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.

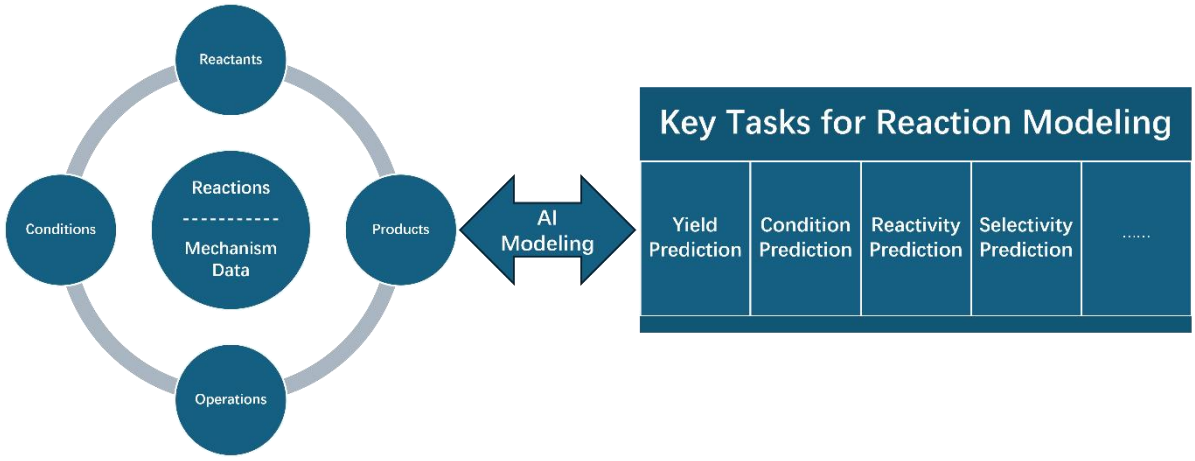

Figure 1. Accurate modeling for chemical reactions needs both standardized experimental data and computed reaction mechanism data

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.