JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Chemical Information

Enhancing Retrosynthetic Reaction Prediction with Deep Learning using Multiscale Reaction Classification

Javier L. Baylon, Nicholas A. Cilfone, Jeffrey R. Gulcher, and Thomas W. Chittenden

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.8b00801 • Publication Date (Web): 14 Jan 2019

Downloaded from http://pubs.acs.org on January 15, 2019

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Enhancing Retrosynthetic Reaction Prediction with Deep Learning using Multiscale Reaction Classification

Javier L. Baylon,[†,‡] Nicholas A. Cilfone,[†,‡] Jeffrey R. Gulcher,[†,¶] and Thomas W. Chittenden[*,§,‡,‖]

†*Computational Statistics and Bioinformatics Group, Advanced Artificial Intelligence Research Laboratory, WuXi NextCODE, Cambridge, MA 02142, USA*

‡*Complex Biological Systems Alliance, Medford, MA 02155, USA*

¶*Cancer Genetics Group, WuXi NextCODE, Cambridge, MA 02142, USA*

§*Computational Statistics and Bioinformatics Group, Advanced Artificial Intelligence Research Laboratory, WuXi NextCODE, Cambridge, MA 02142, US*

‖*Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA 02215, USA*

E-mail: tchittenden@wuxinextcode.com

Abstract

Chemical synthesis planning is a key aspect in many fields of chemistry, especially drug discovery. Recent implementations of machine learning and artificial intelligence techniques for retrosynthetic analysis have shown great potential to improve computational methods for synthesis planning. Herein, we present a multiscale, data-driven approach for retrosynthetic analysis with deep highway networks (DHN). We automatically extracted reaction rules (i.e., ways in which a molecule is produced) from a dataset consisting of chemical reactions derived from U.S. patents. We performed the retrosynthetic reaction prediction task in two steps: first, we built a DHN model to predict which group of reactions (consisting of chemically similar reaction rules) was employed to produce a molecule. Once a reaction group was identified, a DHN trained on the subset of reactions within the identified reaction group, was employed to predict the transformation rule used to produce a molecule. To validate our approach, we predicted the first retrosynthetic reaction step for 40 approved drugs using our multiscale model and compared its predictive performance with a conventional model trained on all machine-extracted reaction rules employed as a control. Our multiscale approach showed a success rate of 82.9% at generating valid reactants from retrosynthetic reaction predictions. Comparatively, the control model trained on all machine-extracted reaction rules yielded a success rate of 58.5% on the validation set of 40 pharmaceutical molecules, indicating a significant statistical improvement with our approach to match known first synthetic reaction of the tested drugs in this study. While our multiscale approach was unable to outperform state-of-the-art rule-based systems curated by expert chemists, multiscale classification represents a marked enhancement in retrosynthetic analysis and can be easily adapted for use in a range of artificial intelligence strategies.

Introduction

Planning chemical synthesis of molecules is a crucial aspect of organic chemistry that has applications spanning many different fields, including drug discovery and material science. The goal of chemical synthesis planning is to derive a pathway, often consisting of multiple reaction steps and reactants, by which a target molecule can be produced.[1] Retrosynthetic analysis is a widely employed technique in chemical synthesis planning, where a target molecule is recursively transformed into simpler precursors, following a reversed chemical reaction.[1–3] This process is carried out until a series of starting precursors (e.g., commercially available molecules) are obtained. An overall synthetic route for the target molecule can subsequently be produced by combining all of the derived reactions from the retrosynthetic analysis. Over the last few decades, numerous approaches have been developed for retrosynthetic analysis that take advantage of novel and emerging computational techniques.[4–6]

Typically, computer-aided retrosynthetic analysis is carried out using reaction rules that consist of a set of minimal transformations (e.g., changes at the reactive center and neighboring bonds and atoms) to characterize a chemical reaction.[7] These reaction rules, which can be encoded by expert chemists[6,8] or automatically extracted from a given dataset,[7,9–17] are used as templates for chemical transformations applied to an input target molecule to derive retrosynthetic precursors. The result of this rule-based approach is a set of reactant molecules that transform into the target product by following the reaction rule. Currently, rule-based systems curated by expert chemists (such as Chematica) outperform entirely data-driven approaches for retrosynthetic analysis in planning the efficient synthesis of target molecules.[18]. However, rule-based systems are inherently limited to the initial set of reaction rules regardless as to whether they are hand coded by expert chemists or automatically extracted from the data. Thus, rule-based systems often struggle to predict retrosynthetic reactions for new target products that are beyond the scope of the transformations rules. To overcome this limitation, several new approaches based on deep learning (DL) have been recently developed for retrosynthetic analysis.[14,17,19] The key advantage of these non-linear statistical-learning approaches over traditional rule-based retrosynthetic methods is that they are able to extract generalizable patterns from large amounts chemical data, such as the molecular context in which a reaction occurs, at a fraction of the computational cost required for traditional rule-based implementations.[14,17,19]

A variety of machine learning and DL techniques have been applied for forward reaction prediction to anticipate the outcome of chemical reactions.[14,15,20–24] These approaches typically employ a variety of statistical learning methods to carry out the prediction task, including automatic identification of the reaction center[22] and classification of the type of reaction.[21] Current DL-based retrosynthetic analysis implementations, including this work, use similar approaches. One important difference in current DL retrosynthetic analysis is the way in which molecules are represented. For example, Liu et al.[19] formulated retrosynthetic reaction prediction as a translation task using a seq2seq architecture, by representing molecules as SMILES[25] strings. In this approach, a target product was encoded as a string of characters extracted from its corresponding SMILES string and converted (e.g., translated) to another sequence of characters, corresponding to reactant SMILES strings. One notable advantage of this approach is it naturally incorporates information about the global environment of molecules is by considering their entire structure (encoded in its SMILES string), instead of an abstracted version of the reaction (i.e., only the reactive center). However, due to the nature of the translation task, the model is prone to predict chemically unfeasible precursors for a target molecule.[19]

In another approach, Segler and Waller formulated the retrosynthetic reaction prediction task as a classification problem,[14,17] by representing molecules as Morgan fingerprints.[26] In this approach, reaction rules (as defined in Fig. 1) are automatically extracted from a large dataset (i.e., the Reaxys chemical database consisting of several millions of reactions[27]) and used as labels to train a multilabel classifier based on deep highway networks.[28] Given an input target molecule encoded as a fingerprint, the model predicts the probability of all the possible reaction rules in the training set. The top predicted rules are then applied to the input target molecule to obtain a set of retrosynthetic precursors. This approach has recently been combined with Monte Carlo tree search to guide the reaction prediction task and derive robust retrosynthetic pathways.[17]

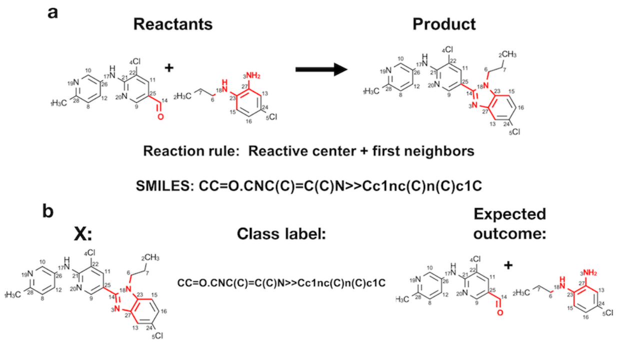
a
Reactants
Product
Reaction rule: Reactive center + first neighbors
SMILES: CC=O.CNC(C)=C(C)N>>Cc1nc(C)n(C)c1C
b
X:
Class label:
CC=O.CNC(C)=C(C)N>>Cc1nc(C)n(C)c1C
Expected outcome:

Figure 1: Schematic representation of reaction rules and how they are used in this work. (a) Reaction rules encode the smallest transformation required to convert reactants into products. To extract these rules, mapped atoms from the reactant and product side were compared to identify the reactive center, which comprises bonds and atoms changed during the reaction (highlighted in red in the reaction).. Starting from the reactive center, atoms and bonds close to the reactive atoms (first neighbors) were identified, and were used to assign the reaction rule (shown as a SMILES string). Note that the depicted reaction was not included in our dataset, but it was obtained from Christ et al.[12] (b) To build multiclass classifiers for retrosynthesis, each product molecule was assigned its corresponding reaction rule as a class label, and dataset was rearranged so it contains product molecules and reaction rules. Classifier models are trained/tested using product molecules as input data, and the expected outcome is a list of reactants obtained by transforming the product molecule based on a predicted class label (reaction rule).

Despite all of the increasingly sophisticated molecular representations developed for statistical machine learning tasks (e.g. latent space representations[29,30] and molecular graph convolutions[31–33]), traditional binary fingerprints are still ubiquitously employed in many cheminformatics approaches, particularly in drug discovery.[34,35] Although fingerprints present several issues (including bit collisions and vector sparsity), their flexibility and ease of computation offer a useful featurization scheme for machine learning applications, especially for reaction prediction as recently exemplified in work by Segler and Waller.[14,17] Thus, DL approaches that take advantage of molecular fingerprints are still highly relevant and valuable to increase the use of DL approaches by expert chemists.

One important aspect of Segler and Waller's work was the size of the initial dataset. It contained as many as 12.4 million reactions, which essentially encompassed all of the knowledge of organic chemistry.[17] As previously stated, an intrinsic limitation of an entirely data-driven retrosynthetic analysis approach is the finite number of reaction rules that a machine can automatically extract from a dataset. These numbers are determined by several factors, including the level of detail of the extracted reaction rule and the diversity of the chemical dataset. This limitation can be aggravated in smaller datasets, found for example in a private electronic lab notebook (ELN), which would necessarily contain less knowledge than the entirety of known reactions in organic chemistry,[17] or even in the publicly available United States patent (USPTO) dataset.[36] Thus, data-driven machine learning rule-based models trained on different datasets would learn different amounts of chemistry, determined by the number of machine-extracted reaction rules, and thus subsequently limited in its predictive scope. Moreover, because of the chemical diversity expected in a typical ELN, a DL model for retrosynthetic reaction prediction can be biased to learn a highly imbalanced type of reaction (one that is highly represented compared to other reactions), if the data is not properly stratified.

To address some of these issues, we present a multiscale approach for retrosynthetic reaction prediction, using machine-extracted reaction rules from USPTO dataset, which contains patented reactions from the last 50 years.[36] Herein, we show that this multiscale statistical learning approach leads to significant increases in classification performance (measured by balanced accuracy) compared to a conventional model trained on machine-extracted rules employed as a control. By stratifying the data by reaction groups during the training phase (the reaction group scale), the model learns patterns over molecules that were obtained by similar chemical reactions (the reaction rule scale), which might be overlooked in a typical machine-extracted rule-based system. We validated our multiscale approach by predicting the first retrosynthetic step for 40 approved small molecules. For these drugs, the multiscale model correctly predicted more known retrosynthetic steps than the conventional model trained on machine-extracted rules. We also show the limitations of our DL approach for retrosynthetic reaction prediction, by using state-of-the-art predictions generated with Chematica for pharmaceutical molecules[18] as ground truth, which our approach was unable to reproduce.

Results and Discussion

Automatic Extraction of Reaction Rules from the U.S. Patent Dataset

To begin, we preprocessed a dataset of chemical reactions aggregated from patents granted in the U.S. between 1976 and 2016, referred to as the USPTO dataset.[36] First, we extracted reactive centers for individual reactions (i.e., the set of atoms and bonds that changed between reactants and products in a given reaction), following an established protocol in order to define reaction rules for multiclass classification.[7,12] Briefly, we algorithmically compared atom properties between mapped atoms in reactants and products, and identified the atoms and bonds that changed in the reaction (e.g., the reactive center). We defined reaction rules (RR) which contained the reactive center and the shell of first-neighboring atoms (Figure 1a). Reaction rules are templates that describe the smallest transformation required to convert reactants into products.[7] By including the first neighboring atoms to the reactive center, we aimed to capture some of the relevant chemical environment in which the transformation associated with the reaction occurs. This approach to derive reaction rules has been employed before to characterize chemical reactions for retrosynthesis.[7,12] In this work, we employed reaction rules as class labels for product molecules to build multiclass classifiers for retrosynthesis (as summarized in Figure 1b.)

Once reaction rules were extracted, the USPTO dataset was rearranged so that each product molecule was matched to the reaction rule associated with its corresponding reaction (Figure 1b). Each reaction rule was employed as a class label for product molecules in the reactions. By arranging the dataset in this manner, a multiclass classifier can be trained on product molecules to predict reaction rules, which in this case correspond to different class labels. Once a reaction rule is predicted, the transformation was applied to the input molecule to generate a list of reactants (Figure 1b).

Our initial rule extraction step resulted in a total of 74,482 unique RRs, however a significant portion of the extracted rules were represented only once (e.g., 54,444 RRs). These reaction rules were discarded since they are associated with very specific reactions, which were not generalizable across the dataset. For the extracted rule set, we defined different cutoffs for the number of times a reaction rule occurs in the dataset (i.e., at least 50, 100, 250, 500 and 1000 times) (Table 1). These numbers were selected to maintain sufficient number of samples per reaction class for statistical power, while keeping a relatively broad set of reaction rules (labels for classification) to encompass diverse chemistries. The resulting datasets contained 855, 462, 225, 129 and 73 unique reaction rules for classification (Table 1). The five resulting datasets were stratified by reaction rule, and split into training (80% of the total reactions) and held-out test sets (20% of the total reactions).

Table 1: Performance of rule-only models. Standard deviation is shown in parentheses. A breakdown of per class performance is presented in Supplementary Table S1.

Reaction Rule Occurs ≥	Average Balanced Accuracy	Total Rules for Classification	Reactions in Test Set
50	0.774 (0.1345)	855	91741
100	0.7863 (0.1293)	462	86297
250	0.8149 (0.1159)	225	79155
500	0.8127 (0.1071)	129	72377
1000	0.8208 (0.0971)	73	64814

Predicting Reactions with Deep Highway Networks Using Different Reaction Rule Set Sizes

To build our multiscale approach, we first employed an unsupervised method (e.g., Taylor-Butina algorithm[37,38]) to group the extracted reaction rules by chemical similarity (Figure 2a). Our model then works via two steps: first a deep neural network (DNN) predicts which *group* of reactions produces a molecule (the reaction group scale), and then a smaller, more focused DNN, trained only on that group of reactions, predicts which rule produces the molecule (the reaction rule scale) (Figure 2c). This is in contrast to the conventional implementation of rule-based retrosynthesis, in which a model is trained to predict a reaction rule from *all* the extracted reaction rules (Figure 2b). Product molecules were represented as fingerprints, and each molecule had an associated reaction rule and reaction group, which were used as labels at different steps in the training phase.

We formulated the task of retrosynthetic reaction prediction as a multinomial classification problem, similar to that proposed by Segler et al.[14,17] Briefly, given an input product molecule (encoded as a molecular fingerprint of 2048 bits), our retrosynthetic DL models predict which reaction rule (i.e. class label) was used to produce the molecule (Figure 2). We employed deep highway network architectures (DHN),[28] based on a combination of a single hidden layer, five highway layers, and a final softmax layer to output class label probabilities (Figure 2b). We initially trained five separate DHN on each of training datasets (Table 1).

It is important to note that our models were trained on the USPTO dataset,[36] which has been employed for forward/backward reaction prediction by other groups.[19,39,40] A critical limitation about using this dataset is that simpler chemistries are likely more represented than more advanced chemical reactions. Training a model on such an imbalanced set could possibly produce misleading results, resulting from the bias of the model towards highly represented chemical reactions. In order to minimize this limitation, we implemented class weights to train our models, and assessed classification performance using balanced accuracies,[41] as summarized in Methods.

The average balanced accuracies on the test set (consisting of 20% of total set) for differ-

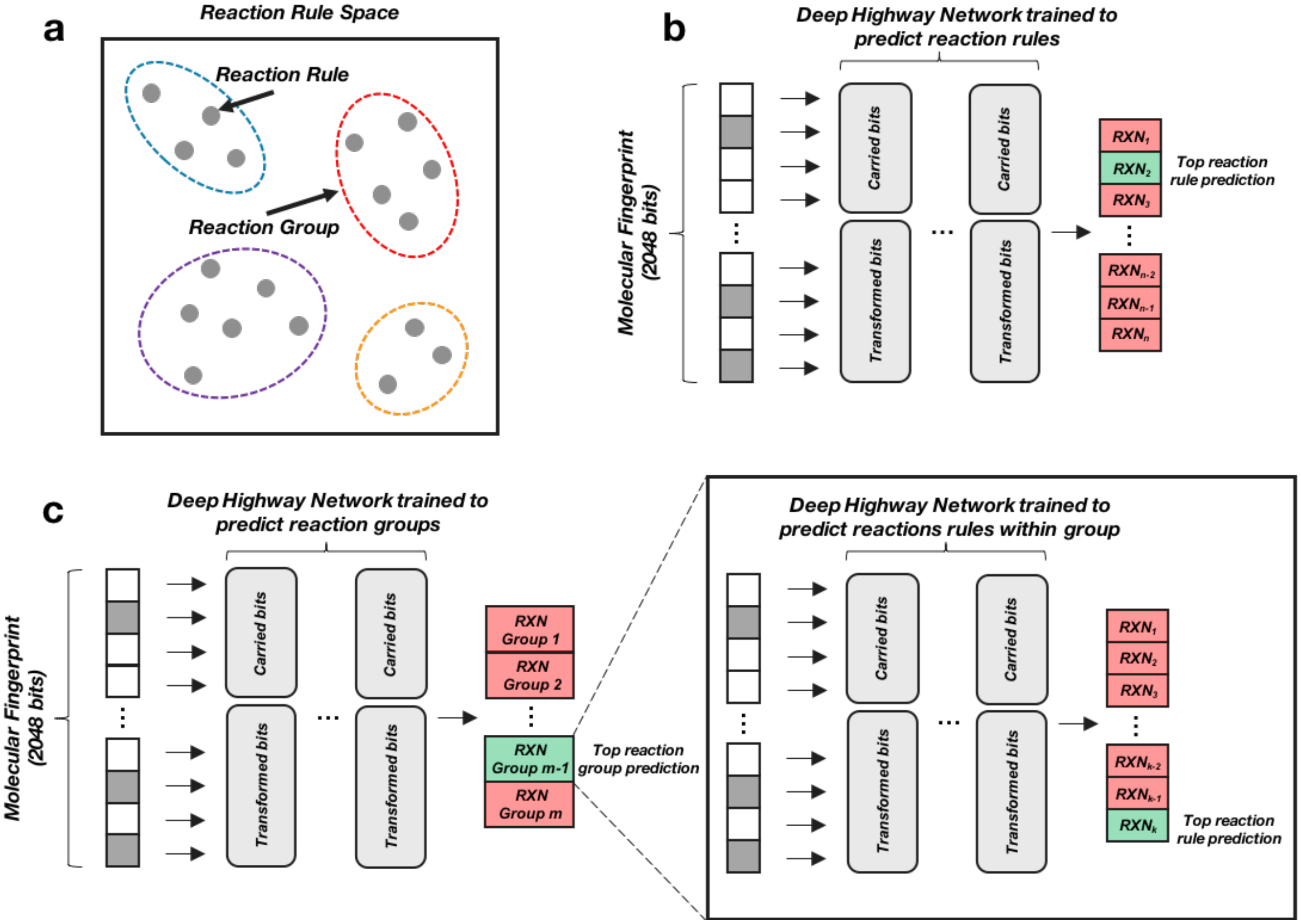
a
Reaction Rule Space
Reaction Rule
Reaction Group
b
Deep Highway Network trained to predict reaction rules
Molecular Fingerprint (2048 bits)
Carried bits
Transformed bits
RXN₁
RXN₂
RXN₃
RXNₙ₋₂
RXNₙ₋₁
RXNₙ
Top reaction rule prediction
c
Deep Highway Network trained to predict reaction groups
Molecular Fingerprint (2048 bits)
Carried bits
Transformed bits
RXN Group 1
RXN Group 2
RXN Group m-1
RXN Group m
Top reaction group prediction
Deep Highway Network trained to predict reactions rules within group
RXN₁
RXN₂
RXN₃
RXNₖ₋₂
RXNₖ₋₁
RXNₖ
Top reaction rule prediction

Figure 2: Summary of multiscale approach for retrosynthetic reaction prediction. (a) Schematic representation of unsupervised reaction rule grouping in our approach. Reaction rules are represented as gray dots, as groups as dashed lines. Reaction grouping was done based on chemical similarity of the reaction rules extracted from the USPTO dataset.[36] Reaction rules and reaction group membership were employed as labels for classification. (b) Schematic representation of a deep highway network (DHN) trained to predict on all extracted reaction rules. (c) Schematic representation of out multiscale approach. A DHN was trained for prediction on all reaction groups, using the same training set data as for the model depicted in (b), but stratified by reaction group. Once a prediction was made at the group scale, another DHN, trained to predict only on rules belonging to the predicted group, was employed to make predictions at the reaction scale.

ent reaction rule occurrence cutoffs ranged from 0.774 to 0.8208 (Table 1). These accuracies are comparable to the performance of other retrosynthetic models. For instance, a reported accuracy of 0.83 with 137 rules using the Reaxys dataset[14] compared to an accuracy of 0.81 for our model trained with 129 rules. It is important to note that we employed a smaller, different dataset than what has been previously reported, thus a direct comparison was difficult. Breaking down the predictive performance of the five models by class label (reaction rule) revealed that the models performed significantly better on some reaction rules than others (balanced accuracies of > 0.99 compared to < 0.5, with a mean range $= 0.4668$ and s.d. $= 0.0386$ over the five models), despite having a similar number of test samples (Supplementary Table S1). We hypothesized that this could be attributed to the chemical diversity of the products within each reaction rule class.

Table 2: Correlation between per class accuracy and the number of product molecule clusters in the each of the five test sets. Cutoff for Taylor–Butina clustering was set to 0.8. A complete breakdown of clusters per label is presented in Supplementary Table S2.

Reaction Rule Occurs ≥	**Total Cluster Number**	**Correlation Coefficient Balanced Accuracy vs. Clusters**
50	10406	-0.4306 (p<0.001)
100	8660	-0.5067 (p<0.001)
250	6979	-0.6619 (p<0.001)
500	5651	-0.6453 (p<0.001)
1000	4557	-0.5912 (p<0.001)

To quantitatively assess this, we clustered product molecules in our five test sets (using the Taylor–Butina algorithm with a cutoff of 0.8[38]) and calculated the correlation between per class accuracy and the number of product molecule clusters in the respective test set (Table 2 and Table S2). We observed a statistically significant negative correlation between per class accuracy and the number of product molecule clusters (ranging from -0.4306 to -0.6619 (p<0.001, by student's t-test), Table 2). For instance, the best performing label in the set with rule occurrence cutoff of 100 (*CC(O)=O.Nc1[nH][n][n][n]1»CC(=O)Nc1[nH][n][n][n]1* with an accuracy of 1.0) had only one cluster, while one of the worst performing labels (*CN(C)C=O»CNC* with an accuracy of 0.5223) had 13 clusters (Figure 3). This suggested that reduced chemical diversity of the products within each reaction rule class was associated with increased model performance (i.e. generalizability vs. specificity) and that our deep highway networks learned more generalizable patterns for classification on chemical subsets with high similarity (characterized by a small number of clusters) (Figure 3).

Thus, one strategy to improve model performance could be to stratify/balance datasets based on molecular similarity (e.g. use an algorithm such as Taylor–Butina to cluster product molecules in the dataset). However, this would require formulating *a priori* assumptions about the dataset that could harm the generalizability and applicability of retrosynthetic reaction prediction. In short, by taking this approach a model could learn to predict reaction rules for a very specific type of product molecule (highly populated cluster), but struggle with others (in less populated clusters) even if they were obtained using the same reaction rule. We hypothesized an alternative route to improve model performance, by building smaller and more focused retrosynthetic models on a smaller number of similar reaction rules (i.e. reaction grouping).

Classifying on Multiscale Reaction Rules Improves Deep Highway Network Performance

To test this hypothesis, we devised a strategy which grouped similar reaction rules together (termed reaction groups), thus creating a multiscale representation of each individual reaction rule (i.e. each reaction rule has group and rule information, Figure 4 and Supplementary Figure S1). This approach is similar to assigning a reaction *type* to

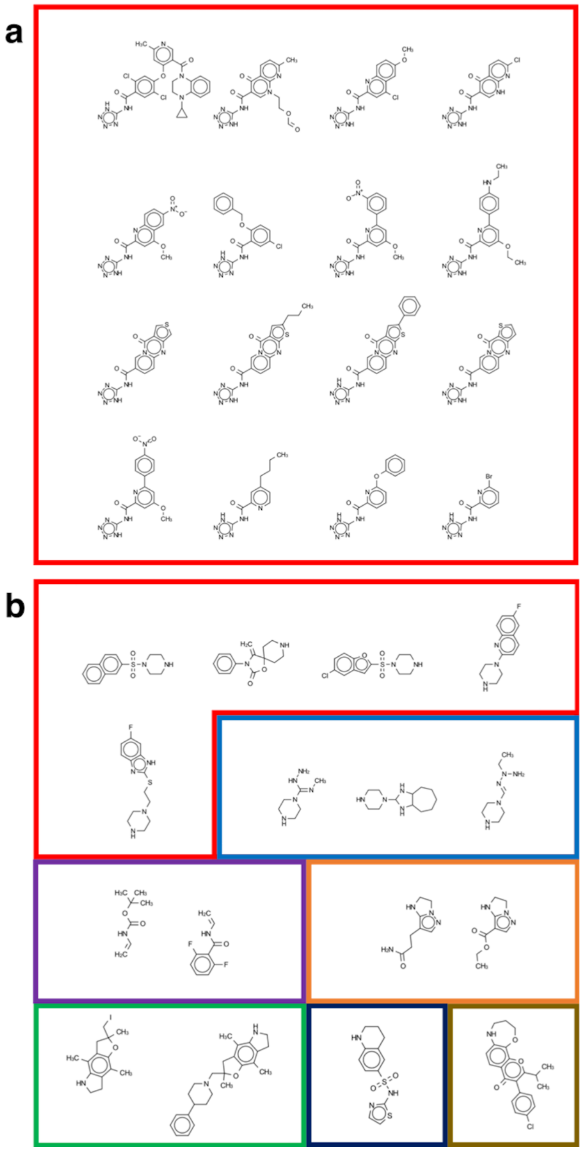

Figure 3: Chemical diversity of product molecules in dataset associated with reaction rule classes. Representative examples of clusters (i.e., most populated) of product molecules in test set with rule occurrence $\geq$ 100 for best (a) and worst (b) performing classes (*CC(O)=O.Nc1[nH][n][n][n]1»CC(=O)Nc1[nH][n][n][n]1* with 1.000 accuracy, and *CN(C)C=O»CNC* with 0.5223 accuracy, respectively). Each colored line represents a subset of members of different clusters. Clusters were obtained using the Taylor–Butina algorithm implemented in RDKit[42] (cutoff = 0.8).

each reaction in the dataset as a preprocessing step. Reaction type assignment is typically performed based on a known, predefined set of reaction types (e.g. using NameRXN, https://www.nextmovesoftware.com/namerxn.html). However, our approach is significantly different as the reaction similarity search is entirely data-driven and can easily be extended to any reaction dataset. To illustrate this, we performed reaction rule grouping by clustering reaction rule fingerprints, built from the reaction rules used in the five derived datasets, with RDKit (using cutoff = 0.7[38]). Reaction rules that were not placed into groups by the clustering algorithm were grouped together into a single group (the last group for each data set, Supplementary Table S3), in order to still consider them in the group model. We employed the same five datasets as in the previous rule-based DHN models described above, however stratification of training (80% of total data) and test data (20% of total data) was based on balancing reaction group number followed by balancing reaction rules.

Table 3: Performance of reaction group classification. Standard deviation is shown in parentheses. A breakdown of per class performance is presented in Supplementary Table S3.

Reaction Rule Occurs ≥	**Average Balanced Accuracy**	**Total Groups for Classification**	**Reactions in Test Set**
50	0.8386 (0.1078)	109	91741
100	0.8516 (0.098)	68	86297
250	0.8407 (0.0955)	32	79155
500	0.8436 (0.0799)	19	72377
1000	0.8517 (0.0605)	14	64814

We then trained DHNs on the multiscale reaction group labels (Figure 2c). Specifically, a DHN was trained to predict which reaction *group* (consisting of similar reactions obtained with reaction clustering) was employed to produce a molecule. Once the reaction group was predicted, another DHN (specific to the reaction group) was trained to predict the corresponding reaction rule (using only the reaction rules within the predicted reaction group,

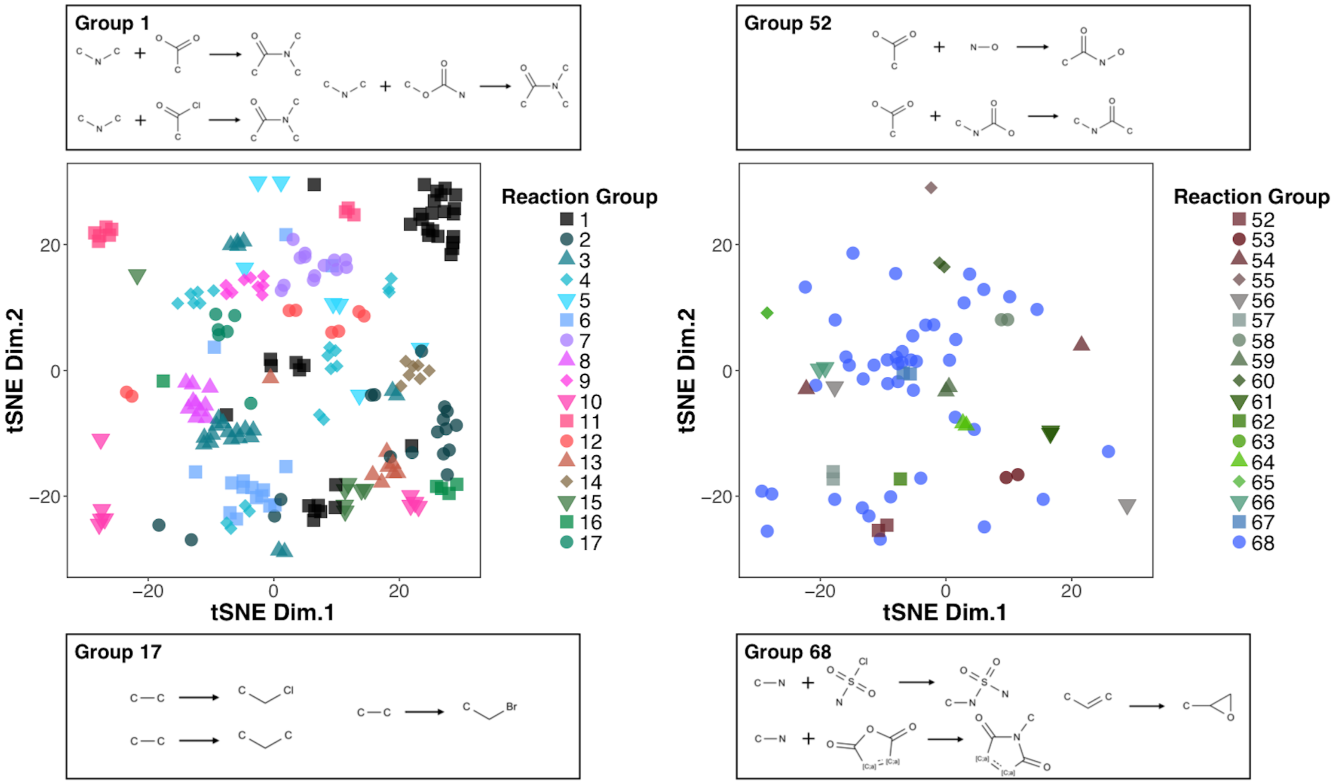

Figure 4: Visualization of identified reaction groups in the multiscale dataset. Distribution of first (left) and last (right) 17 reaction groups obtained from reaction rule clustering. Reaction rule clustering was performed with Taylor–Butina (cutoff = 0.7). The distribution of reaction groups was obtained using t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction over 2048 bits (dimensions) of the binary reaction fingerprints. The insets show structure of some representative reaction rules within the most populated groups. The distribution of the remaining reaction groups is presented in Supplementary Figure S1. To provide chemical meaning to the groups in the plots, complete breakdown of the rules within each group is presented in Table S4.

as opposed to all the derived rules as in our first two models) (Figure 2b). The first step, corresponding to reaction group prediction, was performed for each of the five datasets to select an appropriate balance between reaction rules (contained within the reaction groups) and model performance (measured by balanced classification accuracy) (Table 3 and Supplementary Table S3). Based on this, training of specific reaction groups models was continued using the dataset with reaction rule occurrence $\geq$ 100 (hereafter referred to as multiscale dataset), resulting in 68 additional DHN models, one for each specific reaction group (Table 4).

Overall, classifiers trained using multiscale reaction rules (i.e. with group and rule information) performed significantly better (e.g., mean accuracy increase from 0.7863 to 0.8982) than the rule reaction models (Figure 5 and Figure S2). The DHN models built using the multiscale dataset at the reaction group scale had a mean balanced accuracy of 0.8516 (Table 3) and a mean balanced accuracy of 0.8982 at the reaction rule scale (Table 4). This was a significant improvement compared to the previous DHN reaction-rule classifier for the same dataset, with average balanced accuracy of 0.7863 (Table 1). With our multiscale approach, 384 out of 462 reaction rules (i.e., 83%) showed an accuracy increase (Figure 5 and Supplementary Figure S2). Notably, the smaller classifiers built on reaction rules from the same reaction group achieved near perfect classification in many examples for the multiscale set (50 rules with a balanced accuracy $\geq$0.99, Table S4). The resultant increase in balanced accuracy with the multiscale approach (Figure 5 and Figure S2) indicated that our models learned across all the reaction rules in the training set, instead of just learning highly represented classes (data imbalance). This enhancement in classification performance can be directly attributed to reduced number of classification labels for each model (e.g., from 462 rule labels to 68 group labels for the multiscale dataset). After reaction grouping, the largest number of labels for multinomial classification was 50, and several of the reaction

Table 4: Summary of per label classification performance for the multiscale dataset (with rule occurrence cutoff ≥ 100). Average was taken over per rule balanced accuracies within each reaction group in the test set. Standard deviations are shown in parentheses. A complete breakdown of each group performance, as well as the reaction rules within the group, is presented in Table S4.

Reaction Group	**Average Group Balanced Accuracy**	**Total Rules in Group**	**Reactions in Test Set**
Group 1	0.8422 (0.0855)	45	16282
Group 2	0.8759 (0.0978)	27	16396
Group 3	0.8877 (0.08)	23	3622
Group 4	0.7865 (0.0962)	21	2636
Group 5	0.8792 (0.0764)	19	1993
Group 6	0.9132 (0.0835)	16	3162
Group 7	0.9079 (0.0781)	13	1250
Group 8	0.8659 (0.0922)	10	8537
Group 9	0.8579 (0.1138)	9	500
Group 10	0.8867 (0.0662)	9	2103
Group 11	0.8574 (0.0439)	9	316
Group 12	0.8173 (0.1204)	8	458
Group 13	0.9133 (0.0663)	8	851
Group 14	0.8545 (0.0948)	8	957
Group 15	0.9556 (0.0352)	7	404
Group 16	0.9032 (0.1038)	7	285
Group 17	0.8887 (0.1061)	6	759
Group 18	0.8408 (0.0925)	6	516
Group 19	0.8558 (0.1511)	6	6736
Group 20	0.9366 (0.0354)	6	411
Group 21	0.9548 (0.0444)	5	1007
Group 22	0.8962 (0.0531)	5	317
Group 23	0.861 (0.1377)	5	1364
Group 24	0.8359 (0.0803)	5	307
Group 25	0.9121 (0.069)	5	566
Group 26	0.8947 (0.0783)	5	870
Group 27	0.9786 (0.0162)	5	267
Group 28	0.9602 (0.041)	5	327
Group 29	0.9794 (0.0142)	5	189
Group 30	0.9151 (0.0317)	4	176
Group 31	0.9389 (0.0749)	4	732
Group 32	0.9587 (0.0463)	4	528
Group 33	0.8419 (0.0975)	4	152
Group 34	0.9768 (0.004)	4	140
Group 35	0.8484 (0.0811)	4	308
Group 36	0.9067 (0.064)	4	123
Group 37	0.9954 (0.0055)	4	280
Group 38	0.965 (0.0248)	3	97
Group 39	0.9717 (0.0186)	3	272
Group 40	0.9943 (0.005)	3	136
Group 41	0.9741 (0.0215)	3	120
Group 42	0.7363 (0.0307)	3	65
Group 43	0.9742 (0.02)	3	121
Group 44	0.826 (0.1636)	3	816
Group 45	0.8353 (0.1241)	3	327
Group 46	1.0 (0.0)	3	113
Group 47	0.8184 (0.0572)	3	118
Group 48	0.9441 (0.011)	3	106
Group 49	0.7895 (0.0163)	3	106
Group 50	0.8694 (0.0)	2	85
Group 51	0.84 (0.0)	2	65
Group 52	0.9505 (0.0)	2	127
Group 53	0.9154 (0.0)	2	135
Group 54	0.9772 (0.0)	2	55
Group 55	0.8055 (0.0)	2	81
Group 56	0.9902 (0.0)	2	307
Group 57	0.9815 (0.0)	2	117
Group 58	0.5147 (0.0)	2	68
Group 59	0.8921 (0.0)	2	88
Group 60	0.9857 (0.0)	2	72
Group 61	0.9348 (0.0)	2	57
Group 62	1.0 (0.0)	2	56
Group 63	0.9791 (0.0)	2	49
Group 64	0.708 (0.0)	2	180
Group 65	0.9361 (0.0)	2	60
Group 66	1.0 (0.0)	2	391
Group 67	0.8492 (0.0)	2	78
Group 68	0.9399 (0.0605)	50	6002
Average Balanced Accuracy	**0.8982 (0.0815)**	462	86297

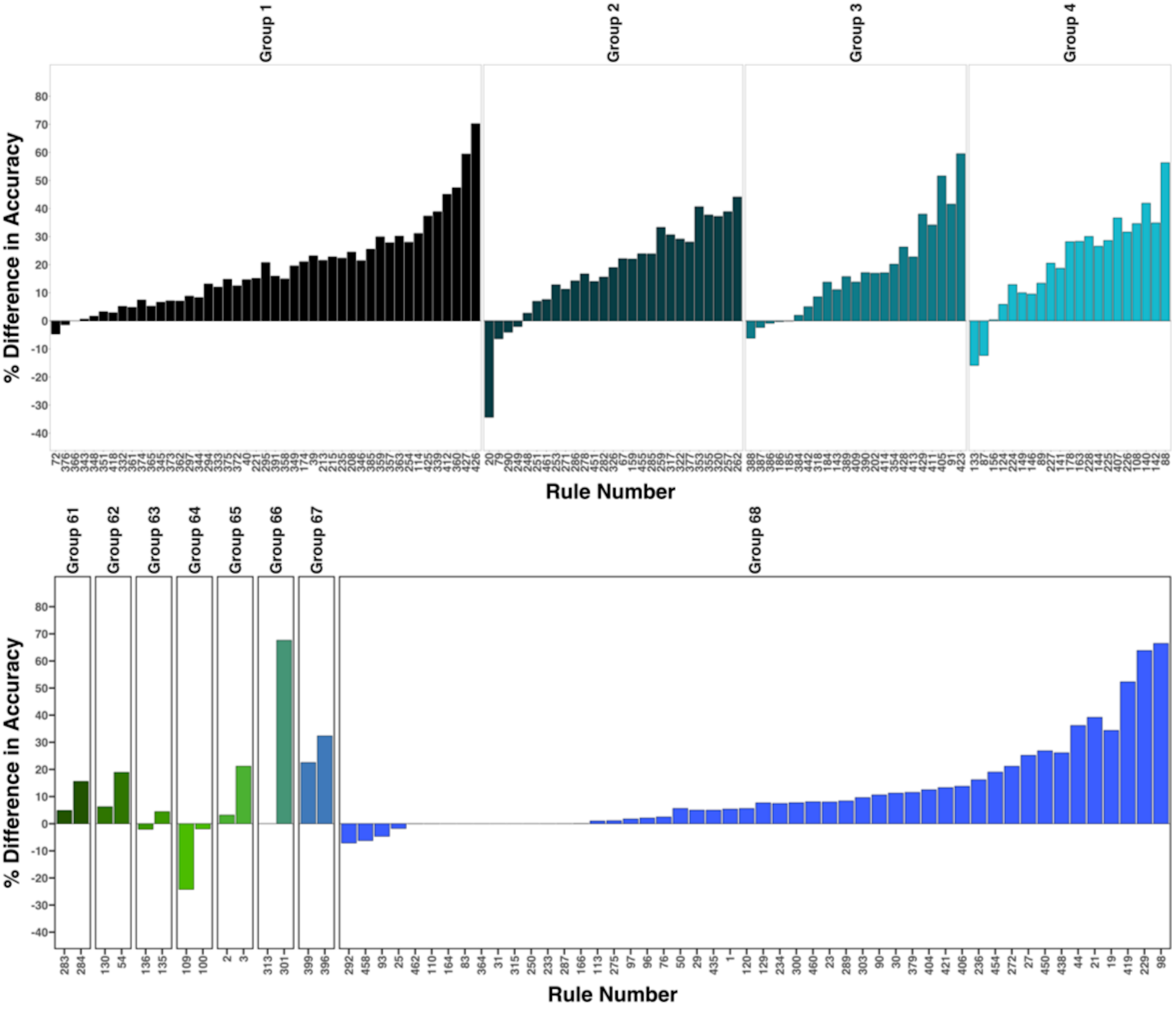

Figure 5: Classification performance increases in the multiscale models. Percentage difference in balanced accuracy for (top) the first four and (b) the last 8 reaction groups derived from the multiscale set (rule occurrence ≥ 100).The difference was taken between accuracies in the multiscale models and all rules models. The difference for the rest of the reaction groups is presented in Supplementary Figure S2. To provide chemically significant information to each label in the plot, the corresponding rules within each group are presented in Supplementary Table S4.

group models were reduced to binomial classifiers (Table 4), which contributed to the observed performance increase in classification.

Predicting Retrosynthetic Reactions with the Multiscale Reaction Rule Models

Our retrosynthetic analysis approach works at two levels. In the first level, DHN classifiers predict a (multiscale) reaction rule employed to make a product molecule, then, at the transformation level, the predicted reaction rule is applied to the molecule using RDKit. Briefly, the predicted reaction rule is loaded into RDKit as a reaction template, and this reaction is performed on the product molecule. Generally, running this reaction will return a list of transformed molecules which correspond to precursors of the product molecule. If the predicted transformation is valid, a list of precursor molecules (reactants) is generated, otherwise an empty list is returned, indicating that the predicted rule did not yield any reactants. By applying the transformation encoded in a predicted reaction rule to a product molecule using RDKit, we ensure that the model not only performs classification, but also provides an end-to-end solution to generate a list of reactants that can be used for chemist for synthesis planning.

Our multiscale approach for reaction prediction outperformed the previous rules-based reaction classification (e.g., 0.8982 vs. 0.7863 accuracy for multiscale and rule-only classification, respectively), which corresponds to the first or upper level of retrosynthetic analysis. To verify the applicability of our multiscale approach for reactant generation (the second or transformation level of our approach), we predicted the first retrosynthetic step of a variety of approved small molecules obtained from DrugBank[43] (Table 5). A detailed description about how these molecules were selected is presented in the Methods section.

Table 5: Summary of small molecules employed for Rule-based and Multiscale model validation. Corresponding literature associated with each predicted reaction is included (patent or journal article). The number of total valid predicted retrosynthetic steps (based on match with known routes) is shown in the last row for each model.

Small Molecule	Rule Model Prediction?	Rule Model Prediction References	Multiscale Model Prediction?	Multiscale Model Prediction References
Abacavir	Yes	J. Org. Chem. 1996	Yes	US5034394A, US6294540B1
Agomelatine	Yes	Tetrahedron Lett. 2014	-	-
Alogliptin	-	-	Yes	WO2010109468A1, US2012029000A1
Alvimopan	Yes	US5250542A	-	-
Aminolevulinic acid	-	-	Yes	Acta Pol. Pharm. 2003, Photochem. Photobiol. 2001
Apremilast	Yes	CN103864670A	Yes	CN103864670A
Armodafinil	Yes	WO2012078800A2, WO2003099774A1	Yes	WO2012078800A2, WO2003099774A1
Asunaprevir	Yes	WO2009085659A1, US8202996B2, WO2012166459A1, US20130129671A1, J. Med. Chem. 2014	-	-
Atomoxetine	Yes	Tetrahedron: Asymmetr. 2002	Yes	Tetrahedron: Asymmetr. 2002
Blonanserin	Yes	US5021421A, JP2013216655A	Yes	US5021421A, JP2013216655A
Cidofovir	Yes	J. Med. Chem. 1989, Nucleos. Nucleot. Nucl. 1989, Drug. Future 1996	Yes	J. Med. Chem. 1989, Nucleos. Nucleot. Nucl. 1989, Drug. Future 1996
Clevidipine	Yes	WO1995012578, US5856346, WO2000031035, US6350877, WO2011130852, WO2011127599	Yes	WO1995012578, US5856346, WO2000031035, US6350877, WO2011130852, WO2011127599
Darifenacin	-	-	Yes	WO2010032225A2, WO2009094957A1, WO2008029257A2
Dasatinib	-	-	Yes	WO2005077945A2, US2012302750A1
Dolasetron	Yes	WO2007072507A2	-	-
Eprosartan	-	-	Yes	US5185351, Drug. Future 1997
Flibanserin	Yes	WO9303016A1, US5576318A, J. Het. Chem. 1981	Yes	WO9303016A1, US5576318A, J. Het. Chem. 1981
Levobetaxolol	Yes	US4311708A	Yes	US4311708A
Levocetirizine	-	-	Yes	WO2009062036A2
Linagliptin	-	-	Yes	WO2006048427A1, US7820815B2
Lomitapide	Yes	WO9626205A1, US5712279A	Yes	WO9626205A1, US5712279A
Mycophenolate mofetil	Yes	Tetrahedron 2003	Yes	Eur. J. Org. Chem. 2013

Continued on next page

Table 5 – continued from previous page

Small Molecule	Rule Model Prediction?	Rule Model Prediction References	Multiscale Model Prediction?	Multiscale Model Prediction References
Mycophenolic acid	-	-	Yes	Tetrahedron 2003
Nateglinide	-	-	Yes	WO2005121071A1, US2013165686A1
Orlistat	Yes	Org. Lett. 2000	-	-
Pemetrexed	Yes	J. Med. Chem. 1992, US5344932A, US5028608A	Yes	J. Med. Chem. 1992, US5344932A, US5028608A
Penciclovir	-	-	Yes	Nucleos. Nucleot. Nucl. 1996, US5075445A, Drug. Future 1989
Pimecrolimus	-	-	Yes	US5912238, US6352998
Pralatrexate	Yes	WO2013164856A1	-	-
Ranolazine	Yes	Org. Process Res. Dev. 2012, US4567264A, WO2009153651A1, US20090318697A1, WO2008047388A2, WO2008139492A2, Tetrahedron Lett. 2013	Yes	Org. Process Res. Dev. 2012, US4567264A, WO2009153651A1, US20090318697A1, WO2008047388A2, WO2008139492A2, Tetrahedron Lett. 2013
Risperidone	-	-	Yes	US4804663A, WO0212200A1, US2002115673A1
Selexipag	Yes	WO02088084A1, US7205302B2, Bioorg. Med. Chem. 2007, J. Am. Chem. Soc. 1952	Yes	WO02088084A1, US7205302B2, Bioorg. Med. Chem. 2007, J. Am. Chem. Soc. 1952
Tamibarotene	-	-	Yes	CN102633673A
Tapentadol	Yes	WO2012089181A1, WO2012089177A1, WO2013090161A1, US2013150622A1, WO2013185928A1	-	-
Tazarotene	Yes	US5089509A	Yes	US5089509A
Telaprevir	-	-	Yes	Chem. Commun. 2010, Synlett 2013
Telaprevir	-	-	Yes	US20050197299A1, Lett. Drug Des. Discov. 2005, Drug. Future 2007
Telmisartan	Yes	J. Med. Chem. 1993, Org. Process Res. Dev. 2007, Drug. Future 1997	Yes	J. Med. Chem. 1993, Org. Process Res. Dev. 2007, Drug. Future 1997
Tezacaftor	Yes	US2009131492A1	Yes	US2009131492A1
Tipranavir	-	-	Yes	J. Am. Chem. Soc. 1997
Voriconazole	-	-	Yes	Org. Process Res. Dev.âĂŃ 2001
Total Valid Predicted Retrosynthetic Steps	**24/41 (58.5 %)**		**34/41 (82.9 %)**	

To compare the performance of the two models at the reactant-generation level, we focused only on the top prediction of the rule-based model, and the top group and top rule prediction of the multiscale model. The transformation in the top prediction of each model was applied to each molecule in our DrugBank set using RDkit. This resulted in a subset of 40 small molecules with known synthetic routes (obtained from Pharmacodia, http://en.pharmacodia.com) for which the top prediction of either model yielded a retrosynthetic transformation that is consistent with a known precluding synthetic step (Table 5). This subset of 40 molecules included a wide variety of small molecules with low similarity (as indicated by low Tanimoto coefficients, Supplementary Table S7), including antiviral and anticancer drugs (e.g., abacavir and dasatinib, respectively). For antiviral drug telaprevir, our multiscale model predicted two different valid retrosynthetic steps using the same multiscale rule. Thus, the total number of predicted retrosynthetic steps used to compare models was 41, including two different valid synthetic steps for telaprevir. Importantly, although U.S. patents have been filed associated with the tested small molecules (e.g., US5034394A for abacavir), these reactions were not included in any of our training/test datasets. Therefore, our models had no *a priori* knowledge of these molecules or their synthetic pathways.

Our multiscale approach produced 34 retrosynthetic predictions out of 41 (82.9% of the total predicted retrosynthetic steps)) that were consistent with known synthetic routes of the tested molecules (Table 5). In contrast, the rule-only model produced 24 predictions out of 41 consistent with known routes (58.5% of the total predicted retrosynthetic steps). This indicated that our multiscale approach not only outperformed the tradition rule-based model at the multinomial classification (upper) level, but also at the reactant-generation (transformation) level. Of the resulting valid predictions, both models made 17 common calls that were consistent with known synthetic steps (Figure 6 and Supplementary Figure S3). For two of these predictions (abacavir and mycophenolate mofetil), the top prediction of each model yielded different, valid retrosynthetic steps (Fig 6a). These transformations included acetylation (i.e., transformation with rule-based prediction for abacavir), and functional group removal (i.e., transforming mycphonelate mofetil to mycophenolic acid by removing 4-(2-hydroxyethyl) morpholinem). For the remaining 15 common predictions, transformations yielded the same retrosynthetic step (Figure 6b and Supplementary Figure S3), which in most cases included the main precursor reactant and reported reagent molecule. Reactions predicted in these cases were diverse, and included functional group protection deprotections and interconversions (i.e., for cidofovir and selexipag, respectively), and heterocycle formation (i.e., for tezacaftor).

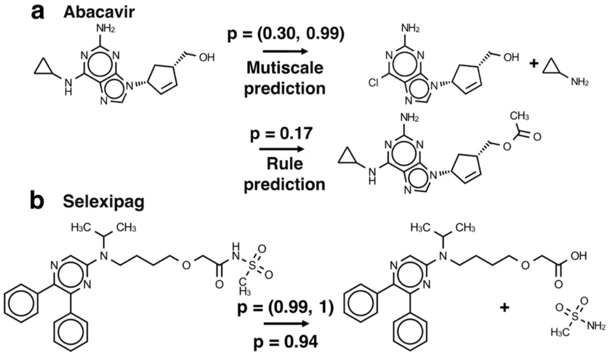

Figure 6: Examples of common retrosynthetic reaction predictions generated with multiscale and rule-based models. (a) Different valid prediction obtained with each model for abacavir. Multiscale probabilities (for reaction group and reaction rule) are shown in parentheses in the figures. (b) For selexipag, both models generated the same valid prediction. The remaining examples are presented in Supplementary Figure S3.

Notably, the multiscale model made 13 calls that yielded a known retrosynthetic step for the tested small molecules, which the rule-based model missed (Supplementary Figure S5). Some of the miscalls by the rule-based model were subtle, missing a functional group in an otherwise similar reactant (i.e., predicting a carbonyl group instead of Br in darifencin, Figure 7), which highlighted the advantage of our multiscale approach in recognizing structural patterns within the product molecules over the model trained on the entire dataset. In contrast, the rule-based model made 7 calls matching known precluding steps that the multiscale model missed (Supplementary Figure S4). In most of these cases, the multiscale model predicted the correct region were the retrosynthetic transformation occurs, but with the incorrect functional group (i.e., protection of OH group in pralatrexate, as in Figure 7).

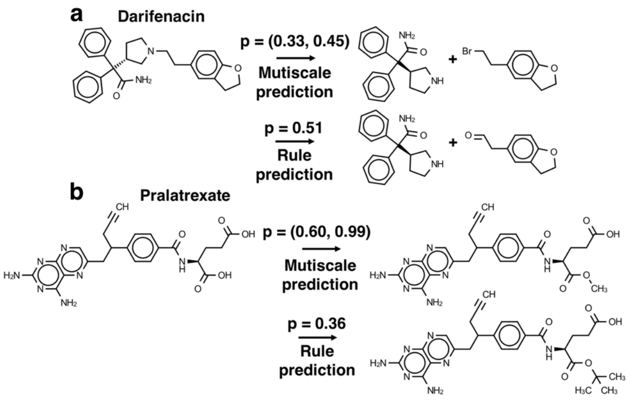

Figure 7: Comparison of reactions predicted with both models. (a) Retrosynthetic reaction that was correctly predicted by the multiscale approach for darifenacin, but miscalled with the rule-only model. (b) Reaction for pralatrexate that was correctly predicted by the rule-based model, but miscalled with the multiscale approach. The rest of the reactions are shown in Supplementary Figs. S4 and S5.

As an additional test for the enhancement provided by out multiscale approach, we performed retrosynthesis based on molecular similarity over the 40 small molecules in the validation set. For this analysis, we employed the algorithm established by Coley and coworkers.[16] This algorithm provided 11 valid predictions that matched ground truth of known synthetic steps (Supplementary Figure S7). In the context of this work, this means that this algorithm has a success rate of 26.8% (11 out of 41 steps), whereas our multiscale approach has a success rate of 82.9%. This comparison indicated the significant enhancement in the retrosynthetic reaction prediction task for pharmaceutical molecules by our approach.

We also tested how the predictions of our multiscale approach compare to the state-of-the-art software Chematica, by employing the eight commercially valuable and structurally diverse targets for which synthetic routes have been generated recently.[18] For this test, we used Chematica predictions recently published and experimentally validated as ground truth to benchmark the performance of our multiscale approach.[18] Although our multiscale approach generated reactants for all the presented molecules, our predictions failed to match the predictions made by Chematica (Supplementary Figure S8). This assessment highlights the limitation of our approach to make predictions for sophisticated molecules in which synthesis includes chemistry that is well beyond the scope of the USPTO dataset used to train our models. To the best of our understanding, this is the first time a DL approach for retrosynthetic reaction prediction has been assessed with Chematica predictions. Based on these findings, we recommend this comparison with Chematica as a robust benchmark for future applications.

Effect of Reaction Rule size in Retrosynthetic Reaction Prediction

In addition to the aforementioned predictions, our multiscale model made 4 partial calls that are consistent with the known precluding step (Figure 8 and Supplementary Figure S6). In these four cases, the multiscale model correctly predicted the addition of a protecting group to the product molecule as part of the retrosynthetic step; however, the group did not fully match the protecting group in the known reaction, and instead resulted in the addition of a closely related functional group (Figure 8 and Supplementary Figure S6).

Specifically, the multiscale model predicted the addition of maleimide instead of phthalimide (i.e., for antidiabetic drug alogliptin, Figure 8), or TMS (Trimethylsilyl) instead of TBS (tert-Butyldimethylsilyl) (Figure S6). This was likely due to an inherent limitation of our rule extraction step, which only considered the reactive center and its first shell of neighboring atoms (Figure 1). Under this scheme, our model was not able to learn on a reaction rule that would yield the addition of a larger functional group (for example, TBS), which would require to extend rule extraction beyond only first neighbors. However, even with this inherent limitation, our multiscale model was able to make a partial prediction that is consistent with the known reaction step. In contrast, the rule-based model was not able to

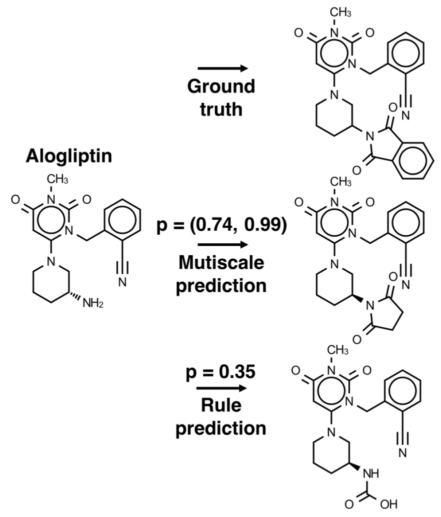

Figure 8: Example of partially correct prediction obtained with the multiscale approach. A retrosynthetic reaction that was partially predicted by the multiscale model. In this case, the functional group addition predicted by the multiscale model is limited by the results of the reaction rule extraction step, which only included reactive center and its first-neighbor atoms. The multiscale model did not know the rule to make the ground truth prediction, which would require more detailed rules from the dataset (e.g., first and second atoms). Other examples are shown in Supplementary Figure S6.

make the correct predictions, even though it was trained with the same set of reaction rules. This observation demonstrated the advantage of our multiscale approach over the rule-only models trained on the same dataset for retrosynthetic reaction prediction.

Conclusion

In this work, we have presented a data-driven, multiscale approach based on DHN and reaction rule classification for retrosynthetic reaction prediction. Our approach performs the reaction prediction task in two steps. First, a DHN is used to predict which reaction group (consisting of reaction rules group by chemical similarity) was used to make a molecule (the reaction group scale). Once a reaction group prediction is made, a more specific model, trained only on reaction rules within the predicted reaction group, is employed to predict a reaction rule (the reaction rule scale). This results in a larger number of DHN models for the multiscale model (determined by the number of reaction groups extracted from the dataset), as opposed to a single DHN model for a rule-based approach. Finally, once a reaction rule is obtained, the transformation is applied to the input molecule to derive chemically viable reactants.

To compare the performance of our multiscale approach to the conventional machine learning rule-based model, we employed a set of approved pharmaceutical small molecules and predicted their first precluding synthetic step. The multiscale model outperforms the conventional rule-based multinomial classification approach, where a model is trained to make predictions over all the reaction rules of the dataset, both at the classification level (the multiscale model has a higher average balanced accuracy), and the reactant-generation level (the multiscale model produces more reactions that match known synthetic routes). This indicated that our multiscale approach clearly enhances the performance of DL for retrosynthetic reaction prediction task relative to all-rules machine learning models.[14,17] Moreover, our multiscale approach significantly outperformed another retrosynthetic approach based on molecular similarity.[16]Importantly, however, our multiscale approach was unable to correctly predict the synthetic steps recently generated with Chematica for pharmaceutical molecules.[18] The inability of our approach to match Chematica predictions highlights an important current limitation of entirely machine-extracted rule systems, in which simpler chemistries are likely more represented than more advanced chemical reactions. One way to alleviate this limitation in the future would be to incorporate reaction rules encoded by expert chemists as class labels, along with corresponding examples of matching chemical reactions to train a DL model.

Overall, due to relative simplicity of molecular featurization used (fingerprints), we believe that out multiscale approach can easily be integrated into current cheminformatic platforms for synthesis planning. The comparison with other machine learning-based approaches indicated that our multiscale approach provided an enhancement in the first component (i.e., the retrosynthetic reaction prediction task) of a DL-based retrosynthetic pipeline. To further expand its usability for this task, our multiscale approach can be coupled with another algorithm, such as Monte Carlo tree search.,[17] to design complete retrosynthetic pathways for pharmaceutical molecules.

As our results show, the size restriction in the rule extraction step might have an impact in the chemical structure of predicted reactants. One future improvement to our model would be to perform a flexible reaction rule extraction step, in which the shell of neighboring atoms around the reactive center is not fixed in size. Reactive centers and reaction rules could be learned, for example, by ”crawling” along the edges of a molecular graph, parametrized by neural networks, as recently proposed.[44] In this flexible reaction rule extraction step, functional groups could be learned by exploring the neighborhood of the reactive center in both sides of a chemical reaction, and comparing reactants and products parametrized as molecular graphs, instead of mapped atoms.

One key limitation of our approach (and other DL-based retrosynthetic analysis approaches[14,17,19]) is the absence of any information in the model about the conditions in which the reaction occurs. This is a key component of any computational technique employed for chemical synthesis planning. In the future, our multiscale model could be integrated with other DL models built for optimizing chemical reaction conditions, for example a deep reinforcement learning model,[45] to predict complete retrosynthetic routes. Another key aspect of models trained on datasets of published literature (like the USPTO dataset) is that they are likely biased towards published chemical reactions that work. One potential way to overcome this inherent model bias would be to build deep reinforcement learning models that also learn on reactions that do not work (e.g., negative results). One caveat, however, is that, to the best of our knowledge, such datasets of negative results are not publicly available.

Methods

Patent Dataset and Reaction Rule Extraction

The set of over one million chemical reactions extracted from United States granted patents from the years between 1976 and 2016 was employed.[36] This dataset is freely available online, and has been commonly used by other groups for forward/backward reaction prediction tasks.[19,39,40] The reactions found in this dataset were preprocessed with RDKit[42] to eliminate reagents, in order for reaction to only contain reactants and products before the rule extraction step,as proposed by Schneider et. al.[40] This step was performed to minimize the possibility of incorrect atom mapping (by taking into account reagents), leading to incorrect reaction rule assignment. After the reagent removal step, reactions were atom mapped using the Indigo toolkit version 1.2.3.[46] The goal of this remapping step was to minimize mapping errors that could be present in the USPTO dataset, as highlighted in its documentation.[36]

Rule extraction was implemented as shown in Fig. 1, employing a strategy described in detail elsewhere.[7,12] The rule extraction step was performed with custom scripts using RDKit.[42] Briefly, for each mapped reaction, the reactive core (e.g., atoms and bonds that change between reactants and products) was identified by comparing the attributes of corresponding mapped atoms. The considered attributes included charge, bond type, valence and number of neighbors, as previously described.[7] Finally, the reactive center was extended to include first neighbors, in order to include more details about the chemical structure of the reactive center. The extracted small and large reaction rules were employed as labels for the reaction classification task. A total of 74,482 unique reaction rules were extracted from the dataset, respectively.

Preparation of Training and Testing Sets for Rule-Based and Multiscale Retrosynthetic Reaction Prediction

Product molecules were encoded as Morgan fingerprints (FP), a form of extended-connectivity fingerprints (ECFP).[26] Each molecule was converted to a 2048-bits FP (of radius 2) and vectorized using RDKit.[42] The resulting vectors were employed as the input data for our deep learning (DL) models. A fingerprint of length 2048 was selected after assessing the classification performance using different lengths while maintaining set sizes that were manageable with our computational resources (Table 6 and Supplementary Table S5).

Table 6: Effect of fingerprint size in classification performance. Subset with best performing model of reaction rules that occurred $\geq$ 1000 was employed for the comparison. Standard deviation is shown in parentheses. A complete breakdown of balanced accuracies per class is presented in Supplementary Table 5.

Fingerprint Size in Bits	**Average Balanced Accuracy**
512	0.7802 (0.1134)
1024	0.8037 (0.1086)
2048	0.8208 (0.0971)

We generated five sub-datesets by defining a cutoff on the number of times that each reaction rule occurs (Table 1). The cutoffs employed were 50, 100, 250, 500 and 1000. These numbers were selected to maintain robustness in the dataset while preserving diversity in the number of reaction rules employed. We note that this numbers are similar to the ones reported by Segler et al.,[14,17] which were obtained from a larger dataset (the Reaxys dataset[27]).

To build DL models for classification on a smaller number of similar reactions for our multiscale approach, we performed grouping of the reaction rules, and use reaction group membership as labels for classification to generate additional sub-datasets for modeling. For this step, each reaction rule was encoded as a difference reaction FP, and the pair-wise Tanimoto similarity matrix[47] of these FPs was built. Grouping (clustering) was performed on this distance matrix with the Taylor-Butina[37,38] method as implemented in RDKit[42] (with a cutoff of 0.7, or distance to cluster center of 0.3). Finally, corresponding group labels were assigned to each product in the five datasets (Table 3). To visually inspect the resulting reaction groups, we employed t-SNE for dimensionality reduction.[48]

.

Finally, the datasets were split 8:2 for training and testing, respectively. For each model, the same datasets were employed, but the data was cut differently, depending on the labels for classification. Data was stratified using reaction rules or cluster labels, in order to take care of the data imbalance in the dataset. We choose to take this approach instead of using a balanced dataset (with same number of samples per class) to account for the data imbalance that is expected in a typical chemical dataset, were some reaction would be more represented than others.

Quantifying Chemical Diversity within Testing Sets

To quantify the degree of chemical diversity within each reaction rule class, we performed Taylor-Butina clustering on the product molecules within each of the five testing sets (one for each rule occurrence cutoff, Table S2). Product molecules were encoded as 2048-bits FP (of radius 2) RDKit.[42] The pair-wise Tanimoto similarity matrix[47] was obtained, and Taylor-Butina clustering[37,38] was performed with RDKit, using a cutoff = 0.8 (distance to cluster center of 0.2). To assess the effect of chemical diversity in classification performance, the correlation coefficient between per class balanced accuracy and number of clusters associated with that class was calculated (Table 2). Briefly, we employed Pearson's correlation coefficient, which is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{1}$$

where n is the number of classes, x_i, y_i are individual values of per class balanced accuracy and number of clusters associated with that class, respectively, and $\overline{x}$, $\overline{y}$ are their mean values, given by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

Pearson's correlation coefficient can take values between -1 and 1, which indicate total negative or positive linear correlation, respectively. In our data, we observed statistically significant negative correlation between class balanced accuracy and number of cluster per class (as summarized in Table 2), which indicates that reduced chemical diversity was associated with increased model classification performance.

Reaction Classification with Deep Highway Networks

The retrosynthetic reaction prediction task was formulated as a multicass classification problem, as proposed by Segler et al.[14] We employed a similar neural network architecture based on a combination of a hidden layer and highway networks.[28] Briefly, highway networks differ from typical neural networks in that they employ gating mechanisms to regulate the flow of information (Fig. 2b). This allows for a portion of unmodified input to be passed across layers together with activation. The same core architecture was employed for the rule-based and mulstiscale models.

Models were built using Keras[49] with the TensorFlow[50] back end. The hidden layer included 2048 neurons (one for each bit of the FP we employed), with an exponential linear unit (ELU)[51] (followed by dropout value of 0.2). This was followed by five highway layers[28] with rectified linear units (ReLU),[52] followed with a dropout value of 0.1. The last layer of the network was a softmax to output class label probabilities. All of the layers in the network had normal initialization. The ADAM optimizer (with learning rate of 0.001) was used to minimize the binary cross entropy loss function for classification. Class weights (determined by the number of samples in each class) were implemented to take care of data imbalance in the training set. As mentioned before, we employed this approach to consider the data imbalance expected in a typical chemical dataset (e.g., some types of chemistries being more represented than other in the USPTO dataset). The number of training epochs for the models was determined by early stopping (with patience of 2), implemented by monitoring the loss on a validation dataset (10% of the training set).

Model Performance Indicators

To asses the classification performance of our models we calculated the balanced accuracy,[41] which is defined as

$$BalancedAccuracy = \frac{(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})}{2} \tag{3}$$

where TP is the true positives, FP is the false positives, TN is the true negatives, and FN is the false negatives generated by a trained model using the test set for each class. We chose this metric in order to account for the data imbalance expected in a chemical reaction dataset (i.e., some types of reactions are more represented than others). In addition to calculating the balanced accuracy per class (reported in Supplementary Tables 1 and 4), we also calculated the average balanced accuracy per model, in order to compare the global performance of the conventional rule-based model and the multiscale approach. Average balanced accuracy was calculated by taken the mean, given by Equation 2 over all per class balanced accuracies generated with different models. To estimate the amount of variation among the balanced accuracies, in addition the average balanced accuracies, we also report the the standard deviation, which is given by

$$s = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{4}$$

where x_i are individual values of per class balanced accuracy, $\overline{x}$ is the average balanced accuracy, and n is the number of classes.

Derivation of Validation Set of Small Molecules

We tested the applicability of our multiscale approach and of the rule-based model on a validation set of 2078 molecules extracted from DrugBank[43] with query terms ”Approved is true” and ”Small molecule is true”. A summary of the molecules employed and the predictions is provided in Supplementary Table 6. Importantly, these molecules were not present in our training sets. The goal of this approach was to employ molecules with known pharmaceutical importance for validation of our approach and comparison to other methods. There was an overlap of 864 molecules for which each model made a reactant prediction, which provided a starting point for comparison of the models. For this step, our multiscale model employed 206 unique rules (out of a total of 462) for its top predictions, while the rule-based model employed 230 unique rules for its corresponding predictions. This indicated that our trained models are able to explore different ”regions” within the chemical space defined by the extracted reaction rules, and are not biased towards a particular corner or reaction. We filtered this initial set of small molecules based on whether or not a published synthesis route was available for the input molecules. Synthetic routes were obtained from pharmacodia (http://en.pharmacodia.com). This step was crucial to provide a ground truth to validate any of the predictions generated by either model. Synthetic route availability further reduced the validation set from 864 to 154 molecules with known and published synthetic routes.

Since the goal of our work was to directly compare the usability of our multiscale approach to generate reactants with respect to the conventional rule-based system, we only report the molecules in the validation set for which the top prediction made by both models resulted in a valid transformation that yielded a list of reactants after applying the reaction transformation with RDKit. This final subset of molecules consisted of the 40 small molecules reported in this work. By inspecting the results of the predictions and comparing them to the ground truth of know synthetic steps for each molecule, we were able to provide a metric of model performance for generating reactants after class label prediction.

Supporting Information Available

Tables S1-S7 include a breakdown of model performance metrics and reaction predictions made by all the models presented in this manuscript.

Table S8 includes a summary of technical terms employed in this work.

Figure S1 is a t-SNE plot for reactions within groups 18 to 51.

Figure S2 is the percentage difference in balanced accuracy for reaction classification in reaction groups 5 to 60.

Figures S3-S6 are additional chemical transformations predicted by the models presented in this work.

Figure S7 are predictions that match ground truth obtained with a molecular similarity algorithm for retrosynthesis, established by Coley and coworkers.[16]

Figure S8 are multiscale prediction results on commercially valuable and/or medicinally relevant targets evaluated with Chematica.[18]

References

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19–38.

(2) Corey, E. J. LXIII.–A Synthesis of Tropinone. *J. Chem. Soc. Trans.* **1917**, *111*, 762–768.

(3) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley, 1989.

(4) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.

(5) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-Aided Synthesis Design: 40 Years on. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 79–107.

(6) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 5904–5937.

(7) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; ; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602.

(8) Fick, R.; Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Design of Synthesis for Heterocyclic Compounds. *Heterocycles* **1995**, *40*, 993–1007.

(9) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Model.* **1995**, *35*, 34–44.

(10) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Model.* **1990**, *30*, 492–504.

(11) Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. *J. Chem. Inf. Model.* **1999**, *39*, 316–325.

(12) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745–1756.

(13) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19*, 357–368.

(14) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. Eur. J.* **2017**, *23*, 6118–6128.

(15) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.

(16) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237–1245.

(17) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.

(18) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *3*, 522–532.

(19) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Nguyen, Q. L.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113.

(20) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540.

(21) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.

(22) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. S. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *CoRR* **2017**, *abs/1709.04555*.

(23) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Vranken, D. V.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442–452.

(24) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(25) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(26) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(27) Chemistry data and chemical literature - Reaxys. `http://www.reaxys.com`, Accessed September 5th, 2017.

(28) Srivastava, R. K.; Greff, K.; Schmidhuber, J. Training Very Deep Networks. *CoRR* **2015**, *abs/1507.06228*.

(29) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35.

(30) Chakravarti, S. K. Distributed Representation of Chemical Fragments. *ACS Omega* **2018**, *3*, 2825–2836.

(31) Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *CoRR* **2015**, *abs/1509.09292*.

(32) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 595–608.

(33) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; ; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(34) Willett, P. Similarity Searching Using 2D Structural Fingerprints. *Methods Mol. Biol.* **2014**, *6*, 133–158.

(35) Franco, P.; Porta, N.; Holliday, J. D.; Willett, P. Molecular Similarity Considerations in the Licensing of Orphan Drugs. *Drug Discov. Today* **2017**, *22*, 377–381.

(36) Patent data. `https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873`, Accessed September 5th, 2017.

(37) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Model.* **1995**, *35*, 59–67.

(38) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and

Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Model.* **1999**, *39*, 747–750.

(39) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39–53.

(40) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56*, 2336–2346.

(41) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. 2010 20th International Conference on Pattern Recognition. 2010; pp 3121–3124.

(42) Landrum, G. A. RDKit: Open-Source Cheminformatics Software. `http://www.rdkit.org`, Accessed September 5th, 2017.

(43) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(44) Urban, G.; Subrahmanya, N.; Baldi, P. Inner and Outer Recursive Neural Networks for Chemoinformatics Applications. *J. Chem. Inf. Model.* **2018**, *58*, 207–211.

(45) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344.

(46) Epam Life Sciences. Indigo Toolkit. `http://lifescience.opensource.epam.com/indigo/index.html`, Accessed September 5th, 2017.

(47) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115–1118.

(48) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2018**, *9*, 2579–2605.

(49) Chollet, F. Keras. `https://keras.io`, 2015; Accessed September 5th, 2017.

(50) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; `https://www.tensorflow.org/`, Software available from tensorflow.org.

(51) Clevert, D.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR* **2015**, *abs/1511.07289*.

(52) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on International Conference on Machine Learning. USA, 2010; pp 807–814.

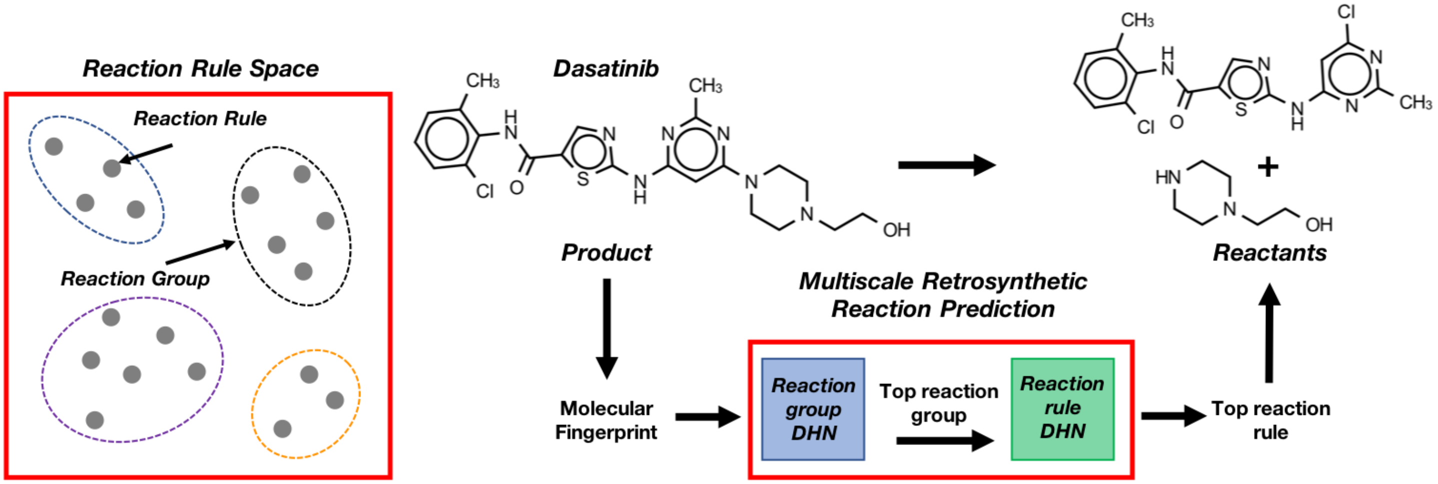

Table of Contents Graphic