Communications in Computational Chemistry
Commun. Comput. Chem., Vol. 7, No. 4, 281–288
doi: 10.4208/cicc.2025.147.01
Regular Article

The Data-Centric Paradigm in Synthetic Chemistry: AI Reaction Modeling Needs More Than Reactant-Product Pairs

Mingjun Yang[1,*] and Peiyu Zhang[1,*]

[1]*Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Floor 3, Sf Industrial Plant, No. 2 Hongliu Road, Fubao Community, Fubao Street, Futian District, Shenzhen 518045, China.*

[*] Corresponding authors: mingjun.yang@xtalpi.com; peiyu.zhang@xtalpi.com

Received on 27 May 2025; Accepted on 10 August 2025

Abstract: AI-driven reaction modeling in synthetic chemistry faces critical data gaps: the scarcity of mechanistic descriptors and inconsistent experimental protocols, leading models trained on sparse reactant-product pairs to falter in tasks like yield prediction, selectivity control, or condition optimization. Recent advances in mechanism-aware data curation, such as hybrid rule-ML frameworks and computational datasets, demonstrate progress but remain limited to small molecules (≤10 heavy atoms) or gas-phase approximations. Concurrently, robot-based high-throughput experimentation (HTE) platforms standardize protocols for a small number of reaction classes yet lack end-to-end traceability, often omitting workup and followed separation and purification steps. To bridge these gaps, we propose a closed-loop framework integrating computational chemistry, robotic HTE, and multimodal AI to resolve critical reaction modeling tasks. From the perspective of future work, the field necessitates expanded collaboration across the community to tackle complex systems, extend HTE to underrepresented reactions, and align data ontologies. Interdisciplinary collaboration is essential to transition from retrospective pattern recognition to mechanism-driven discovery, anchoring AI in datasets that encode *why* reactions succeed, not merely *what* products form.

Key words: yield prediction, reaction mechanism, reaction pathway, reaction dataset.

1. Introduction

The emergence of self-driving laboratories (SDLs) for chemical synthesis, enabled by advances in robotic automation and artificial intelligence (AI)-driven decision-making tools, has created unprecedented opportunities to accelerate molecular discovery [1,2]. While AI models for synthesis planning have demonstrated remarkable capabilities in retro-synthetic route design and forward reaction prediction [3,4], their broader application to critical reaction modeling tasks (e.g., yield prediction, condition selection, and selectivity control) remains hindered by a persistent challenge: the inability of models trained on conventional reaction datasets to generalize reliably beyond their training domains. This "generalizability gap" stems not from algorithmic limitations but from fundamental inadequacies in existing reaction data [5], including sparse mechanistic annotations, inconsistent experimental protocols, incomplete documentation of experimental processes, and inadequate metadata about failed attempts, that collectively obscure the application potential of AI modeling for critical reaction tasks.

Reaction yield prediction, a cornerstone task for improving synthesis efficiency, exemplifies the limitations of current AI modeling approaches. Studies reveal a stark contrast in model generalizability: while HTE datasets enable strong transferability (e.g., transformer models achieving high accuracy using SMILES inputs), models trained on patent, literature, or electronic lab notebook (ELN) data exhibit poor generalization. For instance, transformer-based yield predictors excel on HTE data but fail with

patent-derived reactions due to reporting inconsistencies [6]. Similarly, a benchmark of 41,239 amide coupling reactions from literature by including 2D/3D descriptors and physical properties as modeling features achieved only modest accuracy (R^2=0.395), constrained by reactivity cliffs and measurement variability [7]. Even ELN datasets, such as AstraZeneca's Buchwald–Hartwig reactions, prove challenging for advanced graph neural networks (best R^2=0.266), exposing inherent biases and noise in real-world data [8]. While hybrid models combining DFT features and fingerprints (trained on AbbVie's 24,000+ Suzuki reactions) outperform human chemists in guiding synthesis, their accuracy plummets for complex tasks (R^2=0.137–0.723), underscoring unresolved gaps [9]. Critically, yield discrepancies often stem from undocumented variables: *analytical* vs. *isolated* yields depend on workup protocols, reagent purity, or isolation methods, for which rarely captured in datasets for reaction modeling [10]. These findings highlight that advancing AI-driven reaction modeling demands not just algorithmic innovation but **standardized data** with rigorous experimental metadata to disentangle chemical outcomes from protocol artifacts.

The limitations observed in yield prediction extend to other critical reaction modeling tasks, where incomplete mechanistic and kinetic data constrain model generalizability (Figure 1). For example, machine learning models trained on >10,000 Suzuki-Miyaura couplings from literature failed to outperform simple frequency-based heuristics in predicting optimal solvents or bases, as human reporting biases and the absence of negative data obscured underlying condition-outcome relationships [11]. Similar challenges arise in reactivity and selectivity prediction: most models rely solely on reactant structures and static reaction conditions as inputs, overlooking the role of transition states (TS) in governing reaction pathways. While thermodynamic product stability often guides predictions, kinetic barriers dictated by TS geometries and energies ultimately determine reaction feasibility, site selectivity, and condition sensitivity. Current models, however, lack explicit incorporation of TS characteristics, limiting their ability to resolve competing pathways or guide condition optimization. This disconnect highlights a critical data gap that reaction datasets rarely encode TS descriptors or kinetic profiles, forcing models to infer mechanistic drivers indirectly from sparse reactant-product pairs.

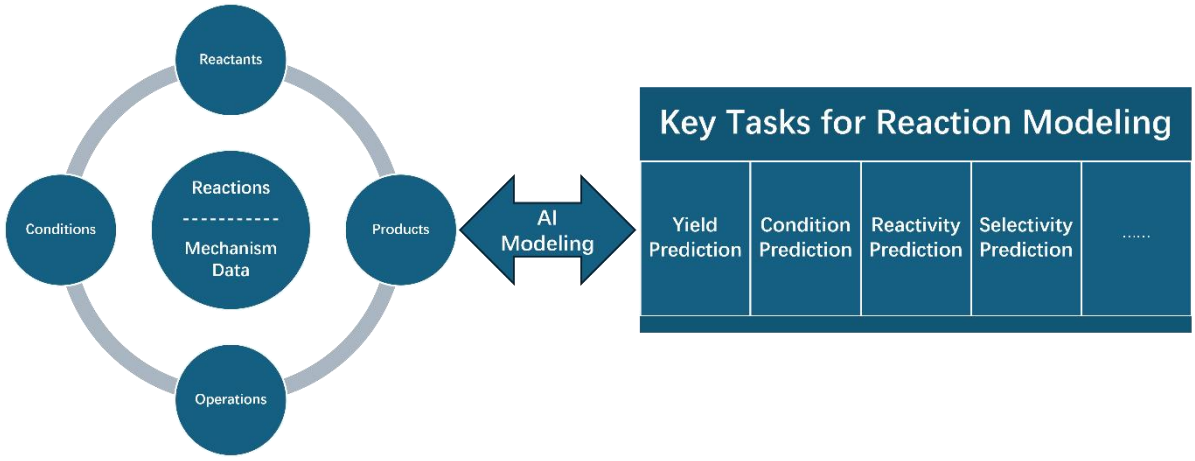


Figure 1. Accurate modeling for chemical reactions needs both standardized experimental data and computed reaction mechanism data

The core challenge for AI-driven reaction modeling lies in two interconnected data deficiencies: (1) the systematic omission of mechanistic descriptors (e.g., transition-state geometries, kinetic profiles, or competing pathway data) in conventional reaction databases, and (2) the inconsistency of experimental protocols across datasets, which conflates chemical causality with procedural artifacts (Figure 1). To bridge this gap, we propose dual strategies: first, curating mechanism-oriented reaction libraries that explicitly encode energetic landscapes, structural and stereoelectronic features along the reaction pathway to consider thermodynamic and kinetic drivers of reactivity; second, leveraging automated HTE to generate protocol-standardized datasets with full traceability of reaction parameters. Combining these approaches enables models to distinguish intrinsic chemical behavior from experimental noise while capturing kinetic bottlenecks that govern selectivity. Achieving this demands interdisciplinary integration of computational chemistry (for mechanism annotation), robotic automation (for protocol reproducibility), and data science (for causality extraction), and thus a collaborative framework to establish "chemically intelligent" datasets as the foundation for reliable AI tools in synthesis planning.

2. Challenges and advances in reaction mechanism data curation

Mechanistic understanding is critical for modeling reaction outcomes, from predicting optimal conditions (temperature, solvent, catalyst loading) to controlling regio-/stereoselectivity and resolving competing pathways. TS geometries govern kinetic feasibility, while electronic interactions between catalysts and substrates (e.g., charge-transfer dynamics) dictate selectivity thresholds. Thermodynamic stability of intermediates and products, combined with kinetic activation barriers, further determines pathway dominance. Despite this foundational role of mechanistic data, its integration into reaction modeling remains constrained by limited access to standardized databases that systematically encode TS descriptors, energetic landscapes, or kinetic profiles. Current datasets predominantly focus on reactant-product pairs [12,13], omitting the multidimensional parameters needed to correlate mechanistic drivers with experimental outcomes.

Efforts to address mechanistic data scarcity are progressing through hybrid strategies that combine rule-based systems with machine learning (ML), enabling the systematic annotation of reaction pathways. **Rule-based approaches**, such as expert systems codified with >1,500 transformation rules (e.g., electron-pushing diagrams and elementary reaction steps), provide chemically grounded frameworks for mechanistic analysis and retrosynthetic validation [14] (Figure 2). These systems, however, face limitations in scalability and coverage of novel reaction types. To overcome these constraints, ML models like **ELECTRO** leverage graph neural networks to infer electron flow directly from atom-mapped reaction data, achieving state-of-the-art accuracy on curated datasets like USPTO while implicitly capturing functional group selectivity [15]. While promising, such models occasionally generate pathways misaligned with expert intuition due to reliance on implicit chemical constraints rather than explicit mechanistic guidance. Bridging this gap, initiatives like **mech-USPTO-31K** demonstrate scalable solutions: using automated template extraction (e.g., MechFinder) and expert curation, they annotate 31,364 reactions with polar mechanisms validated by chemists (74% accuracy) [16]. Similarly, platforms like **Allchemy** integrate mechanistic transforms (e.g., nucleophilic interactions, pericyclic reactions) with physical-organic principles to predict reaction networks and yields (MAE = 7.3–10.5%) [17]. These hybrid frameworks—combining rule-based granularity with data-driven scalability—highlight the potential of mechanistic datasets to enhance predictive models, though challenges persist in stereochemical fidelity and coverage of nonpolar pathways. Collectively, these advances underscore the need for open, standardized resources like mech-USPTO-31K to train next-generation AI while maintaining chemical rigor.

While hybrid rule and ML frameworks address mechanistic annotation gaps, advancing AI models to predict reactivity and selectivity reliably requires datasets that integrate 3D structural coordinates, energetic profiles (activation barriers, enthalpies), and kinetic parameters that are critical for modeling spatial and electronic drivers of reaction outcomes (Figure 2). Recent efforts prioritize computational datasets combining quantum chemistry (QC) calculations with automated TS searches. For example, **Transition1x** [18] leverages nudged elastic band (NEB) methods to generate 9.6 million DFT calculations along reaction pathways, enabling graph neural networks to predict TS geometries and reaction barriers with higher accuracy than equilibrium-geometry datasets (e.g., ANI1x). Similarly, the **Reaction Graph Depth 1 (RGD1)** dataset [19] employs multi-level QC (GFN2-xTB to CCSD(T)-F12) to curate 176,992 validated reactions with TS geometries, activation energies, and enthalpies, spanning diverse C/H/O/N chemistry. By automating TS searches across ~708,000 elementary reactions, RGD1 addresses scalability and chemical diversity limitations of earlier QC datasets, though computational costs restrict its scope to small molecules (≤10 heavy atoms). These datasets highlight three critical advances: (1) explicit encoding of TS geometries and energetic landscapes bridges the gap between static reactant-product pairs and dynamic reaction pathways; (2) automated QC workflows enable systematic TS exploration at scale; (3) multi-level validation (DFT to CCSD(T)) ensures data reliability for training ML potentials. However, challenges persist: gas-phase calculations overlook solvent effects and narrow elemental coverage (H/C/N/O) limits generalizability. Overcoming these limitations demands synergistic integration of more efficient and accurate reaction pathway generation method developed and more computing resources available for these data curation projects.

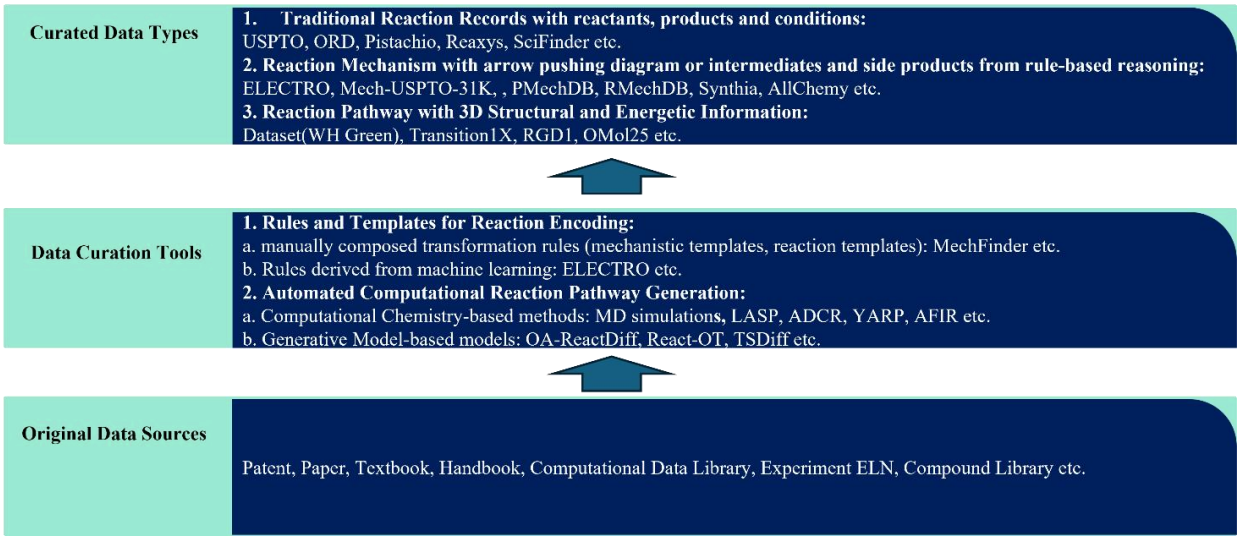


Figure 2. The curation of reaction mechanism data with different methods

To scale reaction pathway data generation, researchers are integrating computational chemistry methods with ML to balance accuracy and efficiency (Figure 2). *Multi-scale frameworks* combine TS search algorithms (e.g., NEB, GSM), molecular dynamics (MD) simulations, and generative ML models to map reaction landscapes hierarchically. For example, the **Yet Another Reaction Program (YARP)** [20] automates TS exploration by enumerating elementary reaction steps (bond breaking/forming), pre-optimizing geometries with GFN2-xTB, and refining pathways with DFT, a workflow enabling the curation of RGD1's 176,992 reactions. Similarly, the automated design of chemical reaction **(ADCR) program** [21,22] combines MD with coordinate-driving methods to handle large systems (>100 atoms) and complex reactions (radicals, transition metals), though computational costs limit its throughput. The large-scale atomistic simulation with neural network (NN) potential (**LASP**) [23]

package combining stochastic surface walking (SSW) with global NN potential to facilitate the PES exploration for a wide range of complex systems including applications to chemical reaction pathway generation. The artificial force induced reaction (**AFIR**) [24] method automates reaction pathway discovery by applying artificial forces between molecular fragments (MC-AFIR for multi-component reactions) or perturbing atom pairs within a molecule (SC-AFIR for isomerization), guided by collision energy parameters to navigate potential energy surfaces. But its performance depends on careful parameter selection and can be computationally demanding for large systems. Recently the study by **FAIR** at meta uses AFIR approach to generate reactive pathways for metal complexes and organic reactions (from RMechDB/PMechDB), producing datasets with 1,436 metal reactions and with elementary reaction steps for radical/polar reactions, which consists of the reactive part of OMol25 database [25]. Meanwhile, ML-driven tools like **OA-ReactDiff** (SE(3)-equivariant diffusion model) [26] and **React-OT** (optimal transport) [27] generate TS geometries with near-quantum accuracy (0.08 Å RMSD) directly from reactant-product pairs, bypassing iterative QC calculations. However, ML models face transferability issues: pretrained machine learning interatomic potentials (MLIPs) require fine-tuning on reactive data to predict TS barriers reliably, and generative models like OA-ReactDiff depend heavily on training data coverage. While these tools accelerate pathway sampling by orders of magnitude, fundamental gaps persist. Gas-phase approximations neglect solvation/entropic effects, and ML efficiency gains are offset by dataset biases (e.g., Transition1x's focus on small molecules). Closing these gaps demands tighter integration of *enhanced sampling*, *multi-fidelity QC workflows*, and *protocol standardization* with end-to-end **benchmark metrics** to ensure datasets capture both thermodynamic and kinetic determinants of reactivity. Only through such synergies can computational frameworks advance from exploratory tools to reliable data sources for AI-driven reaction modeling.

3. Challenges and advances in automated robot-based HTE data curation

Traditional experimental workflows, even when augmented by ELNs, produce datasets riddled with inconsistencies that undermine AI model reliability. Manual experimentation introduces biases (e.g., selective reporting of successful reactions), protocol variability (e.g., undocumented deviations in workup steps), and fragmented metadata (e.g., missing reagent purity or analytical settings), which obscure causal relationships between reaction parameters and outcomes. These limitations are particularly acute for complex tasks like selectivity prediction, where subtle condition-dependent effects dominate. Automated HTE platforms address these gaps by standardizing protocols and digitalizing the end-to-end experimental process. Rapidly increasing studies have been conducted to demonstrate the HTE ability for chemical reactions. In one study, the researchers develop a universal chemical programming language (xDL) to digitize 103 organic synthesis protocols into executable code, enabling their robotic execution on modular "ChemPU" systems [28]. They validate over 50% of these procedures robotically, achieving yields and purities comparable to manual synthesis while demonstrating automated purification via integrated chromatography. By using mobile robots to integrate distributed synthesis (Chemspeed ISynth), analysis (UPLC-MS, benchtop NMR), and specialized equipment (photoreactor), a modular autonomous chemistry platform has been developed with a heuristic decision-maker to process orthogonal NMR and MS data, enabling autonomous discovery and validation of synthetic targets including structurally diverse molecules, supramolecular host-guest systems, and photochemical products [29]. To enhance autonomous lab efficiency of complex experiments, the researchers develop a multi-robot-multi-task scheduling system with constraint programming to optimize concurrent execution of diverse chemical experiments across three robots and 18 stations [30]. Real-world validation with four parallel experiments showed a 40% reduction in total time compared to sequential execution, supporting dynamic task insertion without significant disruption. In addition, some of these researchers also develop ChemAgents by using a hierarchical multiagent AI (*Literature Reader, Experiment Designer, Computation Performer, Robot Operator*) powered by an on-board Llama-3.1-70B LLM to autonomously execute complex multistep chemistry experiments, enabling accelerated discovery with minimal human input [31]. Despite rapid development of robot-based HTE technology together with integrating AI agents and other AI algorithms, current HTE efforts remain narrowly focused with fewer than 5 reaction classes dominating published datasets with public availability.

Systematic exploration of key cross-coupling reactions have been conducted with HTE and datasets are made accessible through repositories like the Open Reaction Database [12] or Github. These HTE studies usually combine with Bayesian or active learning algorithms for feedback-guided reaction condition optimization and reaction space sampling for better prediction of reaction feasibility and yield [32–35]. For Suzuki-Miyaura couplings, studies highlight distinct approaches (Table 1):

(1) A droplet-flow microfluidic system optimized conditions for aryl chlorides and unstable boronic acids via ~400 reactions, using real-time HPLC-UV analysis calibrated via least-squares regression to quantify yields [34].

(2) A nanomole-scale flow platform screened 5,760 reactions (11 ligands, 7 bases, 4 solvents, 5 electrophiles, and 7 nucleophiles) with LCMS-UV analysis, achieving analytical yield quantification for the same one target product and scalable validation [36].

(3) A closed-loop workflow combined machine learning with 528 reactions across 11 substrate pairs, using LCMS-UV to validate model-guided condition recommendations [33].

These efforts demonstrate the integration of automation with machine learning for parameter optimization but reveal persistent challenges in calibration (e.g., isolated vs. analytical yield discrepancies) and scalability (e.g., limited substrate diversity).

The HTE applications extend beyond Suzuki-Miyaura couplings to other reaction classes, though coverage remains sparse (Table 1):

(1) **Buchwald-Hartwig (C–N)**: A study screened 4,608 reactions (15 aryl halides, 4 ligands, 3 bases, and 23 isoxazole additives) with LCMS-UV quantification, identifying inhibitory oxidative addition pathways via random forest modeling [37].

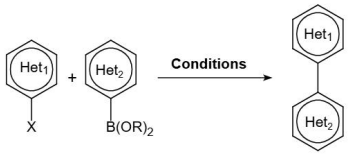
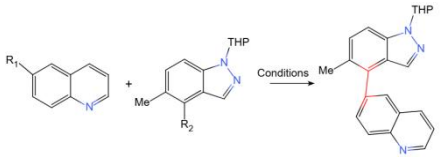
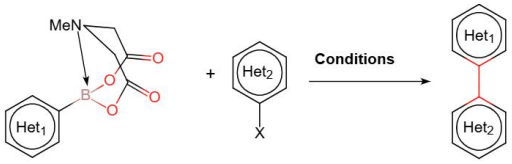
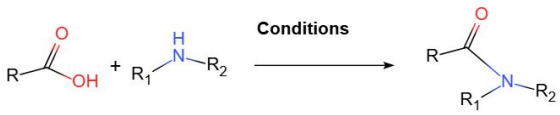
(2) **Amide Coupling:** A Bayesian neural network (BNN) guided 11,669 reactions, using uncalibrated LCMS-UV absorbance ratios to predict feasibility (89.48% accuracy) and prioritize scalable conditions [35].

(3) **Mizoroki-Heck (C–C):** A 384-reaction screen evaluated Pd/Ni catalysts with aryl triflates/iodides, quantifying yields via

GC-FID, NMR, or isolation, and deposited regioselectivity data (>20:1 branched/linear ratios) in the Open Reaction Database [38].

While these studies illustrate HTE's potential for protocol standardization, limitations persist: narrow reaction scope, inconsistent yield validation (e.g., UV calibration vs. isolation), and incomplete mechanistic metadata (e.g., solvent effects on transition states).

Table 1. The publicly available HTE data sets published since 2016.

Reaction Type	Yield Type	Number of Reaction Data	Year of Publish
Suzuki-Miyaura cross coupling: 	HPLC analytic yield with calibration	~400	2016 [34]
Suzuki-Miyaura cross coupling: 	LCMS analytic yield relative to the reference value for one target product	5,760	2018 [36]
Suzuki-Miyaura cross coupling: 	LCMS analytic yield calibrated with standards	~600	2022 [33]
Buchwald-Hartwig Cross Coupling: 	LCMS-UV with internal standard as reference	~3000 (with yield labels)	2018 [37]
Amide Coupling: 	LCMS-UV area without calibration	11,669	2025 [35]
Mizoroki-Heck: 	Reaction yields and regioselectivity quantified by GC-FID, NMR, or isolated yields	384	2023 [38]

4. Conclusion and perspectives

The evolution of AI-driven reaction modeling hinges on resolving two foundational data challenges: (1) the scarcity of mechanistic descriptors that correlate reaction outcomes with kinetic/thermodynamic principles, and (2) the inconsistency of experimental protocols across datasets. Current strategies, including computational datasets encoding transition-state geometries (e.g., RGD1, Transition1x), hybrid rule-ML frameworks (e.g., mech-USPTO-31K), and standardized HTE platforms, demonstrate progress but still face critical gaps. Computational workflows efficiently map reaction pathways yet struggle to incorporate solvation effects and more complicated electronic effects or scale to complex systems, while HTE platforms standardize protocols but remain confined to narrow reaction classes and mostly only cover the reaction and analysis stages leaving the workup, separation and purification stages unquantified.

To bridge these gaps, we advocate for reaction modeling by integrating multimodal data from AI & digitalization enhanced automated HTE reaction data, the reaction text description data, the reaction mechanism diagram data, and the reaction pathway data (Figure 3). In this framework, autonomous robotic platforms execute HTE with end-to-end traceability, recording reaction parameters (e.g., temperature ramps, mixing rates), workup protocols (e.g., extraction steps, drying times), analytical spectra (e.g., inline or offline LCMS, UV, IR, Raman, and NMR etc.), isolation and purification methods and settings, to minimize annotation artifacts and isolate protocol-specific biases. On the one hand, this HTE platform should tightly integrate automatic robots for large-scale and unified-protocol experiments, digital tools for end-to-end data record and management, intelligent computational models for data annotation and decision-making. On the other hand, reaction data collection should be conducted to cover the orthogonal space composed of widely employed reaction types, representative reactants and conditions to not only balance the positive and negative data but also the coverage of critical reaction space. To maximize the impact of this integrated framework, future efforts should prioritize overcoming analytical and resource challenges in developing kinetic, multi-time-point datasets that is essential for capturing rate-dependent mechanistic insights currently obscured by conventional single-yield HTE data [10].

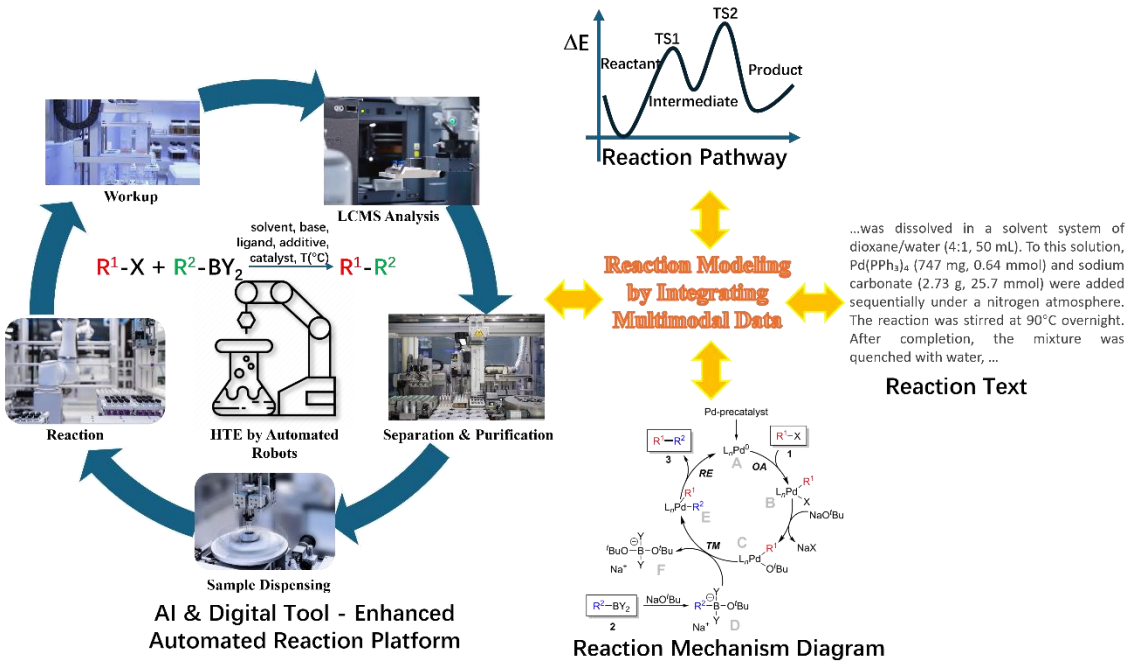


Figure 3. Reaction modeling by integrating multimodal data from AI & digitalization enhanced automated HTE reaction data, the reaction text description data, the reaction mechanism diagram data, and the reaction pathway data.

For modeling the reaction tasks, enhanced predictive performance is observed by combining multi-modal data from HTE (both the process and outcome data), reaction mechanism and pathway (geometric, energetic, electronic data, and reaction diagram), and reaction context (textual protocols and interpretations). To enhance the yield prediction generalization, machine learning model for amide coupling has been developed by embedding intermediate knowledge (specifically, activated acid intermediates formed under specific conditions) into the model's input descriptors, improving the R^2 for both single conditions and rigorous "full substrate novelty" tests [39]. For multi-component reactions (MCRs), the yield prediction model has been developed by integrating mechanistic reaction knowledge from the reaction networks including main reaction pathways, immediate side reactions, and downstream by-product interactions, effectively generalizing to 10 novel MCRs (MAE = 7.3%) with only 20 mechanistically diverse MCRs used as training data [17]. To enable scalable production of optoelectronic materials with minimal catalyst loading, a catalyst-oriented design based on elementary reactions (CODER) strategy is proposed to design Pd catalyst achieving record-breaking turnover numbers (TON=340,000) for triarylamine synthesis [40]. To summarize studies overcoming data

limitations and enabling extrapolative and human-expert-level accuracy, a recent review highlights how embedding chemical knowledge (e.g., mechanistic descriptors, transition-state geometries) into machine learning pipelines significantly enhances performance prediction for organic synthesis including yield, regio-, and stereoselectivity [41]. Collectively, these advances demonstrate a paradigm shift toward multimodal data-guided machine learning as compared to the AI models developed mainly on basis of reactant and product data. However, challenges persist in formalizing complex reaction networks, ensuring cross-reaction transferability, and scaling lab-validated models to industrial workflows.

From the perspective of future work, the field necessitates expanded collaboration across the community to tackle complex systems, extend HTE to underrepresented reactions, and align data ontologies. Only by anchoring AI models in chemically intelligent datasets, those that encode why reactions succeed or fail, not just what reactants transform into products, can synthetic chemistry transition from retrospective pattern recognition to prospective, mechanism-driven discovery.

Acknowledgments

The authors acknowledge funding support from Shenzhen Science and Technology Program (KJZD20240903100304007 and CJGJZD20220517142201004).

References

[1] Jiang X., Luo S., Liao K., Jiang S., Ma J., Jiang J. and Shuai Z., Artificial intelligence and automation to power the future of chemistry. *Cell Rep. Phys. Sci.*, **5** (7) (2024).

[2] Tom G., Schmid S. P., Baird S. G., Cao Y., Darvish K., Hao H., Lo S., Pablo-García S., Rajaonson E. M., Skreta M., Yoshikawa N., Corapi S., Akkoc G. D., Strieth-Kalthoff F., Seifrid M. and Aspuru-Guzik A., Self-driving laboratories for chemistry and materials science. *Chem. Rev.*, **124** (16) (2024), 9633–9732.

[3] Wołos A., Koszelewski D., Roszak R., Szymkuć S., Moskal M., Ostaszewski R., Herrera B. T., Maier J. M., Brezicki G., Samuel J., Lummiss J. A. M., McQuade D. T., Rogers L. and Grzybowski B. A., Computer-designed repurposing of chemical wastes into drugs. *Nature*, **604** (7907) (2022), 668–676.

[4] Szymkuć S., Gajewska E. P., Klucznik T., Molga K., Dittwald P., Startek M., Bajczyk M. and Grzybowski B. A., Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem. Int. Ed.*, **55** (20) (2016), 5904–5937.

[5] Strieth-Kalthoff F., Sandfort F., Kühnemund M., Schäfer F. R., Kuchen H. and Glorius F., Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem. Int. Ed.*, **61** (29) (2022), e202204647.

[6] Schwaller P., Vaucher A. C., Laino T. and Reymond J. L., Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.*, **2** (1) (2021), 015016.

[7] Liu Z., Moroz Y. S. and Isayev O., The challenge of balancing model sensitivity and robustness in predicting yields: A benchmarking study of amide coupling reactions. *Chem. Sci.*, **14** (39) (2023), 10835–10846.

[8] Saebi M., Nan B., Herr J. E., Wahlers J., Guo Z., Zurański A. M., Kogej T., Norrby P. O., Doyle A. G., Chawla N. V. and Wiest O., On the use of real-world datasets for reaction yield prediction. *Chem. Sci.*, **14** (19) (2023), 4997–5005.

[9] Raghavan P., Rago A. J., Verma P., Hassan M. M., Goshu G. M., Dombrowski A. W., Pandey A., Coley C. W. and Wang Y., Incorporating synthetic accessibility in drug design: Predicting reaction yields of Suzuki cross-couplings by leveraging AbbVie's 15-year parallel library data set. *J. Am. Chem. Soc.*, **146** (22) (2024), 15070–15084.

[10] Raghavan P., Haas B. C., Ruos M. E., Schleinitz J., Doyle A. G., Reisman S. E., Sigman M. S. and Coley C. W., Dataset design for building models of chemical reactivity. *ACS Cent. Sci.*, **9** (12) (2023), 2196–2204.

[11] Beker W., Roszak R., Wołos A., Angello N. H., Rathore V., Burke M. D. and Grzybowski B. A., Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic Suzuki-Miyaura coupling. *J. Am. Chem. Soc.*, **144** (11) (2022), 4819–4827.

[12] Kearnes S. M., Maser M. R., Wleklinski M., Kast A., Doyle A. G., Dreher S. D., Hawkins J. M., Jensen K. F. and Coley C. W., The Open Reaction Database. *J. Am. Chem. Soc.*, **143** (45) (2021), 18820–18826.

[13] Mayfield J., Lagerstedt I., Pirie R. and Sayle R., Automated extraction and curation of 15 million reactions., (2023).

[14] Chen J. H. and Baldi P., No electron left behind: A rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.*, **49** (9) (2009), 2034–2043.

[15] Bradshaw J., Kusner M. J., Paige B., Segler M. H. S. and Hernández-Lobato J. M., A generative model for electron paths. *arXiv,* (2019).

[16] Chen S., Babazade R., Kim T., Han S. and Jung Y., A large-scale reaction dataset of mechanistic pathways of organic reactions. *Sci. Data*, **11** (1) (2024), 863.

[17] Szymkuć S., Wołos A., Roszak R. and Grzybowski B. A., Estimation of multicomponent reactions' yields from networks of mechanistic steps. *Nat. Commun.*, **15** (1) (2024), 10286.

[18] Schreiner M., Bhowmik A., Vegge T., Busk J. and Winther O., Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data*, **9** (1) (2022), 779.

[19] Zhao Q., Vaddadi S. M., Woulfe M., Ogunfowora L. A., Garimella S. S., Isayev O. and Savoie B. M., Comprehensive exploration of graphically defined reaction spaces. *Sci. Data*, **10** (1) (2023), 145.

[20] Zhao Q. and Savoie B. M., Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.*, **1** (7) (2021), 479–490.

[21] Yang M., Zou J., Wang G. and Li S., Automatic reaction pathway search via combined molecular dynamics and coordinate driving method. *J. Phys. Chem. A*, **121** (6) (2017), 1351–1361.

[22] Li G., Li Z., Gao L., Chen S., Wang G. and Li S., Combined molecular dynamics and coordinate driving method for automatically searching complicated reaction pathways. *Phys. Chem. Chem. Phys.*, **25** (35) (2023), 23696–23707.

[23] Huang S., Shang C., Kang P., Zhang X. and Liu Z., LASP: Fast global potential energy surface exploration. *WIRes Comput. Mol. Sci.*, **9** (6) (2019), e1415.

[24] Maeda S., Harabuchi Y., Takagi M., Taketsugu T. and Morokuma K., Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces. *Chem. Rec.*, **16** (5) (2016), 2232–2248.

[25] Levine D. S., Shuaibi M., Spotte-Smith E. W. C., Taylor M. G., Hasyim M. R., Michel K., Batatia I., Csányi G., Dzamba M., Eastman P., Frey N. C., Fu X., Gharakhanyan V., Krishnapriyan A. S., Rackers J. A., Raja S., Rizvi A., Rosen A. S., Ulissi Z., Vargas S., Zitnick C. L., Blau S. M. and Wood B. M., The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv,* (2025).

[26] 2024-OA-reactDiffusion-DuanChenru-Transition-State-Generation.

[27] Duan C., Liu G. H., Du Y., Chen T., Zhao Q., Jia H., Gomes C. P., Theodorou A. and Kulik H. J., React-OT: Optimal transport for generating transition state in chemical reactions. *arXiv,* (2024).

[28] Rohrbach S., Šiaučiulis M., Chisholm G., Pirvan P. A., Saleeb M., Mehr S. H. M., Trushina E., Leonov A. I., Keenan G., Khan A., Hammer A. and Cronin L., Digitization and validation of a chemical synthesis literature database in the ChemPU. *Science*, **377** (6602) (2022), 172–180.

[29] Dai T., Vijayakrishnan S., Szczypiński F. T., Ayme J. F., Simaei E., Fellowes T., Clowes R., Kotopanov L., Shields C. E., Zhou Z., Ward J. W. and Cooper A. I., Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, **635** (8040) (2024), 890–897.

[30] Zhou J., Luo M., Chen L., Zhu Q., Jiang S., Zhang F., Shang W. and Jiang J., A multi-robot–multi-task scheduling system for autonomous chemistry laboratories. *Digit. Discov.*, **4** (3) (2025), 636–652.

[31] Song T., Luo M., Zhang X., Chen L., Huang Y., Cao J., Zhu Q., Liu D., Zhang B., Zou G., Zhang G., Zhang F., Shang W., Fu Y., Jiang J. and Luo Y., A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *J. Am. Chem. Soc.*, **147** (15) (2025), 12534–12545.

[32] Dunlap J. H., Ethier J. G., Putnam-Neeb A. A., Iyer S., Luo S. X. L., Feng H., Garrido Torres J. A., Doyle A. G., Swager T. M., Vaia R. A., Mirau P., Crouse C. A. and Baldwin L. A., Continuous flow synthesis of pyridinium salts accelerated by multi-objective Bayesian optimization with active learning. *Chem. Sci.*, **14** (30) (2023), 8061–8069.

[33] Angello N. H., Rathore V., Beker W., Wolos A., Jira E. R., Roszak R., Wu T. C., Schroeder C. M., Aspuru-Guzik A., Grzybowski B. A. and Burke M. D., Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science*, **378** (6618) (2022), 399–405.

[34] Reizman B. J., Wang Y. M., Buchwald S. L. and Jensen K. F., Suzuki-Miyaura cross-coupling optimization enabled by automated feedback. *React. Chem. Eng.*, **1** (6) (2016), 658–666.

[35] Zhong H., Liu Y., Sun H., Liu Y., Zhang R., Li B., Yang Y., Huang Y., Yang F., Mak F. S., Foo K., Lin S., Yu T., Wang P. and Wang X., Towards global reaction feasibility and robustness prediction with high throughput data and Bayesian deep learning. *Nat. Commun.*, **16** (1) (2025), 4522.

[36] Perera D., Tucker J. W., Brahmbhatt S., Helal C. J., Chong A., Farrell W., Richardson P. and Sach N. W., A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, **359** (6374) (2018), 429–434.

[37] Ahneman D. T., Estrada J. G., Lin S., Dreher S. D. and Doyle A. G., Predicting reaction performance in C-N cross-coupling using machine learning. *Science*, **360** (6385) (2018), 186–190.

[38] Isbrandt E., Chapple D., Tu N. T. P., Dimakos V., Beardall A. M., Boyle P., Rowley C., Blacquiere J. and Newman S., Controlling reactivity and selectivity in the Mizoroki-Heck reaction: High throughput evaluation of 1,5-diaza-3,7-diphosphacyclooctane ligands. *Chemistry,* (2023).

[39] Zhang C., Lin Q., Deng H., Yang C., Kong Y., Yu Z. and Liao K., Intermediate knowledge enhanced the performance of N-amide coupling yield prediction model. *Chemistry,* (2024).

[40] Liu H. W., He P., Li W. T., Sun W., Shi K., Wang Y. Q., Mo Q. K., Zhang X. Y. and Zhu S. F., Catalyst-oriented design based on elementary reactions (CODER) for triarylamine synthesis. *Angew. Chem. Int. Ed.*, **62** (44) (2023), e202309111.

[41] Zhang S. Q., Xu L. C., Li S. W., Oliveira J. C. A., Li X., Ackermann L. and Hong X., Bridging chemical knowledge and machine learning for performance prediction of organic synthesis. *Chem. - Eur. J.*, **29** (6) (2023), e202202834.