PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 23914

Exploring the enantioselective synthesis mechanism of ammonium cations in solution using deep learning potential

Hongqiang Cui,[ab] Da Zheng,[ab] Huiying Chu,*[ac] Yan Li*[ac] and Guohui Li *[ac]

Received 6th September 2025,
Accepted 14th October 2025

DOI: 10.1039/d5cp03439f

rsc.li/pccp

Asymmetric synthesis is fundamental to modern organic chemistry. Predicting the stereoselectivity of catalytic reactions in solution remains a challenging problem, as it requires a comprehensive understanding of the underlying mechanistic, energetic, and kinetic factors. To address this challenge, we propose an active learning workflow that integrates iterative cycles of *ab initio* molecular dynamics (AIMD) and deep learning potential molecular dynamics (DLPMD) simulations. To demonstrate its utility, this workflow is applied to investigate the enantioselective synthesis of quaternary ammonium cations catalyzed by 1,1′-bi-2-naphthol scaffolds (BINOL), with a focus on simulations of large molecular systems over extended timescales. The results of the simulations successfully reproduce the experimentally observed chirality of the major product molecules, as confirmed by ^{1}HNMR spectroscopy. Furthermore, these simulations provide detailed insight into the reaction pathways and reveal that the chirality of the primary product is thermodynamically controlled under experimental conditions. Consequently, this workflow offers a promising strategy for exploring complex reaction mechanisms and enhancing the predictive accuracy of asymmetric synthesis in complex solutions using deep learning techniques.

Introduction

Asymmetric synthesis, which enables the selective construction of chiral molecules, is fundamental to modern organic chemistry. Notably, quaternary ammonium cations are widely applied as catalysts and reagents in organic synthesis,[1,2] medicinal chemistry,[3] and polymer materials.[4] Despite their broad applications, achieving precise control over the stereogenicity of the nitrogen atom remains a significant challenge in asymmetric synthesis. Previous work has demonstrated that stereogenic nitrogen centers can be established *via* diastereoselective transfer of chirality from adjacent carbon stereocenters.[5] However, the strong influence of neighboring stereocenters makes direct access to nitrogen only stereogenic compounds challenging. The inherent conformational instability of stereogenic nitrogen centers hinders their enantioselective synthesis. In contrast to carbon stereocenters, which are conformationally and configurationally rigid (Scheme S1a), nitrogen stereocenters in tertiary amines are prone to rapid inversion due to the facile flipping of the nitrogen lone pair. This configurational lability often leads to racemization, which is frequently underestimated in stereoselectivity analyses. Notably, obtaining enantioenriched ammonium cations *via* kinetic resolution[6] or spontaneous resolution[7–10] demonstrates that slowing or preventing nitrogen inversion enables the selective formation of stereogenic nitrogen centers. Consequently, the formation of quaternary ammonium cations prevents this inversion, thereby stabilizing the nitrogen configuration.[11] Building on these principles, previous work[12] demonstrated the high enantioselective synthesis of ammonium stereocenters in a single pot catalyzed by 1,1′-bi-2-naphthol scaffolds (BINOL), using either R-BINOL or S-BINOL (Scheme S1b). Despite these advances, the underlying mechanisms governing the high enantioselectivity in the synthesis of quaternary ammonium cations remain unclear.

In recent decades, transition state (TS) searches[13–19] have been widely employed to explore reaction mechanisms in solution. However, traditional density functional theory (DFT) calculations have inherent limitations when applied to complex reaction systems. Firstly, identifying transition states in large systems comprising thousands or more atoms and in complex reactions relies heavily on expert intuition. Secondly, incorporating solvent effects into DFT calculations of chemical reactions presents significant challenges. Although solvent effects

---

[a] *State Key Laboratory of Chemical Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China. E-mail: chuhy2009@dicp.ac.cn, ghli@dicp.ac.cn*

[b] *University of Chinese Academy of Sciences, Beijing 100049, China*

[c] *Interdisciplinary Research Center for Biology and Chemistry, Liaoning Normal University, Dalian, Liaoning 116029, China. E-mail: liyan1982@lnnu.edu.cn*

on both stable and transition states are commonly treated with continuum solvent models (CSMs),[20,21] such approaches fail to capture specific interactions with explicit solvent molecules. In principle, introducing explicit solvent molecules could address these limitations. However, defining the appropriate number and positions of explicit solvent molecules remains difficult, and both the conformational space and computational cost are significantly increased upon the inclusion of explicit solvent molecules.[22] Thirdly, this approach fails to fully capture the dynamic behavior of the system. While *ab initio* molecular dynamics (AIMD) simulations offer a potential solution to these challenges, their high computational cost restricts the feasible system sizes and simulation timescales.

Recently, researchers have increasingly turned to deep learning (DL) methods.[23–38] DL methods, especially artificial neural networks (NNs), provide the possibility to construct potential energy surfaces (PESs) with the accuracy of the DFT method but with an efficiency comparable to that of force fields. DL methods have been employed to construct PESs in a data-driven manner. In this approach, the PES is derived from a carefully selected training dataset using appropriate descriptors.[24–32,39] Deep learning potentials (DLPs) have shown success in modeling systems such as water,[40] small organic molecules,[41,42] and metal materials.[43] However, the reliability of the DLPs is highly dependent on the completeness of the training datasets.[44,45] Although AIMD simulations are commonly used to generate training datasets,[41,42] they may fail to capture some essential molecular interactions due to the random generation of initial conformations. While notable active learning approaches have been reported in recent years, some studies[46,47] are limited to low-dimensional systems, and extending these methods to complex reaction systems in solution remains a challenge. Furthermore, in previous active learning workflows,[23,35,36,48] for configurations in MD trajectories, the deviation of atomic forces predicted by several DLPs is calculated. If the deviation exceeds a predefined threshold, the corresponding structures are labeled using the DFT method and subsequently added to the training datasets. Nevertheless, some critical but unstable regions remain difficult to sample using conventional MD simulations. Firstly, unstable transition states are "rare events" in MD simulations. Secondly, for large systems, the total charge for cluster structures obtained from MD simulations is not considered when labeling these cluster structures using the DFT method. Consequently, DFT calculations for these structures are infeasible, preventing their inclusion in the training dataset.

In this paper, to address these issues, an integrated iterative workflow (Scheme 1a) combining AIMD and deep learning potential molecular dynamics (DLPMD) simulations is proposed to explore the enantioselective synthesis of quaternary ammonium cations catalyzed by BINOL, with simulations conducted on long timescales and large systems. Specifically, the training datasets are derived from this integrated iterative workflow. Firstly, the DLP model is pre-trained using structures sampled along reaction pathways observed in AIMD and DLPMD simulations, which enables it to accurately describe the reaction energy barriers. Secondly, this pre-trained DLP model is trained against AIMD simulation trajectories to improve the description of stable regions in complex reaction systems. The total charge of the initial conformation is calculated prior to performing AIMD simulations for the eight subsystems (Scheme 1b). Notably, the experimental environments are accurately reproduced in DLPMD simulations. The results of simulations successfully reproduce the experimentally observed chirality of the major product, which is consistent with the ^{1}HNMR spectroscopy results.[12] This study reveals that the chirality of quaternary ammonium cations in experiments is under thermodynamic control. Furthermore, the proposed workflow provides detailed insight into reaction pathways within complex reaction systems comprising thousands or more atoms and offers a promising strategy for accurate modeling of asymmetric synthesis in solution.

Methods

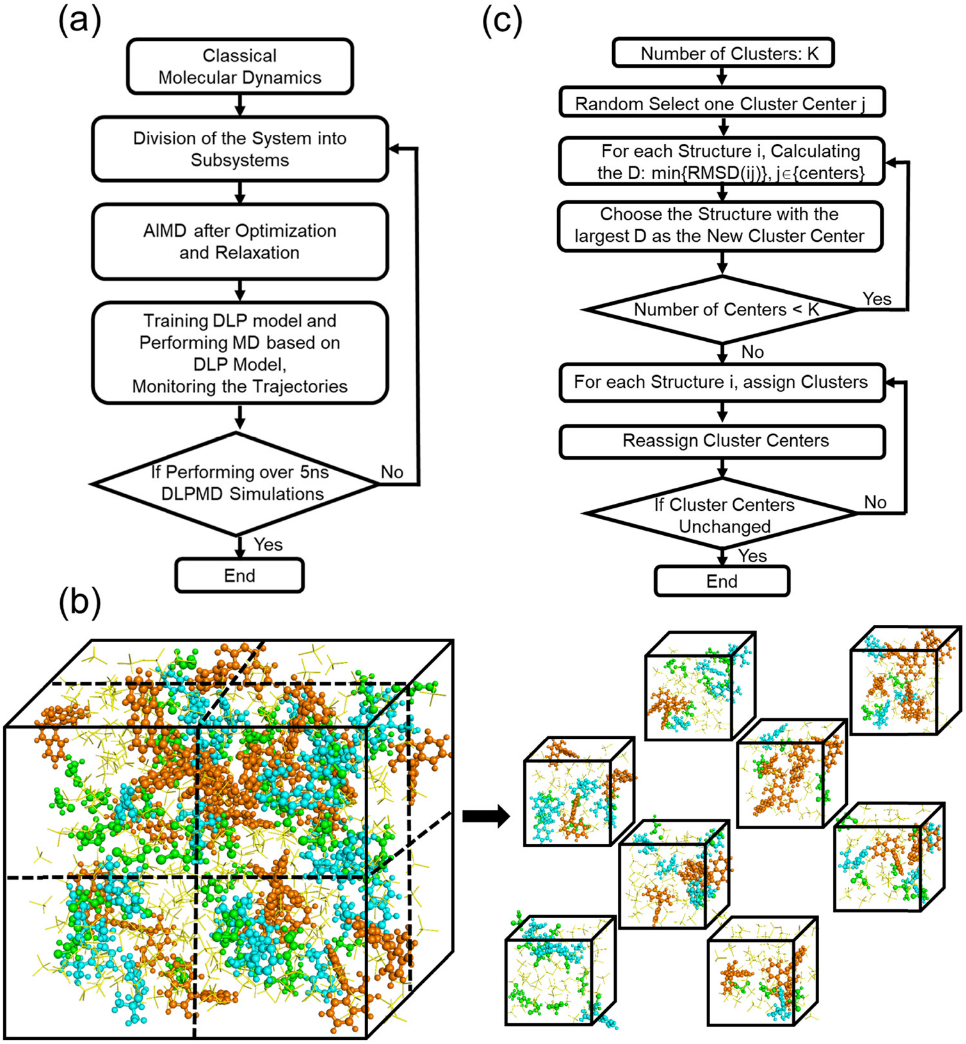
Active learning workflow

The complete active learning flowchart is illustrated in Scheme 1a. The workflow and convergence assessment comprise the following steps: (1) system construction and classical MD simulations; (2) training dataset generation *via* AIMD simulations on divided subsystems; (3) training of the DLP model and DLPMD simulations; and (4) DLP model validation and iterative refinement. Additional adjustments addressing charged species in step (2) are outlined in (a) recalculation of subsystem charges after partitioning, and (b) box adjustment under periodic boundary conditions. The convergence criteria in step (4) encompass (i) the absence of abnormal bond formation or cleavage during DLPMD simulations, and (ii) accurate prediction of energies and forces based on the DLP model. The training workflow proceeds as follows:

(1) System construction and classical MD simulations. The initial full system is constructed based on the experimental ratio of substrate, solvent, and catalyst molecules to reproduce the complex reactive environment of the solution phase under experimental conditions. Equilibrium classical MD simulations with periodic boundary conditions (PBCs) are performed for 100 ns in a cubic box with a side length of 40.0 Å. To improve the computational efficiency of subsequent AIMD simulations while retaining essential intermolecular interactions, the final snapshot of the above MD trajectory is divided into eight smaller cubic systems (20.0 Å × 20.0 Å × 20.0 Å), which are individually subjected to AIMD simulations (Scheme 1b).

(2) Training dataset generation *via* AIMD simulations on divided subsystems. Following geometry optimization and 5 ps NPT MD simulations using the GFN1-xTB method,[49] 2 ps NVT AIMD simulations are performed on each of the eight subsystems in step (1) or step (4) to generate the training dataset. There are no reaction events in MD simulations performed with the GFN1-xTB method or AIMD simulations. Consequently, these simulations were not employed for stereoselectivity or property analyses in this work. Additional simulation details can be found in the section titled "*Ab initio* molecular dynamics simulations".

Notably, the formation and cleavage of chemical bonds during DLPMD simulations often lead to the generation of



Scheme 1 (a) Flowchart of the active learning workflow. (b) Schematic representation of the simulation system. Note: the system contains 24 NR_3 (cyan), 24 BINOL (orange), 48 allyl bromide (green) and 367 trichloromethane (yellow) molecules. (c) Flowchart of the clustering analysis using the K-means method.

cations and anions in the eight subsystems. According to reaction pathways calculated by the DFT method, the charges of these ions must be explicitly specified in the parameter files. The preparation and optimization of these small systems require two critical adjustments:

(a) Recalculation of subsystem charges after partitioning. The net charge of each subsystem must be recalculated following the partitioning of the whole system, based on the sum of charges of ions in each subsystem.

(b) Box adjustment under periodic boundary conditions. Molecular integrity must be preserved during system partitioning, which often leads to irregular, non-cubic subsystem geometries. To prevent artificial molecular contacts under PBCs, the box size needs to be expanded.

(3) Training of the DLP model and DLPMD simulations. A DLP model is trained, and 5 ns DLPMD simulations are performed on the initial full system to explore the reactive behaviors in systems.

(4) DLP model validation and iterative refinement. After DLPMD simulations, the trajectories in step (3) are examined to assess whether the model has converged. The active learning workflow is terminated till the DLP model converges. Otherwise, the trajectory is again divided into eight subsystems (Scheme 1b), and steps (2), (3) and (4) are repeated. In previous studies,[50] the convergence of the DLP model is assessed by monitoring the energies and forces of a 1 ns trajectory every 2.5–25 ps using the DFT method. In the present workflow, the first 5 ns of each trajectory is analyzed, instead of the 1 ns used in previous studies. The DLP model is considered as converged if it satisfies all the following criteria:

(i) The absence of abnormal bond formation or cleavage during DLPMD simulations. Bond formation and cleavage events observed in DLPMD trajectories are examined using DFT calculations performed at the same level of theory as the AIMD simulations. If the DLP model fails to accurately reproduce either stable and unstable states along the reaction pathway, such bond-forming or bond-breaking events are deemed abnormal. Trajectories exhibiting such abnormal events are classified as abnormal, and only the trajectory segments prior to these events are retained as normal. Accurate modeling of reaction pathways is critical for the accuracy of the DLP model. However, DFT calculations frequently fail to converge for abnormal configurations extracted from these trajectories, thereby preventing their inclusion in the training dataset to correct the abnormal bond formation or cleavage observed in DLPMD simulations. In previous work,[51] structures derived from transition states in vacuum are employed to provide the DLP model with information about intrinsic molecular reactivity. Therefore, in this study, structures corresponding to key structures along reaction pathways are labeled using DFT methods in vacuum and used to pre-train the DLP model during the first stage of training, ensuring that the DLP model could capture the reaction energy barrier. The absence of abnormal bond formation or cleavage within the 5 ns trajectory is a prerequisite for ensuring the reliability of the simulation results. The Results and discussion section (the importance of reaction pathways for the accuracy of DLP model) highlights the necessity of pre-training the DLP model using structures in vacuum along the reaction pathways during the first stage of training.

(ii) Accurate prediction of energies and forces based on the DLP model. In previous work,[51] the last frame of a trajectory has been evaluated to determine whether the trajectory should be included into the existing training dataset. Consequently, the last snapshot of the above normal trajectory that passed the first criterion is divided into eight smaller cubic subsystems (Scheme 1b), which are further evaluated using the DFT method. Following geometry optimization and 5 ps NPT MD simulations using the GFN1-xTB method,[49] 2 ps NVT AIMD simulations are performed on each of the eight subsystems to generate a testing dataset. The uncertainty thresholds for energy and force deviations between DFT and DLP predictions are set at 0.05 eV per atom and 0.1 eV Å^{-1}, respectively. The uncertainty threshold values are more stringent than 0.14 eV per atom and 0.12 eV Å^{-1} in previous studies.[23] If the MAEs in energy or force for the test dataset exceed these thresholds, the corresponding AIMD trajectories must be included in the training set. Only when both MAEs fall below their respective thresholds can the trajectory be considered validated. If both criteria are met, the DLP model is regarded as sufficiently accurate and converged.

In this workflow, the accuracy of the DLP model is directly validated against DFT calculations. Since the small systems are derived from the experimental system, the resulting AIMD simulations faithfully reproduce the molecular interactions observed in solution. Capturing these experimental interactions is crucial for improving the fidelity and generalizability of the DLP model. In this study, in the first stage of training, 1202 configurations in vacuum are used to pre-optimize the DLP model and reproduce the energy barriers along reaction pathways. In the second stage, the pre-trained DLP model is refined using both 173 857 configurations obtained from AIMD trajectories and 1202 configurations along reaction pathways to improve its accuracy in describing the stable regions of the reaction landscape. Additional simulation protocols and training details are provided in the Methods section (molecular dynamics simulations with classical force field, *ab initio* molecular dynamics simulations, molecular dynamics simulations based on the deep potential model and deep potential descriptor and training). Notably, only seven cycles are sufficient to reach convergence, demonstrating the efficiency of this workflow.

In total, the workflow consumed approximately 1 307 000 CPU core hours on AMD EPYC 7452 processors and 2400 GPU hours for simulations based on the converged DLP model.

Structural clustering analysis

To elucidate the underlying mechanism of stereoselectivity in this study, the binding modes between BINOL and the product molecule, as well as between BINOL and the substrate molecule, are systematically investigated. Due to the substantial number of binding structures sampled from MD simulations, representative binding structures are selected through structural clustering analysis. The distance cutoff between BINOL and the product or substrate molecule is defined as 8.0 Å, measured between the nitrogen atom and all atoms of BINOL. If more than 12 atoms of BINOL fall within this cutoff, the BINOL is considered to interact with the product or substrate molecule.

To distinguish different binding structures in three-dimensional space, a coordinate system (Schemes 2a and d) is introduced, with the midpoint between atoms 1 and 1′ of BINOL defined as the origin. The three-dimensional space is divided into eight regions, labeled as (1,1,1), (1,1,0), (1,0,1), (0,1,1), (1,0,0), (0,1,0), (0,0,1), and (0,0,0), based on the sign of the nitrogen atom's Cartesian coordinates. For instance, if the coordinates of the nitrogen atom are (x, y, z), and $x > 0, y > 0$, and $z < 0$, the structure is assigned to region (1,1,0). Within each spatial region, the binding structures are further analyzed by evaluating the energy variations along the dihedral angle ω (C3–N–C10′–C10) and χ (C7–N–C6–C10). Each spatial region is subdivided into nine segments (Scheme S2). Representative structures within each segment are selected using the K-means clustering method. The detailed workflow of the K-means method[52,53] is illustrated in Scheme 1c. This workflow proceeds as follows:

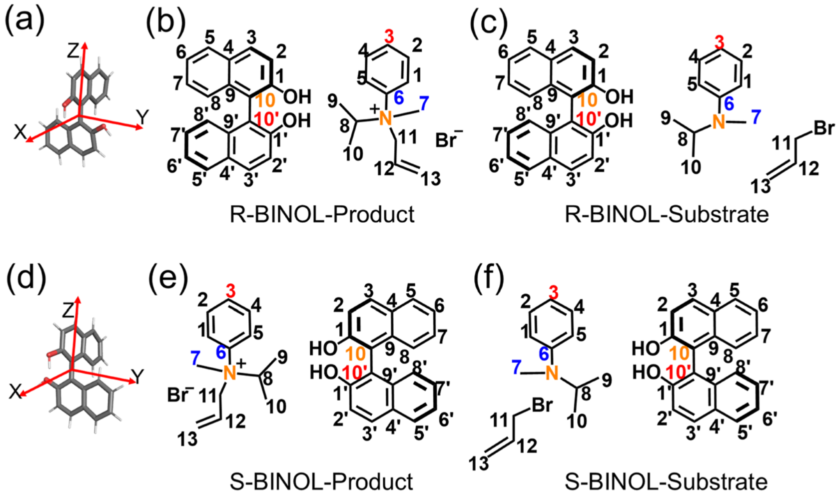
(1) The number of clusters, K, is set to 20 in this study.

(2) A random structure is selected as the initial cluster center, j.

(3) The root mean square deviation (RMSD) between each structure i and existing cluster centers, j, is calculated using eqn (1):

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2\right)} \quad (1)$$

where (x_i, y_i, z_i) and (x_j, y_j, z_j) are the Cartesian coordinates of atoms in structure i and j, respectively, and N is the number of atoms. For each structure i, RMSD_{min} is defined as the minimum RMSD to all known cluster centers.



Scheme 2 Coordinate systems are defined for (a) R-BINOL and (d) S-BINOL, respectively. Binding modes of BINOL (R-BINOL in (b), S-BINOL in (e)) with the product molecule, and binding modes of BINOL (R-BINOL in (c), S-BINOL in (f)) with the substrate molecule. The torsion angles ω (C3–N–C10′–C10) and χ (C7–N–C6–C10) are indicated in red/orange and blue/orange, respectively. Note: COM1 refers to the center of mass of atoms O1, C5′ and C8′ (or O1′, C5 and C8) in BINOL; COM2 refers to the center of mass of atoms C9 and C10 in the NR_3 substrate molecule. Equivalent atoms in BINOL are denoted by primes (*e.g.*, 1 and 1′).

(4) The structure with the largest $RMSD_{min}$ is selected as a new cluster center.

(5) Steps (3) and (4) are repeated until 20 cluster centers are identified.

(6) Each structure is assigned to the cluster with the closest center, as determined by the minimum RMSD.

(7) Within each cluster, the sum of RMSD values between structure i and all others in this cluster is calculated as Si. The structure with the minimum Si value is updated as the new cluster center. Steps (6) and (7) are repeated until the cluster centers converge.

Notably, in the Results and discussion section (binding modes of BINOL with product molecules observed in crystal structures and binding modes of BINOL with product and substrate molecules in solution), NR_4^+ (R) and NR_4^+ (S) species that form a hydrogen bond between the hydroxyl group in BINOL and the bromine atom are referred to as R_Br and S_Br, respectively. In contrast, species lacking this hydrogen bond are denoted R and S. For substrate molecules, those that lead to NR_4^+(R) and NR_4^+(S) formation are labeled Sub_R and Sub_S, respectively.

In binding modes of BINOL with product molecules (R-BINOL-R_Br (or S_Br), R-BINOL-R (or S), S-BINOL-R_Br (or S_Br) and S-BINOL-R (or S)), the coordinates of carbon atoms (1–10, 1′–10′) and oxygen atoms (1, 1′) in BINOL, carbon atoms (3, 6, 7, 8, 11) and nitrogen atom in NR_4^+ (Schemes 2b and e) are used for RMSD calculations. RMSD calculations for binding modes of BINOL with substrate molecules (R-BINOL-Sub_R (or Sub_S) and S-BINOL-Sub_R (or Sub_S)) are performed using the same BINOL atoms, carbon atoms (3, 6, 7, 8) and the nitrogen atom in NR_3, together with either carbon atom C11 or C13 from the allyl bromide molecule (Schemes 2c and f). Between C11 and C13, the carbon atom with the minimum distance to the nitrogen atom is selected for the RMSD calculation. Prior to RMSD calculations in the above cases, all binding structures are aligned based on the heavy atoms of BINOL to eliminate translational and rotational differences.

Equivalent atoms (Scheme 2) can lead to structurally identical complexes being misclassified as distinct during cluster analysis. To address this, the twenty lowest-energy binding structures are selected for further cluster analysis. Specifically, K-means clustering[52,53] is applied using root mean square deviation (RMSD) values that treat equivalent atoms as indistinguishable. An RMSD cutoff of 1.5 Å is applied to define structural similarity within each cluster.

For the binding modes of BINOL with both product and substrate molecules (Scheme 2), only the carbon atoms (1–10, 1′–10′) and the oxygen atoms (1 and 1′) in BINOL are used for RMSD calculations. The symmetry equivalent atoms in BINOL are treated as identical during RMSD evaluation. The maximum RMSD value allowed between structures within the same cluster is set to 1.5 Å. The MDTraj[54] is employed for clustering analysis. Prior to RMSD calculations accounting for equivalent atoms, all binding structures are aligned by superimposing the carbon atoms (3, 6, 7, and 8) and the nitrogen atom of NR_4^+ or NR_3 to eliminate translational and rotational differences.

Molecular dynamics simulations with classical force field

A total of 24 NR_3 molecules, 24 1,1′-bi-2-naphthol scaffolds (R-BINOL or S-BINOL), and 48 allyl bromide are solvated in a 40.0 Å × 40.0 Å × 40.0 Å cubic simulation box under PBCs. The simulation box contains 367 chloroform molecules. All MD simulations are performed at 323 K using a Nose–Hoover thermostat with a damping constant of 10 fs. Furthermore, 100 ns simulations are performed with a time step size of 1.0 fs in the NVT ensemble. All simulations are executed using the general AMBER force field (GAFF),[55] as implemented in the Gromacs-2019 software.[56] AM1-BCC methods[57] are used to assign partial charges for coulombic interactions in GAFF. Long range electrostatic interactions and van der Waals (vdW) interactions are both computed using the Particle Mesh Ewald (PME) method with an interpolation order of 4 and grid spacing of 0.16 nm. The cutoff distance of 1.2 nm is used for both real-space electrostatic and vdW interactions.

Ab initio molecular dynamics simulations

All AIMD simulations are performed using cp2k-7.2 software.[58] The Multiwfn software facilitates the generation of input files for AIMD simulations.[59,60] The initial full system is partitioned into eight smaller cubic subsystems (20.0 Å × 20.0 Å × 20.0 Å). Each subsystem is optimized using xtb-6.5.0[61] at the GFN1-xTB level of theory.[49] The box length is then increased from 20.0 Å to 22.0 Å to eliminate abnormal intermolecular contacts at the boundaries. To ensure reliability, 5 ps simulations are performed with a time step size of 0.5 fs in the NPT ensemble at the GFN1-xTB level of theory, with PBCs. An isotropic pressure of 1 bar is applied. Equilibration is typically achieved within 5 ps. Subsequent to optimization and NPT simulations, 2 ps simulations are performed with a time step size of 0.5 fs in the NVT ensemble with PBCs. All MD simulations are conducted at 323 K using canonical sampling through velocity rescaling (CSVR) thermostat with a damping constant of 10 fs. Energies and forces are computed using the DZVP-MOLOPT-SR-GTH basis set[62] along with GTH-Perdew–Burke–Ernzerhof (PBE) pseudopotentials.[63–65] The exchange–correlation energy was described with the PBE functional.[66,67]

The 5 ps NPT simulation of 400–500 atoms performed on 64 CPU cores of an AMD EPYC 7452 processor requires approximately 3.5 hours of wall time. The subsequent 2 ps NVT AIMD simulation of 400–500 atoms performed on 64 CPU cores of an AMD EPYC 7452 processor requires approximately 17–24 hours of wall time.

Molecular dynamics simulations based on the deep potential model

Molecular dynamics simulations on the initial full system. Metadynamics (MTD) simulations are performed using LAMMPS[68] and PLUMED.[69] Collective variables (CVs) based on coordination numbers, as defined in eqn (2), are employed to explore the reaction mechanism:

$$\mathrm{CN}(i,j) = \frac{1-\left(\frac{d_{ij}}{d_0}\right)^p}{1-\left(\frac{d_{ij}}{d_0}\right)^q} \tag{2}$$

where d_{ij} is the distance between atoms i and j, d_0 is the cutoff distance, and p and q are high-power integers used to distinguish the coordinated and non-coordinated states. The values of d_0 are taken as 2.0 Å, 3.0 Å and 5.0 Å, respectively, for N–C bond, COM1–Br interactions and COM1–COM2 interactions, and COM1 and COM2 are defined as shown in Scheme 2. The values $p = 6$ and $q = 12$ are consistently used for all CVs. CV1 is the coordination number between the nitrogen atom in NR_3 and carbon atoms (11 and 13) in allyl bromide, describing the breaking and formation of the N–C bond. CV2 is the coordination number between the COM1 and bromine atom, describing the interactions between the BINOL and allyl bromide molecules. CV3 is the coordination number between the COM1 and COM2, describing the interactions between the BINOL and NR_3 molecules. Well-tempered metadynamics (WTMTD)[70] simulations are performed using the DLP model. The system is simulated for 5 ns with a 1 fs time step in the NVT ensemble. All MD simulations are conducted at 323 K using a Nose–Hoover thermostat with a damping constant of 10 fs. Gaussian hills are added along CV1, CV2, and CV3 to enhance sampling. For CV1 and CV2, Gaussian hills (height w = 1.0 kcal mol^{-1} and width δ_s = 0.2) are added every 1 ps of simulation. The bias factor is set as 15.0. For CV3, Gaussian hills (height w = 1.0 kcal mol^{-1} and width δ_s = 0.2) are added every 0.5 ps of simulation. The bias factor is set as 5.0.

The 5 ns NVT DLPMD simulation of 3755 atoms is conducted using 4 A100 GPUs combined with 4 CPU cores, with a total wall time of approximately 60 hours.

Molecular dynamics simulations on the test system. A total of 1 NR_3 molecule and 2 allyl bromide molecules are solvated in a 15.0 Å × 15.0 Å × 15.0 Å cubic simulation box under PBCs. The simulation box contains 27 chloroform molecules. MTD simulations are performed using LAMMPS[68] and PLUMED.[69] CVs based on coordination numbers as defined in eqn (2) are employed to explore the reaction mechanism. The values of d_0 are set to 2.0 Å and 2.5 Å, respectively, for the N–C bond and C–Br bond. The values $p = 6$ and $q = 12$ are consistently used for all CVs. As shown in Scheme 2, CV1 is the coordination number between the nitrogen atom in NR_3 and carbon atoms (C11 and C13) in allyl bromide, describing the breaking and formation of the N–C bond. CV2 is the coordination number between the carbon atoms (C11 and C13) and bromine atom in allyl bromide, describing the breaking and formation of the C–Br bond. WTMTD[70] simulations are performed using both DLP models trained with and without structures along reaction pathways. The system is simulated for 20 ns with a 1 fs time step in the NVT ensemble. All MD simulations are conducted at 323 K using a Nose–Hoover thermostat with a damping constant of 10 fs. To enhance sampling, Gaussian hills are added along the defined CVs during simulations. For CV1 and CV2, Gaussian hills (height w = 1.2 kcal mol^{-1} and width δ_s = 0.2) are added every 0.5 ps of simulation. The bias factor is set as 8.0. The PESs of this system are presented in the Results and Discussion section (the importance of reaction pathways for the accuracy of DLP model).

The 20 ns NVT DLPMD simulation of 179 atoms is conducted using 4 A100 GPUs combined with 4 CPU cores, with a total wall time of approximately 35 hours.

Deep potential descriptor and training. DLP models are trained using the DeePMD-Kit package (version 2.1.5)[71] with the se_atten model.[72] The cutoff is set to 7.0 Å with a smoothing function starting at 2.0 Å. The embedding network consists of three hidden layers with sizes (25, 50, 100) and the fitting network also comprises three hidden layers with sizes (240, 240, 240). The training process minimizes the following loss

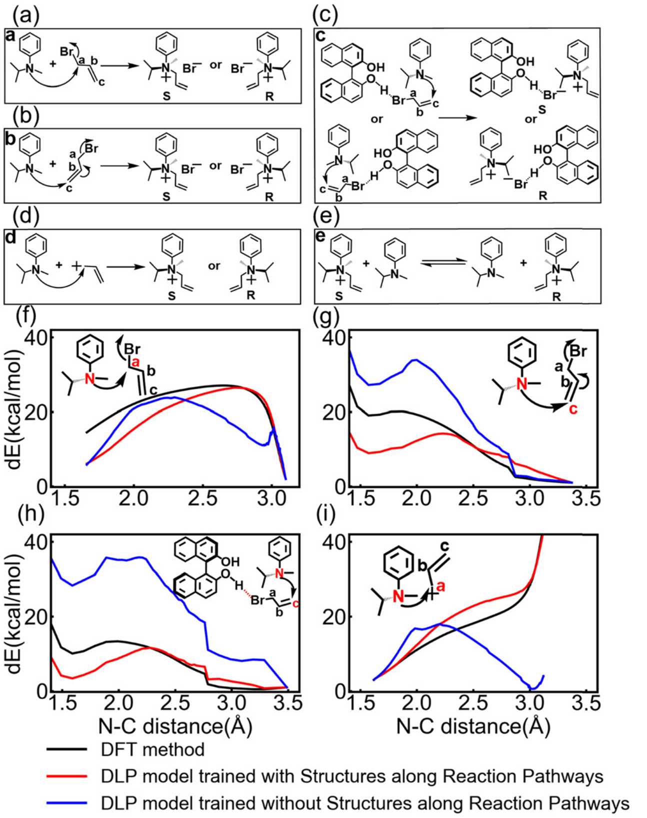


Fig. 1 Five types of reaction pathways: (a–e). (a) Nucleophilic attack by nitrogen on carbon a, leading to bromide departure. (b) Attack on carbon c with bromide departure. (c) Similar to (b), but assisted by hydrogen bonding between bromine and BINOL's hydroxyl. (d) Attack on an allylic carbocation. (e) Intermolecular transfer of the allylic carbocation between two NR_3 molecules. Energetics along the N–C distances are calculated using the DLP model for NR_4^+(S) structures optimized at the PBE/DZVP-MOLOPT-SR-GTH level of theory. (f–i) Reaction coordinate diagrams corresponding to (a–d), respectively. Note: MD simulations reveal that pathway e proceeds *via* a stepwise, non-concerted pathway involving initial N–C bond cleavage and subsequent bond formation, both shown in (d), through a carbocation intermediate. Thus, the energy profile in (i) accurately reflects the associated barrier. In Fig. 1f–i, the horizontal axis represents the distance between the nitrogen and carbon atoms, both labeled in red.

function, as shown in eqn (3)

$$L(p_\varepsilon, p_f, p_v) = \frac{p_\varepsilon}{N}\Delta\varepsilon^2 + \frac{p_f}{3N}|\Delta F_i|^2 + \frac{p_v}{9N}\|\Delta v\|^2 \tag{3}$$

where N is the number of atoms. $\Delta\varepsilon$, $|\Delta F_i|$, and $\|\Delta v\|$ represent the deviations of energy, atomic force, and virial tensors between the DLP model and DFT calculations, respectively.

For the DLP model trained with structures along reaction pathways, the training is conducted in two stages. In the first stage, the prefactors of p_ε, p_f and p_v change gradually from 0.05 to 5, 1000 to 1, and 0.01 to 1, respectively. In previous work,[51] the structures generated from the transition state of the system in vacuum are used to provide information about intrinsic reactivity for training the DLPs to explore chemical processes in explicit solvents. To reproduce the energy barriers along reaction pathways, the DLP model from the final iteration is fine-tuned for 1 × 10^5 steps using a training dataset that includes only structures corresponding to four representative types of reaction pathways. In the second stage, it is observed that assigning a large energy prefactor can lead to instability in MD simulations within tens of picoseconds. Therefore, the pre-tuned DLP model is further trained for 5 × 10^5 steps using a dataset that includes both AIMD trajectories and reaction pathway structures, with fixed prefactors p_ε, p_f, and p_v set to 0.05, 1000, and 0.01, respectively. As shown in Fig. 1f–i and Fig. S1a–d and Fig. 2, the pre-tuned DLP model is essential for maintaining the reaction barrier on the energy landscape. Furthermore, the stable states predicted by the DLP model are further optimized in the second stage. The accuracy of the final DLP model is summarized in the Results and discussion section (the accuracy of the DLP model).

For the DLP model trained without structures along reaction pathways, the training is conducted in a single stage using a dataset that includes only AIMD trajectories, with training parameters identical to those used in the second stage of the DLP model trained with structures along reaction pathways.

The first stage of the training process for the DLP model trained with structures along reaction pathways was performed using an A100 GPU and 1 CPU core, requiring approximately 1 hour of wall time. The second stage was also performed on the same hardware, requiring approximately 16 hours. Similarly, the single-stage training process of the DLP model trained without structures along reaction pathways required approximately 16 hours on the same hardware.

Results and discussion

The importance of reaction pathways for the accuracy of the DLP model

Although the DLP model trained solely on AIMD data can detect reaction events, it fails to accurately describe the resulting species, for example, predicting stable carbocations that contradict experimental observations.[73–79] Sampling based only on AIMD provides insufficient coverage of the transition states, leaving the model lacking essential information about the intrinsic molecular reactivity. Therefore, if the DLP model cannot correctly describe reaction pathways discovered during DLPMD simulations, the corresponding structures must be incorporated for pretraining to provide essential reactivity information and enable accurate modeling of reaction dynamics in solution.[51] To address these limitations, reaction events are monitored and validated using the DFT method during DLPMD simulations. As shown in Fig. 1 and 2, inclusion of these pathway structures is critical for accurately describing reaction barriers and the overall PES, without requiring laborious TS sampling in solution. Specifically, five distinct categories of N–C bond formation pathways (Fig. 1a–e) are identified through the iterative workflow integrating AIMD and DLPMD simulations. Notably, MD simulations reveal that pathway e (Fig. 1e) proceeds *via* a stepwise, non-concerted mechanism involving initial N–C bond cleavage and subsequent bond

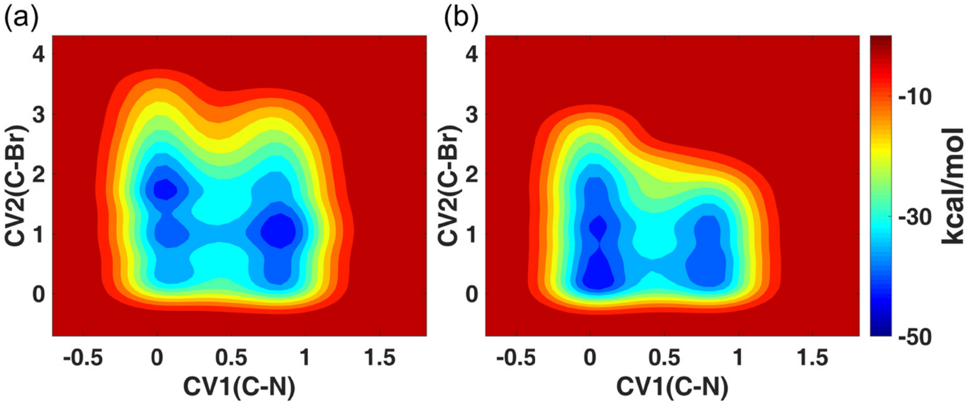


Fig. 2 Free energy profile for the synthesis of NR_4^+ based on the DLP model trained (a) with structures along reaction pathways and (b) without structures along reaction pathways. Note: the convergence of the simulations was checked, as shown in Fig. S3a and b.

formation, both illustrated in Fig. 1d, through a carbocation intermediate. Thus, only structures along reaction pathways a–d (Fig. 1a–d) are included in the training datasets. The relative energies of structures along reaction pathways a–d (Fig. 1f–i and Fig. S1a–d) reveal that these structures are of vital importance to the accuracy of the DLP model. Although structures along reaction pathway e (Fig. 1e) are not included in training datasets, the converged DLP model still reproduces the energy barrier for this pathway (Fig. S2a). To evaluate this effect, two types of DLP models are constructed for comparison. As shown in Methods section (deep potential descriptor and training), the first DLP model is trained in two stages and incorporates structures along reaction pathways, while the second DLP model is trained in a single stage using only AIMD trajectories, with training parameters consistent with the second stage of the first model. In the second stage, the DLP model is primarily trained based on atomic forces using fixed prefactors, which increases the mean absolute errors (MAEs) in energy predictions relative to DFT calculations. Therefore, this study places particular emphasis on the detailed analysis of relative energies among these structures.

Significant differences are observed between the two types of DLP models. Notably, the reaction pathways leading to NR_4^+(R) (Fig. S1a–d) are mirror images of those leading to NR_4^+(S) (Fig. 1f–i). Thus, the analysis focuses exclusively on the reaction pathways towards NR_4^+(S). Firstly, the DLP model trained without structures along reaction pathways predicts unphysical stable states (Fig. 1f and i) for the pathways a and d (Fig. 1a and d) and overestimates the energy barriers (Fig. 1g and h) for the pathways b and c (Fig. 1b and c), yielding values of ∼33.0 and ∼34.8 kcal mol^{-1}, respectively. In contrast, the DLP model trained with structures along reaction pathways eliminates these unphysical stable states and more accurately predicts energy barriers: ∼13.3 and ∼10.9 kcal mol^{-1}, compared to ∼19.2 and ∼12.9 kcal mol^{-1} obtained from DFT calculations. Moreover, the quaternary ammonium salt ($NR_4^+Br^-$) is predicted to be more stable by the DLP model trained with structures along reaction pathways than by the DFT method (Fig. 1f–h). These discrepancies primarily stem from the composition of the training dataset: 1202 structures corresponding to key reaction pathways labeled using DFT calculations in vacuum *versus* 173 857 structures derived from AIMD trajectories performed with explicit solvent molecules. Previous studies have shown that DLP models are trained separately for the same reaction process in implicit and various explicit solvents, since solvation strongly influences both reaction energetics and structural fluctuations.[51] It is normal that the same reaction pathway exhibits different energy barriers in vacuum *versus* in the presence of explicit solvents. However, it remains challenging for a single DLP model to accurately describe the same reaction pathway consistently in both vacuum and solution phases. The DLP model is pre-trained based on vacuum DFT-labeled structures to capture key reactive features, including energy barriers. The pre-trained model is then refined using both the 1202 vacuum structures and the 173 857 AIMD-derived configurations. Because the AIMD data vastly outnumber the vacuum structures, the final DLP model is more strongly influenced by solvation effects. This enhances its accuracy in reproducing solvent-stabilized intermediates and transition states, while slightly underestimating intrinsic reaction barriers compared with DFT calculations in vacuum. Secondly, both the DFT method and the DLP model trained with structures along reaction pathways consistently identified the pathway c (Fig. 1c) as the most favorable. In contrast, the model trained without structures along reaction pathways incorrectly predicted the pathway d (Fig. 1d) as the preferred pathway, despite being physically implausible based on known mechanistic insights. Specifically, the lifetimes of carbocation intermediates typically range from nanoseconds to picoseconds, as determined by laser flash photolysis, azide clock experiments, and theoretical calculations.[73–79] However, as shown in Fig. 1i, the allylic carbocation is erroneously identified as a stable species by the DLP model trained without the inclusion of structures along the reaction pathways, which contradicts experimental evidence. To further examine the stability of the allylic carbocation, MD simulations are conducted using both DLP models. The two models yield completely different PESs (Fig. 2), once again highlighting the critical importance of including reaction pathway structures during training.

In both Fig. 2a and b, five representative states are observed:

1. When CV1 is ∼ 0.06 and CV2 is ∼ 1.73, the system contains one NR_3 molecule and two allyl bromide molecules.

2. When CV1 is ∼ 0.06 and CV2 decreases to ∼1.05, the dissociation of one allyl bromide molecule generates an allylic carbocation and a bromide anion.

3. When CV1 is ∼ 0.05 and CV2 is ∼ 0.35, both allyl bromides dissociate, yielding two allylic carbocations and two bromide anions.

4. When CV1 reaches ∼0.81 and CV2 is ∼ 1.05, the system comprises one NR_4^+ cation, one bromide anion, and one allyl bromide.

5. When CV1 is ∼0.81 and CV2 is ∼0.35, the system consists of one NR_4^+ cation, two bromide anions and one allylic carbocation.

In the DLP model trained with structures along reaction pathways, allyl bromide, the NR_3 molecule, and the NR_4^+ cations are correctly identified as the most stable species in solution. In contrast, in the model trained without these pathway structures, the allylic carbocation is incorrectly described as the most stable species, conflicting with experimental results.

The main reaction process and product molecules in MD simulations

To elucidate the underlying mechanism responsible for the experimentally observed high enantioselectivity, DFT calculations and five independent MD simulations are performed for reaction systems catalyzed by R-BINOL and S-BINOL, respectively. The MD results clearly demonstrate that NR_4^+(S) species predominate in systems with R-BINOL, while NR_4^+(R) is more prevalent with S-BINOL (Fig. 3a). These results are consistent with the absolute configuration (AC) of the major product, as determined experimentally *via* chiral nuclear magnetic resonance (NMR).[12] This well-established technique enables AC assignment based on chemical shifts sensitive to chiral environments.[12,80] Detailed tracking of the reactants and products, as well as population statistics for product molecules, is provided in Fig. S4, with the number of observed reaction events summarized in Table S1. Among the five reaction pathways, reaction pathway b (Fig. 1b) emerges as the dominant one in MD simulations (Fig. 3b), whereas both pathways b and c (Fig. 1b and c) are predicted to be competitive in DFT calculations (Fig. S5). This discrepancy is primarily attributed to the limitations of implicit solvent models in DFT calculations, which do not capture the explicit solvent effects present in MD simulations. A critical distinction between pathways b and c (Fig. 1b and c) lies in the formation of a hydrogen bond between the bromide ion and the hydroxyl group of the BINOL catalyst, which is essential for pathway c (Fig. 1c). However, radial distribution functions (RDFs, Fig. 3c) of chloroform molecules around the bromine atom (Br–H) reveal the presence of a strong hydrogen bond at approximately 3.3 Å between the bromide and chloroform molecules. This hydrogen bonding competes with and effectively inhibits the formation of a hydrogen bond between the bromide and the hydroxyl group in BINOL. The DLP model further supports this observation by accurately reproducing the energetic barrier associated with this hydrogen bond formation (Fig. S2b and c). Therefore, the absence of pathway c (Fig. 1c) in MD simulations is not due to artifacts of the potential energy surface (*e.g.*, unphysical stable states), but rather due to solvation effects that hinder the formation of the key hydrogen bond required for pathway c (Fig. 1c).

Increasing the occurrence of pathway c (Fig. 1c) would enhance the selectivity toward NR_4^+(S) or NR_4^+(R), depending on the chirality of the BINOL catalyst. Since the bromide-hydroxyl hydrogen bond is critical for pathway c (Fig. 1c), reducing solvent polarity is expected to facilitate its formation. This would favor the reaction pathway c (Fig. 1c) and potentially

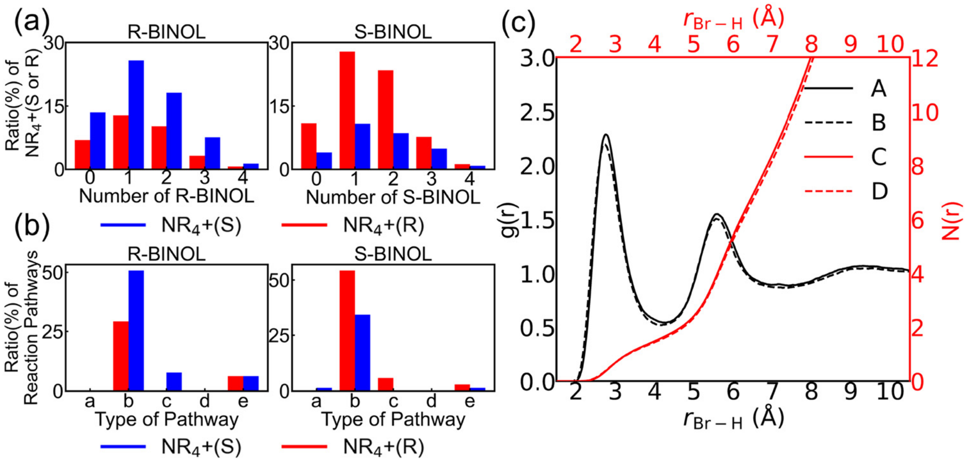


Fig. 3 (a) Ratios of NR_4^+(S) (blue) and NR_4^+(R) (red) observed in all five MD simulations catalyzed by R-BINOL or S-BINOL. (b) Ratios of five reaction pathways a–e (Fig. 1a–e) yielding NR_4^+(S) and NR_4^+(R). (c) Radial distribution functions (RDFs) of Br–H interactions between bromide and chloroform molecules for systems catalyzed by BINOL. Note: the horizontal axis represents the number of BINOL molecules surrounding each product in (a) and reaction pathway types in (b). In (c), panels A and B show $g(r)$ and panels C and D show $N(r)$, from simulations catalyzed by R-BINOL (A and C) and S-BINOL (B and D) using the DLP model.

increase the enantiomeric excess. To evaluate this hypothesis, DFT calculations are performed using implicit solvent models with varying dielectric constants. As shown in Fig. S6a, the absolute binding energy[81–83] between allyl bromide and solvent molecules decreases as solvent polarity decreases. This suggests that in weakly polar solvents, BINOL interacts more favorably with allyl bromide. Notably, when the dielectric constant drops below 4.71, the decrease in binding energy becomes less pronounced, indicating a possible threshold effect.

Further analysis of TS energies shows that in highly polar solvents such as dimethyl sulfoxide (DMSO) and methanol, the barrier for TS2S is lower than that for TS3S by 1.9 and 2.3 kcal mol^{-1}, respectively (Table S2). In contrast, in weakly polar solvents such as chloroform, carbon tetrachloride (CTC), and cyclohexane, the energy barriers for TS3S and TS2S are nearly the same. The binding energies and transition state energies indicate that high solvent polarity suppresses the influence of BINOL by disfavoring the hydrogen bond required for the pathway c (Fig. 1c), thereby reducing the enantioselectivity. However, further decreasing the solvent polarity below that of chloroform does not significantly enhance the binding interaction or reduce the energy barriers. This suggests that the maximum enhancement in enantioselectivity is achieved near this polarity threshold.

Although reaction pathway b (Fig. 1b) dominates in MD simulations, it alone cannot account for the high enantioselectivity observed experimentally, especially when solvent effects and BINOL interactions are excluded. In addition, as shown in Fig. S5a, DFT results show that neither R-BINOL nor S-BINOL significantly lowers the activation barriers for the formation of NR_4^+(S) or NR_4^+(R), suggesting that transition state stabilization is not the primary determinant of enantioselectivity. We therefore propose that BINOL interacts with both the NR_3 substrate and the NR_4^+ product. This interaction may influence the thermodynamic stability of the cationic product, while the binding between BINOL and the NR_3 substrate may regulate the regioselective approach of the allyl bromide toward NR_3, thereby contributing to the observed enantioselectivity. The binding modes of BINOL with both NR_4^+ and NR_3 are discussed in the following sections.

Binding modes of BINOL with product molecules observed in crystal structures

In addition to the predominant reaction pathways observed in simulations, clustering analysis (Scheme 1c) is conducted to investigate the binding modes between BINOL and both the product and substrate molecules. Notably, most product molecules are bound to BINOL within the system (Fig. 3a). Specifically, NR_4^+ (R) and NR_4^+ (S) species that form a hydrogen bond between the hydroxyl group in BINOL and the bromine atom are referred to as R_Br and S_Br, respectively. In contrast, species lacking this hydrogen bond are denoted R and S. For substrate molecules, those that lead to NR_4^+(R) and NR_4^+(S) formation are labeled Sub_R and Sub_S, respectively. Furthermore, the energy profiles for the binding of product and substrate molecules with R-BINOL are presented in Fig. S7–S9, while the corresponding profiles with S-BINOL are presented in Fig. S10–S12. The upper energy profiles are calculated at the PBE/DZVP-MOLOPT-SR-GTH level of theory. The lower panels show results from the DLP model. Notably, the DLP model successfully reproduces the energy landscape calculated at the PBE/DZVP-MOLOPT-SR-GTH level of theory.

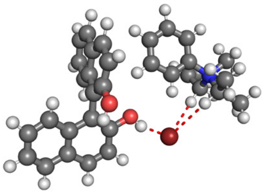
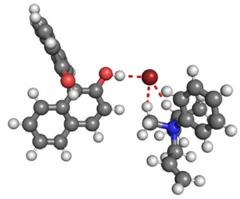
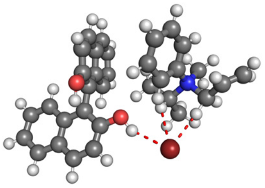
Equivalent atoms (Scheme 2) can lead to structurally identical complexes being misclassified as distinct during cluster analysis. To address this, the twenty lowest-energy binding structures are selected for further cluster analysis. Specifically, K-means clustering[52,53] is applied using root mean square deviation (RMSD) values that treat equivalent atoms as indistinguishable. An RMSD cutoff of 1.5 Å is applied to define structural similarity within each cluster.

Experimental crystal structures[12] demonstrate that BINOL and product molecules are connected *via* a hydrogen bond between the bromine atom and the hydroxyl group of BINOL. The twenty lowest-energy binding structures for both R-BINOL-S_Br and R-BINOL-R_Br are classified into five groups (Fig. S13 and S14), while those for S-BINOL-S_Br and S-BINOL-R_Br are classified into six and three groups (Fig. S19 and S20), respectively. Crucially, Tables 1(a and b) present two distinct binding modes between BINOL and the product molecules. Specifically, in addition to the hydrogen bond between the bromine atom and the R-BINOL hydroxyl group, the first mode involves two additional hydrogen bonds with the isopropyl and propyl groups, whereas the second mode features a hydrogen bond between the bromine atom and the methyl group. In Tables 1(a and b), both binding modes are observed for R_Br, while only the first mode is observed for S_Br in both R-BINOL- and S-BINOL-catalyzed systems. Since R-BINOL is the mirror image of S-BINOL, it is theoretically possible to observe mirror-image binding modes in systems catalyzed by S-BINOL with comparable energy to those in systems catalyzed by R-BINOL. These results support the existence of two distinct binding modes between BINOL and the product molecules. Energetic analysis (Tables 1(a and b)) reveals that the R-BINOL-S_Br (or S-BINOL-R_Br) binding structure is significantly more stable than that of R-BINOL-R_Br (or S-BINOL-S_Br). Moreover, the first mode is consistently more stable than the second. These results are consistent with the experimentally observed enantioselectivity of BINOL.[12]

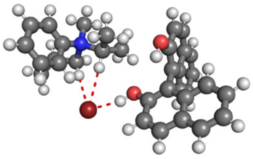
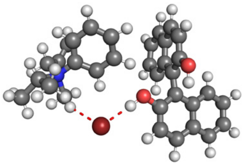
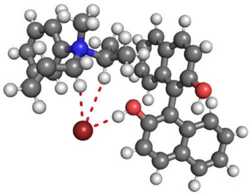
Furthermore, the same binding modes are observed in experimental crystal structures. Specifically, Table 2(a) shows that R-BINOL-S_Br (1) forms two hydrogen bonds: one with the isopropyl group and the other with the propyl group. In contrast, both R-BINOL-R_Br (1) and R-BINOL-R_Br (2) form only one hydrogen bond between the bromine atom and the methyl group. Table 2(b) displays the same binding modes in mirror-image structures corresponding to those in Table 2(a). S-BINOL-R_Br (1) (Table 2(b)) forms two hydrogen bonds with the isopropyl and propyl groups, while S-BINOL-S_Br (1) and S-BINOL-S_Br (2) (Table 2(b)) each form only one single hydrogen bond with the methyl group. These observations demonstrate that DLPMD simulations reliably predict the binding modes identified in the crystal structures. However, despite this agreement in binding modes, the DLP model shows notable discrepancies in computed relative energies. Although it reproduces the

Table 1 Relative energies of BINOL-R_Br and BINOL-S_Br binding structures in MD simulations (a and b)

(a) R-BINOL

Index	R-BINOL-R_Br (1)	R-BINOL-R_Br (2)	R-BINOL-S_Br (1)
Structure			
DFT method	4.74 kcal mol^{-1}	8.79 kcal mol^{-1}	0 kcal mol^{-1}
DLP model	0.67 kcal mol^{-1}	7.05 kcal mol^{-1}	0 kcal mol^{-1}

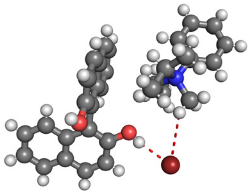
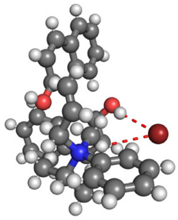
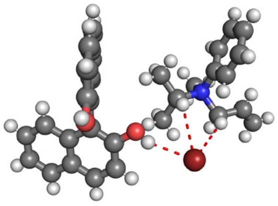
(b) S-BINOL

Index	S-BINOL-R_Br (1)	S-BINOL-R_Br (3)	S-BINOL-S_Br (1)
Structure			
DFT method	0 kcal mol^{-1}	16.58 kcal mol^{-1}	8.28 kcal mol^{-1}
DLP model	0 kcal mol^{-1}	4.00 kcal mol^{-1}	4.16 kcal mol^{-1}

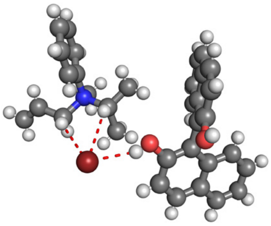
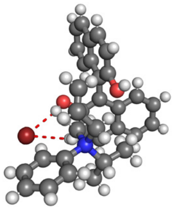
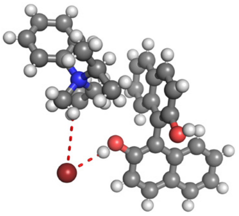
Note: the structure indices in (a) correspond to those in Fig. S13 and S14, while the structure indices in (b) correspond to those in Fig. S19 and S20.

Table 2 Relative energies of BINOL-R_Br and BINOL-S_Br binding structures in experimental crystal structures (a and b)

(a) R-BINOL

Index	R-BINOL-R_Br (2)	R-BINOL-R_Br (1)	R-BINOL-S_Br (1)
Crystal Structure			
DFT method	23.48 kcal mol^{-1}	7.56 kcal mol^{-1}	0 kcal mol^{-1}
DLP model	38.42 kcal mol^{-1}	14.81 kcal mol^{-1}	0 kcal mol^{-1}

(b) S-BINOL

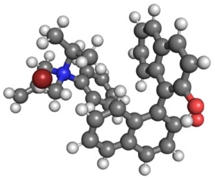
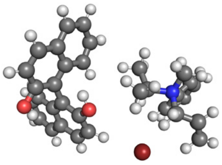
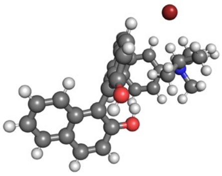
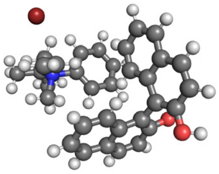
Index	S-BINOL-R_Br (1)	S-BINOL-S_Br (1)	S-BINOL-S_Br (2)
Structure			
DFT method	0 kcal mol^{-1}	8.02 kcal mol^{-1}	23.95 kcal mol^{-1}
DLP model	0 kcal mol^{-1}	15.23 kcal mol^{-1}	38.83 kcal mol^{-1}

Note: the R-BINOL-R_Br binding structures ((a)), which are absent in the experimental data,[12] are the mirror images of the S-BINOL-S_Br binding structures ((b)).

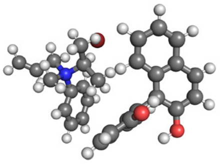
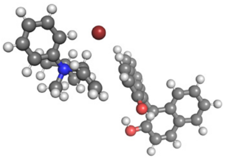
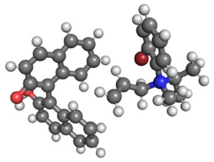
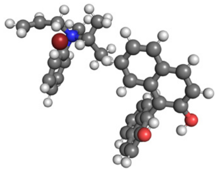
relative energy trends among the structures in Tables 1–4, the computed values deviate significantly from those obtained with the DFT method. As mentioned in the section titled "The Importance of Reaction Pathways for the Accuracy of DLP Model", the training datasets are dominated by structures obtained from AIMD trajectories with explicit solvent molecules. This bias makes it challenging for a single DLP model to accurately describe the same binding structures consistently in both vacuum and solution phases, accounting for the discrepancies between the DFT and DLP results shown in Tables 1–4.

Table 3 Relative energies of BINOL-R and BINOL-S binding structures in MD simulations (a and b)

(a) R-BINOL

Index	R-BINOL-R (1)	R-BINOL-R (2)	R-BINOL-S (1)	R-BINOL-S (2)
Structure				
DFT method	1.05 kcal mol^{-1}	1.70 kcal mol^{-1}	0 kcal mol^{-1}	0.42 kcal mol^{-1}
DLP model	2.42 kcal mol^{-1}	6.05 kcal mol^{-1}	0 kcal mol^{-1}	4.72 kcal mol^{-1}

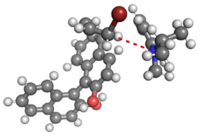
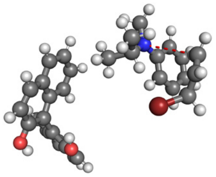
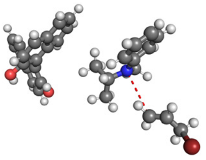
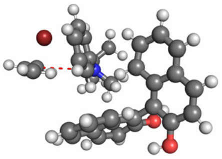
(b) S-BINOL

Index	S-BINOL-R (1)	S-BINOL-R (2)	S-BINOL-S (1)	S-BINOL-S (2)
Structure				
DFT method	0 kcal mol^{-1}	2.05 kcal mol^{-1}	1.65 kcal mol^{-1}	3.12 kcal mol^{-1}
DLP model	0.28 kcal mol^{-1}	0 kcal mol^{-1}	2.44 kcal mol^{-1}	2.60 kcal mol^{-1}

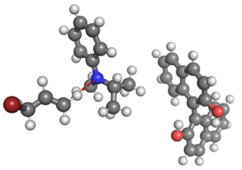
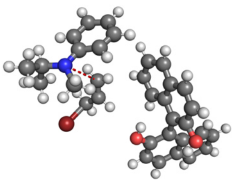
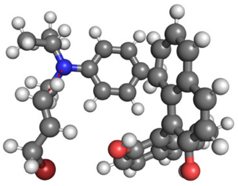
Note: the structure indices in (a) correspond to those in Fig. S15 and S16, while the structure indices in (b) correspond to those in Fig. S21 and S22.

Table 4 Relative energies of BINOL-Sub_R and BINOL-Sub_S binding structures in MD simulations (a and b)

(a) R-BINOL

Index	R-BINOL-Sub_R (1)	R-BINOL-Sub_R (2)	R-BINOL-Sub_S (1)	R-BINOL-Sub_S (2)
Structure				
DFT method	1.44 kcal mol^{-1}	3.70 kcal mol^{-1}	0 kcal mol^{-1}	2.19 kcal mol^{-1}
DLP model	8.65 kcal mol^{-1}	8.02 kcal mol^{-1}	0 kcal mol^{-1}	5.93 kcal mol^{-1}

(b) S-BINOL

Index	S-BINOL-Sub_R (MRS1)^{A}	S-BINOL-Sub_R (1)	S-BINOL-Sub_S (5)
Structure			
DFT method	0 kcal mol^{-1}	4.60 kcal mol^{-1}	4.07 kcal mol^{-1}
DLP model	0 kcal mol^{-1}	5.95 kcal mol^{-1}	2.07 kcal mol^{-1}

Note: the structure indices in (a) correspond to those in Fig. S17 and S18, while the structure indices in (b) correspond to those in Fig. S23 and S24. A. The S-BINOL-Sub_R (MRS1) binding structure is the mirror image of the R-BINOL-Sub_S (1) binding structure, shown in Table 4a.

Binding modes of BINOL with product and substrate molecules in solution

The twenty lowest-energy binding structures for R-BINOL-S and R-BINOL-R are classified into eight and six groups (Fig. S15 and S16), respectively, while those for S-BINOL-S and S-BINOL-R are classified into nine and six groups (Fig. S21 and S22), respectively. Due to the lack of further discernible structural features, only the two lowest-energy binding structures for each case are presented in Tables 3(a and b). Notably, the energies calculated by both the DFT method and the DLP model identify R-BINOL-S (1) as the most stable structure in Table 3(a), while the energies calculated by the DFT method identify S-BINOL-R (1) as the most stable structure in Table 3(b). The relative energies (Tables 3(a and b)) demonstrate that R-BINOL-S (or S-BINOL-R) is more stable than R-BINOL-R (or S-BINOL-S). Therefore, NR_4^+(S) predominates in solution systems catalyzed by R-BINOL, while NR_4^+(R) predominates in solution systems catalyzed by S-BINOL.

Similarly, the twenty lowest-energy binding structures for R-BINOL-Sub_S and R-BINOL-Sub_R are classified into nine and seven groups (Fig. S17 and S18), respectively. The twenty lowest-energy binding structures for S-BINOL-Sub_S and S-BINOL-Sub_R are classified into eight and seven groups (Fig. S23 and S24), respectively. The two lowest-energy binding structures in systems catalyzed by R-BINOL are shown in Table 4(a). For S-BINOL, firstly, although the energy of S-BINOL-Sub_S(1), as calculated by both the DFT method and the DLP model (Fig. S23), is lower than that of S-BINOL-Sub_R binding structures (Fig. S24), S-BINOL fails to regulate the regioselective approach of the allyl bromide toward the NR_3 for S-BINOL-Sub_S(1) and S-BINOL-Sub_S(2) (Fig. S23). Secondly, S-BINOL-Sub_S (3, 4, 6, 7 and 8) proceeds through an unfavorable pathway a (Fig. 1a), which hinders the reaction towards NR_4^+(S). Thirdly, S-BINOL-Sub_R(MRS1), the mirror-image of R-BINOL-Sub_S (1) is identified by both the DFT method and the DLP model in this study. As a result, S-BINOL-Sub_R (MRS1), S-BINOL-Sub_R (1) and S-BINOL-Sub_S (5) are presented in Table 4(b). Energies calculated by both the DFT method and the DLP model consistently indicate that R-BINOL-Sub_S (1) is the most stable structure in Table 4(a), while S-BINOL-Sub_R(MRS1) is the most stable structure in Table 4(b). The relative energies (Tables 4(a and b)) demonstrate that R-BINOL-Sub_S (or S-BINOL-Sub_R) binding structure is more stable than the R-BINOL-Sub_R (or S-BINOL-Sub_S). Therefore, Sub_S is dominant in solution systems catalyzed by R-BINOL while Sub_R is dominant in solution systems catalyzed by S-BINOL. These results suggest that, for substrate molecules, the probability of forming NR_4^+(S) is higher than that of forming NR_4^+I in solution systems catalyzed by R-BINOL while the probability of forming NI(R) is higher than that of forming NR_4^+(S) in solution systems catalyzed by S-BINOL.

Table 5 The energy and force errors for the DLP model using the test dataset

(a) R-BINOL

Simulation index	Number of structures	MAEs Energy (eV per atom)	MAEs Force (eV Å^{-1})
1	1608	0.00941	0.04197
2	1005	0.01385	0.03837
3	1206	0.01440	0.03986
4	1407	0.01433	0.04239
5	1206	0.01657	0.03781

(b) S-BINOL

Simulation index	Number of structures	MAEs Energy (eV per atom)	MAEs Force (eV Å^{-1})
1	1407	0.01084	0.03890
2	804	0.01611	0.03799
3	1407	0.01031	0.04112
4	1206	0.01633	0.04307
5	1236	0.00696	0.04002

Note: if AIMD simulations for structures are not converged, these AIMD simulation trajectories cannot be added to the test dataset.

The major binding modes between BINOL and both product and substrate molecules in solution are consistent with the experimentally observed selectivity of BINOL.[12] These results collectively indicate that the chirality of the main product is under thermodynamic control.

The accuracy of the DLP model

To generate the test dataset, the final snapshot of the 5 ns MD simulation trajectory is divided into eight cubic systems, and 2 ps AIMD simulations are subsequently performed for each of them. More details on the generation of the test datasets are provided in the Methods section (Active Learning Workflow). The MAEs of energies and forces are presented in Table 5. For forces, the MAE values show that the performance of the DLP model on the test dataset is comparable to that of 0.04 eV Å^{-1} (ref. 35) and more accurate than the values in previous studies: 0.26 eV Å^{-1} (ref. 50) and 0.12 eV Å^{-1}.[23] For energies, the MAE values show that the performance of the DLP model on the test dataset is comparable to that of 0.009 eV per atom[84] and significantly lower than that of 0.14 eV per atom[23] reported in previous work. In addition, in the second stage of training, the DLP model is trained with fixed prefactors that emphasize force accuracy. As a result, the analysis primarily focuses on the relative energies of structures calculated by the DLP model. The DLP model successfully reproduces the energy barriers along reaction pathways a–d (Fig. 1f–i and Fig. S1a–d), as well as the relative energies of different binding configurations between BINOL and both product and substrate molecules (Tables 1–4). These results suggest that the enantioselective synthesis mechanisms of ammonium cations in solution can be effectively explored using the DLP model.

Furthermore, the test dataset in Table 5 is divided into two subsets: structures with and without a net charge. The MAEs for neutral structures are lower than those for charged structures (Table S3), indicating that the DLP model performs more accurately on neutral systems. This discrepancy likely arises because electrostatic interactions are not defined in the se_atten model,[72] resulting in larger MAEs for charged structures than for neutral structures.

Conclusions

In this study, an integrated iterative workflow that combines AIMD with DLPMD simulations is developed to investigate the mechanism of the high enantioselectivity of BINOL in the synthesis of quaternary ammonium cations. The key reaction pathways, as well as predominant binding structures of BINOL with both product and substrate molecules, are systematically characterized. These results reveal that the chirality of the major product molecules is governed by thermodynamic rather than kinetic control. To assess solvent effects, DFT calculations employing implicit solvent models with varying dielectric constants are performed, highlighting that solvent polarity significantly impacts the reaction mechanism. However, the DFT calculations fail to fully reproduce the experimentally observed stereoselectivity due to the prohibitive computational costs in capturing the complexity of the reaction environment. In contrast, the proposed workflow enables the mechanistic exploration of asymmetric synthesis in complex solution environments by leveraging long-timescale and large-system MD simulations. It provides more comprehensive mechanistic insights into reactions under realistic solvation conditions. In addition, the active learning workflow converges rapidly, requiring only seven cycles, demonstrating the efficiency of this workflow. Consequently, the workflow holds promise for the accurate prediction of major products in asymmetric synthesis and provides an efficient tool for catalyst and solvent screening. To further advance the methodology, future implementations should incorporate explicit counterions in AIMD simulations to properly neutralize local charges, addressing limitations identified by the MAE analysis (Table S3). Additionally, developing new deep learning potentials that explicitly model electrostatic interactions will be essential to improve accuracy when simulating charged systems.

Author contributions

G. H. L. conceived and supervised the project. H. Q. C. developed the code. H. Q. C., D. Z. and Y. L. carried out the calculations. H. Q. C. prepared the original draft, which was further reviewed and edited by Y. L. and H. Y. C. All authors analyzed and discussed the results as well as assisted in manuscript preparation.

Conflicts of interest

The authors declare no competing financial interests.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information: more details about computational methods, additional Schemes, Figures and Tables. See DOI: **https://doi.org/10.1039/d5cp03439f**.

The source code has been deposited in our GitHub repository **https://github.com/cui-hq/AIMD_DLPMD**.

Acknowledgements

This work was supported by the National Key R&D Program of China (2023YFF1204903 (H. Y. C.)) and the LiaoNing Revitalization Talents Program (No. XLYC2402023).

References

1 K. Ohmatsu and T. Ooi, *Top. Curr. Chem.*, 2019, **377**, 22.

2 D. Y. Qian and J. W. Sun, *Chem. – Eur. J.*, 2019, **25**, 3740–3751.

3 N. Nikfarjam, M. Ghomi, T. Agarwal, M. Hassanpour, E. Sharifi, D. Khorsandi, M. A. Khan, F. Rossi, A. Rossetti, E. N. Zare, N. Rabiee, D. Afshar, M. Vosough, T. K. Maiti, V. Mattoli, E. Lichtfouse, F. R. Tay and P. Makvandi, *Adv. Funct. Mater.*, 2021, **31**, 27.

4 C. E. Diesendruck and D. R. Dekel, *Curr. Opin. Electrochem.*, 2018, **9**, 173–178.

5 D. R. Brown, R. Lygo, J. McKenna and B. G. Hutley, *J. Chem. Soc. B: Phys. Org.*, 1967, 1184–1195.

6 E. Wedekind and O. Wedekind, *Ber. Dtsch. Chem. Ges.*, 1907, **40**, 4450–4456.

7 W. J. Pope and S. J. Peachey, *J. Chem. Soc., Trans.*, 1899, **75**, 1127–1131.

8 E. Havinga, *Biochim. Biophys. Acta*, 1954, **13**, 171–174.

9 R. G. Kostyanovsky, K. A. Lyssenko, O. N. Krutius and V. R. Kostyanovsky, *Mendeleev Commun.*, 2009, **19**, 19–20.

10 V. Y. Torbeev, K. A. Lyssenko, O. N. Kharybin, M. Y. Antipin and R. G. Kostyanovsky, *J. Phys. Chem. B*, 2003, **107**, 13523–13531.

11 U. H. Dolling, P. Davis and E. J. J. Grabowski, *J. Am. Chem. Soc.*, 1984, **106**, 446–447.

12 M. P. Walsh, J. M. Phelps, M. E. Lennon, D. S. Yufit and M. O. Kitching, *Nature*, 2021, **597**, 70–76.

13 C. J. Cerjan and W. H. Miller, *J. Chem. Phys.*, 1981, **75**, 2800–2806.

14 J. Baker, *J. Comput. Chem.*, 1986, **7**, 385–395.

15 A. Heyden, A. T. Bell and F. J. Keil, *J. Chem. Phys.*, 2005, **123**, 14.

16 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.

17 V. Asgeirsson, B. O. Birgisson, R. Bjornsson, U. Becker, F. Neese, C. Riplinger and H. Jónsson, *J. Chem. Theory Comput.*, 2021, **17**, 4929–4945.

18 E. Weinan, W. Q. Ren and E. Vanden-Eijnden, *J. Phys. Chem. B*, 2005, **109**, 6688–6693.

19 P. M. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.

20 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.

21 G. Scalmani and M. J. Frisch, *J. Chem. Phys.*, 2010, **132**, 15.

22 G. Norjmaa, G. Ujaque and A. Lledós, *Top. Catal.*, 2022, **65**, 118–140.
23 J. Z. Zeng, L. Q. Cao, M. Y. Xu, T. Zhu and J. Z. H. Zhang, *Nat. Commun.*, 2020, **11**, 9.
24 Y. X. Zuo, C. Chen, X. G. Li, Z. Deng, Y. M. Chen, J. Behler, G. Csanyi, A. V. Shapeev, A. P. Thompson, M. A. Wood and S. P. Ong, *J. Phys. Chem. A*, 2020, **124**, 731–745.
25 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 4.
26 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 4.
27 R. K. Lindsey, L. E. Fried and N. Goldman, *J. Chem. Theory Comput.*, 2017, **13**, 6222–6229.
28 M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K. R. Müller and K. Burke, *Nat. Commun.*, 2020, **11**, 11.
29 L. F. Zhang, J. Q. Han, H. Wang, R. Car and E. Weinan, *Phys. Rev. Lett.*, 2018, **120**, 6.
30 D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Phys. Rev. Mater.*, 2018, **2**, 16.
31 S. Chmiela, H. E. Sauceda, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2018, **9**, 10.
32 K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K. R. Müller, *31st Annual Conference on Neural Information Processing Systems (NIPS) Long Beach*, CA, 2017.
33 G. L. Zhang, Y. Zhang, L. Li, J. Y. Zhou, H. L. Chen, J. W. Ji, Y. R. Li, Y. Cao, Z. H. Xu and C. Pian, *Interdiscip. Sci.*, 2024, **16**, 712–726.
34 Y. W. Liu, R. S. Zhang, Y. N. Yuan, J. Ma, T. F. Li and Z. X. Yu, *Interdiscip. Sci.*, 2024, **16**, 741–754.
35 M. Y. Yang, U. Raucci and M. Parrinello, *Nat. Catal.*, 2023, **6**, 829–836.
36 Y. Z. Zhang, H. D. Wang, W. J. Chen, J. Z. Zeng, L. F. Zhang, H. Wang and E. Weinan, *Comput. Phys. Commun.*, 2020, **253**, 11.
37 M. Y. Yang, L. Bonati, D. Polino and M. Parrinello, *Catal. Today*, 2022, **387**, 143–149.
38 S. Wille, H. Y. Jiang, O. Bünermann, A. M. Wodtke, J. Behler and A. Kandratsenka, *Phys. Chem. Chem. Phys.*, 2020, **22**, 26113–26120.
39 L. F. Zhang, J. Q. Han, H. Wang, W. A. Saidi, R. Car and W. E, *32nd Conference on Neural Information Processing Systems (NIPS) Montreal*, CANADA, 2018.
40 L. F. Zhang, H. Wang, R. Car and E. Weinan, *Phys. Rev. Lett.*, 2021, **126**, 7.
41 T. Devergne, T. Magrino, F. Pietrucci and A. M. Saitta, *J. Chem. Theory Comput.*, 2022, **18**, 5410–5421.
42 A. S. Raman and A. Selloni, *J. Phys. Chem. A*, 2022, **126**, 7283–7290.
43 W. Jiang, Y. Zhang, L. Zhang and H. Wang, *Chin. Phys. B*, 2021, **30**, 050706.
44 J. A. Keith, V. Vassilev-Galindo, B. Q. Cheng, S. Chmiela, M. Gastegger, K. R. Mueller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.
45 F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Chem. Rev.*, 2021, **121**, 9759–9815.
46 P. O. Dral, A. Owens, S. N. Yurchenko and W. Thiel, *J. Chem. Phys.*, 2017, **146**, 11.
47 D. P. Hu, Y. Xie, X. S. Li, L. Y. Li and Z. G. Lan, *J. Phys. Chem. Lett.*, 2018, **9**, 2725–2732.
48 N. Artrith and J. Behler, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2012, **85**, 13.
49 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
50 S. H. Zhang, M. Z. Makos, R. B. Jadrich, E. Kraka, K. Barros, B. T. Nebgen, S. Tretiak, O. Isayev, N. Lubbers, R. A. Messerly and J. S. Smith, *Nat. Chem.*, 2024, **16**, 727–734.
51 H. W. Zhang, V. Juraskova and F. Duarte, *Nat. Commun.*, 2024, **15**, 11.
52 L. Bai, J. Y. Liang and C. Y. Dang, *Knowl. Based Syst.*, 2011, **24**, 785–795.
53 D. Arthur and S. Vassilvitskii, *18th Symposium on Discrete Algorithms*, New Orleans, LA, 2007.
54 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L. P. Wang, T. J. Lane and V. S. Pande, *Biophys. J.*, 2015, **109**, 1528–1532.
55 J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
56 H. Bekker, H. J. C. Berendsen, E. J. Dijkstra, S. Achterop, R. Vondrumen, D. Vanderspoel, A. Sijbers, H. Keegstra, B. Reitsma and M. K. R. Renardus, *4th International Conference on Computational Physics (PC 92) Prague*, Czech Republic, 1992.
57 A. Jakalian, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.
58 T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borstnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöss, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack and J. Hutter, *J. Chem. Phys.*, 2020, **152**, 47.
59 T. Lu and F. W. Chen, *J. Comput. Chem.*, 2012, **33**, 580–592.
60 T. Lu, *J. Chem. Phys.*, 2024, **161**, 41.
61 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, 49.
62 J. VandeVondele and J. Hutter, *J. Chem. Phys.*, 2007, **127**, 9.
63 S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 1703–1710.
64 C. Hartwigsen, S. Goedecker and J. Hutter, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 3641–3662.
65 M. Krack, *Theor. Chem. Acc.*, 2005, **114**, 145–152.
66 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
67 J. P. Perdew, K. Burke and M. J. P. R. L. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396–1396.
68 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.
69 G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Comput. Phys. Commun.*, 2014, **185**, 604–613.
70 A. Barducci, G. Bussi and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 4.

71 H. Wang, L. F. Zhang, J. Q. Han and W. E, *Comput. Phys. Commun.*, 2018, **228**, 178–184.

72 D. Zhang, H. Bi, F. Dai, W. Jiang, X. Liu, L. Zhang and H. Wang, *npj Comput. Mater.*, 2024, **8**, 58–65.

73 A. Kumar, S. Mondal and S. Banerjee, *J. Am. Chem. Soc.*, 2021, **143**, 2459–2463.

74 S. Jaarinen, J. Niiranen and J. Koskikallio, *Int. J. Chem. Kinet.*, 1985, **17**, 925–930.

75 D. Watanabe and H. O. Hamaguchi, *Chem. Phys.*, 2008, **354**, 27–31.

76 Y. Chiang and A. J. Kresge, *J. Am. Chem. Soc.*, 1985, **107**, 6363–6367.

77 J. P. Pezacki, D. Shukla, J. Lusztyk and J. Warkentin, *J. Am. Chem. Soc.*, 1999, **121**, 6589–6598.

78 R. P. Pemberton and D. J. Tantillo, *Chem. Sci.*, 2014, **5**, 3301–3308.

79 M. M. Toteva and J. P. Richard, *J. Am. Chem. Soc.*, 1996, **118**, 11434–11445.

80 C. Aroulanda and P. Lesot, *Chirality*, 2022, **34**, 182–244.

81 A. Lee, J. H. Yang, J. H. Oh, B. P. Hay, K. Lee, V. M. Lynch, J. L. Sessler and S. K. Kim, *Chem. Sci.*, 2023, **14**, 1218–1226.

82 W. H. Duan, Q. Wang and F. Collins, *Chem. Sci.*, 2011, **2**, 1407–1413.

83 P. Çarçabal, E. J. Cocinero and J. P. Simons, *Chem. Sci.*, 2013, **4**, 1830–1836.

84 G. Zhou, N. Lubbers, K. Barros, S. Tretiak and B. Nebgen, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**(27), e2120333119.