Retrosynthetic Accessibility Score (RAscore) - Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning

Amol Thakkar,1,2* Veronika Chadimová,1 Esben Jannik Bjerrum,1 Ola Engkvist,1 and Jean-Louis Reymond2*

1 Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Gothenburg, 431 50, Sweden.

2 Department of Chemistry and Biochemistry, University of Bern, Bern, CH-3012, Switzerland.

*Corresponding authors: amol.thakkar@dcb.unibe.ch, jean-louis.reymond@dcb.unibe.ch

Abstract

Computer aided synthesis planning (CASP) is part of a suite of artificial intelligence (AI) based tools that are able to propose synthesis to a wide range of compounds. However, at present they are too slow to be used to screen the synthetic feasibility of millions of generated or enumerated compounds before identification of potential bioactivity by virtual screening (VS) workflows. Herein we report a machine learning (ML) based method capable of classifying whether a synthetic route can be identified for a particular compound or not by the CASP tool AiZynthFinder. The resulting ML models return a retrosynthetic accessibility score (RAscore) of any molecule of interest, and computes 4,500 times faster than retrosynthetic analysis performed by the underlying CASP tool. The RAcore should be useful for the pre-screening millions of virtual molecules from enumerated databases or generative models for synthetic accessibility and produce higher quality databases for virtual screening of biological activity.
Introduction

Artificial intelligence (AI) in chemical discovery has been driving improvements in the tools available to the chemical community. This has occurred primarily in the areas of de novo generation of new chemical entities (NCE), toxicology/bioactivity, and computer aided synthesis planning (CASP). The question as to which molecule to make and how to make it, is at the center of chemical discovery programs across academia and a range of industries, ranging from agrochemicals to pharmaceuticals. Typically virtual screening (VS) workflows have been used to decide which compounds to make, starting from generated, enumerated, commercial or public datasets which are then filtered using a variety of statistical and physics based modelling until the search space is refined (Figure 1). The question and decision of how to make a given set of compounds is left to a team of chemists at the end of the VS workflow, prior to synthesis in the laboratory.

Figure 1: Example of a virtual screening (VS) workflow. The synthesis of compounds is typically considered at the end of the workflow as a final selection criteria, and it is at this point CASP is also used to filter compound libraries to synthesizable compounds. Our method allows for pre-screening of compounds that may be synthetically accessible by CASP enabling use earlier in the VS workflow (green).

CASP has emerged as a method by which compounds can be filtered in the VS workflow, and during optimization cycles throughout the generative modelling process. These can be used at
two potential stages of the generation process, either to bias the generation process or as a post-hoc filter after the molecules have been generated.\textsuperscript{11} Given a target compound, CASP can predict each step of the synthesis pathway from commercially available building blocks. This makes it suitable for the \textit{in silico} filtering of large compound libraries, and has been demonstrated by Gao and Coley for the case of generated compounds.\textsuperscript{11} However, despite the vast amount of progress that has contributed to making the prediction of full synthetic routes computationally tractable,\textsuperscript{6, 12-14} to the extent that some predictions may be made within a minute.\textsuperscript{5, 6} The scale at which predictions must be conducted for large compound libraries consisting of several million or even billions of compounds can still be limiting.

Herein, we propose the retrosynthetic accessibility score (RAscore) that enables rapid estimation of synthetic feasibility as determined from the predictions of CASP. We investigate a machine learning classifier for retrosynthetic accessibility (RA) assessment called RAscore, trained on the outcomes generated from a CASP tool, which we have shown can increase the speed at which synthetic accessibility can be estimated, and separate compounds for which retrosynthetic routes can be found by CASP better than existing synthesis scores, such as the synthetic accessibility score (SAscore), synthetic complexity score (SCscore), and synthetic bayesian accessibility (SYBA).\textsuperscript{15-18} The SAscore and SYBA are estimations of synthetic feasibility based on the occurrence of molecular fragments in public databases, whereas SCscore is learned from a reaction corpus, with the underlying assumption that products are more complex than their constituent reactants. Used in combination with the previous scores the RAscore should enable pre-screening of compounds that can be later subjected to full retrosynthetic analysis, and enables CASP to be used at earlier stages of a VS workflow.
Methods

**AiZynthFinder – A Tool for Computer Aided Synthesis Planning**

AiZynthFinder is a template-based retrosynthetic planning tool based on the methodology of Segler and Waller.\(^4,18\) It consists of a neural network policy, which determines which reaction to use at a given retrosynthetic step, with Monte-Carlo tree search, as reported in our previous studies.\(^19\) The code, data, and models are open source and available to the public: [https://github.com/MolecularAI/AiZynthFinder](https://github.com/MolecularAI/AiZynthFinder). The reaction transforms have been extracted from the US patent office extracts (USPTO) and used to train the model by which retrosynthetic expansion was conducted.\(^20\) Models on Reaxys and proprietary datasets have been examined in our previous studies, but have been omitted in this study due to their proprietary nature.\(^19\)

AiZynthFinder considers retrosynthetic routes to be solved if the precursors or building blocks are commercially available. Therefore, as stopping criteria we use the ACD catalogue,\(^21\) Enamine building block set,\(^22\) and AstraZeneca internal database. These are available from the respective vendors with the exception of the AstraZeneca internal catalogue. In place of the vendors mentioned here, the AiZynthFinder GitHub repository contains a set of compounds extracted from the ZINC database,\(^23\) as highlighted in our previous work.\(^18\)

**Retrosynthesis Prediction for Training Set Generation**

Training and test datasets were generated by randomly sampling 200,000 compounds from ChEMBL,\(^24\) and 100,000 compounds each from GDBChEMBL and GDBMedChem.\(^25,26\) The compounds were subsequently subjected to retrosynthetic analysis using AiZynthFinder, and labelled as solved or unsolved. The time limit to search for retrosynthetic routes was set as 3 minutes per target compound, with a maximum of seven steps, a maximum of two hundred iterations, and expansion of fifty actions at each stage of the search as determined by the policy network up to a cumulative cutoff threshold of 0.995.

**Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility**

Estimation of retrosynthetic accessibility (RA) was framed as a binary classification problem. We trained a series of classifiers on the retrosynthetic predictions of AiZynthFinder using the label generation method stated previously. The trained classifier predicts whether or not a given compound is synthetically accessible as found by CASP.
We examined the following classification algorithms: a) a feed forward neural network classifier, b) XGBoost classifier, and c) random forest classifier. For each algorithm 2048 dimensional counted extended connectivity fingerprints were used with a radius set to 3 (ECFP6), and ECFP6 counts with features as generated by RDKit. In total six different models were trained for each dataset, ChEMBL, GDBChEMBL, and GDBMedChem. SAcore, SCcore, and SYBA are continuous scores for complexity, thus we trained a classifier for each score for comparative purposes, where the score was used as the sole descriptor. For the score-based classifiers we used a feed forward neural network and logistic regression. The scores used as descriptors were calculated using RDKit and the models published by the authors of the corresponding publications.

Scikit-Learn was used to train the random forest model, XGBoost for the XGB classifier, and Keras with Tensorflow for the feed forward neural networks. In each case the models where wrapped within an objective function using the Optuna framework for hyperparameter optimisation. All models with the exception of the feed neural network were optimized using a five-fold cross validation. The framework used to train the classifiers and models are available at https://github.com/reymond-group/RAscore, and can be used for any binary classification problem.

Each model was optimized with the Optuna hyperparameter optimization framework to find the optimal parameter set. In the case of the feed forward neural network, we treated the number of layers, the size of the layers, the activation function, the dropout rate, and the learning rate as hyperparameters, to find the optimal architecture within the bounds of the starting criterion as given in the supporting information.

There was no overlap of compounds between training, validation, and test sets. This was determined by computing the InChI-keys of the compounds in the two sets and using the Python built-in set methods to find the intersection. We did not check whether a compound was present in the training data used to train AiZynthFinder, however this is likely not to influence the performance of multi-step retrosynthesis, as the reaction datasets only consider single steps and is supported by our previous studies.
Average Linkage as a Method for Evaluating Machine Learning Based Classifiers

We assessed model performance by computing how well solved and unsolved routes are separated using the concept of average linkage. Average linkage is a statistical method by which the distance between two clusters are treated as the average distance between all pairs of items, where each member of the pair belongs to one of the two clusters. In this instance, the two clusters are solved and unsolved compounds as determined by AiZynthFinder. The average linkage or separation between solved and unsolved compounds was determined by min-max scaling the values of each score such that they were normalized between 1 and 0 using the Scikit-Learn MinMaxScaler. The absolute pairwise distances where computed, and the average of the distances taken to yield a value that corresponds to the separation of the clusters as shown in (Figure 2).

![Diagram of average linkage]

\[
\text{average linkage} = \frac{1}{n_{\text{solved}} n_{\text{unsolved}}} \sum_{i=1}^{n_{\text{solved}}} \sum_{j=1}^{n_{\text{unsolved}}} D(x_{\text{solved},i}, x_{\text{unsolved},j})
\]

**Figure 2:** Illustrates the computation of the average linkage. The average linkage is a method by which the distance between two clusters are treated as the average distance between all pairs of items, where one member of the pair belongs to each cluster.
Results and Discussion

Route Statistics from the Generation of Labels for Machine Learning Classifiers

Initially training and test datasets were generated by randomly sampling 200,000 compounds from ChEMBL, and 100,000 compounds each from GDBChEMBL and GDBMedChem. The two are subsets of the GDB17 database. ChEMBL was chosen to represent a selection of bioactive molecules and the GDB subsets chosen to be more challenging owing to their differing structural and physiochemical property distribution. The compounds were subsequently subjected to retrosynthetic analysis using AiZynthFinder, and labelled as solved or unsolved.

Figure 3: Statistics gathered for the retrosynthesis predicted during the label generation process for each dataset ChEMBL, GDBChEMBL, and GDBMedChem. The statistics are shown for all compounds sampled: 200,000 from ChEMBL, and 100,000 from each of the GDB subsets. a) The percentage of solved compounds as a function of the number of steps, b) The dataset coverage as a function of the number of steps. c-f) Histograms depicting the distribution of the compounds in each dataset for each of the currently used scores. For SAscore and SCscore the lower the score the less complex and easier to synthesise a given compounds, whereas for SYBA positive values indicate easy to synthesise compounds and negative values hard to synthesise. g-h) The percentage of solved compounds as a function of each of the currently used score as computed for each bin in the histogram.

Figure 3 shows statistics gathered for the predicted retrosynthetic routes during the label generation process. The percentage of solved routes increases monotonically, and the rate at which routes are solved decreases with the number of steps for each dataset. This is most
noticeable for compounds requiring synthetic routes between 5 and 7 steps, where we observe a significant increase in the dataset coverage (Figure 3b), but no corresponding increase in the percentage of solved compounds (Figure 3a). ChEMBL has the highest percentage of solved compounds, whereas GDBMedChem and GDBChEMBL are consistently lower.

We observed a correlation between the percentage of solved compounds and the SAscore,\textsuperscript{17} SCscore,\textsuperscript{15} and SYBA,\textsuperscript{16} as well as SMILES length (Figure 3g-j), which are in agreement with the results obtained by Coley and Gao.\textsuperscript{11} In the case of SAscore and SCscore, the lower the score the more likely it is that a synthetic route can be obtained for a compound, as found for all datasets. The ChEMBL sample exhibits a lower range of SAscore than the GDBMedChem and GDBChEMBL samples (Figure 3c), which may explain the higher percentage of solved compounds in ChEMBL as compared to GDBMedChem and GDBChEMBL (Figure 3a).

However, for SCscore (Figure 3d) the GDB subsets exhibit a lower range of scores in comparison to the ChEMBL sample. This the inverse of the distribution we obtain for SAscore and can be rationalized by considering the assumptions made in the SCscore model. The SCscore is based on reactions rather than molecular fragments and assumes that the products of a reaction are more complex than the reactants. In most cases the products are also larger than the reactants, thus the assumption for SCscore falters for the GDB subsets because of their restricted size as shown by the difference in SMILES length between the ChEMBL and GDB subsets (Figure 3f and Figure 3j). This is further supported by the lower percentage of solved routes for the GDB subsets (Figure 3f and Figure 3j).

In the case of SYBA, the higher the score the more likely it is that a route can be found, negative values indicate hard to synthesize compounds. The distribution shown for SMILES length reflects the fact that the GDB subsets are skewed towards smaller molecules, and with lower heavy atom counts than those found in ChEMBL by virtue of the rules used in their enumeration.\textsuperscript{25, 26} The Figure 3j reveals that the rate at which compounds can be solved falls off much more rapidly with SMILES length for the GDB subsets than for ChEMBL.
Attempts at using SAscore, SCscore and SYBA

**Figure 4**: Histograms computed for the test set of ca. 20,000 ChEMBL compounds showing whether a retrosynthetic route could be found by CASP for a given compound (green) or not (red), and their distributions across each of the scores in current use. There is no threshold value at which the current scores are able to separate compounds that can be solved by CASP (green) from those that cannot (red). This highlights how the scores have potential for misuse in generative modelling and filtering sets of compounds.

We assessed the existing scores SAscore, SCscore, and SYBA for their ability to distinguish between compounds that could be solved by CASP and those that could not (Figure 4). These scores have often been used to filter or estimate the synthetic accessibility of large datasets of virtual compounds.\textsuperscript{34-36} However, we have found that there is no threshold value at which the scores can be set that clearly separates compounds that can and cannot be solved by CASP, as shown by the overlapping histograms. This was observed for all datasets examined in this study (refer to SI). Thus, there is potential for them to be misused when filtering large virtual libraries. To resolve this issue, we propose that the existing scores be used alongside the classifiers trained in this study to determine whether a synthetic route can be found, and how difficult it may be to realize the route in the wet lab.

**Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility**

The overlaps shown in Figure 4, demonstrate the need to be able to differentiate between compounds that can and cannot be synthesized. Therefore, we trained a series of ML based classifiers to determine whether a given compound could be solved by CASP. A selection of the results obtained for the trained classifiers are shown in Error! Reference source not found.
In each case the classifiers outperform the existing scores which were used as a baseline (SAscore, SCscore, and SYBA) both in terms of the AUC and average linkage with respect to their ability to classify compounds as solved or unsolved. When using the existing scores as descriptors to train the classifiers, we observed a marginal improvement in comparison to the score itself. A more significant improvement in classifier performance was obtained when using ECFP6 counted vectors as molecular descriptors, both with and without features. The feed forward neural network (NN) based models consistently outperformed random forest and showed comparable performance to gradient boosting methods (XGB).

We identified that the following classifiers were consistently the top three models across each of the datasets: feed forward neural networks using ECFP6 counts, feed forward neural networks using ECFP6 counts with features, and XGBoost using ECFP6 counts. For the RA score we chose the top performing classifier for separating the compounds as determined by the average linkage. We also identified a GDB specific classifier which we term GDBscore in the same manner. The GDBscore classifier was trained on the GDBChEMBL dataset, the classifier trained on GDBMedChem was found to have equivalent performance.

**Table 1:** Outlines the top 3 classifiers trained for each dataset alongside their corresponding metrics.a)

| Dataset       | Model            | Descriptor                                      | ROC-AUC | Accuracy | Precision | Recall | Average Linkage |
|---------------|------------------|-------------------------------------------------|---------|----------|-----------|--------|-----------------|
| ChEMBL        | NN (RAscore)     | ECFP6 counts with features                      | 0.93    | 0.90     | 0.92      | 0.95   | 0.69            |
|               | NN               | ECFP6 counts                                    | 0.94    | 0.90     | 0.92      | 0.95   | 0.68            |
|               | XGB              | ECFP6 counts                                    | 0.95    | 0.91     | 0.92      | 0.96   | 0.65            |
|               | NN               | SAscore                                         | 0.85    | 0.81     | 0.84      | 0.92   | 0.37            |
|               | NN               | SCscore                                         | 0.61    | 0.75     | 0.61      | 1.00   | 0.27            |
|               | NN               | SYBA score                                      | 0.74    | 0.78     | 0.78      | 0.97   | 0.21            |
|               | Baseline         | SAscore                                         | 0.15    | -        | -         | -      | 0.17            |
|               | Baseline         | SCscore                                         | 0.39    | -        | -         | -      | 0.22            |
|               | Baseline         | SYBA                                            | 0.74    | -        | -         | -      | 0.17            |
| GDBChEMBL     | NN (GDBscore)    | ECFP6 counts                                    | 0.93    | 0.87     | 0.76      | 0.73   | 0.64            |
|               | NN               | ECFP6 counts with features                      | 0.94    | 0.88     | 0.78      | 0.74   | 0.63            |
|               | XGB              | ECFP6 counts                                    | 0.94    | 0.89     | 0.81      | 0.73   | 0.61            |
|               | Baseline         | SAscore                                         | 0.11    | -        | -         | -      | 0.26            |
|               | Baseline         | SCscore                                         | 0.38    | -        | -         | -      | 0.14            |
|               | Baseline         | SYBA                                            | 0.72    | -        | -         | -      | 0.17            |
| GDBMedChem    | NN               | ECFP6 counts                                    | 0.93    | 0.88     | 0.75      | 0.64   | 0.64            |
|               | NN               | ECFP6 counts with features                      | 0.94    | 0.89     | 0.77      | 0.66   | 0.63            |
|               | XGB              | ECFP6 counts                                    | 0.94    | 0.89     | 0.78      | 0.64   | 0.61            |
|               | Baseline         | SAscore                                         | 0.13    | -        | -         | -      | 0.22            |
|               | Baseline         | SCscore                                         | 0.39    | -        | -         | -      | 0.14            |
|               | Baseline         | SYBA                                            | 0.70    | -        | -         | -      | 0.17            |
Prediction Time
The importance of training ML based classifiers rather than simply predicting the full retrosynthetic pathway becomes clear when examining Error! Reference source not found.. Full retrosynthetic route prediction of the ChEMBL sample of 200,000 compounds to a set of commercially available building blocks took approximately 239 CPU days on a single machine with 8 CPUs and 64 GB of RAM. Parallelization is not possible on a single machine under the current implementation, however, it is possible to split the compounds over several cores and distribute the workload over several machines as has been done in this study. In comparison 77 minutes were required for classifying retrosynthetic accessibility using RAscore. The increase in prediction speed by ca. 4,500 times opens up the possibility of estimating the retrosynthetic accessibility of virtual compounds, for instance in drug discovery projects, and for the scoring of compounds resulting from generative models earlier in the virtual screening workflow. Similar increases in prediction time are also observed for the GDB subsets (Table 1).

Table 2: Percentage of solved compounds for each dataset and the run time\(^a\) required using AiZynthFinder.

|                          | ChEMBL | GDBChEMBL | GDBMedChem |
|--------------------------|--------|-----------|------------|
| Percentage Solved        | 75.21  | 25.54     | 20.79      |
| Size                     | 200,000| 100,000   | 100,000    |
| AiZynthFinder Run Time (days) | 239    | 149       | 151        |
| Score Run Time (mins)    | 79\(^b\) | 30\(^c\)  | 30\(^c\)   |

\(^a\) Expressed in days taken on a single machine with 8 CPUs and 64 GB of RAM, rounded to the nearest day. The time taken in minutes for the neural network classifier with ECFP6 counted fingerprints is also given for comparative purposes. The neural network classifier, RAAscore, is able to reproduce the results obtained from AiZynthFinder in a fraction of the time taken to predict full retrosynthetic routes.

\(^b\)RAAscore

\(^c\)GDBscore
Gao and Coley previously published the results of running retrosynthetic analysis with ASKCOS for a series of datasets consisting of both published and generated compounds. We tested our trained classifier on the dataset used by Gao and Coley to determine the applicability domain of the classifier and gauge how well the ASKCOS predictions could be reproduced. We also used AiZynthFinder to predict retrosynthetic routes to the same set of compounds to establish whether the classifier could reproduce the underlying CASP tool.

For each dataset AiZynthFinder marginally outperforms ASKCOS, and is most striking for the GDB17 sample (Figure 5). This is because AiZynthFinder only considers retrosynthetic analysis, whereas ASKCOS additionally factors in reaction prediction which enables pruning of unfeasible or low probability retrosynthetic pathways. Furthermore, as the reaction prediction models are trained on published chemistry, and the majority of GDB17 compounds are unpublished or dissimilar to published compounds, the pathways suggested are likely to be pruned resulting in the lower percentage of solved compounds for ASKCOS. Another difference that should be considered when comparing the two models is the building blocks available to
each respective model. This can affect the ability of the CASP tool to find retrosynthetic routes and influences whether or not a compound is labelled as solved.

We found that the feed forward neural network classifier trained on ChEMBL that we term RAscore, overestimates the synthetic accessibility of GDB17 in comparison to ASKCOS and AiZynthFinder. This is also observed for the other datasets examined, however the extent to which RAscore overpredicts is less striking. To replicate the GDB17 dataset, we use GDBscore, which is a classifier trained on GDBChEMBL and find we can better reproduce the underlying AiZynthFinder synthesis planning tool. The MOSES dataset is based on the ZINC Clean Leads collection and GaucaMol is based on the ChEMBL database, both are used for evaluating distribution learning algorithms for drug discovery. The overprediction on both ZINC and the prediction in line with the MOSES dataset is surprising considering the compounds originate from the same database. However, this may be rationalized considering the samples differ in their distribution, and have been obtained from different collections within the ZINC database.

The overprediction on the Sheridan et. al. dataset can be seen as positive as all compounds in the dataset were previously synthesized at Merck. In addition, the prediction in line with the GuacaMol set, implies that the classifier performs well on ChEMBL like compounds by virtue of the underlying training data.

Examples – Limitations of RAscore arising from CASP

We examined the test set from our ChEMBL sample for compounds within a Tanimoto similarity of 0.8 or greater. Some examples of pairs of compounds are shown in Figure 6. In the pairs shown one compound was unsolved by our retrosynthetic tool and the other labelled as solved. For each example we show that the topology is largely unchanged and only small edits have been made to the functionality of the molecule. The change in outcome with minor changes in functionality highlight a limitation of our CASP tool and likely other template based CASP tools. The templates suggested for disconnections are unable to account for subtle changes in the reaction center, thus the appropriate precursors were not able to be enumerated. This arises because the molecular graph underlying the template does not match that of the substrate, thus there is no substructure match. These examples are not ‘true’ negatives in the sense that they cannot be experimentally realised in the wet-lab and are only negative in relation to the ability of
the CASP tool to conduct a retrosynthetic analysis. Some examples of such compounds which have led to poor separation of solved/unsolved compounds are shown in Figure 6.

**Figure 6:** Examples of pairs of compounds from the test set that are similar to each other (Tanimoto > 0.8), where a retrosynthetic route could be found for one example in the pair but not the other. In each case only a slight modification of the compound leads to a change in the outcome from the CASP tool, consider a) changing a methyl ether to an ethyl ether, b) addition of two ortho-methyl groups on the terminal phenyl ring, c) substitution of a cyclohexane moiety for a cyclopropane, and d) a change in substitution pattern and ring morphology, leads to a change in outcome from solved to unsolved.

To understand why the solved/unsolved test cases were not easily separable, consider the examples in Figure 6. In the case of similar compounds, both solved and unsolved compounds are scored as synthetically feasible with values tending towards 1.0, despite our CASP tool not having found a synthetic route. The example in Figure 6a, is a case for which the methyl ether can be solved while the ethyl ether is not solved. The RAscore learns that such minor changes to functionality are feasible, thus assigns a score in line with what would be expected by synthetic chemists. This is an artefact of the subset of compounds from ChEMBL the model was trained on, and examples in which the model misclassifies compounds as synthesizable can also be found. Similar substitutions are shown in Figure 6b and Figure 6c, whereby the algorithm failed to suggest retrosynthetic disconnections leading to commercially building blocks.

In most cases the most similar molecule in the training set was below a Tanimoto value of 0.8 (SI), and potentially requires a different synthetic strategy as compared to the compounds shown for the test set (Figure 6). This raises another limitation of CASP tools, which can be overcome
by RAscore. The performance of the CASP tool is limited by the number and type of building blocks available. In some cases it may be that the building blocks necessary are not included in the database underlying the CASP tool, but are in fact available from other vendors. Furthermore, it can also be the case that similar building blocks are available that a medicinal chemist may consider for functionalization. In these cases the RAscore is able to learn that it is likely that two analogues are synthetically accessible despite a retrosynthetic route having not been found.

To exemplify the aforementioned arguments consider the predicted routes shown in Figure 7. If we again take the case of the methyl and ethyl ethers of the compound shown in Figure 6a, and examine the routes predicted for each, Figure 7a and Figure 7b respectively, we observe large differences in the predicted route in terms of the synthetic strategy used, thus step count. The methyl ether (Figure 7a) can be synthesised in one step from commercially available building blocks. In this case the unprotected compound is available from commercial vendors, and no further syntheses are suggested. In the case of the ethyl ether, the algorithm proposes a different starting disconnection to form the piperazine ring through substitution and subsequent cyclisation, rather than Boc-protection. This leads to a synthetic route whereby none of the resulting intermediates are commercially available.

Thus, similar compounds with largely unchanged topology can have considerably different synthetic routes predicted for them, and as each route prediction is treated independently the algorithm does not learn that similar compounds have the potential to be synthesized via similar routes. Whilst a chemist may consider first synthesizing the scaffold, and subsequently functionalising it to yield the desired analogues, our CASP tool is currently unable to take into account such considerations. This is further exemplified in Figure 7c and Figure 7d, whereby different synthetic routes necessitate different starting materials. The synthetic route proposed in Figure 7d can be used to synthesize the compound in Figure 7c with only a slight variation in how the building blocks are initially synthesized.
Figure 7: Synthetic routes predicted by AiZynthFinder. We show that small variations in the queried compound lead to considerably different synthetic routes, thus require different building blocks. This is a limitation of CASP tools that models such as RAscore may be able to overcome.

Conclusions

Herein we have built on the improvements in AI driven CASP in recent years by combining the predictions made with our CASP tool, AiZynthFinder, with ML, to train a classifier returning a retrosynthetic accessibility score (RAscore). RAscore addresses the challenge of classifying compounds as synthetically feasible and is orders of magnitude faster than full retrosynthetic analysis by CASP, and with comparable performance. The RAscore demonstrates potential for rapid pre-screening of compounds for synthetic accessibility, enabling enrichment of synthetically feasible chemical space. Whereas previous synthetic accessibility and complexity based scores have potential for misuse when filtering large virtual libraries, as a result of being unable to determine a threshold value (Figure 4), we resolve this issue by proposing that the existing scores be used alongside the RAscore to determine whether a synthetic route can be found, and how difficult it may be to realize the route in the wet lab.

In addition, we highlight inherent limitations to be aware of in the RAscore arising from the performance and applicability of the underlying CASP tool, namely: 1) availability of building blocks, 2) different synthetic strategies toward the same scaffold, and 3) route predictions are treated independently to each other. The concept presented herein can be extended to any CASP tool and the predictions it generates. The score will be made available under an MIT license at: https://github.com/reymond-group/RAscore

Declarations

Availability of Data and Materials

The score will be made available under an MIT license at, as well as instructions on how to access the datasets and the framework for training the classifiers: https://github.com/reymond-group/RAscore.

AiZynthFinder is open source and is available under an MIT license at: https://github.com/MolecularAI/AiZynthFinder
The dataset used to assess the applicability domain can be found at: https://github.com/wenhao-gao/askcos_synthesizability/tree/master/results/dataset.csv

**Competing Interests**
The authors declare that they have no competing interests.

**Funding**
A. Thakkar received support from AstraZeneca, the Swiss National Science Foundation (SNF), and the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 676434, “Big Data in Chemistry” (“BIGCHEM,” http://bigchem.eu). V. Chadimová was sponsored by the AstraZeneca R&D graduate programme.

**Author contributions**
A. Thakkar designed, conducted the research, and wrote the manuscript. V. Chadimová performed preliminary modelling studies. E.J. Bjerrum contributed ideas to the project. O. Engkvist, and J.L. Reymond supervised the project and assisted in writing the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**
The authors would like to thank NextMove software for providing access to NameRxn for atom-mapping. Samuel Genheden for valuable feedback and discussions regarding the manuscript and analysis. The Molecular AI group at AstraZeneca and Reymond group at the University of Bern for their support.
References

1. Segler, M. H. S.; Koge, T.; Tyrchan, C.; Waller, M. P., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science 2018, 4 (1), 120-131.
2. Thomas, B.; Josep, A.-P.; Hongming, C.; Christian, M.; Christian, T.; Ola, E.; Kostas, P.; Atanas, P., Reinvent 2.0 – an AI Tool for De Novo Drug Design. ChemRxiv 2020.
3. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S., Deeptox: Toxicity Prediction Using Deep Learning. Frontiers in Environmental Science 2016, 3, 80.
4. Segler, M. H. S.; Preuss, M.; Waller, M. P., Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. Nature 2018, 555, 604.
5. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F., A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. Science 2019, 365 (6453), eaax1566.
6. Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; Desjarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F., Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. Journal of Medicinal Chemistry 2020.
7. Walters, W. P., Virtual Chemical Libraries. Journal of Medicinal Chemistry 2019, 62 (3), 1116-1124.
8. Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutchkov, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G., Rethinking Drug Design in the Artificial Intelligence Era. Nature Reviews Drug Discovery 2020, 19 (5), 353-364.
9. Chevillard, F.; Kolb, P., Scubidoo: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. Journal of Chemical Information and Modeling 2015, 55 (9), 1824-1835.
10. Zhavoronkov, A.; Ivanenko, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsova, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A., Deep Learning Enables Rapid Identification of Potent Ddr1 Kinase Inhibitors. Nature Biotechnology 2019, 37 (9), 1038-1040.
11. Gao, W.; Coley, C. W., The Synthesizability of Molecules Proposed by Generative Models. Journal of Chemical Information and Modeling 2020.
12. Ihlenfeldt, W.-D.; Gasteiger, J., Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. Angewandte Chemie International Edition in English 1996, 34 (23-24), 2613-2633.
13. Corey, E.; Long, A.; Rubenstein, S., Computer-Assisted Analysis in Organic Synthesis. Science 1985, 228 (4698), 408-418.
14. Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A., Computer-Aided Synthesis Design: 40 Years On. Wiley Interdisciplinary Reviews: Computational Molecular Science 2012, 2 (1), 79-107.
15. Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F., Scscore: Synthetic Complexity Learned from a Reaction Corpus. Journal of Chemical Information and Modeling 2018, 58 (2), 252-261.
16. Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D., Syba: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. Journal of Cheminformatics 2020, 12 (1), 35.
17. Ertl, P.; Schuffenhauer, A., Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *Journal of Cheminformatics* **2009**, *1* (1), 8.
18. Genheden, S.; Thakkar, A.; Chadimova, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J., Aizynthfinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. *ChemRxiv* 2020.
19. Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J., Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chemical Science* **2020**, *11* (1), 154-168.
20. Lowe, D. Chemical Reactions from Us Patents (1976-Sep2016). [https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873](https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873) (accessed Apr 31, 2018).
21. [https://www.acdlabs.com/index.php](https://www.acdlabs.com/index.php).
22. Enamine. [https://enamine.net/building-blocks](https://enamine.net/building-blocks).
23. Zinc. [http://zinc.docking.org/rings/subsets/](http://zinc.docking.org/rings/subsets/) (accessed Aug 27, 2019).
24. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutuwo, P.; Atkinson, F.; Bellis, L. J.; Cibirián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magaríños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The Chembl Database in 2017. *Nucleic acids research* **2017**, *45* (D1), D945-D954.
25. Bühlmann, S.; Reymond, J.-L., Chembl-Likeness Score and Database Gdbchembl. *Frontiers in Chemistry* **2020**, *8* (46).
26. Awale, M.; Sirockin, F.; Stiefl, N.; Reymond, J.-L., Medicinal Chemistry Aware Database Gdbmedchem. 2019, 38 (8-9), 1900031.
27. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742-754.
28. Rdkit: Open-Source Cheminformatics; [Http://Www.Rdkit.Org](http://Www.Rdkit.Org).
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E., Scikit-Learn: Machine Learning in {P}Ython. *Journal of Machine Learning Research* **2011**, *12*, 2825-2830.
30. Abadi, M.; Agarwal, A.; Barham, P.; Eugene Brevdo; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; 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. Software available from tensorflow.org.
31. [https://www.vsc.com](https://www.vsc.com).
32. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, Association for Computing Machinery: Anchorage, AK, USA, 2019; pp 2623–2631.
33. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D., Inchi, the Iupac International Chemical Identifier. *Journal of Cheminformatics* **2015**, *7* (1), 23.
34. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database Gdb-17. *Journal of Chemical Information and Modeling* **2012**, *52* (11), 2864-2875.
35. 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 Central Science 2018, 4 (2), 268-276.
36. Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J., De Novo Generation of Hit-Like Molecules from Gene Expression Signatures Using Artificial Intelligence. Nature Communications 2020, 11 (1), 10.
37. Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C., Guacamol: Benchmarking Models for De Novo Molecular Design. Journal of Chemical Information and Modeling 2019, 59 (3), 1096-1108.
38. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. J. a. e.-p. Molecular Sets (Moses): A Benchmarking Platform for Molecular Generation Models 2018, p. arXiv:1811.12823. https://ui.adsabs.harvard.edu/abs/2018arXiv181112823P (accessed November 01, 2018).
39. Irwin, J. J.; Sterling, T.; Mindsinger, M. M.; Bolstad, E. S.; Coleman, R. G., Zinc: A Free Tool to Discover Chemistry for Biology. Journal of Chemical Information and Modeling 2012, 52 (7), 1757-1768.
40. Sheridan, R. P.; Zorn, N.; Sherez, E. C.; Campeau, L.-C.; Chang, C.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O’She, P. D., Modeling a Crowdsourced Definition of Molecular Complexity. Journal of Chemical Information and Modeling 2014, 54 (6), 1604-1616.

Table of Contents Graphic