applied sciences

MDPI

Article

Learning Hierarchical Representations for Explainable Chemical Reaction Prediction

Jingyi Hou [1,2,*] and Zhen Dong [3]

[1] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
[2] Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China
[3] College of Mathematics and Computer Science, Yan'an University, Yan'an 716000, China
* Correspondence: houjingyi@ustb.edu.cn

Abstract: This paper aims to propose an explainable and generalized chemical reaction representation method for accelerating the evaluation of the chemical processes in production. To this end, we designed an explainable coarse-fine level representation model that incorporates a small amount of easily available expert knowledge (i.e., coarse-level annotations) into the deep learning method to effectively improve the performances on reaction representation related tasks. We also developed a new probabilistic data augmentation strategy with contrastive learning to improve the generalization of our model. We conducted experiments on the Schneider 50k and the USPTO 1k TPL datasets for chemical reaction classification, as well as the USPTO yield dataset for yield prediction. The experimental results showed that our method outperforms the state of the art by just using a small-scale dataset annotated with both coarse-level and fine-level labels to pretrain the model.

Keywords: deep learning; reaction classification; yield prediction; data augmentation

check for updates

Citation: Hou, J.; Dong, Z. Learning Hierarchical Representations for Explainable Chemical Reaction Prediction. *Appl. Sci.* **2023**, *13*, 5311. https://doi.org/10.3390/app13095311

Academic Editors: Rung-Ching Chen, Goutam Chakraborty and Christine Dewi

Received: 14 March 2023
Revised: 22 April 2023
Accepted: 23 April 2023
Published: 24 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The development of machine learning techniques and the support of big data computing power have brought about revolutions in organic chemistry research [1–4]. With the assistance of machine-learning based computational models, chemists' materials, time in tasks such as chemical reaction prediction, and synthesis planning would cost less. It also brought the revolution of the high-throughput experimentation for drug and material discovery, as well as advantages for quick evaluation of the chemical processes in production [5–8]. Among these areas, learning how to represent a chemical reaction for simulating the corresponding experiments is one of the fundamental topics to be researched.

There are some studies focusing on the chemical reaction representation for various tasks, such as reaction classification and yield prediction. Schwaller et al. [9] exploited the Transformer-based model with a BERT classifier, which is widely used in the field of natural language processing to learn chemical reaction representations for classifying reactions, and they further adopted the similar network on the yield prediction task [10]. The deep model, which was used, achieves significant performance improvements on both tasks than traditional methods, but large amounts of training samples are required. Without any training data, Probst et al. introduced the differential reaction fingerprint based on circular n-grams of the reaction SMILES and hash functions for chemical reaction representation in [11] and could successfully accomplish the reaction classification and yield prediction tasks with the representations. Zeng et al. extended the scope of available data and proposed a deep-learning framework to learn representations from both biomedical texts and molecule structures in [12]. Schwaller et al. [13] presented data augmentation strategies to learn reaction representations for increasing yield prediction and achieved better performances than physics-based descriptors even in a small dataset. Although good

performances are achieved by large-scale annotated and augmented data, there still remain some prospects about the explanations and generalizations to be explored.

In this paper, we focus on another issue of the deep-learning based chemical reaction representation. The issue is inherent in the reaction data, which is the intra-class variance and the inter-class similarity. As we know, the completeness of a reaction is determined by certain atoms and bonds, and the remaining chemical structures are even unnecessary in the reaction [14–16]. Therefore, these structures can be quite different in the same class and usually much longer and more complex than the key atoms and bonds, which leads to the intra-class variance problem. In addition, even if we manage to enable the representation model to pay attention to these key atoms and bonds, there still exists the inter-class similarity problem. For example, the "isocyanate + amine reaction" and the "Fischer-Speier esterification" both belong to acylation which is the reaction of generating acyl group (-C=O). They are different in more specific atoms, such as the ester group (-COO-) and the amine ($-NH_2$). The deep learning reaction representation model is more likely to fail when the training data are imbalanced or new out-of-distribution data appear in the inference procedure.

To conquer this problem, we explicitly leverage a small amount of expert knowledge (i.e., the coarse-level reaction annotations) to model the coarse- and fine-level reaction representation for capturing both inter-class discriminative and intra-class descriptive information. On one hand, we built a new coarse-fine level sequence encoding model for chemical reaction representation. A sparse attention mechanism [17] was applied to extract and encode the most salient features of each reaction that were able to be coarsely classified. An attention mechanism with suppression was then used to highlight fine-level representations by constraining the impact of the coarse-level features. Accordingly, the model was forced to learn more discriminative and descriptive information with good explanations and generalizations. On the other hand, we designed a probabilistic data augmentation with a contrastive learning strategy to further improve the generalization of our method. The data augmentation was implemented by calculating the changing probabilities of the atoms and the bonds according to the coarse-fine level attention. To avoid introducing additional noises, atoms, and bonds with low correlation to the reaction, which are more likely to be changed for data augmentation. We conducted experiments on the Schneider 50k [18] and the USPTO 1k TPL [9] datasets for chemical reaction classification, as well as the USPTO yield [19] dataset for yield prediction. The experimental results show the effectiveness of the proposed method.

The main contributions of this paper lie in two aspects: Firstly, a coarse-fine level representation of a chemical reactions is proposed. The representation is able to capture both of the inter-class discriminative and intra-class descriptive information about the chemical reactions for multiple tasks, such as reaction classification and yield prediction. Secondly, the designed probabilistic data augmentation method is implemented by using a contrastive learning strategy, which makes it feasible to learn the coarse-fine level chemical reaction representations with only a few annotated training data.

The remaining parts are organized as follows: Section 2 reviews the related literature of our work, including research on chemical reaction representation and contrastive learning. In Section 3 we elaborate on the model structure, the loss functions, and the learning strategy of our method. Experimental results and discussions are presented in Section 4, and Section 5 concludes the whole paper.

2. Related Work

2.1. Chemical Reaction Representation

In recent years, several studies utilizing machine-learning techniques to represent chemical reactions have been conducted for various tasks, such as reaction classification and yield prediction. These studies can be roughly divided into two categories: molecular fingerprints based and graph neural network based.

The molecular fingerprints are mathematical representations capturing the structural and electronic features of the chemical reactions and can be used to train machine learning models. Probst et al. [11] introduced the differential reaction fingerprint method which generates a unique fingerprint representation capturing the changes in molecular structures for each reaction, and successfully adopted the representations of both the reaction classification and yield prediction tasks. Compared with traditional methods, this approach has the potential to accelerate the discovery and optimization of new chemical reactions with less time cost. Zeng et al. [12] developed a deep model that links molecule structure and biomedical text to learn chemical reaction representation and achieved a level of results comparable to human professionals. Schwaller et al. [13] investigated various data augmentation strategies to learn reaction representations for enhancing the yield-prediction performance and providing a better understanding of the associated uncertainties. Gilmer et al. [20] applied neural message passing algorithms to model the interactions between atoms in a molecule and predict various molecular properties such as energy, stability, and reactivity, and showed a high accuracy and efficiency compared to other state-of-the-art methods.

Another category is using the graph-based neural network to encode molecular structures and their associated properties. Schwaller et al. learned chemical reaction representations for the tasks of reaction classification [9] and yield prediction [10] via a Transformer-based deep model with a BERT classifier and achieved superior performances on both tasks, in comparison to traditional methods. Kwon et al. [21] utilized molecular graphs and reaction condition graphs to model the reaction and estimate the yield of reaction products by introducing uncertainty distributions. Saebi et al. [22] captured the graph structures of chemical reactions and their associated properties to predict yield.

Even though large achievements have been made by these methods, in comparison to traditional hand-crafted representations, these studies are limited either in the requirements of large amounts of annotated training data, or by a lack of interpretability and generalizability. In contrast, this paper proposes an explainable coarse-fine level representation model which only needs a small amount of easily available expert knowledge regarding the reaction classification and yield tasks.

2.2. Contrastive Learning

Contrastive Learning is an unsupervised learning method which aims to learn feature representations by comparing the differences among training samples. The main idea is to make similar samples closer together and dissimilar samples as far apart as possible in the feature space by comparing similar sample pairs from various data augmentations. Recently, contrastive learning has been a research hotspot in deep learning and widely applied in many fields, such as computer vision [23,24] and natural language processing [25,26]. However, only a few methods of chemical reaction representation using contrastive learning are proposed. Wang et al. [27] fully exploited the similarity and dissimilarity between molecules to perform contrastive learning for high-quality-molecule representation, which can be used in various applications, such as drug screening and molecular design. Wen et al. [28] represented chemical reactants as images, and utilized contrastive pretraining to improve the learning performance as its usage is in the image domain. Motivated by the excellent performance of contrastive learning on pretraining deep models, we augmented training data by introducing the contrastive learning strategy to improve the generalization ability of our model.

3. Method

The schema of the coarse-fine level chemical-reaction representation is illustrated in Figure 1. In the following subsections: we introduce the complete model structure of the proposed method, which contains a Transformer-based feature extraction module and a coarse-fine level sequence encoding module as shown in Figure 1a; then, we present the objective functions and the strategy for optimizing the model, i.e., the strategy of coarse-fine

level contrastive learning and the objectives of reaction classification and yield prediction as shown in Figure 1b.

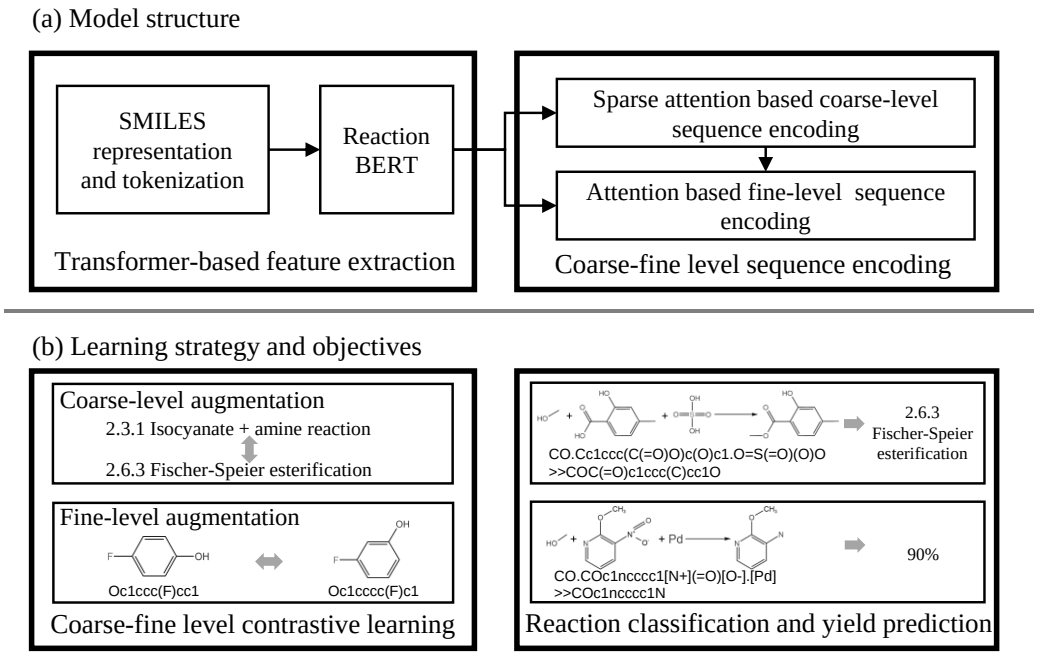


Figure 1. Schematic overview of our method for chemical reaction representation. The entire model structure of the coarse-fine level chemical-reaction representation is shown in (**a**). The Left part of (**b**) illustrates the data augmentation for coarse-level contrastive learning using two examples. The right part of (**b**) shows two examples of reaction classification and yield prediction, respectively.

3.1. Model Structure

3.1.1. Transformer-Based Feature Extraction

We follow [9,10] and use the Transformer-based model, Reaction BERT, as our backbone model to effectively obtain highly abstract information from the chemical reactions for the following relative shallow feed-forward operation. Specifically, we firstly represent each reaction according to the SMILES format [29], which is a sequential representation, and tokenize these sequences into N SMILES tokens. The tokenized SMILES sequence is then fed into the Reaction BERT to derive a transformer-based feature matrix $\boldsymbol{X}^t = \left[\boldsymbol{x}^t_{global}, \boldsymbol{x}^t_1, \boldsymbol{x}^t_2, \ldots, \boldsymbol{x}^t_N\right] \in \mathbf{R}^{d\times(N+1)}$, where $\boldsymbol{x}^t_n \in \mathbf{R}^d$ represents the d-dimensional transformer-based feature vector of the n-th SMILES token, and $\boldsymbol{x}^t_{global}$ is the feature of a virtue node representing the global information of the reaction. The learning of the global feature is task related. For example, this feature fed into a classifier represents the reaction category information, or the feature fed into a regression model represents the reaction yield information.

3.1.2. Coarse-Fine Level Sequence Encoding

The architecture overview of the coarse-fine level sequence encoding is shown in Figure 2, and we elaborate it as follows: Since reactions often occur between different functional groups in which the atoms are always adjacent when the reactions are represented by SMILES, we use a 1D CNN to extract the local context information from the transformer-based local features $\left[\boldsymbol{x}^t_1, \ldots, \boldsymbol{x}^t_N\right]$ to fully exploit the functional group-level representations. To map the global feature $\boldsymbol{x}^t_{global}$ into the same space with the local features, we employ a nonlinear fully connected feed forward network (FFN) to obtain the "local context" feature of the global node. Accordingly, we denote the extracted local context features by $\boldsymbol{x}^l = \left[\boldsymbol{x}^l_{global}, \boldsymbol{x}^l_1, \boldsymbol{x}^l_2, \ldots, \boldsymbol{x}^l_N\right]$.

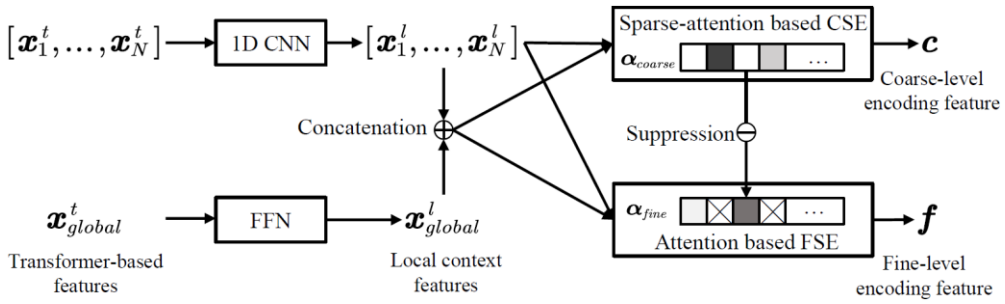
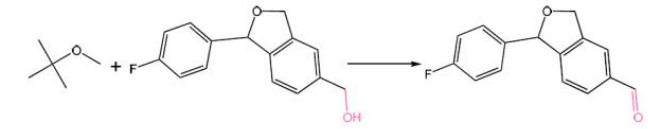
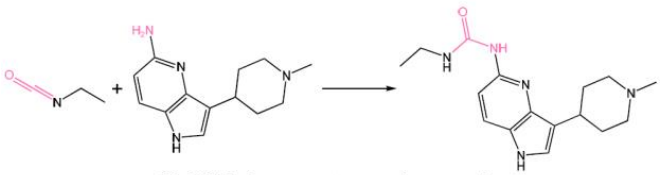


Figure 2. Architecture of the coarse-fine level sequence encoding module.

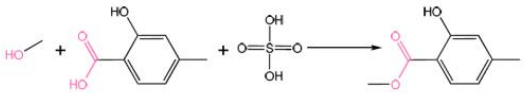
For more explainable modeling and representative encoding, we divide the following model into two branches to separately conduct Coarse-level Sequence Encoding (CSE) and Fine level Sequence Encoding (FSE). The coarse-level represents the molecules with specific bonds that are altered in the reaction, such as the "C-C bond formation" and the "Oxidations". The fine level represents the molecules in the reaction as specific functional groups or atoms, such as the "Alcohol to aldehyde oxidation" and the "Fischer-Speier esterification". Figure 3 illustrates examples of the coarse-level and fine-level differences between different reactions. Figure 3a is an oxidation reaction and Figure 3b,c are acylation reactions, where the bond of "-OH" alters into "=O" in Figure 3a and the bonds of "=C=O" are both altered in Figure 3b,c. Figure 3b is different from Figure 3c because Figure 3b contains "-NH_2" and "-N=C=O" in reactants but Figure 3c contains "-COO-" in the product.



(a) 8.1.4 Alcohol to aldehyde oxidation



(b) 2.3.1 Isocyanate + amine reaction



(c) 2.6.3 Fischer-Speier esterification

Figure 3. Examples of coarse-level and fine-level differences. The pink color remarks the main changes during the reactions. (**a**) is different from (**b**,**c**) at the coarse level. (**b**,**c**) are different at the fine level.

For CSE, we apply a sparse attention mechanism to encode the sequence into the coarse-level encoding feature $\boldsymbol{c}$, given by

$$\boldsymbol{c} = \left[\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_N^l\right] \cdot \boldsymbol{\alpha}_{coarse}, \tag{1}$$

where $\boldsymbol{\alpha}_{coarse}$ represents the attention coefficient vector at the coarse level. The $\boldsymbol{\alpha}_{coarse}$ is obtained by optimizing

$$\arg\min_{\boldsymbol{\alpha} \in \boldsymbol{B}^N} \| \boldsymbol{\alpha} - \widetilde{\boldsymbol{\alpha}}_{coarse} \|, \tag{2}$$

where $\widetilde{\boldsymbol{\alpha}}_{coarse} = [\widetilde{\boldsymbol{\alpha}}_1, \ldots, \widetilde{\boldsymbol{\alpha}}_N]^\top$, $\widetilde{\boldsymbol{\alpha}}_i$ is calculated via a two-layer non-linear fully connected network with $\left[\boldsymbol{x}^t_{global}, \boldsymbol{x}^t_i\right]$ as input, and $\boldsymbol{B}^N = \left\{\boldsymbol{b} \in (\mathbf{R}^+)^N \middle| 1^\top \boldsymbol{b} = 1\right\}$ is the N-dimensional simplex in which each vector has positive elements with total summation of 1. The optimization depicted in Equation (2) is solved by projecting $\widetilde{\boldsymbol{\alpha}}_{coarse}$ onto $\boldsymbol{B}^N$.

Similarly, for FSE, we use an attention mechanism to encode the sequence into fine-level encoding feature $\boldsymbol{f}$, given by

$$\boldsymbol{f} = \left[\boldsymbol{x}^l_1, \ldots, \boldsymbol{x}^l_N\right] \cdot \boldsymbol{\alpha}_{fine}, \tag{3}$$

where $\boldsymbol{\alpha}_{fine}$ represents the attention coefficient vector at the fine level. Since coarse-level features are not discriminative within the same coarse-level category for representing fine-level information, we suppress the coarse-level aware local context features in the fine-level encoding and the attention coefficient is calculated by

$$\boldsymbol{\alpha}_{fine}[i] = \begin{cases} 0 & \boldsymbol{\alpha}_{coarse}[i] > 0 \\ \phi([\boldsymbol{x}^t_{global}, \boldsymbol{x}^t_i]), & \boldsymbol{\alpha}_{coarse}[i] = 0 \end{cases}, \tag{4}$$

where $\phi(\cdot)$ is a two-layer non-linear fully connected network and $\boldsymbol{\alpha}_{fine} \in \boldsymbol{B}^N$. In this work, the calculation of $\boldsymbol{\alpha}_{fine}$ is implemented by masking the corresponding elements of the non-linearly projected coefficients, where $\boldsymbol{\alpha}_{coarse}[i] > 0$ with $-\infty$ and using the Softmax activation to normalize the coefficient vector.

3.2. Learning Strategy and Objectives

3.2.1. Coarse-Fine Level Contrastive Learning

In order to improve the generalization ability of the proposed model, we apply the contrastive learning strategy. In addition, we design specific data augmentation strategies for coarse-level and fine-level contrastive learning, respectively. We first introduce the contrastive learning and then describe the coarse-fine level data augmentation as follows.

Contrastive learning is effective for preserving the generalization of deep learning models by maximizing the lower bound of the mutual information (MI) of the input and the output representation. The lower bound is calculated by the MI of the output representations extracted by the deep models from two augmented views of the original input. To alleviate the collapse caused by large variations between the positive pairs that especially and frequently occur in coarse-level learning, we use an asymmetric mutual information neural estimation loss [30] to separately optimize our model at the coarse level and the fine level. Since the objective calculations at the two levels are the same, we take the coarse-level processing as an example and show the loss function of the j-th instance in the training set:

$$L_{CL_coarse}(j) = -\frac{1}{2}(h^\top_{pred}(\boldsymbol{c}_j)\hat{\boldsymbol{c}}_p - h^\top_{pred}(\boldsymbol{c}_p)\hat{\boldsymbol{c}}_j) + \lambda_{CL} \log(\textstyle\sum_{\boldsymbol{c}_n \in \mathrm{Neg}(\boldsymbol{c}_j)} \exp(\boldsymbol{c}_j^\top \boldsymbol{c}_n)), \tag{5}$$

where h_{pred} is the prediction head following [30,31], $\boldsymbol{c}_p \in \mathrm{Pos}(\boldsymbol{c}_j)$ is the sequence encoding feature of the augmented view of the j-th instance, and variables with hat symbols denote that the stop gradient operations are applied to them.

For the data augmentation, we take inspiration from the method [28] which manages to augment data by randomly changing atoms or bonds outside the reaction centers because these atoms and bonds are relatively not important in the reaction. In this paper, we design a probabilistic data augmentation for the SMILES representation. For the i-th token in the SMILE sequence, we define p_i to represent the probability of changing this token in the augmentation operation. The larger p_i means the farther token is away from the reaction center, and vice versa. We apply two kinds of changes to the SIMILES tokens. First, the rotation of the molecule SMILES randomization [32,33] is introduced and p_i is the probability of conducting the rotation operation on the left part of the i-th token in the

molecule. Second, we randomly move the branched chain to other places with the same 1-step local context in the molecule, and $\sum_{i \in \mathbf{S}} p_i$ is the probability of whether to move the branched chain, where $\mathbf{S}$ denotes the set of the token indices of the corresponding branched chain. An example of the second kind can be found in the box of "Fine-level augmentation" in Figure 1b.

In this work, the previously introduced attention coefficient vectors $\boldsymbol{\alpha}_{coarse}$ and $\boldsymbol{\alpha}_{fine}$ measure the importance of the token in the chemical reaction, and we believe that tokens with larger coefficient values are more likely to be the reaction centers. The probability p_i can be computed according to the attention coefficients. Specifically, for the data augmentation at the coarse level, p_i is calculated by

$$p_i = \text{Softmax}(1 - \boldsymbol{\alpha}_{coarse}[i], \tau), \tag{6}$$

where $\text{Softmax}(\cdot)$ is the Softmax operation and τ denotes the temperature. During training, we gradually decrease the value of the temperature. Besides, we add a data augmentation strategy that directly regards two reactions in the same coarse-level but different fine-level categories as a positive pair. An example can be seen in the box of "Coarse-level augmentation" in Figure 1b. For the data augmentation at the fine level, p_i is calculated by

$$p_i = \text{Softmax}(1 - \boldsymbol{\alpha}_{coarse}[i] - \boldsymbol{\alpha}_{fine}[i], \tau). \tag{7}$$

3.2.2. Reaction Classification and Yield Prediction

For the reaction classification task, we apply an ensemble learning way with three kinds of classifiers. The first one directly classifies the transformer-based feature, $\boldsymbol{x}^t_{global}$, with a Softmax operation optimized according to the cross-entropy loss, L_{CLS_global}. The second one learns a Softmax classifier with the temperature τ at the coarse level and generates a multinomial distribution, $\boldsymbol{y}_{coarse} \in \mathbf{R}^{C_{coarse}}$, and the value of m-th class is calculated by

$$\boldsymbol{y}_{coarse}[m] = \text{Softmax}(h_{CLS_coarse}(\boldsymbol{c})[m], \tau), \tag{8}$$

where $h_{CLS_coarse}(\cdot)$ denotes a two-layer non-linear fully connected network and C_{coarse} represents the number of coarse-level classes. The classifier is optimized according to the cross-entropy loss, L_{CLS_coarse}. The third one learns C_{coarse} classifiers at the fine level, each of which is sensitive to the fine-level classes belonging to a specific coarse-level class. These classifiers output prediction matrix $\boldsymbol{Y}_{fine} \in \mathbf{R}^{C_{fine} \times C_{coarse}}$, and the probability of n-th fine class of the final prediction $\boldsymbol{y}_{fine}$ is

$$\boldsymbol{y}_{fine}[n] = \text{Softmax}((\boldsymbol{y}_{fine} \cdot \boldsymbol{y}_{coarse})[n]). \tag{9}$$

Note that during the pre-training procedure, we replace the $\boldsymbol{y}_{coarse}$ by the one-hot ground-truth coarse-level labels in the first epoch for steady optimization, and then set τ in Equation (9) to 1 in the following training epochs. The total loss of reaction classification is formulated as

$$L_{RC} = \lambda_1 \left(L_{CLS_{coarse}} + L_{CLS_{fine}} \right) + \lambda_2 L_{CLS_global} + L_{CL_coarse} + L_{CL_fine}, \tag{10}$$

where λ_1 and λ_2 are hyperparameters. At the inference procedure, we select the category index with the largest value as the final result.

For fine tuning on the reaction classification dataset without coarse-level annotations, we remove the component of L_{CLS_coarse} in the loss function and gradually decrease the value of τ in Equation (9).

For the yield prediction, we adapt the model pre-trained by the reaction classification task and fine tune the adapted model for yield prediction. The fine-tuning process is just to concatenate the transformer-based feature $\boldsymbol{x}^t_{global}$, the coarse-level encoding feature c, and the fine-level encoding feature f, after the l2-normalization. Then a regression model is

used to learn the yield given the concatenated feature with the mean squared error (MSE) loss function.

4. Experiments

4.1. Datasets

The proposed method is evaluated on three publicly available datasets, i.e., the Schneider 50k [18] and the USPTO 1k TPL [9] datasets for reaction classification, as well as the USPTO yield [19] dataset for yield prediction. Table 1 briefly summarizes these datasets.

Table 1. Datasets used in this paper.

Datasets	Schneider 50k [18]	USPTO 1k TPL [9]	USPTO Yield [19]
Task	reaction classification	reaction classification	yield prediction
Reaction Number	50k	445k	500k
Class Number	9 super classes 50 template classes	1000 template classes	-
Split Strategy	10k for training 40k for testing	90% for training 10% for testing	Random split
		USP-few: 32 per class for training	Time split: data in and before 2012 for training, after 2012 for testing
Metric	Accuracy, CEN [34], and MCC [35,36]	Accuracy, CEN [34], MCC [35,36], and F1-score	R^2

The Schneider 50k dataset contains 50k reactions collected from the granted United States patents. The reactions are annotated with 50 template labels and 9 superclass labels using a substructure-based expert system. We use the superclass labels as the coarse- level labels in our method. We follow the data split in [18] to use 10k reactions for training and the remaining 40k reactions for testing. The performance is evaluated by the Accuracy, CEN [34], and MCC [35,36] scores, respectively.

The latter two datasets are both derived from the USPTO database [19], and both consisted of about 500k chemical reactions with different annotations. The USPTO data are highly noisy and biased.

The USPTO 1k TPL dataset is divided into a training and validation set and a test set, where the former contains 90% data and the latter contains 10% data [9]. The performance on the reaction classification task is evaluated by the Accuracy, CEN, and MCC scores, respectively. We also conduct more challenge tests on this dataset by only selecting 32 training samples per class, named USP-few, and use the macro F1 score for evaluations following [12].

The USPTO yield dataset contains a gram set and a sub-gram set. Following [10], we used two split methods, i.e., the random split and the time split, to validate the prediction performance of the proposed method. The time split denoted that data published in and before 2012 are used for training and validation, and the remaining were used for test. Since the data are quite noisy, the output yield of the data is smoothed as

$$\overline{y}_i = \frac{1}{5}\left(\sum_{y_i^{NN} \in NN(y_i)} y_i^{NN} + 2y_i\right), \tag{11}$$

where y_i and $\overline{y}_i$ represent the yield values of the i-th data before and after the smoothing operation, respectively, and $NN(y_i)$ denotes the set of the 3-nearest neighbors of y_i. The performance on the yield prediction task is evaluated by the R^2 score.

4.2. Implementation Details

We use the RDKit [37] to tokenize the SMILES representation. To implement the 1D CNN in our model, three 1D convolutional layers with residual structures are applied. The kernel size of the convolutional operators is set to 3 and the number of channels is set to 256. The activation function is the Leaky ReLU. The Adam optimizer is applied, and the chemical reaction model and the fine-tuned yield prediction model are trained over 10 and 8 epochs with the batch size equaling 64, respectively. Considering that the Schneider 50k dataset has coarse-level annotations, we use it as the pretraining dataset in this paper.

4.3. Results of Reaction Classification

Table 2 shows the comparison results on the Schneider 50k and the USPTO 1k TPL datasets for reaction classification. Two state-of-the-art methods are compared: Reaction BERT [9] and DRFP [11]. The Reaction BERT applies a Transformer-based model to learn chemical reaction representations and uses a BERT classifier to distinguish different reactions. The DRFP method maps a chemical reaction SMILES into a binary representation by hashing the symmetric difference of the two circular molecular n-gram sets generated from the molecules on the left and right of the reaction arrow, respectively. From the point of view of machine learning, representations of DRFT are hand-crafted features relying on domain prior to any statistical learning. The results of BERT and DRFT in Table 2 are reported in their original paper where CEN and MCC of Reaction BERT on Schneider 50k dataset are not reported. We follow the same experimental settings as the two methods and report the results of our model. Our method outperforms the compared methods. The performance improvement between ours and the backbone Reaction BERT verifies the effectiveness of the proposed coarse-fine level contrastive-pretraining mechanism. Both ours and Reaction Bert are better than DRFT on all three metrics, which demonstrates that representations learned from data can characterize unobvious but useful information on chemical reaction SMILES for the classification task.

Table 2. Reaction classification results on the Schneider 50k and the USPTO 1k TPL datasets.

Datasets	**Schneider 50k**			**USPTO 1k TPL**		
Metrics	Accuracy	CEN	MCC	Accuracy	CEN	MCC
Reaction BERT	0.985	-	-	0.989	0.006	0.989
DRFP	0.956	0.053	0.955	0.977	0.011	0.977
Ours	0.988	0.011	0.987	0.991	0.005	0.990

Table 3 shows the comparison results of the USP-few dataset for reaction classification. The compared KV-PLM [12] establishes the connections between molecule structures and biomedical text to represent chemical reactions. Our method can achieve comparable results with the KV-PLM. Similar to KV-PLM, our method leverages external information (i.e., the coarse-level reaction annotations) to model the coarse- and fine-level reaction representation for capturing both inter-class discriminative and intra-class descriptive information. The phenomenon that both KV-PLM and ours significantly outperform the backbone Reaction BERT method reveals the positive effect of external knowledge on describing molecule structures.

Table 3. Reaction classification results on the USP-few dataset.

Methods	**Reaction BERT**	**KV-PLM**	**Ours**
Marco F1	0.790	0.856	0.873

To further study the classification performance of our model on chemical reaction representation, a further analysis was conducted on misclassified samples in the Schneider 50k dataset. Specifically, we countered the true negative and false negative samples on the Schneider 50k dataset, and calculated their proportions in each category, as shown in

Figure 4. It can be found that the misclassified samples are not evenly distributed across all categories, but rather, show certain regular characteristics. The false positive proportion in the "2.6.1 Ester Schotten-Baumann" category is the highest, indicating that more samples were incorrectly classified into this category. One possible reason is that the features of this category are more dispersed and therefore the model assigns this category to a larger area in the sample space, making it difficult to accurately express this type of chemical reaction. The true negative rate is highest in the "1.8.5 Thioether synthesis" category, indicating that many misclassified samples come from this category. Another noteworthy point is that in the "1.7.9 Williamson ether synthesis" category both the true negative and false positive rates are high, indicating that our model has relatively poor classification performance in this category.

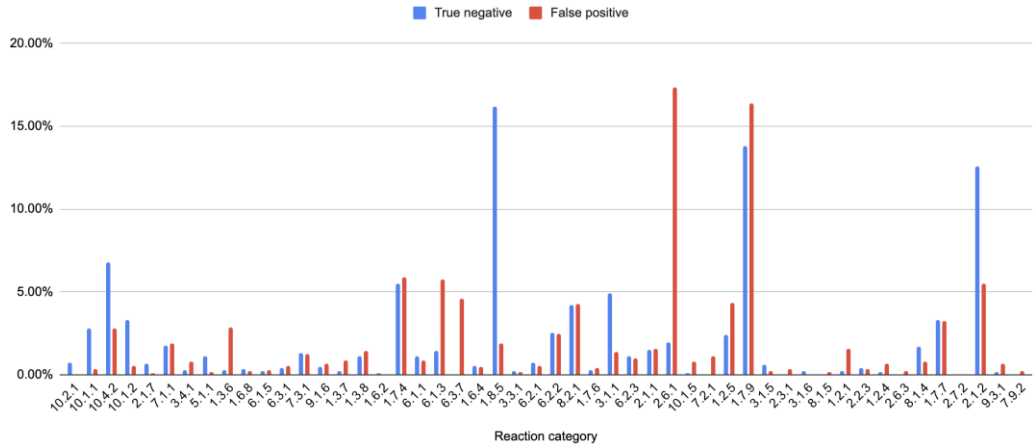


Figure 4. Proportions of true negative and false positive samples of each class on the Schneider 50k dataset.

We also visualize the several examples of attention coefficients, $\boldsymbol{\alpha}_{coarse}$ and $\boldsymbol{\alpha}_{fine}$, as shown in Figure 5. The SMILES tokens in red denote the mainly changed atoms and bonds during the reaction, which are the essential information and should be encoded into the reaction representation. The darker colored squares in the following two rows represent higher weights for the corresponding atoms and bonds. We observe that our coarse-level and fine-level attention mechanisms can actually preserve the discriminative and descriptive information on reactions, which helps to improve the classification accuracies.

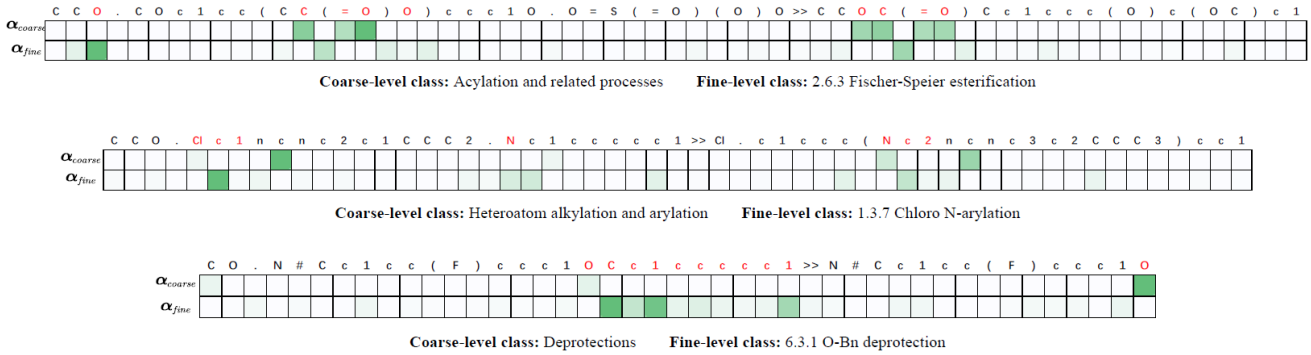


Figure 5. Examples of the coarse-level and the fine-level attention maps. The SMILES tokens in red denote the mainly changed atoms and bonds during the reaction.

To further show the effectiveness of the learned attention coefficients, we searched the neighbors of each chemical reaction according to their $\boldsymbol{\alpha}_{coarse}$ and $\boldsymbol{\alpha}_{fine}$. To be specific, the attention coefficient vectors are normalized to have same length, and then 4 kinds of distances, i.e., the Canberra, Mahalanobis, Euclidean, and Cosine distances, on both

the coarse- and fine-level coefficients between chemical reactions are calculated, which generates 8 metrics represented as "coarse_c", "coarse_m", "coarse_e", "coarse_o", "fine_c", "fine_m", "ine_e", and "fine_o", respectively. With these distance metrics, the top N nearest neighbors were selected, and the rank correlation of coarse and fine coefficients under different metrics were calculated. The correlation matrix is shown in Figure 6. We can conclude from the figure that the coarse-level feature plays a leading role in chemical reaction representation, and the coarse- and fine-level features are complementary.

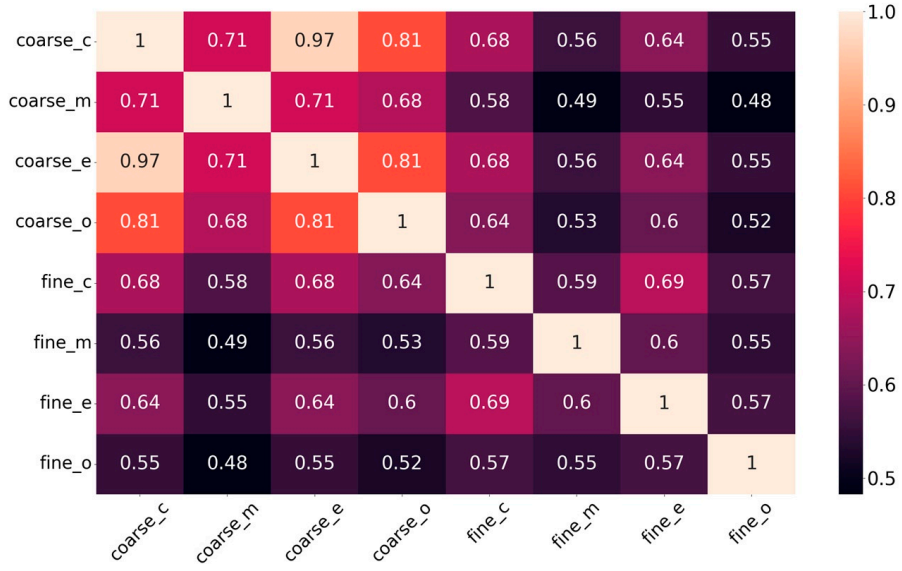


Figure 6. Kendall rank correlation matrix between coarse- and fine-level coefficients under various distance metrics.

4.4. Results of Yield Prediction

Table 4 shows the comparison results of yield prediction performances on the USPTO yield dataset. Note that the DRFP method [11] does not provide the results of the time split. The Yield BERT [10] uses a similar deep network as [9] with an extra regression layer for yield prediction. "Ours" denotes that our model is pretrained on the Schneider 50k and the USPTO 1k TPL datasets. "Ours (w/o pretraining)" denotes that our model is directly trained on the yield prediction data without being pretrained for reaction classification. From the results, we notice our method, no matter with pre-training or not, can almost achieve better performances than the other two methods on both gram scale and sub-gram scale. It shows that the representations learned by the proposed method can fully convey the intra-class similarity and discover the inter-class diversity. The performance difference between "Ours (w/o pretraining)" and "Ours" shows that pretraining is effective, which indicates that coarse-level annotations are valuable for the yield prediction task.

Table 4. Comparison results of R2 scores on the USPTO yield dataset.

Methods	**Yield BERT**		**DRFP**		**Ours (w/o Pretraining)**		**Ours**	
Splits	Random	Time	Random	Time	Random	Time	Random	Time
Gram scale	0.117	0.095	0.130	-	0.125	0.097	0.129	0.099
Sub-gram scale	0.195	0.142	0.197	-	0.198	0.146	0.200	0.147

4.5. Ablation Study

In order to evaluate the rationality of each proposed mechanism in our method, we perform an ablation study on the USP-few dataset. The mechanisms contain Coarse-level Sequence Encoding (CSE), Fine-level Sequence Encoding (FSE), probabilistic Data Augmentation (DA), the first data augmentation strategy (DA #1, i.e., the rotation operation), the

second data augmentation strategy (DA #2, i.e., the branch chain moving operation), and the third data augmentation (DA #3, i.e., intra fine-level classes switching for coarse-level augmentation), and the pretraining on the Schneider 50k dataset.

Table 5 shows the experimental results.

Table 5. Ablation study on the USP-few dataset. The "✓" and "-" represent the method with and without the corresponding mechanism, respectively.

#	CSE	FSE	Probabilistic DA	DA #1	DA #2	DA #3	Pretraining	Macro F1
1	-	-	-	-	-	-	-	0.790
2	-	✓	✓	✓	✓	-	✓	0.802
3	✓	✓	-	-	-	-	✓	0.787
4	✓	✓	-	✓	✓	✓	✓	0.835
5	✓	✓	✓	-	✓	✓	✓	0.823
6	✓	✓	✓	✓	-	✓	✓	0.867
7	✓	✓	✓	✓	✓	-	✓	0.872
8	-	✓	✓	✓	✓	-	-	0.793
9	✓	✓	✓	✓	✓	✓	✓	0.873

(1) Our full model in the 9-th row achieves the highest F1-score, which demonstrates that all the designed parts in our model help to discover discriminative and descriptive information for chemical reaction representation.
(2) The 1st row in the table is our backbone model, i.e., the reaction BERT model.
(3) The model of the 2nd row which removes the CSE module and the DA #3 for coarse-level augmentation is thus disabled. The improvement of the performance might mainly come from the data augmentation strategies. Compared with the 9th row, the result also suggests the great importance and usefulness of our coarse-level representation learning.
(4) We banned all the data augmentation methods and the contrastive learning in the 3rd row and observe that the result is even worse than the backbone models. It might be because the scale of Schneider 50k is too small to train a model to generalize well and lead to the negative transfer learning phenomenon. Therefore, it is important for the effectiveness of our model to conduct contrastive learning and data augmentation.
(5) The 4th row indicates that randomly conducting data augmentation achieves a lower F1-score than augmenting data conditioned on the atoms' importance in the reaction. The performance difference shows the superiority of our probabilistic DA mechanism.
(6) The 5th, 6th, and 7th rows show the results of disabling the 3 DA mechanisms, respectively. Comparing the results with the result of our full model in the 9th row, we can conclude that the DA #1 is the most effective method and the effectiveness of the DA #3 is limited.
(7) The 8th row indicates that the pretraining procedure is removed, and the coarse-level related mechanisms are thus removed without any coarse-level annotations being available. Compared with the 2nd and 3rd rows, the results further emphasized the necessity of our data augmentation and contrastive learning.

4.6. Disscussions

According to the above experimental results, our method outperforms other methods mainly due to:

(1) The coarse-fine level feature representation mechanism. In our model, a sparse attention is first applied to extract and encode the most salient features in coarse level, and an attention with suppression is then used to highlight fine-level representations by constraining the impact of the coarse-level features. The coarse-fine level representation is able to fully discover the useful information for the specific tasks.
(2) The contrastive learning-based data augmentation. Our probabilistic data augmentation calculates the changing probabilities of the atoms and the bonds in the reaction according to the coarse-fine level attention, which generates more accurate data for

contrastive pretraining without too much noise. The contrastive learning enforces the model to capture both the inter-class discriminative and intra-class descriptive information.

(3) The data-driven end-to-end training manner. Based on the coarse-level feature representation and the contrastive data augmentation, it is feasible to train our model in an end-to-end manner with only a small amount of annotated data. The advantage of the end-to-end training is to learn appropriate representations for various tasks, such as the reaction-classification and yield-prediction task.

5. Conclusions

Learning chemical reaction representation is one of the fundamental research topics of high-throughput automation experimentation. We have presented a coarse-fine level contrastive pretraining method for chemical reaction representation. The proposed method is able to learn the inter-class discriminative and intra-class descriptive reaction representation from a few well-annotated training data that are generalized well to various large-scale data. The discriminative and descriptive representation is learned by the sparse attention based coarse-level and attention based fine-level sequence encoding, and the generalization is further accomplished by our probabilistic data augmentation with a contrastive learning strategy. Experimental results on the chemical reaction classification and yield prediction tasks have shown the effectiveness of the proposed method.

Author Contributions: Conceptualization, J.H. and Z.D.; methodology, J.H.; software, J.H.; validation, J.H.; formal analysis, J.H.; investigation, J.H. and Z.D.; resources, J.H.; data curation, J.H.; writing—original draft preparation, J.H.; writing—review and editing, Z.D.; visualization, J.H.; supervision, J.H.; project administration, Z.D.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under grant No. 2020YFC1523200, the Natural Science Foundation of China under grant No. 62106021, and the Interdisciplinary Research Project for Young Teachers of USTB (Fundamental Research Funds for the Central Universities) No. FRF-IDRY-21-021.

Data Availability Statement: This research used publicly available datasets and no new data were created. The Schneider 50k [18], the USPTO 1k TPL [9], and the USPTO yield [19] dataset were used. The Schneider 50k dataset can be accessed through https://pubs.acs.org/doi/10.1021/ci5006614 (accessed on 23 March 2023) and downloaded via https://ndownloader.figstatic.com/files/3848755 (accessed on 23 March 2023). The USPTO 1k TPL dataset can be accessed via https://rxn4chemistry.github.io/rxnfp/ (accessed on 23 March 2023), and the USPTO yield dataset can be accessed through https://rxn4chemistry.github.io/rxn_yields/uspto_data_exploration/ (accessed on 23 March 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramakrishnan, R.; Dral, P.O.; Rupp, M.; von Lilienfeld, O.A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2015**, *2*, 140022. [CrossRef]
2. Goh, G.B.; Hodas, N.O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307. [CrossRef]
3. Coley, C.W.; Barzilay, R.; Green, W.H.; Jaakkola, T.S.; Jensen, K.F.; Kang, S.H. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377. [CrossRef] [PubMed]
4. Segler, M.H.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120–131. [CrossRef]
5. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [CrossRef]
6. Ma, J.; Sheridan, R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V.; Team, D.D. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040.
7. Raccuglia, P.; Elbert, K.C.; Adler, P.D.F.; Falk, C.; Wenny, M.B.; Mollo, A.; Zeller, M.; Friedler, S.A.; Schrier, J.; Norquist, A.J.; et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76. [CrossRef] [PubMed]
8. Mater, A.C.; Coote, M.L. Deep learning in chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559. [CrossRef] [PubMed]

9. Schwaller, P.; Probst, D.; Vaucher, A.C.; Nair, V.H.; Kreutter, D.; Laino, T.; Reymond, J. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **2021**, *3*, 144–152. [CrossRef]
10. Schwaller, P.; Vaucher, A.C.; Laino, T.; Reymond, J. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2021**, *2*, 15016. [CrossRef]
11. Probst, D.; Schwaller, P.; Reymond, J.L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit. Discov.* **2022**, *1*, 91–97. [CrossRef]
12. Zeng, Z.; Yao, Y.; Liu, Z.; Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat. Commun.* **2022**, *13*, 862. [CrossRef]
13. Schwaller, P.; Vaucher, A.C.; Laino, T.; Reymond, J.L. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. In Proceedings of the NeurIPS Workshop on Machine Learning for Molecules, Virtual, 6–12 December 2020.
14. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T.; Nair, V.H. "Molecular Transformer": A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583. [CrossRef]
15. Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 27–37.
16. Coley, C.W.; Barzilay, R.; Green, W.H.; Jaakkola, T.S.; Jensen, K.F.; Kang, B. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3427–3436. [CrossRef]
17. Hou, J.; Wu, X.; Wang, R.; Luo, J.; Jia, Y. Confidence-Guided Self Refinement for Action Prediction in Untrimmed Videos. *IEEE Trans. Image Process.* **2020**, *29*, 6017–6031. [CrossRef] [PubMed]
18. Schneider, N.; Lowe, D.M.; Sayle, R.A.; Landrum, G.A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39–53. [CrossRef]
19. Lowe, D. Chemical reactions from US patents (1976-Sep2016). *Figshare* **2017**. [CrossRef]
20. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 11–15 August 2017; pp. 1263–1272.
21. Kwon, Y.; Lee, D.; Choi, Y.S.; Kang, S. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *J. Cheminform.* **2022**, *14*, 2. [CrossRef] [PubMed]
22. Saebi, M.; Nan, B.; Herr, J.; Wahlers, J.; Wiest, O.; Chawla, N. Graph neural networks for predicting chemical reaction performance. *Chemrxiv.Org* **2021**. [CrossRef]
23. Jung, C.; Kwon, G.; Ye, J.C. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 18260–18269.
24. Wang, X.; Du, Y.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; Han, X. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **2023**, *83*, 102645. [CrossRef]
25. Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; Huang, J. Vision-language pre-training with triple contrastive learning. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 15671–15680.
26. Rethmeier, N.; Augenstein, I. A Primer on Contrastive Pretraining in Language Processing: Methods, Lessons Learned, and Perspectives. *ACM Comput. Surv.* **2023**, *55*, 1–17. [CrossRef]
27. Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287. [CrossRef]
28. Wen, M.; Blau, S.M.; Xie, X.; Dwaraknath, S.; Persson, K.A. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* **2022**, *13*, 1446–1458. [CrossRef]
29. Weininger, D. SMILES, a chemical language and information system. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
30. Lu, Y.; Wen, L.; Liu, J.; Liu, Y.; Tian, X. Self-Supervision Can Be a Good Few-Shot Learner. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 25–27 October 2022; pp. 740–758.
31. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 15750–15758.
32. Arús-Pous, J.; Johansson, S.V.; Prykhodko, O.; Bjerrum, E.J.; Tyrchan, C.; Reymond, J.L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **2019**, *11*, 71. [CrossRef] [PubMed]
33. Lambard, G.; Gracheva, E. Smiles-x: Autonomous molecular compounds characterization for small datasets without descriptors. *Mach. Learn. Sci. Technol.* **2020**, *1*, 025004. [CrossRef]
34. Wei, J.; Yuan, X.; Hu, Q.; Wang, S. A novel measure for evaluating classifiers. *Expert Syst. Appl.* **2010**, *37*, 3799–3809. [CrossRef]
35. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophyica Acta (BBA)-Protein Struct.* **1975**, *405*, 442–451. [CrossRef]

36. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367–374. [CrossRef]
37. Landrum, G.; Tosco, P.; Kelley, B.; Riniker, S.; Gedeck, P.; Schneider, N.; Vianello, R.; Dalke, A.; Schmidt, R.; Cole, B.; et al. *rdkit/rdkit: 2019 03 4 (Q1 2019) Release*; OpenAIRE: Athens, Greece, 2019. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.