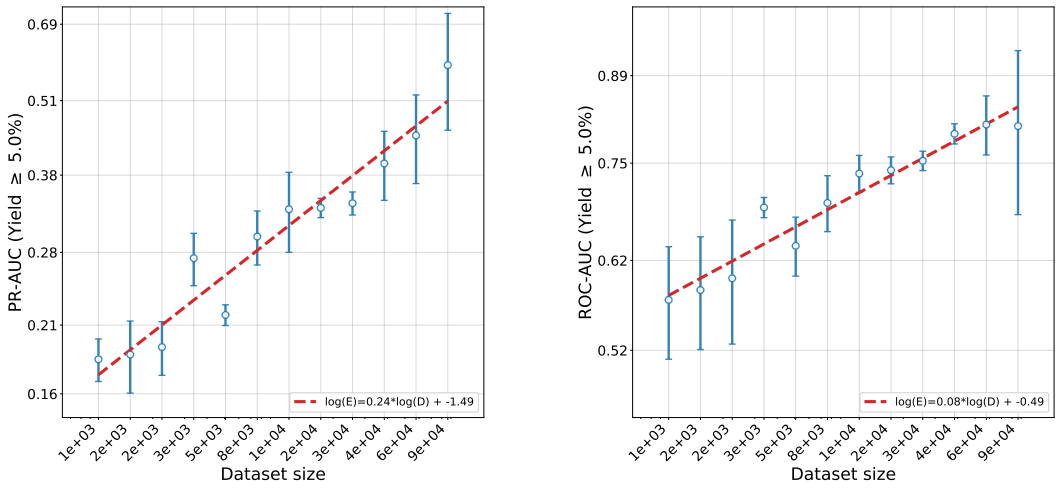
On the Buchwald-Hartwig coupling dataset ($N \approx 30{,}000$ reactions), Chemprop achieves a PR-AUC of $35\% \pm 14\%$ and ROC-AUC of $66\% \pm 5\%$, while UniReact achieves a PR-AUC of $34\% \pm 12\%$ and ROC-AUC of $66\% \pm 7\%$. For the amide coupling dataset ($N \approx 45{,}000$ reactions), both models perform similarly: UniReact achieves a PR-AUC of $69.5\% \pm 13\%$ and ROC-AUC of $74.5\% \pm 14\%$, while Chemprop achieves a PR-AUC of $69.8\% \pm 8\%$ and ROC-AUC of $75.3\% \pm 2\%$.

We hypothesize that UniReact and Chemprop exhibit complementary inductive biases. Motivated by this, we also compare the performance of UniReact to an ensemble of Chemprop and UniReact. We average predictions of all models with all hyperparameter configurations. We observe that the ensemble outperforms both models across all datasets.

4.2 Scaling laws for reaction outcome prediction across three orders of magnitude

The large scale of our datasets raises a natural research question: does scaling to the order of $100,000$ individual reactions improve generalization and robustness?

To investigate this, we trained UniReact on the Suzuki coupling subset of our dataset, varying training set sizes from approximately $1,000$ to $100,000$ reactions. For each training configuration, we optimized the learning rate from the set $\{1 \times 10^{-4}, 2 \times 10^{-4}, 4 \times 10^{-4}\}$. Performance was averaged over 4 random seeds with repeated train-test splits. We evaluated on reactions with both unseen substrates and unseen products. The results are summarized in Fig. 3.



(a) PR-AUC vs. dataset size.

(b) ROC-AUC vs. dataset size.

Figure 3: **Scaling laws on Suzuki coupling across three orders of magnitude of dataset size.** Performance improves smoothly with data scale across three orders of magnitude. Points show mean with $\pm$ one s.d. error bars; red dashed lines are power-law fits in log–log space.

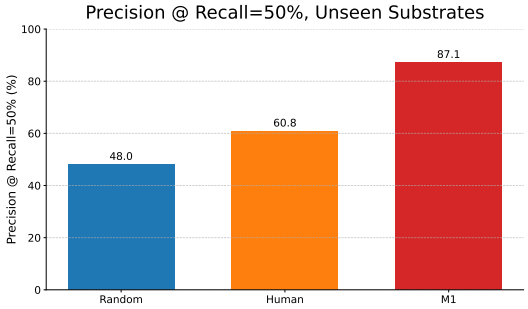


Figure 4: **Superhuman performance at discriminating failures vs. successful Suzuki coupling reactions.** Chemists are routinely asked to prioritize reactions based on their likelihood of success, e.g., when preparing quotations for a synthesis. We compare precision at 50% recall for a random baseline, the average prediction of three organic chemists, and Chemprop.

Performance as measured by PR-AUC increased from 20% to 50%, representing a $2.5\times$ improvement in precision across different recall values. ROC-AUC increased from 57% to 81%.

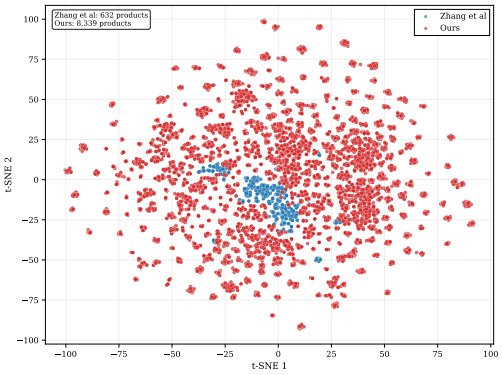
4.3 Superhuman performance in classifying Suzuki coupling reactions

Next, we evaluated our models against human experts. Three PhD-level synthetic chemists were asked to classify Suzuki coupling reactions as achieving 10% or higher LC/MS yield. Their predictions were averaged across all participants.

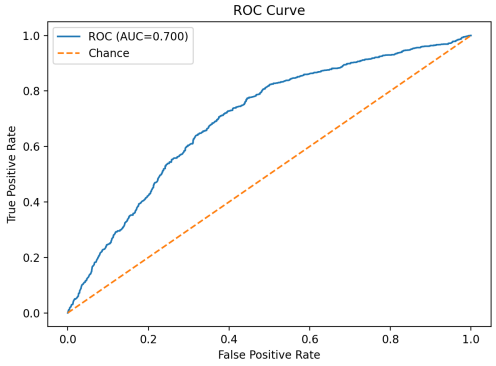
We evaluated a set of 100 Suzuki coupling reactions, randomly sampled from our dataset and balanced to ensure approximately 50% were successful.

We use precision at 50% recall as our metric. Chemists are routinely asked to prioritize reactions based on their likelihood of success, e.g., when preparing quotations for a synthesis. This metric reflects that scenario by measuring the ability to prioritize reactions more likely to succeed when allowed to reject 50% of reactions.

Figure 4 shows the results. Chemprop outperforms the chemists, achieving 87.1% precision compared to 60.8% for humans, representing a $1.7\times$ improvement over the random baseline.



(a) t-SNE of product chemical space. Blue: [Zhang et al., 2025]; red: Ours.



(b) Zero-shot ROC on the external dataset.

Figure 5: **Zero-shot transfer to an external HTE dataset.** Left: products from Zhang et al. [2025] occupy a region largely contained within the product manifold from our HTE campaign. Right: zero-shot performance of a model trained on our amide-coupling subset on the external dataset.

4.4 Zero-shot transfer to an external HTE dataset

Finally, we tested the transferability of our models to an external HTE dataset. We used the dataset from [Zhang et al., 2025], which contains $47,000$ amide coupling reactions. The dataset contains over 90 different condition sets, which exceeds the number of condition sets in our dataset. We performed closest matching based on similarity of conditions.

We trained Chemprop on our amide-coupling subset and evaluated it zero-shot on the external dataset. Figure 5 summarizes the results: the external products largely lie within the manifold spanned by our products, and the model achieves ROC-AUC of 70% without any fine-tuning.

5 Limitations

The scale and design of our dataset introduce several important limitations that should be considered when applying our data and models. Here, we briefly outline the key limitations and discuss their potential impact on model applicability.

First, our yield estimates are not based on product standards, as synthesizing standards at this scale is not practical. Consequently, products with unusual absorbance profiles may be misclassified. While our yield estimation is reliable for distinguishing failures from successes at low yield thresholds, it is not suitable for precise quantitative yield prediction.

Second, our reactions are performed under conditions that differ from those used in larger-scale synthesis, most notably at much lower concentrations (typically at least 10 times lower) and with less efficient mixing. This leads to lower success rates for many reaction classes.

Despite these and other limitations, the dataset remains highly valuable. As with large language models pretrained on Internet data, we expect most downstream applications to benefit from additional fine-tuning. To further test the transferability of our models, we have also demonstrated that our models can transfer to external data sources.

6 Conclusions

We scaled microliter high-throughput experimentation to $200{,}000$ reactions across three workhorse reaction classes with emphasis on product diversity. This breadth enables models that robustly generalize to unseen substrates and products, with 2–$3\times$ gains in PR-AUC over random and ROC-AUC in the 0.7–0.85 range. We report the first scaling laws for reaction-outcome prediction spanning three orders of magnitude in data, and we demonstrate superhuman prioritization on Suzuki (precision

87.1% at 50% recall vs. 60.8% for PhD chemists). We also show the first demonstration of zero-shot transfer to an external HTE dataset.

We also introduced a molecular-attention Transformer that surpasses a graph-based model on the largest subset of the dataset and shows complementary inductive biases to the graph-based model.

These results support our thesis: scaling unbiased HTE is a practical path to robust reaction-outcome prediction—enabling models that exceed human intuition about existing chemistry and ultimately help discover novel chemical reactions.

Looking ahead, our main priorities are: (i) applying our methodology to a rarely used but promising reaction class; (ii) scaling up by another order of magnitude; (iii) continually improving dataset quality, particularly by refining yield estimation; and (iv) automatically extracting new chemical knowledge from the dataset, such as understanding side-product reactivity and the relationships between structure, conditions, and reactivity.

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, 2018a. doi: 10.1126/science.aar5169.

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, 2018b. doi: 10.1126/science.aar5169.

Nicholas H. Angello, Nathan S. Eyke, Wenhao Cui, Andrew A. Wankowicz, David Caramelli, Abigail G. Doyle, and Klavs F. Jensen. Closed-loop optimization of general reaction conditions for heteroaryl suzuki–miyaura coupling. *Science*, 378(6618):399–405, 2022. doi: 10.1126/science.adc8743.

David C. Blakemore, Ian Churcher, and David C. Rees. The importance of synthetic chemistry in the pharmaceutical industry. *Science*, 2018. doi: 10.1126/science.aat0805. URL `https://www.science.org/doi/abs/10.1126/science.aat0805`.

David G. Brown and Jonas Boström. Analysis of past and present synthetic methodologies on medicinal chemistry: Where have all the new reactions gone? *Journal of Medicinal Chemistry*, 59 (10):4443–4458, 2016. doi: 10.1021/acs.jmedchem.5b01409. URL `https://pubs.acs.org/doi/10.1021/acs.jmedchem.5b01409`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Alyssa Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCann, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. URL `https://arxiv.org/abs/2005.14165`.

Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Organic chemistry. nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53, 2015. doi: 10.1126/science.1259203.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Alison Frontier. Not voodoo: Demystifying synthetic organic chemistry, 2025. URL `https://www.chem.rochester.edu/notvoodoo/pages/how_to.php?page=experiment`. Accessed: 2025.

Julian Götz, Euan Richards, Iain A. Stepek, Yu Takahashi, Yi-Lin Huang, Louis Bertschi, Bertran Rubi, and Jeffrey W. Bode. Predicting three-component reaction outcomes from 40,000 miniaturized reactant combinations. *Science Advances*, 11(22):eadw6047, 2025. doi: 10.1126/sciadv.adw6047. URL https://www.science.org/doi/abs/10.1126/sciadv.adw6047.

Hiroki Hayashi, Satoshi Maeda, and Tsuyoshi Mita. Quantum chemical calculations for reaction prediction in the development of synthetic methodologies. *Chemical Science*, 14:11601–11616, 2023. doi: 10.1039/D3SC03319H.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: Machine learning package for chemical property prediction. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2023-00vcg-v2. URL https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/64bbe0a6b053dad33ab29040/original/chemprop-machine-learning-package-for-chemical-property-prediction.pdf.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Steven M. Kearnes, Ryan L. Maser, Michael Wleklinski, Anna Kast, William H. Green, Klavs F. Jensen, and Connor W. Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021. doi: 10.1021/jacs.1c09820.

Emma King-Smith, Louise Bernier, Simon Berritt, and et al. Probing the chemical 'reactome' with high-throughput experimentation data. *Nature Chemistry*, 16(4):633–643, 2024. doi: 10.1038/s41557-023-01393-w.

Shane W. Krska, Daniel A. DiRocco, Spencer D. Dreher, and Michael Shevlin. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Accounts of Chemical Research*, 50(12):2976–2985, 2017. doi: 10.1021/acs.accounts.7b00428.

Paul D. Leeson, A. Patricia Bento, Anna Gaulton, Anne Hersey, Emma J. Manners, Chris J. Radoux, and Andrew R. Leach. Target-based evaluation of "drug-like" properties and ligand efficiencies. *Journal of Medicinal Chemistry*, 64(11):7210–7230, 2021. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.1c00416. URL https://doi.org/10.1021/acs.jmedchem.1c00416.

Babak Mahjour, Juncheng Lu, Jenna Fromer, Nicholas Casetti, and Connor Coley. Ideation and evaluation of novel multicomponent reactions via mechanistic network analysis and automation. ChemRxiv Preprint, Version 3, September 2024. Working paper.

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. Relative molecule self-attention transformer. *Journal of Cheminformatics*, 16(1):3, 2024. doi: 10.1186/s13321-023-00789-7. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00789-7.

Steven M. Mennen, Carolina Alhambra, C. Liana Allen, Mario Barberis, Simon Berritt, Thomas A. Brandt, Andrew D. Campbell, Jesús Castañón, Alan H. Cherney, Melodie Christensen, David B. Damon, J. Eugenio De Diego, Susana García-Cerrada, Pablo García-Losada, Rubén Haro, Jacob Janey, David C. Leitch, Ling Li, Fangfang Liu, Paul C. Lobben, David W. C. MacMillan, Javier Magano, Emma McInturff, Sebastien Monfette, Ronald J. Post, Danielle Schultz, Barbara J. Sitter, Jason M. Stevens, Iulia I. Strambeanu, Jack Twilton, Ke Wang, and Matthew A. Zajac. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Organic Process Research & Development*, 23(6):1213–1242, 2019. doi: 10.1021/acs.oprd.9b00140.

Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9:203–214, 2010. doi: 10.1038/nrd3078. URL https://www.nature.com/articles/nrd3078.

RCSB PDB. Pdb-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Science*, 31:129–140, 2022. doi: 10.1002/pro.4200.

RCSB PDB. Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic Acids Research*, 53:D564–D574, 2025. doi: 10.1093/nar/gkae1091.

Priyanka Raghavan, Alexander J. Rago, Pritha Verma, Majdi M. Hassan, Gashaw M. Goshu, Amanda W. Dombrowski, Abhishek Pandey, Connor W. Coley, and Ying Wang. Incorporating synthetic accessibility in drug design: Predicting reaction yields of suzuki cross-couplings by leveraging abbvie's 15-year parallel library data set. *Journal of the American Chemical Society*, 146(22):15113–15125, 2024. doi: 10.1021/jacs.4c00098. URL https://pubs.acs.org/doi/10.1021/jacs.4c00098. Open Access.

Mehdi Saebi, Wenhao Gao, Łukasz Maziarka, and Connor W. Coley. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 14, 2023. doi: 10.1039/D2SC06041H.

Michael Shevlin. Practical high-throughput experimentation for chemists. *ACS Medicinal Chemistry Letters*, 8(6):601–607, 2017. doi: 10.1021/acsmedchemlett.7b00165. URL https://pubs.acs.org/doi/10.1021/acsmedchemlett.7b00165.

Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesús 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. doi: 10.1038/s41586-021-03213-y.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Paul Eiden, Hua Gao, Alberto Guzman-Perez, Terra Hopper, Bryan Kelley, Matthias Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Chonghuan Zhang, Qianghua Lin, Chenxi Yang, Yaxian Kong, Zhunzhun Yu, and Kuangbiao Liao. Intermediate knowledge enhanced the performance of the amide coupling yield prediction model. *Chemical Science*, 16:11809–11822, 2025. doi: 10.1039/D5SC03364K. URL https://pubs.rsc.org/en/content/articlelanding/2025/sc/d5sc03364k. Open Access.

Gengmo Zhou, Zhifeng Gao, Qiaoyu Ding, Zhen Zheng, Hongteng Zhang, Wei Xu, Zhongli Wei, Lu Zhang, Guolin Ke, Zhen Dong, Yu Zheng, Fan Yang, Jie Yang, Junchi Yan, Jun Zhou, Wei Fan, Ruiqi Wang, Xipeng Qiu, Hao Cheng, Shuguang Cui, Junbo Zhang, Zhiyong Liu, Zhihong Ma, Weiping Jia, Peng Xie, Jianwen Gao, Quanquan Gu, H. Eugene Stanley, Wei Li, Jinbo Xu, Jun Zhang, Jun Zhu, Jian Wang, Jun Wang, Yixue Li, Yang Yu, Weinan Zhang, Ming Chen, Rui Jiang, Jian Wang, Jun Wang, Yixue Li, Yang Yu, Weinan Zhang, Ming Chen, and Rui Jiang. Unimol: A universal 3d molecular representation learning framework. *Advances in Neural Information Processing Systems*, 37:1–12, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/53923bb44655a7defb31c7744c01b62b-Paper-Conference.pdf.