Supporting Information:

Unified Deep Learning Model for Multitask Reaction Predictions with Explanation

Jieyu Lu† and Yingkai Zhang*,†,‡

†*Department of Chemistry, New York University, New York, New York 10003, United States*

‡*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China*

E-mail: yingkai.zhang@nyu.edu

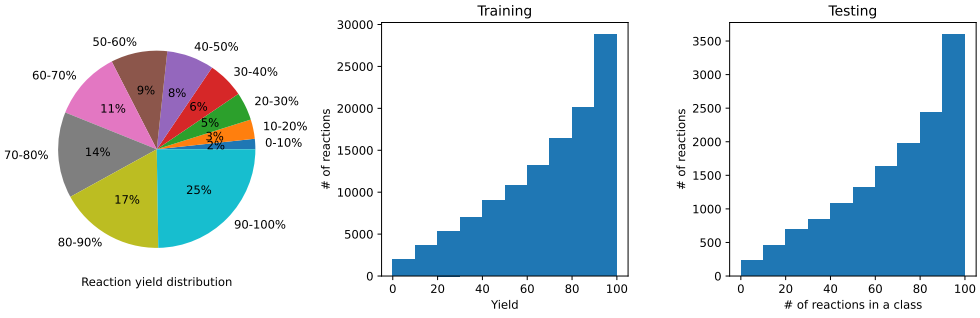

Figure S1: Reaction yield distribution on USPTO_500_MT. It is biased toward high yield reactions, given the fact that people are less likely to report low yield or even failed reactions.

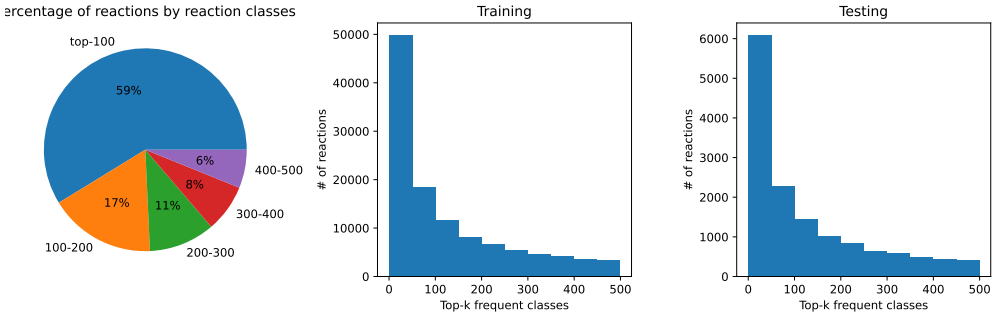

Figure S2: Reaction class distribution in USPTO_500_MT. Even though many sparse reaction classes have been removed, the reaction classes are still unbalanced. The top-100 most frequent reaction classes account for 59% reactions in reaction data sets.

Table S1: Results for different tokenizations on USPTO_500_MT. The bold entries highlight the best-performing approach. Both atom-level tokenization and SELFIES tokenization end up with more than 2,000 token types with long-tailed distributions. In our experiments, we only kept the top 994 most frequent tokens and 6 task-specific prompts, leading to a word embedding layer of input size 1,000. It fully covers $> 99.5\%$ reactions. By contrast, character-level tokenization only creates 74 token types, we ended up with using a word embedding layer of input size 100 for these 74 tokens, 6 prompting tokens and 20 placeholders.

Task type	Forward	Retrosynthesis	Reagents	Classification	Yield
Metrics	Top-1 accuracy	Top-1 accuracy	Top-1 accuracy	Accuracy	R
Character-level	**97.5%**	**72.9%**	**24.9%**	**99.4%**	0.46
Atom-level	96.3%	70.4%	20.0%	99.3%	0.46
SELFIES	76.5%	56.6%	17.0%	97.6%	0.46

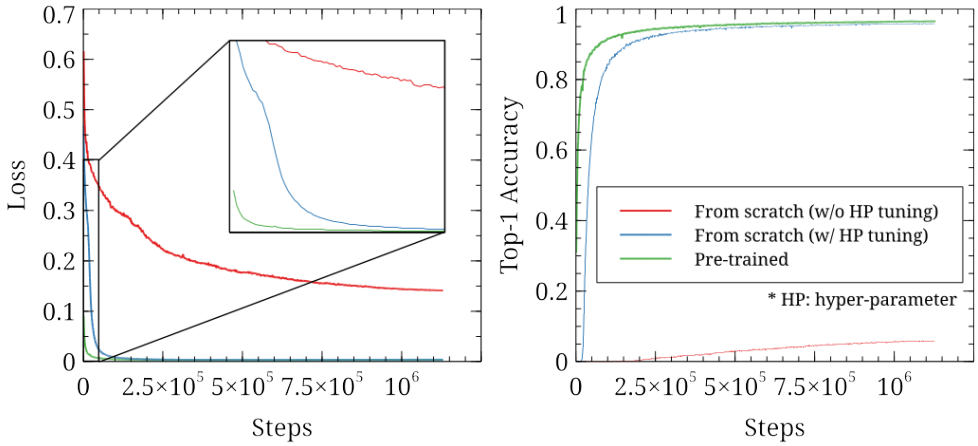

Figure S3: Learning curves (Left) and accuracy plot (right) for forward reaction prediction in validation set. Red curve represents model trained from scratch without hyper-parameter tuning. It has a low top-1 accuracy ($< 10\%$). Blue curve stands for model trained from scratch with careful fine-tuning using various batch sizes and learning rates. It has close final performance as our pretrained model Green), with slower converging speed.

Table S2: Results for multitask training using combined loss functions. To further explore the multitask nature of T5Chem, we built a combined model with all output layers (molecular generation head, regression head and classification head) trained together. We used weighted sum of individual tasks as the final loss for this combined model. We also explored various learning rates for different layers and number of fully connected output layers, but they perform worse than our two multitask models. The best combined model was obtained by applying 1.5× weights to sequence to sequence tasks and 1.0 for regression and classification tasks.

Model type	Forward	Retrosynthesis	Reagents	Classification	Yield
Metrics	Top-1 accuracy	Top-1 accuracy	Top-1 accuracy	Accuracy	R
Sub-models	**97.5%**	**72.9%**	**24.9%**	**99.4%**	0.46
Combined	96.0%	67.6%	19.4%	99.4%	**0.47**

Table S3: Results for data augmentation for sequence-to-sequence tasks. We explored 5-fold data augmentation using non-canonical SMILES, and found that data augmentation did not improve performance on USPTO_500_MT data set in terms of top-k accuracy (even slightly worse results), but it did improve SMILES validity.

	USPTO_500_MT			USPTO_500_MT (Augmented)		
	Product	Reactants	Reagents	Product	Reactants	Reagents
Top-1 Accuracy (%)	**97.5**	**72.9**	**24.9**	96.9	71.4	21.7
Top-2 Accuracy (%)	**98.8**	**86.4**	**33.2**	98.6	85.9	29.6
Top-5 Accuracy (%)	**99.4**	94.6	**43.8**	**99.4**	**95.1**	40.4
Top-1 SMILES invalidity (%)	0.19	0.16	**0.01**	**0.11**	**0.04**	**0.01**
Top-2 SMILES invalidity (%)	10.24	0.77	0.04	**1.56**	**0.13**	**0.03**
Top-5 SMILES invalidity (%)	21.47	2.91	**0.04**	**4.64**	**0.47**	**0.04**