Towards out-of-distribution generalizable predictions of chemical kinetic properties

Zihao Wang*
CSE, HKUST
zwanggc@cse.ust.hk

Yongqiang Chen*
CSE, CUHK
yqchen@cse.cuhk.edu.hk

Yang Duan, Weijiang Li
CS, UIUC
{yangd4,wl13}@illinois.edu

Bo Han
CS, HKBU
bhanml@comp.hkbu.edu.hk

James Cheng
CSE, CUHK
jcheng@cse.cuhk.edu.hk

Hanghang Tong
CS, UIUC
htong@illinois.edu

Abstract

Machine Learning (ML) techniques have found applications in estimating chemical kinetic properties. With the accumulated drug molecules identified through "AI4drug discovery", the next imperative lies in AI-driven design for high-throughput chemical synthesis processes, with the estimation of properties of unseen reactions with unexplored molecules. To this end, the existing ML approaches for kinetics property prediction are required to be Out-Of-Distribution (OOD) generalizable. In this paper, we categorize the OOD kinetic property prediction into three levels (structure, condition, and mechanism), revealing unique aspects of such problems. Under this framework, we create comprehensive datasets to benchmark (1) the state-of-the-art ML approaches for reaction prediction in the OOD setting and (2) the state-of-the-art graph OOD methods in kinetics property prediction problems. Our results demonstrated the challenges and opportunities in OOD kinetics property prediction. Our datasets and benchmarks can further support research in this direction. The github repository for code and data can be found in https://github.com/zihao-wang/ReactionOOD.

1 Introduction

In recent years, graph machine learning has been widely used in scientific discovery [Wang et al., 2023, Zhang et al., 2023] and gained particular success in chemistry [Gilmer et al., 2017, Jumper et al., 2021, Mullowney et al., 2023]. The underlying rationale is the long-standing structure-property relationship [Mihalić and Trinajstić, 1992] in chemistry. For example, Graph Neural Networks (GNNs) can efficiently encode information at both the molecular structure level and the atom level within a molecule, which reveal compelling properties of the molecule [Gilmer et al., 2017]. Such methods yield efficient, cheap, but still effective predictions of the properties of *unseen molecules* before expensive experiments or computations, which can serve as valuable reference information for drug discovery [Mullowney et al., 2023].

One of the *next* questions to be answered after the discovery of a proper but unseen molecule is *How to efficiently obtain unseen molecules through chemical synthesis.* In contrast to the molecule property estimation problem that concerns a *single* molecule, chemical synthesis processes involve the proper arrangement of various reactions that encompass *multiple* molecules under optimal conditions. Therefore, the very first step of achieving a high-throughput synthesis of unseen molecules is to estimate the properties of chemical reactions [Warr, 2014], especially kinetics properties that

*Equal contribution.

NeurIPS 2023 Workshop in AI for Scientific Discovery: From Theory to Practice.

describe the "rate" of reactions. This prediction task is expected to be Out-Of-Distribution (OOD) generalizable so that the kinetic properties of OOD reactions with unseen molecules can be well predicted.

Recently, machine learning methods have been applied to predict the kinetic properties of reactions [Heid and Green, 2021]. In existing studies, chemical reactions are assumed to be Independently and Identically Distributed (IID), and models are trained and tested within random splits [Heid and Green, 2021, Stuyver and Coley, 2022, Heid et al., 2023]. However, results from such IID assumptions provide little credible insight into the performances of existing ML methods in OOD reaction property prediction. Meanwhile, existing theoretical and empirical studies for OOD generalization on graphs [Ji et al., 2022, Gui et al., 2022], are restricted to problems with a single graph. How OOD methods perform reaction properties prediction with multiple molecules is still unknown.

To fill this gap, this paper discusses the out-of-distribution generalization issue when applying machine learning methods to the prediction of chemical reaction properties. We propose three levels of OOD shifts for ML-based reaction prediction: Structure OOD, Mechanism OOD, and Conditional OOD. Then, we reorganize recent reaction kinetic databases [Johnson et al., 2022] and create a comprehensive dataset in three levels of OOD. Furthermore, we empirically justify the performance of state-of-the-art kinetic property prediction produced by state-of-the-art OOD methods for general and graph inputs. Our results demonstrated that there remain huge ID-OOD performance gaps under different distribution shifts in chemical reactions for existing OOD methods.

2 Related works

Increasing efforts have been made to devise machine learning approaches for various aspects of chemical reaction systems [Davies, 2019, Stocker et al., 2020, Meuwly, 2021, Strieth-Kalthoff et al., 2022], such as reaction classification [Schwaller et al., 2021b, Burés and Larrosa, 2023], reaction optimization [Felton et al., 2021], atom mapping [Schwaller et al., 2021a], and the most fundamentally, reaction property prediction [Heid and Green, 2021]. With the burst of chemical reaction data [von Rudorff et al., 2020, Spiekermann et al., 2022, Johnson et al., 2022, Choi, 2023, Stuyver et al., 2023, Zhao et al., 2023], the Graph Neural Network (GNN) based methods [Heid and Green, 2021, Stuyver and Coley, 2022, Heid et al., 2023] are demonstrated its clear advantage over traditional methods by leveraging the structure of reactants and products.

Out-of-distribution shift is one of the long-standing problems in machine learning [Vapnik, 1991, Quinonero-Candela et al., 2008, Shen et al., 2021]. Recently, OOD generalizable graph neural networks have been discussed extensively [Bevilacqua et al., 2021, Zhu et al., 2021, Wu et al., 2022b,a, Chen et al., 2022]. When it comes to scientific discovery, out-of-distribution generalization capabilities enable machine learning methods to find more reliable discoveries from existing observations. A thorough investigation of the intersection of OOD and drug discovery can be found at Ji et al. [2022] and Gui et al. [2022]. However, as we will identify in the incoming parts, the out-of-distribution shifts for chemical reactions are radically different from those with existing graph OOD settings [Gui et al., 2022], and existing OOD methods do not perform well.

3 Preliminary

3.1 Chemical reactions and kinetic property prediction

A chemical reaction $\mathfrak{R}$ is described by the reactants $r_1, \ldots, r_m$, products $p_1, \ldots, p_n$, and the conditions $c_1, \ldots, c_l$.[2]

$$r_1 + \cdots + r_m \overset{c_1,\ldots,c_l}{\Longrightarrow} p_1 + \cdots + p_n. \tag{1}$$

Multiple molecules (including atoms, ions, and other species) are involved in one reaction, which differs from molecule property prediction tasks where only one molecule structure is considered.[3]

[2] In chemistry literature, different arrow types indicate different reaction types. We use $\Longrightarrow$ for simplicity.

[3] In this paper, we assume reactions adhere to the law of conservation of matter, ensuring both atoms and electrons are equally represented on either side of the reaction equation.

Table 1: Two types of distribution shifts. $[X]$ denotes the domain index of the input X.

Distribution shifts	$P([X])$	$P(Y\|[X])$
Covariate shift	$P^{\text{train}}([X]) \neq P^{\text{test}}([X])$	$P^{\text{train}}(Y\|[X]) = P^{\text{test}}(Y\|[X])$
Concept shift	$P^{\text{train}}([X]) = P^{\text{test}}([X])$	$P^{\text{train}}(Y\|[X]) \neq P^{\text{test}}(Y\|[X])$

Each molecule r_i, p_j is considered as a molecule graph, with atoms as attributed nodes and bonds as attributed edges. For a chemical reaction $\mathfrak{R}$, $R[\mathfrak{R}]$ ($P[\mathfrak{R}]$) denotes the graph of reactants (products) as the union of the reactant (product) molecule graphs. $C[\mathfrak{R}]$ denotes the set of conditions. $\mathcal{G}_R$ ($\mathcal{G}_P$) and $\mathcal{C}$ denote the space of graphs for reactants (products) and the space of conditions, respectively.

Kinetics properties investigated in existing literature include *reaction barrier* (activation energy) [von Rudorff et al., 2020, Spiekermann et al., 2022, Stuyver et al., 2023, Zhao et al., 2023] and *rate constants* [Heid and Green, 2021]. As one of the simplest mathematical descriptions of the rate constant, the Arrhenius equation reads

$$k = A \exp\left(\frac{-E_a}{RT}\right), \tag{2}$$

where k is the rate constant, which is modeled by a function of the Arrhenius factor A, the reaction barrier (activation energy) E_a, the universal gas constant R, and the absolute temperate T. Though Equation (2) only holds under mild conditions for some reactions, we could still summarize that, the rate constant k is jointly described from three aspects: (1) the inherent property of the reaction, such as the barrier E_a [Donahue, 2003], (2) the conditions, such as the temperature T, and (3) other parameters A to fit the empirical data.

In this paper, we also consider the reaction barrier E_a and the rate constant k as the target of prediction. Let $\mathcal{Y}$ be the space of the target property. The target Y of reaction $\mathfrak{R}$ is computed by the learnable function $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{X} = \mathcal{G}_R \times \mathcal{G}_P \times \mathcal{C}$ is the input space.

3.2 OOD shifts

Out-of-distribution generalization is one of the key topics in machine learning research. It considers the shifts of the joint distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$ during training and testing phase, $P^{\text{train}}(X, Y) \neq P^{\text{test}}(X, Y)$ [Gui et al., 2022]. Applying the conditional probability formula,

$$\underbrace{P^{\text{train}}(Y|X)}_{\text{Concept}}\underbrace{P^{\text{train}}(X)}_{\text{Covariate}} \neq \underbrace{P^{\text{test}}(Y|X)}_{\text{Concept}}\underbrace{P^{\text{test}}(X)}_{\text{Covariate}}. \tag{3}$$

This decomposition introduces two basic types of OOD shifts: covariate shift and concept shift, where one part of the distribution is changed while the other part is fixed.

One common way to characterize distribution shifts in high-dimensional input space is to properly separate the spaces into domains. Let $\mathcal{D}$ be the set of domains and $\mathcal{X}$ be the input space. For each input X, we define its domain index $[X]$. Table 1 summarizes the covariate shift and concept shift with the domain index.

Throughout this paper, the domain index $[X]$ is equivalent to an equivalent relation $\sim$. Specifically, for $X', X \in \mathcal{X}$, $[X'] = [X]$ if and only if $X' \sim X$. Then, it suffices to define **the equivalent relation** over the input space, which will naturally lead to the definition of the domain index and then two kinds of OOD shifts.

4 Three levels of OOD shifts in kinetic property prediction

This section defines three levels of OOD settings and details the OOD domains by constructing the **equivalent relations** in the input space $\mathcal{X} = \mathcal{G}_R \times \mathcal{G}_P \times \mathcal{C}$. The definition of equivalent classes, so as the OOD shifts, can be decomposed into the subspaces of $\mathcal{X}$. Structure OOD (S-OOD) establishes the equivalent relation in the subspace $\mathcal{G}_R \times \mathcal{G}_P$ while Condition OOD (C-OOD) concerns the equivalent relation in the subspace $\mathcal{C}$. Mechanism OOD (M-OOD) considers the shifts in general chemical reaction space $\mathcal{X}$ with the expertise of chemists. Figure 1 indicates the different levels of out-of-distribution generalization.

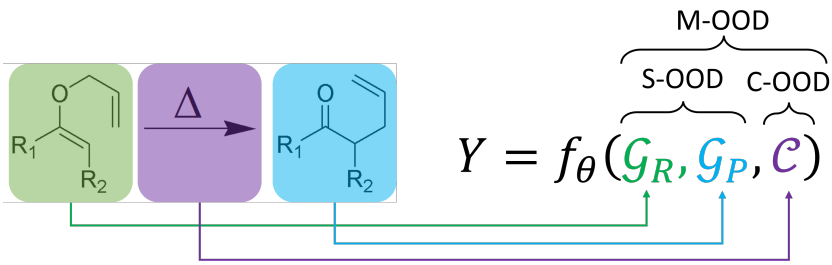

Figure 1: Illustration of different levels of reaction OOD. S-OOD and C-OOD focus on the distribution shifts in the subspace of the input. M-OOD focuses on the distribution shifts in the entire reaction space.

4.1 Structure OOD (S-OOD)

The first level of OOD settings is about the graphs for reactants and products, which are directly related to the graph machine learning methods such as graph neural networks.

Definition 1 (Structure OOD)**.** *Let $\sim$ be an equivalent relation in $\mathcal{G}_R \times \mathcal{G}_P$ and $[X]$ be the equivalent class containing X induced by $\sim$. We say $\sim$ defines a structure OOD setting.*

By instantiating the equivalent relation $\sim$, we can easily define several OOD settings. We begin with the basic cardinality $|G|$ of graph G.

Example 1 (S-OOD by total atom number)**.** *Consider the equivalent relation $\sim$, such that $R_1 \times P_1 \sim R_2 \times P_2$ if and only if $|R_1| = |R_2|$, i.e., the total number of atoms in two reactions are equal. Then the induced S-OOD setting is about the atom number of the largest reactant.*

Besides, the size of each reactant molecule also plays an important role. For example, the number of carbon atoms in an organic compound is related to its phase status in the ambient temperature and pressure, thus affecting the chemical reactions. Inspired by this, we can also define another type of S-OOD.

Example 2 (S-OOD by the atom number of the largest reactant)**.** *Consider the equivalent relation $\sim$, such that $R_1 \times P_1 \sim R_2 \times P_2$ if and only if $\max_{r_i \in R_1} |r_i| = \max_{r_j \in R_2} |r_j|$, i.e., the total number of atoms in two reactions are equal. Then the induced S-OOD setting is about the total atom number.*

Furthermore, it is also possible to define the S-OOD by considering the scaffold of molecules.

Example 3 (Structure OOD by the scaffold of the first reactant)**.** *Let $\sim_S$ be the equivalent relation for two molecular graphs with the same scaffold. Consider the equivalent relation $\sim$ such that $R_1 \times P_1 \sim R_2 \times P_2$ if and only if $r_{11} \sim_S r_{21}$ for $r_{11} \in R_1$ and $r_{21} \in R_2$. Then the induced OOD setting is about the first reactant scaffold.*

One can easily induce a more complex S-OOD setting by using the scaffold equivalence.

Example 4 (Structure OOD by the mutual scaffold of all reactants)**.** *Let $\sim_S$ be the equivalent relation for two molecular graphs with the same scaffold. Consider the equivalent relation $\sim$ such that $R_1 \times P_1 \sim R_2 \times P_2$ if and only if $\exists r_i \in R_1, r_j \in R_2, r_i \sim_S r_j$. Then the induced OOD setting is about the reactant scaffold.*

We could see that basic properties of molecules such as atom number and scaffold can define various types of S-OOD for chemical reactions, which is more complex than the situation for single molecule property predictions [Ji et al., 2022].

4.2 Condition OOD (C-OOD)

Focusing on the subspace $\mathcal{C}$ of $\mathcal{X}$, C-OOD investigates the generalization problem of the same reaction with respect to the reaction condition. Typical examples of C-OOD include rate-temperature dependency and rate-temperature-pressure dependency.

It should be stressed that Equation (2) is not always sufficient to describe how the rate constant k changes with temperature. It turns out that the reaction rate cannot be summarized by a uni-

Table 2: Summary of datasets. † indicates the database processed by Heid et al. [2023].

OOD level	Target	Domain	Database	# RXN	# Smp	Reference
S-OOD	Barrier	Total atom number (Eg. 1)	E2 & S_N2†	3625	3625	von Rudorff et al. [2020]
			RDB7†	23852	23852	Spiekermann et al. [2022]
			Cycloaddition†	5269	5269	Stuyver et al. [2023]
		First reactant scaffold (Eg. 3)	RDB7†	23852	23852	Spiekermann et al. [2022]
			Cycloaddition†	5269	5269	Stuyver et al. [2023]
C-OOD	Rate constant	T	RMG Lib. T	29161	87340	Johnson et al. [2022]
		(T, P)	RMG Lib. TP	3444	113695	
M-OOD	Barrier	Mechanism	RMG Family	12129	12129	

fied formulation. Chemists employ various forms of formulas to describe the dependency between the rate constant and conditions through data-driven approaches such as Chebyshev polynomial fitting [Heal, 1999] or analytical approaches like the Lindemann mechanism [Lindemann et al., 1922]. Other types of rate constant formulas can be found in Johnson et al. [2022].

Due to the complex nature of the dependency between the rate constant and conditions, C-OOD generalizability is particularly important if one wants to obtain the optimal condition of a reaction from machine learning models. Meanwhile, the equivalent relation $\sim$ for C-OOD is easier to construct since the most common conditions are temperature and pressure, which are all real-valued.

4.3 Mechanism OOD (M-OOD)

Another important intuition for characterizing chemical reaction systems, in particular for organic reactions, is the reaction mechanism. Reaction mechanisms are expert knowledge about how reactions happen and are examined in long-standing chemistry research. Typical examples include the E2 and S_N2 mechanisms [von Rudorff et al., 2020] that are usually considered as widely existing competing mechanisms.

Similar to the definition of S-OOD, M-OOD is also defined by the equivalent relationship, but in the general chemical reaction space $\mathcal{X}$ with an expert-defined relationship $\sim_M$. [4]

Definition 2 (Mechanism OOD)**.** *Let $\sim_M$ be the mechanism equivalence relationship in $\mathcal{X}$, such that $\mathfrak{R}_1 \sim_M \mathfrak{R}_2$ if and only if $\mathfrak{R}_1$ and $\mathfrak{R}_2$ follow the same established mechanism. We say $\sim_M$ defines a M-OOD setting.*

Reaction mechanisms also heavily rely on the molecular structure of reactants and products. However, M-OOD is motivated very differently from S-OOD by size equivalence or scaffold equivalence. M-OOD follows the well-established understanding and explanations in chemistry research. In this work, we define the M-OOD dataset by the kinetic database built by Johnson et al. [2022]. Detailed introduction can be found in Section 5.

5 Datasets and benchmarks

This section presents the details of constructing the benchmark to evaluate OOD generalizability of machine learning methods. We introduce how to create the datasets in different settings and the methods to benchmark. The derived collections of datasets is termed as ReactionOOD, which will be updated following to the growing of reaction kinetic databases.[5]

5.1 Dataset construction

Table 2 summarizes the information about the OOD datasets at three levels, with detailed target, domain, database, and statistics. For each database, we reorganize it into one or more domains to form multiple OOD datasets at different OOD levels. It worth mentioning that databases of E2

[4]The reaction of complex molecules can involve multiple reaction mechanisms happening simultaneously. Then $\sim_M$ is not a strict equivalent relation. However, it is always sufficient to conceptually regard the complex reaction as the combination of multiple elementary reactions where each only includes one mechanism.

[5]The version of ReactionOOD derived in this paper is `v0.1`.

& S_N2 [von Rudorff et al., 2020], RDB7 [Spiekermann et al., 2022], and Cycloaddition [Stuyver et al., 2023] has been processed by Heid et al. [2023]. However, the databases provided by Heid et al. [2023] only contain the reaction information (with atoms mapped to SMILES strings[6]) and the target properties. Thus, they can be only used for S-OOD but not C-OOD and M-OOD. We note that the E2 & S_N2 dataset contains reactions with molecules whose scaffold cannot be properly defined, which prevents the scaffold from being an applicable domain index for this dataset.

The well-established and open-sourced RMG kinetic databases [Johnson et al., 2022] are curated for further discussion about the C-OOD and M-OOD[7]. There are two major parts in RMG kinetic databases, i.e., REACTION FAMILY and REACTION LIBRARY. The REACTION FAMILY database contains 74 reaction mechanism families so that it can be used for the M-OOD level. The REACTION LIBRARY database contains 3,956 non-Arrhenius reactions and 3,444 pressure-dependent reactions out of 32,645 reactions in total, providing a data source for the investigation in the C-OOD level.

We use internal tools provided in the RMG databases to fetch the SMILES strings of reactants and products respectively, and integrate them to form the SMILES representation of reactions. Then, the atom mapping is extracted by RXNMapper[8] [Schwaller et al., 2021a]. For REACTION FAMILY database, we only choose the reactions with Arrhenius kinetics and set the target as the activation energy with standard units. This forms the "RMG Family" dataset at the M-OOD level in Table 2. For REACTION LIBRARY database, we further split them into two categories: reactions that depend solely on temperatures and reactions that depend on both temperature and pressure. We also filter out reactions whose valid temperature range does not contain 300K to avoid the effect of singular reactions. Then the "RMG Lib. T" and "RMG Lib. TP" datasets at C-OOD level in Table 2 are obtained by enumerating temperature in $[300K, 1,300K, 2,300K]$ and pressure in $[10^7, 2 \times 10^7, \ldots, 10^8]$. Rate constants at specific conditions are obtained by tools in RMG [Johnson et al., 2022] and conditions that are out of valid ranges are discarded.

For each dataset in Table 2, we follow the scheme of Gui et al. [2022] and Ji et al. [2022] to create In-Distribution (ID) and OOD splits for a comprehensive examination of the generalization abilities of existing approaches.

5.2 Methods

Feature engineering. To handle multi-molecule inputs of the kinetic property prediction task, we follow the established practice of condensed graph representation [Heid and Green, 2021] and use the graph feature extractor in a recent pipeline[9] [Heid et al., 2023]. For C-OOD settings, the temperature and pressure value is concatenated into the feature vector.

Backbone GNN for kinetic property prediction. All methods use the identical backbone GIN [Xu et al., 2019] with the virtual node trick. Noted that the backbone network in this paper is different from the one used in Heid et al. [2023]. However, our evaluation shows that the GIN backbone reaches similar MAE performances as [Heid et al., 2023] in the random split setting. Therefore, the vanilla Empirical Risk Minimization (ERM) [Vapnik, 1991] training is the baseline non-OOD method for comparing to other OOD methods

OOD methods. In addition to the ERM, we consider OOD methods developed for both Euclidean and graph data: (1) **Euclidean OOD methods** include IRM [Arjovsky et al., 2019], VREx [Krueger et al., 2021], GroupDro [Sagawa* et al., 2020], DANN [Ganin et al., 2016], Coral [Sun and Saenko, 2016], and (2) **graph OOD methods** include CIGA [Chen et al., 2022], GSAT [Miao et al., 2022] and DIR [Wu et al., 2022b]. We follow the same evaluation protocol and hyperparameter settings as in Gui et al. [2022]. Though the hyperparameters are tuned for single graph problems, they are still applicable to kinetic property prediction because the CGR feature extraction conceptually reformulates the graphs for reactants and products as one graph. We report the ID and OOD RMSE results of models selected according to the best ID and OOD validation performance. Each score is averaged by 10 runs with different random seeds.

[6]The Simplified Molecular-Input Line-Entry System (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

[7]https://github.com/ReactionMechanismGenerator/RMG-database

[8]https://github.com/rxn4chemistry/rxnmapper

[9]https://github.com/chemprop/chemprop

Table 3: OOD generalization performance on Cycloaddition and RDB7 dataset in two S-OOD shifts.

METHODS	CYCLOADDITION - SCAFFOLD (S-OOD)				CYCLOADDITION - TOTAL ATOM NUMBER (S-OOD)			
	COVARIATE		CONCEPT		COVARIATE		CONCEPT	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	4.78±0.07	5.80±0.40	5.04±0.06	5.57±0.08	4.12±0.07	5.23±0.29	4.30±0.11	6.00±0.13
IRM	13.73±0.59	16.60±1.05	17.09±0.26	18.43±0.44	17.64±0.16	17.06±0.30	23.04±0.22	22.46±0.22
VREX	5.50±0.06	6.41±0.86	5.17±0.05	5.76±0.06	4.84±0.07	**4.99**±0.16	5.40±0.14	6.65±0.19
GROUPDRO	4.88±0.05	**5.39**±0.55	**5.02**±0.07	5.58±0.06	4.24±0.07	5.31±0.35	4.19±0.07	6.04±0.09
DANN	4.76±0.07	5.93±0.32	5.03±0.11	5.55±0.13	**4.11**±0.10	5.17±0.23	4.22±0.07	**5.94**±0.10
CORAL	**4.75**±0.07	6.03±0.53	5.03±0.09	**5.53**±0.11	4.17±0.07	5.25±0.20	4.26±0.11	6.02±0.07
CIGA	5.25±0.39	5.49±0.71	5.15±0.36	5.90±0.33	4.52±0.30	5.12±0.28	5.08±0.45	6.67±0.47
DIR	5.06±0.36	5.75±0.85	6.25±0.71	7.18±1.06	5.12±0.33	5.55±0.38	5.13±0.48	6.67±0.42
GSAT	4.81±0.07	6.02±0.17	5.04±0.08	5.59±0.09	4.17±0.13	5.90±0.13	**4.10**±0.08	5.97±0.09

METHODS	RDB7 - SCAFFOLD (S-OOD)				RDB7 - TOTAL ATOM NUMBER (S-OOD)			
	COVARIATE		CONCEPT		COVARIATE		CONCEPT	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	9.49±0.09	22.90±1.08	10.08±0.10	**13.67**±0.10	**0.31**±0.00	**0.32**±0.02	0.33±0.01	0.47±0.01
IRM	58.59±0.60	71.15±6.74	65.27±0.24	62.66±0.41	0.31±0.00	0.32±0.02	0.93±0.01	1.02±0.02
VREX	13.06±0.17	22.74±0.89	11.88±0.15	15.52±0.18	0.31±0.00	0.33±0.01	0.50±0.01	0.65±0.02
GROUPDRO	10.31±0.09	22.54±1.16	10.08±0.07	13.77±0.11	0.31±0.00	0.32±0.02	0.33±0.00	0.47±0.01
DANN	9.44±0.10	**22.71**±0.78	10.00±0.15	13.88±0.53	0.34±0.01	0.47±0.01	0.33±0.00	0.46±0.01
CORAL	9.40±0.12	22.97±0.91	10.01±0.12	13.72±0.12	0.31±0.00	0.32±0.01	0.33±0.00	0.47±0.01
CIGA	10.48±0.86	25.05±1.74	10.99±1.09	14.23±0.86	0.32±0.01	0.39±0.04	0.36±0.01	0.49±0.03
DIR	10.27±0.82	24.68±2.42	13.37±1.71	16.82±1.48	0.32±0.01	0.28±0.03	0.35±0.01	0.47±0.02
GSAT	**9.27**±0.11	22.72±0.30	**9.88**±0.19	13.79±0.75	0.32±0.00	0.38±0.01	**0.31**±0.00	**0.32**±0.01

Table 4: OOD generalization performance on E2 & S_N2 and RMG with S-OOD and M-OOD shifts.

METHODS	E2 & S_N2 - SIZE (S-OOD)				RMG FAMILY - MECHANISM (M-OOD)			
	COVARIATE		CONCEPT		COVARIATE		CONCEPT	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	3.37±0.05	4.88±0.14	4.40±0.14	**4.21**±0.06	29.13±0.39	118.44±6.16	32.30±0.68	**53.25**±0.60
IRM	19.5±0.53	20.6±1.40	27.9±0.04	20.7±0.05	101.60±0.31	137.54±0.63	93.38±0.06	116.44±0.60
VREX	3.65±0.06	5.11±0.20	20.2±1.11	18.3±1.46	36.77±0.62	115.78±4.64	61.57±1.38	82.38±1.69
GROUPDRO	3.41±0.03	4.90±0.19	**4.26**±0.08	4.21±0.06	**28.11**±0.40	**114.56**±2.77	**31.90**±0.65	53.47±0.68
DANN	3.38±0.03	4.90±0.16	4.30±0.07	4.22±0.05	29.05±0.39	116.41±2.78	32.03±0.39	52.91±0.61
CORAL	**3.36**±0.04	**4.85**±0.12	4.33±0.09	4.21±0.09	29.09±0.39	116.09±3.87	31.91±0.59	52.85±0.66
CIGA	3.69±0.34	4.94±0.36	4.59±0.47	4.39±0.31	43.83±0.87	134.81±9.19	38.82±1.70	64.93±1.57
DIR	3.72±0.20	5.02±0.23	4.52±0.16	4.38±0.19	44.15±0.56	131.66±11.57	39.40±1.29	65.01±2.38
GSAT	3.39±0.06	5.07±0.15	4.34±0.06	4.22±0.06	30.97±0.99	117.06±3.34	32.56±0.85	53.31±0.68

6 Results and findings

Table 5.2 and the left half of Table 5.2 present the RMSE results for S-OOD settings, including graph size shifts on E2 & S_N2, Cycloaddition, and RDB7 datasets and scaffold shifts on Cycloaddition, and RDB7 datasets. The right half of Table 5.2 presents the M-OOD results of the mechanism shift on the RMG Family dataset. Table 5.2 presents the C-OOD results for T and TP shifts on the RMG Lib. dataset. For Table 5.2 and Table 5.2, the kinetic property to be predicted is the reaction barrier. For Table 5.2, the kinetic property to be predicted is the rate constant. For all tables, the best performance is boldfaced. If there are ties between ERM and any OOD methods, the ERM performance is boldfaced. Then we discuss two findings and other observations.

Finding 1: OOD kinetic property prediction is a challenging OOD problem. At first glance, it can be found that there remain huge ID-OOD performance gaps across different levels and types of OOD shifts and datasets. Neither OOD methods developed for Euclidean data nor OOD methods developed for graph data can outperform the vanilla ERM approaches consistently and significantly. The results also align with existing observations in realistic data and distribution shifts [Gulrajani and Lopez-Paz, 2021, Ji et al., 2022, Gui et al., 2022]. The performances of IRM are quite bad possibly due to the high requirements of IRM in learning invariant features under non-linear data [Chen et al., 2022].

Finding 2: M-OOD shift is significantly harder than the S-OOD shift. Moreover, it can be found that the ID-OOD performance gap is largely signified under mechanism OOD shifts (M-OOD), especially compared to the widely studied S-OOD shifts. The RMSE performances of M-OOD are An order of magnitude larger than those for S-OOD. Table 5.2 directly demonstrated this absolute

Table 5: OOD generalization performance in RMG Lib. with C-OOD shifts.

	RMG LIB. T - T (C-OOD)				RMG LIB. TP - (T, P) (C-OOD)			
	COVARIATE		CONCEPT		COVARIATE		CONCEPT	
METHODS	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	**3.28**±0.01	8.40±0.08	**3.70**±0.02	**6.76**±0.05	2.58±0.01	7.35±0.24	**2.92**±0.02	6.34±0.02
IRM	3.41±0.02	8.45±0.08	7.16±0.03	9.04±0.02	4.30±0.13	12.50±0.78	9.13±0.03	15.43±0.06
VREX	3.36±0.01	8.92±0.13	7.27±0.04	8.51±0.17	**2.55**±0.01	7.52±0.98	9.55±0.13	13.10±0.73
GROUPDRO	3.29±0.02	8.45±0.10	3.76±0.04	6.83±0.05	2.57±0.01	7.39±0.24	2.95±0.02	6.41±0.04
DANN	3.46±0.03	8.79±0.08	3.73±0.04	6.82±0.03	2.59±0.01	**7.16**±0.05	2.93±0.02	6.37±0.03
CORAL	3.29±0.02	**8.37**±0.06	3.75±0.04	6.77±0.05	2.58±0.01	7.36±0.16	2.94±0.03	6.39±0.03
CIGA	3.71±0.09	8.95±0.10	4.01±0.05	7.12±0.18	2.56±0.08	7.51±0.19	2.97±0.05	**6.16**±0.04
DIR	3.74±0.07	9.27±0.05	3.96±0.06	7.07±0.08	3.20±0.08	8.23±0.44	3.11±0.07	6.57±0.11
GSAT	4.01±0.06	9.48±0.09	3.82±0.05	7.06±0.11	3.03±0.06	8.26±0.42	3.08±0.03	6.58±0.07

gap by a huge contrast, where the E2 & S_N2 dataset has only two reaction mechanisms but the RMG Family dataset contains 74 reaction mechanisms.

Miscellaneous observations. We can also observe an exception for graph size concept shifts in E2 & S_N2, that OOD performances of various methods are generically better than ID performances. One conjecture is that the distribution of molecule sizes could limit the strength of the potential concept shifts. This distinct size distribution aligns with our earlier finding that scaffolds are not always identifiable for certain molecules. For concept shifts, the ID-OOD performance gaps are mainly caused by spurious correlations across different environments or domains. For images, such spurious correlations mainly lie between the background and the object in the images. For molecules, such spurious correlations are widely observed between scaffolds and the critical functional groups [Hu et al., 2020, Koh et al., 2021]. However, for small molecules that even do not have scaffolds, we suspect there is limited room for the existence of spurious correlations. Nevertheless, it remains challenging for the OOD generalization across different graph sizes, as demonstrated by the graph size covariate shifts.

7 Conclusion and future works

In this work, we identified various distribution shifts exist in chemical reactions and curated a new benchmark to examine the performance of multiple OOD generalization approaches. The results show that there remain huge ID-OOD performance gaps across different distribution shifts. Therefore, it calls for better graph machine learning approaches to tackle the OOD regression challenge for facilitating the chemical proper prediction.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Beatrice Bevilacqua, Yangze Zhou, and Bruno Ribeiro. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*, pages 837–851, 2021.

Jordi Burés and Igor Larrosa. Organic reaction mechanism classification using machine learning. *Nature*, 613(7945):689–695, 2023.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, 2022.

Sunghwan Choi. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nature Communications*, 14(1):1168, 2023.

Ian W Davies. The digitization of organic synthesis. *Nature*, 570(7760):175–181, 2019.

Neil M Donahue. Reaction barriers: Origin and evolution. *Chemical reviews*, 103(12):4593–4604, 2003.

Kobi C Felton, Jan G Rittig, and Alexei A Lapkin. Summit: benchmarking machine learning methods for reaction optimisation. *Chemistry-Methods*, 1(2):116–122, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Mache Learning Research*, 17:59:1–59:35, 2016.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

G Roger Heal. Evaluation of the function p (x), used in non-isothermal kinetics, by a series of chebyshev polynomials. *Instrumentation Science & Technology*, 27(5):367–387, 1999.

Esther Heid and William H Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9): 2101–2110, 2021.

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. 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, 2020.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, Houtim Lai, Shaoyong Xu, Jing Feng, Wei Liu, Ping Luo, Shuigeng Zhou, Junzhou Huang, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery – A Focus on Affinity Prediction Problems with Noise Annotations. *arXiv preprint*, arXiv:2201.09637, 2022.

Matthew S Johnson, Xiaorui Dong, Alon Grinberg Dana, Yunsie Chung, David Farina Jr, Ryan J Gillis, Mengjie Liu, Nathan W Yee, Katrin Blondal, Emily Mazeau, et al. Rmg database for chemical property prediction. *Journal of Chemical Information and Modeling*, 62(20):4906–4915, 2022.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning,*, pages 5637–5664, 2021.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826, 2021.

Frederick A Lindemann, Svante Arrhenius, Irving Langmuir, NR Dhar, J Perrin, and WC McC Lewis. Discussion on the radiation theory of chemical action. *Transactions of the Faraday Society*, 17:598–606, 1922.

Markus Meuwly. Machine learning for chemical reactions. *Chemical Reviews*, 121(16):10218–10239, 2021.

Siqi Miao, Miaoyuan Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. *International Conference on Machine Learning*, 2022.

Zlatko Mihalić and Nenad Trinajstić. A graph-theoretical approach to structure-property relationships, 1992.

Michael W Mullowney, Katherine R Duncan, Somayah S Elsayed, Neha Garg, Justin JJ van der Hooft, Nathaniel I Martin, David Meijer, Barbara R Terlouw, Friederike Biermann, Kai Blin, et al. Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, pages 1–22, 2023.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021a.

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. *Nature machine intelligence*, 3(2):144–152, 2021b.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Kevin Spiekermann, Lagnajit Pattanaik, and William H Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data*, 9(1):417, 2022.

Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T Margraf. Machine learning in chemical reaction space. *Nature communications*, 11(1):5505, 2020.

Felix Strieth-Kalthoff, Frederik Sandfort, Marius Kühnemund, Felix R Schäfer, Herbert Kuchen, and Frank Glorius. Machine learning for chemical reactivity: The importance of failed experiments. *Angewandte Chemie International Edition*, 61(29):e202204647, 2022.

Thijs Stuyver and Connor W Coley. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. *The Journal of Chemical Physics*, 156(8), 2022.

Thijs Stuyver, Kjell Jorner, and Connor W Coley. Reaction profiles for quantum chemistry-computed [3+ 2] cycloaddition reactions. *Scientific Data*, 10(1):66, 2023.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, volume 9915, pages 443–450, 2016.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1991.

Guido Falk von Rudorff, Stefan N Heinen, Marco Bragato, and O Anatole von Lilienfeld. Thousands of reactants and transition states for competing e2 and s2 reactions. *Machine Learning: Science and Technology*, 1(4):045026, 2020.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne J. Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun K. Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47 – 60, 2023.

Wendy A Warr. A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Molecular Informatics*, 33(6-7):469–476, 2014.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022a.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022b.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, Yuqing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael M. Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi S. Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess E. Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint*, arXiv:2307.08423, 2023.

Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A Ogunfowora, Sanjay S Garimella, Olexandr Isayev, and Brett M Savoie. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data*, 10(1):145, 2023.

Qi Zhu, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Shift-robust gnns: Overcoming the limitations of localized graph training data. *Advances in Neural Information Processing Systems*, 34:27965–27977, 2021.