Autonomous, multi-property-driven molecular discovery: from predictions to measurements and back

Authors: Brent A. Koscher[1]†, Richard B. Canty[1]†, Matthew A. McDonald[1]†, Kevin P. Greenman[1], Charles J. McGill[1], Camille L. Bilodeau[1], Wengong Jin[2], Haoyang Wu[1], Florence H. Vermeire[1], Brooke Jin[1], Travis Hart[1], Timothy Kulesza[1], Shih-Cheng Li[1], Tommi S. Jaakkola[3], Regina Barzilay[3], Rafael Gómez-Bombarelli[4], William H. Green[1], Klavs F. Jensen[1]*

Affiliations:

[1]Department of Chemical Engineering, Massachusetts Institute of Technology; Cambridge, USA.

[2]Broad Institute of MIT and Harvard; Cambridge, USA

[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; Cambridge, USA.

[4]Department of Materials Science and Engineering, Massachusetts Institute of Technology; Cambridge, USA.

*Corresponding author. Email: kfjensen@mit.edu

†These authors contributed equally to this work.

Abstract:

A closed-loop, autonomous molecular discovery platform driven by integrated machine learning tools was developed to accelerate the design of molecules with desired properties. Two case studies are demonstrated on dye-like molecules, targeting absorption wavelength, lipophilicity, and photo-oxidative stability. In the first, the platform experimentally realized 312 unreported molecules across three automatic iterations of molecular design-make-test-analyze cycles while exploring the structure–function space of four rarely reported scaffolds. In each iteration, the property-prediction models which guided the exploration learned the structure–property space of diverse inexpensive scaffold derivatives realized through using multi-step syntheses. Conversely, the second study exploited property models trained on a chemical space with pre-existing examples to discover 6 top-performing molecules within the structure-property space. By closing the molecular discovery cycle of prediction, synthesis, measurement, and model retraining, the platform demonstrates the potential for integrated platforms to automatically understand a local chemical space and discover functional molecules.

Main Text:

The development of function-focused molecules for applications in medicine (*1*, *2*), materials (*3*–*5*), and sustainability (*6*, *7*) directly depends on the rate of molecular discovery. Molecular discovery consists of a series of design-make-test-analyze (DMTA) cycles in which molecular candidates are iteratively proposed and verified towards a functional molecule. Ongoing efforts to accelerate individual components of the DMTA cycle include developments in predictive modeling and advancements in chemical automation. Candidate structures can be quickly generated using molecular generation workflows based on genetic algorithms(*8*, *9*), reinforcement learning (*10*–*12*), or conditional generation (*13*, *14*). Plausible reaction pathways to realize those candidates can be proposed using retrosynthetic planning packages, such as ASKCOS (*15*, *16*), Synthia (*17*, *18*), IBM RXN (*19*), and AiZynth (*20*). The properties of candidates can be predicted with quantitative structure-activity relationship models or statistical models such as Chemprop, a message-passing neural network (*21*). To execute predicted reactions, chemical automation advances have developed reconfigurable high-throughput platforms (*16*, *22*, *23*) with hardware to access increasingly diverse reaction conditions (*24*) to discover novel chemical transformations (*25*, *26*), optimize reactions (*27*, *28*), and develop functional materials (*5*, *7*).

An integrated platform that directly leverages the capabilities and advantages of chemical prediction and automation advancements would be capable of autonomous closed-loop molecular discovery across general chemical spaces. Previous feedback-guided chemistry synthesis platforms have been designed for enzyme-assisted carbohydrate synthesis (*29*), iterative cross-coupling reactions (*30*, *31*), and reaction optimization (*23*, *32*–*34*); however, these are narrow and well-defined chemical spaces. These demonstrations illustrate the benefits of directly connecting some of the DMTA components together, but all of the DMTA stages have not been coupled into a single automated platform (*35*). Fully coupling the DMTA cycle requires a platform to propose diverse and useful candidates, recommend reaction pathways, synthesize and isolate selected candidates, characterize isolated molecules, and refine model predictions based on experimental results (Figure 1A), all automatically without human intervention.

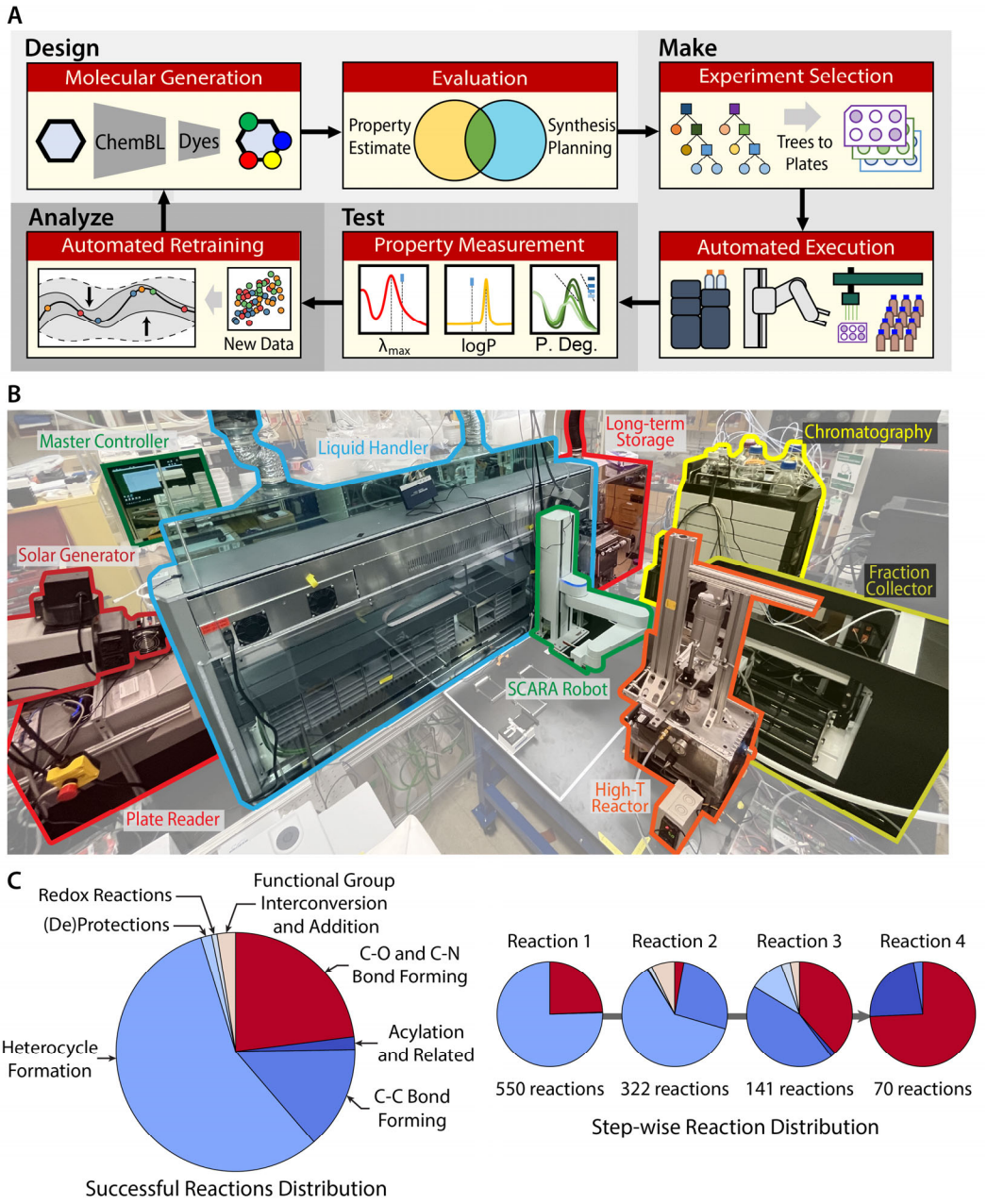

Figure 1. Overview of the developed integrated platform and reactions predicted and successfully realized platform. **(A)** Illustrations of the key workflow components of the integrated platform and how each component fits into the DMTA cycle, in which the properties tested are wavelength of maximum absorption (λ_{max}), partition coefficient (logP), and rate of photo-oxidative degradation (P. Deg.). **(B)** Layout of the physical hardware of the experimental platform. **(C)** Breakdown of the different reaction classes, as classified by ASKCOS, executed by the platform automatically during the exploration case study presented. The four smaller pie charts show the change in the reaction distribution throughout the different steps of the executed reaction pathway. The first reaction executed tends to be a heterocycle formation and the final reaction is often derivatization by forming a C–C (dark blue) C–N, or C–O (both red) bond. Not all selected targets require the same number of reactions in the predicted reaction pathway, leading to the distribution of reaction numbers (a detailed breakdown of reactions is given in the supplementary materials fig. S5).

We demonstrate such a platform (Figure 1B) and two molecular discovery use-cases: (1) exploration of unknown chemical spaces and (2) exploitation of known chemical spaces. We engineered a flexible platform that is able to accept and adjust workflows as-needed, automatically, without human intervention beyond providing consumables and error recovery. Platform flexibility was crucial to execute multi-step reaction pathways containing a breadth of reaction classes (Figure 1C) to realize candidates from molecular generation. In the case studies, we demonstrate the platform capabilities by targeting molecular dyes, a model system with diverse chemistry and non-trivial molecular properties. The platform attempted over 3,000 reactions with >1,000 yielding the predicted reaction product, completing multi-step reaction pathways for 318 molecules without CAS registry numbers. For each of these molecules, the absorption spectrum, water-octanol partitioning, and photo-oxidative stability were automatically predicted and measured.

Front-end predictions.

Before experiments can be executed, the platform must: generate candidates, predict candidate performance, plan reaction pathways, and select candidates to synthesize (Figure 1A, design). The individual models for each necessary prediction have been demonstrated previously and can be integrated to form an experimental plan. To generate candidate structures the platform uses a hierarchical graph-completion model (*36*). Compared to other molecular generation approaches (*8–14*), graph-completion proposes more conservative structures. The graph-completion model takes an input scaffold and completes the structure with learned motifs, encoding and decoding full motifs at once rather than atom-by-atom. The generative model was first pre-trained on ChEMBL (*37*) to learn general organic chemistry rules and then fine-tuned on a curated set of dye molecules (SM2) to complete input scaffolds with diverse dye-like structural motifs.

Synthesis pathways for generated candidates are planned using ASKCOS (*15*, *16*), which makes template-based retrosynthesis recommendations using reaction templates extracted from large databases such as Reaxys or Pistachio. The variety of reactions covered by ASKCOS enables reaction pathways to be planned for the diverse structures generated during molecular generation. The pathway planner builds synthesis pathways and discards pathways that do not end in buyable reagents within the allotted computation time. While reaction planning is time-intensive, pre-filtering generated candidates with a synthesizability metric (*38–41*) does not increase reaction planning throughput (fig. S2). Instead, retrosynthesis planning serves to determine which generated candidates are considered synthesizable. Reaction conditions are recommended with a template-free neural network model (*42*). To check for and correct common problematic condition recommendations (failing to recommend all required reagents for well-known reactions or recommending outdated catalysts when modern and improved catalysts are known), we defined a set of rules for the 100 most popular chemical transformations (SM3). Synthesis pathways requiring platform-inaccessible conditions or more than four reactions are discarded, resulting in 10–20% of generated molecules having platform-executable synthesis pathways, depending on the input scaffold.

The dye-like properties of the remaining synthesizable candidates were evaluated using a set of Chemprop property prediction models: wavelength of maximum absorption, water-octanol partitioning, and photo-oxidative degradation rate. Chemprop models are lightweight, so an ensemble can be automatically retrained as data are collected (SM4), and fast, so predictions can be made from an ensemble of models (*21*, *43*). The ensemble variance is used as a proxy for model

confidence to inform molecule selection. Each of the models were initially trained on datasets of different sizes: 21,000 experimental absorption maxima from literature, 23,000 experimental partition coefficient values from literature, and 85 experimental photo-oxidative degradation rates measured in-house. Due to the small size of the third dataset, we used a random forest for the photodegradation rate model (SM4).

To select candidates to synthesize from a large number of options we consider: predicted property values, prediction ensemble variances, molecular diversity, cost of buyable starting materials, number of reaction steps, and constraints on batch-wise executable reaction conditions. These factors were leveraged in exploration to select practical molecular targets to quickly improve property prediction models and in model exploitation to select molecules with optimized predicted absorption, lipophilicity, and photo-oxidative stability. The same factors are evaluated in both demonstration use-cases but with different relative weightings, allowing the platform to probe chemical spaces that will meet exploration or exploitation goals. The weighting details in each use-case are described below and in the supporting information (SM5).

Experimental execution and automation.

Platform-selected reaction pathways from the front-end predictions are automatically translated into synthesis and characterization workflows to be experimentally executed. When candidates are selected, the platform considers compatible reaction pathways when batching tasks for well plates (jobs; Figure 2A). Predictions for the synthetic pathways and properties of interest for each well plate are converted into high-level goal-oriented tasks (*e.g.*, reaction preparation, Figure 2) with specific details on a well-by-well basis. Four independent workers execute workflows in parallel to accomplish these tasks, and a master controller orchestrates tasks over a local network (Figure 2B). Currently, the four workers are a liquid handler, high-performance liquid chromatography (HPLC) instrument, robotic arm, and the special processes unit which manages storage, a high-temperature reactor, a plate-reader, and custom hardware.

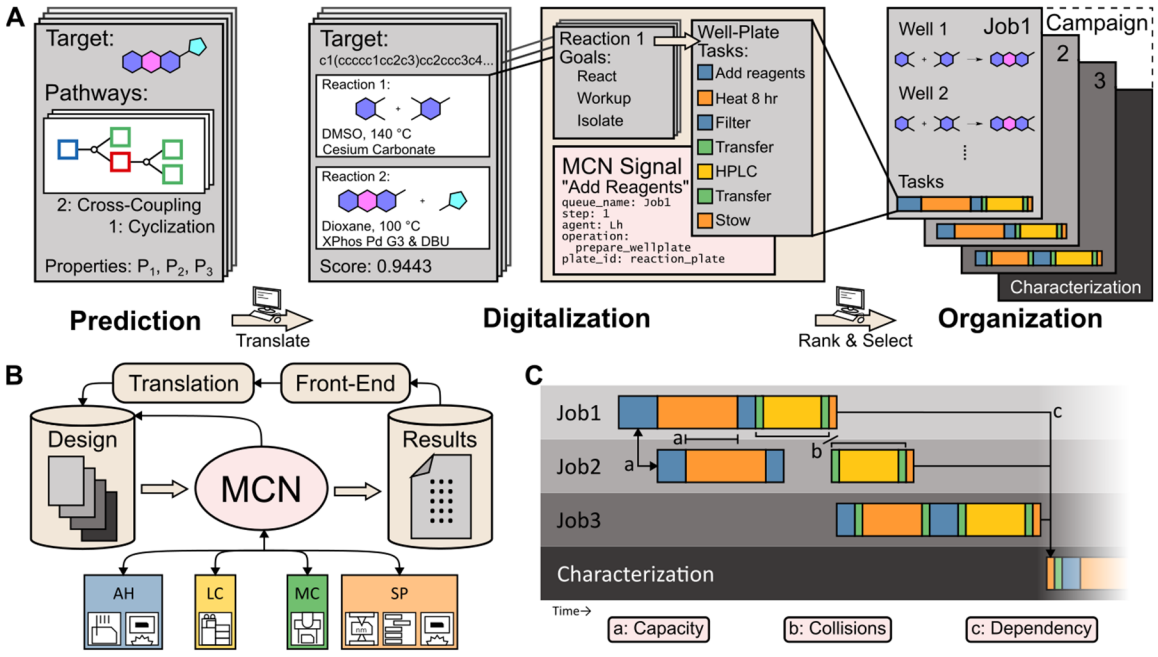

Figure 2. A schematic of the organization of jobs and their coordination across the platform. **(A)** Predicted reaction pathways are translated into a set of multi-step goals and organized into lists of tasks at the well plate level; the jobs required are collated into a campaign. **(B)** The master control

network (MCN), databases, and machine learning mediate the information-flow from a campaign to the platform workers to experimental data. Worker legend: the liquid handler and heater-shaker reactors (AH, blue), the HPLC (LC, yellow), the master controller and robotic arm (MC, green), and the plate-reader, storage unit, and high-temperature reactor (SP, orange). **(C)** The initial scheduling of jobs in accordance with the scheduling algorithm: honoring the capacity of a worker to perform multiple tasks at once, the collision of resources between groups of tasks, and the dependencies between jobs.

The workers are supported by two databases: an experimental design database for executing operations and an experimental results database for recording synthesis and property information. Workers access the design database, which manages platform resources, to make real-time decisions on how to accomplish their assigned tasks, and then record outcomes and molecular properties in the results database (SM7). The master controller avoids well plate collisions between workers and adjusts to real-time changes in the workflows (Figure 2C; further details on the control system and performance metrics are reported in fig. S6–7). Other control architectures, such as those implemented in Chemputer (*44*, *45*) and ChemOS (*46*), were considered; however, the need to operate in parallel with 96-well plates, handle characterization, and modify workflows on-the-fly motivated the design of a custom-built architecture. The result is a control system capable of performing multiple, unrelated jobs in parallel and only requiring human intervention for non-automatable error-recovery (SM7).

Modular combinations of synthesis, isolation, and characterization hardware automatically execute multi-step reaction pathways and measure properties on an as-needed basis. Resources needed by each worker to accomplish their tasks are coordinated by the master controller without pausing on-going operations, except human intervention to resupply consumable materials. A liquid handler robotic arm moves well plates within the liquid handler and collaborates with an intersystem robotic arm to move well plates between instruments. Reactions are prepared and executed in 96-well plates for ease of parallelization compared to flow chemistry and to have adequate material for multi-step synthesis compared to other higher-density well plate formats (*47–49*). Automation of specialized reaction conditions was achieved with custom platform devices (SM11). After reactions have been executed and prepared for HPLC, the reactions are analyzed by analytical HPLC-MS and identified products are isolated by semi-preparative HPLC (HPLC automation was achieved with a vendor-supplied application programming interface, further discussed in SM9). Pre-defined characterization and data processing routines (SM9-10) are automatically performed on isolated products and the data added to the results database to be used when retraining the Chemprop models. For the demonstration chemistry presented, the automatic characterizations included the optical absorption spectra via a plate reader, partition coefficients extracted from calibrated HPLC retention times, and photo-oxidative degradation rates measured in a custom-built solar degradation device (SM11). Furthermore, additional properties can be measured with new hardware or workflows, such as biological assays, and parallel implementation allows new hardware to work synchronously with existing hardware.

Exploration case study.

Molecular exploration of chemical spaces beyond current datasets aims to identify productive spaces and determine whether molecules within those spaces have a set of desired properties. Property predictions in these unexplored chemical spaces are uncertain because relevant examples are not present in their training sets to ground predictions, complicating molecule selection. To

improve the property-prediction models, the platform first proposes a large number of candidates, experimentally realizes several low-cost examples, and uses that information to anchor future predictions. This approach allows the models to understand the chemical space without the platform investing excessive time and resources into synthetically complex, predicted high-performing molecules when simpler molecules are available. With models making faithful predictions, a further evaluation can narrow down the potential chemical space to a small number of hits for further exploration. This approach to molecular discovery is typical of identifying promising hit chemical spaces in hit-to-lead workflows.

To demonstrate molecular exploration, we task the platform to explore the properties of unexplored heterocyclic dye-like molecules. To find candidate scaffolds we filtered ring structures that appear in the ZINC database (SM1) since many molecular dye families are based around conserved, conjugated heterocyclic scaffolds (analogous to xanthenes, coumarins, *etc.*). Although cyclization reactions are critical to form heterocyclic structures, they are a challenge to retrosynthesis planners because multiple bonds form simultaneously (*50*). We manually inspected ASKCOS recommended cyclization reactions for established ring-forming templates and selected four rarely reported heterocyclic structures to serve as exploration scaffolds (Figure 3A). The graph-completion model generated a large set of candidates by allowing any combination of C–H bonds on each scaffold to be functionalized (fig. S1). For molecules with recommended reaction pathways, the generality of ASKCOS leads to several reaction classes recommended as productive derivatization strategies. Some familiar reaction classes, such as palladium-catalyzed coupling reactions, are often recommended for their utility, but less common classes, such as aldol or Knoevenagel condensations and aminations, are also recommended (see Figure 1C, fig. S5). Initial property-prediction models, trained on experimental datasets extracted from literature or measured in-house, with no examples of our scaffolds of interest, were used to predict performance across the candidates (Figure 3B, round 0). For exploration, the candidates that are valued are affordable, easy to synthesize, have properties that approach our target property space (redder absorbing, lipophobic, and photo-oxidatively stable), and have sizable prediction model ensemble variance (fig. S3, SM5). This approach favors diverse structural functionalizations that move the derivatives' predicted properties towards our desired property space, as opposed to random exploration.

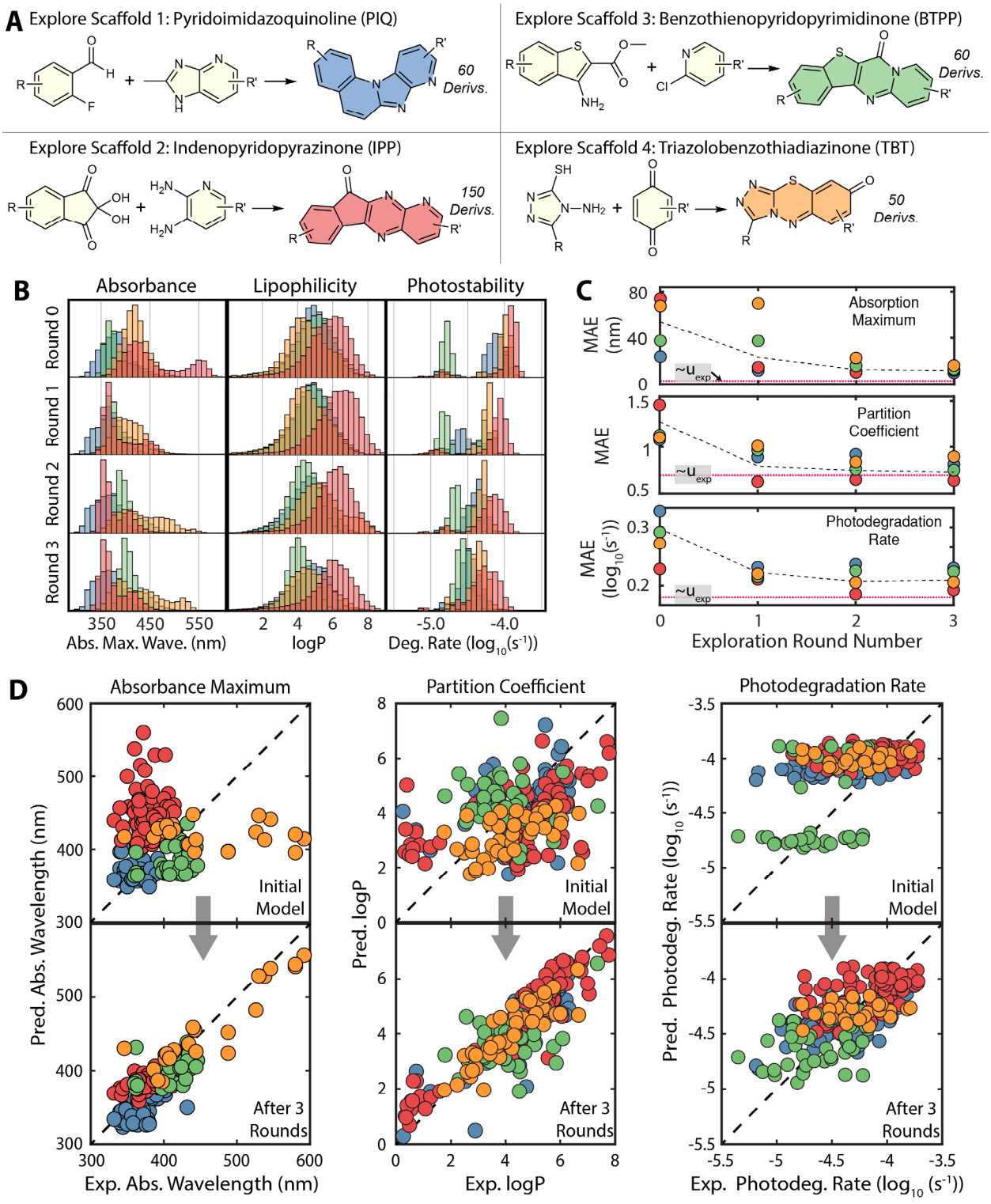

Figure 3. Model exploration case study. **(A)** Key cyclization reactions needed to access the ring-containing scaffolds explored in the exploration case study. The shaded colors correspond to the remaining sub-figures. **(B)** Changes in the predicted property space using the general property model (Round 0) and following each round of exploration (Round 1–3). **(C)** Changes in the models' mean absolute errors (MAEs) for each version of the absorption maximum, partition coefficient, and photo-oxidative degradation rate models evaluated using cross-validation. Colors

correspond to scaffolds in (A) and dashed black lines correspond to full dataset MAE values. The experimental uncertainty (u_{exp}) is shown in each plot as a magenta dashed line. **(D)** Comparison of the model predictions and experimental values from the initial general models and following three rounds of exploration for each of the property-prediction models of interest.

By iteratively realizing and measuring diverse derivatives, the platform learns about the structure–property space for the scaffolds of interest, improving the property prediction quality. Initial property model prediction agreement with experimental values was subpar, with MAEs of 52 nm, 1.3, and 0.30 $\log_{10}(s^{-1})$ for wavelength of maximum absorbance, water/octanol partition coefficient, and rate of photo-oxidative degradation respectively (Figure 3C, round 0). One round of exploration provided a few examples of each scaffold (10–45 depending on the scaffold) which grounded the property models, shifting the predicted property space (Figure 3B) and improving the models' MAEs (Figure 3C). To investigate the impact of adding additional data, the platform performed two additional rounds of chemical exploration, using the same selection criteria but updating property models after each round. While the degree of improvement varied, the impact of new data diminished for all three property models with each round of exploration for all four scaffolds (Figure 3C). After three rounds of exploration—adding 110, 140, and 90 molecules in each round, respectively—the property models further improved for all scaffolds (Figure 3C–D) with the overall MAEs decreasing to 11 nm, 0.7, and 0.21 $\log_{10}(s^{-1})$. Across the scaffolds, 720 molecules were selected to be realized with an average pathway length of 2.6 reactions—requiring the platform to process 37 reaction well plates and 181 additional well plates for workup and characterization—resulting in 312 unreported molecules realized and characterized. For each of the scaffolds, the predicted property spaces (Figure 3B) shift towards ground-truth values as the property models learn about the structure-property spaces (Figure 3D).

Although the models' predictions improved for each scaffold, there was little change in predictive ability across more general chemical spaces, demonstrating the local nature of these structure–property spaces. By verifying a handful of experimental examples to ground model predictions, candidate scaffolds that perform well can be more confidently considered for further examination. Initial ungrounded property models inconsistently overpredict (such as IPP absorption maxima or TBT photodegradation rate, Figure 3) or underpredict (such as TBT or BTPP absorption maxima, Figure 3) the real performance of the candidate scaffolds depending on the local structure–property space. Even though dye-like molecules served as our model system due to properties that are learnable from structure, the initial property models did not extrapolate well to the unexplored selected scaffolds. This exemplifies the challenge of only relying on generative models and property predicting models to directly guide molecular discovery in unknown chemical spaces without experimental verification. To meet this challenge, we demonstrate a platform that automatically designs and executes experiments without human intervention, beyond providing consumables. Such an integrated platform can verify chemical spaces of interest by first focusing on generally improving the property predicting models with relevant local examples before attempting challenging and expensive molecular targets.

Exploitation case study.

Property predicting models can be exploited to selectively realize molecules predicted to be top-performers in chemical spaces containing pre-existing data. Model exploitation naturally follows molecular exploration and is analogous to the transition from hit identification to lead optimization. Exploitation can be achieved on the same integrated platform as exploration by

reweighting how candidates are valued. Predicted top-performers are typically more synthetically complex and expensive, so rather than realizing several low-cost candidates only a few confidently predicted top-performers are selected.

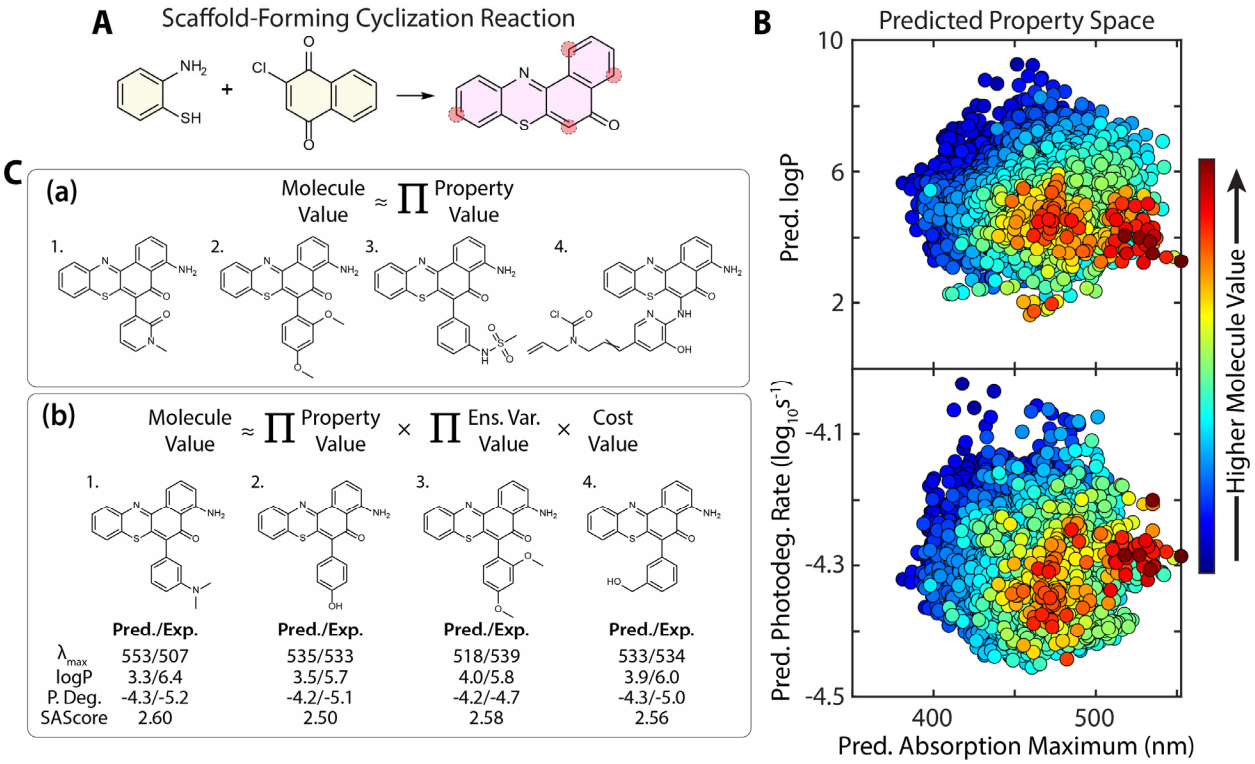

Figure 4. Model exploitation case study. **(A)** Key cyclization reaction needed to access the ring-containing scaffold explored in the exploitation case study. Sites that were accessible with commercially available reagents are highlighted in red. **(B)** Predicted property space of the generated candidates for partition coefficient versus absorption maximum (top) and photodegradation rate versus absorption maximum (bottom). Color of each point reflects the candidate value to the exploitation workflow. **(C)** Two approaches to valuing candidates, (a) considering only the predicted property values and (b) considering the predicted property value, ensemble variances, and cost. The top four most valuable molecules are shown for each valuation method.

To demonstrate model exploitation, we chose to optimize a benzophenothiazinone (BPT) scaffold (Figure 4A), a promising scaffold with available starting materials, acceptable cyclization yields, a modest number of existing data points, and structural similarities to other common dye families. We tasked the platform to synthesize molecules with maximized absorption wavelengths, minimized partition coefficients, and minimized photo-degradation rates simultaneously (Figure 4B). While tempting to only consider predicted property values when evaluating candidates (Figure 4C), molecular generation is known to propose unusual structures (*41*) that may be beyond the scope of property prediction models. By considering model ensemble variance, synthetic accessibility score (*38*), and total reagent cost during candidate selection, the influence of out-of-scope molecular substructures can be minimized (Figure 4C), thus valuing realistic candidates and shifting values away from the optima defined by properties alone (Figure 4B, fig. S4). The relative

importance of these considerations can be weighted depending on acceptable tolerances to unusual chemical structures and prediction uncertainties in the candidate space.

The reaction pathways for the generated BPT derivatives all contain an established cyclization reaction involving the condensation of a halogenated naphthoquinone and a substituted 2-aminobenzenethiol (Figure 4A). Several potentially functionalizable sites on BPT were not accessible within a feasible number of reactions or from purchasable chemicals (Figure 4A). From the generated derivatives, the platform automatically selected 10 of the highest-valued derivatives (Figure 4B) to be realized, a selection of which are shown in Figure 4C (remaining structures in SM13). Due to the air sensitivity of 2-aminobenzenethiol derivatives, the platform was unable to realize some candidates that were attempted (fig. S13). For the 6 candidates that were successfully realized and characterized, the predicted absorption maxima showed good agreement with experimental values (11 nm MAE), consistently underpredicted logP values by 2 log-units, and overpredicted the photodegradation rate by 0.5 log-units. The BPT scaffold is structurally similar to phenoxazine dyes such as Nile blue, which was investigated for targeted delivery of photogenerated singlet oxygen (*51*), but is >10 fold more photo-oxidatively stable than Nile blue. Moreover, the molecules are fluorescent (fig. S9), suggesting the family might have additional photochemical activity to explore. The same integrated platform was able to both explore unknown structure–property spaces searching for hit structures and exploit known structure–property spaces optimizing lead structures.

Conclusion.

Discovery of molecules with targeted properties has historically been driven by manual experimentation, chemists' intuition, and an understanding of mechanisms and first principles; we have co-opted recently developed tools that assist chemists with the DMTA cycle to automatically explore chemical space and exploit known chemical structures. In the classic approach to the molecular discovery cycle, a hit molecule is altered to understand structure-function relationships; the platform we demonstrate does this without manual experimentation. The platform proposed, synthesized, and characterized 318 unreported dye-like molecules, spread across four exploration scaffolds and one exploitation scaffold. Dye-like molecules represent a compelling property space and chemical space to explore and exploit; a variety of chemical transformations are needed to access structures with a range of properties. The modular architecture of the integrated platform we demonstrate allows additional modules and workflows to be added as-needed for different properties, such as biological activity, without impacting existing platform capabilities. Future iterations of the platform will benefit from improvements in predictive capabilities, particularly reaction fidelity, condition recommendation, molecular generation, and analytical tools, such as structural elucidation via HPLC detectors. The on-going development of closed-loop integrated platforms is a promising path to continue accelerating molecular discovery.

References and Notes

1. K. R. Campos, P. J. Coleman, J. C. Alvarez, S. D. Dreher, R. M. Garbaccio, N. K. Terrett, R. D. Tillyer, M. D. Truppo, E. R. Parmee, The importance of synthetic chemistry in the pharmaceutical industry. *Science (80-.).* **363** (2019), doi:10.1126/science.aat0805.
2. T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, S. K. Saikin, Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces*. **11**, 24825–24836 (2019).
3. M. Abolhasani, E. Kumacheva, The rise of self-driving labs in chemical and materials

sciences. *Nat. Synth.* (2023), doi:10.1038/s44160-022-00231-0.

4. E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev, B. Maruyama, Autonomous experimentation systems for materials development: A community perspective. *Matter*. **4**, 2702–2726 (2021).

5. B. P. MacLeod, F. G. L. Parlane, A. K. Brown, J. E. Hein, C. P. Berlinguette, Flexible automation accelerates materials discovery. *Nat. Mater.* **21**, 722–726 (2022).

6. J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C. J. Eom, J. R. Lunger, D. J. Zheng, R. R. Rao, S. Muy, J. C. Grossman, K. Reuter, R. Gómez-Bombarelli, Y. Shao-Horn, Human- and machine-centred designs of molecules and materials for sustainability and decarbonization. *Nat. Rev. Mater.* **7**, 991–1009 (2022).

7. S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, C. J. Brabec, Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Adv. Mater.* **32** (2020), doi:10.1002/adma.201907801.

8. E. W. Lameijer, T. Bäck, J. N. Kok, A. P. Ijzerman, Evolutionary algorithms in drug design. *Nat. Comput.* **4**, 177–243 (2005).

9. Y. Kwon, S. Kang, Y. S. Choi, I. Kim, Evolutionary design of molecules based on deep learning and a genetic algorithm. *Sci. Rep.* **11**, 1–11 (2021).

10. M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, 1–15 (2018).

11. Z. Zhou, S. Kearnes, L. Li, R. N. Zare, P. Riley, Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **9**, 1–10 (2019).

12. S. K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, K. J. Thomas, S. Blackburn, C. W. Coley, J. Tang, S. Chandar, Y. Bengio, Learning to navigate the synthetically accessible chemical space using reinforcement learning. *37th Int. Conf. Mach. Learn. ICML 2020*. **PartF16814**, 3626–3637 (2020).

13. S. Kang, K. Cho, Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **59**, 43–52 (2019).

14. P. C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, E. J. Bjerrum, Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2**, 254–265 (2020).

15. ASKCOS (available at https://github.com/ASKCOS/).

16. C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. John Hart, T. F. Jamison, K. F. Jensen, A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science (80-.).* **365** (2019), doi:10.1126/science.aax1566.

17. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P.

Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem*. **4**, 522–532 (2018).

18. B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, Computational planning of the synthesis of complex natural products. *Nature*. **588**, 83–88 (2020).

19. P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).

20. S. Genheden, A. Thakkar, V. Chadimová, J. L. Reymond, O. Engkvist, E. Bjerrum, AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **12**, 1–9 (2020).

21. 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).

22. S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, L. Cronin, Organic synthesis in a modular robotic system driven by a chemical programming language. *Science (80-.).* **363** (2019), doi:10.1126/science.aav2211.

23. A. C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, T. F. Jamison, Reconfigurable system for automated optimization of diverse chemical reactions. *Science (80-.).* **361**, 1220–1225 (2018).

24. S. Hessam, M. Craven, A. I. Leonov, G. Keenan, L. Cronin, A universal system for digitization and automatic execution of the chemical synthesis literature. *Science (80-.).* **370**, 101–108 (2020).

25. A. F. Zahrt, Y. Mo, K. Y. Nandiwale, R. Shprints, E. Heid, K. F. Jensen, Machine-Learning-Guided Discovery of Electrochemical Reactions. *J. Am. Chem. Soc.* **144**, 22599–22610 (2022).

26. D. Caramelli, J. M. Granda, S. H. M. Mehr, D. Cambié, A. B. Henson, L. Cronin, Discovering New Chemistry with an Autonomous Robotic Platform Driven by a Reactivity-Seeking Neural Network. *ACS Cent. Sci.* **7**, 1821–1830 (2021).

27. J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams, A. G. Doyle, A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **144**, 19999–20007 (2022).

28. M. Christensen, F. Adedeji, S. Grosser, K. Zawatzky, Y. Ji, J. Liu, J. A. Jurica, J. R. Naber, J. E. Hein, Development of an automated kinetic profiling system with online HPLC for reaction optimization. *React. Chem. Eng.* **4**, 1555–1558 (2019).

29. T. Li, L. Liu, N. Wei, J. Y. Yang, D. G. Chapla, K. W. Moremen, G. J. Boons, An automated platform for the enzyme-mediated assembly of complex oligosaccharides. *Nat.*

Chem. **11**, 229–236 (2019).

30. T. C. Wu, A. Aguilar-Granda, K. Hotta, S. A. Yazdani, R. Pollice, J. Vestfrid, H. Hao, C. Lavigne, M. Seifrid, N. Angello, F. Bencheikh, J. E. Hein, M. Burke, C. Adachi, A. Aspuru-Guzik, A Materials Acceleration Platform for Organic Laser Discovery. *Adv. Mater.* **2207070** (2022), doi:10.1002/adma.202207070.

31. J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, Synthesis of many different types of organic small molecules using one automated process. *Science (80-.).* **347**, 1221–1226 (2015).

32. A. M. K. Nambiar, C. P. Breen, T. Hart, T. Kulesza, T. F. Jamison, K. F. Jensen, Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. *ACS Cent. Sci.* **8**, 825–836 (2022).

33. B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6** (2020), doi:10.1126/sciadv.aaz8867.

34. M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, Data-science driven autonomous process optimization. *Commun. Chem.* **4**, 1–12 (2021).

35. B. Goldman, S. Kearnes, T. Kramer, P. Riley, W. P. Walters, Defining Levels of Automated Chemical Design. *J. Med. Chem.* **65**, 7073–7087 (2022).

36. W. Jin, R. Barzilay, T. Jaakkola, Hierarchical Generation of Molecular Graphs using Structural Motifs. *Proc. - 2021 2nd Int. Conf. Big Data Artif. Intell. Softw. Eng. ICBASE 2021*, 543–546 (2020).

37. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, 1100–1107 (2012).

38. P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 1–11 (2009).

39. C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).

40. A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist, J. L. Reymond, Retrosynthetic accessibility score (RAscore)-rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **12**, 3339–3349 (2021).

41. W. Gao, C. W. Coley, The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).

42. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).

43. G. Scalia, C. A. Grambow, B. Pernici, Y. P. Li, W. H. Green, Evaluating Scalable

Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **60**, 2697–2717 (2020).

44. P. S. Gromski, J. M. Granda, L. Cronin, Universal Chemical Synthesis and Discovery with 'The Chemputer.' *Trends Chem.* **2**, 4–12 (2020).

45. A. J. S. Hammer, A. I. Leonov, N. L. Bell, L. Cronin, Chemputation and the Standardization of Chemical Informatics. *J. Am. Chem. Soc.* **1**, 1572–1587 (2021).

46. L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, ChemOS: An orchestration software to democratize autonomous discovery. *PLoS One*. **15**, 1–18 (2020).

47. L. Gao, S. Shaabani, A. Reyes Romero, R. Xu, M. Ahmadianmoghaddam, A. Dömling, 'Chemistry at the speed of sound': automated 1536-well nanoscale synthesis of 16 scaffolds in parallel. *Green Chem.*, 1380–1394 (2023).

48. A. Osipyan, S. Shaabani, R. Warmerdam, S. V. Shishkina, H. Boltz, A. Dömling, Automated, Accelerated Nanoscale Synthesis of Iminopyrrolidines. *Angew. Chemie*. **132**, 12523–12527 (2020).

49. S. W. Krska, D. A. DiRocco, S. D. Dreher, M. Shevlin, The Evolution of Chemical High-Throughput Experimentation to Address Challenging Problems in Pharmaceutical Synthesis. *Acc. Chem. Res.* **50**, 2976–2985 (2017).

50. A. Thakkar, N. Selmi, J. L. Reymond, O. Engkvist, E. J. Bjerrum, "ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **63**, 8791–8808 (2020).

51. M. Li, J. Xia, R. Tian, J. Wang, J. Fan, J. Du, S. Long, X. Song, J. W. Foley, X. Peng, Near-Infrared Light-Initiated Molecular Superoxide Radical Generator: Rejuvenating Photodynamic Therapy against Hypoxic Tumors. *J. Am. Chem. Soc.* **140**, 14851–14859 (2018).

52. Y. Mo, Y. Guan, P. Verma, J. Guo, M. E. Fortunato, Z. Lu, C. W. Coley, K. F. Jensen, Evaluating and clustering retrosynthesis pathways with learned strategy. Chem. Sci. 12, 1469–1478 (2021).

53. K. P. Greenman, W. H. Green, R. Gómez-Bombarelli, Multi-fidelity prediction of molecular optical peaks with deep learning. Chem. Sci. 13, 1152–1162 (2022).

54. J. F. Joung, M. Han, M. Jeong, S. Park, Experimental database of optical properties of organic compounds. Sci. Data. 7, 1–6 (2020).

55. C. W. Ju, H. Bai, B. Li, R. Liu, Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. J. Chem. Inf. Model. 61, 1053–1065 (2021).

56. V. Venkatraman, R. Raju, S. P. Oikonomopoulos, B. K. Alsberg, The dye-sensitized solar cell database. J. Cheminform. 10, 1–9 (2018).

57. V. Venkatraman, L. K. Chellappan, An open access data set highlighting aggregation of dyes on metal oxides. Data. 5 (2020), doi:10.3390/data5020045.

58. F. H. Vermeire, W. H. Green, Transfer learning for solvation free energies: From quantum chemistry to experiments. Chem. Eng. J. 418, 129307 (2021).

59. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082 (2018).

60. I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-De-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I. V. Tetko, Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. J. Comput. Aided. Mol. Des. 25, 533–554 (2011).

61. Free online access to experimental and predicted chemical properties through the EPA's CompTox Chemistry Dashboard, doi:https://doi.org/10.23645/epacomptox.5179045.v1.

62. Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY, 1994.

63. Y. Chung, F. H. Vermeire, H. Wu, P. J. Walker, M. H. Abraham, W. H. Green, Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. J. Chem. Inf. Model. 62, 433–446 (2022).

64. M. H. Abraham, W. E. Acree, The solubility of liquid and solid compounds in dry octan-1-ol. Chemosphere. 103, 26–34 (2014).

65. M. H. Abraham, R. E. Smith, R. Luchtefeld, A. J. Boorem, R. Luo, W. E. Acree, Prediction of Solubility of Drugs and Other Compounds in Organic Solvents. J. Pharm. Sci. 99, 1500–1515 (2010).

66. S. Martel, F. Gillerat, E. Carosati, D. Maiarelli, I. V. Tetko, R. Mannhold, P. A. Carrupt, Large, chemically diverse dataset of log P measurements for benchmarking studies. Eur. J. Pharm. Sci. 48, 21–29 (2013).

67. MobleyLab/SAMPL6: SAMPL6 Part II - Release the evaluation results of logP predictions.

68. C. H. Camp, PyMCR: A python library for multivariatecurve resolution analysis with alternating regression (MCR-AR). J. Res. Natl. Inst. Stand. Technol. 124, 1–10 (2019).

Acknowledgments:

The authors thank the Shimadzu Innovation Center for their creation of the application programming interface which allowed programmatic control of the HPLC-MS and fraction collector. The authors thank the ASKCOS development team, particularly Max Liu, Mark Murnin, Mike Fortunato, and Thomas Struble for help with ASKCOS and the ASKCOS API. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing high-performance computing resources that have contributed to the research results reported within this report. The authors thank Rabab Alrufayi for help gathering the rules for

implementation of reaction condition recommendations correction, and Prajwal Tumkur for developing Arduino code and the custom circuit board for the high-temperature thermal reactor. The authors thank Austin Croke for help scaling up reaction workup.

Funding:

DARPA Accelerated Molecular Discovery (AMD) program under contract HR00111920025 and the MIT Consortium, Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS). K. P. G. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1745302.

Author contributions:

Designed, developed, and supervised construction of the platform – BAK, RBC, MAM, KFJ

Developed, maintained, and assembled the robotic platform – BJ, TH, TK

Developed and implemented platform control and automation software – BAK, RBC, MAM

Designed and developed property prediction models - KPG, CJM, HW, FHV, S-CL

Developed automated property prediction model retraining – KPG, CJM

Advised in the development of property prediction models – RG-B, WHG

Designed and developed graph-completion model – CLB, WJ

Advised on the development of graph-completion model – TSJ, RB, KFJ

Developed chemistry, developed assays, and performed experiments – BAK, RBC, MAM

Interpretation of results – All authors

Prepared the manuscript – BAK, RBC, MAM

Edited the manuscript – All authors

Conceptualized, supervised the project, and secured funding – TSJ, RB, RG-B, WHG, KFJ

Competing interests: Authors declare no competing interests.

Data and materials availability: All data and code used in this study are freely available in the main text, supplementary materials, with the exception of the Shimadzu API, data from the CAS Content Collection, and Tecan Spark API which are proprietary and unable to be publicly shared.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S11

Tables S1 to S4

References (*52–68*)