JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Article

Smart Reaction Templating: A Graph-Based Method for Automated Molecular Dynamics Input Generation

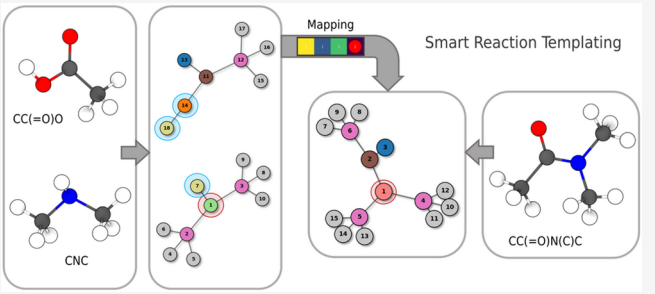
Julian Konrad* and Robert Meißner*

Cite This: J. Chem. Inf. Model. 2025, 65, 6038−6047

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurately modeling chemical reactions in molecular dynamics simulations requires detailed pre- and postreaction templates, often created through labor-intensive manual workflows. This work introduces a Python-based algorithm that automates the generation of reaction templates for the LAMMPS REACTION package, leveraging graph-theoretical principles and subgraph isomorphism techniques. By representing molecular systems as mathematical graphs, the method enables the automated identification of conserved molecular domains, reaction sites, and atom mappings, significantly reducing manual effort. The algorithm was validated on three case studies: poly addition, poly condensation, and chain polymerization, demonstrating its ability to map conserved domains, identify reaction-initiating atoms, and resolve challenges such as symmetric reactants and indistinguishable atoms. Additionally, the generated templates were optimized for computational efficiency by retaining only essential reactive domains, ensuring scalability and consistency in high-throughput workflows for computational chemistry, materials science, and machine learning applications. Future work will focus on extending the method to mixed organic−inorganic systems, incorporating adaptive scoring mechanisms, and integrating quantum mechanical calculations to enhance its applicability.



INTRODUCTION

Modeling chemical systems presents a significant challenge due to the intricate interplay of atoms and their interactions. At the core of this complexity lies the necessity to describe chemical reactions, based on the electronic configurations. Quantum mechanics (QM) provides a robust foundation for this task, offering a highly precise framework to capture the behavior of atoms and their electrons. Methods such as ab initio and density functional theory (DFT) have proven to be invaluable tools in this domain.[1,2] However, due to the immense computational demands, these approaches are limited to relatively small systems, typically comprising no more than a few hundred atoms.

To study chemical phenomena at larger spatial and temporal scales, molecular mechanics is an alternative. Simply speaking, this method simplifies atomic interactions into a "bead-and-spring" model, where intramolecular interactions are approximated as simple linear springs, and nonbonded interactions are represented using potential functions.[3] While less detailed than quantum mechanical methods, molecular mechanics enables the efficient simulation of systems containing thousands or even millions of atoms, facilitating insights into macroscopic properties and dynamic behavior.

One downside of such molecular mechanics (MM) models, which rely on fixed parameters for bonds, angles, dihedrals, and charges is that they are typically unsuitable for capturing chemical reactivity due to their static nature. To overcome this limitation, approaches such as the charge equilibration method[4] and reactive force fields have been developed. These methods enable dynamic charge adjustments and the simulation of bond formation and breaking on the basis of some smooth bond-order dependent functions, thereby allowing for modeling of chemical reactions. ReaxFF, in particular, is a versatile tool capable of handling a wide range of reactive systems, but it often requires substantial computational resources and careful parametrization to align with experimental data or high-level quantum mechanical calculations.[5,6] Similarly, tailor-made reactive force fields offer precise solutions for specific systems but demand labor-intensive calibration processes to ensure their accuracy and reliability.[7]

Another approach that requires less parametrization from quantum mechanical calculations is the Smooth Topology Transfer method, as described by Meißner et al.[8] This method enables a seamless transition between the topologies of reactants and products by blending molecular forces and energies in a controlled manner. While this approach reduces the need for

Received: March 5, 2025
Revised: May 2, 2025
Accepted: May 21, 2025
Published: June 6, 2025

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

extensive quantum mechanical data and incorporates only QM energy correction terms, defining the reactant and product topologies remains a labor-intensive process.

The REACTION package in LAMMPS[9] provides functionality for simulating chemical reactions in molecular systems by utilizing pre- and postreaction templates to define the topologies of reactants and products.[10−12] The process involves mapping individual atoms between these templates and defining reaction criteria, such as distance thresholds or specific bonding conditions. While fix bond/react facilitates the modeling of reactions in large-scale simulations, preparing accurate templates remains a significant challenge. These templates must account for bonds, angles, dihedrals, constraints and special neighbor rules to ensure the consistency of the system's topology.

To address the challenges of template preparation, AutoMapper was developed as a Python-based tool to automate much of the workflow for fix bond/react simulations in LAMMPS.[13] AutoMapper significantly reduces the manual effort required by generating reaction templates from LAMMPS input files. However, users must still manually specify which bonds are formed or broken during the reaction.

A recent study introduced the fully automated algorithm PX−MDsim for the cross-linking of polyamides.[14] This platform, based on the PXLink framework, automates the entire cross-linking simulation process − from input preparation and initial system construction to force field generation, functional group identification, and charge distribution updates. However, its applicability is limited to monomers containing amino and carboxyl groups. Similarly, the reactive algorithm proposed in Provenzano et al.[15] was developed specifically for a single thermoset system, and is therefore restricted to a specific monomer and reaction chemistry.

Recent advancements have explored the use of machine learning for atom mapping in chemical reactions from large reaction data sets.[16,17] These models can accurately predict atom correspondences in a wide range of reactions but require extensive training data and may struggle to generalize beyond the types of reactions seen during training. Other approaches have been developed to define reaction networks and explore chemical space.[18,19] While powerful, these methods often assume that reactants and products contain the same number of atoms, which limits their applicability to systems involving fragmentation, recombination, or addition/elimination reactions.

To enable high-throughput screening of materials and generate diverse structural data sets for machine learning applications, it is essential to fully automate the mapping process between reactants and products. More importantly, our approach imposes no restrictions on specific monomers or types of chemical reactions, making it broadly applicable. In this work, we present a Python-based algorithm that relies exclusively on the definition of LAMMPS data files. These files can be conveniently generated from simple SMILES[20] strings using tools such as LigParGen,[21] Open Babel,[22] and Moltemplate.[23]

By representing molecules as graphs, the algorithm employs graph isomorphism techniques, optimization, and neighborhood analysis to automatically map reactant and product atoms and identify reaction sites. This approach eliminates much of the manual effort traditionally associated with preparing reaction templates and ensures consistent and scalable workflows for complex simulations and data generation tasks.

METHODS

Molecular structures are represented as graphs $G = (V, E)$, where V is the set of nodes (atoms) and E is the set of edges (bonds). Each node $v \in V$ is assigned attributes such as atomic mass $m(v)$, atom type $t(v)$, atomic charge $q(v)$, and a component identifier $c(v)$, which specifies the molecule to which the atom belongs. Edges $e = (v, u) \in E$ connect nodes to represent bonds determined by the bonding information in the input data.

For chemical reactions, the reactants and products are represented as unified graphs, $G_{reac} = (V_{reac}, E_{reac})$ and $G_{prod} = (V_{prod}, E_{prod})$, respectively. These graphs are constructed by combining the individual molecular graphs of the reactants G_r and products G_p:

$$V_{reac} = \bigcup_{i=1}^{n_{reac}} V_{r,i},\ E_{reac} = \bigcup_{i=1}^{n_{reac}} E_{r,i}$$
$$V_{prod} = \bigcup_{j=1}^{n_{prod}} V_{p,j},\ E_{prod} = \bigcup_{j=1}^{n_{prod}} E_{p,j} \quad (1)$$

where n_{reac} and n_{prod} denote the number of reactant and product molecules, and $G_{r,i} = (V_{r,i}, E_{r,i})$ and $G_{p,j} = (V_{p,j}, E_{p,j})$ are the graphs of individual reactant and product molecules.

The construction of $G_{r,i}$ and $G_{p,j}$ involves parsing input data to extract atomic properties $m(v)$, $t(v)$, $q(v)$, and $c(v)$ for each atom. Based on this information, edges $E_{r,i}$ or $E_{p,j}$ are established according to bonding information. To ensure consistency and avoid conflicts, globally unique node identifiers are assigned when combining individual graphs into G_{reac} and G_{prod}. The component identifier $c(v)$ ensures that nodes remain associated with their respective molecules during the mapping process:

$$c(v) = k, \quad \text{if } v \in V_{r,k} \text{ or } v \in V_{p,k} \quad (2)$$

where k corresponds to the index of the reactant or product molecule.

This unified representation of reactants and products as G_{reac} and G_{prod} allows for efficient computation of subgraph alignments, which are essential for identifying reaction sites and transformations. By retaining the modularity of individual molecular graphs, the method ensures scalability and accuracy in handling complex chemical systems.

The identification of structurally conserved domains between the reactants and products is achieved through subgraph isomorphism detection. This step determines correspondences between parts of the molecular graphs that remain chemically unchanged during the reaction. In chemical terms, isomorphic subgraphs are substructures, i.e. molecular fragments, that retain the same atomic types, bonding patterns, and local connectivity before and after the reaction. Crucially, subgraph isomorphism will only succeed in identifying exact structural matches − any atoms or bonds altered by the reaction will no longer satisfy the isomorphism criteria and will therefore not be considered part of the conserved domain.

Subgraph isomorphism detection is performed using the VF2 algorithm via the GraphMatcher class from NetworkX.[24] A subgraph isomorphism is defined as a bijection $f: V'_r \to V'_p$, where $V'_r \subseteq V_{reac}$ and $V'_p \subseteq V_{prod}$, such that

$$(v, u) \in E_{reac} \Rightarrow (f(v), f(u)) \in E_{prod} \quad (3)$$

and node attributes such as atomic mass and type are preserved:

$$m(v) = m(f(v)),\ t(v) = t(f(v)),\ \forall\ v \in V'_r \quad (4)$$

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

The subgraph isomorphism search begins with the largest connected component of the reactant graph, which is then queried within the product graph. To ensure that also smaller fragments − potentially arising from the fragmentation of the reactant structure − are identified, a minimum fragment size is enforced via the parameter n. In this work, $n = 4$ is proposed, corresponding to 1−4 interactions, which represent the maximum relevant intramolecular distances in typical molecular mechanics force fields.

The subgraph isomorphism search operates by iteratively reducing the reactant graphs, beginning with the largest graph G_r. In each iteration, nodes are removed from G_r based on a distance criterion starting from each node, and the remaining graph is decomposed into a collection of connected components. Each of these connected components is then tested for isomorphism against the product graph G_{prod}. The algorithm seeks the maximum possible mapping size by comparing the components and their corresponding subgraphs in G_{prod}. If a valid isomorphism is not found in the current iteration, the number of nodes removed is increased, resulting in smaller connected components of G_r.

It might be necessary to perform multiple iterations of subgraph isomorphism search. Since the reaction can occur at the center of a reactant molecule as shown in Figure 2, it results in two conserved domains that may not be fully captured in a single iteration, as they are no longer connected but are separated by the reaction site.

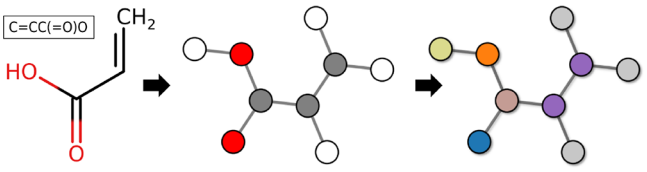


Figure 1. Schematic illustrating the transformation from a chemical structure or SMILES (left) to a network graph. In both the middle and right panels, bonds are represented as edges E between nodes V. The middle panel shows the graph with atoms as nodes, colored according to their atomic mass (element). The right panel displays the graph with nodes colored by the atom type, which in this case is assigned based on the local chemical environment defined by the force field. Atom types could also be derived directly from the molecular structure − e.g., using algorithms that infer types from bonding patterns and atomic elements, independent of a specific force field. This transformation effectively maps the molecular structure to a graph representation, capturing both atomic and bonding information, which is essential and sufficient for the subsequent mapping procedure.

In each iteration, the largest conserved subgraph is identified. For each reactant molecule and product molecule, the unmapped portions of the reactant and product graphs are extracted. These subgraphs exclude all nodes and edges already matched in previous iterations. Initially, these subgraphs exclude no nodes or edges, as the exclusion only occurs after the mapping. Among the unmapped portions, a mapping is sought to maximize the size of the conserved domain:

$$f^* = \arg\max_f |V'_r \cap f(V'_p)| \tag{5}$$

where f represents all possible bijections preserving node attributes and connectivity. The result of this step is a mapping of the largest structurally conserved subgraph between the reactant and product graphs. The remaining unmapped portions

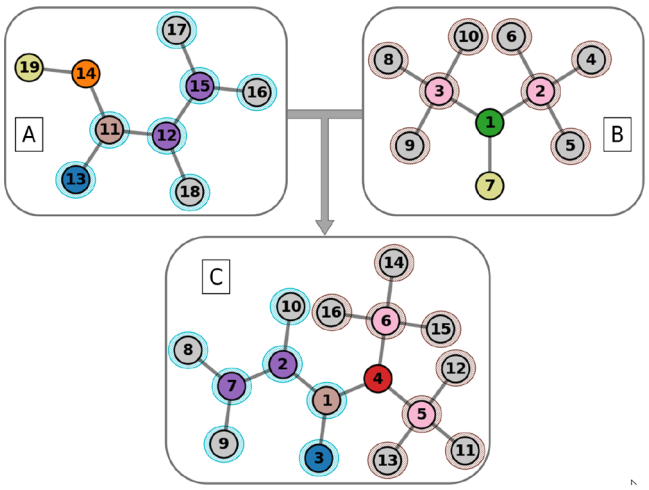


Figure 2. Illustration highlights the conserved domains of graph A and B as they map to graph C. Specifically, nodes 14 and 19 of graph A, and nodes 1 and 7 of graph B, are shown as exceptions, while the remaining fragments of both graphs can be perfectly mapped onto graph C. While the complete substructure of graph A is considered in one iteration, for graph B, two iterations are necessary due to the reaction site (nodes 1 and 7), which fragment the conserved domains into nodes 3,5,9,10 and 2,4,5,6, respectively.

of the graphs are then processed in subsequent iterations to identify smaller conserved subgraphs.

The mappings f^* from all iterations are combined to form the overall correspondence between the reactant and product graphs in conserved domains. This ensures that each node in the reactant graph is mapped to at most one node in the product graph and that the mapping encompasses all conserved domains, regardless of their size or separation. This approach guarantees completeness by systematically processing both large and small conserved domains. Two iterations are sufficient to ensure that conserved domains on both sides of the reaction site are fully mapped and that fragmented components, which might otherwise be missed, are included in the final mapping of the conserved domains.

After identifying conserved substructures using subgraph isomorphism, the remaining unmapped nodes, which correspond to the reaction site, are aligned based on a similarity score that incorporates structural and chemical properties, as well as the local molecular environment.

The similarity score S for a pair of unmapped nodes $\tilde{v}_r \in \tilde{V}_{reac}$ and $\tilde{v}_p \in \tilde{V}_{prod}$ is computed iteratively as

$$S(\tilde{v}_r, \tilde{v}_p) = \frac{1}{|V^*|}\left(w_m \cdot \delta_m + w_t \cdot \delta_t + \frac{1}{d}\sum_{k=1}^{d} w_n \cdot (N_m^{(k)} + N_t^{(k)})\right) \tag{6}$$

where k represents the current neighborhood depth in the graph, with d being the smaller of the longest shortest paths from $\tilde{v}_r$ to any node in V_r or $\tilde{v}_p$ to any node in V_p, ensuring a consistent maximum depth in both graphs. The weights w_m, w_t, and w_n control the relative importance of mass, type, and neighborhood similarity in the score and are normalized such that their sum equals 1, ensuring a balanced contribution of mass, type, and neighborhood similarity. The terms δ_t and δ_m are indicator functions that are 1 if the atom types and masses of the nodes $\tilde{v}_r$ and $\tilde{v}_p$ are identical, and 0 otherwise. The terms $N_t^{(k)}$ and $N_m^{(k)}$ measure the neighborhood similarity in terms of atom type and mass, respectively, and are computed as

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

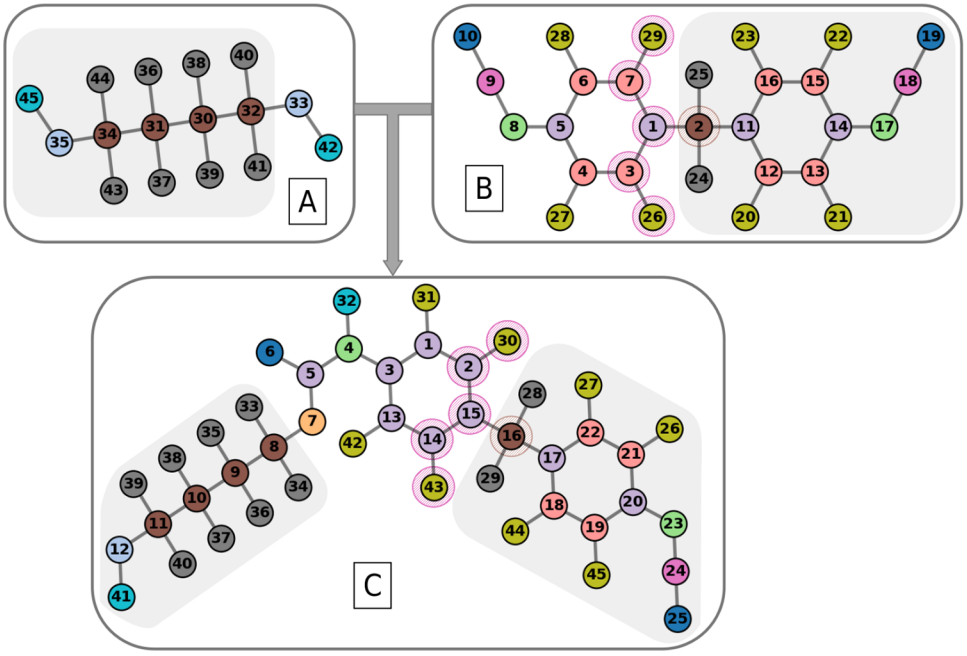


Figure 3. This example illustrates the path consistency, as demonstrated by the paths between nodes 2−26 and 2−29 in graph B (highlighted in pink), and nodes 16−30 and 16−43 in graph C. These paths are indistinguishable because the included nodes have identical similarity scores. The paths originate from node 2 in graph B and node 16 in graph C (highlighted in brown), which serve as correctly mapped anchor points from the subisomorphic conserved domain (gray). If path consistency is not maintained from these anchor points, the mapped nodes are swapped with their indistinguishable counterparts to achieve path consistency in G_{reac} and G_{prod}.

$$N_t^{(k)} = \sum_j \delta_t^{(k)}(j),\ N_m^{(k)} = \sum_j \delta_m^{(k)}(j) \tag{7}$$

where j indexes the neighboring nodes of $\tilde{v}_r$ and $\tilde{v}_p$ at depth k. The normalization factor $|V^*|$ represents the number of nodes considered in the iterative process and is given by

$$|V^*| = \max(|V_{reac}^{(k)}|, |V_{prod}^{(k_{max})}|) \tag{8}$$

Once the similarity scores are computed, the cost matrix $\mathbf{C}$ is constructed as

$$C_{ij} = -S(\tilde{v}_r^i, \tilde{v}_p^j) \tag{9}$$

where each entry represents the negative similarity score between nodes $\tilde{v}_r^i$ and $\tilde{v}_p^j$. To ensure numerical stability, the matrix is normalized. Finally, the optimal one-to-one mapping is determined by solving the linear sum assignment problem:

$$f^* = \arg\min_f \sum_{(i,j)\in f} C_{ij} \tag{10}$$

where f represents the optimal node correspondence between $\tilde{V}_{reac}$ and $\tilde{V}_{prod}$, and the summation runs only over the assigned pairs.

If a molecule is composed of structurally identical paths, certain subgraph traversals become indistinguishable (Figure 3). Thus, isomorphic node anchors p serve as fixed reference points in the mapping process. Iterative adjustments are made to maximize structural consistency by swapping identical nodes and re-evaluating their associations based on bond connectivity in G_{reac} and G_{prod}. Subsequently, paths originating from the reassigned nodes are evaluated to ensure connectivity consistency. For a path $\{p_1, p_2, \cdots, p_k\}$ originating from a node p_1, the reassignment is refined iteratively to maximize structural alignment:

$$\mathcal{A}_{opt} = \arg\max_f \sum_k S^*(p_k, f(p_k)) \tag{11}$$

where f represents the node mapping, and S^* is a similarity function based on node attributes and connectivity. This process guarantees a complete and accurate mapping, ensuring alignment consistency across all conserved domains and reaction-site nodes.

To ensure consistent alignment between nodes in the reactant and product graphs, discrepancies in neighborhood connectivity are addressed when G_{prod} contains more nodes than G_{reac}. This scenario typically arises in reduction reactions, where hydrogen atoms are not explicitly treated and are instead created in the product (Figure 4). During the initial mapping, these newly created hydrogen atoms may be mistakenly mapped to existing atoms, especially when the atoms are in a chemically similar environment. The adjustment process addresses this by comparing the sizes of the neighborhoods of aligned nodes (v_r, v_p).

Let $\mathcal{N}(v_r)$ and $\mathcal{N}(v_p)$ denote the neighborhoods of v_r and v_p, respectively. A mismatch is identified if:

$$|\mathcal{N}(v_r)| \neq |\mathcal{N}(v_p)| \tag{12}$$

To resolve these mismatches, hydrogen-like nodes are swapped. The algorithm identifies candidates for swapping by iterating over the neighbors of v_r and v_p. The sets of such nodes are denoted as

$$\mathcal{N}_{swap}(v_r) = \{v_r' \in \mathcal{N}(v_r) | m(v_r') \approx 1.008\},\ \mathcal{N}_{swap}(v_p) = \{v_p' \in \mathcal{N}(v_p) | m(v_p') \approx 1.008\} \tag{13}$$

where $m(x)$ represents the mass of node x. Pairs of nodes (v_r', v_p') from these sets are identified for swapping, ensuring that connectivity remains consistent.

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

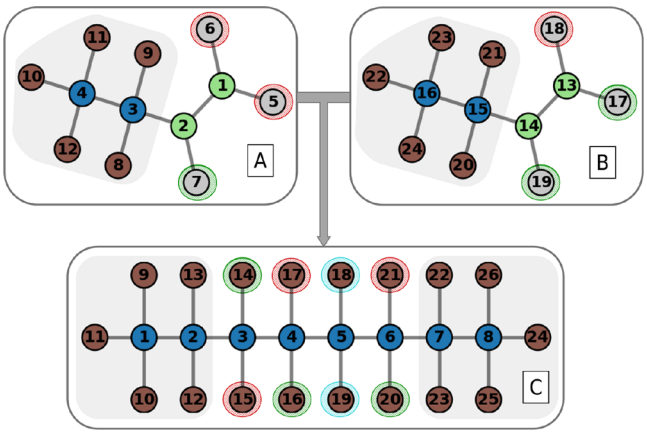


Figure 4. Nodes underlined in green indicate correct mappings, while red nodes represent incorrect mappings, and cyan nodes are those excluded from the mapping. Comparing the neighborhoods of nodes 5 and 3 in graph C reveals inconsistencies when aligned with their counterparts, nodes 1 and 14, respectively. These differences become evident when considering hydrogen-like neighbors: node 5 in graph C is associated with nodes 18 and 19, whereas in the product, newly introduced nodes 15 and 21 appear in the corresponding mappings. This mismatch highlights the need for neighborhood-based adjustments during the mapping process.

Reaction sites are determined by analyzing changes in bond connectivity across molecular subgraphs. To identify newly formed bonds, each aligned node pair (v_r, v_p) is evaluated based on the component labels (eq 2) of its neighboring nodes:

$$C(v) = \{c(u) | u \in \mathcal{N}(v)\} \tag{14}$$

where $\mathcal{N}(v)$ is the set of neighbors of node v, and $c(u)$ denotes the component label of node u. The set $C(v)$ characterizes the local molecular environment by indicating which distinct components are connected to v. A node v_p is considered to participate in a new bond if its neighboring components differ from those of its aligned counterpart in the reactant graph:-

$$\mathcal{C}(v_p) \nsubseteq \mathcal{C}(v_r)$$

Among all such nodes, the primary reaction site is selected based on the largest change in eigenvector centrality:

$$|\Delta\tilde{C}| = |\tilde{C}(v_p) - \tilde{C}(v_r)| \tag{16}$$

where centrality $\tilde{C}(v)$ is defined as

$$\tilde{C}(v) = \frac{1}{\lambda} \sum_{u \in \mathcal{N}(v)} \tilde{C}(u) \tag{17}$$

with λ being the largest eigenvalue of the adjacency matrix.[25] A large $|\Delta\tilde{C}|$ indicates that a bond has been formed or broken in a structurally significant domain of the molecule, highlighting it as the main reactive center of the transformation.

The algorithm produces two primary outputs, tailored for efficient use in simulations. First, it generates reactant and product templates that specify the atomic structures, bonds, angles, and other interactions for G_{reac} and G_{prod}. Second, the algorithm produces mapping, identifies the reaction sites, and accounts for any atoms that are created or deleted. To minimize template size, only the reaction site and its neighboring atoms, determined by a distance cutoff of four in the graph representation, are included. This cutoff based on nearest neighbors is chosen to capture all interactions affected by the reaction found in common molecular force fields, as these intramolecular interactions might alter within the reaction. The reduction is achieved by calculating the shortest path distances from the reaction site nodes and selecting atoms within the defined cutoff distance. Edges corresponding to these nodes are also identified, ensuring the inclusion of all relevant structural and interaction information.

For instance, in a representative example involving a double-linked epoxy species consisting of BFDGE and DETDA,[7,8] the full molecular graph includes 117 nodes and 123 edges. After reduction, the template comprises only 29 nodes and 28 edges - capturing all reactive and neighboring atoms while significantly reducing the size. In terms of performance, template generation for these larger molecules typically takes about 1 min on a standard desktop computer. Although this is slower than some atom mapping heuristics, the primary computational cost in our workflow lies in the molecular dynamics (MD) simulations that use these templates. As such, the one-time cost of generating reaction templates is justified by their downstream utility in enabling long-time, reactive MD simulations. Moreover, the template generation process scales reasonably with the size of the reactive fragment – which is not necessarily the full molecule – because only the specified reaction center and its local environment are considered. Even for larger biomolecules such as proteins or polymers, the effective graph sizes remain small as the user is expected to manually define the reacting units (e.g., two amino acids or two monomers).

The applicability of the algorithm was exemplified by investigating a selection of polymer classes and chemical reactions. Specifically, the algorithm was applied to the poly addition of butanediol (BD) and methylenediphenylisocyanate

Table 1. SMILES Strings for the Reactions under Study, Reflecting the Reactants (r) and the Product (p)[a]

reaction (FF)	MOL	SMILES
poly addition (OPLS-AA)	r_1 (MDI)	O=C=Nc2ccc(Cc1ccc(N=C=O)cc1)cc2
	r_2 (BD)	C(CCO)CO
	p_1	O=C=Nc2ccc(Cc1ccc(NC(=O)OCCCCO)cc1)cc2
poly condensation (GAFF)	r_1 (HT)	O=C(O)C1CC(C(=O)O)CC(C(=O)O)C1
	r_2 (MPD)	Nc1cccc(N)c1
	p_1	Nc2cccc(NC(=O)C1CC(C(=O)O)CC(C(=O)O)C1)c2
chain-polymerization (GAFF)	r_1	C=CCC
	r_2	C=CCC
	p_1	CCCCCCCC

[a]The LAMMPS data files, derived from these strings, were parameterized using the corresponding force field.

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

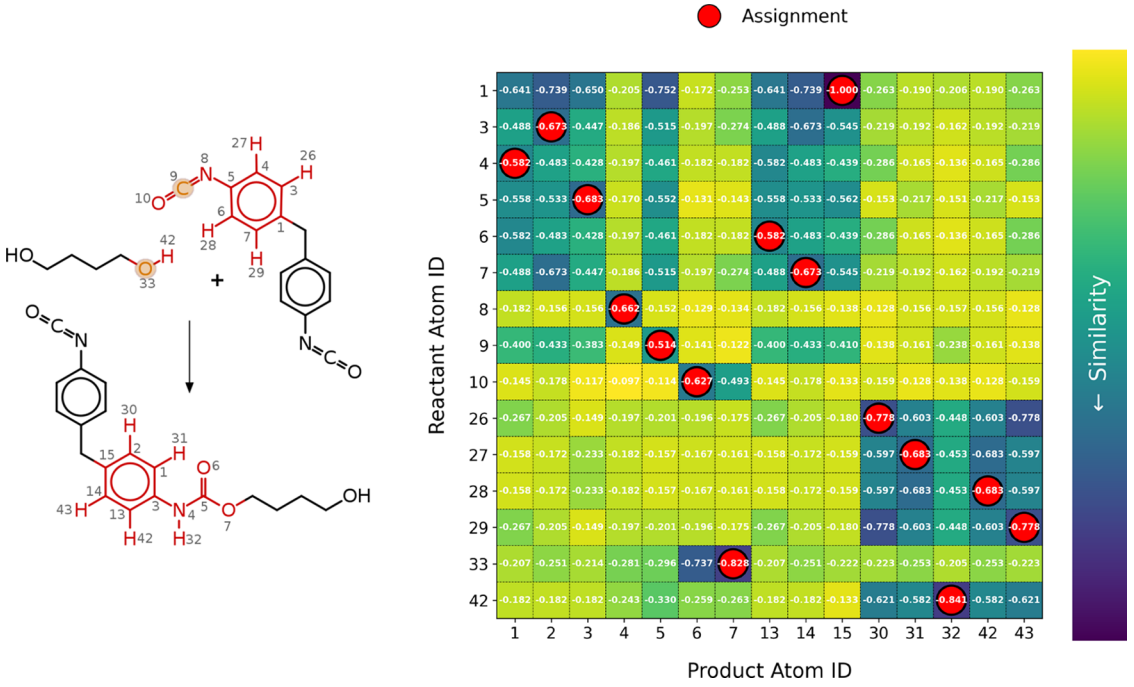


Figure 5. Mapping process for the poly addition of BD and MDI (left) is depicted. The chemical structures of the reactants and product distinguish between conserved domains (black) and the nodes that were mapped based on the similarity criterion (red). The atom labels correspond to the nodes of the similarity matrix on the right. Final assignments are shown in red.

(MDI), the poly condensation of cyclohexane-1,3,5-tricarboxylic acid (HT) and 1,3 phenylenediamine (MPD),[26] and a alchemical type of chain-polymerization between two 1-buthene molecules to octane. The molecular SMILES strings and force fields used are summarized in Table 1.

Based on these inputs, the initial LAMMPS data files were generated applying the desired force fields. For the OPLS-AA force field,[27] LigParGen[21] was employed, while GAFF (General Amber Force Field)[28] parameters were generated using OpenBabel[22] in combination with the Antechamber module of AmberTools.[29] To achieve accurate charges, a RESP fit was performed following the parametrization process.[30]

Subsequently, the algorithm was used to generate the reactant and product templates, along with the necessary mapping files for the simulations. A full simulation, including cross-linking to bulk polymer structure, was performed for the poly condensation between HT and MPD, using LAMMPS. 240 HT molecules and 320 MPD molecules, comprising in total 10,233 atoms, were equilibrated in a simulation box for 3 ns at room temperature, followed by heating to the cross-linking temperature of 600 K for 2 ns. Cross-linking was then performed for 5 ns using the REACTION package in LAMMPS,[10] with NVE/limit applied for 500 fs, restricting the displacement of reaction sites to 0.0015 Å per fs. During the simulation, the hydroxyl groups of the carboxylic acids were constrained using the SHAKE algorithm to enhance stability.[31] Constant temperature and pressure were maintained using the Nosé-Hoover thermostat and barostat with a damping time τ of 100 fs and a pressure damping time τ of 1000 fs at 1 atm.[32] Periodic boundary conditions were applied with a time step of 0.5 fs. Lennard-Jones interactions and electrostatics were calculated using a cutoff of 12 Å, while the long-range electrostatic interactions were treated using the particle−particle particle-mesh (PPPM) method with a *k*-space accuracy of 0.0004 kcal/mol ·Å.[33]

RESULTS & DISCUSSIONS

The weighting parameters were empirically determined to provide an optimal balance for capturing a diverse range of reaction types. The initial parameter values were set to $w_m = 0.5$, $w_t = 0.25$, and $w_n = 0.25$. It is worth noting that extreme parameter settings, such as $w_m = 0$, can lead to unintended consequences. For instance, without any contribution from the mass identity term, atom mapping may become unstable − potentially allowing atoms to be mapped across different elements. Even with equal weighting ($w_m = w_m = w_m = \frac{1}{3}$), the influence of element identity may be diluted − despite being the most chemically critical feature − potentially resulting in mappings that overlook fundamental chemical principles. Similarly, extreme values for the other weights (e.g., $w_t = 0$ or $w_n = 0$) may reduce the sensitivity of the similarity score to important structural or contextual features, potentially degrading the quality of the atom mapping. In general, while the default values offer a robust balance, caution should be exercised when deviating significantly from them, as essential information is contributed by each component to the mapping process.

The versatility of the algorithm is demonstrated through the poly addition reaction between butanediol (BD) and methylenediphenylisocyanate (MDI). In the initial step, subgraph isomorphism successfully identified conserved molecular domains, mapping 30 out of 45 nodes, as shown in black in Figure 5. The conserved domains are notably smaller than expected since the aromatic ring not being directly involved in the reaction. However, the force field's atom type definition distinguishes between isocyanate and urethane substitutions, leading to the assignment of the aromatic ring to the reactive domain. Consequently, the definition of conserved molecular domains depends on the provided input data and force field, emphasizing the algorithm's precision in isolating chemically invariant domains.

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

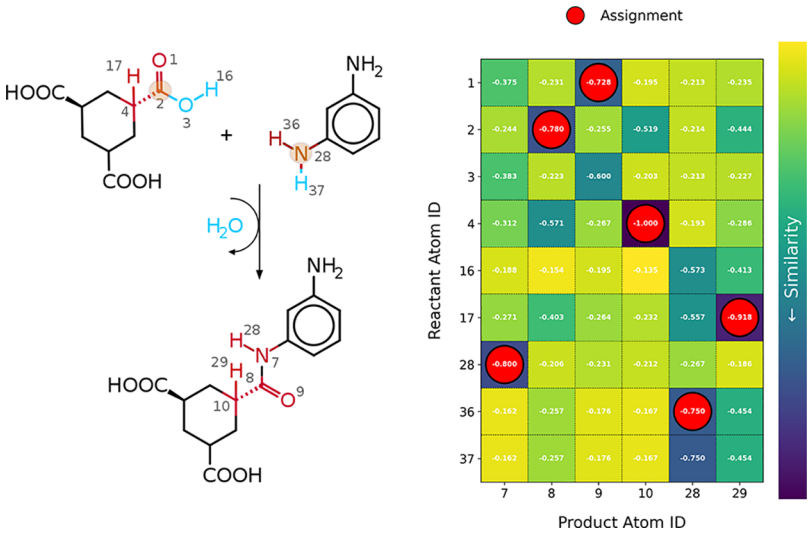


Figure 6. Representation of the poly condensation reaction between HT and MDP (left) shows conserved domains in black, mapped similar atoms in red, and atoms that are eliminated in blue. Automatic detection of the reaction-initiating atoms is highlighted in orange. The atom indices correspond to the node numbers of the graph representation of the molecules. The calculated similarity scores (right) for the nodes from G_{reac} and G_{prod}, along with the assignment via linear sum assignment, illustrate the convergence of the algorithm.

The subsequent alignment, based on similarity scores, accurately mapped nodes corresponding to heavy atoms (C, N, and O) directly involved in the reaction, as highlighted in red in Figure 5. The expected conserved domain of the aromatic ring was correctly assigned, demonstrating the robustness of the algorithm even with more complex force fields definitions and large reactive domains that seemingly contain conserved domains. Furthermore, the algorithm correctly matches indistinguishable paths in the graph structure. For instance, the para-phenyl substitution in the product allows two paths from atom ID 3 to 15, either through atoms 1 and 2 or through atoms 13 and 14. The algorithm also successfully identified the proton transfer from the hydroxyl group of BD to the urethane group of the product. Reaction-initiating atoms, including the reactive sites of the hydroxyl and isocyanate groups, were automatically detected and are highlighted in orange. To enhance the efficiency of subsequent simulations, the final `LAMMPS` templates were stripped to include only 31 atoms, representing the essential reactive domains and their immediate environment. These optimized templates are provided in the Supporting Information. This example highlights the algorithm's adaptability in accurately handling cases involving atom rearrangements into newly formed chemical groups and indistinguishable paths, ensuring precise node alignments and robust mapping in complex reaction systems.

In Figure 6, the reactants and products for the poly condensation reaction between HT and MDP are illustrated. Conserved molecular structures, identified through subgraph isomorphisms are highlighted in black with 34 of 40 nodes. The remaining atoms participating in the chemical reaction, are labeled with their respective indices from the graphs G_{reac} and G_{prod}.

The similarity scores for the unmapped nodes were calculated, and the algorithm achieved an optimal assignment, as highlighted in red in Figure 6, accurately mapping the acidic group to the newly formed amide bond. The water elimination process was identified and is represented by blue atoms. Reaction-initiating atoms were automatically detected and highlighted in orange. The final reaction template, consisting of 27 atoms, was derived by minimizing the structural representation to include only the reaction site and its immediate environment, reducing computational complexity while preserving chemical accuracy. The templates are provided in the Supporting Information. Figure 7 shows the corresponding reactant and product templates.

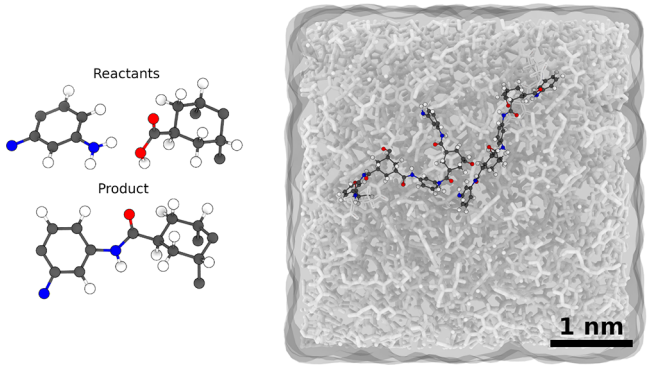


Figure 7. Templates of the reactants and the product are molecular fragments of the reacting molecules (left). Their size is minimized and accounts for intramolecular interactions that change during the reaction. The illustration depicts a cross-linked polymer structure (right) with an edge length of approximately 50 Å. The highlighted molecular strands represent the cross-linking.

To validate the algorithm-generated templates, a mixture of HT and MPD was prepared at an exact stoichiometric ratio, achieving an initial density of 1.26 g/cm^3. The cross-linking process achieved a degree of cross-linking (η) of 93%. The final polymer structure exhibited a density of 1.20 g/cm^3, close to the experimental value of 1.40 g/cm^3.[26] This deviation can be attributed to simulation constraints such as the limited box size (l_{box} = 48 Å) and the simplified reaction conditions. Nonetheless, the results confirm that the algorithm enables accurate modeling of polymer structures starting directly from SMILES strings, as depicted in Figure 7.

Importantly, the algorithm can also handle an alternative reaction description where water is explicitly included as an

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

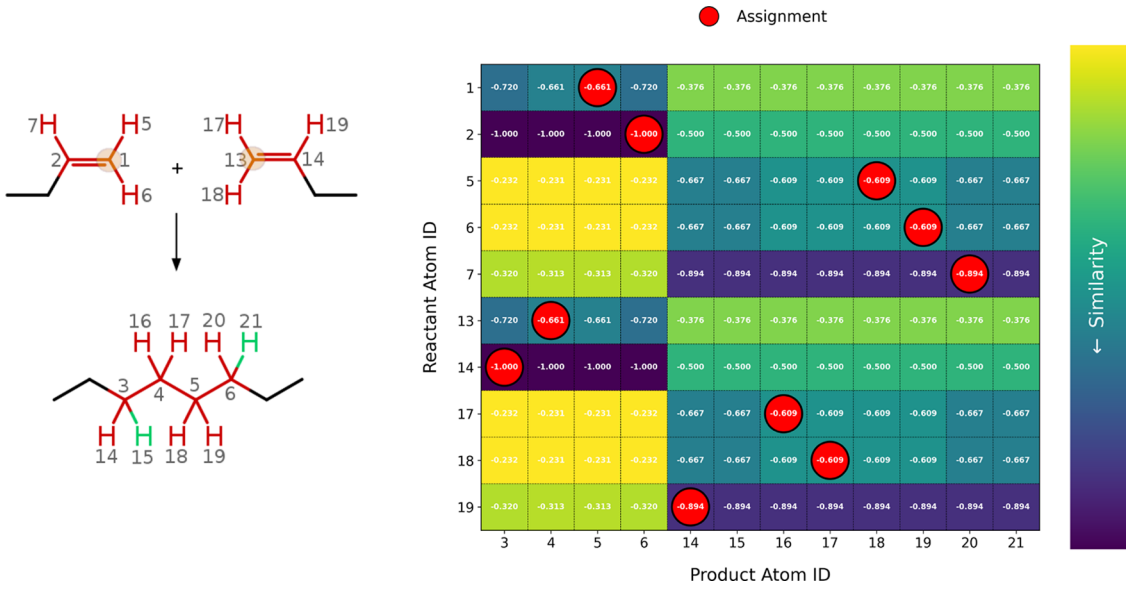


Figure 8. Left panel illustrates the chain polymerization reaction type, where the ethylene group (black) is identified as the only conserved domain. The similarity matrix reveals ambiguities for several atoms based on the calculated similarity scores. The correct atom mapping was achieved through path-dependent and neighborhood-based swapping, as shown by the atomic labels in the chemical structures. The green hydrogen atoms are introduced during the alchemical reduction process, and the reaction-initiating atoms are highlighted in orange.

additional product. This formulation provides a more explicit representation of condensation reactions, and the algorithm continues to successfully identify the reactive site while accurately mapping both the reactant and product molecules. This highlights the versatility of the algorithm in accommodating various reaction schemes, such as $A + B \rightarrow C$ and $A + B \rightarrow C + D$. For instance, in Figure 6 product D represents the water molecule removed in the reaction. The output data corresponding to this setup are provided in the Supporting Information.

The reduction of 1-butene to octane, as an example of an alchemical chain reaction, demonstrates the algorithm's ability to handle symmetrical reactants and complex atom mapping scenarios. This reaction involves the formation of a cross-link between the double bonds of two butene molecules, resulting in a saturated octane product. As shown in Figure 8, 14 out of the 26 total nodes were identified as conserved domains using subgraph isomorphism search. These conserved domains primarily correspond to the ethylene group, which remains unaltered during the reaction and is highlighted in black.

Using similarity scores, the algorithm initially mapped the nodes via linear sum assignment. However, ambiguities arose due to the indistinguishability of equivalent atoms between the two identical reactants (e.g., Node 2 and Node 14). This challenge was addressed through an iterative refinement process, where graph paths originating from isomorphic anchors were analyzed, allowing for the swapping of indistinguishable atoms and ensuring consistent mapping.

The mapping refinement also accounted for newly formed hydrogen atoms, which exhibited high similarity to existing hydrogen atoms. Neighborhood analysis proved crucial in resolving these ambiguities by evaluating structural and connectivity changes within the molecular graphs. Specifically, the assignment of atoms 5 and 6 to atoms 16 and 17, as well as atom 7 to either 14 or 15, could only be achieved through neighborhood analysis. This allowed for the clear identification of atoms 15 and 21 as those created during the reaction.

The final mapping is shown in Figure 8 (left), where atomic labels confirm the successful resolution of the potential misalignments. The reaction-initiating atoms highlighted in orange and the hydrogen additions in green. The bond rearrangements at the double bonds, were accurately identified. The final reaction template, comprising all 24 atoms, was derived as all atoms of the molecules are directly involved in the reaction and is available in the Supporting Information.

While in general the similarity scoring parameters (w_m, w_t, and w_n) can be used with the proposed default values, their adjustment may be necessary for reactions with fundamentally different chemistry or higher structural complexity. The observed variability in conserved domains also underscores the role of the force fields in accurately defining atom types and local environments. In reactions where specific atom types or bonding interactions are not well-defined by the chosen force field, the algorithm may face challenges in accurately identifying conserved domains and reaction sites. Consequently, the choice of force field becomes a crucial factor in determining the success of the algorithm with default parameters. Additionally, the current method may exhibit limitations in systems involving significant symmetry, conformational flexibility, or alchemical atom creation, where the alignment of molecular paths becomes more challenging.

CONCLUSIONS

In this work, a graph-based algorithm for the automated generation of reaction templates for the LAMMPS package REACTION was successfully developed and demonstrated. Leveraging graph-theoretical principles and subgraph isomorphism techniques, this method eliminates the labor-intensive manual processes traditionally required in reaction template preparation. By solely relying on LAMMPS input data, our approach ensures scalability, consistency, and adaptability across diverse reaction types.

The algorithm was validated on multiple case studies, including poly addition, poly condensation, and chain-polymerization reactions. These examples showcased the ability to accurately map conserved domains, identify reaction-initiating sites, and handle ambiguous molecular alignments through

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047

iterative refinement processes. Notably, the tool demonstrated high efficiency in reducing template size while preserving crucial molecular interactions, providing streamlined outputs for large-scale simulations.

While the presented framework focused on polymerization reactions, it has the potential to be extended to any organic reaction. Its flexibility is evident in the adjustable similarity scoring parameters, which can be tailored to capture the nuances of different chemical systems. Moreover, the algorithm's graph-based design inherently supports integration with machine learning pipelines for high-throughput material discovery and structural data generation.[34,35]

Additionally, the algorithm could be employed to bias unreacted species into reaction-prone configurations, effectively preconditioning systems for subsequent simulations with reactive force fields, such as the method described by Winetrout et al.[36] and applied in work of Odegard et al.[37] to epoxies. This hybrid approach may enhance sampling efficiency and reaction discovery in complex chemical systems, further expanding the utility of the framework. While the algorithm is primarily designed for mapping chemical reactions and their corresponding molecular structures, it could serve a similar purpose in exploring reaction pathways using metadynamics. However, it remains to be clarified how such templates could be translated into collective variables. One potential avenue involves leveraging the reaction templates to bias the system along the RMSD between reactants and products. This could be implemented using the Smooth Topology Transfer method proposed by Meißner et al.,[8] enabling enhanced sampling across reactive states and facilitating a form of thermodynamic integration.

Despite its strengths, certain limitations warrant further investigation. For instance, the accuracy of reaction site identification is influenced by the quality of the force fields employed, particularly regarding their ability to distinguish interactions between different atom types based on local atomic environments. Addressing this dependency through adaptive or transferable scoring mechanisms could enhance the algorithm's robustness. Additionally, improving its capability to handle highly symmetrical or complex molecules could broaden its application scope.

Looking ahead, this tool opens new opportunities for automating reaction modeling in computational chemistry and materials science. Future work may explore coupling this algorithm with quantum mechanical calculations or extending it to simulate reactive systems in mixed organic−inorganic environments. Furthermore, integrating this method with cloud-based platforms could facilitate collaborative, large-scale simulation workflows.

In conclusion, the algorithm represents a significant step toward automating and democratizing the preparation of reaction templates for molecular simulations. By bridging the gap between molecular mechanics and high-throughput computational workflows, it has the potential to accelerate innovation in fields ranging from polymer science to drug discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

All data and scripts supporting the findings of this study are openly available at the `Templater` repository via https://collaborating.tuhh.de/m-29/software/templater. This includes the full source code, data sets, and relevant documentation. The version of the data related to this study is available on the `published` branch.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00445.

Processed LAMMPS data files, output data, and final reaction templates for all examples discussed in the paper (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Julian Konrad − *Institute for Interface Physics and Engineering, Hamburg University of Technology, 21073 Hamburg, Germany;* orcid.org/0000-0002-8466-9291; Email: julian.konrad@tuhh.de

Robert Meißner − *Institute for Interface Physics and Engineering, Hamburg University of Technology, 21073 Hamburg, Germany; Institute of Surface Science, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany;* orcid.org/0000-0003-1926-114X; Email: robert.meissner@hereon.de

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00445

Author Contributions

J.K.: Conceptualization, methodology, investigation, data analysis, writing—original draft and final manuscript. R.M.: Supervision, writing—review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) − Project number 525597740.

■ REFERENCES

(1) Leach, A. R. *Molecular Modelling - Principles and Applications*; Pearson Education: Amsterdam, 2001.

(2) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons: New York, 2016.

(3) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(4) Rappe, A. K.; Goddard, W. A. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(5) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.

(6) Chenoweth, K.; van Duin, A. C. T.; Goddard, W. A. ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation. *J. Phys. Chem. A* **2008**, *112*, 1040−1053.

(7) Konrad, J.; Meißner, R. H.; Bitzek, E.; Zahn, D. A Molecular Simulation Approach to Bond Reorganization in Epoxy Resins: From Curing to Deformation and Fracture. *ACS Polymers Au* **2021**, *1*, 165−174.

(8) Meißner, R. H.; Konrad, J.; Boll, B.; Fiedler, B.; Zahn, D. Molecular Simulation of Thermosetting Polymer Hardening: Reactive Events Enabled by Controlled Topology Transfer. *Macromolecules* **2020**, *53*, 9698−9705.

(9) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, No. 108171.

(10) Gissinger, J. R.; Jensen, B. D.; Wise, K. E. Modeling chemical reactions in classical molecular dynamics simulations. *Polymer* **2017**, *128*, 211−217.
(11) Gissinger, J. R.; Jensen, B. D.; Wise, K. E. REACTER: A Heuristic Method for Reactive Molecular Dynamics. *Macromolecules* **2020**, *53*, 9953−9961.
(12) Gissinger, J. R.; Jensen, B. D.; Wise, K. E. Molecular modeling of reactive systems with REACTER. *Comput. Phys. Commun.* **2024**, *304*, No. 109287.
(13) Bone, M. A.; Howlin, B. J.; Hamerton, I.; Macquart, T. AutoMapper: A python tool for accelerating the polymer bonding workflow in LAMMPS. *Comput. Mater. Sci.* **2022**, *205*, No. 111204.
(14) Peng, Y.; Zhang, C.; Wu, M.; Bu, G.; Fan, K.; Chen, X.; Liang, L.; Zhang, L. PX-MDsim: a rapid and efficient platform for large-scale construction of polyamide membranes via automated molecular dynamics simulations. *RSC Adv.* **2025**, *15*, 5906−5915.
(15) Provenzano, M.; Bellussi, F. M.; Fasano, M.; Chávez Thielemann, H. Atomistic Modeling of Cross-Linking in Epoxy-Amine Resins: An Open-Source Protocol. *ACS Appl. Polym. Mater.* **2025**, *7*, 4876−4884.
(16) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **2021**, *7*, No. eabe4166.
(17) Chen, S.; An, S.; Babazade, R.; Jung, Y. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nat. Commun.* **2024**, *15*, 2250.
(18) Wen, M.; Blau, S. M.; Spotte-Smith, E. W. C.; Dwaraknath, S.; Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chemical Science* **2021**, *12*, 1858−1868.
(19) Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chemical Science* **2022**, *13*, 1446−1458.
(20) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(21) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331−W336.
(22) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(23) Jewett, A. I.; Stelter, D.; Lambert, J.; Saladi, S. M.; Roscioni, O. M.; Ricci, M.; Autin, L.; Maritan, M.; Bashusqeh, S. M.; Keyes, T.; Dame, R. T.; Shea, J.-E.; Jensen, G. J.; Goodsell, D. S. Moltemplate: A Tool for Coarse-Grained Modeling of Complex Biological Matter and Soft Condensed Matter Physics. *J. Mol. Biol.* **2021**, *433*, No. 166841.
(24) Cordella, L.; Foggia, P.; Sansone, C.; Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2004**, *26*, 1367−1372.
(25) Bonacich, P.; Lloyd, P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks* **2001**, *23*, 191−201.
(26) Arthur, S. D. Structure-property relationship in a thin film composite reverse osmosis membrane. *J. Membr. Sci.* **1989**, *46*, 243−260.
(27) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(28) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(29) Case, D. A.; et al. AmberTools. *J. Chem. Inf. Model.* **2023**, *63*, 6183−6191.
(30) Woods, R.; Chappelle, R. Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *Journal of Molecular Structure: THEOCHEM* **2000**, *527*, 149−156.
(31) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(32) Evans, D. J.; Holian, B. L. The Nose−Hoover thermostat. *J. Chem. Phys.* **1985**, *83*, 4069−4074.
(33) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles*; Hilger: Bristol, 1988.
(34) Baudis, S.; Behl, M. High-Throughput and Combinatorial Approaches for the Development of Multifunctional Polymers. *Macromol. Rapid Commun.* **2022**, *43*, No. e2100400.
(35) Hayashi, Y.; Shiomi, J.; Morikawa, J.; Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput. Mater.* **2022**, *8*, 222.
(36) Winetrout, J. J.; Kanhaiya, K.; Kemppainen, J.; in 't Veld, P. J.; Sachdeva, G.; Pandey, R.; Damirchi, B.; van Duin, A.; Odegard, G. M.; Heinz, H. Implementing reactivity in molecular dynamics simulations with harmonic force fields. *Nat. Commun.* **2024**, *15*, 7945.
(37) Odegard, G. M.; Patil, S. U.; Deshpande, P. P.; Kanhaiya, K.; Winetrout, J. J.; Heinz, H.; Shah, S. P.; Maiaru, M. Molecular Dynamics Modeling of Epoxy Resins Using the Reactive Interface Force Field. *Macromolecules* **2021**, *54*, 9815−9824.

https://doi.org/10.1021/acs.jcim.5c00445
J. Chem. Inf. Model. 2025, 65, 6038−6047