Supplementary material for the paper "Reagent Prediction with a Molecular Transformer Improves Reaction Data Quality"

Mikhail Andronov,[*,†] Varvara Voinarovska,[‡] Natalia Andronova,[¶] Michael Wand,[†,§] Djork-Arné Clevert,[‖] and Jürgen Schmidhuber[†,⊥]

†*IDSIA, USI, SUPSI, 6900 Lugano, Switzerland*

‡*Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich – Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), 85764 Neuherberg, Germany*

¶*Via Berna 9, 6900 Lugano, Switzerland*

§*Institute for Digital Technologies for Personalized Healthcare, SUPSI, 6900 Lugano, Switzerland*

‖*Machine Learning Research, Pfizer Worldwide Research Development and Medical, Linkstr.10, Berlin, Germany*

⊥*AI Initiative, KAUST, 23955 Thuwal, Saudi Arabia*

E-mail: mikhail.andronov@idsia.ch

1 Data

Fig. S1 shows the t-SNE maps of both USPTO 50K and the Reaxys test set used in the paper. We used parametric t-SNE, which ensures that the 2D embeddings of similar examples from the two datasets are close if the examples are similar. The overlap between the maps is not

perfect but significant.

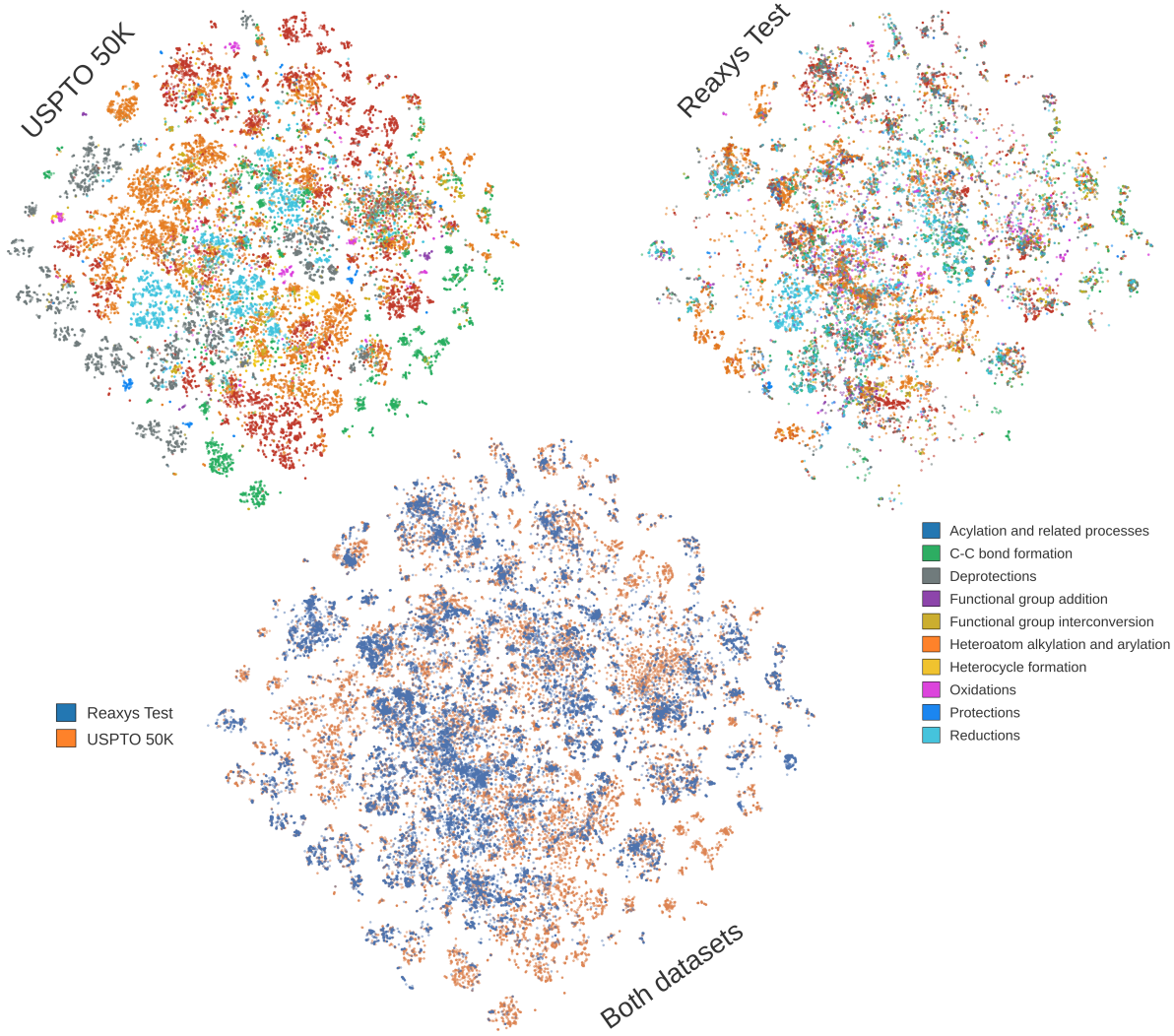

Figure S1: t-SNE maps for reactions in USPTO 50K and Reaxys test. The points which lie close together represent similar reactions. The absolute coordinates of the points have no physical meaning.

2 Reagent prediction

The performance of the reagent prediction model across reaction classes and reagent roles in the Reaxys test set is given in Fig. S2.

Figure S2: Comparison of the reagent prediction top-1 exact match accuracy across all reaction classes and reagent roles. On the top, the overall comparison of the test set is given. In the middle, for nonempty ground truth only. On the bottom, the table shows the number of nonempty ground truth strings for each reagent role and reaction type in the Reaxys test set.

In each reaction class, only a few reagent roles are usually represented. For example, oxidizers rarely appear in C-C bond formation, and catalysts are not very frequent in FGA. At the same time, solvents are usually listed for any class of reactions. The table in the middle of Fig. S2 takes into account only the performance of the model in correctly predicting the presence of reagents, not the general presence and absence. It can be seen that the performance is pretty low when predicting the presence of rare reagent roles for any type of reaction.

3 Statistical tests

To prove that the improvement of the model trained on the new data compared to the baseline Molecular Transformer is statistically significant, we employed McNemar's test. McNemar's test is usually applied to test if one binary classifier performs better than another. If one has some test set T and two classifiers F_1 and F_2, then one can build a 2×2 contingency table (Table S1):

Table S1: A contingency table is needed to perform McNemar's test. Each entry contains the number of test examples in T for which either F_1 or F_2 give correct or incorrect binary predictions.

	F_1 incorrect	F_1 correct
F_2 incorrect	A	B
F_2 correct	C	D

The McNemar's test statistic is calculated as in Eq. S1:

$$x = \frac{(|B - C| - 1)^2}{B + C} \tag{S1}$$

Under the null hypothesis, this statistic has a χ^2-distribution with one degree of freedom if both B and C are large enough, i.e. one hundred and more. The null hypothesis is that the performance of the models F_1 and F_2 is in fact the same and any apparent difference is accidental. Therefore, we can reject the null hypothesis with p-value > 0.05 if x exceeds

~ 3.83, and with p-value > 0.01 if x exceeds ~ 6.6.

In our experiments, we treat the product prediction models like binary classifiers, which are either correct if the generated SMILES sequence of the product is correct, or incorrect otherwise. We denote the baseline Molecular Transformer as "MT base" and the Molecular Transformer trained on the data with altered reagents as "MT new". They are compared both in mixed and separated settings on both USPTO MIT and the Reaxys test set. The corresponding contingency tables are tables S2-S5.

Table S2: The contingency table for the product models tested on the Reaxys test set in the separated setting.

Reaxys test	MT new, separated, fail	MT new, separated, correct
MT base, separated, fail	11751	3444
MT base, separated, correct	3123	78411

The McNemar's test statistic for table S2 is equal to 15.6.

Table S3: The contingency table for the product models tested on the Reaxys test set in the mixed setting.

Reaxys test	MT new, mixed, fail	MT new, mixed, correct
MT base, mixed, fail	12768	4623
MT base, mixed, correct	3650	75688

The McNemar's test statistic for table S3 is equal to 114.2.

The McNemar's test statistic for table S4 is equal to 11.3.

The McNemar's test statistic for table S5 is equal to 22.3.

Table S4: The contingency table for the product models tested on USPTO MIT in the separated setting.

USPTO MIT	MT new, separated, fail	MT new, separated, correct
MT base, separated, fail	3086	1230
MT base, separated, correct	1068	34616

Table S5: The contingency table for the product models tested on USPTO MIT in the mixed setting.

USPTO MIT	MT new, mixed, fail	MT new, mixed, correct
MT base, mixed, fail	3448	1460
MT base, mixed, correct	1215	33877

One can see that for all four tables the value of McNemar's statistic is sufficiently high to reject the null hypothesis with p-values less than 0.01. This allows concluding that the improvement in product prediction performance when the models are trained on the data with altered reagents is statistically significant.