"Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models[†]

Philippe Schwaller,[‡] Theophile Gaudin,[‡] David Lanyi, Costas Bekas and Teodoro Laino

[a] IBM Research, Zurich, Switzerland.
E-mail: {phs,tga,dla,bek,teo}@zurich.ibm.com
‡ P. S. and T. G. contributed equally to this work.

1 Predictions on recent patent reaction

Using the model trained with stereochemical on Lowe's data, containing reactions from granted patents until September 2016, we predicted the 15418 reactions of the Pistachio database[1,2]. To have a time split we selected all reactions from 2017 with a yield of more than 50% and a single product. Reactions that had the same reactants as a reaction in the training set were filtered out and prediction input duplicates, as well as reactions with incomplete product atom mappings, were removed. The pistachio database was extracted from patents with a similar, but improved workflow, compared to the open source Lowe database. Overall, a top-1 prediction accuracy of 0.60 was achieved. Table 1 shows an overview of the results. It can be seen that more than 97% of the predicted SMILES were valid according to RDKit and that the mean confidence of the invalid predictions was low with 0.41.

The following sections display examples from correctly predicted reactions belonging to diverse subclasses, an example of a falsely predicted reaction with low confidence an one of a falsely predicted reaction with a high confidence.

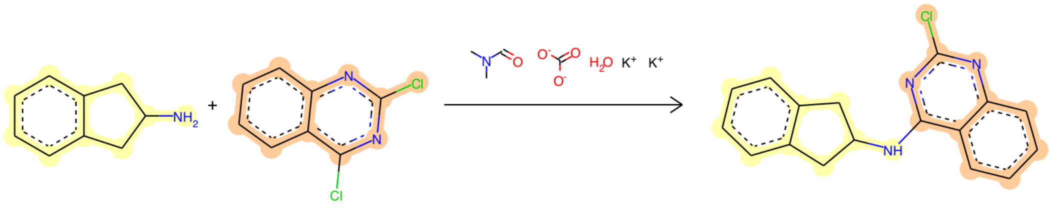
Table 1 Prediction details, classified using the super classes proposed by[3]

	Count	Accuracy	Mean confidence
Pistachio2017	15418	0.60	0.83
- Classified	11817	0.70	0.87
- Heteroatom alkylation and arylation	2702	0.73	0.87
- Acylation and related processes	2601	0.82	0.91
- Deprotections	1232	0.69	0.86
- C-C bond formation	329	0.56	0.79
- Functional group interconversion (FGI)	315	0.54	0.84
- Reductions	1996	0.72	0.87
- Functional group addition (FGA)	1090	0.72	0.88
- Heterocycle formation	310	0.58	0.84
- Protections	868	0.53	0.84
- Oxidations	339	0.41	0.80
- Resolutions	35	0.34	0.73
- Unrecognized	3601	0.27	0.68
Invalid SMILES	429	0.00	0.41
With stereochemistry	4103	0.48	0.76
Without stereochemistry	11315	0.64	0.85

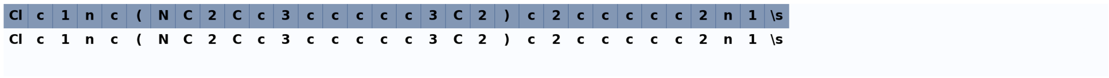
2 Correct predictions

Chloro N-arylation

Namerxn	1.3.7	**Patent**	US20170001976A1	**Yield**	66%
Reactants	Cl c 1 n c (Cl) c 2 c c c c c 2 n 1 . N C 1 C c 2 c c c c c 2 C 1				
Reagents	A_O A_CN(C)C=O A_[K+] A_O=C([O-])[O-]				
Products	Cl c 1 n c (N C 2 C c 3 c c c c c 3 C 2) c 2 c c c c c 2 n 1				
Prediction	Clc1nc(NC2Cc3ccccc3C2)c2ccccc2n1				
Confidence	1.00				**True**

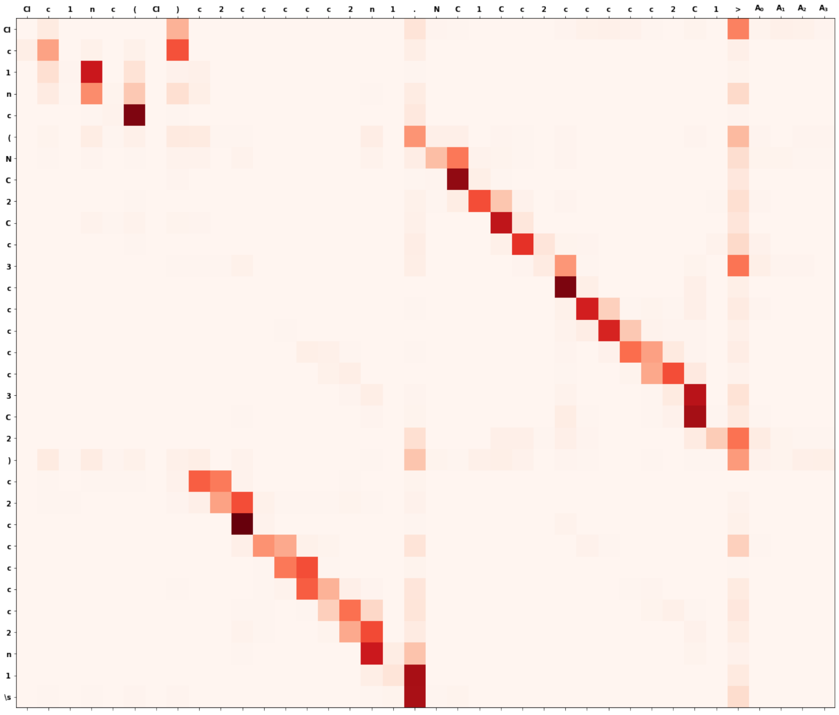


(a) Reaction plotted with rdkit[4]



Cl	c	1	n	c	(	N	C	2	C	c	3	c	c	c	c	c	3	C	2	)	c	2	c	c	c	c	c	2	n	1	\s
Cl	c	1	n	c	(	N	C	2	C	c	3	c	c	c	c	c	3	C	2	)	c	2	c	c	c	c	c	2	n	1	\s

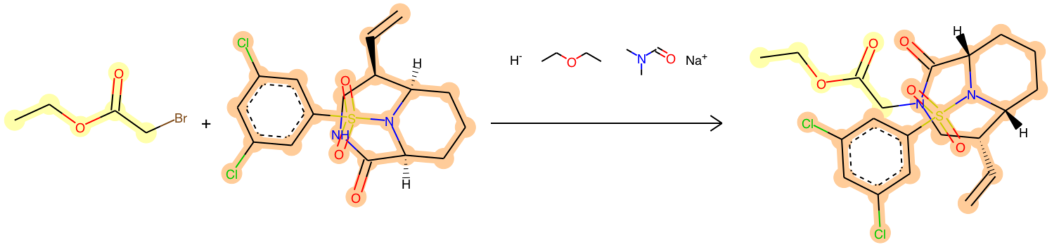
(b) Predicted output compared with token probabilities to true output



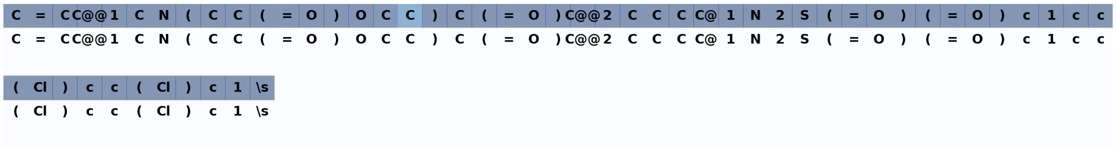
(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

Bromo N-alkylation

Namerxn	1.6.2	**Patent**	US20170002003A1	**Yield**	53%
Reactants	C = C [C@@H] 1 C N C (= O) [C@@H] 2 C C C [C@H] 1 N 2 S (= O) (= O) c 1 c c (Cl) c c (Cl) c 1 . C C O C (= O) C Br				
Reagents	A_CN(C)C=O A_[H-] A_CCOCC A_[Na+]				
Products	C = C [C@@H] 1 C N (C C (= O) O C C) C (= O) [C@@H] 2 C C C [C@H] 1 N 2 S (= O) (= O) c 1 c c (Cl) c c (Cl) c 1				
Prediction	C=C[C@@H]1CN(CC(=O)OCC)C(=O)[C@@H]2CCC[C@H]1N2S(=O)(=O)c1cc(Cl)cc(Cl)c1				
Confidence	0.76			**True**	



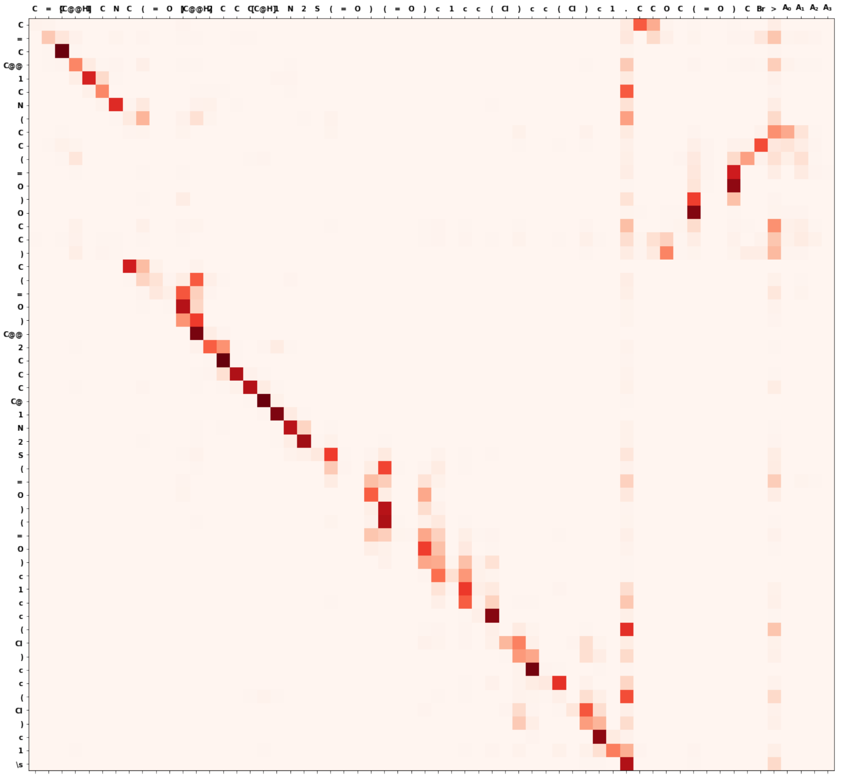
(a) Reaction plotted with rdkit[4]



C	=	C	C@@	1	C	N	(	C	C	(	=	O	)	O	C	C	)	C	(	=	O	)	C@@	2	C	C	C	C@	1	N	2	S	(	=	O	)	(	=	O	)	c	1	c	c
C	=	C	C@@	1	C	N	(	C	C	(	=	O	)	O	C	C	)	C	(	=	O	)	C@@	2	C	C	C	C@	1	N	2	S	(	=	O	)	(	=	O	)	c	1	c	c

(	Cl	)	c	c	(	Cl	)	c	1	\s
(	Cl	)	c	c	(	Cl	)	c	1	\s

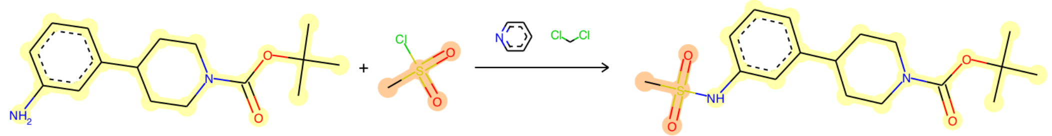
(b) Predicted output compared with token probabilities to true output



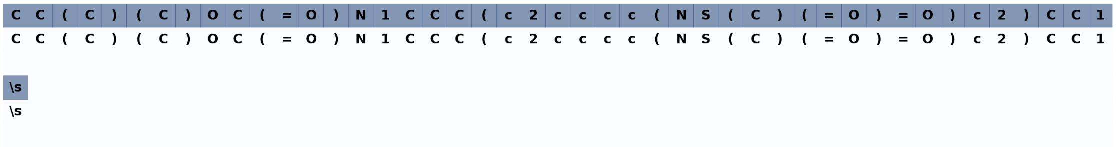
(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

Sulfonamide Schotten-Baumann

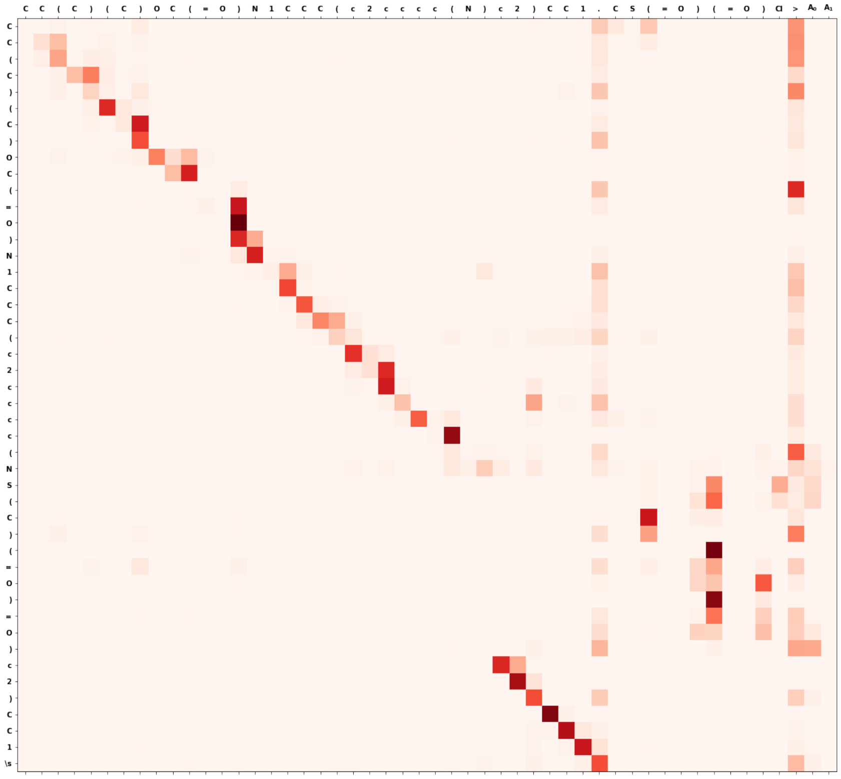
Namerxn	2.2.3	Patent	US20170001978A1	Yield	97%
Reactants	C C (C) (C) O C (= O) N 1 C C C (c 2 c c c c (N) c 2) C C 1 . C S (= O) (= O) Cl				
Reagents	A_c1ccncc1 A_ClCCl				
Products	C C (C) (C) O C (= O) N 1 C C C (c 2 c c c c (N S (C) (= O) = O) c 2) C C 1				
Prediction	CC(C)(C)OC(=O)N1CCC(c2cccc(NS(C)(=O)=O)c2)CC1				
Confidence	1.00			**True**	



(a) Reaction plotted with rdkit[4]


C C (C) (C) O C (= O) N 1 C C C (c 2 c c c c (N S (C) (= O) = O) c 2) C C 1
C C (C) (C) O C (= O) N 1 C C C (c 2 c c c c (N S (C) (= O) = O) c 2) C C 1
\s
\s

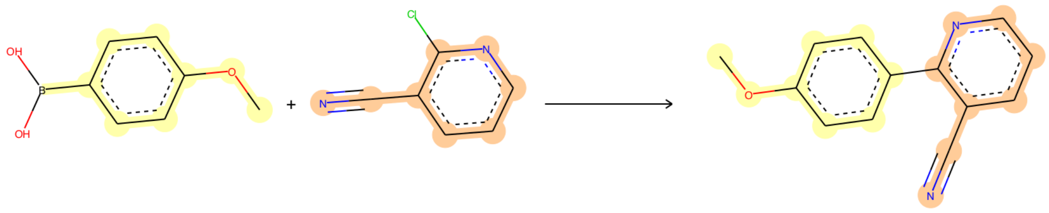
(b) Predicted output compared with token probabilities to true output



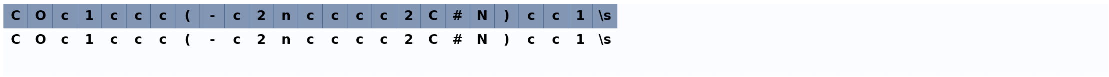
(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

Chloro Suzuki-type coupling

Namerxn	3.1.6	**Patent**	US20170001964A1	**Yield**	95%
Reactants	C O c 1 c c c (B (O) O) c c 1 . N # C c 1 c c c n c 1 Cl				
Reagents					
Products	C O c 1 c c c (- c 2 n c c c c 2 C # N) c c 1				
Prediction	COc1ccc(-c2ncccc2C#N)cc1				
Confidence	1.00				**True**

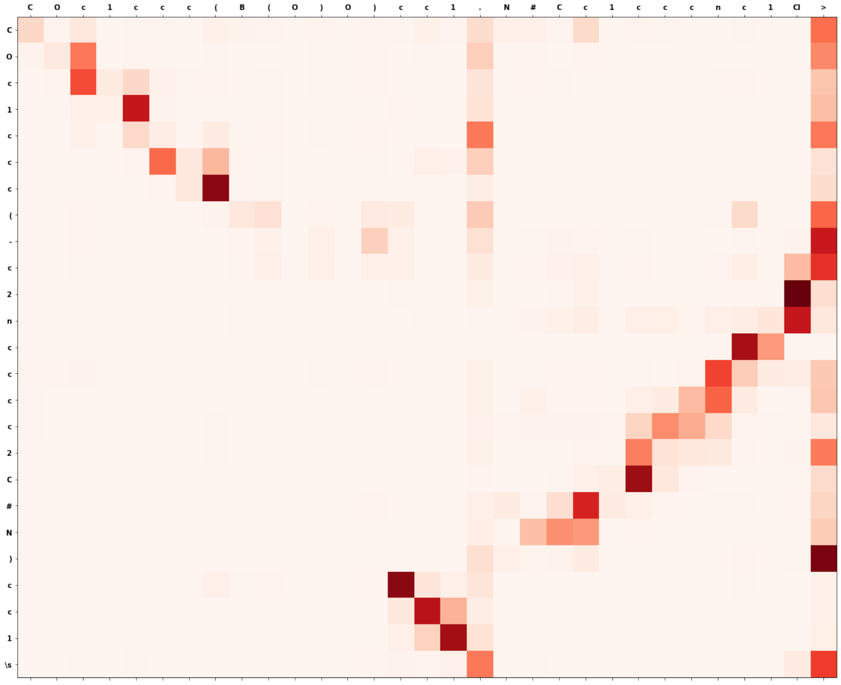


(a) Reaction plotted with rdkit[4]



C	O	c	1	c	c	c	(	-	c	2	n	c	c	c	c	2	C	#	N	)	c	c	1	\s
C	O	c	1	c	c	c	(	-	c	2	n	c	c	c	c	2	C	#	N	)	c	c	1	\s

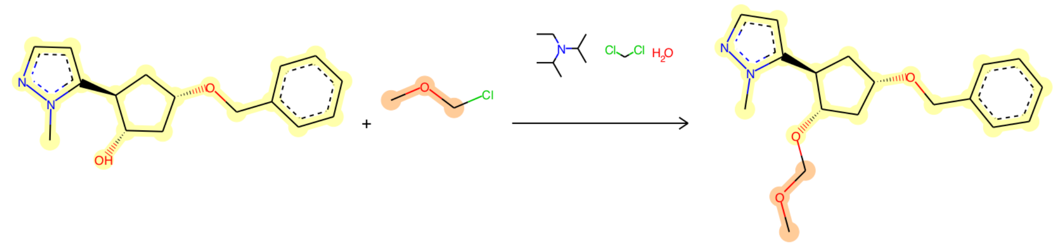
(b) Predicted output compared with token probabilities to true output



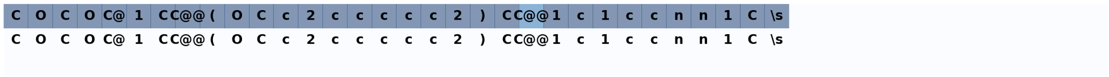
(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

O-MOM protection

Namerxn	5.3.6	Patent	US20170001984A1	Yield	62%
Reactants	C O C Cl . C n 1 n c c c 1 [C@H] 1 C [C@H] (O C c 2 c c c c c 2) C [C@@H] 1 O				
Reagents	A_O A_CCN(C(C)C)C(C)C A_ClCCl				
Products	C O C O [C@H] 1 C [C@@H] (O C c 2 c c c c c 2) C [C@@H] 1 c 1 c c n n 1 C				
Prediction	COCO[C@H]1C[C@@H](OCc2ccccc2)C[C@@H]1c1ccnn1C				
Confidence	0.78				True

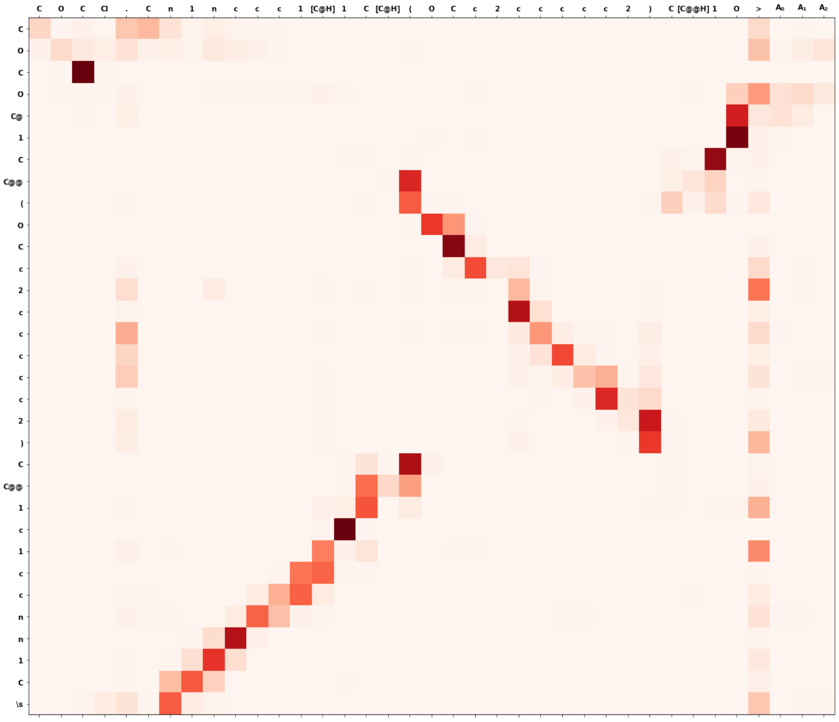


(a) Reaction plotted with rdkit[4]



```
C  O  C  O C@ 1  CC@@ (  O  C  c  2  c  c  c  c  c  2  )  CC@@1  c  1  c  c  n  n  1  C  \s
C  O  C  O C@ 1  CC@@ (  O  C  c  2  c  c  c  c  c  2  )  CC@@1  c  1  c  c  n  n  1  C  \s
```

(b) Predicted output compared with token probabilities to true output

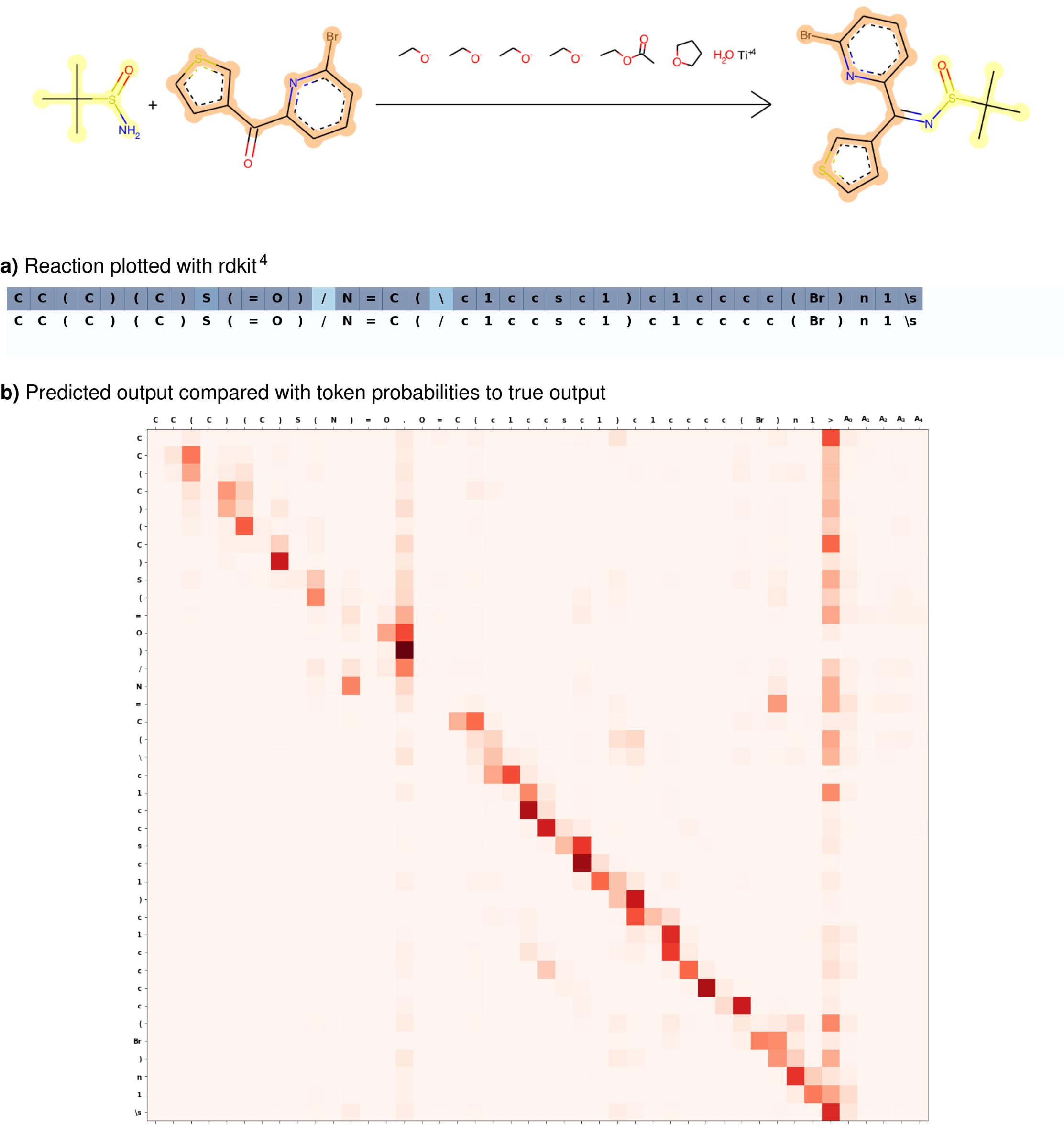


(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

3 Example of a false prediction with low confidence

Ketone reductive imination

Namerxn	1.2.6	**Patent**	US20170001990A1	**Yield**	61%
Reactants	C C (C) (C) S (N) = O . O = C (c 1 c c s c 1) c 1 c c c c (Br) n 1				
Reagents	A_CC[O-] A_C1CCOC1 A_O A_[Ti+4] A_CCOC(C)=O				
Products	C C (C) (C) S (= O) / N = C (/ c 1 c c s c 1) c 1 c c c c (Br) n 1				
Prediction	CC(C)(C)S(=O)/N=C(1̧ccsc1)c1cccc(Br)n1				
Confidence	0.30				**False**


Br
S
N
O
NH2
O
O
O
O
O
O
O
H2O Ti+4
Br
N
S
O
N
S
C C (C) (C) S (= O) / N = C (\ c 1 c c s c 1) c 1 c c c c (Br) n 1 \s
C C (C) (C) S (= O) / N = C (/ c 1 c c s c 1) c 1 c c c c (Br) n 1 \s
C C (C) (C) S (N) = O . O = C (c 1 c c s c 1) c 1 c c c c (Br) n 1 > A₀ A₁ A₂ A₃ A₄
C C (C) (C) S (= O) / N = C (\ c 1 c c s c 1) c 1 c c c c (Br) n 1 \s

(a) Reaction plotted with rdkit[4]

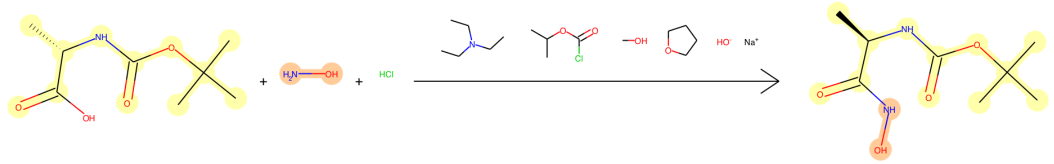
(b) Predicted output compared with token probabilities to true output

(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

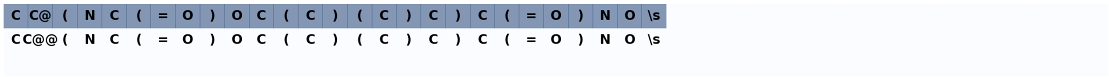
4 Example of a false prediction with high confidence

Carboxylic acid + amine condensation

Namerxn	2.1.2	**Patent**	US20170002015A1	**Yield**	55%
Reactants	C [C@H] (N C (= O) O C (C) (C) C) C (= O) O . N O				
Reagents	A_C1CCOC1 A_CC(C)OC(=O)Cl A_CCN(CC)CC A_Cl A_[Na+] A_[OH-] A_CO				
Products	C [C@@H] (N C (= O) O C (C) (C) C) C (= O) N O				
Prediction	C[C@H](NC(=O)OC(C)(C)C)C(=O)NO				
Confidence	0.98				**False**

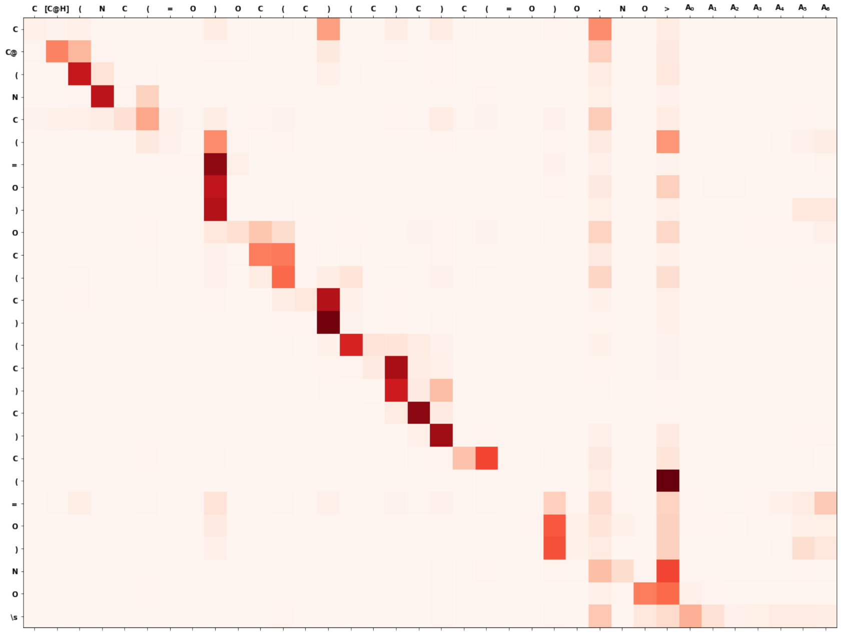


(a) Reaction plotted with rdkit[4]



C	C@	(	N	C	(	=	O	)	O	C	(	C	)	(	C	)	C	)	C	(	=	O	)	N	O	\s
C	C@@	(	N	C	(	=	O	)	O	C	(	C	)	(	C	)	C	)	C	(	=	O	)	N	O	\s

(b) Predicted output compared with token probabilities to true output



(c) Attention weight matrix. Input tokens horizontal, output tokens vertical.

Notes and references

1 N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli and G. A. Landrum, *J. Med. Chem.*, 2016, **59**, 4385–4402.

2 https://www.nextmovesoftware.com/pistachio.html.

3 J. S. Carey, D. Laffan, C. Thomson and M. T. Williams, *Org. Biomol. Chem.*, 2006, **4**, 2337–2347.

4 G. Landrum, B. Kelley, P. Tosco, S. Riniker, Gedeck, N. Schneider, R. Vianello, A. Dalke, S. Alexander, S. Turk, M. Swain, B. Cole, JP, Strets123, JLVarjo, A. Pahl, P. Fuller, G. Doliath, M. Wójcikowski, D. Cosgrove, G. Sforna, M. Nowotka, J. H. Jensen, J. Domański, D. Hall, N. O'Boyle, W.-G. Bolick, Nhfechner and S. Roughley, *Rdkit/Rdkit: 2017_09_1 (Q3 2017) Release*, 2017, https://zenodo.org/record/1004356#.Wd3LDY6l2EI.