Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 304

Democratizing machine learning in chemistry with community-engaged test sets

Jason L. Wu,[ab] David M. Friday,[ab] Changhyun Hwang,[bc] Seungjoo Yi,[bd] Tiara C. Torres-Flores,[bc] Martin D. Burke,[abefg] Ying Diao,[abc] Charles M. Schroeder[abcd] and Nicholas E. Jackson *[ab]

Machine learning (ML) is increasingly central to chemical discovery, yet most efforts remain confined to distributed and isolated research groups, limiting external validation and community engagement. Here, we introduce a generalizable mode of scientific outreach that couples a published study to a community-engaged test set, enabling post-publication evaluation by the broader ML community. This approach is demonstrated using a prior study on AI-guided discovery of photostable light-harvesting small molecules. After publishing an experimental dataset and in-house ML models, we leveraged automated block chemistry to synthesize nine additional light-harvesting molecules to serve as a blinded community test set. We then hosted an open Kaggle competition where we challenged the world community to outperform our best in-house predictive photostability model. In only one month, this competition received >700 submissions, including several innovative strategies that improved upon our previously published results. Given the success of this competition, we propose community-engaged test sets as a blueprint for post-publication benchmarking that democratizes access to high-quality experimental data, encourages innovative scientific engagement, and strengthens cross-disciplinary collaboration in the chemical sciences.

Received 19th September 2025
Accepted 18th November 2025

DOI: 10.1039/d5dd00424a

rsc.li/digitaldiscovery

1 Introduction

Machine learning (ML) has become a transformative tool across the chemical sciences,[1] enabling advances in molecular property prediction,[2] reaction optimization,[3] polymer science,[4,5] materials discovery,[6] and beyond. By identifying complex patterns in high-dimensional data, ML allows chemists to accelerate hypothesis generation, reduce experimental workloads, and uncover relationships that may elude traditional scientific paradigms.[7] As datasets grow in size and complexity, and as computational tools become more accessible, ML is increasingly positioned as a core competency of modern chemical research.[8] However, realizing the full potential of ML in chemistry demands robust, reproducible benchmarks and strong collaboration across experimental and computational domains throughout the world.

Despite the technological momentum of ML methods, there is a growing awareness that the scientific community must develop more inclusive and engaging ways of connecting with the public. ML presents a rare opportunity: its widespread accessibility, intuitive appeal, and applicability across disciplines make it an ideal entry point for engaging students, hobbyists, and educators alike. Public enthusiasm for ML is high, yet structured avenues for meaningful participation in the chemical sciences are limited. Simultaneously, the need for effective scientific communication has never been more urgent in a global landscape shaped by rapid technological change and distrust of expertise; the ability to convey the significance and impact of scientific discoveries is critical. Traditional modes of scientific dissemination, such as peer-reviewed publications and conference presentations, often fall short in reaching broad audiences. Taken together, these considerations motivate the need for innovative frameworks that not only explain research outcomes, but actively invite participation, deepen trust, and demonstrate the relevance of scientific work to societal challenges.

In this work, we introduce a new model of scientific community engagement by directly interfacing experimental chemistry with community-engaged test sets for ML. Building

[a]*Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. E-mail: jacksonn@illinois.edu*
[b]*Molecule Maker Lab, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
[c]*Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[d]*Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[e]*Carle R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
[f]*Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
[g]*Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

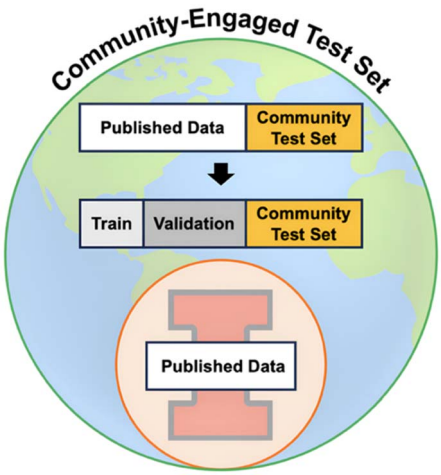

Fig. 1 Schematic of the community-engaged test set paradigm. During our previous study of light-harvesting small molecules, we performed in-house ML to predict photostability of our dataset.[9] After publishing our results, we hosted a global hackathon for photostability prediction by synthesizing and characterizing an additional community test set. Participants were provided with our published dataset (42 molecules) to train their ML models, which were then evaluated on the unseen community test set photostability values (7 molecules).

on a prior study of small-molecule photostability, we leveraged automated block chemistry to construct a blinded test set consisting of newly synthesized compounds and hosted a public Kaggle competition in which participants predicted degradation properties of the new molecules using open-source tools and training data from the prior study (Fig. 1). This approach, referred to as community-engaged test sets, invites broad participation in post-publication model validation while fostering dialogue between experimental chemistry and the global community. Here, we describe the design, execution, and outcomes of our approach, and propose this model as a scalable framework for democratizing access to data, amplifying scientific visibility, and accelerating innovation in chemistry.

2 Closed-loop transfer of small molecule photostability

We previously reported the integration of closed-loop experiments with physics-based feature selection and supervised learning, known as closed-loop transfer (CLT), to yield chemical insights together with the optimization of objective functions (Fig. 2a).[9] CLT was used to uncover the molecular properties dictating the small molecule photostability, a ubiquitous photophysical property for which general chemical design principles are lacking.[10] Key to this campaign was the automated block-based synthesis of conjugated small molecules comprised of three building blocks using an iterative cycle of deprotection (D), coupling (C), and purification (P) steps (Fig. 2b). We found that using Suzuki coupling conditions identified *via* a previous AI-guided closed-loop process[11] (GC1; Fig. 2b), and newly discovered anhydrous slow-release coupling conditions[12] (GC2; Fig. 2b) maximized the synthetic hit rate. Subsequently, the photobleaching lifetime (T_{80}), defined as the time required for the observed absorbance spectrum to decay to 80% of its initial value under constant irradiation, was measured for the small molecules *via* solution-based photodegradation in a solar irradiation cell (Fig. 2c). In tandem with synthetic efforts, we trained interpretable ML models drawing from physics-based features to generate hypotheses relating molecular features to photostability. Through this approach, we generated 114 features for our molecules using time-dependent density functional theory (TDDFT) calculations and RDKit. We then trained every combination of 4-feature support vector regression models with the radial basis function kernel (~2.5 million total 4-feature models) to predict the T_{80} of our 44-molecule dataset. The best 4-feature model (heteroatoms, rotatable bonds, TDOS 3.9, TDOS 2.5) predicted T_{80} of our dataset with an R^2 value of 0.95 (Fig. 2d). An important outcome of the interpretable ML method was the finding that high-energy triplet density of states (TDOS) is a primary determinant of T_{80}, a key contributor to overall molecular photostability. This campaign successfully generated a high-quality experimental organic chemistry dataset and demonstrated that a simple, interpretable supervised ML model using physics-based features can provide new fundamental chemical insights into complex molecular systems.

3 Democratization of chemical discovery

3.1 Community-engaged test set

The quality, quantity, and diversity of available data impose an upper limit on the accuracy and generality of any ML model.[13] Organic chemistry datasets generally struggle to maximize all three of these requirements due to time, resource, and experimental design constraints. For example, high-throughput experimental datasets are large, but only focus on a narrow scope,[14] whereas literature-derived databases are also large but lack quality control and balance of reported results.[15] Although smaller, high-quality, functional chemistry datasets exist such as Burke and Aspuru-Guzik's 413 organic laser molecule dataset[16] and Tong's 56 aluminum complex dataset,[17] the availability of chemical datasets remains limited for establishing community ML standards. Furthermore, most ML-ready chemical datasets are locked behind paywalls or buried in opaque online archives, which makes the barrier for engagement insurmountable for the average ML data scientist. Given the lack of publicly available experimental organic chemistry datasets and the room for growth of small-data ML, we envisioned that our photostability dataset would be of interest to the ML community.

Hackathons, events where computer scientists collaboratively build projects over a short, intense period of time, have become a low-cost, high-reward outreach strategy.[18] Recent competitions such as Nomad2018 Predicting Transparent Conductor,[19] Novozymes Enzyme Stability Prediction,[20] Predicting Molecular Properties,[21] and Bristol-Myers Squibb –

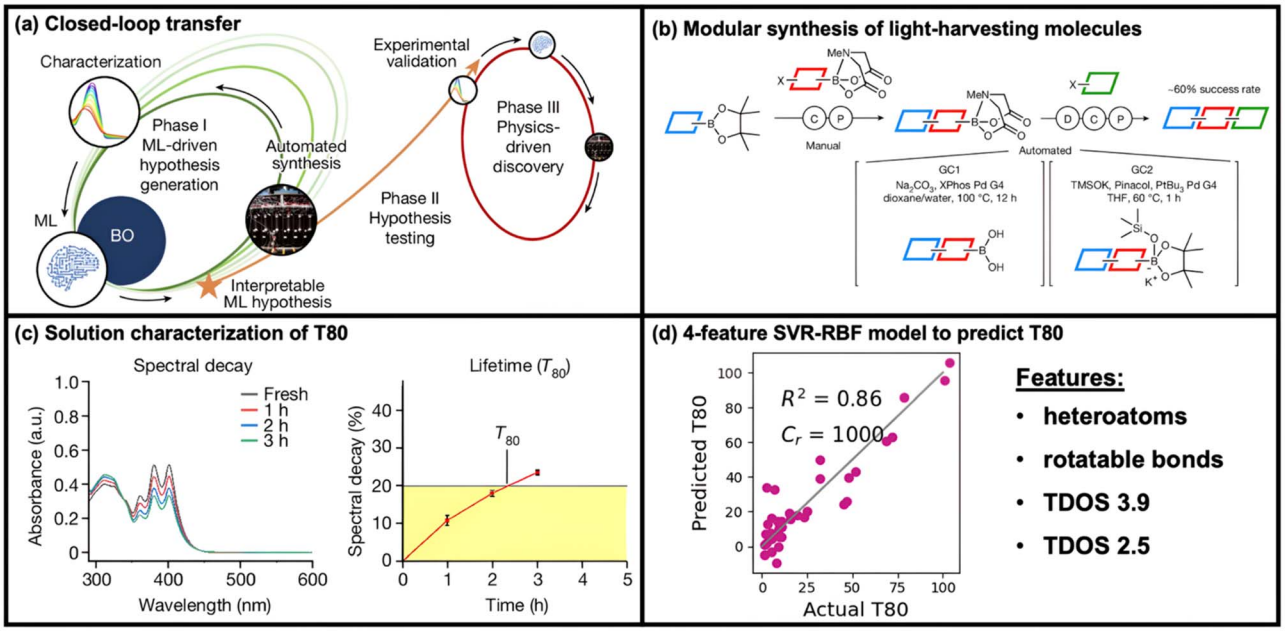

Fig. 2 Overview of our closed-loop transfer (CLT) campaign to optimize and understand photostability in light-harvesting small molecules. (a) The CLT method integrates physics-based feature selection and supervised learning with closed-loop optimization. (b) Automated synthesis used for roughly 60% of the molecules. (c) Solution characterization of the time-dependent spectral decay of absorption, which is used to calculate T_{80}. (d) Our highest performing 4-feature support vector regression with the radial basis function (SVR-RBF) T_{80} model trained on all 44 molecules from the campaign.

Molecular Translation[22] underscore the increasing intersection of ML and chemistry. Moreover, a biennial competition that has garnered success is the Critical Assessment of Protein Structure Prediction (CASP) competition.[23] For the CASP competition, hundreds of research groups attempt to predict the three-dimensional structure of a variety of proteins from only its amino acid sequence. Notably, the winners of CASP in 2018 and 2020 were AlphaFold and AlphaFold2 respectively, demonstrating the importance of hosting worldwide ML competitions.[24] Given the success of the global hackathon strategy in engaging participation and inventing scientific breakthroughs, we envisioned hosting our own hackathon based on our relatively small experimental photostability dataset. Unlike previous versions, our hackathon would be the first to study structure–function relations of small organic molecules, a fundamental challenge in academia and industry. The absence of such competitions in the organic chemistry domain is largely attributed to the lack of modular synthetic methods as well as automated characterization methods. The emergence of automatable, AI-friendly block chemistry[12,16,25–29] and many advances in automated characterization have opened the door for launching such competitions.

A potential challenge in hosting a hackathon using only the photostability dataset from the prior study is the lack of hidden data to evaluate public ML models, so any models trained on the published data would be prone to overfitting. We therefore created a new test set specifically intended for evaluating community ML models in order to conduct a global ML competition for molecular photostability. To this end, we leveraged automated block chemistry to prepare nine additional light-harvesting small molecules consisting of aryl and heteroaryl moieties and measured their photostability exactly as reported in the previous campaign (Fig. 3, see SI for synthesis details). The nine new molecules exhibited a broad range of T_{80} values (Fig. 3b). Ester bearing molecule (H) and extended bipyridine (I) were added to the training set to balance chemical diversity in the train/test split, leaving the remaining seven molecules to comprise the community-engaged test set. It should be noted that four light-harvesting molecules from the initial campaign were omitted from the dataset due to their T_{80} being too low to characterize.

3.2 Kaggle competition

With a training set (42 molecules) and test set (7 molecules) in hand, we hosted a competition on Kaggle, an established platform that has facilitated a diverse array of data science contests.[30] For each molecule in the dataset, we provided 144 chemical features calculated from RDKit and TDDFT as well as the corresponding SMILES string (Fig. 4).[31] Participants were allowed to use as many or as few of the provided features for their models, and they were free to generate additional features from the SMILES strings. For the test set, the same set of features was provided, and participants were tasked with predicting the associated T_{80} values. The submissions were evaluated based on the mean squared log error (MSLE) between the model predictions and experimental T_{80} values of the test set molecules. As a reference, the 4-feature SVR-RBF model from our previous campaign predicted the community test set photostability with an MSLE of 3.051. This high MSLE value stems

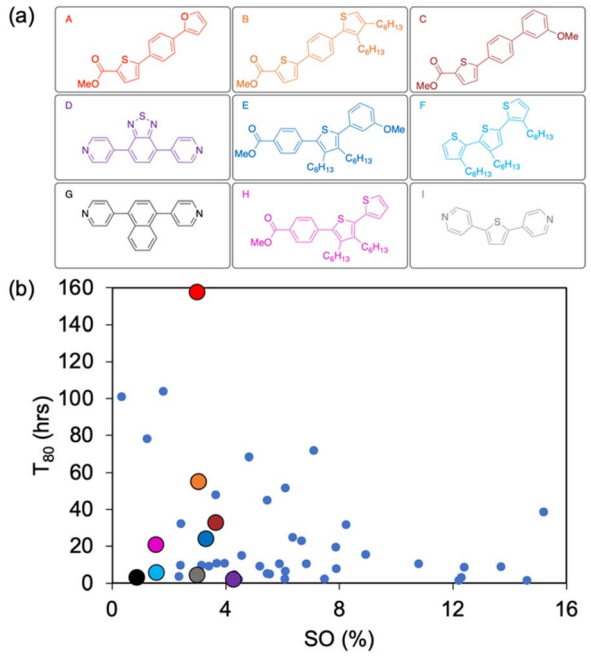

Fig. 3 (a) Nine new light-harvesting molecules were synthesized and characterized to serve as the community-engaged test set for the hackathon. (b) Plot of the T_{80} and spectral overlap (SO) of the nine new light-harvesting small molecules. Molecules H and I were incorporated into the train dataset to balance out the chemical diversity in the train/test split.

from our previous campaign's emphasis on interpretability rather than performance, as incorporating more features would have improved accuracy at the expense of losing interpretability and simple hypothesis generation.

We ran the competition between March 24, 2025 and April 26, 2025. At the conclusion, we received a total of 729 submissions from 522 entrants. It is important to note is that our competition only provided a total of $150 in prizes (compared to other competitions that gave out up to $50 000), yet we were still able to garner hundreds of participants in only one month. These outcomes suggest a strong community interest in participating in the scientific discovery process, even when engaging a small but practically important chemistry dataset.

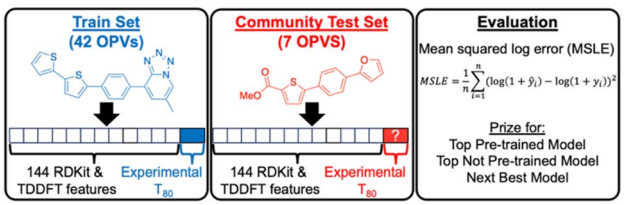

Fig. 4 Overview of the Kaggle competition. Participants were provided a training set of 42 molecules, which included the SMILES string, RDKit features, and TDDFT features and were tasked with predicting the T_{80} of seven test molecules. Model performance was evaluated on the MSLE between predictions and experimental T_{80} of the test set. We provided a total of $150 in prizes to incentivize participation.

Among the many excellent submissions, we highlight a few of the highest performing submissions. The top performing model, trained by a user who wished to remain anonymous, predicted the photostability of the test set with an MSLE of 1.026 (Fig. 5a). For this user's strategy, they first generated every possible RDKit feature to supplement the 144 features we provided. Then, they used the SelectFromModel class of scikit learn to select the top 35 features when trained on $\log(T_{80})$ rather than T_{80}. Finally, they found that the SVR with a linear kernel predicted the T_{80} with the lowest MSLE. Unlike our four-feature model, one of the features they included was "fr_pyridine," which is the number of pyridine rings in the molecule. Based on Fig. 3, all the bipyridines exhibited low T_{80} values, so their model correctly identified that as the number of bipyridine rings increases, the T_{80} decreases. This novel insight exemplifies the utility of a blinded test set that extends into a chemical space beyond that reported in the original published work.

The second-best model was trained by Valterri Valo (29 years old, Finland, Data Science/ML consultant), who like the top performing user, added over 100 RDKit features as well as 100 Morgan Fingerprints to the original dataset. Interestingly, Valterri augmented the data by adding methyl groups or replacing halogens on the original molecules while keeping the same T_{80} values to generate 74 new molecules. He ultimately chose 13 features and trained an XGBoost model to produce an MSLE of 1.208 (Fig. 5a).

The best pretrained model was implemented by Nikita Sharma (22 years old, India, Computer Science Undergraduate), who used the seyonec/ChemBERTa-zinc-base-v1 (ref. 32) model to embed each molecule, and trained a Ridge regressor on the embeddings to produce an MSLE of 1.760 (Fig. 5a). Interestingly, the pretrained model did not lead to the best results, which supports the observation that issues arise when model complexity outweighs the small dataset sizes that proliferate the chemical sciences.

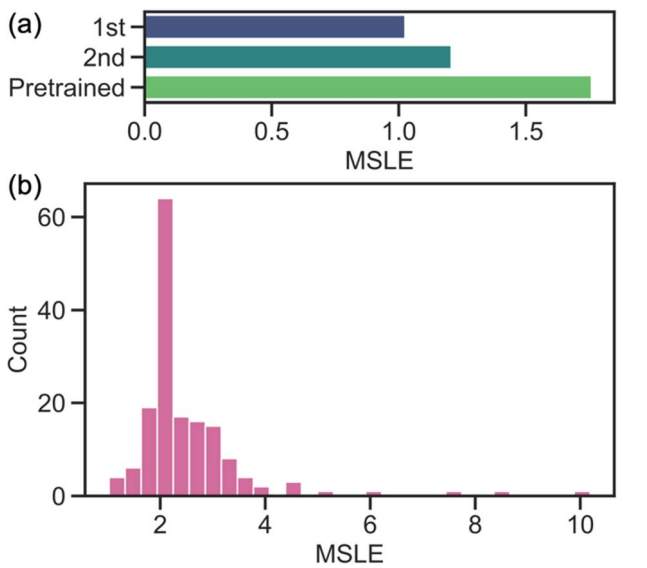

Fig. 5 (a) Results of the top performing models from the Kaggle competition. (b) Distrubution of scores for all submissions from the Kaggle competition.

Overall, our Kaggle competition was successful in engaging community scientists to improve upon our previously best-performing model. We were delighted to see that our participants used a variety of strategies such as using a log transformation of the T_{80} measurements or augmenting the data with chemically modified molecules. In addition, the participants uncovered new scientific trends in our dataset such as the negative correlation between number of pyridine rings and T_{80}.

3.3 Potential for generalization

As the science community enters a new, automation-centric era of innovation, the amount of high-quality data will only increase. To fuel consistent scientific breakthroughs, it is essential for ML techniques across scientific domains to keep pace with the experimental technologies.

We envision the community-engaged test set paradigm as a new direction for future community engagement and scientific outreach in the natural sciences. Instead of publishing the entire data set produced from a synthetic campaign, research groups could (for example) withhold a small test set (~10% of the data) for a public Kaggle completion. A small amount of work to clean up the dataset into Kaggle's ML-accessible format and clearly explain features and targets paves the way for broad engagement and ML-driven discovery in chemistry. Alternatively, as automatable block chemistry has increased accessibility to the synthesis and thus testing of new small molecules with a wide range of useful functions, the bar is lowered for researchers to generate post-publication datasets for community testing.

We provide a playbook for hosting your first Kaggle competition in the SI section. From a broad perspective, hosting a Kaggle competition would give attention to the initial publication and garner interest for future advances on the topic. In this way, the community-engaged test sets paradigm serves to democratize scientific discovery and align the objectives of experimental science and ML.

Provided the success of this Kaggle competition, it is interesting to consider future adaptations to our approach to further engage community interest. An obviously fruitful direction for future competitions is to integrate experimental design and validation more cohesively into the competition objective. For example, rather than asking participants to regress over the community-engaged test set, we could task competitors with training models and directly suggesting the next best experiments to run (*i.e.* the most informative molecule to synthesize on the grounds of exploration and exploitation of the design space). Subsequently, our automated synthesis robots could synthesize the suggested molecules and validate the hypotheses generated from the Kaggle competition participants. This future paradigm would concurrently allow the participants to directly contribute to the research and strengthen the chemical interpretation of the ML models. Such improved outreach strategies moving forward will aim to further increase democratization of ML and chemistry within the broader community.

4 Conclusions

By creating an experimental test set explicitly for hosting a community hackathon, we developed a new paradigm for the democratization of ML in chemistry. Unlike previous ML efforts that primarily involve one researcher performing ML on in-house data, our global hackathon enabled hundreds of researchers to tackle the grand challenge of predicting small molecule photostability, a fundamental task that lacks a complete molecular scale understanding in chemistry. The competition successfully uncovered new strategies such as log transformation of target data and data augmentation by functional group modification. Additionally, this competition engaged a diverse audience drawing from various countries, professions, and ages. By successfully bridging the gap between experimental chemistry and computer science, we envision that the creation of a community-engaged test set will become a standard for future community engagement and scientific progress in the natural and applied sciences.

Author contributions

Jason Wu: data curation, writing – original draft, writing – review & editing. David Friday: conceptualization, writing – review & editing. Changhyun Hwang: data curation, writing – review & editing. Seungjoo Yi: data curation, writing – review & editing. Tiara Torres-Flores: data curation, writing – review & editing. Martin Burke: conceptualization, writing – review & editing, supervision, project administration, funding acquisition. Ying Diao: conceptualization, writing – review & editing, supervision, project administration, funding acquisition. Charles Schroeder: conceptualization, writing – review & editing, supervision, project administration, funding acquisition. Nicholas Jackson: conceptualization, writing – review & editing, supervision, project administration, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

The datasets used for this article can be found at: https://www.kaggle.com/competitions/molecular-machine-learning. The code repository can be found at the following link: https://github.com/TheJacksonLab/Community-Engaged-Test-Sets as well as Zenodo *via* the following DOI: https://doi.org/10.5281/zenodo.17632410.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00424a.

Acknowledgements

This work was supported by the Molecule Maker Lab Institute, an AI Research Institutes program supported by the US National Science Foundation under grant no. 2019897 and grant no. 2505932.

Notes and references

1 V. Zuin Zeidler, *Science*, 2024, **384**, eadq3537.
2 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *J. Chem. Inf. Model.*, 2024, **64**, 9–17.
3 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
4 S. Kim, C. M. Schroeder and N. E. Jackson, *ACS Polym. Au*, 2023, **3**, 318–330.
5 A. J. Gormley and M. A. Webb, *Nat. Rev. Mater.*, 2021, **6**, 642–644.
6 C. M. Collins, L. M. Daniels, Q. Gibson, M. W. Gaultois, M. Moran, R. Feetham, M. J. Pitcher, M. S. Dyer, C. Delacotte, M. Zanella, C. A. Murray, G. Glodan, O. Pérez, D. Pelloquin, T. D. Manning, J. Alaria, G. R. Darling, J. B. Claridge and M. J. Rosseinsky, *Angew. Chem., Int. Ed.*, 2021, **60**, 16457–16465.
7 J. Fang, M. Xie, X. He, J. Zhang, J. Hu, Y. Chen, Y. Yang and Q. Jin, *Mater. Today Commun.*, 2022, **33**, 104900.
8 Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins and Q. Zhou, *J. Chem. Inf. Model.*, 2021, **61**, 3197–3212.
9 N. H. Angello, D. M. Friday, C. Hwang, S. Yi, A. H. Cheng, T. C. Torres-Flores, E. R. Jira, W. Wang, A. Aspuru-Guzik, M. D. Burke, C. M. Schroeder, Y. Diao and N. E. Jackson, *Nature*, 2024, **633**, 351–358.
10 S. Alem, S. Wakim, J. Lu, G. Robertson, J. Ding and Y. Tao, *ACS Appl. Mater. Interfaces*, 2012, **4**, 2993–2998.
11 N. H. Angello, V. Rathore, W. Beker, A. Wolos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski and M. D. Burke, *Science*, 2022, **378**, 399–405.
12 W. Wang, N. H. Angello, D. J. Blair, T. Tyrikos-Ergas, W. H. Krueger, K. N. S. Medine, A. J. LaPorte, J. M. Berger and M. D. Burke, *Nat. Synth.*, 2024, **3**, 1031–1038.
13 N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, *Nat. Chem.*, 2021, **13**, 505–508.
14 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 and J. W. Bode, *Sci. Adv.*, 2023, **9**, eadj2314.
15 W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke and B. A. Grzybowski, *J. Am. Chem. Soc.*, 2022, **144**, 4819–4827.
16 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. Liu, X. Tang, M. Mamada, W. Wang, T. Tsagaantsooj, C. Lavigne, R. Pollice, T. C. Wu, K. Hotta, L. Bodo, S. Li, M. Haddadnia, A. Wołos, R. Roszak, C. T. Ser, C. Bozal-Ginesta, R. J. Hickman, J. Vestfrid, A. Aguilar-Granda, E. L. Klimareva, R. C. Sigerson, W. Hou, D. Gahler, S. Lach, A. Warzybok, O. Borodin, S. Rohrbach, B. Sanchez-Lengeling, C. Adachi, B. A. Grzybowski, L. Cronin, J. E. Hein, M. D. Burke and A. Aspuru-Guzik, *Science*, 2024, **384**, eadk9227.
17 X. Wang, Y. Huang, X. Xie, Y. Liu, Z. Huo, M. Lin, H. Xin and R. Tong, *Nat. Commun.*, 2023, **14**, 3647.
18 D. O. Shkil, A. A. Muhamedzhanova, P. I. Petrov, E. V. Skorb, T. A. Aliev, I. S. Steshin, A. V. Tumanov, A. S. Kislinskiy and M. V. Fedorov, *Molecules*, 2024, **29**, 1826.
19 Nomad2018 Predicting Transparent Conductors, https://kaggle.com/nomad2018-predict-transparent-conductors, accessed 21 July 2025.
20 Novozymes Enzyme Stability Prediction, https://kaggle.com/novozymes-enzyme-stability-prediction, accessed 30 May 2025.
21 Predicting Molecular Properties, https://kaggle.com/champs-scalar-coupling, accessed 30 May 2025.
22 Bristol-Myers Squibb – Molecular Translation, https://kaggle.com/bms-molecular-translation, accessed 30 May 2025.
23 Home – Prediction Center, https://predictioncenter.org/, accessed 21 July 2025.
24 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
25 J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse and M. D. Burke, *Science*, 2015, **347**, 1221–1226.
26 J. W. Lehmann, D. J. Blair and M. D. Burke, *Nat. Rev. Chem.*, 2018, **2**, 0115.
27 M. Trobe and M. D. Burke, *Angew. Chem., Int. Ed.*, 2018, **57**, 4192–4214.
28 D. J. Blair, S. Chitti, M. Trobe, D. M. Kostyra, H. M. S. Haley, R. L. Hansen, S. G. Ballmer, T. J. Woods, W. Wang, V. Mubayi, M. J. Schmidt, R. W. Pipal, G. F. Morehouse, A. M. E. Palazzolo Ray, D. L. Gray, A. L. Gill and M. D. Burke, *Nature*, 2022, **604**, 92–97.
29 T. Tyrikos-Ergas, S. Agiakloglou, A. J. LaPorte, W. Wang, C.-K. Chan, C. E. Wells, C. K. Rakowski, R. I. Hammond, J. Qiu, J. D. Raymond, T. Vieira, J. Limanto, M. N. Feiglin, D. J. Blair and M. D. Burke, *Angew. Chem., Int. Ed.*, 2025, **64**, e202509974.
30 Kaggle, https://www.kaggle.com/, accessed 24 July 2025.
31 Molecular Data Machine Learning, https://kaggle.com/molecular-machine-learning, accessed 25 July 2025.
32 S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.