Predicting Three-Component Reaction Outcomes from 40k Miniaturized Reactant Combinations

Julian Götz[1], Euan Richards[1], Iain A. Stepek[1†], Yu Takahashi[1‡], Yi-Lin Huang[1§], Louis Bertschi[2], Bertran Rubi[2¶], Jeffrey W. Bode[1*]

Affiliations:

[1] Department of Chemistry and Applied Biosciences, ETH Zürich; 8093 Zürich, Switzerland.

[2] Molecular and Biomolecular Analysis Service (MoBiAS), Department of Chemistry and Applied Biosciences, ETH Zürich; 8093 Zürich, Switzerland.

*Corresponding author. Email: bode@org.chem.ethz.ch

† Present address: Chemical Development, Novo Nordisk; 2880 Bagsværd, Denmark.
‡ Present address: Modality Research Laboratories, Kyowa Kirin Co.; 411-8731 Shizuoka, Japan.
§ Present address: Quality Control, Bachem AG; 4416 Bubendorf, Switzerland.
¶ Present address: Biologics Process Development, LCMS R&D, Lonza AG; 3930 Visp, Switzerland.

Abstract

Efficient drug discovery relies on accessing diverse small molecules expediently and reliably. Improvements to reliability through machine learning predictions are hampered by poor availability of high-quality reaction data. Here, we introduce an on-demand synthesis platform based on a three-component reaction that delivers drug-like molecules overnight. Miniaturization and automation enable the execution and analysis of 50,000 reactions on a 3 μL scale with distinct substrates, producing the largest public reaction outcome dataset. With machine learning, we accurately predict the result of unknown reactions and analyze the impact of data set size on model training. This study advances the on-demand synthesis of drug-like molecules through concatenating chemoselective reactions and provides a sufficiently large data set to critically evaluate emerging machine learning approaches to predicting chemical reactivity.

Introduction

Accurately predicting the outcome of organic reactions through data-driven methods remains an unmet challenge, chiefly due to the paucity of large, unbiased data sets that cover a meaningful range of molecular properties and structural diversity (*1*, *2*). Molecular assembly processes involving three or more components are uniquely suited to generating large libraries of organic molecules from a modest number of building blocks, rendering them ideal for producing the experimental data needed for evaluating and improving computational approaches to the prediction of reaction outcome. For example, well studied multi-component reactions including Ugi condensations (*3*, *4*) constitute the basis for work at the forefront of miniaturized library generation (*5*).

Inspired by such processes, our group has reported Synthetic Fermentation – the assembly of specialized building blocks into functionally and stereochemically rich oligomers without the need for exogenous reagents (*6*, *7*). We recognized that similar processes could be leveraged to yield more drug-like molecules through three-component couplings, enabling a miniaturized approach to the production of hundreds of thousands of compounds from a limited selection of starting materials. Importantly, this approach presented the opportunity to generate an extensive data set comprising the outcome for tens of thousands of unique reactant combinations, thereby establishing an ideal data set for machine learning assisted predictions of successful reactant combinations. Turning to high-throughput experimentation to provide tailored data sets relevant to the prediction problem at hand has become a standard practice in the field (*8*–*11*), but it remains limited to a few thousand reactions in total and rarely more than a few dozen different substates (*12*). Assembling a data set on the order of 10s of thousands of reactions would enable testing of unconfirmed hypotheses in the field, including what quantity of data is necessary for predictive models and the dependence of extrapolation ability on model type and data set size.

In this study, we develop an on-demand reaction platform that uses three-component assembly to conduct 50,000 reactions on microliter scale, each using a unique combination of substrates. Analyzing all reaction mixtures using liquid chromatography high-resolution mass spectrometry (LC-HRMS), we track the formation of eight different products for each combination, resulting in a reaction outcome data set of unprecedented size comprising ~40,000 reactions and ~320,000 discrete products after filtering for multiple quality control criteria. Using machine learning, we accurately predict the reaction outcome for unseen reactant combinations. By investigating model performance at different data set sizes, we can derive general conclusions for data efficiency of machine learning models in reaction prediction and the relative importance of generating more data over developing better models.

Reaction development and automation

In this implementation of Synthetic Fermentation, we developed a three-component reaction where a potassium acyltrifluoroborate (KAT) initiator (**I**) reacts with one equivalent of an isoxazolidine monomer (**M**) bearing a protecting group that reveals an alpha-keto acid. This reacts with a thiohydrazide (TH) or aminobenzenethiol (ABT) terminator (**T**) to give the final product (Fig. 1A, fig. S1). Notably, the reaction proceeds in aqueous buffer without additional reagents or catalysts and yields benign byproducts, allowing direct-to-biology screens (*6*, *13*). We curated a building block collection initially comprising 78 initiators, 74 monomers, and 41 terminators (fig. S2–4). Three different initiator types (aliphatic, aromatic, heteroaromatic), four different monomer types (alpha/beta-substituted vis-à-vis the nitrogen, *cis*-substituted, quaternary), and two terminator types (ABT, TH) give rise to an on-demand library of ~236,000 products (Fig. 1B) that we call “PRIME library” – i.e. the reactions to produce these structures can be executed immediately from the available stock of 193 building blocks. Our preliminary exploration of multi-component assembly of these products revealed that the reaction usually affords the desired product (**A**), but several byproducts were observed if the reaction stalls ahead of the final oxidative decarboxylation (product **B**) or undergoes intramolecular cyclization instead (**C**). In certain cases, the **I-T** condensation product (**D**) and the terminator dimer (**E**) are also detectable. We further observed stalled **I-M** intermediate (**F**) and the corresponding carboxylic acid **G**. Finally, an **M-T** elimination product (**H**) was formed in some test reactions (Fig. 1C). The various

reaction outcomes, particularly the distinct reaction pathways leading to desired structure **A** versus alternative product **C**, rendered this system well suited for both high-throughput experimentation and as a testing ground for modern reaction prediction.

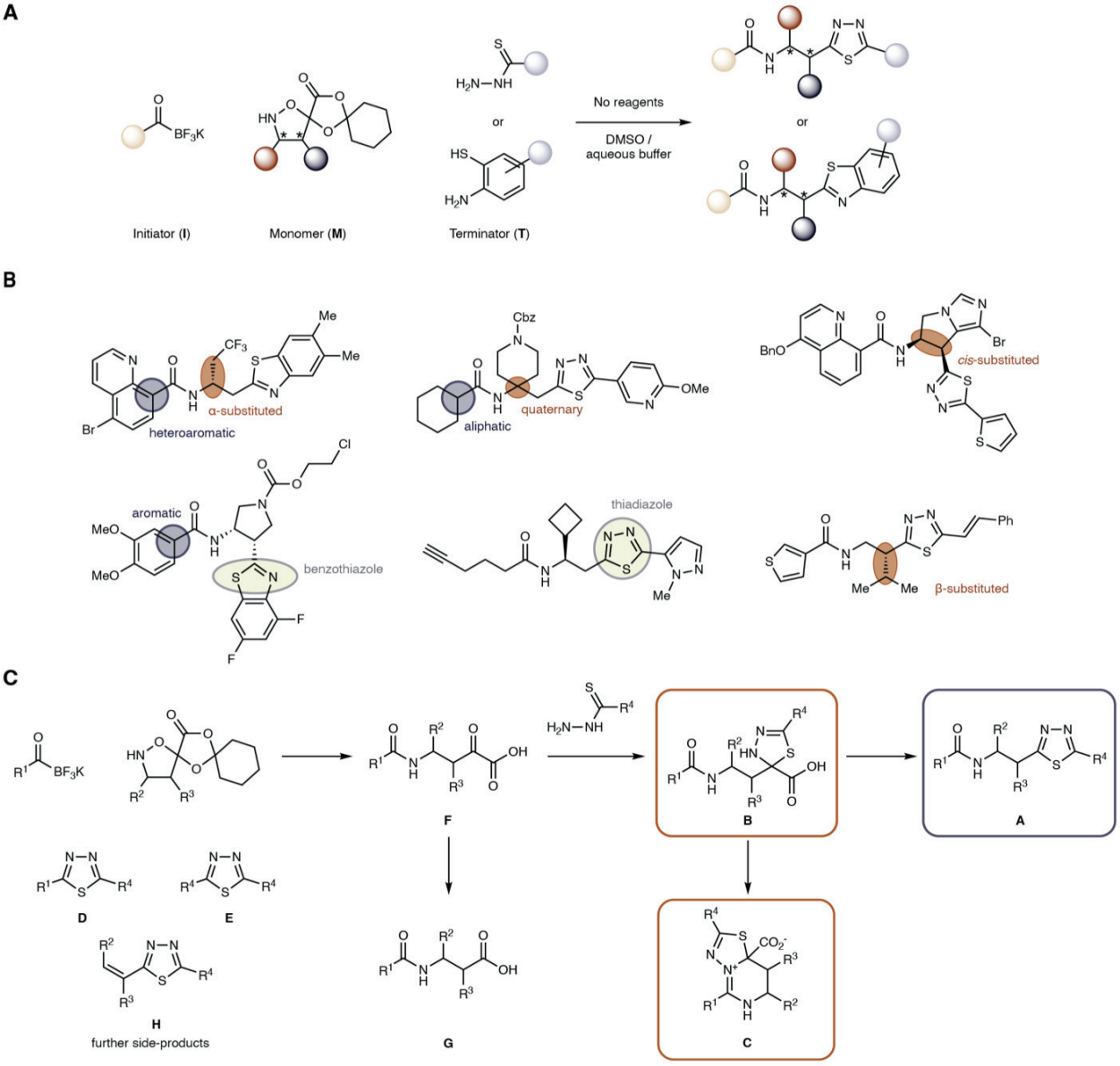

Fig. 1. Synthetic Fermentation. (**A**) A potassium acyltrifluoroborate initiator (**I**) reacts with an isoxazolidine monomer (**M**) to produce an alpha-keto acid that subsequently condenses with a thiohydrazide or aminobenzenethiol (ABT) terminator (**T**) to form the final product. The reaction sequence proceeds without additional reagents or catalysts. (**B**) Examples of products obtained through Synthetic Fermentation with some common structural features highlighted. (**C**) The desired product **A** is formed in most reactions. Several side-products were also identified. In some cases, the reaction stalled before the final oxidative decarboxylation at product **B**. In some cases, an intramolecular condensation led to alternative, bicyclic product **C**. Further products that are occasionally observed are the intermediate alpha-keto acid (**F**) and its oxidation product **G**, the **I-T** condensation product **D**, the **T-T** condensation product **E**, and the **M-T** elimination product **H**. The analogous products formed with ABT terminators are show in fig. S1.

To make the reaction amenable to high-throughput synthesis and a wider range of substrates, we modified our preferred conditions for Synthetic Fermentation, including changing the solvent system from *t*-BuOH to DMSO/1 M aq oxalic acid and reducing the concentration of building block solutions to 50 mM to ensure solubility. Most building blocks are stable for months at –20 °C in DMSO (**I**, **M**) or in the reaction solvent (**T**). In order to reduce the consumption of building blocks and disposables, we adopted acoustic dispensing (*4*, *14*) for the automated delivery of reaction components and were pleased to find that it performed well at preparing the 384-well plates used for the reactions, with each three-component reaction conducted in 3.3 μL of 9:1 DMSO/aq oxalic acid. Even at this low volume, no notable evaporation was observed during incubation at 60 °C.

Execution and evaluation of 50,000 reactions

Library syntheses were initially conducted in batches of 1,920 reactions using 16 initiators, 12 monomers and 10 terminators per run, which required 22 h for setup and conduction. Improvements in the automated workflow, namely employing incubators within the automated system and conducting the previously manual preparation of source plates on a programmable robotic liquid handler (see Movie S1), allowed us to prepare two batches or 3,840 reactions in a 24 h period. The full automation workflow is outlined in Fig. 2A. Stock solutions of building blocks were prepared freshly or used from previous batches and aliquots were analyzed by liquid chromatography mass spectrometry (LC-MS) for quality control. The source plates were transferred to an automation system where six 384-well synthesis plates served as the targets for acoustic dispensing (see Movie S2). The plates containing initiator and monomer mixtures were incubated at 60 °C for 3 h, followed by addition of terminator solutions by acoustic dispensing and further incubation at 60 °C overnight (16 h). Dispensing, incubation, and all auxiliary steps (plate transfers, centrifuging, sealing/de-sealing) proceeded without human intervention on the automated system.

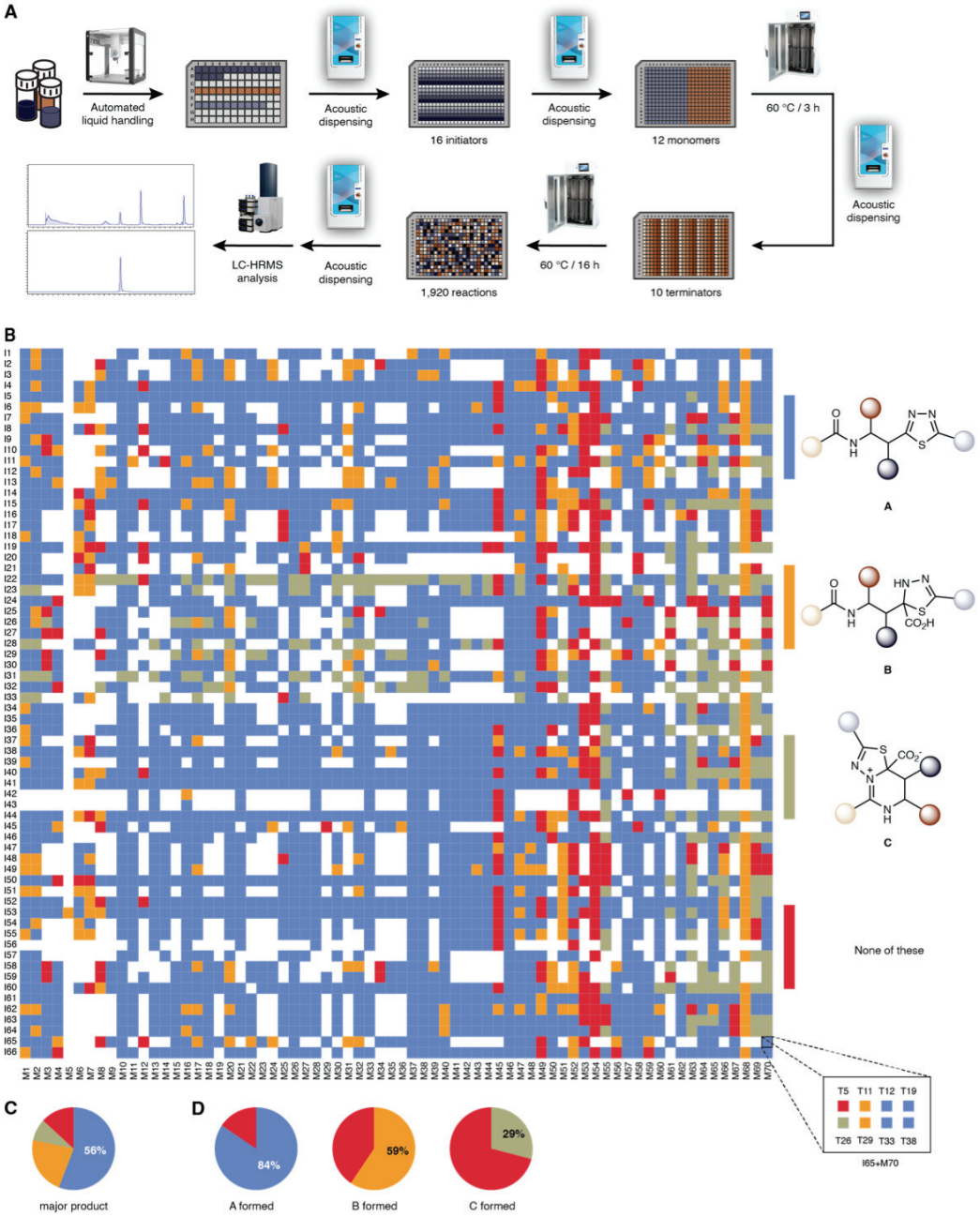

Fig. 2. Automated conduction and analysis of 50,000 reactions. (**A**) The synthesis workflow begins with the preparation of the source plate on an OT-2 liquid handler. From the source plate, solutions of **I** and **M** are distributed to six 384-well reaction plates by acoustic dispensing and incubated within the automated system. **T** solutions are added by acoustic dispensing and incubation is continued. Aliquots are removed for subsequent LC-HRMS analysis. (**B**) Heatmap of major product (between **A**, **B**, and **C**) across all initiators and monomers, averaged over all terminators tried with this combination. The "None" class is assigned if none of the three products was present. For the combination of **I65** and **M70**, the square is expanded into all reactions with different terminators making up this square. White squares indicate that the combination was not attempted. For heatmaps for the formation of individual products see fig. S7–9. An interactive version of this heatmap, showing the outcome for individual reactions is available at https://jugoetz.com/synferm-heatmap. (**C**) Product **A** was the major product in 56%, product **B** in 23%, and product **C** in 9% of all reactions. The remaining 13% afforded none of these products. (**D**) Average observation ratio for each product (**A**: 84%, **B**: 59%, **C**: 29%).

Analyzing 50,000 different reactions – each with distinct starting materials, products, and side-products – presents a formidable task. The automated system was used to prepare plates for LC-HRMS analysis by aliquoting 30 nL of the reaction solution and diluting 1:1000 with a MeCN solution containing fenofibrate as internal standard. Valuing data quality over speed, we required resolution of all expected products by chromatography, leading to an LC-HRMS runtime of 14 min per sample, including column washes (see fig. S5 for an example trace). With a dual column setup, we could reduce the effective time to 7 min/sample or about eight months of instrument time in total. Extensive scripting for the determination of required sum formulae of all expected products, including from conceivable deprotection reactions, using Python/RDKit, and near complete automation of instrument operation kept human involvement to a minimum. The peak areas for all sum formulae were determined using a VBScript within Bruker Compass Data Analysis and normalized by dividing by the peak area of the internal standard (fenofibrate). Due to challenges of normalizing product responses for 50,000 different reaction combinations, the reaction outcome was assessed with a binary label depending on the presence or absence of a product or side product in the LC-HRMS trace. The major product was determined by scaling the peak area according to the relative ionizability of the product type and comparing the scaled value across products (see Supporting Information). For inclusion in the data set, each data point was checked against several quality criteria using collected metadata such as building block LC-MS, transfer logs from acoustic dispensing, and the product LC-HRMS. These rigorous protocols ensured high data quality without experimental replicates and led to the exclusion of around 10,000 data points from the final data set. By conducting a subset of the reactions in duplicate (377 reactions), we observed good reproducibility (98% accuracy for product A, 73% balanced accuracy for the major product; see fig. S6). The outcome for ~40,000 reactions determined to be operationally successful is shown in Fig. 2B (see fig. S7–9 for individual products). An interactive version of this heatmap, showing the outcome for individual reactions is available at https://jugoetz.com/synferm-heatmap. In 84% of reactions, the expected product **A** is observed, and it is the major product in 56% of reactions. Product **B** is observed irregularly with few building blocks strongly favoring it. Formation of product **C** is more likely for aliphatic initiators (**I22**–**I33**) or monomers bearing a quaternary carbon (**M55**–**M70**). Only few building blocks afford none of these products. For example, monomers with a 2-pyridyl or 4-pyridyl substituent in the beta-position (**M53**, **M54**) appear to degrade under the established reaction conditions.

Prediction of reaction outcome

Having collected the reaction outcome data, we turned to machine learning to predict the outcome for arbitrary combinations of building blocks. Our use of a three-component reaction has the consequence of machine learning models confronted with four distinct prediction problems of increasing complexity, i.e. predicting for reactions with three, two, one, or zero previously seen building blocks (see Fig. 3A). We trained and evaluated models independently for all four problems, which we refer to according to the number of independent dimensions between training and test data. The “0D split” corresponds to zero independent dimensions or all three building blocks being present in the training data, whereas in the 3D split none of the three building blocks would be present in the training data (Fig. 3B). The splitting method is described in detail in our previous work (*15*) and in the Supporting Information. Similar approaches have been used elsewhere (*16*–*19*) to prevent information leakage. For each problem, we tried several different representations, spanning structural fingerprints (FP), RDKit

properties, or one-hot encoding (OHE) of reactants, or graph encoding of the reaction (CGR) (*20*). We tested model architectures comprising several graph neural networks (GNN) (*21–24*), a feed-forward network (FFN), gradient boosting (XGB) (*25*), and logistic regression. Hyperparameters were optimized for each combination. The best classifier for each problem was selected by the area under the precision-recall curve (AUPRC) obtained on validation data and computationally cheaper methods were favored if score differences were not significant.

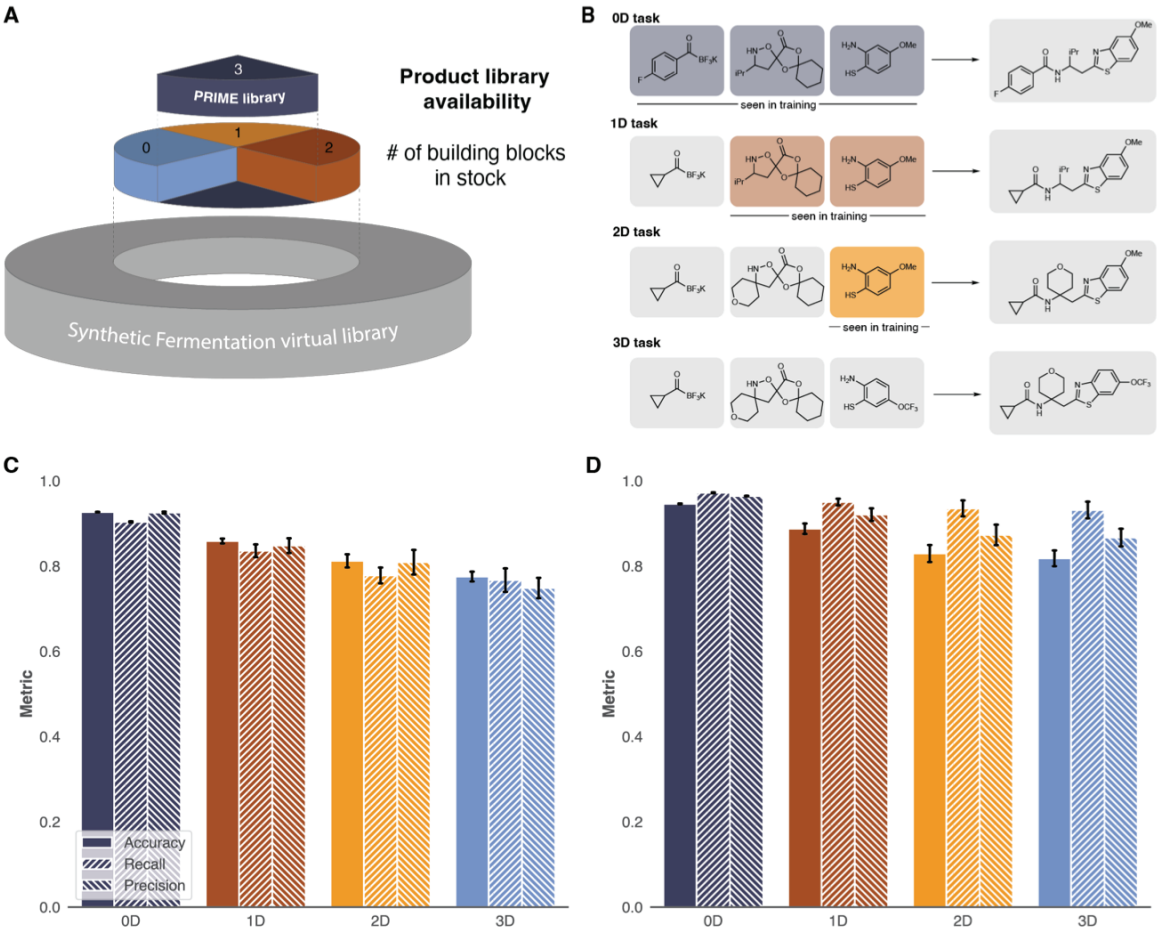

Fig. 3. Data-driven prediction of reaction outcome. (**A**) Different library spaces are differentially difficult to predict. The Synthetic Fermentation virtual library can be subdivided based on the number of new building blocks needed to synthesize a VL member. With the number of new building blocks, the difficulty of predictions increases. The training data is a subset of the PRIME library. (**B**) Data splits created to simulate the different prediction problems faithfully. Building blocks in colored boxes are contained in the training data for each task, building blocks in grey boxes are not contained in the training data. For example, a 1D split would leave out building blocks on one dimension (here initiators) but contain all building blocks for the other two dimensions (monomers, terminators) in both training and test data. Using multiple folds, the three possible permutations (for 1D and 2D split) of reactant types to be left out are investigated at the same time. (**C**) Performance (accuracy, precision, and recall) of the best model on each task, averaged over all products **A**–**C** and (**D**) for product **A**. The selected models were FFN/OHE for the 0D problem and XGB/FP for the other problems. Error bars indicate the standard error of the mean over nine random repetitions.

We found that for the 0D split (i.e. all building blocks have been seen), the FFN/OHE model was superior, significantly outperforming all other combinations. For the 1D, 2D, and 3D splits the XGB/FP models were optimal, though the computationally more expensive directed message passing neural

network (D-MPNN)/CGR combination showed similar performance on most problems (see fig. S10 for all models). Evaluating the selected models on test data, we found the 0D model to have an accuracy of 93%±0.2% for predicting products **A**–**C**. The 1D model still showed a good accuracy of 86%±1.6%. The 2D model was correct for 81%±4.4% of reactions and the 3D model for 78%±3.3%. These metrics and the respective precision and recall scores as well as metrics for product **A** only are shown in Fig. 3.

While these results are satisfactory, it is worth highlighting their limitations. For product **A** specifically, which was formed in more than 4 out of 5 reactions, only the 0D and 1D models are meaningfully more accurate than a naïve classifier that always predicts that the expected product is formed. The monotonous decline of accuracy from the 0D to the 3D model is expected and reflects the relative difficulty of learning chemical information in contrast to learning the combinatorial structure of the data set. Still, on the 3D split, the "chemistry-aware" XGB/FP is significantly and markedly better than the chemistry-unaware FFN/OHE model which (by design of the split) is only as good as random guesses.

Experimental validation

In addition to the computational validation, models were tested on prospective tasks reflecting real-world use cases (Fig. 4A). As we use the 0D model to predict reaction outcomes for the PRIME library, (we tested the precision of this model by preparing one plate of compounds predicted to be synthesizable. Out of 149 valid reactions that were predicted to yield product **A** by the model, it was found in 147 of them, corresponding to a 99% precision score in this prospective evaluation (Fig. 4B). For the secondary objectives, predicting formation of products **B** and **C**, the prospective accuracy was 87% and 86%, respectively.

The second task reflects a question that may arise when using Synthetic Fermentation for compound optimization. If an initial hit from the PRIME library was promising but analogues are desired for optimization, one may synthesize new initiators not contained in the PRIME library. To check whether these building blocks are worth procuring, one would ask the 1D model for the likely reaction outcome. To test this scenario, we curated a diverse set of twelve initiators that were not used in the library synthesis. These new initiators were reacted with random monomers and terminators from the library. For 593 valid reactions, the model successfully predicted whether product **A** is formed for 533 (90%) of them (Fig. 4C). Notably, most errors were false negatives and in only six cases was product **A** not detected despite being predicted to form, corresponding to 98.9% precision. As the main objective in this scenario is avoiding costly synthesis of unproductive building blocks, the model proved highly suitable.

The strong prospective performance, combined with the previous computational evaluation demonstrates practical usefulness of the models.

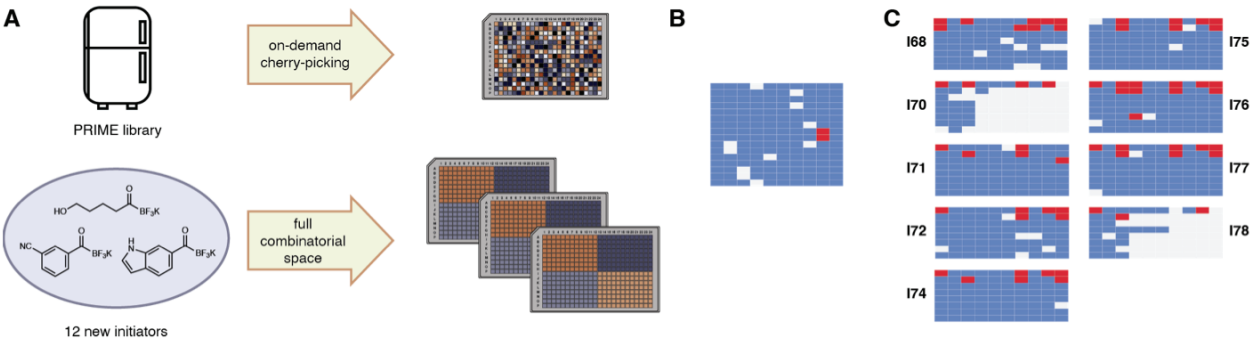

Fig. 4. Experimental validation. (**A**) Tasks posed in experimental validation. One plate was created by cherry-picking compounds from the in-stock PRIME library that the 0D model predicted to work. Three plates were created based on 12 new initiators previously not contained in the PRIME library, paired with 10 previously used monomers and 8 previously used terminators. Reaction outcomes for these plates were predicted with the 1D model (**B**) Heatmap of reaction outcome for the cherry-picked plate. Blue squares indicate product **A** was found (i.e. the 0D model was correct). Red squares indicate it was not found. Grey squares were excluded from the analysis because they failed previously established quality control criteria. (**C**) Heatmap showing whether the model predicted the correct outcome for the plates with new initiators, grouped by initiator. Blue squares indicate a correct prediction, red squares an incorrect prediction. Grey squares were excluded from the analysis because they failed previously established quality control criteria.

Data and model requirements for reaction prediction

The unprecedently large data set gives the opportunity to investigate the influence of data set size on the performance of different model types. For each of the problems outlined above, we created truncated splits with a reduced number of data points in the training set. We selected the three model types which showed best performance for the full data set (FFN/OHE, XGB, D-MPNN). For the chemistry-aware models we used either structural features (FP for XGB and CGR for D-MPNN) or additionally property features (RDKit). The five model types were trained (including full hyperparameter tuning) for all truncated splits.

On the 0D split (Fig. 5A) i.e. a pure interpolation task, the performance for all model types converges above 2,000 data points. Despite the small difference, the FFN/OHE model is significantly better than all other model types at large data set sizes. While the predictive power keeps rising even up to ~30,000 training data points, working with a much sparser data set of a few thousand samples yields satisfactory models. The differentiation below 2,000 samples, where XGB models outperform neural network methods, hints at the higher data efficiency of the former. On the 1D split (Fig. 5B), a similar picture emerges. For small training data sets, XGB is the best choice, but with increasing data set size (>4,000 samples), the advantage over neural networks vanishes. Surprisingly, the chemistry-unaware model (FFN/OHE) is indistinguishable from the chemistry-aware methods up to training data set sizes of ~8,000 samples. On the 2D split (Fig. 5C), scores are lower across model types, but the advantage of XGB models for small data sets is reproduced and extends to even larger sizes. The chemistry-unaware model becomes unfavorable at smaller sizes (a few thousand data points). Lastly, on the 3D split (Fig. 5D), the chemistry-unaware model gives random predictions (as expected by design). The markedly increased variance of the results leads to no consistent differentiation among the chemistry-aware models. Performance remains moderate, but much better than chance up to the maximum data set size.

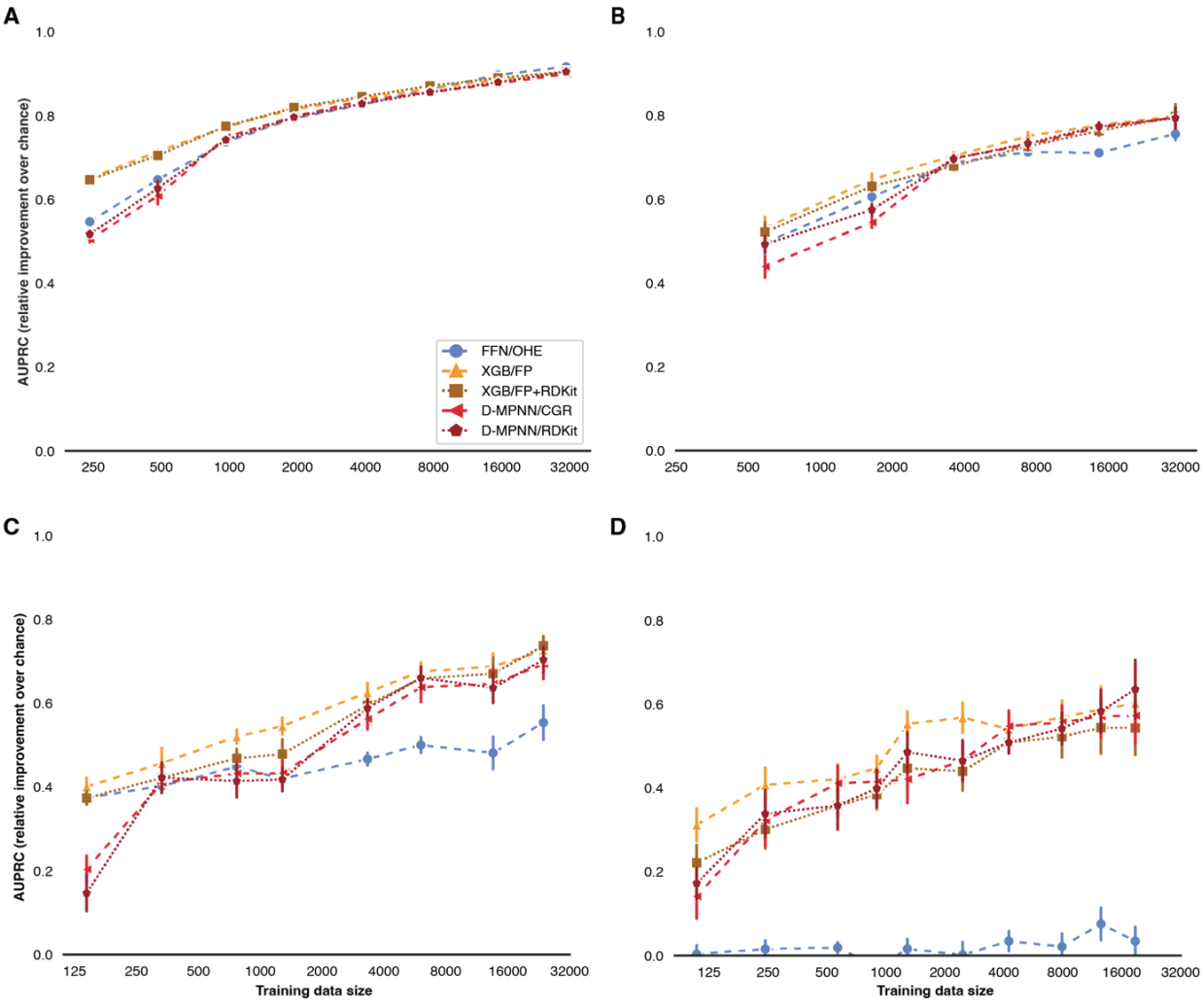

Fig. 5. Data efficiency of machine learning models. Relative improvement over the chance level for the area under the precision-recall curve (AUPRC) on test data depending on the size of the data set used in training. As relative improvement is given, a random classifier has an expected value of 0 and a perfect classifier has an expected value of 1. All error bars correspond to the standard error of the mean over 9 random repetitions. (**A**) On the 0D split, differentiation is small but significant at large data set sizes. Below a few thousand training samples, XGB markedly outperforms even a simple neural network and outperforms GNNs by a wider margin, irrespective of supplementary featurization with RDKit features. (**B**) On the 1D split, XGB models are best for most data set sizes, but GNNs catch up for large training data sets. Only at large sizes, the chemistry-aware models outperform the chemistry-unaware model (FFN/OHE). (**C**) On the 2D split, XGB models are preferred at all training data set sizes, but GNNs can outperform the chemistry-unaware model after a few thousand data points. (**D**) On the 3D split, the chemistry-unaware model is not better than chance, while chemistry-aware models are moderately predictive and do not consistently differentiate in performance.

These observations show that while predicting within a combinatorial reaction data set (equivalent to the 0D split) is possible even from a sparsely sampled combinatorial space (i.e. approximately 1% of ~200,000 possible data points), there is no reason to use chemistry-aware models in such cases. Even when extrapolating for one of the three reactants (1D split), chemistry-awareness became relevant only beyond 10,000 training data points in our case, which is an order of magnitude larger than previously examined data sets. If chemistry-aware models are used for the 0D and 1D split, we

saw only disadvantages for more complicated GNNs over XGB, especially at the small data set sizes that are common in chemistry.

Conclusion

In summary, we have developed an automated, microscale synthesis platform for the on-demand synthesis of vast numbers of drug-like compounds. Based on a three-component reaction, currently 220,990 VL members are immediately accessible from 188 curated building blocks. We conducted and analyzed 50,000 reactions affording distinct products, generating the first public reaction data set of this magnitude. We used this data to train machine learning classifiers for reaction outcome with excellent accuracy within the PRIME library and notable generalization ability to new building blocks. The usefulness of the classifiers was corroborated by prospective experimental validation. The previously unprecedented size of this data set allowed us to study the data requirements of different machine learning models under a set of interpolation and extrapolation tasks, revealing that reaction outcome prediction problems are best approached as interpolation tasks, but that sparsely (~1%) sampling the combinatorial space suffices for training predictive models, substantiating prior conjectures (*26*). For either interpolation or extrapolation tasks, we find that simpler models outperform frequently used neural network based approaches at data set sizes previously available in the field. Combined, our observations suggest that the optimal approach to reliable reaction outcome predictions is building simple, local models with sparse data sets; elaborate model architectures and featurization will not suffice to construct general models of chemical reactivity, at least not without new modes of data collection.

References and Notes

1. W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **144**, 4819–4827 (2022).

2. M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle, N. V. Chawla, O. Wiest, On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **14**, 4997–5005 (2023).

3. I. Ugi, C. Steinbrückner, Über ein neues Kondensations-Prinzip. *Angew. Chem.* **72**, 267–268 (1960).

4. F. Sutanto, S. Shaabani, C. G. Neochoritis, T. Zarganes-Tzitzikas, P. Patil, E. Ghonchepour, A. Dömling, Multicomponent reaction–derived covalent inhibitor space. *Sci. Adv.* **7**, eabd9307 (2021).

5. Z. Wang, S. Shaabani, X. Gao, Y. L. D. Ng, V. Sapozhnikova, P. Mertins, J. Krönke, A. Dömling, Direct-to-biology, automated, nano-scale synthesis, and phenotypic screening-enabled E3 ligase modulator discovery. *Nat. Commun.* **14**, 8437 (2023).

6. Y. L. Huang, J. W. Bode, Synthetic fermentation of bioactive non-ribosomal peptides without organisms, enzymes or reagents. *Nat. Chem.* **6**, 877–884 (2014).

7. I. A. Stepek, T. Cao, A. Koetemann, S. Shimura, B. Wollscheid, J. W. Bode, Antibiotic Discovery with Synthetic Fermentation: Library Assembly, Phenotypic Screening, and Mechanism of Action of β-Peptides Targeting Penicillin-Binding Proteins. *ACS Chem. Biol.* **14**, 1030–1040 (2019).

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

9. N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, M. D. Burke, Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **378**, 399–405 (2022).

10. J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Püntener, K. A. Mack, M. S. Sigman, Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **145**, 110–121 (2023).

11. D. F. Nippa, K. Atz, R. Hohler, A. T. Müller, A. Marx, C. Bartelmus, G. Wuitschik, I. Marzuoli, V. Jost, J. Wolfard, M. Binder, A. F. Stepan, D. B. Konrad, U. Grether, R. E. Martin, G. Schneider, Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nat. Chem.* **16**, 239–248 (2024).

12. E. King-Smith, S. Berritt, L. Bernier, X. Hou, J. L. Klug-McLeod, J. Mustakis, N. W. Sach, J. W. Tucker, Q. Yang, R. M. Howard, A. A. Lee, Probing the chemical 'reactome' with high-throughput experimentation data. *Nat. Chem.* **16**, 633–643 (2024).

13. J. G. Hubert, I. A. Stepek, H. Noda, J. W. Bode, Synthetic fermentation of β-peptide macrocycles by thiadiazole-forming ring-closing reactions. *Chem. Sci.* **9**, 2159–2167 (2018).

14. B. Hadimioglu, R. Stearns, R. Ellson, Moving Liquids with Sound: The Physics of Acoustic Droplet Ejection for Robust Laboratory Automation in Life Sciences. *J. Lab. Autom.* **21**, 4–18 (2016).

15. J. Götz, M. K. Jackl, C. Jindakun, A. N. Marziale, J. André, D. J. Gosling, C. Springer, M. Palmieri, M. Reck, A. Luneau, C. E. Brocklehurst, J. W. Bode, High-throughput synthesis provides data for predicting molecular properties and reaction success. *Sci. Adv.* **9**, eadj2314 (2023).

16. A. F. Zahrt, J. J. Henle, S. E. Denmark, Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **22**, 586–591 (2020).

17. A. Vall, S. Hochreiter, G. Klambauer, "BioassayCLR: Prediction of biological activity for novel bioassays based on rich textual descriptions" in *ELLIS Machine Learning for Molecule Discovery Workshop 2021* (2021).

18. N. I. Rinehart, R. K. Saunthwal, J. Wellauer, A. F. Zahrt, L. Schlemper, A. S. Shved, R. Bigler, S. Fantasia, S. E. Denmark, A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. *Science* **381**, 965–972 (2023).

19. R. Joeres, D. B. Blumenthal, O. V. Kalinina, DataSAIL: Data Splitting Against Information Leakage. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.11.15.566305.

20. E. Heid, W. H. Green, Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **62**, 2101–2110 (2022).

21. T. N. Kipf, M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks." in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (2017).

22. W. L. Hamilton, R. Ying, J. Leskovec, Inductive Representation Learning on Large Graphs. arXiv arXiv:1706.02216 [Preprint] (2018). https://doi.org/10.48550/arXiv.1706.02216.

23. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

24. Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, M. Zheng, Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **63**, 8749–8760 (2020).

25. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016; http://arxiv.org/abs/1603.02754), pp. 785–794.

26. P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman, C. W. Coley, Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **9**, 2196–2204 (2023).

Acknowledgments

We thank Dr. Cara Brocklehurst and Dr. Kian Tan for fruitful discussions. We acknowledge Michael Meier of the Molecular and Biomolecular Analysis Service (MoBiAS) of the Department of Chemistry and Applied Biosciences at ETH Zürich for mass spectrometry. We acknowledge Dr. Eva Riegler, Dr. Tijmen Booij, and David Keller of NEXUS at ETH Zürich for their help with high-throughput experiments. We are grateful to Gabor Erös, Aaron Dumas, Sara Da Ros, Alberto Gálvez, Anne Schuhmacher, Matthias Tanriver, Dominik Schauenburg, Philipp Schilling, Sizhou Liu, Dino Wu, Dmitry Mazunin, Fabio Masero, Jonathan Hubert, Tomoya Shiro, and Yi-Chung Dzeng for contributing materials.

Funding

Novartis Global Scholars Program (JWB) and ETH Zürich (JWB)

Author contributions

Conceptualization: JG, IAS, YLH, JWB

Methodology: JG, ER, IAS, YT, YLH, LB, BR

Software: JG, LB

Formal Analysis: JG

Investigation: JG, ER, IAS, YT, YLH, LB

Data Curation: JG, ER, IAS, YT, LB

Visualization: JG

Validation: JG, ER, LB

Funding acquisition: JWB

Project administration: JG, BR, JWB

Supervision: JWB

Writing – original draft: JG, JWB

Writing – review & editing: JG, ER, IAS, YT, YLH, LB, BR, JWB

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

The data set will be made available upon publication in a peer-reviewed journal. Code is available from GitHub at https://github.com/jugoetz/library-generation (data processing and analysis) and https://github.com/jugoetz/synferm-predictions (machine learning). Versions archived at the time of publication are available from Zenodo at https://doi.org/10.5281/zenodo.11121449 and https://doi.org/10.5281/zenodo.11121456.