# Reaction Design from Artificial Intelligence Guided Potential Energy Surface Exploration: Suzuki Coupling from Theory to Experiment

*Zi-Xing Guo[1], Jin-Peng Tang[1], Zhen-Xiong Wang[1], Qi-Ming Liang[1], Si-Cong Ma[2], Cheng Shang[1],*

*Lin Chen[1]*, Zhi-Pan Liu[1,2]**

[1]Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Science, Department of Chemistry, Fudan University, Shanghai 200433, China

[2]Key Laboratory of Synthetic and Self-Assembly Chemistry for Organic Functional Molecules, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

## Abstract

Reaction design from first principles has been a grand challenge in chemistry. Here we develop a theoretical framework to explore uncharted chemical reaction space, which relies on a general diffusion generative model (DGM) to generate three-dimension structure of reaction intermediates and transition states, and global neural network potential calculations for fast potential energy surface optimization. A high-order pair-reduced equivalent massage passing neural network model is developed to meet the general (elements ≤ 86) and high precision (<0.05 Å) purpose of DGM, which can distinguish sensitively elements, bonds and the spatial arrangement of atoms. By bootstrapped training over 437,952 molecule datasets, including drug-like molecules and metal-containing complexes (especially Pd-phosphine), our DGM can now be utilized for molecule structure generation in general (accessible from www.laspai.com). Taking a challenging Suzuki-Miyaura coupling reaction with a chunky reactant as the target, we generate the complete reaction profile of the catalytic reaction for 81 different Pd-P catalysts. By using the reaction kinetics data as descriptors, we screen the high activity ligands and perform synthetic experiments to verify theoretical prediction, which leads to the identification of a new ligand that reaches 81.1 % reaction yield. This work establishes a new paradigm of reaction design by a fast, high precision reaction pathway generation from 2D graph.

## 1. Introduction

It has been a long history for human beings in developing understanding of chemical reactions and further the ability to identify reaction rules.[1–3] To design reaction, a major challenge is to establish a quantitative structure-activity relationship, which requires an efficient exploration of the vast reaction space containing reactants and catalysts, if present.[4–7] In transition-metal-catalyzed organic reactions, the cornerstone of modern chemical synthesis, this difficulty is manifested in the rich variables relating to auxiliary ligands, metal species and reactant types, which are prohibitively expensive to enumerate *via* both experiment and first principles calculations.[8–11] Instead, reaction design via descriptor engineering has been pursued in recent years[12–15] and many descriptors[16,17], such as buried volume[18–20], bite angle[21–23], Tolman electronic parameter[24] and ligand replacement energy[25] were proposed to envelop the parameter space of reaction into simply accessible quantities. These designed descriptors, while being informative, are often reaction-dependent and lack of clear physical interpretation in the context of reaction kinetics. For a general-purpose reaction design, better approaches are desirable that can be applied to different reaction data with good model transferability, and ideally, be rooted rigorously in kinetics theory.

In order to establish a robust structure-activity relationship, it is essential to equip with high-quality rection data. Compared to traditional approaches to accumulate reaction data via either experimental synthesis or first principles calculations, deep generative models emerged in recent years offered a fast and cost-effective route.[26–31] By learning available three-dimensional (3D) structure databases with known atom composition and bond connectivity, these models as represented by GeoDiff based on Diffusion-based stochastic denoising framework, more generally, diffusion generative model (DGM), can achieve fast 3D structure generation[32–36] from two-dimensional (2D) molecular graph, outperforming traditional methods that rely heavily on empirical parametrization and semi-empirical approximations.[37–39] Specifically, GeoDiff algorithm can achieve the mean root-mean-square-error (RMSE) of structure with 0.863 Å and 0.209 Å on GEOM-DRUG and GEOM-QM9 datasets, which primarily focus on conformation of small to medium-sized organic molecules (typically below 40 atoms) with only C, H, O, and N elements. However, due to the shortage of 3D structures from literature, there is no DGM available for complex chemical systems, particularly those with transition-metal elements. For transition metal-catalyzed organic reactions, the current theoretical datasets, such as CATCO group's curated database of >400 transition metal pathways (spanning 12 metals from Ti to Hg)[26,40] and Gensch *et al.* compiled dataset containing 1,558 monodentate phosphine structures and their $Ni(CO)_3$-bound forms[15], are either limited in the ligand diversity (*e.g.*, type and conformation) or in the metal diversity, which is mainly due to the high computational costs of first principles calculations to explore the huge reaction space.

Even if the reactant/product is available, common diffusion-based stochastic denoising framework still meets difficulty in predicting the transition state (TS) of complex reactions. This is because common DGMs typically utilize a sophisticated dual-encoder architecture[32,34]: a SchNet as the spatial encoder for unconditional geometric modeling and a Graph Isomorphism Network (GIN)[41] message passing for processing graph-conditioned structural data, where the TS information is not present explicitly. Beyond the GeoDiff model, the OA-ReactDiff model utilizes a SE(3) equivariant neural network (LEFTNet) on fully connected graph containing initial state (IS), TS and final state (FS) structures during vector space construction, learning the Transition1x database, which contains 11,961 gas-phase organic reactions (limited to H, C, O, and N elements with maximum 7 heavy atoms)[42], and achieves a mean RMSE of 0.183 Å for generating the TS structure for 1,073 test elementary

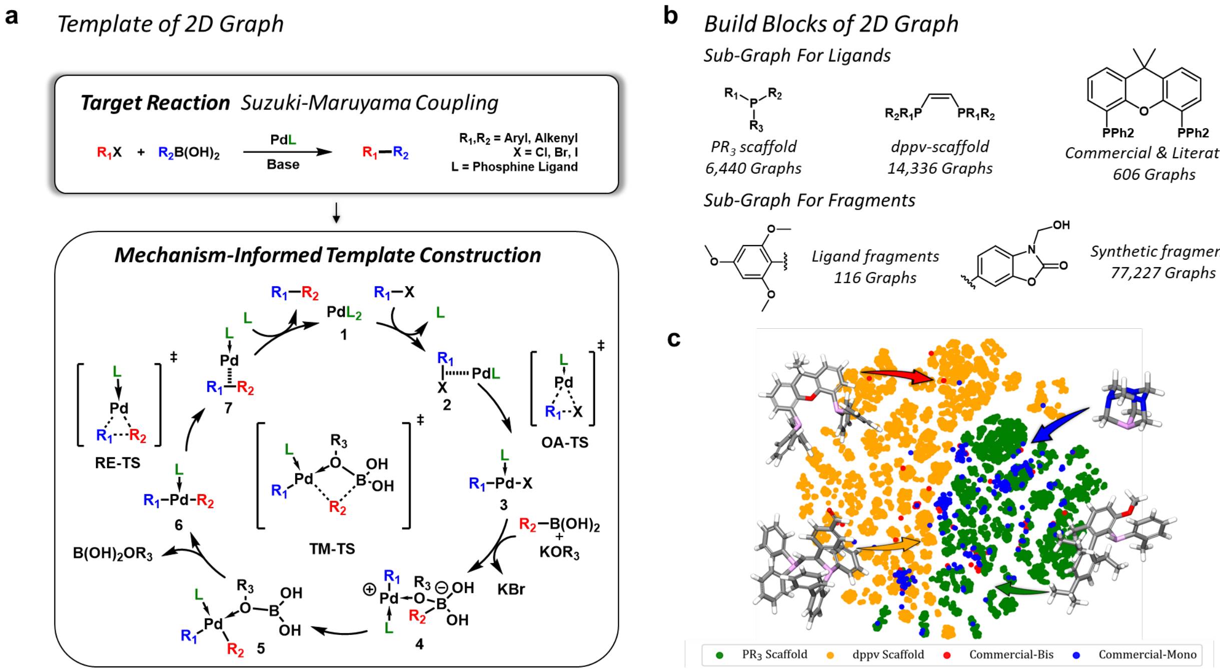

**Figure 1 Construction Pd-P database from 2D graphs of Suzuki-Miyaura coupling Reaction.** (**a**) Template-driven exploration of chemical space using 2D graph representations of Suzuki-Miyaura cross-coupling reactions. (**b**) Type of molecular building blocks, including molecules and fragments, represented by 2D graphs. **(c)** Chemical feature visualization *via* t-SNE projection of ligand molecular Morgan fingerprints.

reactions. This framework, however, encounters scalability bottlenecks when handling large systems (*e.g.*, greater than 100 atoms) due to the O($n^2$) computational complexity of full connected graphs.

Here we develop a BDGM-PES method, acronym for Bootstrapping Diffusion Generative Model coupled with Potential Energy Surface (PES) exploration, which is a computational framework for exploring chemical spaces of reaction through an automated self-regressive pipeline. BDGM-PES utilizes a DGM to generate 3D structures of reaction intermediates and TSs directly from their 2D molecular graphs. This capability leverages a high-order pair-reduced neural network incorporating edge and time information (HPNN-ET), which facilitates the generation of large molecules (>100 atoms) with atomic-level precision (RMSE < 0.05 Å) at remarkable efficiency. Furthermore, BDGM-PES incorporates our group's global neural network potential (G-NN) and a single-ended TS searching algorithm to fast explore the PES[43,44]. Taking Pd-catalyzed Suzuki-Miyaura reaction[45,46] as the target system, we systematically generated 3.7 million conformations of 96,000+ organic molecules (*avg.* 96.6 atoms) across Pd/P/C/H/O/N/Cl/Br/B/K ten elements using BDGM-PES method. Using the trained DGM model, we obtained complete pathways for 81 different ligands, from which the high catalytic activity of ligands is screened out using the descriptors directly from the reaction profile and then confirmed by experiment. This work establishes a paradigm for automated large-scale structural exploration, pathway construction, and activity prediction in transition metal catalysis.

## 2. PdP5 dataset

Before we introduce BDGM-PES method, we summarize the PdP5 dataset created in this work, as summarized in Figure 1, which comprises 3D structures of 96,793 organopalladium complexes derived from five fundamental coordination scaffolds, L, PdL, PdLRX, PdL$R_1R_2$, and PdLRBorate (detailed in Supplementary Table S1). The dataset construction involves rigorous sampling of 3.7 million distinct 3D molecular conformations. From the PdP5 dataset, we randomly selected a small Pd-2000 dataset for benchmarking purpose.

The 3D structures of PdP5 dataset are all generated from 2D molecule graphs using our DGM method that allows fast template-based substitution along the best-regarded reaction pathway of the classic organometallic Suzuki-Miyaura reaction, as shown in Figure 1a. It proceeds through three elementray steps in the coupling reaction[47–49], namely, oxidative addition (OA), transmetallation (TM), and reductive elimination (RE). Our 2D library thus includes 14 distinct graphs: intermediates **1**-**7**, reactants ($R_1X$, $R_2B(OH)_2$), product ($R_1R_2$), ligand (L), and TSs (OA-TS, TM-TS, RE-TS), among which the five templates L, PdL, PdLRX (**3**), PdLBorate (**5**), and PdL$R_1R_2$ (**6**) are the representatives defining the chemical space. As shown in Figure 1b, the ligand types considered include both template-derived ligands ($PR_3$ and dppv) and commercial ligands:  6,987 mono-phosphine ligands (435 commercial and 6,552 $PR_3$ scaffold) and  9,690 bis-phosphine ligands (170 commercial / 9,520 dppv scaffold). They can be better visualized through t-distributed stochastic neighbor embedding (t-SNE) clustering algorithm in Figure 1c: template-derived ligands form many densely packed clusters (orange/green regions), spaning over a continous chemical space, while commercial ligands disperse in the space (red/blue points), reflecting heterogeneous nature of the commercial ligands. The methodology for assembling fragments into scaffolds is illustrated in detail in Supplementary Figure S1.

## 3. The BDGM-PES method and Pd-complex generator

Our BDGM-PES aims to solve two fundamental challenges in reaction design: (1) structure data accumulation of large molecular systems; 2) uncharted chemical space exploration. The former is *via* a universal generative model for both minima

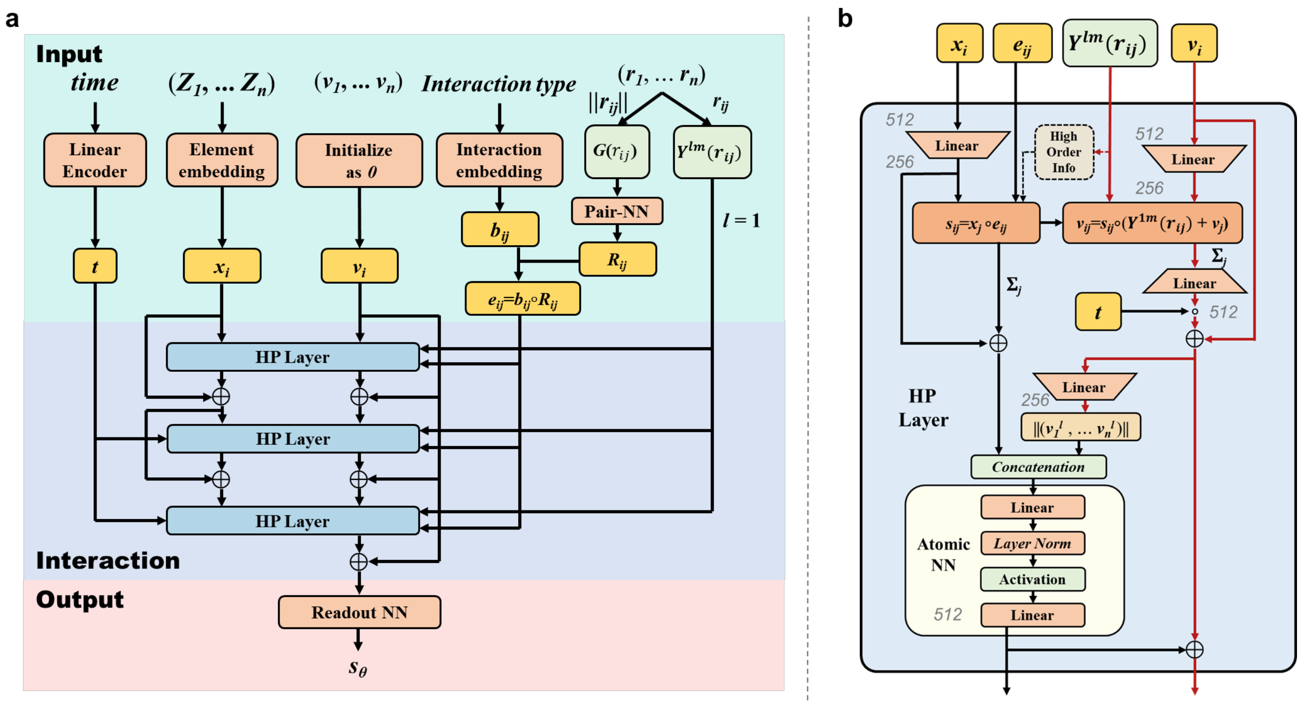

**Figure 2. The architecture of HPNN-ET neural network.** (a). The architecture of HPNN-ET model. (b). The details of HP layer.

and TS structure generation and the latter utilizes an active learning workflow facilitated by G-NN potential. In this work, we show that BDGM-PES can generate automatically reaction pathways involving complex transition metal-ligand interaction and non-covalent interactions (such as metal-$\pi$ interaction and ions in catalytic systems) for the first time, thereby enabling reaction design of organometallic catalysis systems.

### 3.1 HPNN-ET architecture of DGM

As the core of BDGM-PES method, an SE(3)-equivariant machine-learning model, namely high-order pair-reduced neural network with edge and time information (HPNN-ET), is developed for 3D structure generation. Compared with previous DGM, our HPNN-ET has two key features: (1) a unified information fusion strategy in gathering 2D and 3D information. (2) an enhanced equivariant architecture for spatial discrimination between structures. More details on the diffusion process algorithm used in HPNN-ET training and sampling can be found in Methods. Here we briefly introduce the architecture and highlight the major advances.

Figure 2a illustrates the HPNN-ET architecture, which unifies and updates all molecular information in one model. It incorporates three connected modules: input processing, interaction *via* message passing, and output generation. The input module systematically integrates three heterogeneous data streams through embedding layers: atomic-specific scalar $x_i$ and vector $v_i$ features on atomic node $i$; pairwise interactions, $v_{ij}$, $b_{ij}$, $R_{ij}$, $e_{ij}$, between node $i$ and $j$; temporal diffusion dynamics ($t$), where $x_i$ and bond-type indicator $b_{ij}$ initiate from embedding layers using atomic number and bond type (*e.g.*, single bond) information, respectively; while $v_i$ vectors are initialized as zero vectors; the radial basis function transforms interatomic distances into continuous embeddings $R_{ij}$; the conditional and unconditional information are fused into $e_{ij} = b_{ij} \circ R_{ij}$. The temporal diffusion component $t \in [0,1]$ indicate the process of diffusion. All inputs undergo progressive update through three cascaded interaction blocks that accomplish the information fusion. The architecture ends with an output layer that generates atom-specific vector $v_i$ that further form the score function $s_\theta \in \mathbb{R}^{N_{atom}\times 3}$ for the generative process.

The interaction block constitutes three high-order pair-reduced (HP) layers, performing message passing to update the node feature $x_i$ and vector $v_i$ simultaneously, as detailed in Figure 2b. The message passing is a pairwise operation, thus computationally demanding, to collect the geometrical information of atomic pairs using the interaction information $e_{ij}$ and the spherical function $Y^{lm}(\overrightarrow{\boldsymbol{r}_{ij}})$. To enhance the geometry discrimination between structures, we utilize high-order spheric function $Y^{lm}(\boldsymbol{r}_{ij})$ with $l_{max}$ up to 6. To enhance the performance of message passing, we reduce the dimension of atomic features $x_i$ and vector parameters $v_i$ with a shrinking linear layer upon entering interaction block, but restore the dimension with an expanding linear layer once the pair operation is finished. By this way, the parameter space size of HPNN will not be compromised at the expense of the demanding pair operation since the subsequent atomic NN remains at the full dimension $x_i$. In practice, the pair dimension (PD) can shrink to half or even quarter.

Table 1 shows HPNN-ET-based DGM performance for creating 3D structure of Pd-coordinated complexes (maximum 233 atoms, average 120.4 atoms) together those of popular algorithms, including RDKit *via* traditional cheminformatics approaches[50], local environment prediction models (GeoMol)[27], and diffusion-based DGMs using Cartesian coordinates (GeoDiff)[32] or torsional angles (TorsionalDiff)[36]. All models are trained on the same Pd-2000 dataset. For HPNN-ET-based DGM, results from three architectural variants are listed, which differ in the maximum order of spherical harmonic (A: $l_{max}$=1, B: $l_{max}$=1, C: $l_{max}$=6) and the shrink of PD (B: quartic reduction, A/C: half reduction).

As shown in the table, our HPNN-ET model achieves the best performance across all evaluation metrics. HPNN-ET-A achieves

**Table 1 Comparison of 3-D structure generation on Pd-phosphine complexes datasets from current models.**

| Model | Description | Param | Precision[a] (Å) | | Matching Score [b] | Efficiency[c] | Extrapolation[d] |
|---|---|---|---|---|---|---|---|
| | | Num | mean | median | (%) | (s) | (%) |
| RDKit | Empirical | None | 0.27 | 0.24 | 98.7 | 0.68 | 80 |
| GeoMol | Local Environment | 75K | 0.35 | 0.38 | 2.2 | 10.14 | × |
| GeoDiff | EGNN, $l_{max}$ = 1 | 0.8M | 0.85 | 0.84 | 91.4 | 4.38 | 0.0 |
| TorsionalDiff | Torsional Space | 1.0M | 0.55 | 0.52 | 31.2 | 1.09 | 0.0 |
| HPNN-ET-A[e] | $l_{max}$ = 1, PD = 256 | 5.4M | 0.08 | **0.04** | **100** | 0.60 | **100** |
| HPNN-ET-B[e] | $l_{max}$ = 1, PD = 128 | 3.8M | 0.10 | 0.05 | **100** | **0.35** | **100** |
| HPNN-ET-C[e] | $l_{max}$ = 6, PD = 256 | 7.55M | **0.07** | **0.04** | **100** | 0.94 | **100** |
| HPNN-ET-General[f] | $l_{max}$ = 6, PD = 256 | 7.55M | 0.16 | 0.07 | **100** | 0.94 | **100** |

[a] Precision is measured by the RMSE of atom coordinate between generated molecular structures and their fully relaxed counterparts.

[b] Matching Score is measured by the topology matching ratio, which quantifies the ratio of 3D molecular structures that maintain topological isomorphism with their corresponding 2D molecular graphs.

[c] The benchmark was conducted on an RTX 4090 GPU, except GeoMol's results from its CPU-based model, on EPYC 9474F CPUs (2 sockets,96 cores).

[d] Extrapolation topology correctness, similar to the Matching Score, is tested on a Pd-coordinate complex containing 362 atoms that are not utilized in training dataset (Supplementary Figure S2).

[e] Different HPNN-ET models. The pair-dimension (PD) is set to 256, 128, 256 for A, B, C respectively. A, B utilize only L=1 in spheric function, while C take advantage of higher order of spheric with L up to 6.

[f] The general-purpose (element≤86) HPNN-ET model trained on 467,757 molecules.

high geometry precision (median RMSE of 0.04 Å) while maintaining high accuracy (described by topology matching score, 100%) and generation efficiency (0.60 s per structure). Notably, HPNN-ET-B, the variant with the PD reduced to a quarter, achieves the highest time efficiency (0.35 s per structure), outperforming even the empirical RDKit method (0.68 s) while still maintaining a high geometric precision (0.05 Å). A detailed benchmark for the trade-off on PD is shown in Supplementary Table S2. HPNN-ET-C achieves the highest precision (0.04 Å) thanks to the incorporation of higher-order spherical harmonics ($l_{max}$=6), but its generation time (0.94 s) increases by 3-fold compared to HPNN-ET-B, suggesting the use of $l_{max}$=6 can be applied to the scenarios that very high geometry precision is required. RDKit demonstrates the second-best precision 0.24 Å with 98.7% matching score. For the other three ML-models, TorsionalDiff achieves moderate precision (0.52 Å) but suffers from low matching (31.2%), suggesting that torsional space parameterization may not adequately constrain metal coordination geometries; GeoDiff prioritizes structure topology (91.4%) at the expense of precision (0.84 Å). The performance of GeoMol is poor with a low matching 2.17%.

Furthermore, we note that HPNN-ET model can treat well large molecular systems beyond the training set, the extrapolation ability, as listed in Table 1. All HPNN-ET variants achieve the perfect topology matching (100%) for large Pd-coordinated complexes (362 atoms, see Supplementary Figure S2), and have high computational efficiency as exemplified by HPNN-ET-B that accomplishes the task in only 1.2 seconds. In contrast, other ML models exhibit catastrophic failure in this challenging scenario, showing a complete prediction breakdown (0.0% correctness); and RDKit demonstrates relatively better performance with 80% correctness, although being still much worse than HPNN-ET models.

We have also performed ablation analysis on HPNN-ET to evaluate the impact of the unified information fusion strategy. It turns out that the strategy strongly enhances the precision (0.07 Å to 0.04 Å), while also improve the generation speed by 8% compared to a separated processing of unconditional and conditional information (Supplementary Table S3). The implementation of temporal encoding (*t*) and pairwise interaction encoding ($e_{ij}$) improve precision by 43% and 57%, respectively; the inclusion of $b_{ij}$ dramatically improves the topology matching from 6.7% to 100% (Supplementary Table S4).

The high-performance architecture of HPNN-ET allows us to generate a general-purpose DGM for molecules. For this purpose, we curated a dataset of 437,952 molecules comprising 94,697 Pd-phosphine ligands from this work, 273,501 drug-like molecules from GEOM-DRUG[51], and 69,754 structures containing all elements ≤ 86 from the PubChem database (detailed in Supplementary Information B) and trained the general DGM. From this dataset, we randomly selected 1,000 molecules covering all 86 elements for testing. The model achieved 99.5% matching scores not only on this representative test set but also 100% on the Pd-2000 benchmark. Moreover, it attained a median RMSE of 0.07 Å for precision on Pd-2000 structures (exemplified in Supplementary Figure S3). This general-purpose DGM model is openly accessible at website www.laspai.com.

Now we turn to solve the TS structure generation issue. In this work, we develop a dual-phase constraint-guided generation algorithm for TS generation, exploiting the transferability of machine learning model from minima to TS. This is inspired by Hammond's postulate[52], which posits structural similarity between TS and the associated minima (IS or FS). In dual-phase constraint-guided generation algorithm, we can first generate the TS-like 3D structure, the guessed TS structures, using the DGM trained on minima geometries. The molecular graph of TS is directly utilized as the input, along with an additional constraint biasing the molecule towards TS during sampling. Details for constrained generation of TSs can be found in Methods. This enables a fast TS-like structure generation without requiring TS structure data *a priori* as training set, achieving the *zero-shot* TS generation. The guessed structure is further optimized towards TS through the Constrained Broyden Dimer (CBD) algorithm, a single-ended TS searching algorithm developed by our group.[43,53] In the second phase, we search for TS exactly based on these guessed TS structures, and by incorporating the validated TS structures into training set, we retrain the DGM to directly generating the TS without constraint. This dual-phase constraint-guided generation strategy achieves both efficiency and chemical accuracy in TS dataset

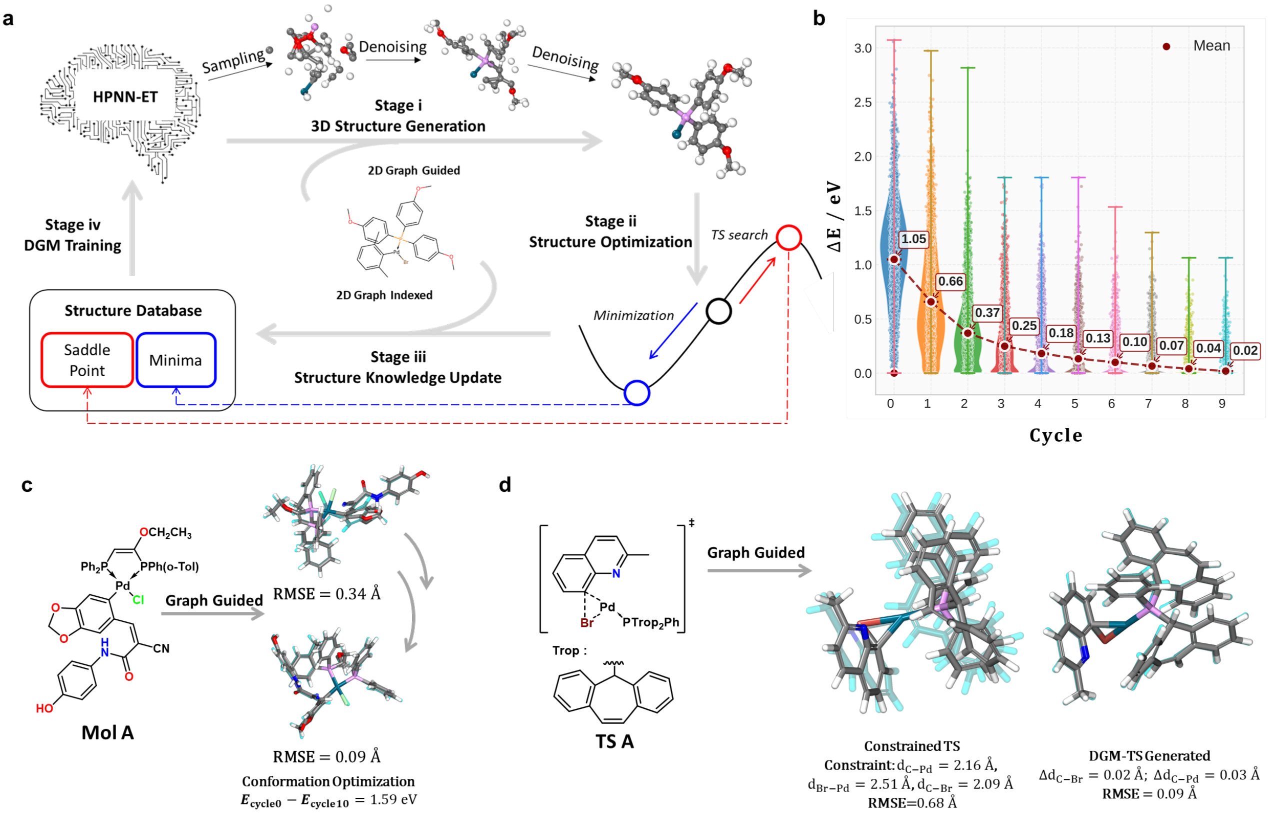

**Figure 3. Active learning workflow of the BDGM-PES framework. a.** Schematic architecture of the BDGM-PES framework. **b.** evolution of ΔE for 10,694 molecules in PdLRX dataset along the active learning cycles. The x-axis represents the iteration cycle, where "cycle 0" corresponds to the initial structures before training, and subsequent cycles represent the updated structures through of the BDGM-PES process. The y-axis indicates the relative energy (ΔE, in unit eV), which evaluates the instability of the most stable conformer in each cycle compared to that in the final (10th) cycle. **c.** Structural refinement of Mol A through successive active learning cycles. **d.** Transition state generation for TS A (left) by distance constraint (middle) and DGM-TS (right). The generated result is in solid color, while that optimized after CBD algorithm is represented by blue shading.

construction. Compared to existing approaches employing disparate architectural frameworks for minima generation (*e.g.*, utilizing dual embedding for IS and FS graphs)[26,31,34], our methodology unifies TS and minima generation within a single framework, achieving a powerful generalized generative model addressing both minima and TS generation.

### 3.2 BDGM-PES framework

Figure 3a illustrates the workflow of BDGM-PES framework, which implements a closed-loop architecture for autonomous chemical space exploration. It contains four stages in cycle: (i) 3D structure generation guided by 2D molecule graph, (ii) structure optimization on PES, (iii) structure knowledge update, and (iv) DGM training.

Stage i generates moleculear 3D structure that satisfies the molecular topology in 2D graph. The structure generator is RDKit initially and then switches to DGM in the subsequent cycles. This stage establishes the bidirectional mapping between 2D molecular graphs and their 3D conformers, ensuring structural consistency for subsequent structure optimization (stage ii) and DGM training (stage iv).

Stage ii utilizes the global NN (G-NN) potential for PES optimization of DGM-generated 3D structures, which serves to identify true minimum or TS structure. Our previous work has shown that G-NN calculation is several orders of magnitude faster than DFT while maintaining the high accuacy in energy and structure. Supplementary Table S5 and S6 benchmark the structure accuacy of G-NN optimized structure compared to DFT, which shows the median RMSE is below 0.02 Å forminima.

Stage iii implements an adaptive knowledge cumulation through integrating novel conformers that are predicted poorly by Stage i as compared to the optimized geometry at Stage ii. Stage iii executces repeatedly Stage i generation and Stage ii optimization, archiving new conformations systematically and thus evolving the chemical knowledge base.

Stage iv establishes an energy-driven generative training paradigm. The thermodynamically most stable 3D conformation is selected for each 2D molecular graph as the canonical structural label to update the DGM. With the retrained DGM, one can repeat Stage i-to-iv to expand the knowledge base and obtain a better DGM.

The iterative learning process of BDGM-PES can systematically improve the quality of generated structures towards uncharted chemical space, driving them closer to the most stable minima on PES. In general, for metal-containing complexes, the initial 3D geometry from RDKit deviates largely from the optimized structure on PES (RMSE = 0.34 Å), and notably, the Pd complex adopts an energetically unfavorable tetrahedral configuration (ΔE = +1.59 eV) instead of the most stable square-planar geometry. RDKit-generated structures also miss some important geometry motifs, such as Pd-π interactions in the "B-ring" of Buchwald ligands[54–57] (exemplified in Supplementary Figure S4). Figure 3b illustrates the relative energy (ΔE) evolution of 10,694 palladium-containing PdLRX scaffold molecules over ten iterative cycles of BDGM-PES. ΔE quantifies the molecular instability by calculating the energy

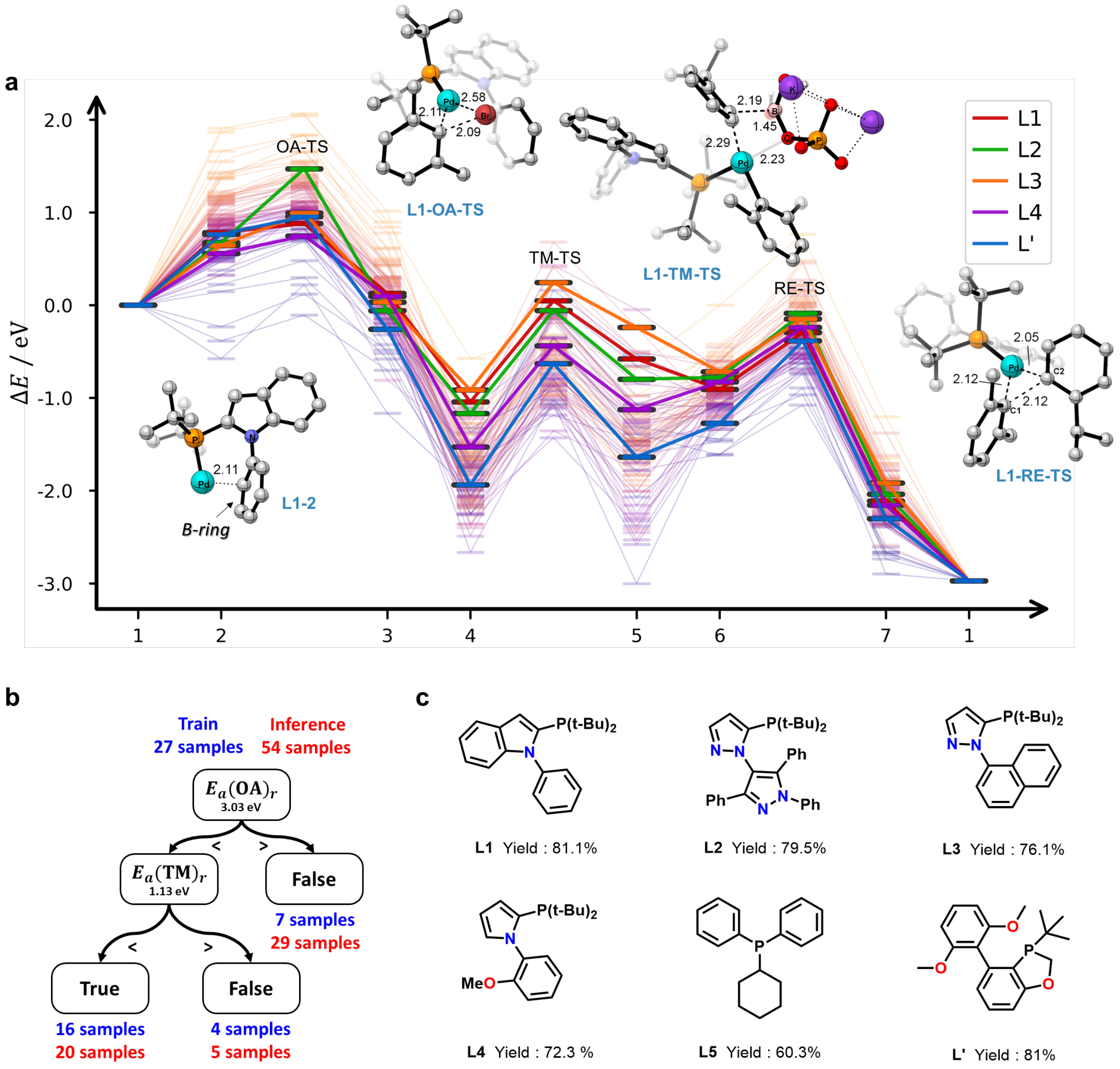

**Figure 4. Rational ligand design for Suzuki-Miyaura reaction through BDGM-PES method. a.** Reaction profile screening for 81 types of ligands for target C-C coupling reaction: 2-bromo-m-xylene coupling with 2-isopropylphenylboronic acid. The structure of species 2 for **L1** is noted as L1-2. The structure of TSs for OA, RE, and TM, are exemplified by the structures of **L1**, noted as L1-OA-TS, L1-TM-TS, and L1-RE-TS, respectively. **b.** Illustration of the CART tree in classifying the reactivity of ligands. The ligands are set to true if the ligand have final productivity greater than 40%. **c.** The structure and reaction yield of phosphine ligands (L': from literature [14] that reaches yield 81%)

difference between the most stable conformer in the given cycle and that in the final (10th) cycle. As shown, there is an average of 1.05 eV energy drop per molecule after ten cycles. The most substantial improvement occurred in the first cycle (-0.40 eV), followed by progressively smaller gains (-0.29 eV in cycle 2, -0.12 eV in cycle 3), before converging (ΔE ≤ 0.02 eV after cycle 9). Similar iterative refinement is found for other scaffold molecules, where the average energy of L, PdL, $PdLR_1R_2$, and PdLBorate datasets drop by 0.10 eV, 0.86 eV, 1.07 eV, and 1.38 eV, respectively. For example, Figure 3c shows the 3D structure evolution of molecule A, a Pd(II) complex comprising a ddpv chelating ligand, chloride ($Cl^-$), and an aryl fragment. Its large energy reduction (1.59 eV) after 10 cycles is mainly caused by the Pd coordination transformation from tetrahedral to square-planar pattern. The final generated geometry reaches an RMSE of 0.09 Å per atom compared to the G-NN structure.

Equipped with the general DGM trained on extensive molecule structures, we employ our dual-phase constraint-guided generation algorithm to produce high-quality 3D TS structures. As illustrated in Figure 3d, the algorithm applies geometric constraints to the three atoms in the reaction center during generation: d(Pd–C) = 2.16 Å, d(Pd–Br) = 2.51 Å, and d(C–Br) = 2.09 Å. This constrained DGM generates an initial TS structure (TS A) possessing the correct bond topology. This initial structure is then refined through relaxation (keeping Pd, C, and Br fixed) and a subsequent optimization through CBD algorithm, to produce the TS. A true TS is confirmed by identifying a single imaginary frequency corresponding to the reaction coordinate. However, comparing this initial TS guess to the reference TS geometry reveals a significant deviation, with an RMSE of 0.68 Å (see Figure 3d). To address this limitation, we incorporated the generated TS data into further training, yielding an enhanced model termed DGM-TS. DGM-TS substantially improves TS generation quality. Key bond distance errors in the generated TS (TS A) are minimized: Δd(C-Br) = + 0.03 Å and Δd(C–Pd) = + 0.02 Å, while the overall RMSE decreases dramatically to 0.09 Å. The effectiveness of this two-stage approach (constrained generation followed by DGM-TS refinement) is demonstrated on a larger test set of 2414 molecules (detailed in Supplementary Figure S5). The constrained generation stage successfully identified the correct TS topology in 87.0% of cases. Subsequent refinement using DGM-TS significantly increased the success rate to 98.6%. These results underscore the strong *zero-shot* capability inherent in our BDGM-PES architecture.

**4. Reactivity prediction and Experiment verification**

Taking BDGM-PES, we can now in-silico explore unknown reaction space to design desirable catalysts. Here we select Suzuki-Miyaura coupling as the target to search for the best Pd-P catalysts for 2-bromo-m-xylene and 2-isopropylphenylboronic acid with $K_3PO_4$ as the base. The main purpose of BDGM-PES method is to screen the ligand space of Pd-P catalysts to achieve the highest activity. In total, 81 phosphine ligands are considered, among them 27 from the existing experiment by Doyle *et al.* with the activity ranging from 0 to 81 % conversion[14], and 54 from commercially available phosphine ligands where no previous experiment results are available.

Using the BDGM-PES, we generated reaction pathways for all 81 Pd-P catalysts, with their corresponding energy profiles presented in Figure 4a. All structures, including the key TSs (OA-TS, RE-TS, and TM-TS), were initially generated and optimized using BDGM-PES, followed by final optimization at the PBE-D3 level of DFT (see Methods).

Representative TS structures for ligand L1 are shown in Figure 4a insets, namely L1-OA-TS, L1-TM-TS and L1-RE-TS. The L1-OA-TS exhibits a three-centered TS where the Pd atom coordinates with the Br atom while simultaneously interacts with the aromatic ring's C atom; key bond distances here being d(Pd-C) = 2.11 Å, d(Pd-Br) = 2.58 Å, and d(C-Br) = 2.09 Å. Next, the L1-TM-TS involves a 4-membered ring (O-B-C-Pd atoms), characterized by Pd-C bond formation and C-B bond breaking, with the O atom coordinating to Pd. The involved TS distances are d(Pd-C) = 2.29 Å, d(B-C) = 2.19 Å, d(B-O) = 1.45 Å, and d(Pd-O) = 2.23 Å. Finally, the L1-RE-TS, also a three-centered state, leads to the C-C bond formation in the aromatic rings and the cleavage of two Pd-C bonds, and the critical TS bond lengths are d(Pd-C1) = 2.12 Å, d(Pd-C2) = 2.05 Å, and d(C1-C2) = 2.12 Å.

We emphasize that different ligands follow similar structural scaffolds, but the individual geometries and energies vary significantly due to differences in ligand properties, as summarized in the reaction profiles Figure 4a. The energy span for an individual states can be up to several eV. For instance, OA-TS energies average 1.13 eV, ranging from 0.28 eV (lowest) to 2.07 eV (highest). One must bear in mind that these profiles reflect only the reaction kinetics in the gas phase, which partly explains the large energetic heterogeneity as the homogeneous reaction conditions should be important in controlling reaction rates. Nevertheless, this comprehensive reaction kinetic data provides valuable kinetics descriptors, offering the foundation for guiding reaction design.

From the reaction profile, we can obtain six energy-based descriptors, *i.e.*, the forward (f) and reverse (r) reaction barriers for three elementary steps, *i.e.*,$E_a(\mathrm{OA})_f$, $E_a(\mathrm{OA})_r$, $E_a(\mathrm{TM})_f$, $E_a(\mathrm{TM})_r$, $E_a(\mathrm{RE})_f$, $E_a(\mathrm{RE})_r$. The value of these descriptors can be found in Supplementary Table S7. Using these descriptors and 27 existing experimental datapoints, we built a binary Classification and Regression Tree (CART) model to identify key descriptors governing catalytic activity. After training on experimental data, a tree depth of 2 proves sufficient for catalyst classification. The decision tree first splits ligands at $E_a(\mathrm{OA})_\mathrm{r}$ = 3.03 eV. Ligands exceeding this reverse oxidative addition barrier are classified as inactive, excluding 7 inactive catalysts correctly. The remaining ligands are further partitioned at $E_a(\mathrm{TM})_\mathrm{r}$ = 1.13 eV based on transmetallation energy analysis, again correctly screening out 4 additional inactive cases. The terminal node predicts 16 ligands as active (predicted yield > 40%, as shown in Supplementary Table S8). This simple model achieves high fidelity with only 2 ligands incorrectly predicted as active (88.9% correctness). Using this optimized 2Depth CART model, we screened 54 commercially available P-ligands and predicted 20 as active catalysts (Supplementary Table S9).

To validate these predictions, we experimentally tested 9 ligands with low price and easy access on market. All our experiments employed $Pd_2(dba)_3$ (3 mol%) and $K_3PO_4$ under standardized conditions (all experiments detailed in Supplementary Information C). Our experiment results reveal that 7 out of the 9 ligands can achieve >40% yield, *i.e.*, 77.8% prediction correctness. Especially, this screening revealed five novel active P-ligands (**L1**-**L5**) yielding > 60% product. Particularly remarkable are P(t-Bu)$_2$-TripyrazPh (**L1**), P(t-Bu)$_2$-IndolPh (**L2**), and P(t-Bu)$_2$-Naph (**L3**) with yields of 81.1%, 79.5%, and 76.1%, respectively. It is interesting that, interestingly, all **L1**, **L2**, **L3**, **L4**, as well as **L'**, the ligand with highest activity in the literature, all belong to the bulky Buchwald-type family, which are widely recognized as excellent choices for C-C and C-N coupling reactions.[58–61] These ligands possess key structural features that critically influence catalytic behavior: their large size promotes the formation of highly reactive mono-dentate PdL species by positioning the "B ring" (Figure 4a, L1-2) within the metal's first coordination sphere, thereby sterically preventing the binding of multiple ancillary ligands. Furthermore, they stabilize the unsaturated Pd center through Pd-arene interactions involving the biaryl π-system, exhibiting a pseudo-bidentate mode. These characteristics align well with our computational results.

(1). $E_a(\mathrm{OA})_\mathrm{f}$ for **L1**-**L4**, and **L'** (0.88, 1.47, 0.99, 0.65, and 0.96 eV, respectively) fall within a moderate range (0.28–2.07 eV across all ligands studied), suggesting that optimal catalytic activity requires barriers that are neither excessively high nor prohibitively low.

(2). Intermediates **4** and **5** for ligands **L1-L4** exhibit unusually high relative energies (low stability) compared to other ligands. Intermediate **4** energies rank at positions 3 (-1.04 eV), 7 (-1.17 eV), 2 (-0.91 eV), 23 (-1.53 eV), and 51 (-1.94 eV) for **L1-L4**, **L'** respectively, while intermediate 5 energies take positions 7 (-0.58 eV), 11 (-0.80 eV), 3 (-0.24 eV), 22 (-1.13 eV), and 57(-1.64). Nevertheless, their intermediates **4** and **5** are still more stable than preceding states **1** and **3**, preventing the reverse reaction to occur but facilitating the forward reaction, effectively smoothing the whole energy profile. Too stable intermediate 4 and 5 will obviously inhibit the reaction by stopping at subsequent RE reaction (**6**->**7**). This is attributed to the Buckwald-type ligands' ability in preventing the formation of high-coordination with other ligand/bases (>1).

The moderate OA barriers together with the destabilization of intermediates explain the high catalytic performance observed for **L1**-**L4**, demonstrating the kinetics descriptors generated from BDGM-PES do enable physics-informed rational ligand screening.

## Conclusion

To recap, we develop a general theoretical framework, namely BDGM-PES, for reaction design by combining DGM of structure generation, PES calculation for structure evaluation and optimization and active learning for exploring uncharted chemical space. Our DGM utilized a HPNN-ET model, an enhanced equivariant architecture with unified information fusion strategy. The PES optimization is based on G-NN potential which can achieve fast and accurate structure relaxation and TS search. Taking Pd-catalyzed Suzuki reaction as example, we demonstrate that BDGM-PES is an automated and highly efficient method to generate new molecule structures, reveal reaction pathways and design new catalysts. Specifically, we achieve the following.

(i) A general DGM (accessible from http://www.laspai.com/) based on HPNN-ET framework covering all elements from the first four periods of the periodic table (H to Xe, Z =1-54), trained on Pd-database, GEOM-DRUG, and metal-containing compounds from PubChem (total 467,757 structures). This enables fast, accurate and autonomous molecular structure generation: high topology correctness (100%), high geometry precision (deviation < 0.05 Å), as well as high computation efficiency (0.35s per structure) for large metal-containing complexes (e.g. more than 100 atoms).

(ii) Reaction pathway generation with optimized reaction intermediates and TSs, leading to 81 Pd-catalyzed Suzuki C-C coupling energy profiles for reaction design.

(iii) Identification of 7 active P-ligand from experiment for 2-bromo-m-xylene coupling with 2-isopropylphenylboronic reaction following theoretical predictions, among them the best ligand achieves >80% yield (L1: 81.1%).

## Method

### Details of diffusion process described by stochastic differential equation (SDE)

We follow the variance exploding stochastic differential equation (VE-SDE) to construct the DGM.[33] The model is trained on the forward diffusion process of VE-SDE, as shown in Eq1.

$$\mathrm{d}\boldsymbol{x} = g(t)\mathrm{d}\boldsymbol{w} \qquad (\mathbf{1})$$

where $\boldsymbol{x} \in \mathbb{R}^{N_{atom}\times 3}$ represents the atomic coordinates; $\boldsymbol{w}$ is the standard Wiener process and $g(\cdot)$ is the diffusion coefficient of $\boldsymbol{x}(t)$; $t$ describes the diffusion process and increases from 0 to 1. VE-SDE sets $g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$, where $\sigma^2(t)$ is the variance of stochastic gaussian noise at $t$. The cumulation of noise in gaussian form from 0 to $t$ lead to the distribution of $\boldsymbol{x}$ is denoted as $p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0)$, from which the $\nabla_x \mathrm{log} p_{0t}(\boldsymbol{x}(t))|\boldsymbol{x}(0)$ can be obtained, which is crucial in the reverse process of diffusion (*vide infra*). We use the score function $\boldsymbol{s_\theta}(\boldsymbol{x}, t)$ to approximate $\nabla_x \mathrm{log} p_t(\boldsymbol{x})$ via Eq2 by minimizing the expectation value $\mathbb{E}_t$:

$$\begin{aligned}\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbb{E}_t\{a(t)\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\mathbf{x}(0)}[\|\boldsymbol{s_\theta}(\boldsymbol{x}(t), t) \\ - \nabla_x \mathrm{log} p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0)\|_2^2]\}\end{aligned} \qquad (2)$$

where $a(t) \propto \sigma^{-2}(t)$ weights the loss across time. This score function $s_\theta$ is defined through a neural network, *i.e.*, HPNN-ET in this work. HPNN-ET satisfy two key requirements: (1) preserve SE(3) equivariance for atom coordinates to maintain geometric consistency; and (2) integrate correctly molecular 2D graph information to ensure proper stoichiometry and bonding connectivity.

In 3D conformation generation, VE-SDE takes the reverse process starting from noise, as shown in Eq 3,

$$\mathrm{d}\boldsymbol{x} = [-g^2(t)\boldsymbol{s_\theta}(\boldsymbol{x}(t), t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \qquad (3)$$

, where $\bar{\mathbf{w}}$ is a standard Wiener process when time flows backwards from 1 to 0.

In generating TS structure, we introduce a constrained generation process to incorporate chemical heuristics directly into the sampling procedure. Specifically, we define a set of atoms $\boldsymbol{\mathcal{N}_{mod}}$, where chemical knowledge, such as reaction center geometry or typical bond-length/bond-angle constraints for a TS, is applied.

$$\mathrm{d}\mathbf{x_i} = \begin{cases} \boldsymbol{dx} & \boldsymbol{if\ i \notin \mathcal{N}_{mod}} \\ \mathrm{d}\boldsymbol{x}_{\mathbf{mod}} & \boldsymbol{if\ i \in \mathcal{N}_{mod}} \end{cases} \qquad (4)$$

As shown in Eq. (4), for atoms are not present in $\boldsymbol{\mathcal{N}_{mod}}$, the update follows the normal reverse diffusion in Eq. (3). For atoms in $\boldsymbol{\mathcal{N}_{mod}}$, the update is modified to $\mathrm{d}\boldsymbol{x}_{\mathrm{mod}}$ which is defined in Eq. (5):

$$\mathrm{d}\boldsymbol{x}_{\mathbf{mod}} = \big(1 - \lambda(t)\big)\mathrm{d}\boldsymbol{x} + \lambda(t)h(\boldsymbol{x}_{\mathrm{mod}}) \qquad (5)$$

The term $\lambda(t)$ serves as a time-dependent factor to modulate how much of the displacement in the diffusion process is contributed by chemical heuristics versus the gradient ($\boldsymbol{s_\theta}$). Typically, $\lambda(t)$ ascends to 1 at the end of denoising. Meanwhile, $h(\boldsymbol{x}_{\mathrm{mod}})$ defines how the heuristic-informed displacement is introduced. $h(\boldsymbol{x}_{\mathrm{mod}})$ can be formulated using correlation function by scanning the reaction coordinates (e.g. reacting bond lengths, bond angles, and dihedral angles). For instance, in a reaction where bonds are making or breaking, we describe the reaction mode by pair distance ($d_{ij}$), as defined in Eq. 6

$$h(\boldsymbol{x}_{\mathbf{mod}})_i = \sum_j(|\mathbf{x}_{\boldsymbol{i}} - \mathbf{x}_{\boldsymbol{j}}| - d_{ij}) \cdot \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|} \qquad (6)$$

Using this strategy, we can bias the generated conformations according to the reaction mode, thereby boosting the model's expressive power. In practice, one can train on minima structures (*e.g.*, the IS and FS) and use simple heuristics to guide the generation of TS structures.

### High-order Pair-reduced with Edge and Time Neural Network

The HPNN-ET architecture follows the standard message-passing neural network (MPNN)[62] framework that has been widely utilized in generating machine learning potentials, while introducing critical innovations in feature representation and update mechanisms. As illustrated in Figure1a, our model processes atomic systems through three primary stages: (1) hierarchical feature abstraction, (2) multi-modal message passing, and (3) iterative co-update of invariant and equivariant representations.

The system begins with a 3D atomic structure that is projected into a hybrid 2D-3D graph representation. Each atom (node) is initialized with invariant features $a_i \in \mathbb{R}^{L_A\times 1}$ derived from elemental properties (*e.g.*, atomic number), while equivariant features are initialized as $\boldsymbol{v_i} = \vec{\mathbf{0}} \in \mathbb{R}^{L_A\times 3}$. The $L_A$ denotes the length of atomic embedding, also the node dimension. Neighbor selection employs a dual-criterion approach, where atom $j$ is considered adjacent to central atom $i$ if it satisfies either $\boldsymbol{n}$ th order adjacency in the 2D molecular graph, or partial proximity within radius $r_c$ (typically $\boldsymbol{n}$ =7, $r_c$ =5.0 Å)

Edge attributes $e_{ij}$ synthesize multiple interaction modalities of conditional and spatial through Hadamard product of $b_{ij}$ and $R(r_{ij})$, as shown in Eq 7, where the bond-type embedding $b_{ij}$ is derived from chemical interaction types determined jointly by 2D-graph adjacency and bond type; the spatial encodings $R(r_{ij})$ is from radial basis functions, as shown in Eq 8, where k indexes a series of gaussian centers to discretize the pair distance. In addition, we may also include the high order angular information *via* real spherical harmonics $Y_{lm}(\theta,\phi)$ with *l* > 1 using Eq 9 ($\oplus$ represents concatenate operation, *Nei* represents the neighbor of atom *i*).

$$e_{ij} = b_{ij} \circ R(r_{ij}) \tag{7}$$

$$R_{ij}^{k} = \exp\left(-\frac{\left(r_{ij}-r_k\right)^2}{2\sigma^2}\right) \tag{8}$$

$$x_i = x_i + \boldsymbol{W_l}\left\|\sum_j^{Nei}\boldsymbol{W_{l,m}}\boldsymbol{Y^{lm}}(\boldsymbol{r_{ij}})\right\|_2 \oplus \boldsymbol{W_1}\left\|\sum_j^{Nei}\boldsymbol{W_{1,m}}\boldsymbol{Y^{1m}}(\boldsymbol{r_{ij}})\right\|_2 \tag{9}$$

Our architecture involves the co-update of invariant ($x_i$) and equivariant ($\boldsymbol{v_i}$) atomic features. As illustrated in Figure 1a, this process unfolds through three iterative cycles (denoted as c in the following Eqs) through interaction block, implemented with residual connections. A unified information fusion strategy in gathering 2D and 3D information, as well as temporary information ($t \in \mathbb{R}^{L_A}$) into $x_i$ and $\boldsymbol{v_i}$. The interaction block typically includes three components, message-construction, message aggregation, and feature integration.

**Message Construction.** At each iteration $c$, neighboring atom pairs $(i, j)$ generate through Eq 10 and 11.

$$s_{ij}^{c} = \boldsymbol{W}^{\downarrow}x_j \circ \boldsymbol{W}e_{ij} \tag{10}$$

$$\boldsymbol{v_{ij}^{c}} = s_{ij}^{c} \circ (\boldsymbol{Y^{1m}}(\boldsymbol{r_{ij}}) + \boldsymbol{Wr_{ij}}) \tag{11}$$

The invariant pair information $s_{ij}^{c}$ is produced by multiplying reduced $x_j$ and $e_{ij}$ via the shrinking linear layer $\boldsymbol{W}^{\downarrow}$. In Eq 10, $s_{ij}$ is then broadcast to multiply with the equivariant vector constructed from $\boldsymbol{Y^{lm}}(\boldsymbol{r_{ij}})$ and $\boldsymbol{Wr_{ij}}$ in Eq. 11.

**Message Aggregation.** Each atom accumulates information from its neighborhood through Eq 12 and 13.

$$\boldsymbol{h_{i,v}^{c}} = (\boldsymbol{W_t}t + b_t) \circ (\boldsymbol{W}^{\uparrow}\textstyle\sum_j^{Nei}\boldsymbol{v_{ij}^{c}}) + \boldsymbol{v_i^{c}} \tag{12}$$

$$h_{i,x}^{c} = \ \left\|W^{l}\boldsymbol{h_{i,v}^{c}}\right\|_2 \oplus \left(\textstyle\sum_j^{Nei}s_{ij}^{c} + \boldsymbol{W}^{\downarrow}x_j\right) \tag{13}$$

In Eq. 11, the expanding linear layer ($\boldsymbol{W}^{\uparrow}$) update $\boldsymbol{v_i^{c}}$ once the pair operation is finished, and the time-dependent modulation is through a normalized temporal embedding $t \in (0,1)$ , with linear embedding through $\boldsymbol{W_t}$ and $b_t$ (Eq 12). A varibale angular interaction scheme is developped to incorporate both equivariant and invariant information. The former is conducted through normalization, as shown in Eq 13 leading term on the left side, and the latter is formulated in Eq 13 after the contatenation operation $\oplus$.

**Feature Integration.**

The central atom updates its features by synthesizing accumulated messages with temporal dynamics through Eq 14 and Eq 15.

$$x_i^{c+1} = NN\left(h_{i,x}^{c}\right) \tag{14}$$

$$\boldsymbol{v_i^{c+1}} = x_i^{c+1} + \boldsymbol{h_{i,v}^{c}} \tag{15}$$

The $NN(\cdot)$ denotes the atomic NN layer (depicted in Figure 2b), which sequentially incorporates operation of linear layer, layer normalization[63], activation function, and linear layer. $x_i^{c+1}$ and $\boldsymbol{v_i^{c+1}}$ is produced after these above interactions.

The final atomic coordinates are predicted through an equivariance-preserving readout module that maintains geometric consistency with the SE(3) symmetry group. As depicted in Eq. (16), the equivariant feature vector $\boldsymbol{v_i} \in \mathbb{R}^{L_A \times 3}$ undergoes dimension reduction *via* a learnable projection matrix $\boldsymbol{W}^{\downarrow}$ to finally produce score function $\boldsymbol{s_{\theta,}}(\boldsymbol{x}, t) \in \mathbb{R}^{1\times 3}$

$$s_{\theta,i} = \boldsymbol{W}^{\downarrow}\boldsymbol{v_i} \tag{16}$$

**DFT calculations**

All DFT calculations were performed by using the plane wave VASP[64–66] code, where electron–ion interaction was represented by the projector augmented wave pseudopotential. The exchange-correlation functional utilized was the spin-polarized GGA-PBE[67,68]. The kinetic energy cutoff was set at 450 eV. The first Brillion zone k-point sampling utilized the Gamma-centered mesh grid. The energy and force criterion for convergence of the electron density and structure optimization were set at $10^{-5}$ eV and 0.05 eV Å$^{-1}$, respectively.

## AUTHOR INFORMATION

**Corresponding Authors**

*E-mail: lin_c@fudan.edu.cn (L. C.)

*E-mail: zpliu@fudan.edu.cn(Z.-P. L.)

**ORCID**

Zhi-Pan Liu: 0000-0002-2906-5217

## ACKNOWLEDGMENTS

This work received financial support from the National Science Foundation of China (12188101, 22033003, 91745201, 91945301), the Fundamental Research Funds for the Central Universities (20720250005, 20720220011), the National Key Research and Development Program of China (2024YFA1509600), and the Tencent Foundation for XPLORER PRIZE.

## Notes

The authors declare no conflict of interest.

**References**

1. Evans, M. G. & Polanyi, M. Inertia and driving force of chemical reactions. *Trans. Faraday Soc.* **34**, 11–24 (1938).
2. Nørskov, J. K., Studt, F., Abild-Pedersen, F. & Bligaard, T. *Fundamental Concepts in Heterogeneous Catalysis*. (Wiley-Blackwell, 2014). doi:10.1002/9781118892114.
3. Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59**, 96–103 (1937).
4. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **120**, 1620–1689 (2020).
5. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
6. Reid, J. P. & Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2**, 290–305 (2018).
7. Foscato, M. & Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **10**, 2354–2377 (2020).
8. Morán-González, L., Burnage, A. L., Nova, A. & Balcells, D. AI Approaches to Homogeneous

Catalysis with Transition Metal Complexes. *ACS Catal.* **15**, 9089–9105 (2025).

9. Nandy, A. *et al.* Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **121**, 9927–10000 (2021).
10. Renom-Carrasco, M. & Lefort, L. Ligand libraries for high throughput screening of homogeneous catalysts. *Chem. Soc. Rev.* **47**, 5038–5060 (2018).
11. Keith, J. A. *et al.* Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **121**, 9816–9872 (2021).
12. Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
13. Wu, K. & Doyle, A. G. Parameterization of phosphine ligands demonstrates enhancement of nickel catalysis via remote steric effects. *Nat. Chem.* **9**, 779–784 (2017).
14. Newman-Stonebraker, S. H. *et al.* Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301–308 (2021).
15. Gensch, T. *et al.* A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **144**, 1205–1217 (2022).
16. Durand, D. J. & Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **119**, 6561–6594 (2019).
17. Fey, N., Orpen, A. G. & Harvey, J. N. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal–phosphorus bond. *Coord. Chem. Rev.* **253**, 704–722 (2009).
18. Poater, A. *et al.* SambVca: A web application for the calculation of the buried volume of N-heterocyclic carbene ligands. *Eur. J. Inorg. Chem.* 1759–1766 (2009) doi:10.1002/ejic.200801160.
19. Falivene, L. *et al.* Towards the online computer-aided design of catalytic pockets. *Nat. Chem.* **11**, 872–879 (2019).
20. Clavier, H. & Nolan, S. P. Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chem. Commun.* **46**, 841–861 (2010).
21. Van Leeuwen, P. W. N. M., Kamer, P. C. J., Reek, J. N. H. & Dierkes, P. Ligand Bite Angle Effects in Metal-catalyzed C−C Bond Formation. *Chem. Rev.* **100**, 2741–2770 (2000).
22. Kranenburg, M. *et al.* New Diphosphine Ligands Based on Heterocyclic Aromatics Inducing Very High Regioselectivity in Rhodium-Catalyzed Hydroformylation: Effect of the Bite Angle. *Organometallics* **14**, 3081–3089 (1995).
23. Gensow), M.-N. B. (née, Freixa, Z. & Leeuwen, P. W. N. M. van. Bite angle effects of diphosphines in C–C and C–X bond forming cross coupling reactions. *Chem. Soc. Rev.* **38**, 1099–1118 (2009).
24. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal–phosphorus bond. *Coord. Chem. Rev.* **253**, 704–722 (2009).
25. Ma, S. *et al.* Data-driven discovery of active phosphine ligand space for cross-coupling reactions. *Chem. Sci.* **15**, 13359–13368 (2024).
26. Makoś, M. Z., Verma, N., Larson, E. C., Freindorf, M. & Kraka, E. Generative adversarial networks for transition state geometry prediction. *J. Chem. Phys.* **155**, 024116 (2021).
27. Ganea, O.-E. *et al.* GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles. Preprint at https://doi.org/10.48550/arXiv.2106.07802 (2021).
28. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.
29. Westermayr, J., Gilkes, J., Barrett, R. & Maurer, R. J. High-throughput property-driven generative design of functional organic molecules. *Nat. Comput. Sci.* **3**, 139–148 (2023).
30. Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminformatics* **10**, 31 (2018).
31. Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nat. Commun.* **14**, 1168 (2023).
32. Xu, M. *et al.* GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation. Preprint at http://arxiv.org/abs/2203.02923 (2022).
33. Song, Y. *et al.* Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at https://doi.org/10.48550/arXiv.2011.13456 (2021).
34. Kim, S., Woo, J. & Kim, W. Y. Diffusion-based generative AI for exploring transition states from 2D molecular graphs. *Nat. Commun.* **15**, 341 (2024).
35. Morehead, A. & Cheng, J. Geometry-complete diffusion for 3D molecule generation and optimization. *Commun. Chem.* **7**, 150 (2024).
36. Jing, B., Corso, G., Chang, J., Barzilay, R. & Jaakkola, T. S. Torsional Diffusion for Molecular Conformer Generation. in (2022). doi:10.48550/arXiv.2206.01729.
37. Vainio, M. J. & Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. 13.
38. Riniker, S. & Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
39. Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **57**, 1747–1756 (2017).
40. Kraka, E., Zou, W., Tao, Y. & Freindorf, M. Exploring the Mechanism of Catalysis with the Unified Reaction Valley Approach (URVA)—A Review. *Catalysts* **10**, 691 (2020).
41. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How Powerful are Graph Neural Networks? Preprint at https://doi.org/10.48550/arXiv.1810.00826 (2019).
42. Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **9**, 779 (2022).
43. Shang, C. & Liu, Z.-P. Constrained Broyden Minimization Combined with the Dimer Method for

Locating Transition State of Complex Reactions. *J. Chem. Theory Comput.* **6**, 1136–1144 (2010).

44. Xie, X.-T. *et al.* LASP to the Future of Atomic Simulation: Intelligence and Automation. *Precis. Chem.* (2024) doi:10.1021/prechem.4c00060.
45. Martin, R. & Buchwald, S. L. Palladium-Catalyzed Suzuki−Miyaura Cross-Coupling Reactions Employing Dialkylbiaryl Phosphine Ligands. *Acc. Chem. Res.* **41**, 1461–1473 (2008).
46. Miyaura, Norio. & Suzuki, Akira. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. *Chem. Rev.* **95**, 2457–2483 (1995).
47. Braga, A. A. C., Morgon, N. H., Ujaque, G. & Maseras, F. Computational Characterization of the Role of the Base in the Suzuki−Miyaura Cross-Coupling Reaction. *J. Am. Chem. Soc.* **127**, 9298–9307 (2005).
48. Pérez-Rodríguez, M. *et al.* C−C Reductive Elimination in Palladium Complexes, and the Role of Coupling Additives. A DFT Study Supported by Experiment. *J. Am. Chem. Soc.* **131**, 3650–3657 (2009).
49. Braga, A. A. C., Ujaque, G. & Maseras, F. A DFT Study of the Full Catalytic Cycle of the Suzuki−Miyaura Cross-Coupling on a Model System. *Organometallics* **25**, 3647–3658 (2006).
50. RDKit: Open-source cheminformatics. https://www.rdkit.org.
51. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **9**, 185 (2022).
52. Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **77**, 334–338 (1955).
53. Shang, C. & Liu, Z.-P. Constrained Broyden Dimer Method with Bias Potential for Exploring Potential Energy Surface of Multistep Reaction Process. *J. Chem. Theory Comput.* **8**, 2215–2222 (2012).
54. Barder, T. E., Biscoe, M. R. & Buchwald, S. L. Structural Insights into Active Catalyst Structures and Oxidative Addition to (Biaryl)phosphine−Palladium Complexes via Density Functional Theory and Experimental Studies. *Organometallics* **26**, 2183–2192 (2007).
55. Yin, J., Rainka, M. P., Zhang, X.-X. & Buchwald, S. L. A Highly Active Suzuki Catalyst for the Synthesis of Sterically Hindered Biaryls: Novel Ligand Coordination. *J. Am. Chem. Soc.* **124**, 1162–1163 (2002).
56. Reid, S. M., Boyle, R. C., Mague, J. T. & Fink, M. J. A Dicoordinate Palladium(0) Complex with an Unusual Intramolecular η-Arene Coordination. *J. Am. Chem. Soc.* **125**, 7816–7817 (2003).
57. Barder, T. E., Walker, S. D., Martinelli, J. R. & Buchwald, S. L. Catalysts for Suzuki−Miyaura Coupling Processes: Scope and Studies of the Effect of Ligand Structure. *J. Am. Chem. Soc.* **127**, 4685–4696 (2005).
58. Rataboul, F. *et al.* New Ligands for a General Palladium-Catalyzed Amination of Aryl and Heteroaryl Chlorides. *Chem. – Eur. J.* **10**, 2983–2990 (2004).
59. Zapf, A. *et al.* Practical synthesis of new and highly efficient ligands for the Suzuki reaction of aryl chlorides. *Chem. Commun.* 38–39 (2004) doi:10.1039/B311268N.
60. Surry, D. S. & Buchwald, S. L. Dialkylbiaryl phosphines in Pd-catalyzed amination: a user's guide. *Chem. Sci.* **2**, 27–50 (2010).
61. Ruiz-Castillo, P. & Buchwald, S. L. Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chem. Rev.* **116**, 12564–12649 (2016).
62. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for Quantum chemistry. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 1263–1272 (JMLR.org, Sydney, NSW, Australia, 2017).
63. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer Normalization. Preprint at https://doi.org/10.48550/arXiv.1607.06450 (2016).
64. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
65. Kresse, G. & Hafner, J. *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
66. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
67. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
68. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).