

Generative discovery of chemical reaction networks

Yunhong Han^{1,*}, Gaopeng Ren^{1,*}, Yan Liu^{1,*}, Ronghua Zhang¹, Huangjiang Huang¹, Ruiqi Zhang^{1,2}, Zhaoxi Yu^{1,3}, Chuncheng Liu¹, Zhao Li¹, Xiaohan Lin¹, Jiaxu Wang¹, Yuxuan Qin¹, Meijie Wang¹, Guoshun Mei¹, Doudou Wang¹, Taoyong Cui¹, Haojun Jia^{1,†}, Chenru Duan^{1,†}, Qiyuan Zhao^{1,†}

¹Deep Principle, Hangzhou, Zhejiang, China.

²Peking University Shenzhen Graduate School, Shenzhen, Guangdong, China.

³Key Laboratory of Theoretical and Computational Photochemistry of Ministry of Education, College of Chemistry, Beijing Normal University, Beijing 100875, China.

*These authors contributed equally to this work.

Chemical kinetics determines molecular stability, reaction selectivity and the evolution of reacting mixtures, but discovering the underlying reaction networks remains challenging. Here we introduce ReactNexus, a generative framework for kinetic-guided network exploration starting solely from reactant graphs. Generative models propose transition states and predict the full reaction trajectories connecting reactants and products, while calculations using a machine-learned interatomic potential verify connectivity and evaluate electronic barriers. On an in-distribution benchmark of 10,000 generated candidates, 80.8% connect to the input reactant after density functional theory (DFT) optimization and intrinsic reaction coordinate verification. The resulting networks uncover lower-barrier cooperative pathways and competition between catalyst release and catalyst-bound branches. In the examined carbocation cascades, coverage of known product precursors saturates after limited expansion and distinct entries converge onto shared intermediates, indicating a compact accessible reaction space. Network-derived candidate structures further assist mass-spectral interpretation of a complex nucleophilic aromatic substitution reaction. Together, these results establish a generative approach to reaction-network discovery and mechanistic analysis without predefined products or reaction templates.

Date: October 26, 2026

†Correspondence:

Haojun Jia: haojunjia@deepprinciple.com

Chenru Duan: duanchenru@deepprinciple.com

Qiyuan Zhao: zhaoqiyuan@deepprinciple.com

Available soon: <https://mira.deepprinciple.com>



Deep Principle

1 Introduction

Starting from a set of reactants, reaction-network prediction seeks to discover accessible products and intermediates, identify the elementary pathways connecting them and characterize the barriers governing their competition.^{1–6} Such networks provide a mechanistic basis for assessing molecular stability, understanding catalytic selectivity and interpreting experimental reaction outcomes.^{7–11} Their construction requires discovering species and connections as exploration proceeds: each new intermediate can become a starting point for further reactions.^{12,13} This task becomes increasingly demanding as molecular size, conformational flexibility and the number of interacting components increase. More reactive sites, component combinations and coupled bond rearrangements must be considered, while an omitted pathway can alter the mechanistic picture and redirect subsequent network growth.¹⁴ The challenge is therefore to discover relevant competing pathways and establish their connections and barriers within a finite computational budget.

Existing approaches differ primarily in how they propose reactions. Template- and rule-based generators, including Reaction Mechanism Generator (RMG), efficiently construct networks within encoded reaction families and can incorporate kinetic estimates, but the scope of exploration is limited to those families.^{15–17} Reactive dynamics, including the *ab initio* nanoreactor, discovers unanticipated

transformations without predefined products; enhanced sampling accelerates rare events, followed by pathway refinement to locate transition states (TSs) and determine barriers. However, the computational cost of reactive trajectories and the difficulty of sampling competing pathways limit systematic exploration as chemical complexity increases.^{6,18,19} Single-ended and stochastic potential energy surface searches, such as artificial force induced reaction (AFIR) and stochastic surface walking (SSW), likewise discover reactions without specified products.^{4,20} Broad exploration nevertheless requires repeated searches across alternative pathways and conformations,^{21–23} and may require user-defined restrictions on reactive atoms or reaction coordinates to make larger systems tractable.^{4,24} Graph-based methods such as ZStruct and Yet Another Reaction Program (YARP) instead enumerate candidate products through generic bond edits, extending beyond established reaction templates.^{3,25,26} Their candidate space nevertheless grows combinatorially with molecular size, permitted bond changes and combinations of reacting components.¹² Across these approaches, the challenge is to propose competing pathways efficiently enough to sustain network growth in complex systems.

Machine learning offers a new perspective on reaction-network exploration through learned potentials and generative models.^{2,27,28} Machine-learned interatomic potentials (MLIPs), increasingly trained and applied to reactive systems,^{29–33} reduce the cost of energy and force evaluations for simulations and pathway searches.^{34–38} In parallel, diffusion and flow-based models generate three-dimensional TS geometries from molecular structures or reaction graphs.^{39–42} These models adopt a double-ended formulation, requiring both reactants and products to be specified before TS generation. Their use in network exploration therefore requires an upstream source of candidate reactions, such as graph-based enumeration.^{41,43} Generating unknown reaction networks requires models that propose pathways from reactants alone and identify new species that can undergo further reactions. Developing and evaluating this capability requires reaction-path data spanning competing transformations and diverse chemical environments. Widely used datasets such as Transition1x and RGD1 mainly cover small organic systems, leaving larger molecules and multicomponent environments less well represented.^{44,45} The high cost of intrinsic reaction coordinate (IRC) calculations at the density functional theory (DFT) level also limits systematic verification of reactant–TS–product assignments, leaving uncertainty that can affect the reliability of models trained on these data for reaction-network exploration.^{44,45}

Here we introduce ReactNexus, a framework for generating chemical reaction networks directly from reactants, without predefined products, reaction templates or prior enumeration of candidate reactions. Single-step benchmarks establish the basis for iterative, kinetic-guided exploration (Fig. 1). In multicomponent and structurally complex systems, the generated pathways uncover lower-barrier cooperative transformations and conformational alternatives (Fig. 2), while catalytic networks reveal pathways to catalyst regeneration and competing routes to potential catalyst deactivation (Fig. 3). Distinct carbocation entries converge onto shared intermediates, and coverage of known product precursors stabilizes after limited expansion. The pathways also revise stepwise mechanisms inferred from classical rearrangement templates, supported independently by targeted DFT calculations (Fig. 4). These computationally generated networks provide structural hypotheses that assist mass-spectral interpretation of a complex nucleophilic aromatic substitution reaction (Fig. 5).

2 Results

2.1 Generative network construction and single-step validation

Method overview. ReactNexus grows a network by coupling pathway proposals to endpoint identification and energetic evaluation (Fig. 1a). From an input reactant set, Nexus-Sampler proposes TS geometries without specifying products, and Nexus-Simulator predicts full trajectories from these geometries, enabling endpoint comparison with the input reactants and identification of distinct products. Nexus-Verifier evaluates energetics and supports pathway refinement using Carina, a reactive MLIP based on the Equiformer-V3 architecture⁴⁶ and trained on over 100 million energy–force data points from our reaction dataset (Fig. 1b).

Simulated trajectories and Carina single-point energies support rapid network exploration without TS optimization or IRC calculations. When further verification is required, Carina TS optimization is performed, followed by Simulator prediction or IRC calculations on the Carina potential energy surface and endpoint verification. To reduce the computational cost of these calculations, TS candidates are first clustered by geometrical similarity, and up to k representatives per cluster are sent to Simulator, where k is configurable. Simulated trajectories that do not recover the input reactants at either endpoint, or show no net change in bond connectivity between the endpoints, are excluded from further calculation. The remaining candidates are grouped by endpoint connectivity and mapped bond changes. Within each group, Carina single-point energy (SPE)-based screening and farthest-point sampling select low-energy, geometrically diverse representatives, allowing multiple TS arrangements and pathways connecting the same reactants and products to be retained.

After each round of calculations, further network exploration requires selecting which accepted product sets to expand. Selection uses a persistent global candidate pool and a kinetic-guided policy adapted from previous work.¹² Candidates are ranked using local electronic barriers along their best available complete routes from the starting reactants, with the largest barrier contributing most. The configurable branching factor n sets the number of highest-priority eligible candidates selected from this pool per round. Their complete product sets, including co-products and spectators, become new Sampler inputs. The global pool is updated and reranked only after calculations for all inputs selected in the current round are complete. New routes can update priorities, while previously discovered unexpanded candidates remain eligible. Screening, identity assignment and navigation are detailed in Sections 4.4, 4.7 and 4.8.

Reaction dataset. To establish a reaction database with broad chemical coverage, we sampled approximately half a million molecular seeds from PubChem⁴⁷ with an emphasis on chemical diversity. The seed set included both open-shell and closed-shell species, retaining neutral, cationic and anionic charge states as recorded in PubChem. Candidate reactions were then generated through generic bond edits. The core dataset spans 11 elements (H, B, C, N, O, F, Si, P, S, Cl and Br) and covers reactions with up to 15 heavy atoms. Conformational sampling and GFN2-xTB pre-explorations supplied candidate TSs for DFT refinement (Fig. 1b).

To increase the representation of bimolecular reactions, we incorporated 52,169 bimolecular reactions from RGD1⁴⁵ and Reaction-QM.⁴⁸ Water-catalysed reactions were added from previous studies of hydrogen cyanide and glucose reaction networks.^{12,14} Additional 53,315 in-house calculations on macrocycle decomposition, click chemistry and polymer-chain initiation extended the dataset to reaction systems containing up to 25 heavy atoms. These supplemental reactions entered the DFT stage directly: their existing TS geometries were reoptimized and subjected to IRC calculations at the same level of theory. Together, these sources cover multicomponent reactions and configurations away from equilibrium, providing reactant and product structures, optimized TSs, trajectories obtained from IRC calculations, and energy (E), force (F) and Hessian (H) labels. In total, the database contains 1,113,366 reaction records and 130,325,537 configurations with energies; forces and Hessians are available for 124,296,911 and 4,972,426 configurations, respectively. These populations overlap and are not additive. Enumeration is used to construct training examples offline; network generation at inference does not require enumerated products. Database construction and training/evaluation overlap are described in Section 4.1; composition and chemical-space analyses are provided in Supplementary Section S1. We named the resulting database RD11E, reflecting its coverage of 11 elements.

Single-step generation and verification. Network growth first requires proposals that connect the input reactants to new products. For the benchmark, 1,000 reactions were selected from the test set such that their reactant molecules were distinct and absent from the training set. Using only the corresponding reactant graphs, Sampler generated ten TS candidates per reactant, yielding 10,000 candidates in total. Of these candidates, 93.6% yield a valid first-order TS after DFT optimization at the same level of theory used for database construction; IRC calculations recover the input reactant

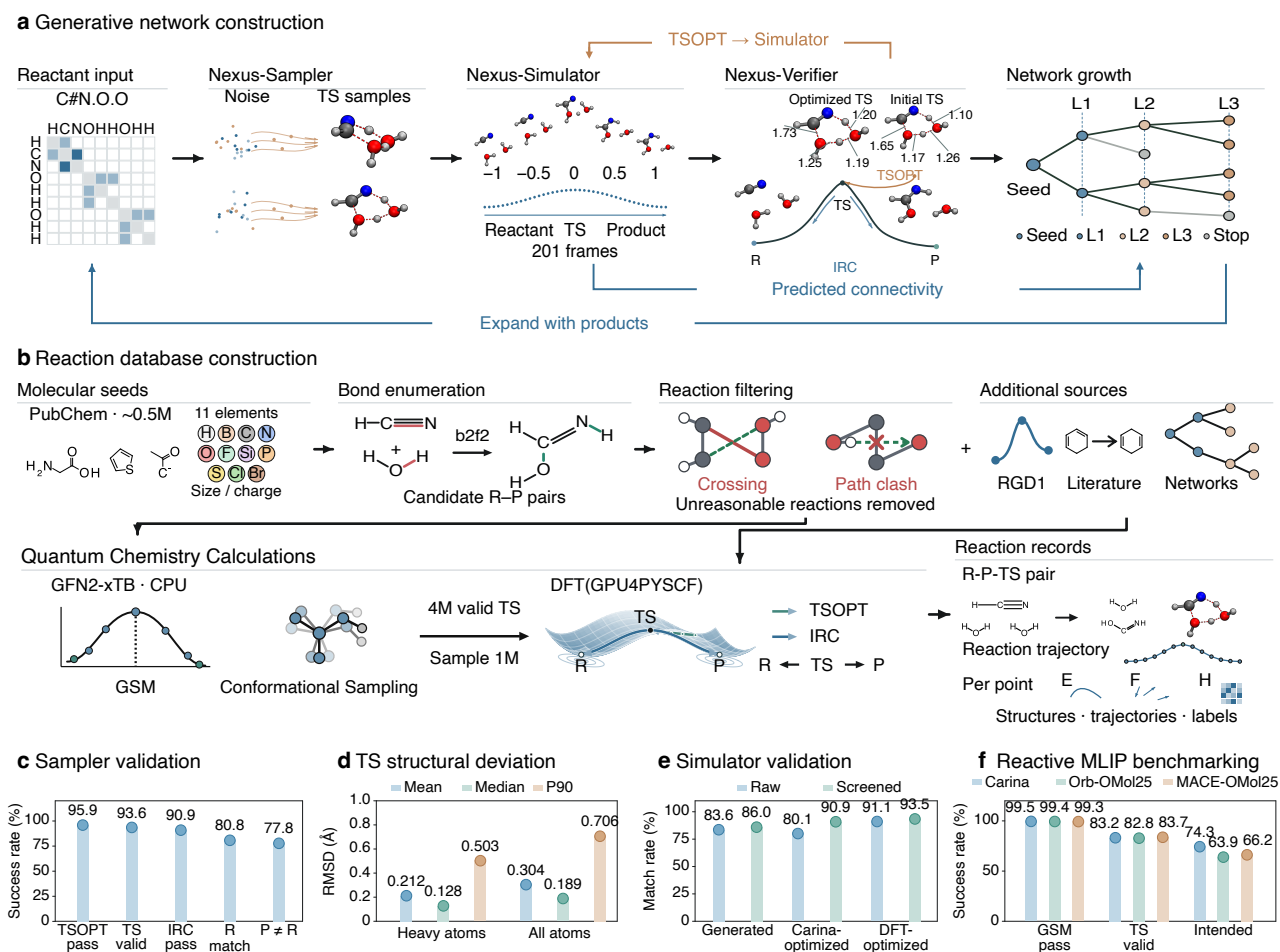


Fig. 1 | Generative reaction-network construction and single-step validation. (a) Reactant graphs condition Nexus-Sampler to generate candidate transition-state (TS) geometries; Nexus-Simulator predicts full reaction trajectories and their endpoints from sampled TSs. Nexus-Verifier uses Carina, a reactive MLIP, for single-point energy (SPE)-based screening, TS refinement and intrinsic reaction coordinate (IRC) calculations. The modules support different levels of refinement, with accepted product sets serving as inputs for further generation. A 201-frame trajectory prediction and an example path obtained from an IRC calculation are shown. Distances are in Å. (b) Database construction from molecular seeds and generic bond edits via conformational sampling, GFN2-xTB pre-explorations and DFT refinement. Additional reaction sources enter directly at DFT TS reoptimization and IRC calculations. Records retain structures and trajectories together with energies (E), forces (F) and Hessians (H). (c) Outcomes of DFT TS optimization, frequency analysis and IRC calculations for 10,000 Sampler-generated TS structures. R match denotes recovery of the input reactant; R ≠ P denotes distinct endpoint connectivities. All percentages use the initial 10,000 TS structures. (d) Displacements from generated TS structures to corresponding DFT-optimized geometries for 9,360 valid TSs; P90 denotes the 90th percentile. (e) Agreement of both Simulator-predicted endpoints with the corresponding DFT IRC endpoints for a common set of 8,622 TS structures, using Sampler-generated, Carina-optimized and DFT-optimized TS inputs. *Raw* includes all 8,622 input structures; *Screened* retains those satisfying R match and R ≠ P based on Simulator predictions. (f) GSM convergence, valid-TS formation and intended-reaction recovery using Carina, Orb-OMol25 and MACE-OMol25 for the same 1,000 reference reactions, with optimized reactant and product endpoints supplied.

for 80.8% and connect it to a distinct product for 77.8% of the initial candidates (Fig. 1c). These are the *valid-TS*, *input-reactant match* and *productive-connection* rates, respectively. The latter directly measures the fraction of proposals that provide a potential new network connection. For the 9,360 valid TSs, median RMSDs between generated and corresponding DFT-optimized geometries are 0.128 Å for heavy atoms and 0.189 Å for all atoms (Fig. 1d). Differences in task formulation and required inputs limit direct comparison with related generative models; contextual comparisons with MolGEN, React-OT and GoFlow^{41,42,49} are provided in Supplementary Section S5.4.

Of the 10,000 Sampler-generated TS candidates, 8,622 had completed DFT IRC calculations and converged Carina optimizations. This common subset was used to compare Simulator predictions from the Sampler-generated, Carina-optimized and DFT-optimized geometries. *Raw* agreement measures

the fraction of predictions matching both endpoints obtained from DFT IRC calculations across all 8,622 candidates: 83.6%, 80.1% and 91.1%, respectively (Fig. 1e). During network growth, only TS candidates whose simulated endpoints connect the input reactants to a distinct product are retained for expansion. To reflect this selection in the benchmark, *Screened* agreement uses the number of retained predictions in each scenario as its denominator, giving 86.0%, 90.9% and 93.5%, respectively. These are conditional agreements within the selected connections. The lower Raw agreement for Carina-optimized inputs is partly due to the inclusion of geometries that fail the strict Carina TS criterion (7.06%). The comparison against DFT IRC endpoints also includes differences between the Carina and DFT potential energy surfaces. Among 7,967 candidates with accepted IRC calculations on both potentials, Simulator predictions match the corresponding Carina IRC endpoints in 90.7% of cases. Population definitions and complete results are provided in Supplementary Section S5.

The same 1,000 reactions were used to benchmark reactive MLIPs following the procedure of ReactBench.³⁰ Reactant and product inputs were taken from geometries obtained by optimizing the corresponding IRC endpoints. Carina recovers both specified endpoints by IRC for 74.3% of reference reactions, compared with 63.9% for Orb-OMol25 and 66.2% for MACE-OMol25. Higher rates of GSM convergence and valid-TS formation are also obtained with Carina (Fig. 1f).^{34,50}

2.2 Reaction-network coverage and conformational refinement

The decomposition of γ -ketohydroperoxide (KHP) is an established benchmark for automated reaction discovery, with characterized pathways against which to assess coverage of competing decomposition channels and their mechanisms.^{3,26,49,51} ReactNexus generated a two-layer network by expanding the five first-layer product nodes reached through the lowest-barrier pathways (branching factor $n = 5$; Fig. 2a). The recovered chemistry includes the classical Korcek sequence through a five-membered cyclic peroxide, dehydration and direct fragmentation. The relative kinetic ordering of the previously reported channels agrees with earlier YARP-based network exploration, apart from enolization (purple in Fig. 2a), which becomes more competitive through a hydroxyl-assisted intramolecular hydrogen-transfer mechanism (Fig. 2b).²⁶ Retaining the decomposition co-products also allows the generated water and formic acid to participate in subsequent reactions involving proton transfer, revealing water- and acid-catalysed pathways.⁵²

These pathways illustrate how cooperative proton transfer can provide lower-barrier mechanisms connecting the same reactants and products (Fig. 2c). For acetaldehyde enolization, formic acid uses different oxygen atoms to donate and accept protons, coupling these transfers to C–O bond-order reorganization before being regenerated. This b4f4 pathway has a Carina-computed electronic barrier of 26.9 kcal/mol, compared with 45.9 kcal/mol for the b3f3 alternative, in which a single acid oxygen performs both roles. Similarly, addition of formic acid to acetaldehyde proceeds through a b3f3 pathway coupling proton transfer with C–O bond formation and bond-order reorganization, with a barrier of 16.4 kcal/mol, compared with 54.1 kcal/mol for the b2f2 bond-exchange pathway to the same product. Each comparison uses the same optimized reactant reference. Such cooperative mechanisms involve several coupled reactive coordinates and can fall outside restricted bond-edit enumeration or be difficult to locate through single-ended searches. ReactNexus samples these cooperative mechanisms without prescribing their bond-change patterns or proton-transfer sequences.

For larger, conformationally flexible molecules, each reaction step presents numerous competing pathways and a vast TS conformational space that is difficult to sample comprehensively. This complexity increases during multistep exploration, as each newly formed intermediate opens further competing pathways and multiplies the number of possible reaction sequences. The unexpected Ireland–Claisen (IC) rearrangement cascade observed during studies towards the synthesis of the natural product curcuscione C was mechanistically elucidated by Lee and colleagues using trial-and-error calculations and AFIR searches.⁵³ It subsequently served as a benchmark for automated TS searches and reaction conformational sampling, with reactant and product structures supplied for the individual steps.^{54,55} To control the cost of exploring this larger system, one closed-shell input was selected for expansion per round ($n = 1$), using the barrier-based score of its best complete route from the starting reactant.

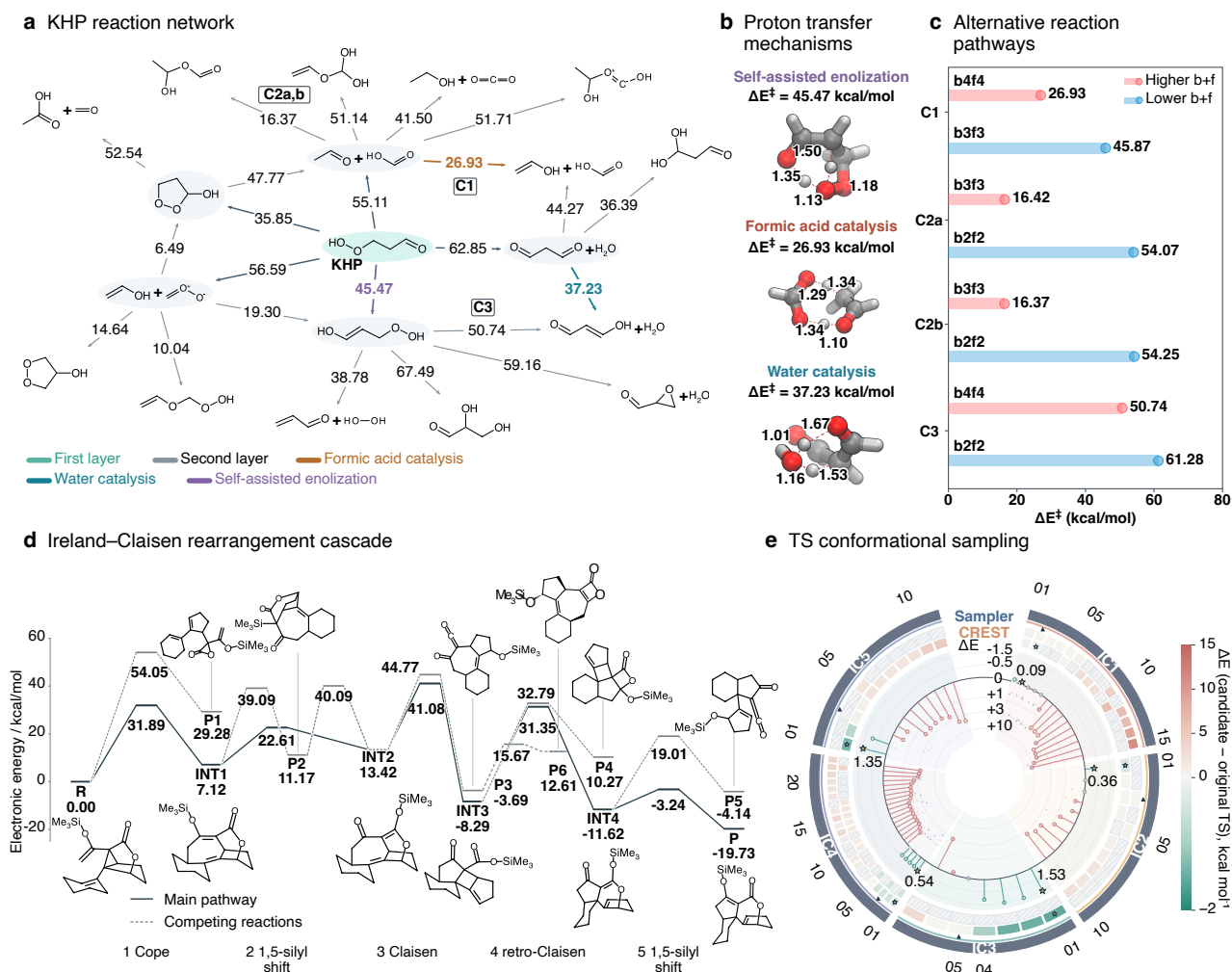


Fig. 2 | Mechanism discovery and conformational refinement in reaction-network exploration. (a) Two-layer γ -ketohydroperoxide (KHP) network, expanding the five first-layer product nodes reached through the lowest-barrier pathways ($n = 5$). Co-products remain available for subsequent reactions. (b) Representative TSs for self-assisted, formic-acid-assisted and water-assisted proton transfer. Colours correspond to highlighted pathways in a; distances are in Å. (c) Alternative pathways connecting the same reactants and products. Labels $b_n f_n$ count broken and formed bond-order units, including explicit hydrogen transfers. (d) Five-step Ireland-Claisen (IC) reference sequence and selected competing connections. One closed-shell input is selected per global-pool expansion ($n = 1$); the eleventh expanded input, INT4, supplies the connection to P. Dark solid and grey dashed profiles denote the reference sequence and competing reactions. Platforms show TS and product reference energies relative to the initial R. (e) TS conformational coverage from Sampler candidates and constrained sampling after Carina refinement (226 records in 66 geometrical groups). Tracks show source-specific group minima and points individual records; triangles and stars mark initial reference and selected TSs. ΔE is relative to the lowest TS retained for each reaction in the initial exploration. Central values give barrier reductions with the corresponding reactant reference fixed. All energies are Carina-computed electronic energies in kcal/mol.

Selection draws from the entire persistent candidate pool, retaining competing branches for later exploration. After 11 rounds of expansion, including exploration of competing branches, the network reaches the final product P (INT5) through the five-step sequence of Cope and Claisen/retro-Claisen rearrangements and 1,5-silyl migrations (Fig. 2d). The competing pathway from INT2 couples O-to-O silyl transfer with C–O cleavage to form a ketene-containing product (P3). From INT3, silyl transfer coupled to intramolecular C–O bond formation yields a β -lactone (P4), providing an alternative to the retro-Claisen rearrangement. A distinct ring closure from P3 forms the four-membered lactone P6, while the P2 branch supplies a connection back to INT2. Competing chemistry therefore includes both new skeletons and routes rejoining the reference sequence. The selected profiles compare reaction pathways using the selected reactant conformers as energy references; stereochemical differences between successive calculations and the separately explored P6 diastereomer are distinguished in Section 4.8.

For a given reaction connection, distinct TS conformations can yield different barrier estimates. Two complementary approaches broadened the set of TS seeds beyond the original screened representatives. First, Sampler candidates previously screened out to reduce computational cost were recovered for supplementary refinement. The full set of 540 candidates assigned to the five reference transformations was taken forward for Carina TS optimization and IRC calculations without reapplying the original multistage screening. Second, constrained conformational sampling generated additional seeds around an initially verified TS with the reactive core restrained.^{56,57} All restraints were released before Carina TS refinement and subsequent IRC calculations (Fig. 2e). The two approaches populate both shared and distinct geometrical groups. Three of the five lowest retained TSs were obtained from constrained sampling, and two from additional refinement of Sampler candidates. With the reactant reference held fixed for each step, these searches lower the Carina-computed electronic barriers by 0.09–1.53 kcal/mol relative to the lowest TSs retained for the corresponding reactions in the initial exploration. Combining the two approaches reduces reliance on a single TS conformation when assigning the barriers used for network exploration. Sampling settings, geometrical grouping and numerical diagnostics are detailed in Supplementary Section S7.1. For comparison with earlier work, B3LYP-D3 single-point calculations with continuum n-hexane solvation were performed on the Carina-optimized structures. The comparison and computational differences are provided in Supplementary Section S7.2.

2.3 From multicomponent reactants to catalyst regeneration

Multicomponent catalysis couples substrate activation and bond formation to product release and catalyst regeneration. With several components present, the catalyst can interact with different substrates, while newly formed intermediates and co-products open additional reaction pathways. Exploring the resulting network requires identifying pathways to product formation and catalyst regeneration, together with competing reactions of the substrates and catalyst. Proline-mediated aldol addition and phosphine-catalysed allenolate–acrylate coupling provide contrasting modes of substrate activation for examining this challenge. Both networks were explored from two substrates and an organocatalyst (Fig. 3).

The proline-mediated aldol network was seeded with acetone, benzaldehyde and proline (L0-1; Fig. 3a). Exploration identified covalent adducts involving all three components [L1-2(S)], followed by formation of an acetone-derived iminium and benzaldehyde hydrate (L2-3). Coupled proton and hydroxyl-group transfer from benzaldehyde hydrate to the acetone–proline intermediate regenerates benzaldehyde and yields the carbinolamine (L3-3). Subsequent dehydration forms the iminium intermediate and releases water (L4-1), followed by proton transfer to form the enamine (L5-3). Benzaldehyde therefore participates in catalyst activation as well as serving as the subsequent C–C coupling partner. Enamine C–C addition gives L6-2(S,E/Z) and L6-3(R,Z). Subsequent proton transfers and hydration lead to carbinolamines L9-2(S/R), whose C–N cleavage releases the aldol products and regenerates proline [L8-2(S/R)]. The established catalytic sequence is thus recovered within the generated network.^{58–60}

Alongside these three-component activation pathways, the network also contains the conventional acetone–proline carbinolamine-forming pathway to L3-3, with benzaldehyde remaining a spectator. Its electronic TS height is 39.0 kcal/mol, compared with 22.7 kcal/mol for formation of L1-2(S) and 18.1 kcal/mol for a distinct three-component adduct, all relative to the same initial mixture of acetone, benzaldehyde and proline. Sampling acetone, benzaldehyde and proline together therefore accesses both two-component and three-component reactions, with pathway barriers guiding which products are explored next.

A separate route couples the free enol of acetone with benzaldehyde, with free proline serving as a proton shuttle (L7-1). During C–C addition, the carboxylic acid group of proline accepts a proton from the enol and donates one to the aldehyde oxygen, yielding the aldol products and regenerating free proline [L8-2(S/R)]. The network thus connects covalent enamine catalysis with an alternative mode of C–C formation assisted by free proline.

Removing water from L5-3 gives the enamine–benzaldehyde reactant pair of the classical Houk

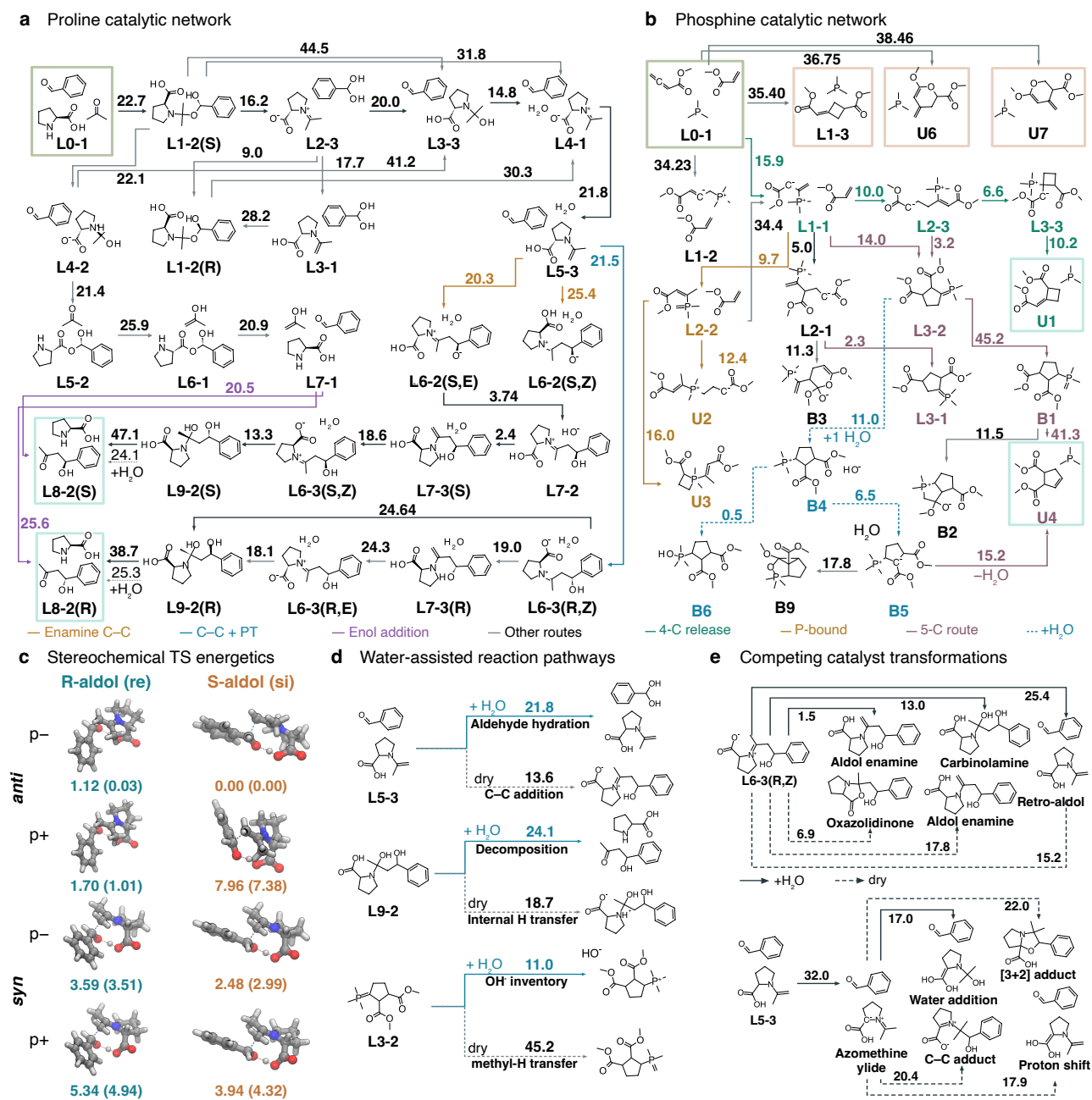


Fig. 3 | Connecting multicomponent reactants to product formation and catalyst regeneration. (a) Proline-mediated aldol network from acetone, benzaldehyde and proline. Orange edges indicate enamine C–C addition, and blue edges indicate acid–water–contact pathways. Purple edges show C–C addition between the free enol of acetone and benzaldehyde, with free proline mediating proton transfer through its carboxylic acid group. Grey edges denote other sampled connections. (b) Reaction network generated from methyl allenolate, methyl acrylate and PMe_3 . Green and muted purple edges indicate four-membered-carbocycle and five-membered-ring pathways, respectively. Ochre edges indicate competing pathways involving catalyst methyl groups and phosphorus-containing rings. Blue denotes supplementary pathways explored with added water. $\text{L3-2} \rightarrow \text{B4} \rightarrow \text{B5}$ transfers a substrate proton through water and regenerates water. After removal of water, further calculations from B5 located the elimination to U4 . (c) Eight R-proline C–C TS geometries classified by syn/anti enamine–carboxylic-acid arrangement, aldehyde re/si face and ring conformation ($p-$: endo; $p+$: exo). Numbers are approximate relative free energies, with relative DFT electronic energies in parentheses, based on $\omega\text{B97X-D4/def2-TZVPD/SMD(DMSO)}$ calculations on Carina-optimized geometries. (d) Changes in sampled pathways from selected intermediates with and without added water. In the phosphonium-ylide example, proton transfer from water to the substrate (11.0) and hydrogen transfer from a phosphine methyl group (45.2) lead to different products. (e) Competing reactions of proline-derived condensation intermediates and a COOH-retaining azomethine ylide. Solid and dashed arrows indicate reactions with and without water, respectively. Arrow values are local Carina-computed electronic barriers (kcal/mol), each relative to the corresponding optimized reactants. L labels denote exploration batch/rank; B, supplementary intermediates; U, unexpanded products.

model.^{59,60} An independent sampling test on this water-free system generated 10,000 TS proposals and recovered all eight reference conformational families for R-proline, whereas seven were observed for S-proline. The complete R-proline set is shown in Fig. 3c; corresponding whole-system mirror pathways for the two catalyst enantiomers are energetically equivalent in an achiral environment. DFT energies and thermal corrections were evaluated on the Carina-optimized geometries. The lowest anti structures lie below the syn alternatives, consistent with the qualitative trend obtained from the TSs manually constructed and optimized by Houk and co-workers.⁶¹ Recovery of these established TS families provides a test of conformational sampling coverage beyond identification of the reaction connection.

The phosphine-catalysed network was seeded with methyl allenolate, methyl acrylate and PMe₃. Phosphine addition followed by C–C bond formation opens several ring-forming pathways (Fig. 3b). One dry branch closes a four-membered carbocycle and regenerates intact neutral PMe₃ (U1; final-step barrier 10.2 kcal/mol), demonstrating autonomous recovery of a catalyst-release pathway. A separate branch reaches the five-membered cyclic phosphonium ylide L3-2, where proton transfer is required before elimination to the classical [3+2] product.

Direct substrate [1,2]-hydrogen transfer from L3-2 has a barrier of 43.1 kcal/mol (Supplementary Section S8.3). This differs from the 45.2 kcal/mol transfer from a phosphine methyl group, which gives a different product (Fig. 3d). Earlier experimental and computational studies identified water-assisted proton transfer in related phosphine-catalysed cycloadditions, motivating extension of the network from this intermediate with one explicit water molecule.^{62,63} Water first protonates the ylide to form B4 (11.0 kcal/mol); the resulting hydroxide then removes the adjacent substrate proton, forming B5 and regenerating water (6.5 kcal/mol). This sequence relocates the carbanion to the carbon adjacent to the phosphonium-bearing carbon, providing the arrangement required for elimination. After removal of the regenerated water, further calculations from B5 locate release of PMe₃ and the five-membered cyclopentene U4 (15.2 kcal/mol), below the lowest sampled competitor B9 (17.8 kcal/mol) relative to the same optimized reactants. The expanded network therefore provides further computational support for the proton-shuttle role of water identified in earlier mechanistic studies.⁶³

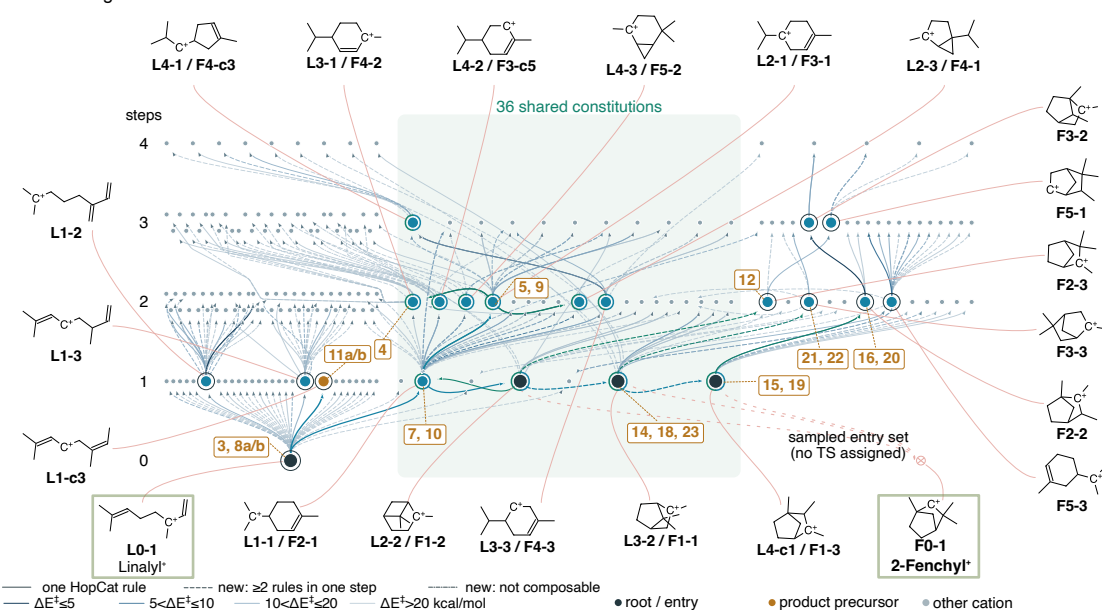
Water-assisted regeneration also competes with reactions of the catalyst itself. From B4, P–O bond formation between hydroxide and the phosphonium centre yields B6 with a barrier of 0.5 kcal/mol, below the proton-transfer route to B5 (6.5 kcal/mol), and subsequent reactions lead to phosphine oxide (Fig. 3b,d). Water therefore opens competing routes towards regeneration and phosphorus oxidation. Other phosphine branches involve catalyst methyl groups or phosphorus-containing rings. Analogous branching appears in the proline network: condensation intermediates undergo retro-aldol cleavage or cyclize to oxazolidinones, while azomethine ylides undergo further C–C addition and cycloaddition (Fig. 3e). Together, these pathways reveal alternative fates of catalytic intermediates alongside regeneration.

2.4 Distinct carbocation entries converge onto a shared rearrangement network

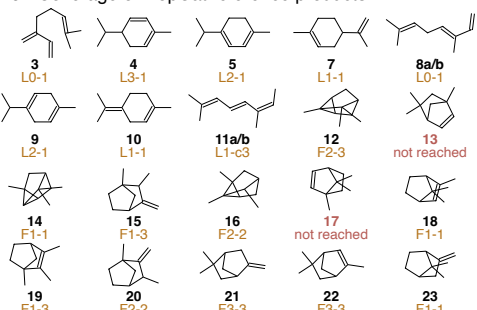
Carbocation rearrangements provide a setting to examine the extent of chemically relevant reaction space. In other applications, HopCat illustrates the scale of formal rearrangement space, with reported networks of 3,404 nodes for a taxanin rearrangement and 909,937 nodes for a mixed forward–retro biosynthetic search.⁶⁴ Here, linalyl and fenchol-derived cations provide contrasting acyclic and bridged entries for assessing how much product-related chemistry can be recovered by kinetic-guided exploration without enumerating every formally allowed branch.^{12,65}

The two explorations converge onto shared cation skeletons (Fig. 4a). Cyclization and subsequent rearrangements from linalyl cation reach α -terpinyl (L1-1/F2-1), 2-pinanyl (L2-2/F1-2) and fenchyl-type (L3-2/F1-1) skeletons. Ring opening and skeletal migration from the fenchol-derived entries also reach α -terpinyl and terpinen-4-yl (L2-1/F3-1) cations. HopCat quenching rules map these cations to the experimental reference products (Fig. 4b), excluding oxidative p-cymene formation from the linalool comparison. The displayed networks contain 153 and 72 nodes after grouping stereoisomers with the same connectivity and recover precursors for 8/8 non-oxidative linalool and 10/12 fenchol reference

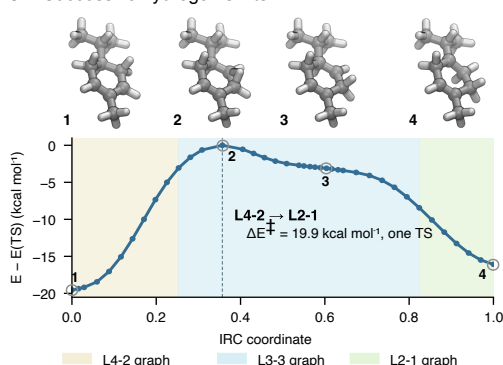
a Convergent carbocation networks



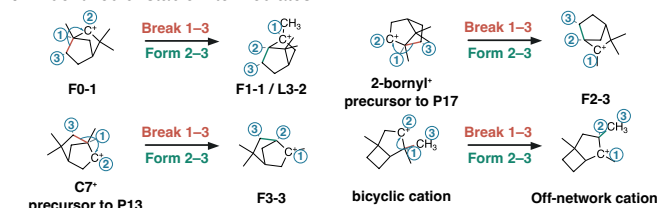
b Coverage of HopCat reference products



c Successive hydrogen shifts



d Identified unstable intermediates



e Rearrangement mechanism classes

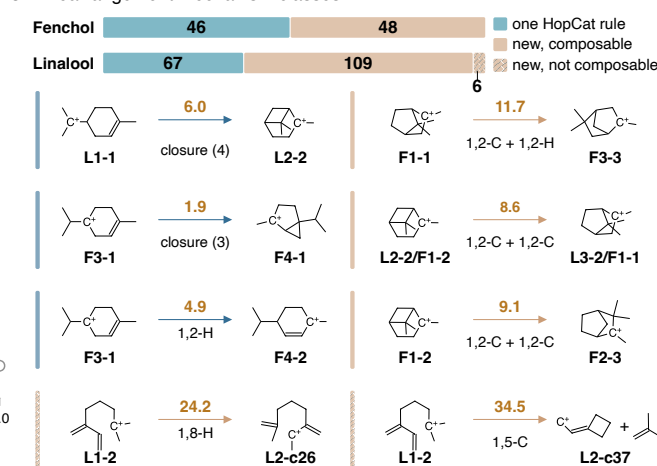


Fig. 4 | Convergence of carbocation rearrangement networks and identification of unstable intermediates.

(a) Reaction networks from linalyl and fenchol-derived cations: 189 nodes, including 36 shared structures drawn once. F0-1 serves only to initialize Sampler-Simulator entry proposals, whose links have no assigned TSs. Stereoisomers with the same connectivity share one node. L/F labels identify the two explorations, and vertical position denotes displayed network layers. Teal outlines mark expansion sites, while pink curves link molecular structures to their corresponding nodes. Edge width and shade encode local barriers. Solid, dashed and dash-dotted edges denote single-rule, composable multiple-rule and non-composable changes under the tested HopCat rule set, respectively. (b) Applying the quenching templates identifies cation precursors for 8/8 non-oxidative linalool and 10/12 fenchol reference products. (c) Carina IRC energy profile and selected structures for L4-2 $\rightarrow$ L2-1. Successive hydrogen shifts connect L4-2 to L2-1 without a separate intermediate minimum. Shaded regions denote connectivity classes. (d) DFT relaxation of F0-1, C7⁺ and 2-bornyl⁺ reproduces their Carina rearrangement outcomes. A sampled bicyclic cation provides an additional control. Red and green bonds mark cleavage and formation; blue arrows indicate inferred electron-pair displacement. (e) Classification of directed source-product pairs before suppression of redundant reverse arrows, with representative local barriers. *New* includes changes composable from multiple rules and changes not composable under the tested rule set. Network barriers and IRC energies are Carina-computed gas-phase electronic energies (kcal/mol). Neutral-product quenching is rule-based and does not establish termination barriers or yields.

products, respectively. This reference-product coverage is already reached within three expansion rounds.

Beyond recovering product-related chemistry, kinetic-guided exploration can test whether intermediates inferred from rearrangement templates persist on the calculated potential-energy surface. For fenchol, the nominal 2-fenchyl cation F0-1 relaxes to F1-1 in both Carina and DFT optimizations (Fig. 4d). Sampler and Simulator were therefore used to generate alternative entry structures, and the retained F1-1, F1-2 and F1-3 structures seeded separate network expansions (Fig. 4a). The HopCat-assigned precursors of the two unreached fenchol reference products, δ -fenchene (13) and bornylene (17), are likewise not retained as distinct nodes: C7⁺ and 2-bornyl⁺ relax to F3-3 and camphenyl cation F2-3, respectively, in both methods. These relaxation outcomes do not exclude product formation through deprotonation coupled to rearrangement. Such routes remain unresolved because quenching is assigned by rules rather than calculated explicitly. A stepwise template sequence can also imply an intermediate that is absent along an explicit reaction path. For L4-2 $\rightarrow$ L2-1, successive hydrogen shifts form a continuous elementary reaction step containing a single TS (Fig. 4c). Along this path, the connectivity associated with the HopCat-proposed intermediate appears only transiently, without forming a separate intermediate minimum. The classification in Fig. 4e places these findings in the broader context of HopCat rules. Most directed net changes match a single encoded rule or can be composed from multiple rules. Low-barrier examples include net carbon and hydrogen migration in F1-1 $\rightarrow$ F3-3 (11.7 kcal/mol) and two carbon migrations in F1-2 $\rightarrow$ F2-3 (9.1 kcal/mol). Only six connections are not composable under the tested rule set. Their barriers are comparatively high (24.2–51.9 kcal/mol), and none lies on an identified reference-product route. This supports the breadth of the tested HopCat rules along the identified product routes, while explicit pathways clarify the stability and connectivity of the intervening species.

2.5 Candidate generation and pathway competition in a multisite reaction

Reaction networks can also aid the interpretation of experimental reaction mixtures by linking candidate species to their possible formation pathways.^{11,66} We examined whether a network generated directly from reactants could help explain unresolved LC–MS signals and guide further experiments. Initial LC–MS measurements of the reaction of 2,6-dichloro-4-methylnicotinonitrile (A) with hydrazine (B) contained signals that could not be explained by the expected monosubstitution products alone, including the M332 mass group (Fig. 5c). Assigning these signals is challenging because both reactants offer multiple reactive sites: A contains two activated C–Cl bonds and a nitrile group, while either nitrogen of hydrazine can participate in further reaction.

Starting from A and B, ReactNexus identifies C2 and C6 substitution to P₂ and P₆, respectively, both in the M182 family, with Carina electronic barriers of 24.4 and 25.4 kcal/mol (Fig. 5a,b). Their remaining C–Cl and N–H groups open two distinct types of continuation. Reaction with another molecule of A gives the M332 diarylation candidates P_{332a-f} , whereas reaction with additional B gives the disubstituted candidate P₁₇₈ from P₆. The network thus connects the initial A + B reaction to products derived from 2A + B and A + 2B. The displayed 1,2-N-diarylation routes to P_{332a-c} lie below the competing 1,1-N-diarylation routes to P_{332d-f} . For P₆ $\rightarrow$ P₁₇₈, the local barrier is 25.4 kcal/mol; the plotted TS level of 11.3 kcal/mol results from alignment to the displayed P₆ level. These second-step profiles use separate reactant references for the added A or B.

The network-derived candidates guided a re-examination of the LC–MS data, revealing an additional signal consistent with the M178 family. This finding prompted follow-up experiments at 100 °C to examine continued substitution of a monosubstituted intermediate. M178 signals were very weak at 70 °C, weak after 1–2 h at 100 °C and strong after 4 h, whereas M332 signals were strong at 70 °C and weaker at 100 °C (Fig. 5c). Thus, computational exploration identified candidate products, directed the search for an additional experimental signal and motivated experiments at a higher temperature. The measurements support access to the continued-substitution product family; individual regioisomers and N-substitution patterns remain to be distinguished experimentally.

The network also connects substitution and nitrile addition to cyclization. Closure from P₂ and

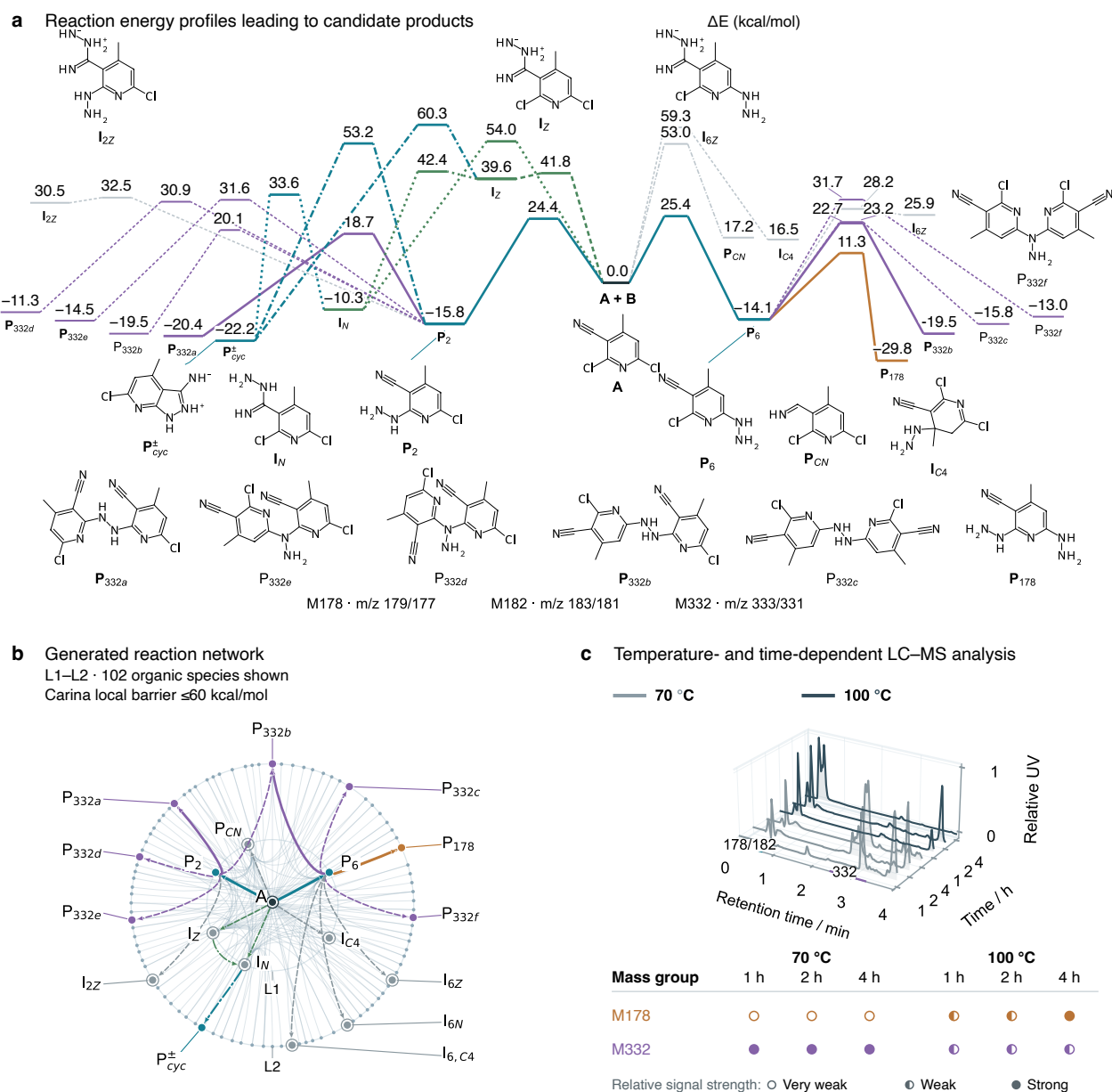


Fig. 5 | Reaction pathways and experimental analysis of a multisite reaction. (a) Gas-phase Carina electronic-energy profiles (kcal/mol) and candidate structures from 2,6-dichloro-4-methylnicotinonitrile (A) and hydrazine (B). Teal, purple and orange pathways indicate the M182 (P_2 , P_6 and $P_{\text{cyc}}^{\pm}$), M332 (P_{332a-f}) and M178 (P_{178}) candidates, respectively. Green highlights routes through I_Z and I_N . P_{332a-c} and P_{332d-f} are 1,2-N- and 1,1-N-diarylation candidates. Second-step reactions use additional A for diarylation and B for further substitution or nitrile addition. Energy profiles are aligned to the displayed parent-product levels; local barriers use each step's own optimized reactant reference. (b) Two-layer network showing 102 organic species. The full network contains 108 species and 122 reaction records with local barriers ≤ 60 kcal/mol. Radial position indicates exploration layer. (c) LC–MS analysis after 1, 2 and 4 h for A (1 equiv.), hydrazine (1.2 equiv.) and K_2CO_3 (3 equiv.) in DMF at 70 or 100 °C. PDA traces at 254 nm are individually normalized. The displayed ion pairs are m/z 179/177, 183/181 and 333/331 for M178, M182 and M332, respectively. The matrix summarizes M178 and M332 signals: open, half-filled and filled circles indicate very weak, weak and strong signals, respectively, for qualitative comparison within each mass group.

routes through the nitrile-addition intermediate I_Z , either directly or through I_N , reach the charge-separated cyclic candidate $P_{\text{cyc}}^{\pm}$. These higher-energy alternatives broaden the mechanistic hypotheses beyond successive substitution. The reported CuI-promoted formation of pyrazolo[3,4-b]pyridines from 2-chloro-3-cyanopyridine and hydrazines provides a precedent for examining catalytic conditions that could favor this class of cyclization.⁶⁷

The initial site preference is sensitive to the energy model and solvation. Although Carina slightly favors C2 substitution, independent DFT single-point calculations with SMD(DMF) on Carina geometries favor C6 substitution, with barriers of 14.25 and 12.93 kcal/mol for P_2 and P_6 , respectively.

These results illustrate a practical role for generative reaction networks in experimental discovery. Unexplained signals motivate exploration beyond the anticipated transformation, while the resulting candidate structures and formation pathways provide hypotheses that can be examined by changing reaction conditions. Here, the network connects additional mass groups to competing continuations of the initial substitution, and temperature- and time-dependent measurements support access to one of these product families. Network generation can therefore guide an iterative exchange between experimental observation and mechanistic hypothesis testing.

3 Discussion

In this work, we establish a generative formulation of reaction-network exploration that constructs unknown networks directly from reactants, without predefined products, reaction templates or enumerated reaction candidates. Pathway generation and verification identify new species and their connections, and kinetic-guided selection uses these discoveries to determine subsequent exploration. ReactNexus thus builds networks through successive mechanistic predictions, including for multicomponent and flexible systems whose relevant reactions are difficult to specify in advance.

Because newly formed species remain available for further reaction, network growth can reveal co-product-assisted proton transfer alongside lower-barrier mechanisms involving cooperative bond changes (Fig. 2). Applied to catalysis, this connects regeneration with competing reactions, complemented by targeted exploration of water-mediated proton shuttles (Fig. 3). Explicit pathways additionally distinguish transient connectivity from stable intermediates within overlapping carbocation networks that recover most reference-product precursors from a compact set of species (Fig. 4). These mechanistic connections provide a basis for linking candidate product families to unexplained LC-MS signals and designing experiments to test their formation (Fig. 5).

Translating these mechanistic hypotheses into quantitative predictions of experimental outcomes requires a more complete treatment of energetics and reaction conditions. Carina currently lacks an implicit-solvent treatment, and thermochemical accuracy has not been a primary focus. Reaction-path training data across solvent environments and validated vibrational free-energy corrections are therefore needed. MLIP Hessians provide approximate corrections, but low-frequency modes require particular scrutiny, together with consistent treatment of conformational populations and concentrations.^{33,68}

The range of reactions that can be explored reliably also depends on how electronic states and stereochemistry are represented. Diradical pathways remain insufficiently sampled and validated, including their competition with closed-shell ring formation. The reactant graphs used here do not impose atom-chirality or alkene E/Z constraints, so both proline enantiomers are sampled even when an enantiopure catalyst is intended. Explicit stereochemical conditioning and broader open-shell validation are therefore needed. Extending the framework to transition metals will additionally require training data spanning their coordination environments and electronic states. We anticipate that generative reaction-network exploration will open new opportunities for testing mechanistic hypotheses, discovering reactions and guiding experimental design.

4 Methods

ReactNexus couples reactant-conditioned TS generation, trajectory prediction and energetic verification with kinetic-guided selection of subsequent inputs. The methods below define the training data, model roles, scientific admission criteria and network-selection procedure. Full model specifications, hyperparameters, algorithms, campaign settings and application-specific controls are collected in the Supplementary Information.

4.1 Reaction database construction and evaluation overlap

Approximately half a million molecular seeds were sampled from PubChem with an emphasis on chemical diversity.⁴⁷ Selection retained neutral, cationic and anionic species, including open-shell and closed-shell systems, rather than generating charge states by protonating or deprotonating each seed. The core systems contained up to 15 heavy atoms drawn from H, B, C, N, O, F, Si, P, S, Cl and Br. Candidate reactions were constructed offline by generic bond edits involving up to three broken and three formed bonds, extending the b2f2 and b3f3 elementary reaction steps used in YARP.^{3,26} A tree-based filter removed unphysical candidates.⁵⁵ Multimolecular examples included pairs of reactants and systems containing explicit catalysts or solvent molecules. Enumeration supplies training examples; it is not required during reactant-only network generation.

Core candidates were characterized with the YARP-derived reaction characterization workflow. Reaction conformational sampling, GFN2-xTB pre-exploration, growing-string calculations, TS optimization and IRC calculations yielded approximately four million reactions with valid TSs.^{55,69} Growing-string calculations used pyGSM in double-ended mode (DE-GSM),^{70,71} and TS optimization and IRC calculations were performed with pysisyphus.⁷² Approximately one million were selected for gas-phase DFT TS optimization and IRC calculations at the ω B97X-D4/def2-TZVPD level using GPU4PySCF 1.4.1.^{73–75} The functional/dispersion specification was `wb97x-d4:wb97x-2008`. Supplemental TSs entered this DFT stage directly and were reoptimized at the same level.

Bimolecular reactions from RGD1 and Reaction-QM supplemented the core data, while previous studies of hydrogen cyanide and glucose supplied water-catalysed reactions.^{12,14,45,48} In-house macrocycle-decomposition, click-chemistry and polymer-chain-initiation calculations extended the upper size limit to 25 heavy atoms. The database contains 1,113,366 reaction records and 130,325,537 configurations with energies; forces and Hessians are available for 124,296,911 and 4,972,426 configurations, respectively. These populations overlap and are not additive. Before training, records with anomalous endpoints, discontinuous energy profiles or severely distorted optimization geometries were removed. Retained reactions require positive TS-to-endpoint energy differences no greater than 150 kcal/mol for both IRC endpoints and both optimized endpoints. The database totals and source breakdown are reported separately from the cleaned training populations in Supplementary Section S1.

All three models use the same reaction-level training–test split, with validation data drawn from the training partition. Evaluation independence additionally requires examining complete reactions and reactant sets. Reaction-level overlap concerns the complete reactant–product connection, including the reverse direction, rather than a shared TS conformer or reaction label alone. Reactant-set overlap concerns the same complete molecular composition and state, including all co-products and spectators, stoichiometry, charge and spin, irrespective of which outgoing reaction is considered. Reuse of an individual molecule or fragment is recorded separately and does not establish either complete-reaction or reactant-set overlap. Correspondence must respect mapped connectivity and electronic-state information, with stereochemical equivalence stated separately.

For the single-step benchmark, 1,000 test reactions were selected with distinct reactant molecules absent from training. Reactions from the KHP network, the IC rearrangement and the proline- and PMe_3 -catalysed systems were not included in the training data.

4.2 Sampler construction and training

Sampler generates all-atom TS geometries conditioned on the complete reactant graph, total charge and spin multiplicity, without product information or a three-dimensional reactant geometry. The graph can contain several disconnected molecular components. A harmonic prior places springs along reactant bonds and keeps the components in a common spatial region.^{76,77} A flow-matching model transforms independent prior samples into TS candidates, enabling different reaction channels and conformations to be generated from the same input.^{78,79}

The model uses the DiTMC transformer, with attention across atom pairs, including pairs from different molecules.⁸⁰ Training aligns the prior geometry to the reference TS, samples noise-smoothed intermediate structures and learns to recover the target TS. Either endpoint graph can condition a training TS, exposing both reaction directions. At inference, the learned flow is integrated from an independent prior sample in 50 Euler steps. The full graph features, prior, learning objective and integration scheme, together with training and sampling algorithms, are specified in Supplementary Section S2. Independent prior draws supply sampling diversity; their frequencies are not interpreted as reaction rates or product populations.

4.3 Simulator construction and inference

Simulator predicts a reaction trajectory from a supplied TS geometry, atomic elements, total charge and spin multiplicity. Its Equiformer-V3 backbone predicts displacements from the TS as a function of a normalized progress coordinate, reproducing the supplied structure at the central frame.⁴⁶ This coordinate describes progress along a path, not physical time. Training uses trajectories obtained from DFT IRC calculations, resampled independently to 100 frames on each side of the central TS. The positive branch is oriented towards the lower-energy endpoint during training to avoid inconsistent directional labels.

The objective combines atomic-position and interatomic-distance errors; perturbed input TS geometries provide tolerance to unoptimized candidates. A dense prediction contains 201 frames, while endpoint screening can evaluate only the TS and two terminal frames. All frames are predicted from the same input without iterative force-based propagation. Generated, Carina-optimized and DFT-optimized structures can therefore be used as inputs. Simulator supplies endpoint assignments and initial geometries for subsequent checks; it does not predict energies. Its architecture, objective, training settings and endpoint-perception procedure are given in Supplementary Section S3.

4.4 Screening and representative selection

Rapid exploration combines Sampler, Simulator and Carina single-point energies without TS optimization and IRC calculations. When further verification is required, selected TS structures are optimized on the Carina potential, followed by Simulator prediction or IRC calculations and endpoint verification. DFT checks are separate operations. To reduce the number of expensive refinements, generated structures first undergo geometrical clustering and representative selection. Near-duplicate candidates are excluded from further processing while their original geometries and selection records are preserved. Up to k geometrically diverse representatives per cluster are sent to Simulator.

Simulated trajectories are retained when exactly one endpoint recovers the input adjacency and the other has different connectivity. Trajectories describing only conformational change or returning to the input connectivity at both ends are excluded, as are trajectories that recover neither input endpoint. Retained candidates are grouped using mapped endpoint states and bond-change signatures. Different mechanisms connecting the same reactants and products are not automatically merged. Within each family, Carina single-point energy (SPE)-based screening and farthest-point sampling select low-energy, geometrically diverse representatives, preserving multiple TS arrangements.

For example, in the phosphine-catalysed reaction network, initial complete-linkage clustering used a 0.15 Å heavy-atom RMSD threshold and $k = 3$. The SPE window extended 50 kcal/mol above

the lowest finite candidate SPE in the entire input batch; it was not a barrier relative to a reactant minimum. Up to five representatives per pathway family were retained, seeded by the lowest-SPE candidate. These operational values are campaign-specific. Refined connections are assigned from their actual calculated endpoints, which can differ from the initial prediction. Supplementary Section S6.1 specifies the selection sequence, distance definitions, settings and separately recorded rescue calculations.

4.5 Carina and stationary-point verification

Carina is an Equiformer-V3 reactive MLIP with charge and multiplicity embeddings.⁴⁶ It was trained on cleaned configurations from reactions in the RD11E training partition using energy and force targets; Hessians were not used as training targets at this stage. Forces and Hessians are obtained by differentiating its predicted energy. Elemental reference energies are added to the learned atomic-energy target to recover total energies. Full architecture, training settings and evaluation populations are provided in Supplementary Sections S4 and S5.

The recorded production protocol uses redundant internal coordinates and restricted-step image rational-function optimization in pysisyphus, with native Gaussian `gau` convergence criteria, up to 50 steps and Hessian recalculation every five steps.⁷² Native convergence and the imaginary-frequency criterion are assessed separately. Production records use a significant imaginary-frequency threshold of -10 cm^{-1} ; the strict single-step benchmark uses exactly one imaginary frequency below -50 cm^{-1} . These populations are not interchangeable. Cycle-limited candidates continued under a provisional policy retain their convergence status and are not relabelled converged TSs.

IRC calculations use the Carina potential in both directions, followed by the endpoint and common-reactant operations recorded for each campaign. Raw terminal frames and optimized endpoints are distinguished, and a terminal frame without a frequency calculation is not assumed to be a minimum. A connectivity change during endpoint optimization is attributed to that relaxation. Numerical settings and provisional handling are detailed in Supplementary Section S6.2; IC conformational supplementation is described in Supplementary Section S7.1.

4.6 Energy references and independent checks

Network barriers are local Carina-computed electronic-energy differences between a TS and the declared optimized common-reactant reference for the same complete molecular composition. Values from independent inputs do not automatically share a global energy zero. For a nonstationary candidate, the difference is a relative electronic height. Non-finite or negative heights are retained for diagnosis but excluded from navigation scoring. Free-energy corrections are reported only where calculated, with their Hessian source, temperature, standard state and solvent treatment specified.

The IC conformational comparison measures TS-energy changes relative to the initially retained lowest-energy TS for each reaction, using the same reactant reference. Reactive-core-constrained sampling follows CREST and the established model-reaction procedure; all restraints are released before final refinement.^{56,57} Independent IC DFT single points use B3LYP-D3/6-311++G(2d,2p) with IEF-PCM hexane on unchanged Carina geometries. They are energetic checks, not DFT-optimized TSs. Similarly, fixed-geometry proline SMD(DMSO) single points and Hessian-based qRRHO corrections provide approximate free energies rather than verified DFT stationary-point barriers.⁶⁸

4.7 Mapped identity and stereochemistry

Endpoint admission compares adjacency matrices from stored geometries under the original atom mapping, retaining every explicit hydrogen and fragment. Exactly one matching endpoint supplies a productive connection; two matching endpoints or no matching endpoint do not. Bond-electron matrices, formal charges, nonbonding electrons, isotopes and total charge/spin are recorded separately.

Consolidating complete molecular states for navigation also checks their electronic state. Equivalent-atom permutations require mapping witnesses; adjacency agreement alone does not establish Lewis-state equivalence or a stationary minimum.

Sampling inputs do not impose stereochemical constraints. Stereocentres and alkene configurations are analysed afterwards using stored three-dimensional coordinates and fixed mapped ligands, distinguishing local parity from CIP naming changes. Node consolidation never supplies an uncalculated stereoisomerization step. Full identity and stereochemistry procedures are specified in Supplementary Section S6.4.

4.8 Global kinetic-guided network growth

After each input calculation, eligible product sets enter a persistent global candidate pool. For a route π from the root r to node v , finite non-negative local electronic heights are sorted in descending order, $b_{(1)} \geq b_{(2)} \geq \dots$. The candidate priority is

$$S(v) = \min_{\pi \in \mathcal{P}_{L_{\max}}(r,v)} \sum_{i=1}^{|\pi|} 10^{-(i-1)} b_{(i)}(\pi). \quad (1)$$

The minimum is taken over admissible routes within the declared reaction-depth bound. Thus the largest height dominates, with decreasing contributions from the remaining heights. This score guides exploration rather than defining a rate constant or activation free energy. Newly discovered routes can lower the priority score of an existing candidate.

The branching factor n specifies the number of unexpanded input groups selected from the entire pool per batch: $n = 5$ for KHP, $n = 1$ for IC and $n = 3$ for the global-top-three catalytic explorations. It is distinct from the per-cluster representative count k . Unselected candidates remain available. All selected inputs must finish their scientific modules before their admitted outgoing edges are incorporated together and the pool is reranked. Failed modules do not count as completed expansions. Compatible archived calculations can be reused only when their parent is causally selected, preventing future edges from influencing earlier decisions.

Each input retains all co-products, spectators, charge/spin and atom mapping. Products from unrelated input sets are not recombined. Explicit water controls, targeted follow-ups and refinements are identified separately. IC combines causal replay with new calculations for competing reactant sets; its five-step chemical path is distinguished from the eleven expanded inputs needed to recover its terminal product. General selection rules are given in Supplementary Section S6.5; IC selection history and catalytic stopping criteria are described in Supplementary Sections S7.3 and S8.4.

Data and code availability

Static images and interactive HTML views of every reaction network reported here are provided in the Supplementary Information. The reaction database and the trained models will be released upon acceptance of the manuscript. ReactNexus will be available shortly for online use and network visualization through the MIRA platform at <https://mira.deepprinciple.com>.

Acknowledgement

We thank Heather J. Kulik for useful and insightful discussions. We acknowledge the NVIDIA DevTech and ALCHEMI teams for assistance with operator optimization to accelerate Equiformer-V3, and Wen Yan and colleagues at ByteDance for supporting the operation and maintenance of high-throughput GPU4PySCF calculations. We also thank the entire Deep Principle team for discussion and support.

References

1. Unsleber, J. P. & Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **71**, 121–142 (2020).
2. Wen, M. *et al.* Chemical reaction networks and opportunities for machine learning. *Nat. Comput. Sci.* **3**, 12–24 (2023).
3. Zhao, Q. & Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.* **1**, 479–490 (2021).
4. Maeda, S. *et al.* Implementation and performance of the artificial force induced reaction method in the GRRM17 program. *J. Comput. Chem.* **39**, 233–251 (2018).
5. Unsleber, J. P., Grimm, S. A. & Reiher, M. Chemoton 2.0: Autonomous Exploration of Chemical Reaction Networks. *J. Chem. Theory Comput.* **18**, 5393–5409 (2022).
6. Wang, L.-P. *et al.* Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **6**, 1044–1048 (2014).
7. Dewyer, A. L., Argüelles, A. J. & Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.* **8**, e1354 (2018).
8. Steiner, M. & Reiher, M. Autonomous Reaction Network Exploration in Homogeneous and Heterogeneous Catalysis. *Top. Catal.* **65**, 6–39 (2022).
9. Seumer, J. *et al.* Practical Computational Exploration of Chemical Reaction Space: Assessing the Chemical Stability of Active Pharmaceutical Ingredients Under Acidic Conditions. *Angew. Chem. Int. Ed.*, e4407622. doi:[10.1002/anie.4407622](https://doi.org/10.1002/anie.4407622) (2026).
10. Oliveira, B. L., Guo, Z. & Bernardes, G. J. L. Inverse electron demand Diels–Alder reactions in chemical biology. *Chem. Soc. Rev.* **46**, 4895–4950 (2017).
11. Abdelgaid, M. *et al.* Identification of Solid-Electrolyte Interphase Species by Joint Characterization of Li-Ion Battery Chemistry by Mass Spectrometry and Electrochemical Reaction Networks. *J. Am. Chem. Soc.* **148**, 17493–17507 (2026).
12. Zhao, Q. & Savoie, B. M. Deep reaction network exploration of glucose pyrolysis. *Proc. Natl Acad. Sci. USA* **120**, e2305884120 (2023).
13. Bensberg, M. & Reiher, M. Concentration-Flux-Steered Mechanism Exploration with an Organocatalysis Application. *Isr. J. Chem.* **63**, e202200123 (2023).
14. Zhao, Q., Garimella, S. S. & Savoie, B. M. Thermally Accessible Prebiotic Pathways for Forming Ribonucleic Acid and Protein Precursors from Aqueous Hydrogen Cyanide. *J. Am. Chem. Soc.* **145**, 6135–6143 (2023).
15. Broadbelt, L. J., Stark, S. M. & Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Ind. Eng. Chem. Res.* **33**, 790–799 (1994).
16. Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).
17. Liu, M. *et al.* Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **61**, 2686–2696 (2021).
18. Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **15**, 2847–2862 (2019).
19. Wang, L.-P., McGibbon, R. T., Pande, V. S. & Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **12**, 638–649 (2016).
20. Shang, C. & Liu, Z.-P. Stochastic Surface Walking Method for Structure Prediction and Pathway Searching. *J. Chem. Theory Comput.* **9**, 1838–1845 (2013).

21. Hayashi, H. *et al.* In silico reaction screening with difluorocarbene for N-difluoroalkylative dearomatization of pyridines. *Nat. Synth.* **1**, 804–814 (2022).
22. Sumiya, Y. & Maeda, S. Rate Constant Matrix Contraction Method for Systematic Analysis of Reaction Path Networks. *Chem. Lett.* **49**, 553–564 (2020).
23. Staub, R., Harabuchi, Y., Seraphim, C., Varnek, A. & Maeda, S. An Accurate and Efficient Reaction Path Search with Iteratively Trained Neural Network Potential: Answering the Passerini Mechanism Controversy. *J. Chem. Theory Comput.* **22**, 422–440 (2026).
24. Liu, Q.-Y., Chen, D., Shang, C. & Liu, Z.-P. An optimal Fe–C coordination ensemble for hydrocarbon chain growth: a full Fischer–Tropsch synthesis mechanism from machine learning. *Chem. Sci.* **14**, 9461–9475 (2023).
25. Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **34**, 1385–1392 (2013).
26. Zhao, Q. & Savoie, B. M. Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces. *Angew. Chem. Int. Ed.* **61**, e202210693 (2022).
27. Unke, O. T. *et al.* Machine Learning Force Fields. *Chem. Rev.* **121**, 10142–10186 (2021).
28. Wang, X., Mao, Y. & Wang, Z. Machine learning approaches for transition state prediction. *Chem Catalysis* **5**, 101458 (2025).
29. Anstine, D. M. *et al.* AIMNet2-rxn: A machine-learned potential for generalized reaction modeling on a millions-of-pathways scale. *Sci. Adv.* **12**, eaea1557 (2026).
30. Zhao, Q. *et al.* Harnessing Machine Learning to Enhance Transition State Search with Interatomic Potentials and Generative Models. *Adv. Sci.* **12**, e06240 (2025).
31. Yuan, E. C.-Y. *et al.* Analytical ab initio hessian from a deep learning potential for transition state optimization. *Nat. Commun.* **15**, 8865 (2024).
32. Cui, T., Han, Y., Jia, H., Duan, C. & Zhao, Q. A Large Scale Molecular Hessian Database for Optimizing Reactive Machine Learning Interatomic Potentials. *Sci. Data* **13**, 37 (2026).
33. Burger, A. *et al.* HIP: Hessian Interatomic Potentials without derivatives. *Nat. Comput. Sci.* doi:10.1038/s43588-026-01066-x (2026).
34. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *Advances in Neural Information Processing Systems 35* 11423–11436 (2022). doi:10.52202/068431-0830. <https://doi.org/10.52202/068431-0830>.
35. Anstine, D. M., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **16**, 10228–10244 (2025).
36. Wood, B. *et al.* UMA: A Family of Universal Models for Atoms. In *Advances in Neural Information Processing Systems 38* 143528–143564 (2025). doi:10.52202/085713-4310. <https://doi.org/10.52202/085713-4310>.
37. Zhang, S. *et al.* Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential. *Nat. Chem.* **16**, 727–734 (2024).
38. Ichino, T., Nabata, H., Matsumoto, K., Hasegawa, T. & Maeda, S. *Systematic and unbiased pathway exploration by artificial force application to a generic neural network potential* ChemRxiv preprint. 2024. doi:10.26434/chemrxiv-2024-9h20v. <https://doi.org/10.26434/chemrxiv-2024-9h20v>.
39. Duan, C., Du, Y., Jia, H. & Kulik, H. J. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nat. Comput. Sci.* **3**, 1045–1055 (2023).
40. Kim, S., Woo, J. & Kim, W. Y. Diffusion-based generative AI for exploring transition states from 2D molecular graphs. *Nat. Commun.* **15**, 341 (2024).

41. Duan, C. *et al.* Optimal transport for generating transition states in chemical reactions. *Nat. Mach. Intell.* **7**, 615–626 (2025).
42. Galustian, L., Mark, K., Karwounopoulos, J., Kovar, M. P.-P. & Heid, E. GoFlow: efficient transition state geometry prediction with flow matching and E(3)-equivariant neural networks. *Digit. Discov.* **4**, 3492–3501 (2025).
43. Sun, Z. *et al.* A Unified Generative Framework for Scalable Chemical Reaction Network Exploration arXiv preprint arXiv:2606.21002. 2026. arXiv: [2606.21002](#).
44. Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **9**, 779 (2022).
45. Zhao, Q. *et al.* Comprehensive exploration of graphically defined reaction spaces. *Sci. Data* **10**, 145 (2023).
46. Liao, Y.-L. *et al.* EquiformerV3: Scaling Efficient, Expressive, and General SE(3)-Equivariant Graph Attention Transformers 2026. arXiv: [2604.09130](#).
47. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
48. Lee, K. *et al.* A Comprehensive Dataset of Chemical Reactions Covering Second and Third Row Elements with Million-Scale Quantum Chemical Calculations. *Sci. Data* **13**, 983 (2026).
49. Tuo, P., Chen, J. & Li, J. Flow matching for reaction pathway generation. *Nat. Commun.* **17**, 8769 (2026).
50. Rhodes, B. *et al.* Orb-v3: atomistic simulation at scale arXiv preprint arXiv:2504.06231. 2025. arXiv: [2504.06231](#).
51. Grambow, C. A. *et al.* Unimolecular Reaction Pathways of a γ -Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **140**, 1035–1048 (2018).
52. Jalan, A. *et al.* New Pathways for Formation of Acids and Carbonyl Products in Low-Temperature Oxidation: The Korcek Decomposition of γ -Ketohydroperoxides. *J. Am. Chem. Soc.* **135**, 11100–11114 (2013).
53. Lee, C. W. *et al.* An Unexpected Ireland–Claisen Rearrangement Cascade During the Synthesis of the Tricyclic Core of Curcusone C: Mechanistic Elucidation by Trial-and-Error and Automatic Artificial Force-Induced Reaction (AFIR) Computations. *J. Am. Chem. Soc.* **141**, 6995–7004 (2019).
54. Young, T. A., Silcock, J. J., Sterling, A. J. & Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles—Application to Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **60**, 4266–4274 (2021).
55. Zhao, Q., Hsu, H.-H. & Savoie, B. M. Conformational Sampling for Transition State Searches on a Computational Budget. *J. Chem. Theory Comput.* **18**, 3006–3016 (2022).
56. Pracht, P., Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169–7192 (2020).
57. Zhao, Q., Singla, V., Hsu, H.-H. & Savoie, B. M. Graphically Defined Model Reactions Are Extensible, Accurate, and Systematically Improvable. *J. Chem. Theory Comput.* **21**, 6572–6583 (2025).
58. List, B., Lerner, R. A. & Barbas, C. F. Proline-Catalyzed Direct Asymmetric Aldol Reactions. *J. Am. Chem. Soc.* **122**, 2395–2396 (2000).
59. Bahmanyar, S. & Houk, K. N. Transition States of Amine-Catalyzed Aldol Reactions Involving Enamine Intermediates: Theoretical Studies of Mechanism, Reactivity, and Stereoselectivity. *J. Am. Chem. Soc.* **123**, 11273–11283 (2001).

60. Bahmanyar, S., Houk, K. N., Martin, H. J. & List, B. Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions. *J. Am. Chem. Soc.* **125**, 2475–2479 (2003).
61. Allemann, C., Um, J. M. & Houk, K. N. Computational investigations of the stereoselectivities of proline-related catalysts for aldol reactions. *J. Mol. Catal. A Chem.* **324**, 31–38 (2010).
62. Xia, Y. *et al.* An Unexpected Role of a Trace Amount of Water in Catalyzing Proton Transfer in Phosphine-Catalyzed (3 + 2) Cycloaddition of Allenates and Alkenes. *J. Am. Chem. Soc.* **129**, 3470–3471 (2007).
63. Liang, Y., Liu, S., Xia, Y., Li, Y. & Yu, Z.-X. Mechanism, Regioselectivity, and the Kinetics of Phosphine-Catalyzed [3+2] Cycloaddition Reactions of Allenates and Electron-Deficient Alkenes. *Chem. Eur. J.* **14**, 4361–4373 (2008).
64. Klucznik, T. *et al.* Computational prediction of complex cationic rearrangement outcomes. *Nature* **625**, 508–515 (2024).
65. De Witt, M. J., Dooling, D. J. & Broadbelt, L. J. Computer Generation of Reaction Mechanisms Using Quantitative Rate Information: Application to Long-Chain Hydrocarbon Pyrolysis. *Ind. Eng. Chem. Res.* **39**, 2228–2237 (2000).
66. Sakas, J., Kitson, E., Bell, N. G. A. & Uhrin, D. MS and NMR Analysis of Isotopically Labeled Chloramination Disinfection Byproducts: Hyperlinks and Chemical Reactions. *Anal. Chem.* **96**, 8263–8272 (2024).
67. Lavecchia, G., Berteina-Raboin, S. & Guillaumet, G. Synthesis and functionalisation of 1H-pyrazolo[3,4-b]pyridines involving copper and palladium-promoted coupling reactions. *Tetrahedron Lett.* **45**, 2389–2392 (2004).
68. Ribeiro, R. F., Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation. *J. Phys. Chem. B* **115**, 14556–14562 (2011).
69. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
70. Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **138**, 184102 (2013).
71. Zimmerman Group. *pyGSM: Thermal and photochemical reaction path optimization and discovery* <https://github.com/ZimmermanGroup/pyGSM> (2026).
72. Steinmetzer, J., Kupfer, S. & Gräfe, S. pysisyphus: Exploring potential energy surfaces in ground and excited states. *Int. J. Quantum Chem.* **121**, e26390 (2021).
73. Caldeweyher, E. *et al.* A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **150**, 154122 (2019).
74. Rappoport, D. & Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **133**, 134105 (2010).
75. Li, R., Sun, Q., Zhang, X. & Chan, G. K.-L. Introducing GPU Acceleration into the Python-Based Simulations of Chemistry Framework. *J. Phys. Chem. A* **129**, 1459–1468 (2025).
76. Jing, B. *et al.* EigenFold: Generative Protein Structure Prediction with Diffusion Models. *arXiv preprint arXiv:2304.02198* (2023).
77. Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. Harmonic Self-Conditioned Flow Matching for Joint Multi-Ligand Docking and Binding Site Design. In *Proceedings of the 41st International Conference on Machine Learning* **235** (2024).

- 78. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. *Flow Matching for Generative Modeling* 2022. doi:[10.48550/ARXIV.2210.02747](https://arxiv.org/abs/2210.02747). <https://arxiv.org/abs/2210.02747>.
- 79. Liu, X., Gong, C. & Liu, Q. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow* 2022. doi:[10.48550/ARXIV.2209.03003](https://arxiv.org/abs/2209.03003). <https://arxiv.org/abs/2209.03003>.
- 80. Frank, J. T. *et al.* Sampling 3D Molecular Conformers with Diffusion Transformers. In *Advances in Neural Information Processing Systems* **38**. arXiv:2506.15378 (2025).