ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2019, 55, 12152

Received 4th July 2019,
Accepted 29th August 2019

DOI: 10.1039/c9cc05122h

rsc.li/chemcomm

Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space

Alpha A. Lee, *[a] Qingyi Yang,[b] Vishnu Sresht,[c] Peter Bolgar,[d] Xinjun Hou,[b] Jacquelyn L. Klug-McLeod[e] and Christopher R. Butler[b]

Predicting how a complex molecule reacts with different reagents, and how to synthesise complex molecules from simpler starting materials, are fundamental to organic chemistry. We show that an attention-based machine translation model – Molecular Transformer – tackles both reaction prediction and retrosynthesis by learning from the same dataset. Reagents, reactants and products are represented as SMILES text strings. For reaction prediction, the model "translates" the SMILES of reactants and reagents to product SMILES, and the converse for retrosynthesis. Moreover, a model trained on publicly available data is able to make accurate predictions on proprietary molecules extracted from pharma electronic lab notebooks, demonstrating generalisability across chemical space. We expect our versatile framework to be broadly applicable to problems such as reaction condition prediction, reagent prediction and yield prediction.

Despite decades of methodological advances, organic synthesis can still be a hurdle in drug discovery. At the crux of organic synthesis are two interrelated challenges: (1) retrosynthesis – given a complex molecule, how can one chemically synthesise it from simpler commercially available molecules *via* a series of reactions? (2) Reaction prediction – given reactants and reagents, what are the possible products? Pioneering attempts at computational reaction prediction and synthesis planning, dating back to the 1960s, relied upon codifying chemical heuristics by manually curating reaction rules.[1,2] However, those rules are limited by either personal expertise or incomplete sourcing of the chemical literature. More recent approaches employ heuristics to automatically extract reaction rules ("templates") directly from data. Those methods have been applied to reaction prediction[3,4] and retrosynthesis.[4–8] However, template-based methods cannot generalise beyond the templates, thus its predictive power is circumscribed. Moreover, automatic template extraction algorithms rely on atom mapping – a scheme that maps atoms in the reactants to atoms in the product. Common atom-mapping tools are themselves based on libraries of expert rules and templates, thus creating a vicious circle that misses important data.

Going beyond templates, recent machine learning approaches have focused on developing template-free approaches based on graph or sequence representation of molecules. Graph-based approaches consider molecules as graphs where atoms are nodes and bonds are edges, and predict graph edits that transform reactants to product.[9–11] However, those approaches require an atom-mapped training dataset, thus implicitly rely on predefined templates. To date, graph-based approaches have been developed for reaction prediction but not retrosynthesis. Sequence-based approaches represent molecules as SMILES strings, and treat reaction prediction and retrosynthesis as a machine translation problem – either translating from the SMILES of the reactants and reagents to the SMILES of the product,[12,13] or *vice versa* for retrosynthesis,[14,15] and completely avoid atom mapping. In particular, some of us showed that the Molecular Transformer model outperforms all methods in the literature in reaction prediction,[13] but retrosynthesis has not been described therein.

Beyond accuracy, dataset bias was recently recognised as a major challenge in validating machine learning models in chemistry applications. Recent pioneering works show that, in the context of virtual screening, machine learning models can achieve high accuracy by memorising the training set if the training and test sets are very similar.[16,17] Within the context of forward reaction prediction and retrosynthesis, the benchmark dataset used in the literature[9–14] is harvested from medicinal chemistry patents, which often report synthesis of highly similar analogues.

In this Letter, we show that Molecular Transformer is a versatile framework that tackles both reaction prediction and retrosynthesis with high accuracy, and predicts plausible disconnections for drug-like molecules. Going beyond accuracy, we demonstrate the generalisability of the Molecular Transformer by showing that a model trained on publicly available data is able

[a] *Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK. E-mail: aal44@.cam.ac.uk*
[b] *Medicine Design, Pfizer Inc., Cambridge, MA 02139, USA*
[c] *Simulations and Modelling Sciences, Pfizer Inc., Cambridge, MA 02139, USA*
[d] *Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK*
[e] *Simulations and Modelling Sciences, Pfizer Inc., Groton, CT 06357, USA*

to make accurate prediction on proprietary molecules and reactions extracted from electronic lab notebooks. To our knowledge, this is the first validation of reaction prediction tools with industrial pharma data, and Molecular Transformer is the only framework that accurately tackles both reaction prediction and retrosynthesis.

The Molecular Transformer model is described in detail in ref. 13 and available on GitHub (https://github.com/pschwllr/MolecularTransformer). In this Letter, we consider 3 datasets: first, the USPTO dataset (referred to hereafter as USPTO), compiled by Lowe by digitising patents.† This dataset contains stereochemical information, and is more challenging than the data commonly used in the literature.[9–11] Second, a set of proprietary electronic lab notebook data from internal medicinal chemistry projects in Pfizer. This dataset contains 147 392 reactions with 2 reactants (referred to hereafter as PF-ELN), split into training/validation/test sets (60%/20%/20%). Third, 50 000 reactions drawn from diverse reaction classes in USPTO (referred to hereafter as USPTO-R), curated by ref. 18 and used in ref. 5 and 14 to benchmark retrosynthesis models. Stereochemical information is removed from USPTO-R, and is randomly split into training/validation/test set (80%/10%/10%). We canonicalised SMILES using the open-source package rdkit.

Reaction prediction and chemical space extrapolation. Previous work by some of us[13] showed that Molecular Transformer outperforms the state-of-the-art on publicly available datasets, with graph neural networks (WLDN)[11] being the second best. However, a looming question is whether the model can achieve a comparable level of accuracy on chemical space and reactions relevant to the pharmaceutical industry. To answer this question, we re-trained the model on Pfizer electronic lab notebook data using the training procedure and hyperparameters reported in ref. 13. Table 1 shows that Molecular Transformer achieves 97% top-1 accuracy. We note that it is a challenge to train WLDN on the full PF-ELN dataset as many reactions cannot fully meet the atom-mapping requirements of WLDN; Molecular Transformer, in contrast, can be trained on incomplete reaction information typically present in private lab notebooks.

The recent literature[17] highlighted the need to control for analogue bias by ensuring that the training set is not trivial chemical analogues of the test set. To this end, we evaluate the model trained on USPTO data and evaluated on PF-ELN test set. The two datasets are drawn from distinct regions of chemical space as only 6% of reactants or products of PF-ELN can be found in the public domain. Fig. 1 shows reactions within PF-ELN and USPTO are similar to each other, whereas reactions are very different across PF-ELN and USPTO. Evaluating the model performance that is trained on USPTO data and evaluated on PF-ELN is thus a reasonable test of generalisability.

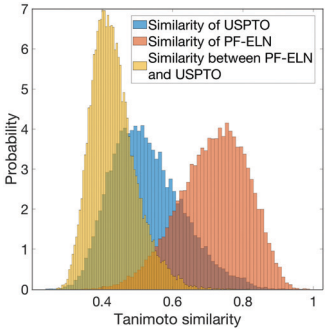

Fig. 1 PF-ELN and USPTO are drawn from distinct chemical spaces. The histogram shows the distribution of similarity between reactions within PF-ELN and USPTO, and between PF-ELN and USPTO. A reaction is represented by concatenating ECFP4 fingerprints of reactants, and similarity is quantified by the mean Tanimoto similarity between the reaction and its 5 nearest neighbours in the dataset. Ambiguity in reactant ordering is avoided by permuting different orderings and taking the highest similarity.

Table 1 shows that, perhaps unsurprisingly, the performance of the model decreases, but the top-2 accuracy is still 80.3% (top-1 accuracy of 69%), which is substantially higher than the accuracy of WLDN (top-1/top-2 accuracy of 50.4%/51.7%). The WLDN model has a preprocessing step that cleans and prunes the USPTO data as well as remove stereochemical information, whereas Molecular Transformer can work directly on USPTO data.

As synthesis is a multistep process where one unsuccessful reaction can derail an entire synthesis campaign, perhaps more important than predicting the correct product is whether the model can accurately calibrate of confidence of its own predictions and return a low confidence when outside its domain of applicability. Molecular Transformer outputs the log-likelihood of a prediction, which can be interpreted as the estimated probability of that prediction being correct. Fig. 2 shows that we can use the log-likelihood to accurately classify whether a prediction is incorrect.

Retrosynthesis prediction. We next consider the inverse problem – given a product, what are plausible reactants? Retrosynthesis is a challenging because unlike reaction prediction, the retrosynthesis problem has no "correct" answer. There are many possible ways to

Table 1 Molecular Transformer accurately predicts products of reactions given reactants and is generalisable across chemical space. The table show the top-*k* accuracy of Molecular Transformer tested on proprietary electronic lab notebook data for different training sets

Model	Training set	Test set	Top-1 [% acc.]	Top-2 [% acc.]	Top-3 [% acc.]
Molecular Transformer	PF-ELN	PF-ELN	97.0	98.5	98.8
Molecular Transformer	USPTO	PF-ELN	69.0	80.3	82.9
WLDN[11]	USPTO	PF-ELN	50.4	51.7	52.2

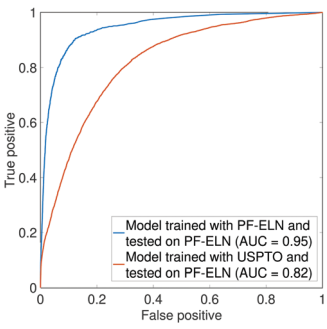

Fig. 2 Molecular Transformer accurately estimate the uncertainty of its own prediction. The quality of the uncertainty estimate is quantified by the receiver operating characteristic of using the log-likelihood to classify whether the model prediction is incorrect.

Table 2 Molecular Transformer for retrosynthesis accurately predicts reactants of reactions given products and outperforms the state-of-the-art. The table show the top-*k* accuracy of Molecular Transformer trained and tested on USPTO-R. The top-2 accuracy is not reported in ref. 5 and 14

Model	Training set	Test set	Top-1 [% acc.]	Top-2 [% acc.]	Top-3 [% acc.]
Molecular Transformer	USPTO-R	USPTO-R	43.8	56.0	60.5
Coley *et al.*[5]	USPTO-R	USPTO-R	37.3	—	54.7
Liu *et al.*[14]	USPTO-R	USPTO-R	37.4	—	52.4

build a molecule from simpler reactants. Following previous work,[5,14] we benchmark Molecular Transformer using a simple criterion: when given the products of reactions in the dataset, do we recover and rank highly the recorded reactants in that dataset without having seen that reaction previously? We train Molecular Transformer with the product SMILES string as the input and the reactant SMILES strings as output.

Table 2 shows that Molecular Transformer outperforms the benchmark model,[5] which uses a template-based approach. Liu *et al.*[14] likewise employs a machine translation approach. However, the attention mechanism in Molecular Transformer captures long-ranged interactions between SMILES tokens thus outperforms Liu *et al.*[14] In Table 2, all models are compared using the USPTO-R dataset, the common benchmark dataset used by the community.

To estimate the effect of bias due to prevalence of chemically similar compounds, we next evaluate the performance of Molecular Transformer trained on USPTO-R and deployed on PF-ELN. Table 3 shows that the top-1 accuracy is still 31.5% despite the training and test sets come from distinct regions of chemical space, demonstrating generalisability. We note that the ASKOS method developed by Coley *et al.*[8] achieves a top-1 accuracy of 27.6% on PF-ELN; we use an offline version of their pretrained model. Unsurprisingly, Molecular Transformer is much more accurate when trained and tested on PF-ELN, achieving top-1 accuracy of 91.0%. ASKOS is not retrained on PF-ELN because of the aforementioned atom-mapping constraint.

Qualitative evaluation of retrosynthesis predictions. Having quantitatively evaluated Molecular Transformer for one-step retrosynthesis, we next qualitatively discuss how Molecular Transformer can be applied to multistep retrosynthesis. We retrain the model combining USPTO-R and PF-ELN training sets, and apply it to 4 drug-like organic molecules (Fig. 3). We assume that substituted benzenes, quinolines and quinazolines are assessable starting materials and consider up to 3 disconnections.

Table 3 Molecular Transformer for retrosynthesis is accurate for chemistries that are industrially relevant, and generalises across chemical space (*c.f.* accuracy of the model trained and tested on USPTO-R reported in Table 2). The table show the top-*k* accuracy of Molecular Transformer with different training set-test set combinations

Model	Training set	Test set	Top-1 [% acc.]	Top-2 [% acc.]	Top-3 [% acc.]
Molecular Transformer	USPTO-R	PF-ELN	31.5	40.5	44.3
Molecular Transformer	PF-ELN	PF-ELN	91.0	94.5	94.5

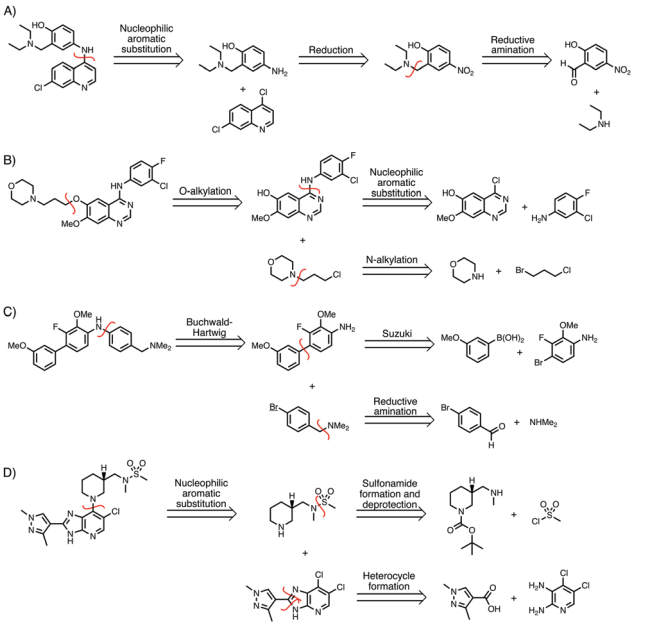

Fig. 3 Example of synthesis routes of bioactive organic molecules predicted by Molecular Transformer.

Fig. 3A shows the predicted synthesis route of amodiaquine, an antimalarial drug. Molecular Transformer first disconnects one of the C–N bonds of the secondary amine. In the corresponding forward step regioselective nucleophilic substitution is expected to happen at position 4 in preference to position 7, as predicted. Following a functional group interconversion, Molecular Transformer proceeds to disconnect the tertiary amine. This is a reasonable sequence of disconnections, as in the forward synthesis the reductive amination will have to be carried out before the reduction of the nitro group to avoid competition between the two amino groups. Reported synthetic routes to amodiaquine[19] introduce the *N*,*N*-diethylaminomethylene group ortho to the hydroxyl group of 4-hydroxyacetanilide by an electrophilic aromatic substitution with *in situ* generated *N*,*N*-diethylmethaniminium ion. By-products arsing from the weak regioselectivity of substitution and from double substitution are often observed.

Fig. 3B shows the predicted synthesis route of gefitinib, an EGFR inhibitor. Molecular Transformer first disconnects the C–O bond of aryl ether. Then the alkyl chain is cleaved off the morpholine moiety to give 1-chloro-3-bromopropane and morpholine as starting materials. Molecular Transformer correctly infers that when using 1-chloro-3-bromopropane as the starting material, in the first reaction bromide will be preferentially replaced. The retrosynthesis of the substituted quinazoline unit continues by the disconnection of the secondary amine. The disconnection places the chlorine atom on the quinazoline ring and the amino group on the benzene ring, accounting for the electronic requirements of nucleophilic aromatic substitutions. This route is similar to the route disclosed in the original patent,[20] although the compound is not in our training set.

Fig. 3C shows the predicted synthesis route of a RAS-effector inhibitor that has been recently reported in the literature.[21]

Molecular Transformer predicts the same sequence of disconnections as it is reported in the literature,[21] although the compound was published very recently and not contained in the training set. This demonstrates the generalisability of Molecular Transformer across chemical space to novel molecular scaffolds.

Fig. 3D shows the predicted synthesis route of an aromatic heterocycle which inhibits Aurora-A kinase.[22] The first disconnection, a regioselective nucleophilic aromatic substitution, gives a tricyclic aromatic heterocycle and a secondary amine. The sulfonamide is disconnected to mesyl chloride and an amine, while accounting for the fact that one of the amines will have to be protected when forming the sulfonamide. Next, one of the 5-membered rings of the aromatic heterocycle is disconnected to a carboxylic acid and a diamine (selective formation of the desired heterocycle is likely, see ref. 23). The major difference between this and the reported synthesis[22] lies in the synthesis of the heterocyclic core. The reported route condenses an *o*-nitroaniline and an aromatic aldehyde *via* reductive cyclisation to assemble the heterocyclic core. Molecular Transformer suggests a qualitatively different but chemically equally plausible method, creating an orthogonal route to the target.

In summary, we show that Molecular Transformer is a versatile framework that tackles both reaction prediction and retrosynthesis. Using typical examples of predicted retrosynthetic routes, we illustrate how Molecular Transformer has inferred the logic of chemical synthesis, paying attention to nuances and subtle selectivities. We further demonstrate that Molecular Transformer is generalisable across chemical space by evaluating its performance trained reactions reported in US patents and tested on proprietary electronic lab notebook reactions, reading across chemical reactivity patterns inferred from one dataset to structurally novel compounds in another dataset. A key advantage of our framework is that it can be easily extended to tackle problems such as predicting reaction condition, reagents and yield.

We wish to highlight two subtleties in retrosynthesis prediction that are beyond this Letter's scope: first, synthetic route prediction is often not a matter of simple interpolation – a small change in the target product can require a significant change in the synthesis route, and algorithms must recognise those retrosynthesis "cliffs". Second, evaluation metrics used in the literature (including this Letter) describe whether the model predicts how a compound was made in the lab. This is not necessarily how the compound could be made or the "best" way to make it, which is the actual objective of retrosynthesis route design. These subtleties perhaps explain why the accuracy increase going from top-1 to top-3 is modest across all machine learning models (*cf.* Table 2).

We thank Chris Helal, Martin Pettersson for feedback on predicted synthetic routes, Gregory Bakken for software engineering support, Philippe Schwaller for insightful discussions, and support from Winton Programme for the Physics of Sustainability (AAL) and Herchel Smith Fund (PB).

Conflicts of interest

QY, VS, XH, JLK and CRB are employees of Pfizer Inc.

Notes and references

† http://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

1 E. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
2 B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos and T. Klucznik, *Chem*, 2018, **4**, 390–398.
3 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
4 M. H. Segler and M. P. Waller, *Chem. – Eur. J.*, 2017, **23**, 5966–5971.
5 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.
6 M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604.
7 J. S. Schreck, C. W. Coley and K. J. M. Bishop, *ACS Cent. Sci.*, 2019, **5**, 970–981.
8 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, eaax1566.
9 W. Jin, C. Coley, R. Barzilay and T. Jaakkola, *Advances in Neural Information Processing Systems*, 2017, pp. 2607–2616.
10 J. Bradshaw, M. J. Kusner, B. Paige, M. H. Segler and J. M. Hernández-Lobato, 2018, arXiv preprint arXiv:1805.10970.
11 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.
12 P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.
13 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, DOI: 10.1021/acscentsci.9b00576.
14 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
15 P. Karpov, G. Godin and I. Tetko, *ChemRxiv*, DOI: 10.26434/chemrxiv.8058464.
16 I. Wallach and A. Heifets, *J. Chem. Inf. Model.*, 2018, **58**, 916–932.
17 J. Sieg, F. Flachsenberg and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 947–961.
18 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.
19 J. H. Burckhalter, F. H. Tbndick, E. M. Jones, P. A. Jones, W. F. Holcomb and A. L. Rawlins, *J. Am. Chem. Soc.*, 1948, **70**, 1363–1373.
20 K. H. Gibson, *Quinazoline derivatives, US Pat.*, US5770599A, 1998.
21 A. Cruz-Migoni, P. Canning, C. E. Quevedo, C. J. R. Bataille, N. Bery, A. Miller, A. J. Russell, S. E. V. Phillips, S. B. Carr and T. H. Rabbitts, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 2545–2550.
22 V. Bavetsias, A. Faisal, S. Crumpler, N. Brown, M. Kosmopoulou, A. Joshi, B. Atrash, Y. Perez-Fuertes, J. A. Schmitt, K. J. Boxall, R. Burke, C. Sun, S. Avery, K. Bush, A. Henley, F. I. Raynaud, P. Workman, R. Bayliss, S. Linardopoulos and J. Blagg, *J. Med. Chem.*, 2013, **56**, 9122–9135.
23 T. Wang, M. A. Block, S. Cowen, A. M. Davies, E. Devereaux, L. Gingipalli, J. Johannes, N. A. Larsen, Q. Su, J. A. Tucker, D. Whitston, J. Wu, H.-J. Zhang, M. Zinda and C. Chuaqui, *Bioorg. Med. Chem. Lett.*, 2012, **2**, 2063–2069.