11 September 2026

Descriptor-Based Pre-Training Improves Reaction Property Prediction

Akshat Shirish Zalte[1,2], Cheng Fang[2], Yiting Zheng[3], Essam Metwally[4], Alan Cheng[4], Alec Glisman[4], Hao-Wei Pang[2]

1. Department of Chemical Engineering, Massachusetts Institute of Technology
2. Modeling and Informatics, Merck & Co., Inc.
3. Modeling and Informatics, Merck & Co., Inc.
4. Modeling and Informatics, Merck & Co., Inc.

Abstract

Pre-training graph neural networks on large unlabeled datasets has substantially improved molecular property prediction, yet pre-training approaches for reaction property prediction remain underexplored. We introduce CheMeleon-Rxn, a pre-trained reaction graph neural network that adapts the CheMeleon framework to condensed graphs of reaction (CGR) and demonstrates that descriptor-regression is an effective pre-training strategy for low-data reaction property prediction. An O(10M)-parameter D-MPNN encoder was pre-trained on #2 million reactions by regressing onto dense reaction descriptors pooled from classical molecular descriptors, then fine-tuned on small reaction property prediction datasets. Across regression tasks spanning gasphase activation energies, reaction enthalpies, experimental yields, and rate coefficients, CheMeleon-Rxn is best or statistically tied-for-best on seven of eight tasks. It outperforms the same encoder trained from scratch, as well as classical-descriptor, fingerprint, and pre-trained-fingerprint baselines. We further tested the pre-training strategy across various graph neural network architectures and found that its benefit holds for other edge-aware backbones such as GINE, and that it is robust to the choice of descriptor set, pooling operation, and other pre-training hyperparameters. We visualized and analyzed the CheMeleon-Rxn embeddings, which reveal chemically coherent structure in the learned reaction representation space. The CheMeleon-Rxn code has been released publicly (https://github.com/MSDLLCpapers/chemeleon-rxn) to encourage adoption and extension by the community. The model weights will be released upon publication.

Descriptor-Based Pre-Training Improves Reaction Property Prediction

Akshat Shirish Zalte,[†,‡] Cheng Fang,[‡] Yiting Zheng,[¶] Essam Metwally,[§] Alan Cheng,[§] Alec Glisman,[*,§] and Hao-Wei Pang[*,‡]

†*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

‡*Modeling and Informatics, Merck & Co., Inc., Cambridge, Massachusetts 02141, United States*

¶*Modeling and Informatics, Merck & Co., Inc., Boston, Massachusetts 02115, United States*

§*Modeling and Informatics, Merck & Co., Inc., South San Francisco, California 94080, United States*

E-mail: alec.glisman@merck.com; hao-wei.pang@merck.com

Abstract

Pre-training graph neural networks on large unlabeled datasets has substantially improved molecular property prediction, yet pre-training approaches for reaction property prediction remain underexplored. We introduce **CheMeleon-Rxn**, a pre-trained reaction graph neural network that adapts the CheMeleon framework to condensed graphs of reaction (CGR) and demonstrates that descriptor-regression is an effective pre-training strategy for low-data reaction property prediction. An $\mathcal{O}$(10M)-parameter D-MPNN encoder was pre-trained on $\sim$2 million reactions by regressing onto dense reaction descriptors pooled from classical molecular descriptors, then fine-tuned on

small reaction property prediction datasets. Across regression tasks spanning gas-phase activation energies, reaction enthalpies, experimental yields, and rate coefficients, CheMeleon-Rxn is best or statistically tied-for-best on seven of eight tasks. It outperforms the same encoder trained from scratch, as well as classical-descriptor, fingerprint, and pre-trained-fingerprint baselines. We further tested the pre-training strategy across various graph neural network architectures and found that its benefit holds for other edge-aware backbones such as GINE, and that it is robust to the choice of descriptor set, pooling operation, and other pre-training hyperparameters. We visualized and analyzed the CheMeleon-Rxn embeddings, which reveal chemically coherent structure in the learned reaction representation space. The CheMeleon-Rxn code has been released publicly (`https://github.com/MSDLLCpapers/chemeleon-rxn`) to encourage adoption and extension by the community. The model weights will be released upon publication.

Introduction

Machine learning has become a central tool for predicting molecular properties, and graph neural networks (GNNs) have emerged as the dominant approach by learning representations directly from molecular structure rather than relying on fixed descriptors.[1–5] Predicting the properties of chemical reactions, however, has proven substantially harder. A reaction is a transformation between molecular states rather than a single static structure, and its key properties (activation energies, reaction enthalpies, rate coefficients, and yields) depend on subtle electronic and steric changes along the reaction coordinate.[6] Compounding this difficulty, labeled reaction datasets are typically small, expensive to generate, and heterogeneous in origin. For example, barrier heights and enthalpies are derived from costly quantum-chemistry calculations, rate coefficients from kinetics measurements, while yields from high-throughput experimentation or noisy text-mined patent records. Purely data-driven deep models tend to perform poorly in the low-data regime, and the development of transferable

reaction representations continues to lag well behind its molecular counterparts.

Because a reaction is not a single molecule, a central question is how to encode the joint transformation of reactants into products as a fixed input for a predictive model. Early approaches typically relied on fixed representations: quantum-chemical descriptors of the participating species, or structure-based fingerprints such as the concatenation or difference of reactant and product Morgan (extended-connectivity) fingerprints.[7–9] The differential reaction fingerprint (DRFP) instead hashes the substructures that change between reactants and products, matching learned fingerprints on yield and classification tasks without any training.[10] In parallel, a learned paradigm emerged from natural-language processing, in which transformer encoders are trained directly on reaction SMILES. Most prominent among these is the RXNFP reaction fingerprint, which produces continuous embeddings that separate reaction classes far more finely than traditional fingerprints.[11] These text-based models apply natural-language methods to reaction SMILES, but many distinct strings can encode the same reaction.

Graph-based encoders address this by operating directly on molecular connectivity, building on the message-passing neural networks that are prevalent in molecular property prediction.[2,12] For reactions, the condensed graph of reaction (CGR) superimposes the atom-mapped reactant and product graphs into a single pseudo-molecule whose atom and bond features encode both the reactant and product states, so that a graph neural network sees the transformation directly.[13,14] Heid and Green[6] paired the CGR with a D-MPNN and reported strong performance across activation energies, enthalpies, rate coefficients, and yields, an approach now available in the Chemprop package.[15] The CGR representation itself is architecture-general and can be combined with other graph neural networks like the graph isomorphism network (GIN)[12] and GIN with edge features (GINE).[16]

Transfer learning has become the dominant remedy for label scarcity in molecular property prediction. Models pre-trained on large unlabeled corpora, whether masked-token transformers over SMILES,[17] 3D-conformation encoders,[18] or graph networks trained with node-

and graph-level self-supervision,[16] learn representations that transfer to small downstream datasets and improve over training from scratch. A recent example is CheMeleon, which regresses a message-passing network onto deterministic molecular descriptors, a low-noise and fully computable signal, to learn transferable molecular representations without experimental labels.[19] This descriptor-regression strategy is appealing because it needs no experimental labels and no hand-designed pretext task as the supervision is generated deterministically from structure.

Reaction-level pre-training, by contrast, has been explored mostly in the text domain. Transformer foundation models such as ReactionT5, pre-trained on the Open Reaction Database,[20] and related multi-task models transfer effectively to yield and product prediction from limited data,[21,22] echoing the RXNFP fingerprints discussed above.[11] For graph-based reaction models the picture is sparser. Wen et al.[23] pre-trained a reaction GNN with unsupervised contrastive learning and improved downstream prediction on small datasets. Descriptor-based pre-training, despite its success on molecules, has not been adapted to reactions, and the CGR D-MPNN that is prevalent in reaction property prediction is still trained from scratch on each downstream task. Closing this gap requires two specific answers: what deterministic descriptor target captures a transformation rather than a single molecule, and whether such a target provides enough signal to pre-train a CGR encoder that transfers across diverse reaction properties.

In this work, we introduce CheMeleon-Rxn, an $\mathcal{O}(10\text{M})$-parameter CGR D-MPNN pre-trained on roughly two million reactions drawn predominantly from the Pistachio dataset.[24] The pre-training target is a set of reaction descriptors we call *Mordred-Rxn descriptors*. These are obtained by pooling per-molecule Mordred descriptors[25] over the reactant and product sides separately and combining them as $[\Sigma_{\text{reac}} \| (\Sigma_{\text{prod}} - \Sigma_{\text{reac}})]$ into a single dense, permutation- and count-invariant reaction descriptor set. The encoder is trained to regress its reaction embedding onto this target under a masked mean-squared-error loss, in which a random subset of descriptor dimensions is held out of the loss at each step. Specifically, in this

paper we (i) introduce the Mordred-Rxn descriptor construction and pre-training procedure; (ii) pre-train and evaluate three CGR message-passing architectures (D-MPNN,[3] GIN,[12] and GINE[16]); (iii) ablate the descriptor set, pooling operation, masking fraction, and fine-tuning strategy; (iv) demonstrate that CheMeleon-Rxn achieves best or statistically tied-for-best performance on seven of eight tasks spanning activation energies, reaction enthalpies, rate coefficients, and experimental yields; and (v) use a tree-based manifold approximation and projection (TMAP) of the learned fingerprints to show that pre-training on Mordred-Rxn descriptors produces chemically coherent reaction representations. Finally, all code and pre-trained checkpoints are made publicly available.

Results

CheMeleon-Rxn Improves Reaction Property Prediction

CheMeleon-Rxn was built in two stages that shared a single CGR encoder. In the pre-training stage, Mordred-Rxn descriptors were computed for each reaction in the corpus by summing per-molecule Mordred descriptors over the reactant and product sides and combining them with a difference-and-concatenate operator (Fig. 1a). The encoder was then trained to regress its reaction-level embedding onto those descriptor targets, with 70% of target dimensions randomly masked per step (Fig. 1b(i)). In the fine-tuning stage, the pre-trained encoder weights were loaded and a fresh prediction head was trained end-to-end on each downstream benchmark (Fig. 1b(ii)).

We first investigated whether pre-training improved downstream prediction relative to training from scratch and whether any improvement came from the pre-training itself rather than from the descriptors it used as a target. We therefore compared CheMeleon-Rxn against five baselines that separated these effects: (1) a D-MPNN trained from scratch; (2) the same D-MPNN with the reaction descriptors concatenated to its learned reaction embedding but without pre-training; (3) the descriptors alone fed to a multilayer perceptron (MLP); (4)

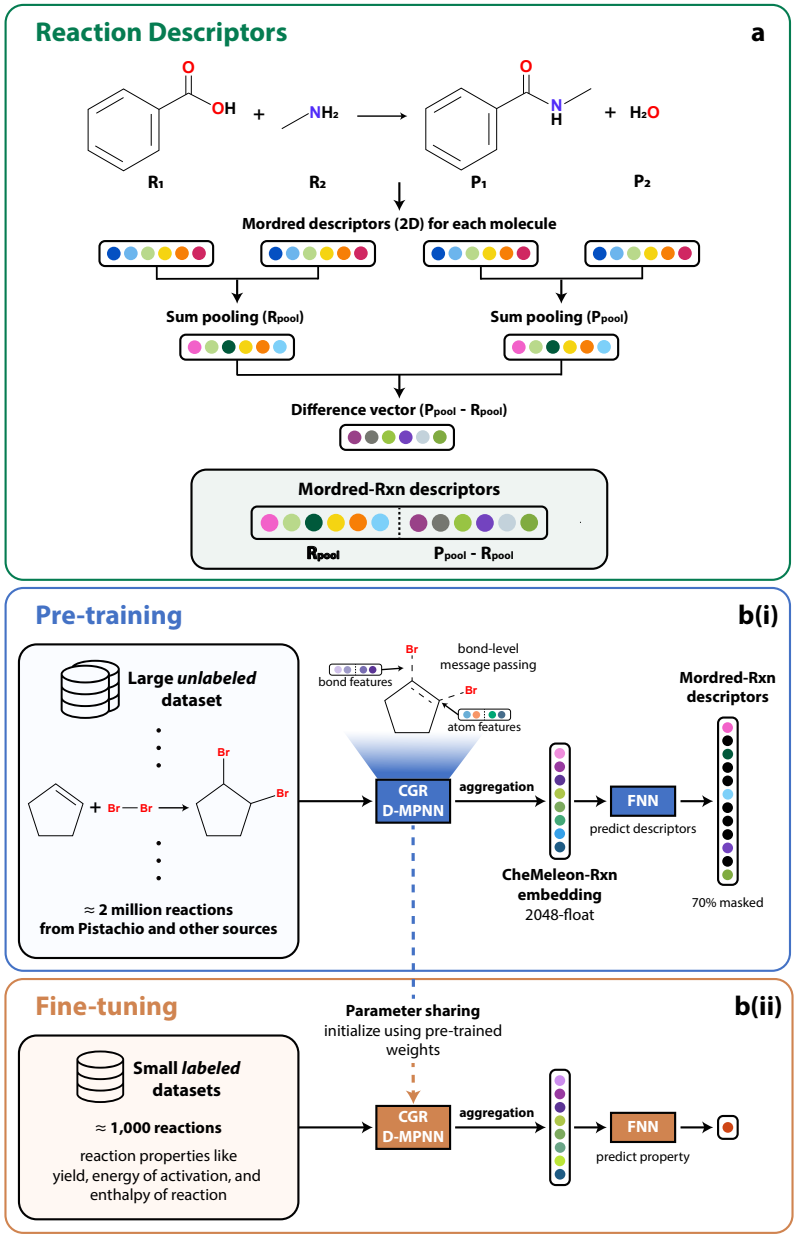


Figure 1: Overview of CheMeleon-Rxn. (**a**) Per-molecule Mordred descriptors are sum-pooled over each side of the reaction and combined with a difference-and-concatenate operator into the 3226-dimensional Mordred-Rxn descriptor. (**b(i)**) A CGR D-MPNN is pre-trained on 2 million reactions to regress its reaction embedding onto this descriptor target, with 70% of target dimensions held out of the loss per step. (**b(ii)**) The pre-trained encoder is then fine-tuned end-to-end with a fresh feed-forward neural network (FNN) on each downstream benchmark.

the Morgan difference fingerprint fed to an MLP; and (5) the pre-trained RXNFP reaction embedding[11] fed to an MLP. All six models were evaluated on eight downstream regression tasks (Table 4) under an identical five-split protocol with a three-model ensemble per split, and were ranked per task by a Tukey honestly-significant-difference (HSD) test on cross-split RMSE, following the comparison protocol of Ash et al.[26] (Fig. 2, Table 1).

Pre-training on classical reaction descriptors yields the lowest mean rank of the six models. CheMeleon-Rxn attains the best or statistically tied-for-best RMSE on seven of the eight tasks and a mean rank of 1.12, compared with 2.00 for the from-scratch D-MPNN and 2.75 or higher for every baseline that consumes the descriptors without pre-training on them (Table 1). The two descriptor-only baselines isolate the target signal: feeding the same Mordred reaction descriptor to an MLP, or concatenating it to the D-MPNN's learned embedding as additional molecular features, both trail the pre-trained model (mean ranks 2.88 and 2.75). The descriptors are therefore useful when they are distilled into the encoder's weights through pre-training, and less so when they are supplied directly as features. The two reaction-fingerprint baselines rank lowest (MorganDiff and RXNFP, mean ranks 5.62 and 5.38, no wins), and the gap is largest on the data-rich RDB7 tasks, where a fixed or externally learned fingerprint cannot adapt to the barrier-height regression the way a fine-tuned graph encoder can.

Table 1: Model comparison across 8 downstream tasks. Mean rank = mean over per-task ranks (Tukey-tied-with-best = 1); win rate = number of tasks on which the model is the best-mean or not significantly different from the best by Tukey HSD at $\alpha = 0.05$.

Model	Win rate (%)	Mean rank
CheMeleon-Rxn	87.5	1.12
D-MPNN	37.5	2.00
D-MPNN + Descriptors	25.0	2.75
Descriptors + MLP	25.0	2.88
IBM RXNFP + MLP	0.0	5.38
MorganDiff + MLP	0.0	5.62

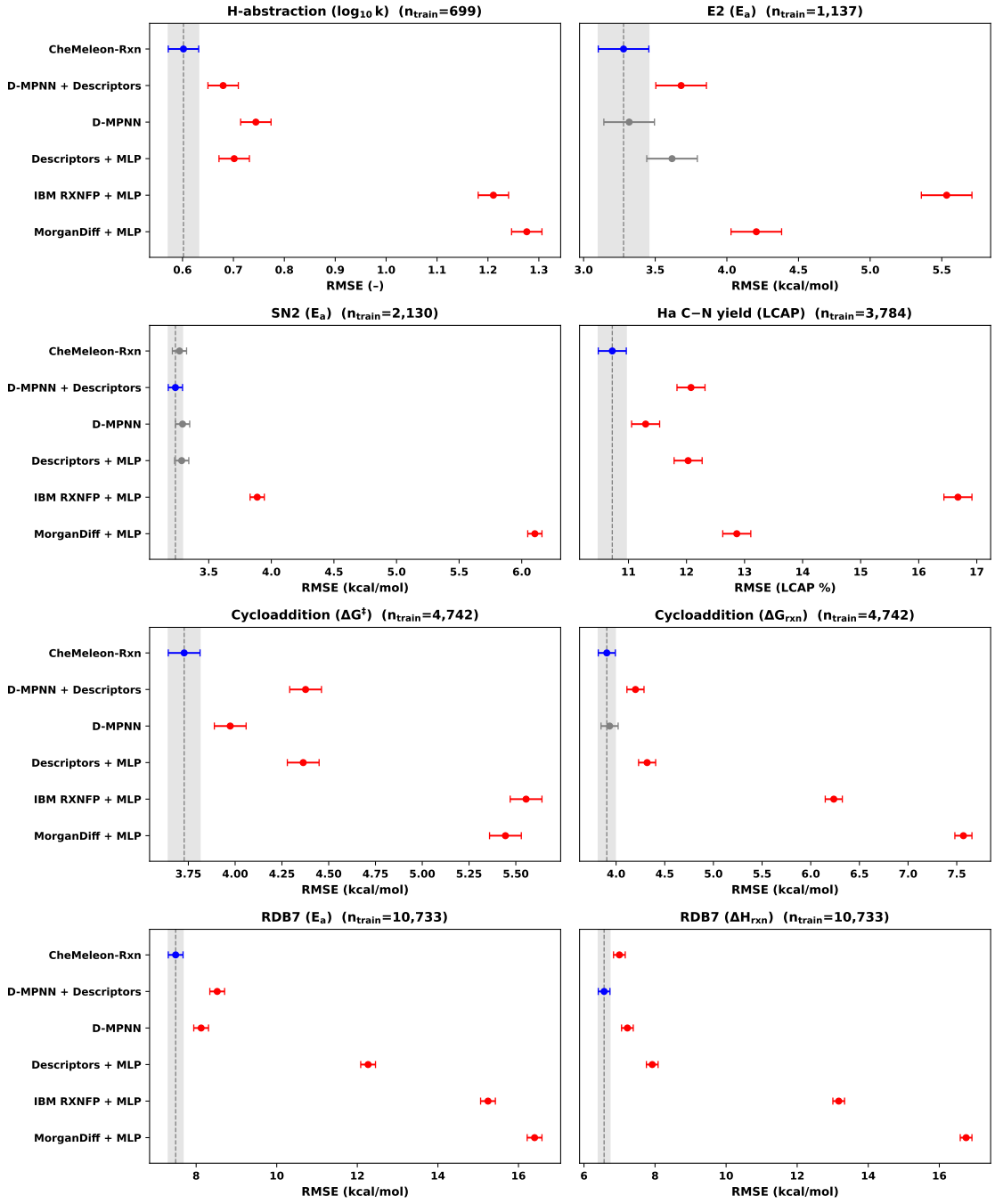


Figure 2: Descriptor-regression pre-training improves reaction property prediction across eight tasks. Each panel shows per-task test RMSE (mean and 95% simultaneous Tukey interval over five splits) for CheMeleon-Rxn and five baselines, ordered by dataset size. The best-mean model is blue, models not significantly different from it (Tukey HSD, $\alpha = 0.05$) are grey, and significantly worse models are red. CheMeleon-Rxn is best or tied-for-best on seven of eight tasks.

GNN backbone generalizability: edge features matter

To test whether the benefit of descriptor-regression pre-training is specific to the D-MPNN or a general property of the recipe, we repeated the full pre-train and fine-tune pipeline on two further graph neural network backbones, GIN[12] and GINE,[16] each in three configurations: trained from scratch, augmented with the reaction descriptors, and pre-trained with Mordred-Rxn descriptors. The resulting nine models were ranked per task exactly as before (Fig. 3, Table 2).

Pre-training improves model performance for the two backbones that consume edge features but not for GIN, whose messages depend only on node features. The pre-trained D-MPNN is the single best model, best-or-tied on all eight tasks, and the pre-trained GINE is close behind (mean rank 1.62, best-or-tied on seven of eight). On GIN, however, pre-training degrades performance: the pre-trained GIN has the worst mean rank of any model (6.88), below even the from-scratch GIN (5.38). The difference between GINE and GIN is that GINE conditions its messages on edge features while GIN does not, which points to bond-level information as the essential ingredient for downstream reaction property prediction. The benefit of descriptor-regression pre-training therefore extends to edge-aware backbones beyond the D-MPNN.

Table 2: Backbone-generality comparison. Three backbones (D-MPNN, GIN, GINE) each in three configurations (from-scratch, augmented with reaction descriptors as input, and pre-trained with CheMeleon-Rxn). Descriptor-regression pre-training helps when the encoder consumes edge features (D-MPNN, GINE) and hurts otherwise (GIN).

Model	Win rate (%)	Mean rank
D-MPNN (pre-trained)	100.0	1.00
D-MPNN	87.5	1.50
GINE (pre-trained)	87.5	1.62
D-MPNN + Descriptors	62.5	2.62
GINE + Descriptors	37.5	4.50
GIN + Descriptors	37.5	4.75
GIN	37.5	5.38
GINE	25.0	5.62
GIN (pre-trained)	25.0	6.88

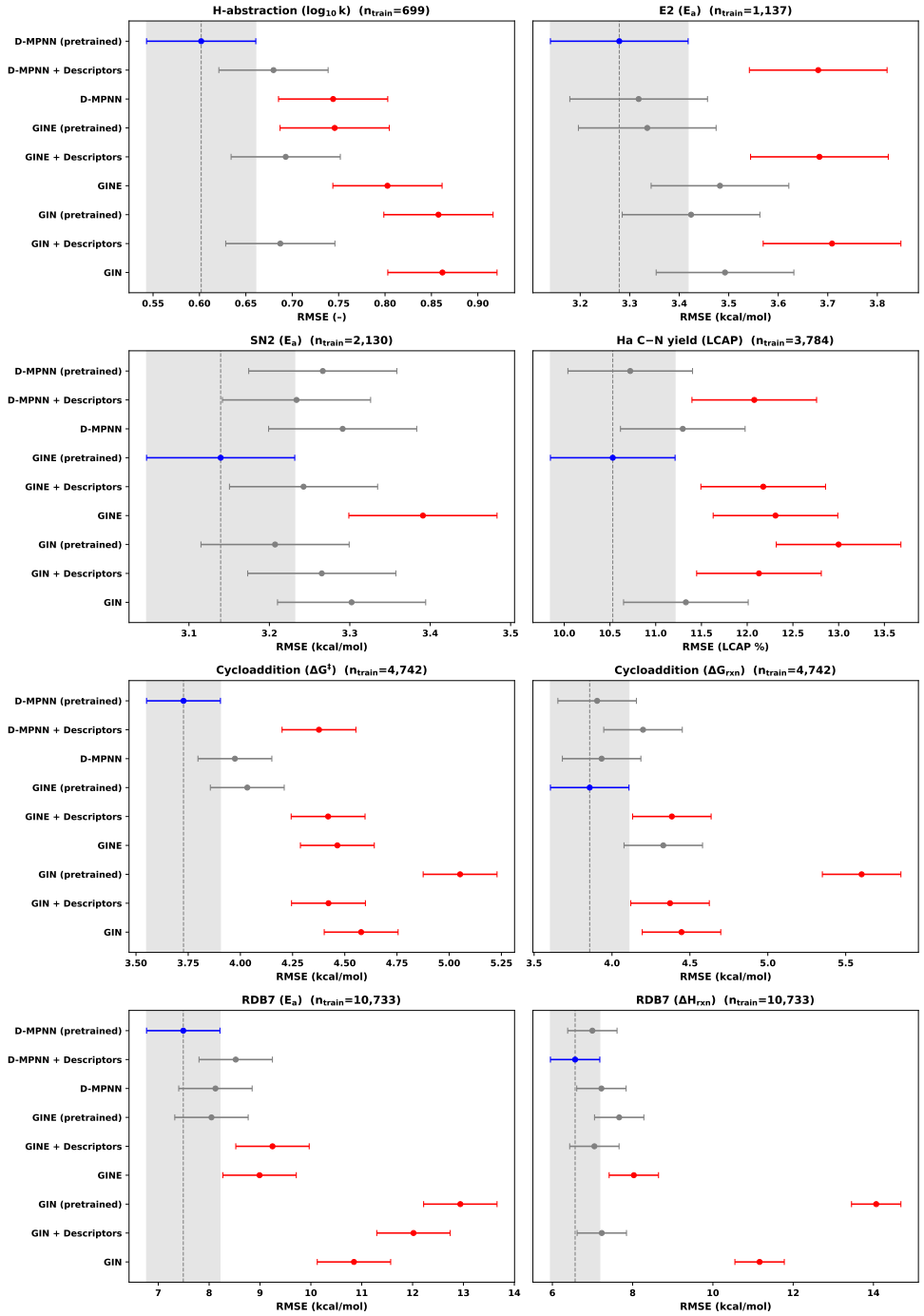


Figure 3: Descriptor-regression pre-training helps edge-aware backbones and hurts node-only ones. Per-task test RMSE (mean and 95% simultaneous Tukey interval over five splits) for three backbones (D-MPNN, GINE, GIN), each from scratch, augmented with the reaction descriptors as input, and pre-trained with Mordred-Rxn descriptors. Colors as in Fig. 2. Pre-training strengthens the edge-aware D-MPNN and GINE but degrades the node-only GIN below its from-scratch baseline.

Based on these results, we fixed the D-MPNN as the reference CheMeleon-Rxn encoder for all remaining experiments. It achieves the best aggregate performance of the nine configurations, and the statistically competitive pre-trained GINE sits within one rank of it, so the choice also rests on two practical considerations: the D-MPNN is the established convention for CGR-based reaction property prediction,[6] and it is directly supported by the widely used Chemprop package,[15,27] lowering the barrier to adopting the pre-trained model.

Effect of Descriptor Set

Having fixed the encoder, we asked how much the choice of descriptor target matters, comparing three pre-training targets built from different classical descriptor sets and pooling operations: the reference Mordred descriptors summed over each side, the same descriptors averaged rather than summed, and the smaller RDKit-2D descriptor set (Table 3). Downstream accuracy is largely unchanged across these choices. Sum and mean pooling of the Mordred descriptors differ by no statistically detectable margin on seven of the eight benchmarks, the exception being the H-abstraction rate task, where sum pooling is better. Notably, pre-training with RDKit-2D descriptors, a smaller and considerably cheaper descriptor set, still transfers successfully to reaction property prediction on seven of the eight benchmarks. We retained the summed Mordred target as the reference configuration, both because no target significantly outperforms it on any task and because summation mirrors the count-sensitive reactant/product featurization used by CGR-based models.[6] For the masking fraction, the 70% setting is best or statistically tied-for-best on seven of the eight tasks by the Tukey test (Supporting Information, Fig. S1), and we used it for the reference configuration.

Effect of Fine-Tuning Strategy

Finally, we asked how the pre-trained encoder is best transferred to a downstream task by comparing five fine-tuning schedules that share the pre-trained weights and differ only in how

Table 3: Descriptor-target choice. Per-task RMSE (mean ± std over five random splits, three inits per split). Best per row (and any entries not significantly different from it by Tukey HSD at $\alpha = 0.05$) underlined.

Task	CheMeleon-Rxn	D-MPNN (pre-trained) Mordred-mean	D-MPNN (pre-trained) RDKit2D
H-abstraction ($\log_{10} k$)	0.602 ± 0.021	0.738 ± 0.046	0.653 ± 0.028
E2 (E_a)	3.279 ± 0.149	3.247 ± 0.145	3.223 ± 0.176
SN2 (E_a)	3.266 ± 0.052	3.260 ± 0.032	3.293 ± 0.084
Ha C−N yield (LCAP)	10.721 ± 0.241	10.653 ± 0.236	10.700 ± 0.192
Cycloaddition ($\Delta G^{\ddagger}$)	3.728 ± 0.053	3.716 ± 0.070	3.731 ± 0.073
Cycloaddition (ΔG_{rxn})	3.905 ± 0.061	3.910 ± 0.106	3.927 ± 0.055
RDB7 (E_a)	7.492 ± 0.053	7.504 ± 0.087	7.863 ± 0.046
RDB7 (ΔH_{rxn})	6.997 ± 0.120	6.971 ± 0.161	7.104 ± 0.126

long the encoder is held frozen before it is unfrozen: (1) fully trainable from the first epoch, (2) frozen for the first 10 epochs, (3) frozen for the first 20 epochs, (4) frozen for the first 50 epochs, and (5) frozen throughout (training only the FNN head on the fixed pre-trained features). The pre-trained encoder behaves as a strong initialization rather than as a fixed feature extractor. Fully trainable fine-tuning is best or tied-for-best on every task, whereas the fully frozen encoder is significantly worse on seven of eight tasks. The loss curves in Fig. 4 highlight the underlying mechanism: on the data-rich tasks the frozen-encoder validation loss plateaus far above every trainable variant, indicating that the pre-trained representation, though a useful starting point, must continue to adapt to reach competitive accuracy. Brief warm-up freezing for 10 or 20 epochs remains statistically tied with full fine-tuning on most tasks, while the transient loss spike at the unfreezing epoch and the degradation of the 50-epoch schedule on the barrier tasks indicate that longer freezing delays the adaptation the encoder ultimately requires. We therefore fine-tuned the full network by default.

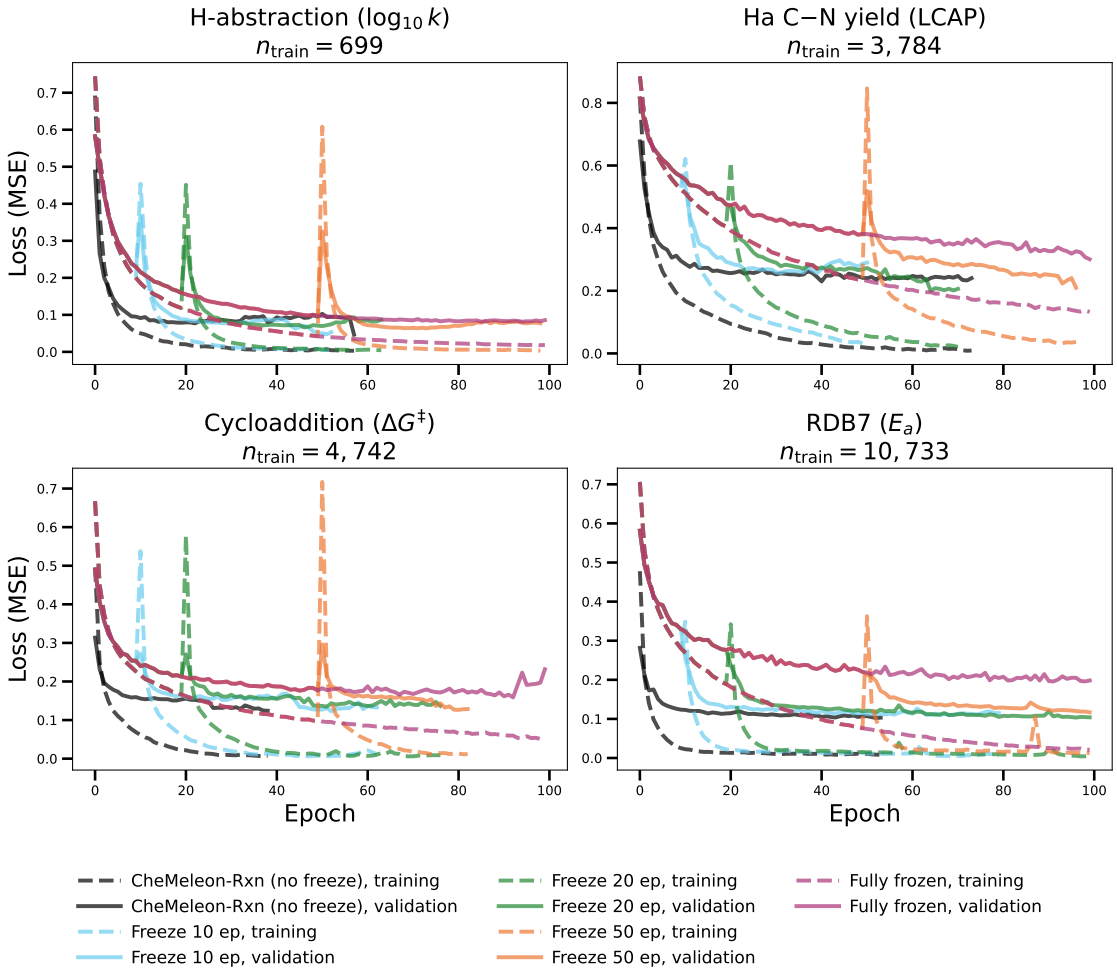


Figure 4: The pre-trained encoder is a strong initialization, not a fixed feature extractor. Training (dashed) and validation (solid) mean-squared-error loss versus epoch, averaged over five splits and three seeds, for five fine-tuning schedules on four representative tasks; panel titles give n_{train}, the combined train and validation set size (the train fold alone is listed as N_{train} in Table 4). The schedules share the CheMeleon-Rxn checkpoint and differ only in how many initial epochs the encoder is frozen (0, 10, 20, 50, or all 100). The fully frozen encoder (only the FNN head trained) plateaus well above the trainable variants on the data-rich Ha C−N and RDB7 tasks; spikes mark the unfreezing epoch.

CheMeleon-Rxn Learns a Chemically Coherent Reaction Representation

Beyond downstream accuracy, we asked whether the pre-trained encoder organizes reaction space in a chemically meaningful way. We embedded fifty thousand Pistachio reactions with the CheMeleon-Rxn fingerprint, the 2048-dimensional reaction embedding produced by the pre-trained D-MPNN encoder before the prediction head, projected them with a TMAP layout, and colored the tree by the eleven NameRxn reaction superclasses (Fig. 5).[24] Chemically related transformations occupy coherent regions of the tree, and the coherence sharpens with granularity. The fraction of a reaction's ten nearest TMAP neighbors that share its label rises from the 11-way superclass level to the 1365-way sub-class level, and at every scale it far exceeds the independent random-label expectation $\sum_i p_i^2$ for label frequencies p_i (Supporting Information, Table S11). Because the neighborhoods are taken in the projected layout rather than in the 2048-dimensional embedding, these purities characterize the fingerprint as visualized. The purity is most striking at fine granularity: a reaction's immediate neighbors are overwhelmingly the same specific transformation type, even though the number of possible sub-classes is large. Coarse-level purity is more modest because the NameRxn superclass axis groups reactions by synthetic purpose rather than by the structural change itself, mixing chemistries whose condensed graphs differ significantly. The TMAP embedding of the fingerprint therefore organizes reactions by the structure of the transformation itself, the property that our descriptor-regression objective rewards.

Discussion

Descriptor-regression pre-training improves reaction property prediction with graph neural networks. CheMeleon-Rxn achieves the best or statistically tied-for-best mean RMSE on seven of the eight tasks, which span reaction types as different as gas-phase barrier heights and high-throughput experimental yields. The same recipe works across these chemically

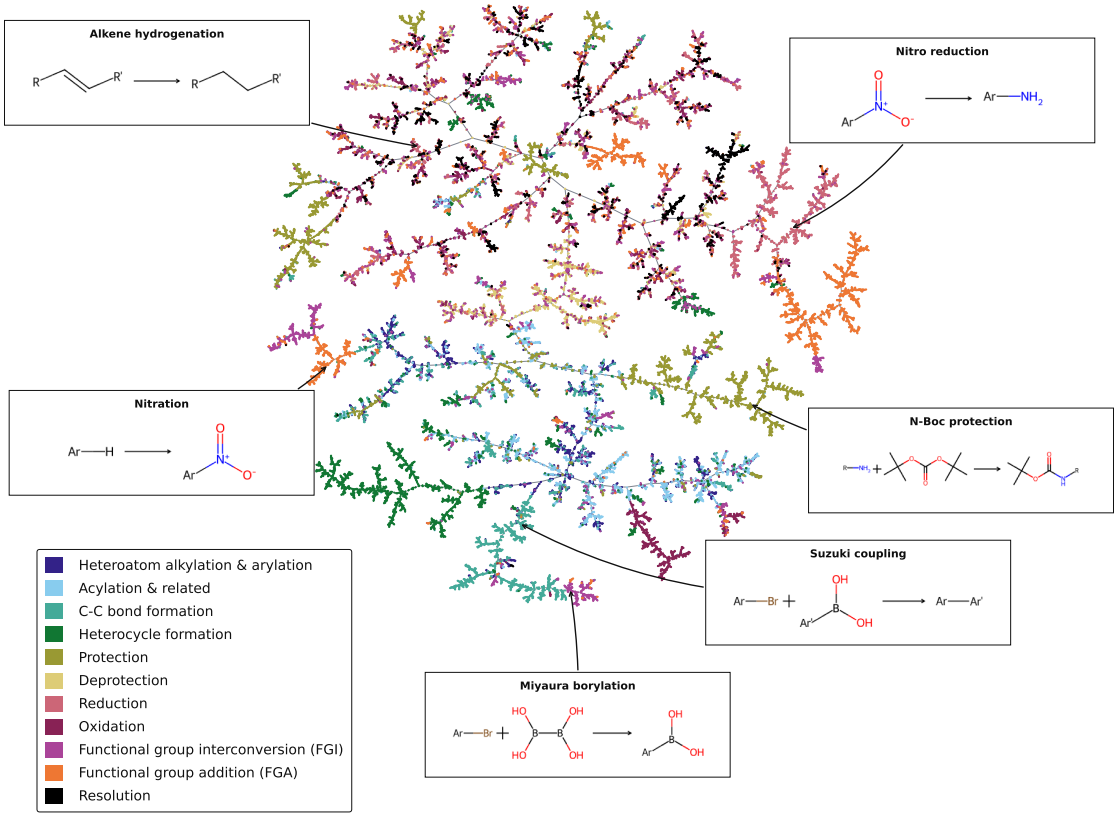


Figure 5: The CheMeleon-Rxn fingerprint induces a chemically coherent reaction space. TMAP layout of fifty thousand Pistachio reactions embedded with the pre-trained D-MPNN fingerprint, colored by the eleven NameRxn superclasses, with callouts for six representative regions. The reaction schemes are textbook archetypes with generic R/Ar groups, not verbatim records, as Pistachio is proprietary.[24]

diverse tasks without task-specific descriptor engineering or a hand-designed pretext task.

Ablation studies confirm that the design choices are secondary rather than critical. How the per-molecule descriptors are pooled over each side of the reaction makes little difference: sum- and mean-pooling of the Mordred descriptors produce statistically equivalent results on seven of the eight tasks, with sum pooling ahead only on the H-abstraction rate task. Similarly, the choice of masking fraction is not critical: the 70% setting is best or statistically tied-for-best on seven of the eight tasks (SI Fig. S1), and we used it for the reference configuration. The choice of descriptor set is also less consequential than expected: pre-training with the smaller RDKit-2D set (210 per-molecule descriptors) still transfers to seven of the eight downstream tasks, and only the RDB7 activation energy task favors the richer Mordred set (1613 descriptors). Together, these results suggest that a wide variety of classical molecular descriptors, pooled and differenced across reaction sides, can serve as a sufficient pre-training signal for reaction graph encoders.

The one design axis that does matter is the architectural choice of whether the encoder consumes edge features. Pre-training improves D-MPNN and GINE, which both incorporate bond-level information into their message-passing updates, and degrades GIN, which ignores edge attributes entirely. We attribute this to the pre-training target, which encodes bond-level chemistry that a node-only message function cannot represent. Hence, pre-training drives GIN to a worse starting point than random initialization. Because a reaction is defined by how its bonds reorganize, an encoder blind to bond features is poorly suited to reaction property prediction, whether trained from scratch or pre-trained. This is consistent with prior evidence that the fidelity of bond featurization governs the representational power of CGR-based reaction encoders.[28]

The improvement is not explained by model capacity. A same-architecture D-MPNN (message-passing hidden 2048, depth 6) trained from random initialization, a capacity-matched scratch baseline, is never significantly better than CheMeleon-Rxn and is substantially worse on several tasks (see Supporting Information, Table S6). The gain is therefore

attributable to the pre-training signal rather than to the larger network. CheMeleon-Rxn trails on only one of the eight tasks, the RDB7 reaction enthalpy, where it sits just behind the D-MPNN that receives the reaction descriptors as additional molecular features.

Looking ahead, there are several natural, future extensions of this work. The descriptor-regression objective is agnostic to the source of reactions: any collection of atom-mapped reaction SMILES yields a valid pre-training corpus, and reaction databases continue to grow rapidly. Including reaction conditions such as solvents, catalysts, and reagents in both the descriptor target and the CGR featurization may be a direct path toward improving yield prediction, where reagents play a decisive role. The descriptor target itself can be made richer by incorporating computed quantum-chemical fragment descriptors,[29] which preserve the low-noise, fully calculable nature of the pre-training signal. Finally, that the benefit transfers to both edge-aware backbones we tested (D-MPNN and GINE) suggests the recipe is not D-MPNN-specific and is a candidate for other edge-conditioned reaction GNNs, though this remains to be tested more broadly. All pre-trained model weights are publicly available on Hugging Face and the full codebase is open-source (URL to be added on release).

Methods

Datasets

Pre-Training Corpus

The pre-training corpus pools eight atom-mapped reaction sources: the gas-phase elementary-reaction sets RGD1,[30] RDB7,[31] and the E2, SN2, and cycloaddition barrier benchmarks,[32,33] the Ha C−N and H-abstraction rate datasets,[9,34] and a large sample of patent reactions from Pistachio (NextMove Software), text-mined from the chemical patent literature.[24] Downstream labels are discarded and only reaction structures are retained. Every source is required to be natively atom-mapped, and each reaction is reduced to reactants and products

only, with agents and reagents stripped. Reactions are restricted to the organic main-group elements (H, B, C, N, O, F, Si, P, S, Cl, Br), and the reactant side is capped at 40 heavy atoms. A cap of 50,000 reactions per NameRxn class (applied at the granular reaction-type level) is imposed on Pistachio so that the most common patent reaction types do not dominate the corpus. The labeled benchmark sources contribute their training and validation reactions as including these downstream training reactions in the pre-training corpus is well supported in the transfer-learning literature.[35] Test reactions from all six downstream benchmarks are identified using map-independent canonical keys and withheld in both forward and reverse directions before pooling. Reactions are deduplicated on those canonical keys, and each retained reaction is augmented with its reverse (products $\rightarrow$ reactants), so that the encoder sees both directions of every transformation. The pooled, deduplicated corpus of 2,215,990 reactions is randomly subsampled to a target of two million reactions. This round target is a heuristic choice and we would not expect pre-training on the full $\sim$2.2 million reaction corpus to materially change the downstream results. These two million reactions comprise around 1.4 million unique molecules, and the classical descriptor target is computed and cached once per unique molecule rather than once per reaction.

Downstream Benchmarks

Six labeled datasets serve as fine-tuning benchmarks, all also contributing their non-test reactions to the pre-training corpus above (Table 4). The E2 and SN2 barrier sets[32] contain 1,264 and 2,361 DFT-computed gas-phase reactions. The cycloaddition dataset[33] (5,269 reactions) provides two targets: activation free energy ($\Delta G^{\ddagger}$) and reaction free energy (ΔG_{rxn}). The RDB7 dataset[31,36] (11,926 reactions) reports activation energies and reaction enthalpies using high-quality quantum chemistry calculations. The Ha C$-$N coupling dataset[9] (4,204 reactions) reports experimental yields, measured as liquid chromatography area percent (LCAP), from a high-throughput Pd-catalyzed cross-coupling screen. The LogRate dataset[34] (778 reactions) contains $\log_{10}$ rate coefficients for gas-phase hydrogen-abstraction reactions. For

Table 4: Datasets used for pre-training and fine-tuning. Pre-training counts are the pooled composition post-augmentation (each reaction and its reverse included), before the pool is randomly subsampled to the two-million-reaction training corpus; the per-source rows therefore sum to the pooled total rather than to the trained corpus size. †Pistachio is a proprietary licensed dataset (NextMove Software) and cannot be redistributed. Benchmark N_{train} refers to the first split.

Dataset	Target(s)	Pre-training	Benchmark	N_{pretrain}	N_{train}	Source
Pistachio†	n/a	✓		1,888,804	n/a	24
RGD1	n/a	✓		283,507	n/a	30
E2	activation energy (E_a)	✓	✓	2,202	1,014	32
SN2	activation energy (E_a)	✓	✓	2,847	1,902	32
CycloAdd	activation free energy ($\Delta G^{\ddagger}$), reaction free energy (ΔG_{rxn})	✓	✓	9,404	4,215	33
RDB7	activation energy (E_a), reaction enthalpy (ΔH_{rxn})	✓	✓	20,936	9,539	31
Ha C−N	yield (LCAP)	✓	✓	6,894	3,364	9
LogRate	rate coefficient ($\log_{10} k$)	✓	✓	1,396	621	34
Pooled total				2,215,990		
Pre-training corpus				**2,000,000**		

each benchmark we construct five splits with a fixed test set (≈10%, held identical across all five splits) and, over the remaining ≈90%, reshuffled train/validation partitions corresponding to an ≈80/10/10 train/validation/test split of the full dataset. Splitting is performed at the group level, keeping each reaction and its reverse (if it exists) together. This prevents leakage of structures to the test set. All splits are shared across every model.

CGR Representation and Reaction Descriptors

Each reaction is represented as a condensed graph of reaction (CGR), the superposition of the atom-mapped reactant and product graphs into a single pseudo-molecule.[6,13,37] Both atoms and bonds carry features describing the reactant and product states, so the graph encodes the full transformation rather than either side alone. We featurize reactions with Chemprop v2.3.1[15] in its REAC_DIFF mode, which represents each atom and bond by concatenating

its reactant-state features with the difference between its product-state and reactant-state features. Explicit atom-mapped hydrogens are retained (`KEEP_H = True`).

Per-molecule Mordred descriptors[25] (1613 2D descriptors, calculated on RDKit[38] molecular graphs) or RDKit-2D descriptors (210 2D descriptors) are computed for every unique molecule in the pre-training corpus and benchmark datasets. To construct a reaction-level descriptor target, each side is pooled independently by an element-wise sum over all constituent molecules' descriptor vectors (Mordred-sum, the default) or an element-wise mean (Mordred-mean). The two pooled vectors are combined with a difference-and-concatenate operator,

$$\mathbf{d}_{\text{rxn}} = \left[\Sigma_{\text{reac}} \,\|\, (\Sigma_{\text{prod}} - \Sigma_{\text{reac}})\right], \tag{1}$$

yielding a 3226-dimensional reaction descriptor for Mordred or a 420-dimensional descriptor for RDKit-2D. This operator is analogous to the REAC_DIFF featurization mode of Chemprop, which feeds the same difference structure as graph-level input to the CGR encoder.[6] Descriptor values are standardized using Welford's online mean and variance computed over the pre-training corpus, and winsorized by clipping to ± 6 standard deviations before z-scoring.

Pre-Training

The CheMeleon-Rxn D-MPNN encoder uses message-passing hidden size 2048 and depth 6 with mean aggregation. A single-layer feed-forward neural network (FNN) prediction head (hidden size 1024, LeakyReLU activation) is attached on top: during pre-training it projects onto the full descriptor target, and at fine-tuning time onto the single downstream property. The deployed fine-tuned model has $\mathcal{O}(10\text{M})$ parameters. The same encoder architecture is used for the GIN[12] and GINE[16] backbone variants in the analysis experiments.

Pre-training minimizes a masked descriptor regression loss. For each training reaction, a random fraction f of the descriptor dimensions is masked and the model predicts only

the visible dimensions; the loss is the mean squared error over those unmasked targets (RandomDropoutMSE loss[19]). The reference configuration uses $f = 0.70$ (70% masked per step). Learning rate follows a Noam warm-up schedule with initial rate 5×10^{-5}, peak rate 3×10^{-4}, final rate 5×10^{-5}, and 10 warm-up epochs over a maximum of 100 epochs. We use early stopping based on the validation loss, terminating training when the validation loss fails to improve for 10 consecutive epochs (patience = 10). The batch size is 1024 reactions. The pre-training corpus is split 90/10 (train/val) by a group-aware scheme that keeps reactions sharing molecular species in the same partition. For pre-training specifically, featurizing the reaction graphs on the fly over the two-million-reaction corpus is made tractable by the accelerated cuik-molmaker backend.[39]

Fine-Tuning

To fine-tune CheMeleon-Rxn, the CGR encoder is initialized with the pre-trained weights and a fresh single-layer FNN prediction head (hidden size 1024, LeakyReLU activation) is attached. We fine-tune the full network end-to-end with a flat learning rate of 10^{-4}, batch size 64, and a maximum of 100 epochs with patience 10. The gradual-unfreeze experiments follow the same setup but hold the encoder frozen for the first $k \in \{10, 20, 50, 100\}$ epochs before releasing it. The learning-rate schedule differs by model group and is summarized in the Supporting Information (Table S3): pre-trained encoders use the flat rate above, whereas the randomly initialized GNN baselines use Chemprop's default Noam schedule.

For each benchmark dataset, five splits are used as described above; within each split, three models are trained with distinct random seeds and their test-set predictions are averaged to produce one ensemble RMSE per split, yielding five Tukey-HSD replicates per task. Pairwise differences in per-split RMSE are tested with the Tukey HSD test at the 5% level; the mean rank reported in the summary tables counts a model's rank as 1 on any task where it is best-mean or statistically tied-with-best.

Baselines

Five baselines frame the performance of CheMeleon-Rxn. The **D-MPNN (scratch)** is a Chemprop CGR D-MPNN with the default architecture (hidden size 300, depth 3) trained from random initialization, following the identical fine-tuning code path as CheMeleon-Rxn to isolate the benefit of pre-training; this is the standard Chemprop-CGR starting point without pre-training.[27] The **D-MPNN + Descriptors** baseline attaches the 3226-dimensional Mordred reaction descriptor to the graph embedding after pooling and before the FNN head (head hidden size 512, depth 2), testing whether the descriptor information helps when supplied as direct input rather than through pre-training. The **Descriptors + MLP** baseline discards the graph encoder and trains an MLP with two hidden layers (hidden size 300) on the Mordred reaction descriptor alone, isolating the downstream predictive value of the classical descriptor. The **MorganDiff + MLP** baseline replaces the Mordred descriptor with a Morgan count fingerprint difference (radius 3, 1024 bits, products minus reactants), fed into the same two-layer MLP; this is a Morgan-based variant of the difference reaction fingerprint of Schneider et al.,[40] a simple and widely used fixed reaction representation. The **IBM RXNFP + MLP** baseline uses the pre-trained BERT reaction fingerprint of Schwaller et al.[11] (256-dimensional continuous embedding, fixed, from the publicly available checkpoint) as the sole input to the same two-layer MLP; this is a widely used learned NLP-based reaction embedding. The GNN baselines (D-MPNN scratch and D-MPNN + Descriptors, along with the GIN and GINE variants of the backbone-generality analysis) are trained with Chemprop v2's default optimizer and learning-rate schedule,[15] while the MLP baselines use a flat 10^{-4} learning rate as in the fine-tuning setup above. All baselines share the five-split, three-seed ensembling and evaluation protocol.

Declaration of Competing Interest

The authors declare no conflicts of interest. A.S.Z., C.F., Y.Z., E.M., A.C., A.G., and H.P., when performing the research, worked for Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

References

(1) David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 2224–2232, 2015.

(2) Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pages 1263–1272, 2017.

(3) Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

(4) Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018. doi: 10.1039/c7sc02664a.

(5) Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet – a deep learning architecture for molecules and

materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

(6) Esther Heid and William H. Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *J. Chem. Inf. Model.*, 62(9):2101–2110, 2022. doi: 10.1021/acs.jcim.1c00975.

(7) David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

(8) Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A structure-based platform for predicting chemical reactivity. *Chem*, 6 (6):1379–1390, 2020. doi: 10.1016/j.chempr.2020.02.017.

(9) Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018. doi: 10.1126/science.aar5169.

(10) Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1(2):91–97, 2022. doi: 10.1039/d1dd00006c.

(11) Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.*, 3(2):144–152, 2021. doi: 10.1038/s42256-020-00284-w.

(12) Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019. ICLR 2019.

(13) Alexandre Varnek, Denis Fourches, Frank Hoonakker, and Vitaly P. Solov'ev. Substructural fragments: an universal language to encode reactions, molecular and

supramolecular structures. *J. Comput.-Aided Mol. Des.*, 19:693–703, 2005. doi: 10.1007/s10822-005-9008-0.

(14) Frank Hoonakker, Nicolas Lachiche, Alexandre Varnek, and Alain Wagner. A representation to apply usual data mining techniques to chemical reactions: Illustration on the rate constant of SN2 reactions in water. *Int. J. Artif. Intell. Tools*, 20(2):253–270, 2011. doi: 10.1142/S0218213011000140.

(15) David E. Graff, Nathan K. Morgan, Jackson W. Burns, Anna C. Doner, Brian Li, Shih-Cheng Li, Joel Manu, Angiras Menon, Hao-Wei Pang, Haoyang Wu, Akshat Shirish Zalte, Jonathan W. Zheng, Connor W. Coley, William H. Green, and Kevin P. Greenman. Chemprop v2: An efficient, modular machine learning package for chemical property prediction. *J. Chem. Inf. Model.*, 66(1):28–33, 2026. doi: 10.1021/acs.jcim.5c02332.

(16) Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

(17) Benedek Fabian, Thomas Edlich, Hélèna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020. NeurIPS 2020 ML for Molecules Workshop.

(18) Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3d molecular representation learning framework. ChemRxiv preprint, 2023.

(19) Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H. Green. Deep learning foundation models for low-data regimes from classical molecular descriptors. *J. Chem. Inf. Model.*, 2026. doi: 10.1021/acs.jcim.6c01546. Advance online publication.

(20) Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The open reaction database. *J. Am. Chem. Soc.*, 143(45):18820–18826, 2021. doi: 10.1021/jacs.1c09820.

(21) Tatsuya Sagawa and Ryosuke Kojima. Reactiont5: A pre-trained transformer model for accurate chemical reaction prediction with limited data. *Journal of Cheminformatics*, 17(1):126, 2025. doi: 10.1186/s13321-025-01075-4.

(22) Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387, 2022. doi: 10.1021/acs.jcim.1c01467.

(23) Mingjian Wen, Samuel M. Blau, Xiaowei Xie, Shyam Dwaraknath, and Kristin A. Persson. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.*, 13:1446–1458, 2022. doi: 10.1039/d1sc06515g.

(24) NextMove Software. Pistachio. `https://www.nextmovesoftware.com/pistachio.html`, 2025.

(25) Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *J. Cheminf.*, 10(1):4, 2018. doi: 10.1186/s13321-018-0258-y.

(26) Jeremy R. Ash, Cas Wognum, Raquel Rodríguez-Pérez, Matteo Aldeghi, Alan C. Cheng, Djork-Arné Clevert, Ola Engkvist, Cheng Fang, Daniel J. Price, Jacqueline M. Hughes-Oliver, and W. Patrick Walters. Practically significant method comparison protocols for machine learning in small molecule drug discovery. *J. Chem. Inf. Model.*, 65(18):9398–9411, 2025. doi: 10.1021/acs.jcim.5c01609.

(27) Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *J. Chem. Inf. Model.*, 64 (1):9–17, 2024. doi: 10.1021/acs.jcim.3c01250.

(28) Akshat Shirish Zalte, Hao-Wei Pang, Anna C. Doner, and William H. Green. RIGR: Resonance-invariant graph representation for molecular property prediction. *J. Chem. Inf. Model.*, 65(20):10832–10843, 2025. doi: 10.1021/acs.jcim.5c00495.

(29) Shih-Cheng Li, Haoyang Wu, Angiras Menon, Kevin A. Spiekermann, Yi-Pei Li, and William H. Green. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *J. Am. Chem. Soc.*, 146(33):23103–23120, 2024. doi: 10.1021/jacs.4c04670.

(30) Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A. Ogunfowora, Srinivasa Garimella, Olexandr Isayev, and Brett M. Savoie. Comprehensive exploration of graphically defined reaction spaces. *Sci. Data*, 10:294, 2023. doi: 10.1038/s41597-023-02043-z.

(31) Kevin Spiekermann, Lagnajit Pattanaik, and William H. Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Sci. Data*, 9(1):417, 2022. doi: 10.1038/s41597-022-01529-6.

(32) Guido Falk von Rudorff, Stefan N. Heinen, Marco Bragato, and O. Anatole von Lilienfeld. Thousands of reactants and transition states for competing E2 and SN2 reactions. *Mach. Learn.: Sci. Technol.*, 1(4):045026, 2020. doi: 10.1088/2632-2153/aba822.

(33) Thijs Stuyver, Kjell Jorner, and Connor W. Coley. Reaction profiles for quantum chemistry-computed [3 + 2] cycloaddition reactions. *Sci. Data*, 10(1):66, 2023. doi: 10.1038/s41597-023-01977-8.

(34) Calvin Pieters and Alon Grinberg Dana. Learning rates: predicting rate coefficients for

hydrogen abstraction reactions. *Digital Discovery*, 5(7):3044–3065, 2026. doi: 10.1039/d6dd00113k.

(35) Kundan Krishna, Saurabh Garg, Jeffrey P. Bigham, and Zachary C. Lipton. Downstream datasets make surprisingly good pretraining corpora, 2022.

(36) Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. *J. Phys. Chem. A*, 126(25): 3976–3986, 2022. doi: 10.1021/acs.jpca.2c02614.

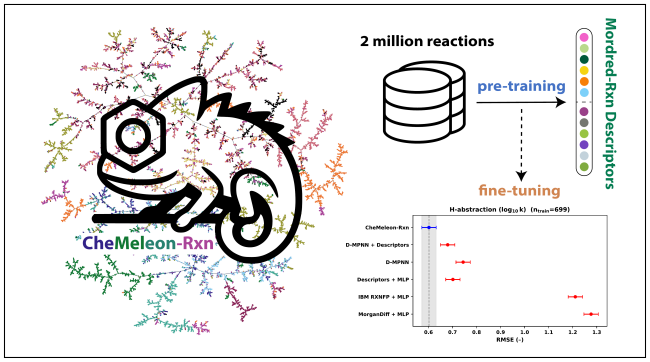
(37) Ramil I. Nugmanov, Ravil N. Mukhametgaleev, Tagir Akhmetshin, Timur R. Gimadiev, Valentina A. Afonina, Timur I. Madzhidov, and Alexandre Varnek. CGRtools: Python library for molecule, reaction, and condensed graph of reaction processing. *J. Chem. Inf. Model.*, 59(6):2516–2521, 2019. doi: 10.1021/acs.jcim.9b00102.

(38) Greg Landrum et al. RDKit: Open-source cheminformatics. URL `https://www.rdkit.org`. Version 2026.03.4.

(39) NVIDIA BioNeMo. cuik-molmaker: Fast featurization of molecules and reactions. `https://github.com/NVIDIA-BioNeMo/cuik-molmaker/`, 2025. Accessed 2026-08-07.

(40) Nadine Schneider, Daniel M. Lowe, Roger A. Sayle, and Gregory A. Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.*, 55(1):39–53, 2015. doi: 10.1021/ci5006614.

TOC Graphic


CheMeleon-Rxn
2 million reactions
pre-training
fine-tuning
Mordred-Rxn Descriptors
H-abstraction (log10 k) (ntrain=699)
CheMeleon-Rxn
D-MPNN + Descriptors
D-MPNN
Descriptors + MLP
IBM RXNFP + MLP
MorganDiff + MLP
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
RMSE (-)