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

Mikhail Andronov,<sup>\*,†</sup> Varvara Voinarovska,<sup>‡</sup> Natalia Andronova,<sup>¶</sup> Michael Wand,<sup>†,§</sup> Djork-Arné Clevert,<sup>∥</sup> and Jürgen Schmidhuber<sup>†,⊥</sup>

†IDSIA, USI, SUPSI, 6900 Lugano, Switzerland

‡Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz
 Munich – Deutsches Forschungszentrum f
ür 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

 $\perp AI$  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.

							,	p		ross classes
	Acylation and related processes	C-C bond formation	Deprotections	Functional group addition (FGA)	Functional group interconversion (FGI)	Heteroatom alkylation and arylation	Heterocycle formation	Oxidations	Protections	Reductions
Catalyst -	0.908	0.514	0.935	0.959	0.841	0.889	0.931	0.895	0.882	
Ox -	0.985	0.996	0.991	0.609	0.953	0.989	0.92	0.797	0.991	0.981
Red -	0.983	0.981	0.919	0.999	0.96	0.804		0.889	0.972	
Acid -	0.842	0.927	0.766	0.888	0.927	0.891		0.921	0.87	0.806
Base -		0.569	0.762		0.819		0.54		0.447	
Unspecified -		0.241	0.488		0.652		0.46	0.447	0.526	0.399
Solvent -	0.318	0.233	0.332	0.511	0.355	0.352	0.276	0.314	0.338	0.329
				Top-1 acc	uracy of pre	dictions by i	roles across	classes for	nonempty g	round truth
Catalyst -	0.426	0.379	0.564	0.0556	0.598	0.445		0.0603	0.415	0.423
Ox -	0.0271	0.109	0.0133	0.494	0.399	0.0636		0.0516	0.0286	0.0347
Red -	0.357	0.509	0.525	1	0.432	0.367	0.5	0.838	0.239	0.458
Acid -	0.454	0.272	0.319	0.347	0.381	0.223	0.0278	0.036	0.213	0.172
Base -	0.36		0.634	0.158		0.594	0.125	0.0266	0.334	0.5
Unspecified -	0.186	0.149	0.269	0.0989	0.265		0.309	0.484	0.2	0.343
Solvent -	0.297	0.214	0.335		0.32	0.341	0.256	0.32	0.346	0.339
							Number	of non-emp	ty ground t	ruth SMILES
Catalyst -	2556	8266	1770		1280	3962		116	786	3387
Ox -	221	46	150	781	286	283		407	35	202
Red -	429	352	2618		88	7307	30	1089	155	
Acid -	2585	827	4200	222	291	2858	36	139	503	1836
Base -	10349	8977	9004	620	1410	15221	40	414	3373	4212
Unspecified -	8056	10067	10768	465	1680	22598	55	1600	2362	8653
Solvent -		10693	15848	2056	3459	25217	82	1860	4131	9870

Top-1 accuracy of predictions by roles across classes

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	В
$F_2$ 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  $\chi^2$ -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	MT new,	MT new,			
test	separated,	separated,			
lest	fail	correct			
MT base,					
separated,	11751	3444			
fail					
MT base,					
separated,	3123	78411			
correct					

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	${\rm the}$	product	$\operatorname{models}$	tested	on	USPTO	MIT	in t	the
separated	settin	g.											

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

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.