PMechRP: Interpretable Deep Learning for Polar Reaction Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

In recent years, deep learning methods have been widely applied to chemical reaction prediction due to the time consuming and resource intensive nature of designing synthetic pathways. However, with the majority of models being trained on the US Patent Office dataset, many proposed architectures lack interpretability by modeling chemical reactions as overall transformations. These models map directly from reactants to products, and provide minimal insight into the underlying driving forces of a reaction. In order to improve interpretrability and provide insight into the causality of a chemical reaction, we train various machine learning frameworks on the PMechDB dataset. This dataset contains polar elementary steps, which model chemical reactions as a sequence of steps associated with movements of electrons. Through training on PMechDB, we have created a new system for polar mechanistic reaction prediction: PMechRP. Our findings indicate that PMechRP is able to provide both accurate and interpretable predictions, with a novel two-step transformer based method achieving the highest top-5 accuracy at 89.9%.

1 Introduction

Two main approaches exist for the prediction of chemical reactions: machine learning based methods, and quantum chemistry based methods [1, 13, 5, 8]. While quantum chemistry models offer detailed prediction of chemical properties, their computational demands render them feasible only for a limited scope of reaction systems, precluding their use for broad-spectrum, high-throughput reaction prediction. Conversely, ML models offer computational efficiency and scalability, making them well-suited for application across larger chemical systems and datasets. Countless ML models have been devised for tasks such as reaction yield prediction [16], reaction classification [14], chemical property prediction [4, 2], and both forward and reverse reaction prediction [6, 20, 3, 10].

Although ML models offer a high-throughput and highly adaptable chemical prediction, a significant drawback lies in their lack of interpretability when compared to quantum chemistry or simulation based methods. The predominant approach of predicting reactions as overall transformations results in a black-box scenario, where predicted products emerge directly from reactants without insight into intermediate transition states. Although these models may achieve high accuracy on datasets like the US Patent Office dataset [11], their outputs pose challenges for organic chemists, who typically reason through chemical synthesis via arrow-pushing mechanisms rather than overall transformations. An example of a overall transformation vs a mechanistic elementary step approach can be seen in Figure 1. The elementary step approach breaks the overall transformation down into a sequence of arrow pushing steps, which illustrate the flow of electrons and the shifting of atoms.

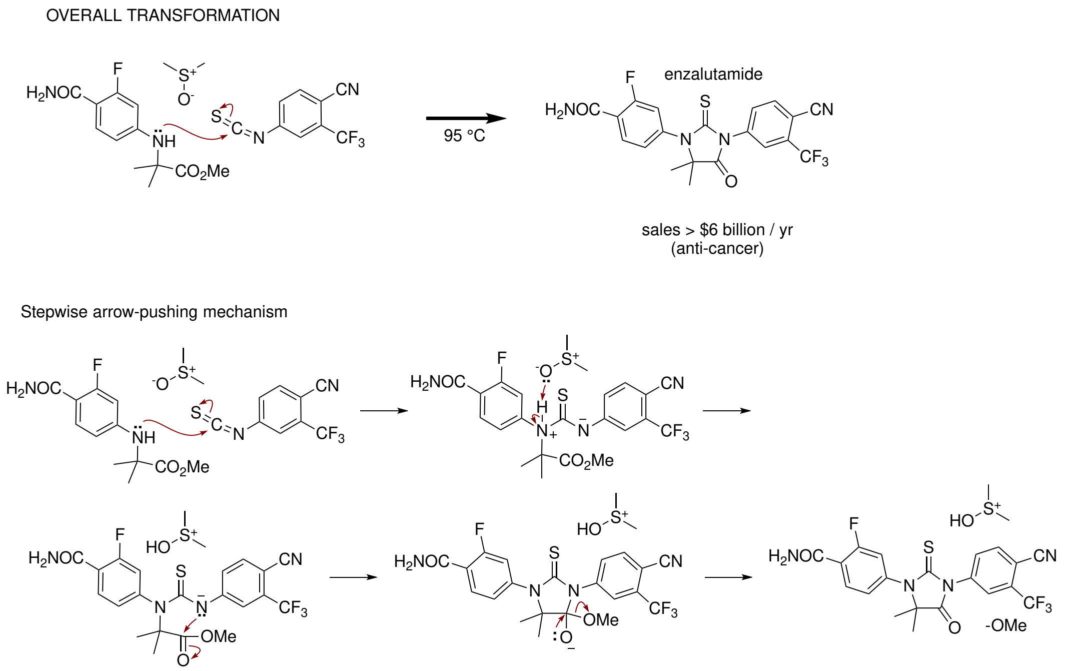


Figure 1: Example of an overall transformation vs an elementary step approach. This is a the final reaction step in the synthesis of enzalutamide, a drug used to treat prostate cancer that generates over $6 billion a year in revenue [21].

By thinking about reactions as occurring through many elementary steps, organic chemists are able to reason about the underlying driving forces of a reaction. When training ML models to forecast elementary step reactions, we effectively guide them to emulate an organic chemists' thought processes, thereby generating predictions that are readily interpretable and serve as practical aids for organic synthesis design.

2 Data

To develop predictive models for polar reaction mechanisms, we undertook training on the recently introduced PMechDB dataset. This dataset comprises more than 12,700 polar elementary steps, each balanced, partially atom mapped, and manually verified by a team of organic chemists. Each reaction represents a single elementary step polar reaction. These entries have been collected through manual curation from a diverse array of chemistry literature and textbooks [19]. These reactions are stored as smiles strings, and notably, the reactions contain arrow pushing information, providing insights into the reactivity of individual atoms within each reaction. Leveraging the manually curated reactions within the dataset, we conducted an 80/10/10 train/val/test split via random sampling from the "manually_curated_all.csv" file. For models which perform cross-validation, the validation data was combined with the training data.

3 Methods

Here we describe two different machine learning approaches for predicting polar elementary step mechanisms. Namely, we describe the reactive atom two-step approach, the single-step seq-to-seq prediction methods, and a spectator focused two-step transformer method.

3.1 Two-Step Prediction

The two-step prediction model comprises distinct phases. Initially, the model undertakes the task of predicting reactive atoms within the given reaction. Subsequently, these identified reactive sites are paired to formulate potential reaction mechanisms, followed by the application of a ranker model to rank the plausibility of these proposed mechanisms. This architectural design yields highly interpretable predictions, enabling a granular understanding of the model's rationale. When generating predictions, users can discern precisely which atoms are deemed reactive, and they can view the precise arrow-pushing mechanism predicted by the model. From the view point of organic chemists, the two-step architecture offers greater transparency compared to single-step approaches, as the arrow pushing mechanism provides justification for why the final products were predicted.

3.1.1 Siamese Architecture

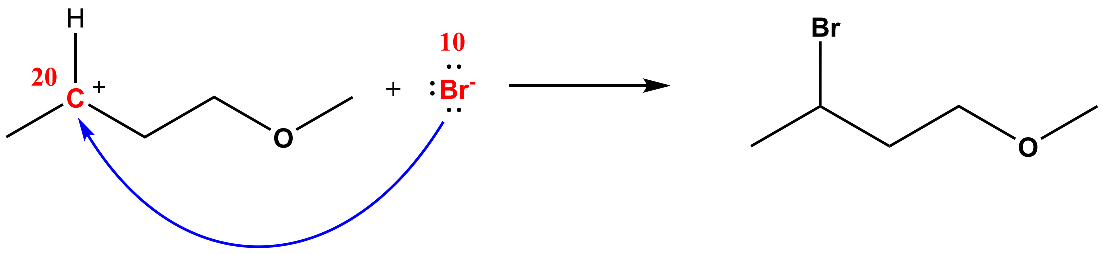
The two-step siamese architecture [6] comprises three distinct models, each serving a specific function. Initially, two separate reactive atom predictor models are instantiated. One model is specifically trained for predicting source atoms, while the other is trained for predicting sink atoms. To train the source and sink models, the electron-donating atom from the intermolecular arrow is labeled as the source atom, while the electron-accepting atom is labeled as the sink atom. This labeling process employs the reactive sites identification method as detailed in [6]. Atoms are represented by continuous vectors derived from predefined atomic and graph-topological features, utilizing a neighborhood of size 3. Subsequently, both source and sink classifiers are trained to categorize these feature vectors accordingly. After the trained reactive atom classifiers predict source and sink atoms, these atoms are paired together to enumerate possible arrow pushing mechanisms. Afterwards, a siamese architecture is used as a plasubility ranker model, which then ranks the plausibility of each potential mechanism to generate a final set of predictions. A visual representation of the source and sink pair is provided in Figure 2.

3.1.2 OrbChain

A polar elementary step reaction Rxn can be modeled as the following: a set of reactant molecules $R = \{r_0, r_1, \ldots, r_n\}$, a set of product molecules $P = \{p_0, p_1, \ldots, p_n\}$, and a set of arrows $\alpha = \{a_0, a_1, \ldots, a_m\}$, which transforms R into P. We consider a molecular orbital (MO) $m_i^{(*)}$ to be associated with four parameters: m = (a,e,n,c), where a represents the atom corresponding to the molecular orbital, e denotes the number of electrons contained in the MO, n corresponds to the atom adjacent to atom a in the case of a bond orbital, and c represents a possible chain of filled or unfilled MOs. Based on the methods described in [6, 9, 18], we model a polar mechanism as an interaction between two reactive molecular orbitals $(m_1^{(*)}, m_2^{(*)})$, where one orbital is the "source" orbital and acts as a nucleophile, while the other orbital is the "sink" orbital and acts as the electrophile. Given atom mapped reactants and products, and A, we can uniquely determine the reactive pair of orbitals in R used to create P. Conversely, given the reactive pair of orbitals $(m_1^{(*)}, m_2^{(*)})$ and the reactants R, we can generate P given R.

3.1.3 Reactive Atom Prediction and Plausibility Ranking

We enumerate over all molecular orbitals found in reactants R, and divide orbitals into reactive and non-reactive orbitals. These positive and negative examples are used to train the source and sink identification models. Rather than directly predicting the reactive MOs, we perform a binary classification prediction on the label of atom a, which is associated with the molecular orbital. We adopt the reactive sites identification method from [6] and represent atoms using continuous vectors becased on predefined graph-topological and physiochemical features. We train two models: a source model and a sink model. The source model predicts a binary classification label for whether or not an atom is a source, while the sink predicts a binary classification for whether or not an atom is a sink. The training data was constructed by extracting the labeled source, and the labeled sink atom from each reaction as positive examples, and then randomly sampling non-source or non-sink examples to use as negative examples.



C[CH+:20]CCOC.[Br-:10]>>C[CH:20](CCOC)[Br:10] 10=20

Figure 2: An example of a simple polar elementary step. The electron pushing arrows can be seen in blue, while the source and sink sites are seen in red. The bromine atom labeled 10 is the source atom. The carbon atom labeled 20 is the sink atom. The corresponding SMILES string and arrow codes can be seen below.

3.2 Plausibility Ranking

Once a set of source atoms and sink atoms are predicted, these two sets are paired together to generate pairs of molecular orbitals. A siamese network is used to rank the resulting molecular orbital pairs to generate the final reaction mechanism predictions.

3.3 Seq-to-seq Prediction

In addition to exploring two-step models, we also explore the performance of text-based models. An exceedingly common representation of chemical reactions is in the form of SMILES strings (simplified molecular-input line-entry system), which is a text-based representation. This representation lends itself towards NLP models such as transformers. These architectures model reaction prediction as a translation problem, wherein they are translating from reactant SMILES to product SMILES. These models have achieved state-of-the-art accuracies when predicting overall chemical transformations. However, these models possess several drawbacks in that they are more difficult to interpret and do not explicitly encode important molecular information such as invariance to atom permutations. This means that the same reaction can be represented by a large number of different SMILES strings, and additional strategies such as data augmentation may be needed to prevent a transformer model from making different predictions for identical sets of reactants.

3.3.1 Molecular Transformer

We utilize the innovative text-based reaction predictor, Molecular Transformer [15], which employs a bidirectional encoder and autoregressive decoder coupled with a fully connected network to generate probability distributions over potential tokens. The pre-trained Molecular Transformers underwent training using various versions of the USPTO dataset. We did not separate reactants and reagents, so the model pre-trained using the USPTO_MIT_mixed dataset was selected and subsequently fine tuned on the PMechDB dataset.

3.3.2 Chemformer

In addition to the molecular transformer, we also adopt the Chemformer model [7], which is another transformer-based reaction predictor. The Chemformer model also employs a bidirectional encoder and autoregressive decoder with a fully connected network to generate probability distributions over potential tokens. The Chemformer model was pre-trained on molecular reconstruction and classification tasks using a dataset of 100M SMILES strings from the ZINC-15 [17] dataset. Afterwards, the model was fine-tuned on various downstream tasks including forward prediction and retrosynthesis. The pre-training substantially improved the model's generalizability and convergence times on downstream tasks, such as USPTO forward prediction, compared to randomly initialized models. We chose to start from the model fine-tuned on USPTO-mixed since reactants and reagents are not separated in the PMechDB dataset. This model was subsequently fine-tuned on the PMechDB dataset for mechanistic-level predictions. The vocabulary of the model was expanded by 66 tokens to account for unseen atoms in the PMechDB dataset.

3.3.3 Two-Step Transformer Architecture

During experiments, all models were observed to exhibit a significant decrease in performance in reaction prediction as the size of the reactants grows. A quantitative analysis of the effects of spectators, and the number of atoms can be found in Figure 5 and Figure 6 respectively. To combat this, we propose a novel two-step architecture for transformers. Firstly, we use the source and sink reactive atom models from the siamese architecture to predict top-2 reactive atoms of the model. Reactant molecules which contain the predicted reactive atoms are considered to be non-spectator molecules. Since we take top-2 predictions from the source and sink models, we predict at most 2 sink molecules, and at most 2 source molecules. Pairing the sinks and sources together, we can have at most 4-unique source-sink combinations. After the combinations are generated, we run a top-5 prediction using our best performing transformer on each combination. Hence a fine-tuned chemformer model was used on each combination, as well as on the original reactants. After generating predictions for the source-sink combinations, the molecules which were deemed as spectators and removed are added back into the predicted products. If there are fewer than 4-unique source-sink combinations, more predictions are made on the original reactants until 5 total predictions are generated. For each reaction, we take the output predictions, canonicalize them, and then perform a simple majority vote with ties being broken randomly.

This architecture takes inspiration from common practices in organic chemistry. Often times when an organic chemist aims to predict the outcome of a set of reactants, they quickly look through all reactant molecules, and filter away molecules which are likely to be spectators or non-reactive, before focusing on a few molecules of interest. By performing a two-step prediction, we are able to first filter away potential spectator ions, then predict the reaction mechanism after reducing the space of possible reactions exponentially. A considerable performance increase was observed after performing this method of ensembling. The results can be seen in Table 3 and Table 4.

3.4 Multi-task learning

Due to the highly related nature of many chemistry prediction tasks, multitask learning can be used to develop robust models which may demonstrate improved learning efficiency and prediction accuracy. T5Chem is one such model, which leverages multitask learning on a transformer architecture to perform 5 different tasks. The T5Chem multi-task transformer architecture is able to perform forward/backwards prediction, reaction yield prediction, reaction classification, and reagents prediction [12]. This architecture was first pretrained with a BERT-like MLM objective on 97 million PubChem molecules. Then, the model was further fine-tuned on 5 different tasks using the USPTO_500_MT dataset. We selected this model, and fine-tuned it using the 80/10/10 split of the manually curated PMechDB reactions.

4 Results and Discussion

4.1 Performance on PMechDB Dataset

We assess the performance of the two-step prediction method, comprising reactive sites identification and plausibility ranking. The top-N accuracy of the reactive sites identification on PMechDB is presented in Table 7. Reactive site identification is considered correct if both the source and sink atom were correctly identified within the top-N predictions of each model.

Table 1: Reactive Atom Classification for Siamese Architecture

Top-1	Top-2	Top-3	Top-5	Top-10
53.8	79.0	86.8	91.8	94.4

The source and sink ranking models are able to predict the reactive atoms with relatively high accuracy. Although the reactive atom models are able to filter down the number of potentially reactive atoms significantly, due to the large number of atoms and aromatic structures contained in the polar reactions, enumerating all possible molecular orbital pairs leads to a large number of possible reaction mechanisms fed into the ranker model. Several reaction fingerprints were used for plausibility ranking. The results can be found in Table 2.

Table 2: Plausibility Ranking for Two-Step Architecture

Model Type	**Top-1**	**Top-2**	**Top-3**	**Top-4**	**Top-5**
reactionFP	39.5	56.3	65.6	70.3	73.0
DRFP	37.3	52.2	60.1	67.1	72.5
rxnfp	35.1	51.3	60.5	66.1	70.0

In order to perform two-step prediction, both reactive site identification and plausibility ranking must be performed. Thus for the best performing two-step model, we use the reactionFP fingerprint for plausibility ranking. Therefore in Table 3, we consider this as the best two-step siamese model. For the Chemformer, MolTransformer, and T5Chem models, we fine-tuned the pretrained models on the PMechDB datset. The results comparing all the trained models can be seen in Table 3

Table 3: Top-N Accuracy of Trained Models

Model Type	**Top-1**	**Top-3**	**Top-5**	**Top-10**
Best Two-Step Siamese	39.5	65.6	73.0	76.6
MolTransformer	59.1	66.3	69.2	70.1
T5Chem	56.6	69.1	73.7	77.5
Chemformer	74.0	84.1	85.2	87.2
Two-Step Transformer	80.6	88.8	89.9	91.0

Although the Siamese two-step model allows for improved interpretability due to its direct prediction of arrows, the models based on Chemformer yield the most accurate predictions, with the two-step transformer model outperforming all other models significantly. The effects of various ensemble sizes can be seen in Table 4.

Table 4: Effects of Ensemble Size on Top-N Accuracy

$ensemblesize$	**Top-1**	**Top-3**	**Top-5**	**Top-10**
2	71.8	85.8	86.9	87.7
3	77.8	87.5	88.5	89.2
4	79.8	88.7	90.0	90.7
5	80.6	88.8	89.9	91.0

4.1.1 Pretraining

Pretraining the Chemformer models made a large difference in performance, the effects of pretraining can be seen in Table 5.

The large increase in performance from the pretraining, indicates overlap between the USPTO dataset and the PMechDB dataset. This is in stark contrast to radical mechanisms, which exhibited lower performance when using a pretrained model [18]. This suggests that radical reactions are underrepresented in USPTO datasets compared to polar reactions, and that pre-trained transformer models would be expected to have higher performance on polar reactions.

4.2 Pathway Search

In addition to predicting single-step elementary reactions, further work is being done to evaluate and improve the model's performance on predicting polar mechanistic pathways. This involves chaining

Table 5: Top-N Accuracy of Chemformer Models

Model Type	**Top-1**	**Top-3**	**Top-5**	**Top-10**
no-pretraining	39.9	55.6	58.7	60.4
pretrained on zinc	74.9	77.0	82.8	84.5
pretrained on zinc and USPTO Mixed	74.0	84.1	85.2	87.2

several elementary steps together to transform a list of starting reactants to a list of target products. An example of a simple two-step mechanism correctly predicted by the ensemble transformer model can be seen in Figure 3.

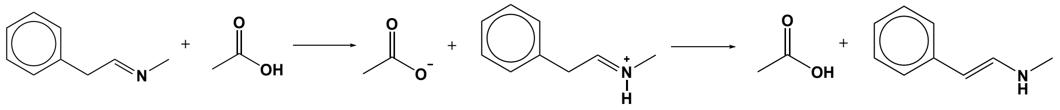


Figure 3: A simple 2-step mechanism correctly predicted by ensemble transformer model.

Although the transformer architectures outperform all other models in single step predictions on the test dataset, the reactions contained in PMechDB are mostly 1-2 reactant reactions, and contain very limited spectator ions. This results in the transformer models having a strong performance on reactions which contain 1-2 reactants, but inconsistent performance on reactions with one or more spectator ions. An example of this can be seen in the following elementary step which contains a spectator benzene ring. 4

When the chemformer model is asked to predict on Step A, it does not recover the correct products, while on Step B with spectators removed, it ranks the products as the top-1 prediction. Interestingly, the two-step transformer model is able to correctly predict this step. Comparing the various methods numerically, the two-step models appear to demonstrate significantly less performance degradation in predicting elementary steps with spectator molecules. Figure 5 demonstrates the top-5 accuracies of the various models as the number of reactant molecules is varied, while Figure 6 demonstrates the top-5 accuracies as the number of atoms contained in the reactants is varied.

The two-step transformer model can be seen to outperform both the chemformer and siamese architectures. When comparing the models, it seems that the number of reactant atoms has a much smaller effect on the prediction accuracy of the transformer models when compared to the siamese architecture. Perhaps this indicates that the transformer models are able to implicitly learn which reactive atoms it should pay attention to without being distracted by large unreactive substructures.

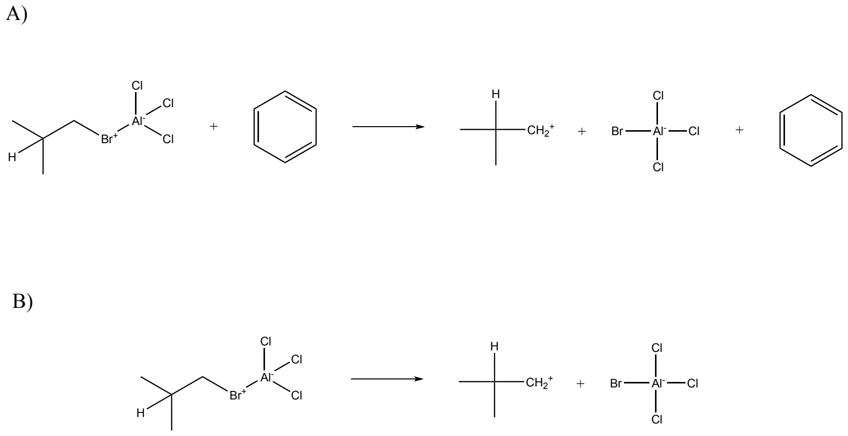


Figure 4: Step A represents the elementary step with the spectator molecule benzene included. Step B represents the elementary step with the benzene ring excluded.

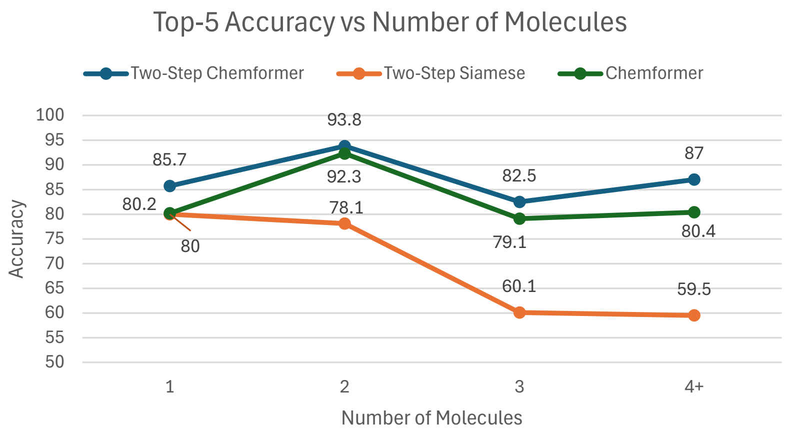


Figure 5: Comparing the top-5 accuracies of both the transformer and two-step models as number of reactant molecules is varied.

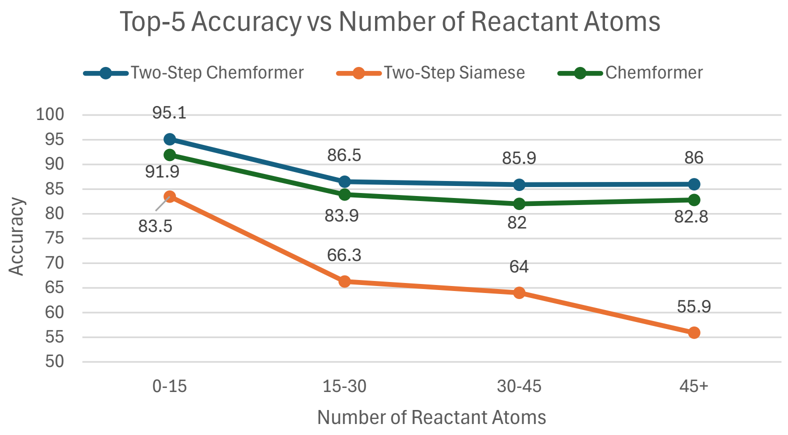


Figure 6: Comparing the top-5 accuracies of both the transformer and two-step models as number of reactant atoms is varied.

Notably, the two-step transformer model strongly outperforms the chemformer model when it views reactions which contain more than 2 reactant molecules. This suggests the first step manages to filter away the spectator ions to some extent and makes the prediction task easier for the transformer model.

5 Limitations

Lastly, we note there are several limitations with the current state of the PMechRP polar reaction system. Firsly, the PMechDB dataset includes less than 13,000 steps. This means the dataset is relatively small for training large architectures, and it may be difficult for these models to generalize well to all forms of experimental chemistry. Secondly, the transformer models directly translate from reactants to products, without generating the arrow pushing mechanisms. Although the elementary step predictions still offer significant interpretability, the two-step siamese method offers greater insight into the causality of a reaction by directly showing the flow of electrons. Additional methods could be developed to predict arrow codes or reactive orbitals using a transformer architecture in order to offer predictions with arrow pushing mechanisms.

6 Conclusion

We developed and compared several reaction prediction systems for polar reaction mechanisms. Through our analysis, we have created the reaction prediction system, PMechRP. This predictor offers a fresh perspective on reaction prediction by specifically targeting polar reactions and operating at the mechanistic reaction level. From the viewpoint of organic chemists, mechanistic level reaction prediction offers immense interpretabiltiy benefits, and has a lot of potential to aid in the prediction of synthetic pathways. We utilized PMechDB datasets to train and develop a wide range of architectures. Our findings demonstrate that the most accurate models are based on a two-step process, where spectators are filtered out to generate a variety of reactants before they are fed into an ensemble transformer architecture. Leveraging PMechDB datasets, our polar predictor marks a significant step towards interpretable reaction prediction.

References

[1] Roman M Balabin and Ekaterina I Lomakina. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *The journal of chemical physics*, 131(7), 2009.

[2] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8):1757–1772, 2017.

[3] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–443, 2017.

[4] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

[5] Larry A Curtiss, Paul C Redfern, and Krishnan Raghavachari. Gaussian-4 theory. *The Journal of chemical physics*, 126(8), 2007.

[6] David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3(3):442–452, 2018.

[7] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

[8] Dora Kadish, Aaron D Mood, Mohammadamin Tavakoli, Eugene S Gutman, Pierre Baldi, and David L Van Vranken. Methyl cation affinities of canonical organic functional groups. *The Journal of Organic Chemistry*, 86(5):3721–3729, 2021.

[9] Matthew A Kayala, Chloé-Agathe Azencott, Jonathan H Chen, and Pierre Baldi. Learning to predict chemical reactions. *Journal of chemical information and modeling*, 51(9):2209–2222, 2011.

[10] Matthew A Kayala and Pierre Baldi. Reactionpredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. *Journal of chemical information and modeling*, 52(10):2526–2540, 2012.

[11] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, 2012.

[12] Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 62(6):1376–1387, 2022.

[13] Gabriel A Pinheiro, Johnatan Mucelini, Marinalva D Soares, Ronaldo C Prati, Juarez LF Da Silva, and Marcos G Quiles. Machine learning prediction of nine molecular properties based on the smiles representation of the qm9 quantum-chemistry dataset. *The Journal of Physical Chemistry A*, 124(47):9854–9866, 2020.

[14] Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital discovery*, 1(2):91–97, 2022.

[15] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[16] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1):015016, 2021.

[17] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

[18] Mohammadamin Tavakoli, Pierre Baldi, Ann Marie Carlton, Yin Ting Chiu, Alexander Shmakov, and David Van Vranken. Ai for interpretable chemistry: Predicting radical mechanistic pathways via contrastive learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Mohammadamin Tavakoli, Ryan J Miller, Mirana Claire Angel, Michael A Pfeiffer, Eugene S Gutman, Aaron D Mood, David Van Vranken, and Pierre Baldi. Pmechdb: A public database of elementary polar reaction steps. *Journal of Chemical Information and Modeling*, 2024.

[20] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling*, 60(1):47–55, 2019.

[21] Ai-Nan Zhou, Bonan Li, Lejun Ruan, Yeting Wang, Gengli Duan, and Jianqi Li. An improved and practical route for the synthesis of enzalutamide and potential impurities study. *Chinese Chemical Letters*, 28(2):426–430, 2017.

A Appendix / supplemental material

In this appendix, we provide additional details about the experiments and models trained.

A.1 Compute Resources

All models were trained using a single NVidia Titan X GPU.

A.2 PMechDB Dataset

Here we provide some Figures 7, 8 displaying the the number of atoms and atom types found in the PMechDB dataset [19]

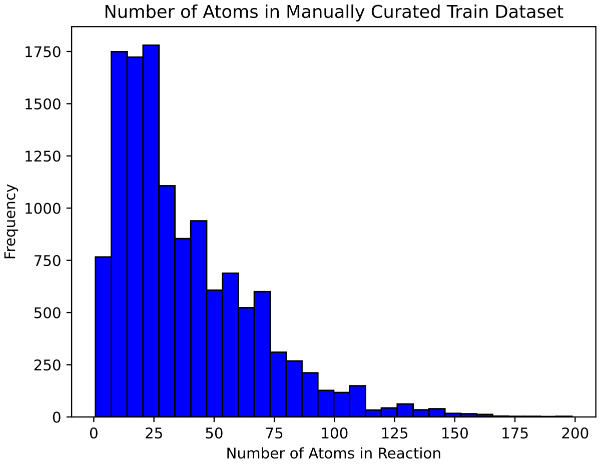


Figure 7: The distribution of the total number of atoms contained in each reaction for the manually curated training dataset.

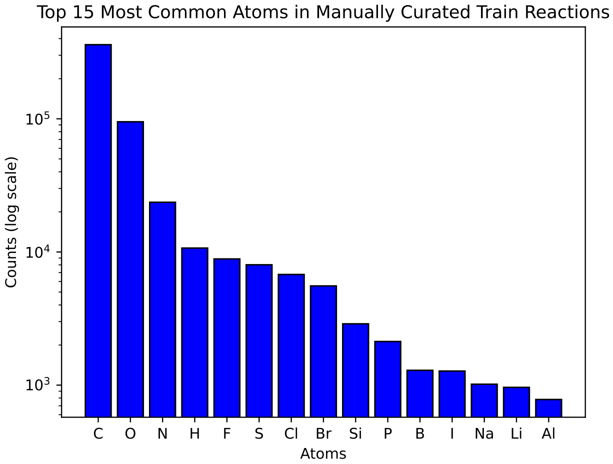


Figure 8: The distribution of atoms for the reactions in the manually curated training dataset.

A.3 Reactive Atom Prediction

A fingerprint of length 800 is constructed for each atom. This fingerprint includes 700 graph-topological features. These features are extracted using a neighborhood of size 3 with the method described in [6]. The remaining features consist of physiochemical properties such as valence number, electronegativity, etc.

The source and sink prediction models are trained using the "manually_curated_all.csv" file, where a 90/10 train/test split was performed. Each training reaction is processed to extract the atom fingerprints, the atom is given a label 1 if it is reactive, and 0 if it is non-reactive. The final output layer performs a binary classification on a reactive atom. The parameters of the source and sink prediction models can be seen below:

Table 6: Source and Sink Model Parameters

Batch Size	Num Layers	Layer Dim	Act	Reg
64	5	512-256-128-164-1	RELU	L2

A.4 Plausibility Ranking

We tested 3 fingerprints. The reactionFP fingerprint is extracted using the features explained in [6] to create a fingerprint of length 3200. For the rxnfp fingerprint, we use the default configuration to create a fingerprint of size 256. We use the DRFP fingerprint with a size of 2048 with the default configuration.

The parameters of the ranker models can be seen below:

Table 7: Source and Sink Model Parameters

Batch Size	Num Layers	Layer Dim	Act	Reg
200	3	360-360-1	Tanh	Dropout (0.5)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: The paper clearly outlines its contributions and scope. All contributions are backed by evaluating the accuracy of the models, and providing tables and plots for the performance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper does discuss the limitations of the work. We analyze and address situations where the model performs poorly, such as on reactions with a large number of spectator ions or atoms.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We clearly describe the architectures used, as well as any modifications made to them. The hyperparameters and dimensions of the models can be found in the appendix. The PMechDB dataset is publically available and can be accessed by any user.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: For the existing models, their codes can be found online at their respective git repositories. For the two-step models which predict reactive atoms, the codes use openeye software, which is a commercial library to do most of the chemoinformatics processing and thus this code cannot be released. Everything else from the paper is publically available including the PMechDB dataset.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the data splits and hyperparameters necessary to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: There are not error bars to report, the models were assessed based on their reaction prediction accuracy. They were evaluated once on the test set, so there are no error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: This information can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: To our knowledge the paper conforms with the code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discussed the ability of the model to be applied to synthetic pathway prediction, which is a very important challenge of chemistry, and the ability of the models to provide interpretable predictions for chemistry.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not have risk for misuse. It simply describes architectures which are useful for specifically predicting elementary step reactions. The models currently have no ability to design synthetic pathways for a target molecule, they must be first provided with a list of reactants to produce a set of products.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Papers are cited. The models used have public access git repos. The PMechDB dataset is governed by the Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND) license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have provided descriptions of the various methods and experiments, as well as their limitations.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This does not apply.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This does not apply.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.