Unassisted Noise-Reduction of Chemical Reactions Data Sets

Alessandra Toniato,[*,†] Philippe Schwaller,[†,‡] Antonio Cardinale,[†,¶] Joppe Geluykens,[†] and Teodoro Laino[†]

†*IBM Research Europe – Zurich, Rüschlikon, Switzerland*

‡*Department of Chemistry, University of Bern, Bern, Switzerland*

¶*Department of Chemistry, University of Pisa, Pisa, Italy*

E-mail: ato@zurich.ibm.com

Abstract

Existing deep learning models applied to reaction prediction in organic chemistry are able to reach extremely high levels of accuracy ($> 90\%$ for NLP- based ones[1]). With no chemical knowledge embedded than the information learnt from reaction data, the quality of the data sets plays a crucial role in the performance of the prediction models. While human curation is prohibitively expensive, the need for unaided approaches to remove chemically incorrect entries from existing data sets is essential to improve the performance of artificial intelligence models in synthetic chemistry tasks. Here we propose a machine learning-based, unassisted approach to remove chemically wrong entries (noise) from chemical reaction collections. Results show that models trained on cleaned and balanced data sets improve the quality of the predictions without a decrease in performance. For the retrosynthetic models the round-trip accuracy is enhanced by 13% and the value of the cumulative Jensen Shannon metric is lowered down to 70% of its original value, while maintaining high values of coverage (97%) and constant class-diversity (1.6) at inference.

1 Introduction

The last decade witnessed a flourishing development and application of several data-driven approaches to synthetic organic chemistry, mainly thanks to the availability of chemical reactions data sets.[2,3] The publicly available United States Patent Office (USPTO)[3] data set, along with proprietary Pistachio[4] and Reaxys[5] fueled the development of several deep learning models and architectures to assist organic chemists in chemical synthesis planning.[1,6–10] Despite efforts in building models that effectively learn chemistry from data, the quality of the data sets remains the primary limitation on performance improvements. The impact of data sets sizes and variability on the performance of computer-assisted synthesis planning tools has been recently investigated by Thakkar et al.[11]. Nevertheless, the influence of having chemically wrong examples in training data sets remains a topic of little research regardless of its relevance and impact in all data-driven chemical applications.

Some deep learning architectures may represent data more effectively than others, but the presence of chemically wrong entries in training data sets has a detrimental effect on all of them. Trying to learn from a large portion of chemically wrong examples affects the way models represent latent chemical rules, biasing the predictions towards unreasonable connections and disconnections. Unfortunately, while we observed a large variety of innovation at the level of mathematical architectures, the development of strategies to remove noise in data sets received little attention. The use of specific rule-based systems, such as the identification of ”wrong chemistry” based on the unsuccessful matching of predefined reaction templates provides a simple approach to remove incorrect chemistry. Still, the failure to map chemical reactions with existing reaction templates may lead to a loss of crucial chemical knowledge for unmatched and potentially relevant examples. On the other hand, the prohibitively expensive effort of humanly curating large data sets composed of millions of entries to create ground truth sets is hindering the development of supervised approaches to identify chemically wrong examples. The need for unaided, automatic and reliable protocols to minimize the loss of meaningful chemical knowledge, while removing errors and noise is of critical importance to bring data-driven chemical synthesis models from an experimental to an industrial readiness level.

Here, we present an efficient technique to reduce noise in chemical reactions data sets with the goal to improve the performance of existing predictive models. In particular, inspired by an approach applied to classification tasks,[12] we designed a new data noise-reduction and balancing strategy which is machine learning(ML)-based and unassisted. The main idea behind the proposed protocol relies on the *catastrophic forgetting*, the tendency of AI models to forget previously learnt events when trained on new tasks. In the context of AI-driven chemical synthesis, the cause of this behaviour can be traced back to a limited overlap between distributions of the chemical reaction features of different training batches. Similar to language models where data points more difficult to learn are likely example of a wrong grammar, the most difficult examples to learn during training of reaction prediction models are likely examples of wrong chemistry when compared to the chemical grammar described by the majority of the data set. Starting from this hypothesis, we inspect each entry of the data set for the number of times it is forgotten during training. The most forgotten examples are then removed in a certain percentage and a new model is trained with the ”cleaned” data sets, carrying a better representation of the latent chemical grammar. We show that this strategy leads to an effective statistically-based noise reduction in chemical data sets. The disclosed protocol can be used to remove chemically wrong examples from large collections of public and proprietary data sets, to improve from simple analytics to the performance and reliability of existing forward and retrosynthesis architectures.[1,9]

2 Results and discussion

2.1 The forgotten events strategy

Learning algorithms achieved state-of-the-art performances in many application domains, including chemical reaction prediction and synthesis design. When training a neural network on new data, there is a tendency to forget the information previously learnt. This usually means new data will likely override the weights that have been learnt in the past, and thus degrade the model performance for the past tasks. This behaviour is known as catastrophic forgetting or catastrophic interference[13] and is one of the main impediment to transfer learning.

The application of the principles of catastrophic forgetting within different epochs of the same training session leads to the definition of ”forgotten” data points learnt in previous epochs.[12] This behaviour may be a symptom of under-representation of the forgotten entries in the data set, outliers of the underlying feature distribution. The outliers may be carriers of some important feature rarely seen relative to its importance, but quite often they are only semantically wrong data points. Assuming the underlying feature distribution of the entire data set is a statistically correct representation of the carried knowledge, one can improve its significance by removing a certain fraction of the more frequently forgotten events. The use of domain specific statistical metrics helps in determining the maximum threshold to remove.

2.2 Forward prediction model noise-reduction

We applied this strategy to the transformer-based[14] forward model by Schwaller et al.[1] trained on the Pistachio data set. The molecular transformer is a sequence-based approach, which casts the reaction prediction as a translation problem. We used the commercial Pistachio data set, compiled using text mining on Patents,[3] properly pre-filtered (see section 4.2 for pre-filtering strategies). During training and at each epoch, we considered only the top-1 prediction out of five beams. Similar to previous transformer studies, we trained long enough to reach convergence (approx. 48 hours). During this time frame we analyzed the forgetting events across 34 epochs. Figure 1 shows the results for the forgetting experiment.

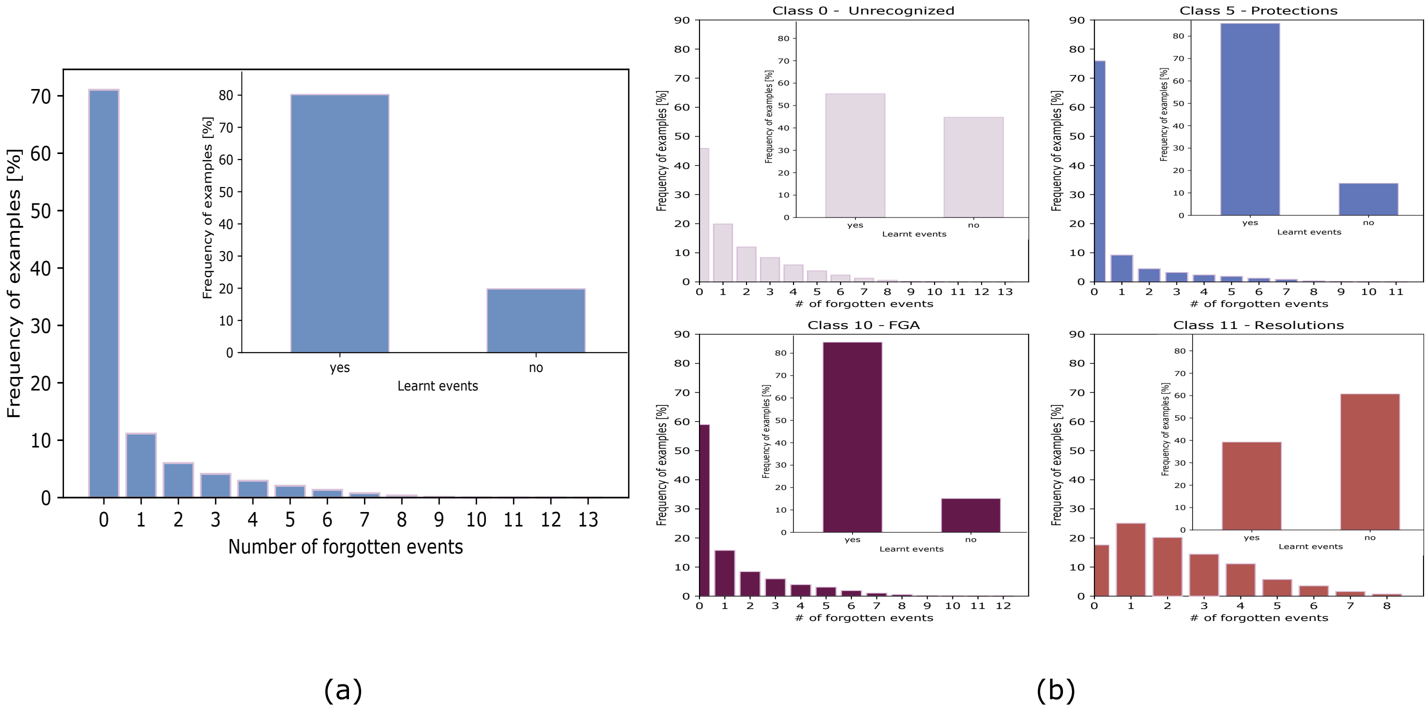


Figure 1: Results for the forgetting forward experiment: (a) Percentage of forgotten examples as a function of forgotten frequency. Notice that here the percentage refers to the number of examples that have been learnt *at least* once. The label ”0” denotes examples once learnt, are never forgotten at later epochs. In the inside plot percentage of events learnt at least once (yes) or never learnt (no). (b) Percentage of forgotten examples as a function of the forgotten frequency, for superclasses: Unrecognized, Protections, Functional Group Addition, Resolutions. In the inside plot the percentage of events learnt and never learnt, divided by macro reaction classes.

Approximately 80% of the entire data set is learnt at least once during training. Out of 80%, 70% of the examples are never forgotten by the transformer across epochs, once they are learnt. As expected, the distribution of forgotten events is quite unequal across different superclasses: in superclass 10 (Functional Group Addition) and 5 (Protections) the number of never forgotten examples is higher in percentage to the class of Unrecognized (class 0) and Resolutions (class 11) reactions (see Figure 1.b). For few classes, such as the Unrecognized (class 0) and the Resolutions (class 11), the distribution of the number of forgotten events shows a less pronounced peak around zero and a significant population of entries that are learnt and forgotten with larger frequencies. These two last superclasses contain most of the wrong chemistry in the Pistachio data set, which is where the model struggles to learn. The large population of chemically wrong examples in these classes originates from the difficulties to consistently text-mine stereochemical information (for Resolutions) or from the difficulties to match a text-mined reaction with existing chemical templates (for Unrecognized).

The cohort of never learnt examples likely includes chemically wrong data as well as chemically correct reactions with features (i.e. reaction templates) rarely seen across the entire data set. The removal of a large section of such elements would cause the loss of important information and, as a consequence, a reduced performance of the model. While there is no possibility to tag each forgotten event as either rare or wrong chemistry, we need to apply precise strategies to minimize loss of rare but important information while removing noise. For this purpose, the set of training samples was first ordered based on the registered number of forgetting events. Starting from those entries which were never learnt and proceeding from the high-end tail of the distribution of forgotten events towards the never forgotten ones, we removed increasingly bigger portions of the data set up to a maximum of 40%. Each reduced set was used to train a new forward model. In Table 1 we report the new models top-1 and top-2 results on a common test set, in comparison with the baseline model (see section 4.6 for baseline and test set details).

Table 1: The top-1 and top-2 accuracies of the models trained with increasing percentages of data set entries removed by the forward forgetting experiment. For the model "forget01perc", 0.1% of the data set was removed, for "forget1perc" the removed percentage was 1%, and so on. All models are compared to the results of the baseline model (full data set) on the same test set.

model	**# of samples**	**top-1**	**top-2**
forget01perc	2 376 480	68.8	74.6
forget1perc	2 355 070	68.9	74.8
forget5perc	2 259 916	68.9	74.5
forget10perc	2 140 973	69.0	74.5
forget15perc	2 022 030	69.0	74.0
forget20perc	1 903 087	69.1	74.1
forget25perc	1 784 144	69.2	74.0
forget30perc	1 665 201	68.1	73.0
forget40perc	1 427 315	66.3	71.0
allmixed (baseline)	2 378 859	68.5	74.2

We notice that the top-1 accuracy is weakly affected by the removal of portions of the data set, until approximately 25% of the data set is removed: for 30% and 40% reductions, the performances start to decrease as a consequence of the loss of meaningful chemical knowledge.

We recently introduced the square root of the cumulative Jensen Shannon divergence (CJSD) to quantify bias in retrosynthetic disconnection strategies.[9] Here we revise its definition using a non-parametric approach, free of any kernel density estimation (see sections 4.4 and 4.5). The revised CJSD improves with the removal of portions of the data set (see SI, Figure 1), as a consequence of an increased similarity between the prediction confidence distributions (lower $\sqrt{\text{CJSD}}$). This is due to the fact that populations across classes in the training set become more balanced as many unrecognized reactions are removed, thus leading the trained models towards a similar confidence performance across all classes (see 4.2 for class population changes for few noise-reduced data sets). Moreover, removing noisy entries improves the confidence of the model in establishing the correctness of a prediction, resulting in all distributions peaking more towards a confidence of 1.0. Some examples of how the cumulative distributions change by removing portions of the data set can be found in section 4.5.

When increasing the number of removed entries, the CJSD monotonically decreases exhibiting a convergence towards unbiased confidence distributions across the different classes. Nonetheless, the model with the best CJSD (40% of the data set removed) was also the one with the lowest top-1 accuracy. The removal of 25% of the data set allowed the model to retain the high value of the top-n accuracy while improving the CJSD. Consistency checks were implemented for the entire removal strategy using different training random seeds to guarantee that the number of forgetting events experienced by each sample was not the result of a random evaluation (see section 4.7 for details).

While the removal of the forgotten events is increasing the significance of the underlying statistical distribution of the chemical reaction features, we still need to demonstrate that the entire process is effectively removing chemically wrong data. In order to asses this, we designed two different experiments where after introducing artificial noise into the data, we analyzed the forgetting frequency of those entries. In the first experiment, all the products of the reactions present in the validation set were shuffled and assigned randomly to a new group of precursors, similarly to Segler et al.[6]. This was considered the ”easy to identify” type of noise. As expected, the noise removal protocol led to the identification of 99% of the chemically wrong examples. This is shown bottom row of Figure 2.a, where the noisy set (7%) falls inside the 10% of the whole training data set never learnt. For the second experiment, the introduced noise was slightly more subtle. All original target products of the validation set, computed with the ”forget-25perc” model, were substituted by the third entry of the top-3 predictions, which is usually wrong. While being more challenging, the forgotten event strategy correctly identified 60% of the chemically wrong entries as never-learnt reactions, and 90% of the introduced noisy data experienced at least 1 forgotten event (Figure 2.a, bottom row). An example of the type of noisy reaction used is given in Figure 2.b.

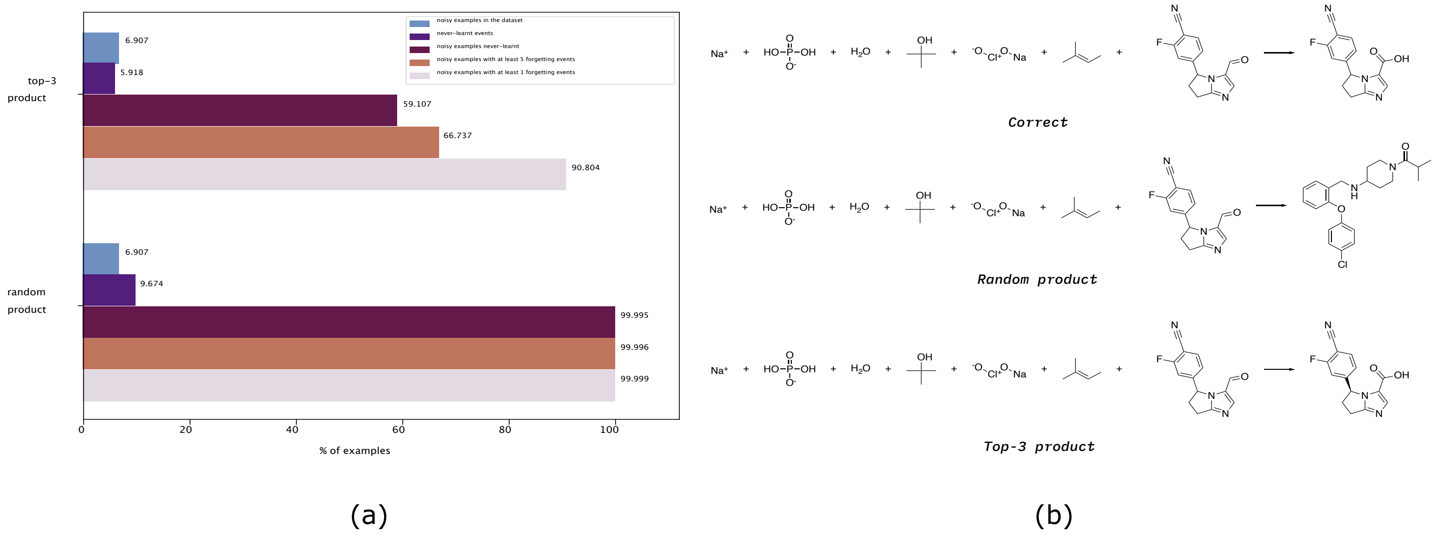


Figure 2: (a) Comparison of the two noisy strategies. In blue the percentage of noise introduced in the data set, in purple the percentage of events which were never learnt, then the percentage of noise found in the never learnt events, in those with down to 5 and down to 1 forgetting events respectively. ”random product” is the noise that is easy to detect: the examples were generated by shuffling the products of the reactions in the validation set.[6] ”top-3 product” was the experiment where the noise identification was more subtle: the target product in the validation set was substituted with the top-3 prediction on the same set. (b) example of the noise introduced in the same reaction SMILE for the two experiments. On top the correct reaction, in the middle the random product assigned and at the bottom the top-3 prediction (which shows a specific stereocenter that the reaction cannot generate).

A final comparison with the baseline of the newly trained forward model on the 25% cleaned data set was performed with the addition of some regularizing strategies. The complete plots are reported in Figure 2 of the Supporting Information.

The similarity metric (CJSD) for cumulative density functions is better for the group of models trained with the clean data set whereas the top-1 accuracy remains high in both baseline and new forward model. Top-1 is reported both for the original test set (where the predictions where hashed to identify tautomers and redox pairs) as well as for the test set removed of all one-precursors reactions (with one reagent/reactant only). These are mostly wrongly text-mined reactions and consequently should be considered not relevant for the current evaluation. All reported metrics show that the cleaned forward model introduces a consistent improvement compared to the baseline.

2.3 Retrosynthetic model noise-reduction

Recently,[9] we introduced a novel statistical retrosynthetic strategy with new ranking metrics based on the use of the forward prediction model, where the corresponding forward reaction prediction confidence plays a crucial role in determining which disconnections are the most effective among the entire set of single-step retrosynthetic predictions.

In this context, the use of the retrosynthetic model[9] in the noise reduction schema, does not lead to any effective noise reduction of the considered data set, as the single-step retrosynthetic model operates only as a prompter of possible disconnections, ranked subsequently by the forward prediction model. Therefore, the noise reduction strategy is effective only when used in combination with the forward prediction model. The corresponding noise-reduced data set is subsequently used to train the single-step retrosynthetic model. Figure 3 shows the performance of the new retro model compared to the baseline (trained with the non cleaned data set) relative to coverage, class diversity, round-trip accuracy[9] and square root of the cumulative Jensen-Shannon divergence, CJSD (see section 4.4 and 4.5 for details).

The coverage is high in both experiments, ensuring the existence of at least one valid

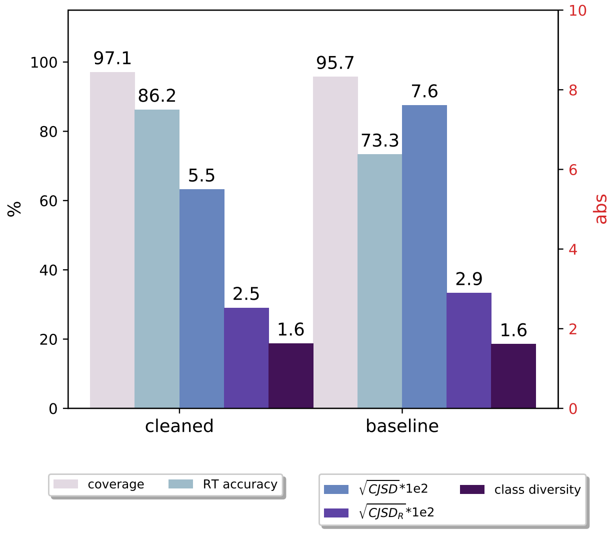


Figure 3: Results for the new retrosynthesis model, trained on the data set cleaned by the forgetting forward. Four metrics are shown: coverage, round-trip(RT) accuracy, $\sqrt{CJSD}$ and class diversity. $\sqrt{CJSD_R}$ excludes the Resolution class.

disconnection. Class diversity does no show any degradation, indicating that the noise-reduction strategy does not affect the diversity of the predictions. The round-trip accuracy on the other hand improved by almost 15%: removing noisy entries allows both forward and retrosynthetic model to learn better the difference between correct and wrong chemistry. In Figure 3 we report also the CJSD, the measure of the bias towards few classes over others. The CJSD decreases substantially, indicating a reduced bias in the confidence distributions across superclasses compared to the baseline.

2.4 Noise-reduced models assessment

We assessed the quality of the improved forward and single-step retrosynthetic models using the same set of chemical reactions used in Schwaller et al.[9]. In addition, we considered few additional retrosynthetic examples particularly challenging for the baseline model.[15]

The evaluation of the new forward model was characterized by the same performance of the original model[1] (see SI, Figure 3). The first set of compounds used to evaluate the performance of the noise-removed retro model, as well as the parameters used to run the retrosynthetic problems are reported in the SI (see SI, Figure 4). For literature references and for retrosynthesis results using the old model we refer to Schwaller et al.[9].

For compound **1** the new model assigns the highest confidence to a one-step retrosynthesis, trading the formation of the tetrazole ring for a commercial precursor carrying the same substructure. While this is an interesting strategy for operational reasons, it utilizes complex and more expensive starting materials. Nonetheless, the subsequent recommended path shows disconnection strategies similar to those of the literature and the old model. Only the conditions for the first retrosynthesis step changes from literature: recommending trimethylsilyl azide instead of sodium azide. Both approaches are chemically valid, with the most effective being decided by costs, risks, environmental impact, yield. For compound **2**, three different synthesis are reported in literature, where the shortest one exploits the opening of the epoxide ring. The improved model recommends an alternative sequence with respect to all those known, which differs from the first one by two aspects. First, two reaction steps are swapped: the attachment of the Grignard compound on the carbonyl group versus the N-alkylation. Moreover, the final ketone (alpha chloride) is not synthesized because found as commercial. Although this new route is using starting materials relatively more complex, the simplicity of the procedure may balance the increased cost of the precursors. Similar to the baseline model, among the other predicted pathways we find the optimal one reported in literature. Moving to compound **3**, unlike the old model the automatic retrosynthesis using the new model did not succeed in providing any retrosynthetic path. A deeper analysis shows that the first disconnection (Diels Alder cycloaddition) has a lower confidence (0.174) compared to the old model (0.362). The removal of forgotten events may have led to the a reduction of a particular class of functional groups, which affects the assessment of the forward reaction confidence in examples containing those specific functional groups. This may be compensated by repopulating chemical reactions data with examples containing functional groups more severely affected by the noise-reduction strategy. For compound **4**, the new model suggests an alternative strategy that avoids the problem of the conjugate reduction of the aldehyde faced by the baseline model. The new model improves the choice of chemoselective strategies while still showing few weakness on the stereoselectivity likely depending on the examples present in the noise-removed data sets. In fact, the choice of a less stabilized phosphonate for the Horner-Wadsworth-Emmons reaction could lead to reduced selectivity for the formation of the double bond 'E'. Similar to the radical bromination in allylic position, that may lead to selectivity issues on the primary carbon. Compound **5** shows identical retrosynthetic pathways both with the baseline and new model. Unlike for the old model, the retrosynthesis of compound **6** proposed by the new model completes with a narrower hyper- graph exploration finding a disconnection choice (alkylation of the ketone in the most substituted position) which, although not optimal in terms of regioselectivity, provides a viable path to the target. In both models, the first step is an ozonolisys, followed by the deprotection of the alcoholic group: these two steps should be inverted as the highly oxidizing conditions of ozonolysis could lead to a partial oxidation of the free hydroxyl group. The retrosynthesis proposed by the new model for compound **7** is made of a single step only. If we choose to exclude the most complex molecules within the ones proposed, we can still identify a complete route. This route does not provide a central disconnection by the opening of the ossiranic ring, as reported in literature, being characterized by peripheral disconnections which lead to a sequential type of synthesis. The route is in any case valid and makes a wise use of protections and deprotections steps to handle chemoselectivity. The commercial compound found at the last step is still quite complex , but unlike the old model the handling of the many chiral centers is satisfactory. For compound **8** the new model proposes an additional retrosynthetic strategy, which shows the enhancement and flexibility introduced by the noise reduced data. This strategy allows for a temporary increase of the complexity of the molecule through the introduction of the carbamic ring. The retrosynthesis for compound **9** performed with the new model terminates successfully and reveals itself to be quick and simple with few obvious improvement points: at step 5 we find a saponification of an ethyl ester to acid, which in the previous steps was transformed into a methyl ester, resulting in a useless interchange of ester groups.

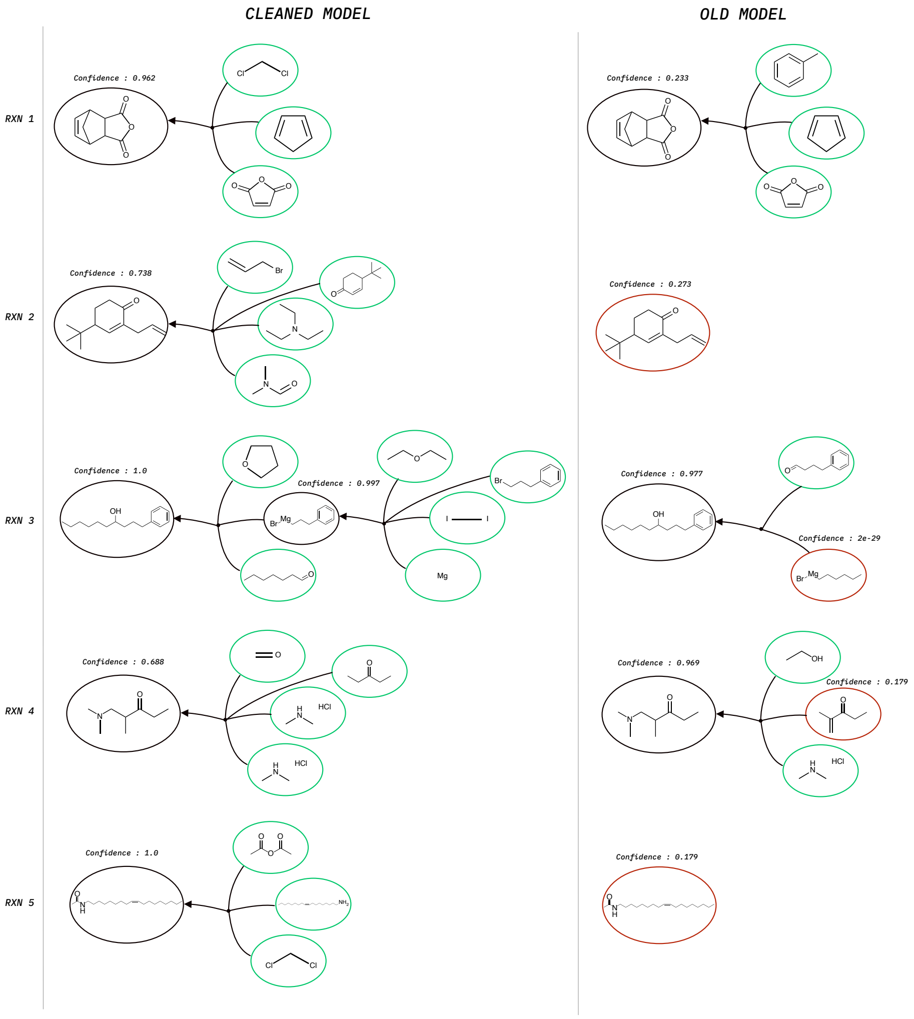


Figure 4: Examples of challenging retrosynthesis for the old model, like the formation of an enolate or the correct synthesis of a reactive organometallic species. On the middle column are reported retrosynthesis for the cleaned model. A tree structure is used to highlight the algorithm structure. On the right, are reported the results for the old model on the same compounds. The precursors highlighted in green are successfully found as commercial, while the red ones are not. Above each non-commercial molecule is also reported the confidence that the model assigns to the set of precursors predicted: if the confidence falls below a certain threshold the step is not carried on (this happens for two of the molecules for the old model). More details on the retrosynthetic algorithm can be found in Schwaller et al.[9]

In Figure 4, we show few additional examples, together with the results for the baseline and improved models. In **RXN 1** is shown the impact of the improved confidence distributions of the new model: the proposed Diels-Alder top-1 prediction has a much higher confidence for the cleaned model (0.962) than for the old one (0.233). **RXN 2** analyses a reaction which involves the formation of an enolate. The cleaned model proposed a correct retrosynthetic strategy, where the old model failed, due to the low level of confidence in all the proposed set of precursors. Another interesting case regards a multistep synthesis which involves the preparation of reactive organometallic species, in particular non commercial Grignard (**RXN 3**): the new model is clearly more confident about the predictions and this allows to complete the retrosynthesis of the compound. The same does not happen in the old model, which does not provide any suggestion after the synthesis of the metal-carbon bond. For what concerns **RXN 4**, the new model successfully completes the retrosynthesis for a Mannich reaction, proposing the correct set of reagents with a reasonable value of confidence. The old model suggested instead a first addition step with a secondary ammine (α,β-unsaturated ketone), which may not be the best choice even if chemically valid. In addition, for the latter model the algorithm stopped at the second step (the preparation of the α,β-unsaturated ketone) due to low confidence values of the proposed disconnections. We also noted that the old model assigned quite low values of confidence (0.179) to reactions like **RXN 5**, where a primary ammine and acetyl anhydride are used, whereas the cleaned model synthesizes successfully the compound with high confidence.

3 Conclusions

In this work we present the first unassisted, machine-learning based technique to automatically remove wrong and noisy chemical reactions from data sets. This methodology provides a statistical alternative to the more tedious human or rule-based curation schemes. We applied the noise reduction strategy to USPTO-based data sets and we used the cleaned data set to retrain both forward, retrosynthetic and classification models. Statistics on the test set revealed an increase in the values of all the significant metrics (round-trip accuracy + 13%, coverage + 1.4%, CJSD down to 60% of its original value for forward model and to 70% for the retro model). We assessed the new model by reviewing several retrosynthetic problems to highlight the improved performances. All the compounds used in our previous works[1,9] were tested: overall the new model is able to either reproduce the results or propose valid alternatives. In most cases the confidence of the new model on the retrosynthetic steps is increased, as highlighted in the last reactions comparison (Figure 4). The results show that it is not the quantity but the quality of knowledge embedded in training data sets which leads to more reliable models. We hope the development of noise-reduction protocol for improving the quality of existing data sets will have an enormous impact on the application of data-driven trained models.

4 Methods

4.1 Transformer model

As in previous works[1,9] we used machine translation architectures to map the chemical syntheses onto the world of machine learning models. The reactants, reagents and products were codified as Simplified molecular-input line-entry system (SMILES)[16,17] strings, tokenized, and fed to the Molecular Transformer,[1] the architecture based on the well known sequence-2-sequence transformer by Vaswani et al.[14]. The hyperparameters of the model were kept fixed throughout all simulations. The transformer is made of a set of encoder layers and a set of decoder layers. The tokens of the input SMILES string are encoded into (learnt) hidden vectors by the encoder. Those vectors are then fed to the decoder to predict the output sequence, one token at a time. Based on the work of Schwaller et al.[9], we decided to set the number of layers in both encoder and decoder to 4 (size 384). The transformer main characteristic is the presence of multi-head attention and the number of these heads was set to 8. Dropout was also included in the model at a rate of 0.1. Adam optimizer was used for loss minimization and the starting learning rate was set to 2. We used the OpenNMT framework[18] and PyTorch[19] to build the models.

4.2 Data

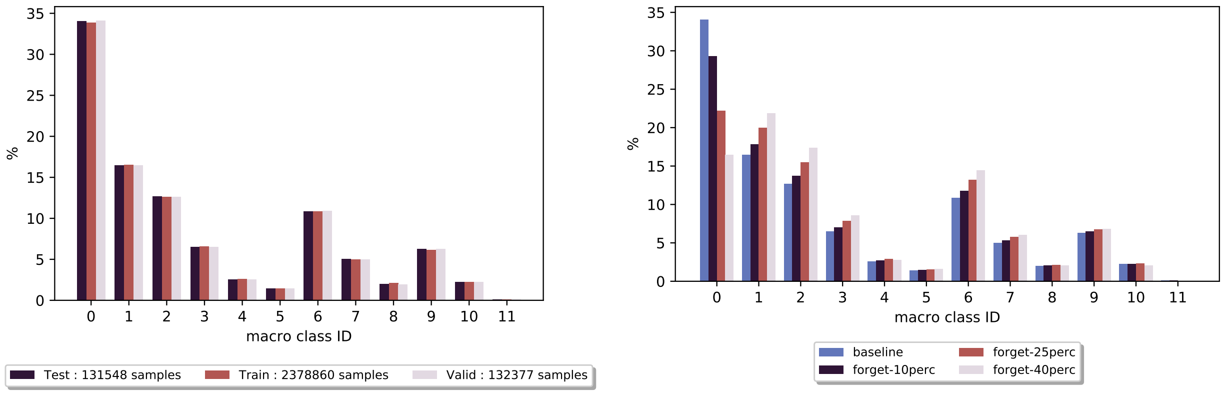
The reaction data set used to experiment on the forgetting cleaning strategy was the non-public Pistachio (release of Nov. 18th 2019), derived from text-mining chemical reactions in US patents. Molecules and reactions are represented as SMILES strings. A first coarse filtering strategy was applied to the raw data set which counted approximately 9.3 million reactions. First, all duplicate reactions were removed. Equal molecules in the same precursor set were made unique and the set was then alphabetically sorted and any duplicate removed. For consistency with the sequential design of retrosynthetic routes, multi product reactions were eliminated and for single product reactions only the largest fragment (as defined by RDKit[20]) was kept. Purification reactions were entirely removed.

The pre-filtered data set was then randomly splitted into training, test and validation set. This brought the data set down to 2‘378‘860 entries for the training set, 132‘376 for the validation set and 131‘547 for the test set (90%-5%-5% splitting). To enable a more exhaustive model evaluation the splitting was performed on unique products selections, to avoid similar examples being present in the training and test or validation. In Figure 5a the percentage of examples present in the three different splits, divided into the 12 macro classes. The same is reported for three different train data sets cleaned by the forward forgetting in Figure 5b to show how the macroclasses get re-balanced. Note that the test and validation sets were not subject to cleaning.

As in previous work,[1] we choose not to make a distinction between reactants and reagents because, even chemically, this splitting is not always well defined and can change with the tool used to make the separation. Each char in the SMILES string was codified as a single token and a new token was introduced, the '$\sim$' in order to model fragment bonds.

4.3 Definition of forgotten events

In the context of chemical reaction and retrosynthesis, the definition of forgotten event is of crucial importance. For forward prediction, given a data set of chemical examples



(a)

(b)

Figure 5: (a) How the different superclasses are populated in the three splits: train, validation and test set. (b) The balancing of the classes with the forgetting experiment: the baseline model is reported together with 3 of the cleaned models.

$A = (x_i, y_i)$, we denote as y_i the predicted product given reactants/reagents x_i. The acc_i^t is a binary variable denoting if example i is correctly predicted at time step t. We identify each time step with an epoch. A $forward$ forgetting event occurs when a product molecule, previously classified as correct at $t-1$, is then wrongly predicted at time step t. On the other side, a learning event occurs when at the next epoch the previously wrong prediction becomes correct:

$$\begin{aligned} \text{forgetting event} \qquad & acc_i^t < acc_i^{t-1} \\ \text{learning event} \qquad & acc_i^t > acc_i^{t-1} \end{aligned}$$

It can also happen that some events are *never learnt* and others are *never forgotten*. The former are labeled as having an infinite number of forgetting events, the latter as having none. While for the forward prediction, the definition of what is ”forgotten” is in line with the one in the cited reference,[12] some modifications need to be made for the retrosynthesis definition.

In fact, for retrosynthetic predictions, the top-N accuracy of the single-step retrosynthetic model is only a measure of how efficient the model is in memorising data rather than extracting contextual knowledge. The goal is not to propose the most commonly reported sets of reactants and reagents that result in a certain target molecule, but to generate many chemically correct sets, with ideally high diversity. A more appropriate metric would be the round-trip accuracy, obtained by applying back a forward prediction to the predicted precursors set and comparing the result with the original product molecule. In this sense, a learnt retro-event occurs when we can recover the product molecule. Following the formalism used for the forward model, we define a data set of chemical examples $D = (x_i, y_i)$ where x_i is a molecule to be synthesized, y_i the target set of precursors. Again $\overline{y}_i$ denotes the model prediction (this time the precursors set). A *retro* forgetting event occurs when a set of precursors, which led back to the product molecule at $t-1$, fails in recovering this at time step t:

$$\begin{aligned} \text{forgetting event} \qquad & RTacc_i^t < RTacc_i^{t-1} \\ \text{learning event} \qquad & RTacc_i^t > RTacc_i^{t-1} \end{aligned}$$

$RTacc_i^t$ denotes the round-trip accuracy in binary form of example i at time step t.

4.4 The metrics for performance evaluation

In order to evaluate a single-step forward and retro synthetic prediction, appropriate metrics need to be designed which do not rely on tedious, manual analysis by a human chemist. Here, we use: top-n accuracy, round-trip accuracy, class diversity and coverage as reported in a previous work.[9] In Schwaller et al.,[9] authors introduced the Jensen-Shannon divergence, while here we turn to its cumulative version explained in detail in the next section.

4.5 Cumulative Jensen-Shannon divergence

Jensen-Shannon divergence is a measure of how similar two or more discrete probability distributions are.

$$JSD(P_0, P_1, ...P_N) = H\left(\sum_{i=0}^{N} \frac{1}{N} P_i\right) - \frac{1}{12} \sum_{i=0}^{N} H\left(P_i\right) \tag{1}$$

Where P_i denotes a probability distribution and $H(P_i)$ the Shannon Entropy for the distribution P_i. However, when we deal with continuous probability density functions the generalization of this formula is not straightforward. The problem lies in the definition of the entropy. The first expression proposed by Shannon[21] to deal with the continuous case was the one of differential entropy .

$$H\left(P\right) = -\int p(x) \log p(x) dx \tag{2}$$

Where the capital letter P denotes the probability distribution and p the probability density. This definition raises many concerns as higlighted in detail by Murali Rao et al.[22], the most important is the fact that it is "inconsistent", in the sense that a probability density function, differently from a probability function, can take values which are greater than 1. As a consequence, entropy for values less than 1 will contribute positively to the entropy, while those greater than one with a negative sign and the points equal to 1 will not contribute at all. This brings to an ambiguous definition of the information content carried by the differential entropy expression. A way to overcome this issue is to define a new version of entropy which defines both continuous and discrete distributions. This quantity is known as Generalized Cumulative Residual Entropy[22].[23]

$$GCRE(P) = -\int_{-\infty}^{\infty} F(X > x) \log F(X > x) dx \tag{3}$$

$$F(X > x) = 1 - P(X < x) \tag{4}$$

The key idea is that now $P(X < x)$ denotes the cumulative distribution and no more the probability density function. The advantages in the use of this definition is duplex. First, there is no more inconsistency: the principle that the logarithm over a distribution is a measure of its information content is preserved. Second, we don't have to rely on probability density functions, which need to be estimated from the data with parametric methods like kernel density. This way, we can easily calculate directly from the data the cumulative distribution function, which is the "real" one representing the data (giving that we have enough observations). If not enough observations are provided, the error in the calculation of the entropy will only related to missing information and will not be dependent on the parameter used to estimate a probability density function. As a consequence, a **cumulative Jensen- Shannon divergence**[23] can be defined to compare cumulative distributions:

$$CJSD(P_0, P_1, ...P_N) = GCRE\left(\sum_{i=0}^{N} \frac{1}{N} P_i\right) - \frac{1}{12} \sum_{i=0}^{N} GCRE\left(P_i\right) \tag{5}$$

With this new metric we were able to compare information content of likelihood distributions extracted from different reaction macro classes. As extensively explained in a previous work,[9] having a model with dissimilar likelihood distributions is equivalent to bias towards specific reaction macro classes over others. The model with the lowest CJSD value will be the one which has the most uniform likelihoods distributions. In Figure 6 can be found the cumulative distribution functions for the baseline and the cleaned forward model divided by the 12 macro classes. These are the distributions of correctly predicted samples from the test set used in the evaluation of the CJSD.

Note that the cumulative distributions for the retrosynthesis were constructed from the forward likelihoods obtained by applying back the forward model to the top-10 set of precursors.

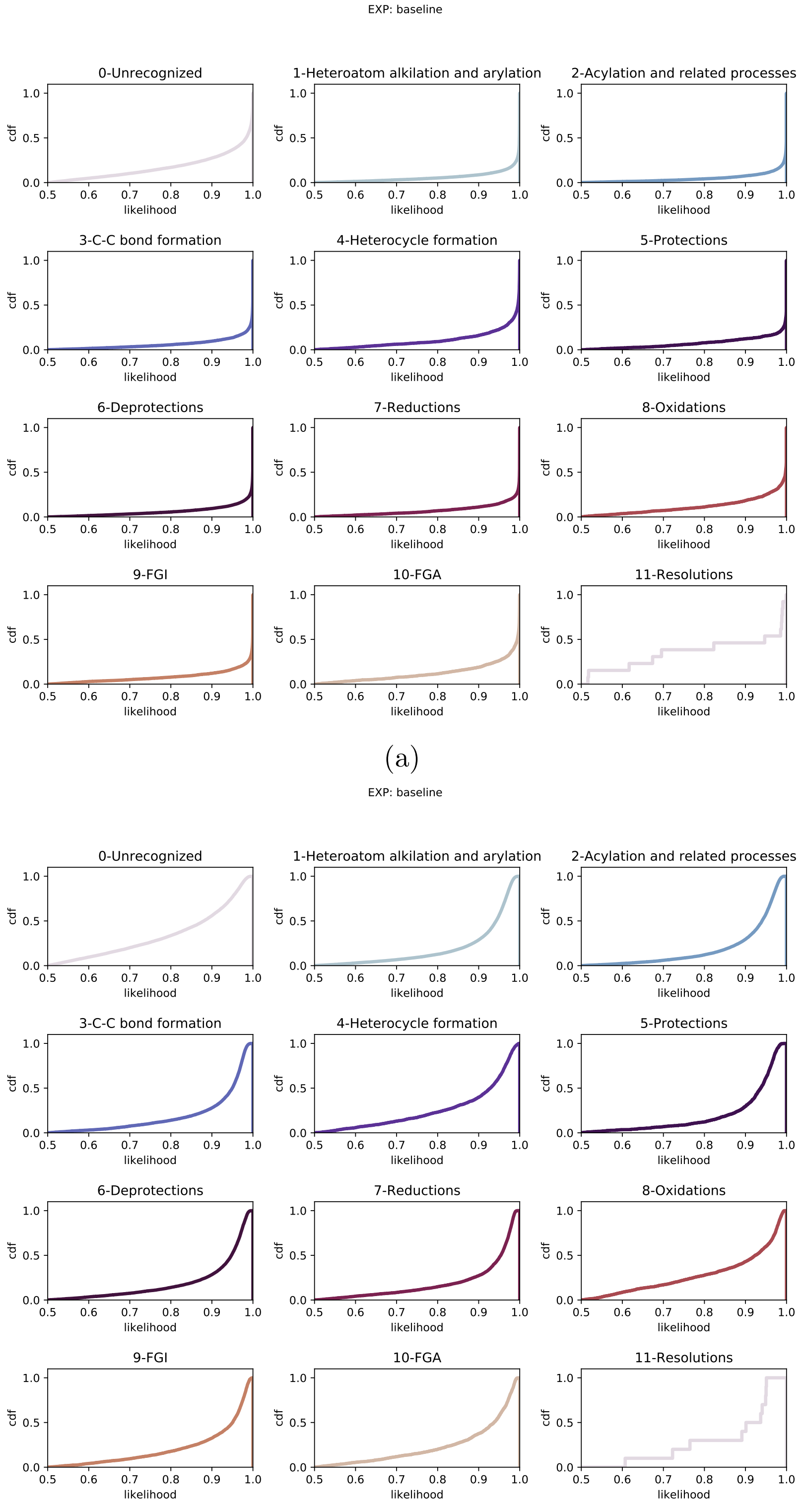


Figure 6: Cumulative distribution functions divided by macro classes of the correctly predicted samples from the test set (forward model). (a) cleaned model: the distributions are more peaked towards 1.0 because removing noisy samples allows the model to be more sure of the predictions. (b) baseline model.

4.6 The baseline model

For the baseline model of both forward and retro prediction, we trained the Molecular Transformer[1] directly on the 2.4 million reactions extracted from the coarse filtering of Pistachio. The forward model used the tokenized smiles of the precursors (reactants and reagents with no distinction), while having the SMILE string of the product molecule as a target. For the retro model the two were swapped.

First of all, we tried to establish a metric of comparison with our "old model", which is the one currently running on the IBM-RXN open source online platform for chemical synthesis prediction.[1,15] These old models (both forward and retro) were tested both on the old and new test set. In Figure 2 the results can be compared to the performances of the new baseline models.

Table 2: Top-n accuracy of old models tested on the old and new test sets. For the forward models the top-1 and the $\sqrt{\text{CJSD}}$ is reported. $\sqrt{\text{CJSD}_R}$ is the cumulative JSD without the Resolutions. For retro models, top-10 is reported along with the four metrics presented in the previous section 4.4: round-trip accuracy (RT), class diversity (CD), coverage (COV), and$\sqrt{\text{CJSD}}$. The latter is computed on the forward likelihood distributions generated by the cleaned forward model on the precursors set.

models	**top-1**	**top-10**	**RT**	$\sqrt{\textbf{CJSD}}$	$\sqrt{\textbf{CJSD}_R}$	**CD**	**COV**
fwd-old: old ts	71.86	-	-	0.066	0.048	-	-
fwd-old: new ts	68.41	-	-	0.123	0.047	-	-
fwd-new: new ts	69.32	-	-	0.095	0.052	-	-
retro-old: old ts	-	21.11	70.08	0.106	0.027	1.5	89.77
retro-old: new ts	-	20.33	71.26	0.106	0.028	1.6	94.68
retro-new: new ts	-	24.25	73.27	0.076	0.029	1.6	95.71

This check on the performances was done in order to make sure that no significant difference arose from using the new release of Pistachio. We notice that indeed top-1 for forward clusters around 70%. If the performance of the old model on the new test set was to be clearly better than the one of the new baseline model, no conclusions on the comparison of the two could have been drawn as the strategy used to extract the two sets was different: an increased performance could have been a sign of a strong presence of examples from the new test set in the training set of the old model. Note that, in order to have a more appropriate comparison of the two models, the new baseline forward model was augmented with randomized reactions of the training set (in equal number to the training examples) as well as reaction extracted from textbooks, the same technique reported in a previous work of Schwaller et al.[1] For the retro model, we reported top-10 accuracy, but more importantly the round-trip accuracy, which is already slightly better for the new baseline model on the new test set compared to *retro-old* on the same set. The coverage follows the same trend, while the class diversity is not affected. It is important to point out that the forward model used for the "back-evaluation" of the retro, was the same for all three models (*retro-old: old ts*, *retro-old: new ts*, *retro-new: new ts*) and is identified with the one called "forget-25perc", the cleaned forward model presented in section 2.2.

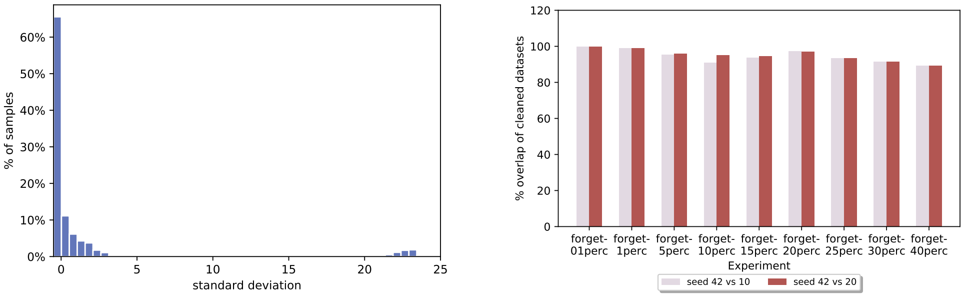
4.7 Statistical analysis of forgotten examples

Concerning the forward forgetting experiment, some statistical evaluations were performed in order to make sure that the computation of how many times an example was forgotten/learnt was not completely random. First, two models were trained with a different seed (10, 20) to the original one used (42) and again the number of forgotten events was computed. These counts were then analyzed across seeds. The table in Figure 7 reports the Pearson coefficient computed for couples of seeds. By being close to one it indicates a high correlation between the numbers. Also the standard deviation though all three seeds behaves well (Figure 7a), with most of the examples concentrating around 0 and some outliers between 20 and 25. The position of the outliers depends on the number chosen (n = 50) to substitute the "∞" denoting the events never learnt (this substitution was necessary to compute the standard deviation). Also the overlap of the removed sets, in all of their percentages shows its stability in oscillating around 100% (Figure 7b).

4.8 Retrosynthetic forgetting strategy: a random failure

To prove the random behaviour when cleaning data set using the single-step retrosynthetic model, we performed a simple statistical experiment. We used the forgetting forward strategy

Seed	**Pearson**
42vs10	0.8812
42vs20	0.8810
10vs20	0.8831

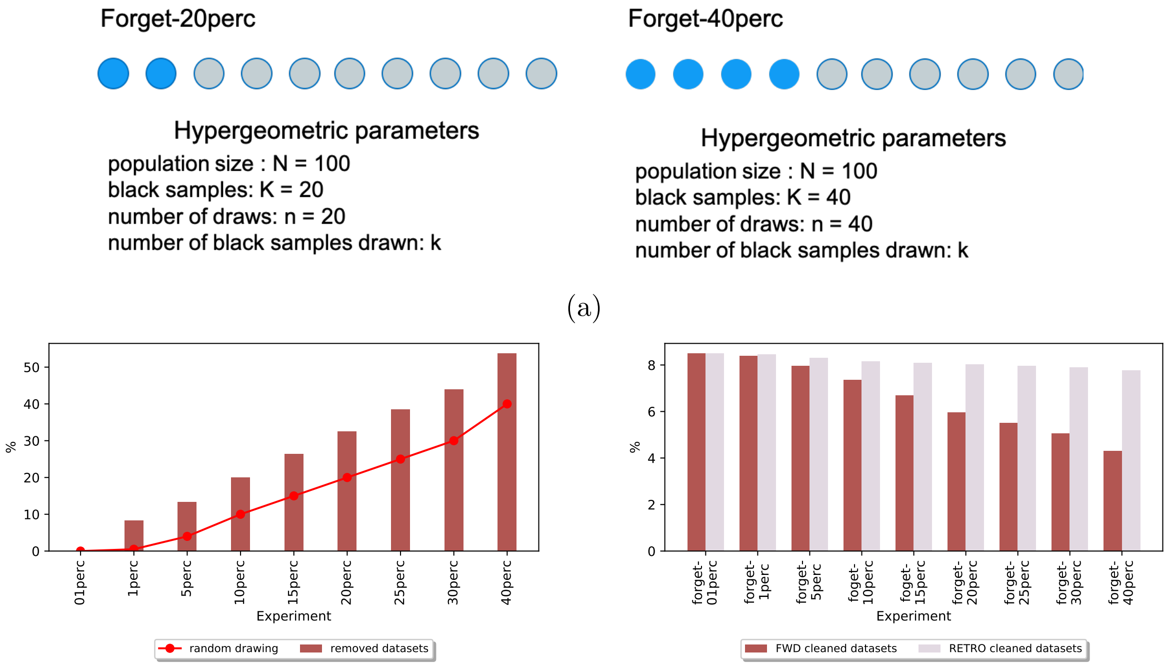


(a) (b)

Figure 7: On the left the Pearson coefficient calculated for couples os seeds. (a) Standard deviation of the number of forgotten events calculated for each sample across the three seeds. The concentration of outliers around 20/25 is due to the fact that never learnt events, nominally ∞ forgotten, were defined as forgotten 50 times in order to compute a stdv. Indeed, a minor percentage of examples in some of the seeds is never learnt and in others is learnt with a high forgetting rate. (b) Overlap between the removed data sets in different percentages across seeds. The overlap oscillates around 100%, indicating the stability of the methods across seeds.

to label all the examples tagged for removal (20% for the "forget-20perc", 25% for the "forget-25perc", etc.) without actually removing them from the original data set. Now that we have a label for the removed examples, if we were to draw some of them (random retro forgetting) from the data set with no replacement we could describe the probability of drawing a "blue" sample with an hypergeometric distribution (8a).

The probability of drawing all the labeled samples would be 0 for all cases and maximal for the percentages which correspond to the red points in Figure 8b. In Figure 8b we have the percentage of samples removed by the retro model which match those removed by the forward model. We see that the behaviour is almost random and maps the red points. Otherwise, we would have expected an overlap close to 100% for all the removed data sets. One more proof of the random behaviour of the forgetting retro can be seen in Figure 8c. The one-precursor reactions, wrong chemical reactions because of missing reagents, are detected and removed by the forgetting forward , but not by the forgetting retro (their percentage remains constant in the data set cleaned by the latter).



(a)

(b)

(c)

Figure 8: (a) statistical experiment: detecting whether the selection of most forgotten examples by the retro model is random. (b) Overlap between the data sets removed by forward and retro models. The red line models how the overlap should be if the retro selection were entirely random. (c) Percentage of one-precursor reactions in the data set cleaned by the forward forgetting and by the retro forgetting. The former is able to identify them, while the latter is not.

Acknowledgement

We thank Matteo Manica for the help with the model deployment. We also acknowledge all the IBM-RXN team for all the insightful suggestions.

References

(1) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. *ACS Central Science* **2019**, *5*, 1572–1583.

(2) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge, 2012.

(3) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017; https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

(4) Nextmove Software Pistachio. http://www.nextmovesoftware.com/pistachio.html, (Accessed Apr 02, 2020).

(5) Reaxys. https://www.reaxys.com, (Accessed Apr 02, 2020).

(6) Segler, M.; Preuss, M.; Waller, M. *Nature* **2018**, *555*, 604–610.

(7) Coley, C. W. et al. *Science* **2019**, *365*.

(8) Schwaller, P.; Laino, T. Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches. *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*. 2019; pp 61–79.

(9) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. *Chemical Science* **2020**, *11*, 3316–3325.

(10) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. *Drug Discovery Today* **2020**,

(11) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. *Chem. Sci.* **2020**, *11*, 154–168.

(12) Toneva, M.; Sordoni, A.; Tachet des Combes, R.; Trischler, A.; Bengio, Y.; Gordon, G. J. *arXiv e-prints* **2018**, arXiv:1812.05159.

(13) Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. `https://doi.org/10.1016/S0079-7421(08)60536-8`, (Accessed Apr 02, 2020).

(14) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. 2017; pp 5998–6008.

(15) IBM RXN for Chemistry. `https://rxn.res.ibm.com`, (Accessed Jan 17, 2020).

(16) Weininger, D. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(17) Weininger, D.; Weininger, A.; Weininger, J. L. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.

(18) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proc. ACL. 2017.

(19) Paszke, A. et al. *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp 8024–8035.

(20) Landrum, G. et al. rdkit/rdkit: 2019_03_4 (Q1 2019) Release. 2019; `https://doi.org/10.5281/zenodo.3366468`.

(21) Shannon, C. E. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

(22) Murali Rao,; Chen, Y.; Vemuri, B. C.; Fei Wang, *IEEE Transactions on Information Theory* **2004**, *50*, 1220–1228.

(23) Nguyen HV., V. J. *Appice A., Rodrigues P., Santos Costa V., Gama J., Jorge A., Soares C. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science. Springer, Cham* **2015**, *9285*.