# CoPriNet: Deep learning compound price prediction for use in de novo molecule generation and prioritization.

Ruben Sanchez-Garcia[1,5,*], Dávid Havasi[2,3], Gergely Takács[2,3], Matthew C. Robinson[8], Alpha Lee[4,8], Frank von Delft[5,6,7,*], Charlotte M. Deane[1,*]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, U.K.

[2] Mcule.com Kft, Bartók Béla út 105-113, Budapest, 1115, Hungary.

[3] Department of Chemical and Environmental Process Engineering, Budapest University of Technology and Economics, Műegyetem rakpart 3, Budapest, 1111, Hungary.

[4] Department of Physics, University of Cambridge, Cambridge CB3 0HE, United Kingdom

[5] Structural Genomics Consortium (SGC), University of Oxford, Oxford OX3 7DQ, U.K.

[6] Diamond Light Source, Didcot OX11 0DE, U.K.

[7] Department of Biochemistry, University of Johannesburg, Johannesburg 2006, South Africa

[8] PostEra Inc., 1209 Orange Street, Wilmington, Delaware 19801, United States.

*Corresponding authors: R. S.-G. (ruben.sanchez-garcia@stats.ox.ac.uk); F. D. (frank.vondelft@cmd.ox.ac.uk); C. M. D. (deane@stats.ox.ac.uk)

## Abstract

Compound availability is a critical property for design prioritization across the drug discovery pipeline. Historically, and despite its multiple limitations, compound-oriented synthetic accessibility scores have been used as proxies for this problem. However, the size of the catalogues of commercially available molecules has dramatically increased over the last decade, redefining the problem of compound accessibility as a matter of budget. In this paper we show that if compound prices are an alternative proxy for compound availability, then synthetic accessibility scores are not effective strategies for assessing availability. Instead, we learn how to predict prices directly from the catalogues. Our approached, CopriNet, is a retrosynthesis-free deep learning model trained on pairs of compound/prices extracted from the Mcule catalogue. CoPriNet is able to provide price predictions that exhibit far better correlation with actual compound prices than any synthetic accessibility measurement. Moreover, unlike standard retrosynthesis methods, CoPriNet is rapid, comparable in execution time to popular synthetic accessibility metrics and thus is suitable for high-throughput experiments including virtual screening and de novo compound generation.

## Introduction

The drug design process can be thought of as a multi-objective optimization problem in which potential drug compounds need to satisfy a wide set of properties from binding affinity and toxicity to selectivity and solubility (Nicolaou and Brown, 2013). One property that is key when developing potential drug molecules is their availability, since no matter how promising a design might be, if it is not available, it is doomed to fail.

In order to estimate compound availability, several computationally calculated synthetic accessibility (SA) scores have been developed. These approaches can be roughly classified as retrosynthesis-based predictions (Coley *et al.*, 2019; Genheden *et al.*, 2020; Gillet *et al.*, 1995; Huang *et al.*, 2011; Ihlenfeldt and Gasteiger, 1996), binary classifiers (Podolyan et al., 2010; Amol Thakkar et al., 2021;

Voršilák et al., 2020), and complexity-based estimations (Ertl and Schuffenhauer, 2009; Coley *et al.*, 2018; Allu and Oprea, 2005; Barone and Chanon, 2001).

Retrosynthesis-based approaches aim to identify suitable synthetic routes for a given molecule using distinct types of search algorithms over databases of building blocks and chemical transformations. State-of-the-art methods (Coley *et al.*, 2019; Genheden *et al.*, 2020; Schwaller *et al.*, 2019a; Dai *et al.*, 2020), which are based on deep learning, are able to integrate information from millions of reactions and building blocks, suggesting feasible synthetic routes for the majority of the benchmarked compounds in a matter of seconds to minutes (Genheden *et al.*, 2020). However, their outputs strongly depend on the employed databases (Voršilák and Svozil, 2017) and they tend to suggest multiple solutions which are difficult to rank (Yiming Mo et al., 2021) and more importantly, even the fastest are computationally demanding and therefore ill-suited for high-throughput computational pipelines (Amol Thakkar et al., 2021).

Binary classifiers are machine learning algorithms trained to distinguish between compounds that are easy or difficult to make. Although the available approaches may differ in terms of learning algorithms (support vector machine, neural network, etc) and compound featurization (descriptors, fingerprints, etc.), it is the definition of the training set, consisting of compounds labelled as easy or difficult to make, that most impacts the behaviour of the methods. Some strategies for compiling training datasets include retrosynthesis (Podolyan et al., 2010; Amol Thakkar et al., 2021), presence in commercial catalogues or virtually edited compounds (Voršilák *et al.*, 2020). Although binary classifiers tend to be much faster than retrosynthesis-based methods, they are also less accurate (Voršilák *et al.*, 2020) and their performance is highly dependent on the training dataset (Amol Thakkar et al., 2021). Binary classifiers also by definition cannot distinguish between different levels of difficulty (Voršilák et al., 2020; Amol Thakkar et al., 2021).

Complexity-based methods try to define an empirical metric under the assumption that complex molecules are more difficult to synthesize (Boda *et al.*, 2007; Hendrickson *et al.*, 1987; Barone and Chanon, 2001). Most methods define complexity as a function of the presence of features deemed to be complex or infrequent such as chiral centres, uncommon moieties, or unusual molecular fragments. One of the most popular measures of SA (Omolabi *et al.*, 2021; Basu *et al.*, 2020; Lu and Li, 2021; Imrie *et al.*, 2021a; Humbeck *et al.*, 2018), SAscore (Ertl and Schuffenhauer, 2009) is a complexity-based method that uses the rarity of fragments found in PubChem (Kim *et al.*, 2016) and a set of predefined properties such as the ring complexity or the number of stereo centres to calculate its score. Another commonly used SA score, SCScore (Coley *et al.*, 2018) employs an indirect estimation of complexity assuming that the complexity of the reactants is never larger than the complexity of the products. Such estimation is obtained using a neural network trained on pairs of products/reactants and is able to capture some changes in complexity that reactions cause.

Due primarily to their simplicity and speed, SAscore and SCScore have been used extensively across drug development pipelines including for compound screening (e.g., Omolabi *et al.*, 2021; Basu *et al.*, 2020; Lu and Li, 2021; Huang *et al.*, 2019), dataset preparation (e.g., Imrie *et al.*, 2021b; Humbeck *et al.*, 2018) and molecule generation/optimization (e.g., Leguy *et al.*, 2020; Zhou *et al.*, 2019; Khemchandani *et al.*, 2020a; Green *et al.*, 2020). SAScore is one of the most popular metrics for biasing or discarding potentially infeasible compounds in methods for computational generation of *de novo* molecules (e.g., Yassine *et al.*, 2021; Imrie *et al.*, 2020; Prykhodko *et al.*, 2019; Leguy *et al.*, 2020; Khemchandani *et al.*, 2020b). However, as described above, SAscore and SCscore are simple approximations for SA and as such, present several limitations. For instance, it is well known that these scores tend to underestimate the SA of difficult compounds that can be synthesized from complex commercially available building blocks (Gao and Coley, 2020; Makara *et al.*, 2021). It has

also been shown that structurally similar compounds, which also tend to have similar complexity-based scores, can require synthetic strategies of different difficulty levels (Gao and Coley, 2020), leading to incorrect SA estimations.

Independent of their nature, the aim of all the methods described above is to computationally filter compounds, ruling out those difficult to make or purchase. This suggest that many users of SA metrics would benefit from a direct estimation of an alternative metric strongly related to compound availability and purchasability: the price of the compounds. The price is a metric of undeniable importance, influencing many of the decisions taken during the drug discovery pipeline, particularly in the early stages, where the cost of the compounds to be experimentally tested is often of central importance.

Current SA measurements exhibit poor correlation with prices, Fukunishi et al. (Fukunishi *et al.*, 2014) found that the Pearson correlation coefficient (PCC) between their SA measurement and the logarithmic sales prices of the compound, in $/mmol, was ~0.3. We have expanded on this work and found that none of four evaluated SA metrics correlates with price. This is perhaps not surprising, since SA scores were never intended to capture price information. However, most methods for automatic compound design try to optimize their molecules against a SA metric, which will lead to the method suggesting many potentially feasible yet prohibitively expensive compounds. While for the hundreds of millions of compounds that are available in the commercial catalogues, price estimation translates to a database search problem, in the age of machine learning molecular generation techniques, many novel drug-like compounds that are not in catalogues can and will be proposed. Consequently, estimating the price of non-catalogue compounds is a problem that will arise.

Compound cost prediction has previously been addressed using retrosynthesis-based methods (Badowski *et al.*, 2019). In their approach, Badowski et al. estimated the cost of a molecule as the cost of the least expensive synthetic route. The cost of each route is computed as the summation of the cost of the initial reactants and the cost of the reactions required to synthesize the molecule, which is defined recursively as the sum of fixed costs associated to the reaction and the cost of each of the reactants corrected by the yield of the reaction. While this formulation captures the different terms involved in the final price, it presents multiple drawbacks. First, since it is based on retrosynthesis, the method is slow to compute. Second, it relies on estimations of reaction yields and fixed costs, information that is only available for a limited number of reactions and that, in many cases, is not in the public domain. Lastly, the assumption that the cheapest retrosynthetic route is the one that determines the final cost does not necessary hold since multiple factors such as reaction success chances are not considered.

With the aim of overcoming these problems, in this manuscript we present a retrosynthesis-free method to obtain price predictions using only the compound itself. Our method is based on a graph neural network (GNN) trained on a dataset of molecule/price pairs collected from the Mcule (Kiss *et al.*, 2012) catalogue (https://mcule.com/). Our approach follows that of SA binary classifiers trained on retrosynthesis predictions: given enough data, machine learning methods should identify patterns in the input molecules that are relevant for the synthetic planning (or the price) without the need to explicitly undergo retrosynthetic decomposition. Although retrosynthesis-based computations tend to be more accurate, our predictions exhibit a far stronger correlation with catalogue prices than any SA metric, with comparable running times to popular SA estimations. Consequently, our approach can be employed as a complementary metric to fast SA estimations for high throughput assays and more importantly, for *de novo* molecule generation, in which the large number of required assessments prevents retrosynthetic-based approaches from being used.

# Results and discussion

## Limitations of synthetic accessibility approximations

We examined the behaviour of four current SA scores, SAscore (Ertl and Schuffenhauer, 2009), SCScore (Coley *et al.*, 2018), SYBA (Voršilák *et al.*, 2020), and RAscore (Amol Thakkar et al., 2021), on two set of molecules, a dataset of purchasable compounds (PC) and a dataset of non-purchasable natural products (NPNP) (Figure 1, a-e). As a first approximation all the PC molecules should be classified as synthetically feasible, and most of them as highly accessible, whereas most of the NPNP compounds should be considered hard to synthesize. As Figure 1 (a-d) shows, none of the methods perfectly separate the two compound sets. However, this is not surprising nor potentially even desired since not all NPNP are synthetically infeasible, nor all purchasable compounds are easy to make. In order to get a better estimation of the actual number of synthetically feasible NPNP compounds we computed ManifoldSA, a pure retrosynthesis-based score (see Methods for more details). This score estimates that ~24% of the NPNP are synthesizable, with 4.6% of those being easily synthesizable (Figure 1 e). Whereas for the PC dataset, ~6% of the compounds were regarded as infeasible despite being commercially available.

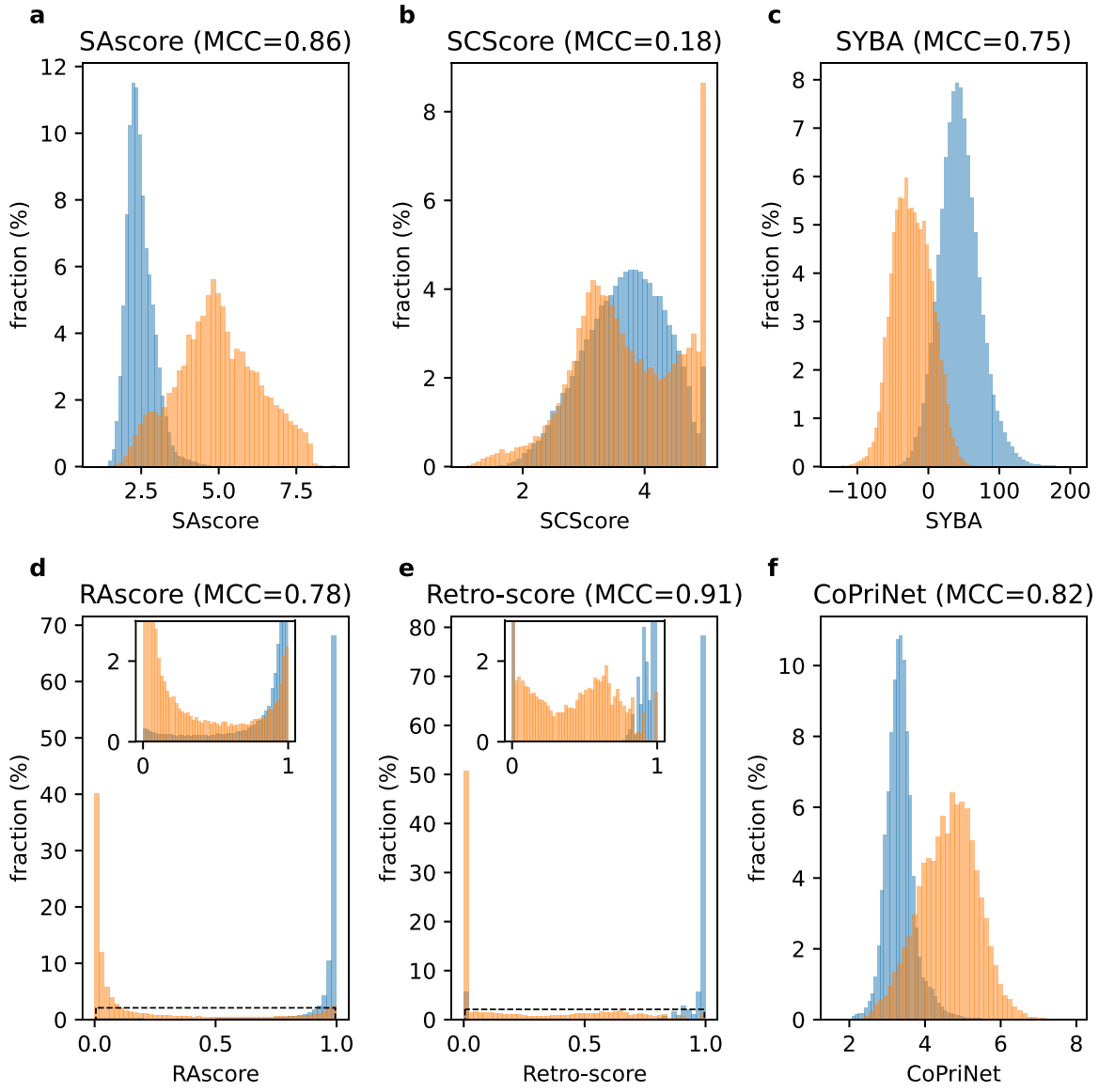


Figure 1. Synthetic accessibility estimations are unable to separate purchasable compounds (PC, blue) from non-purchasable natural products (NPNP, orange). NPNP compounds are expected to exhibit lower synthetic accessibility and larger predicted price. The Matthew's correlation coefficient for each score is displayed in brackets. **a**-**e**) Synthetic accessibility/feasibility score distributions computed with SAscore (Ertl and Schuffenhauer, 2009), SCScore (Coley *et al.*, 2018), SYBA (Voršilák *et al.*, 2020), RAscore (Amol Thakkar et al., 2021), and a retrosynthesis-based score (ManifoldSA, see Methods). **f**) Log-price predictions distribution using CoPriNet.

However, even given this the two distributions should be relatively separable, leading us to conclude that RAscore and SAscore are the best performing of the methods not only because they better separated the two distributions with a Matthew's correlation coefficient (MCC) of 0.86 (SAscore) and 0.78 (RAscore), compared to 0.91 for the ManifoldSA score, but also because their predictions were in better agreement to the ManifoldSA score with Spearman's rank correlation coefficients (SRCC) of 0.79 (SAscore) and 0.77 (RAscore) (see Supplementary Figure 1 and Supplementary Table 1). Despite these levels of correlation only ~45% of the accessible NPNP compounds according to retrosynthesis prediction were detected using SAscore. Moreover, the two metrics showed additional problems beyond accuracy. The SAscore assigned high accessibility values to all the purchasable compounds

(authors recommended a threshold of 6) but also to more than half of the non-purchasable ones, suggesting that the SAscore threshold should be carefully selected depending on the dataset. On the other hand, for the RAscore, due to its binary nature, most of the estimations gravitated around the values of 0 and 1, thus it is not able to obtain a reliable estimation of which of the synthesizable molecules are indeed easy to make. The other two scores performed poorly, the SYBA score exhibited an MCC of 0.75 and SRCC of 0.65, and the SCScore showed even poorer correlations, with MCC and SRCC values below 0.2.

The performance of the different approaches can be influenced by the dataset used. So we also tested their behaviour on the datasets compiled by Gao and Coley (Gao and Coley, 2020) that include typically used catalogues of compounds as well as de novo generated molecules for which retrosynthesis predictions were computed (see Methods). Overall, the SAscore and the RAscore better reproduced the retrosynthesis results (see Supplementary Material Section 3), but the different data subsets offered quite different results. One case of especial interest is the dataset of *de novo* generated molecules that were optimized against several multi-property objective functions (see Supplementary Figures 9-10). In this case, the RAscore score performance drops when the properties used to optimize the molecules do not account for SA. These results are in line with what would be expected for a machine learning approach, since the molecules that are obtained, although biased to replicate catalogue properties, do not necessarily represent viable instances.

The results for the PC and NPNP dataset and those from the Gao and Coley datasets suggest that the SAscore, with all its imperfections, is currently the best heuristics for retrosynthesis-based SA estimation. However, there are also several examples reported in which the SAscore severely underperforms (for visual examples see Supplementary Figure 2). Moreover, retrosynthesis-based methods, despite being computationally demanding, are not perfect at identifying synthetically accessible compounds. The high degree of variability and the fact that the agreement between the different estimations depends on the dataset used, suggests that all methods are imperfect (see Supplementary Figure 11).

## The relationship between price and synthetic accessibility

Though synthetic accessibility is an important criterion, often early in the drug discovery pipeline molecules to be tested are selected based on price, effectively availability and ease of synthesis. Given that, we next examined the relationship between SA metrics and price. We compared the price in the Mcule catalogue for the compounds in the PC dataset to our set of SA scores. All SA metrics had only a weak correlation with price (see Figure 2), with PCC values ranging from 0.16 to 0.35 and SRCC ranging from 0.16 to 0.41. Even a combination of all scores in the form of a linear regression model still performs poorly when trying to predict the price, with a PCC of 0.45. These numbers agree with the value of 0.3 reported by Fukunishi et al. (Fukunishi *et al.*, 2014) and suggest that the synthetic difficulty of a molecule may have only a small impact on the final cost of a compound.

Although this conclusion seems counterintuitive, there are many reasons why this might be the case, for example, compounds that are in high demand will benefit from economies of scale, thus lowering their price regardless of their synthetic accessibility. For the same reasons, it is not unusual to find complicated building blocks at low prices in multiple catalogues, which allows the easy synthesis of otherwise difficult compounds. Nevertheless, while cheap compounds comprise both easy and difficult compounds, it is also true that expensive compounds tend to be less synthetically accessible than their cheaper counterparts (Figure 3).

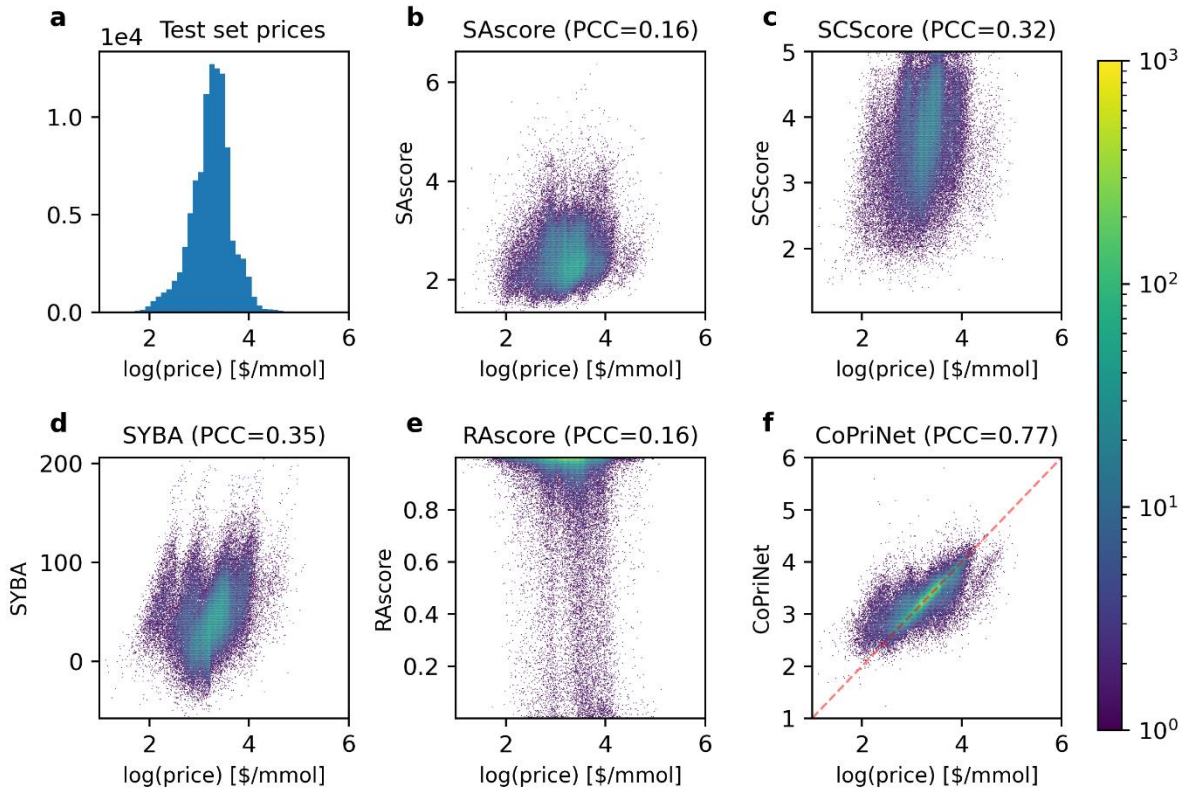


Figure 2. Synthetic accessibility scores correlate poorly with compound price while CoPriNet predictions exhibit better correlation. **a**) Histogram of the natural logarithm of the compound prices of the CoPriNet test set; **b**-**e**) Natural logarithm of the CoPriNet test set compound prices against four different SA scores: SAscore (Ertl and Schuffenhauer, 2009), SCScore (Coley *et al.*, 2018), SYBA (Voršilák *et al.*, 2020) and RAscore (Amol Thakkar et al., 2021); **f**) Natural logarithm of the CoPriNet test set compound prices against CoPriNet predictions for the CoPriNet test set.

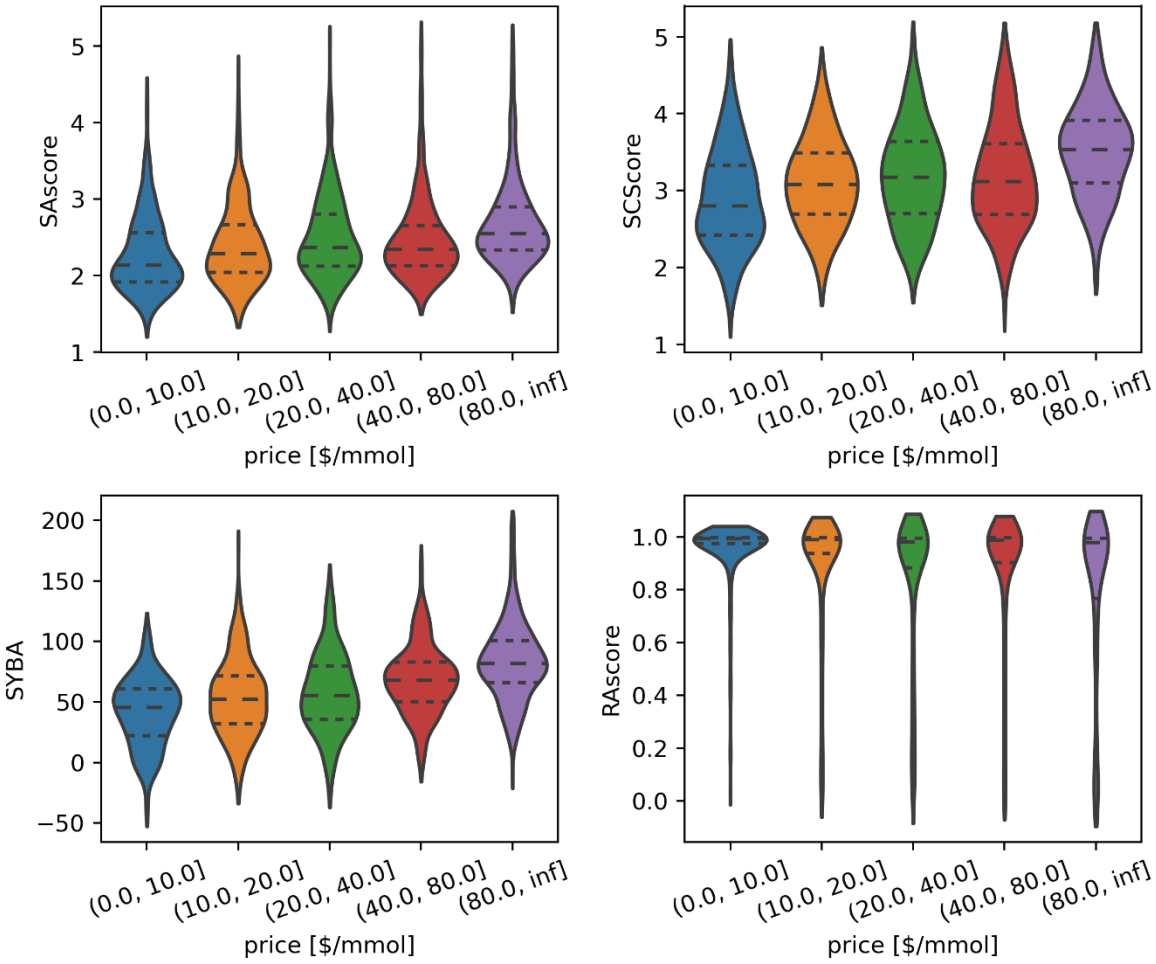


Figure 3. Expensive compounds tend to exhibit larger synthetic accessibility, but the high degree of variability suggests a week relationship between the two variables, making synthetic accessibility scores unhelpful. Distributions of different synthetic accessibility estimations (SAscore (Ertl and Schuffenhauer, 2009), SCScore (Coley *et al.*, 2018), SYBA (Voršilák *et al.*, 2020) and RAscore (Amol Thakkar et al., 2021)) for catalogue compounds of different price ranges. Last price range comprises all compounds with prices above 80$/mmol.

## Price prediction estimation using graph neural networks: CoPriNet

In the same way that the synthetic feasibility of a compound can be predicted without explicitly considering the building blocks and the reaction steps involved, we built a method to predict the compound prices directly from the 2D structure of the compound itself using a GNN trained on pairs of catalogue compounds/prices. Our GNN, termed CoPriNet, is capable of producing more accurate price predictions on our test set than any of the considered SA measurements (Figure 2). CoPriNet achieves a PCC of 0.77 and SRCC of 0.80 which are higher than the best achieved by any of the other methods (Figure 2).

Although it is also true that this approach is not able to compete in terms of accuracy with retrosynthesis-based approaches (see Supplementary Section 5), its running time (~1000 compounds/s on a single GPU) is up to 3 orders of magnitude better than retrosynthesis methods (~1-10 compounds/s). Indeed, this throughput is comparable to fast SA estimations such as RAscore or SYBA, thus is suitable for high-throughput experiments.

## CoPriNet generalizability

The performance of all machine learning methods is strongly influenced by their training set, leading to inaccurate predictions when the studied compounds are quite different to the examples included in the training set (Amol Thakkar et al., 2021). CoPriNet is not an exception and such low generalizability problems may occur.

Given that we can only obtain prices for a tiny fraction of the chemical space that is contained in catalogues and that prices for commercial catalogues are not generally in the public domain, studying compound price prediction generalizability is challenging. Since we only have access to the prices of the molecules in the Mcule catalogue, one of the experiments we can conduct is to ensure that results are consistent independently of the random train/validation split. To do so, we trained CoPriNet on three distinct train/validation partitions, measuring a mean PCC of 0.73 and mean SCC of 0.74 with a standard deviation of 0.04 for PCC and 0.07 for SRCC, showing high consistency.

Next, we assumed that non-purchasable natural products (NPNP) should be more expensive than purchasable compounds (PC) as a way to test for generalizability to non-catalogue compounds. Figure 1 f shows that CoPriNet tends to predict larger prices for the NPNP compounds than for the compounds of the PC dataset. Since the NPNP compounds are quite different from the compounds in the training/validation/test sets used in this work and, on average, they are much more difficult to synthesize than the purchasable compounds, these results suggest generalization capability beyond catalogue compounds.

Finally, we tested generalizability when the method is used on molecules substantially different from the ones in the training set. To do this we evaluated the performance of our method using another test set constructed from virtual compounds included in the Mcule catalogue. These are compounds that are likely to be easily synthesizable from accessible building blocks and for which prices are estimated by the providers according to expected synthetic routes and requirements, thus price distributions tend to be very different. For these compounds, the correlations between all the SA measurements and the price are very poor. CoPriNet predictions tend to systematically underestimate the price of these virtual compounds (see Supplementary Section 6), leading to a poor linear correlation but a reasonable SRCC of 0.56, far better than the one obtained by the best performing SA metric, SCScore, with an SRCC of 0.32. However, since the main purpose of CoPriNet is compound prioritization/optimization, predicting accurate prices is not as important as accurately ranking them, which is not severely affected by the underestimation bias, suggesting that CoPriNet may still be effective even on challenging datasets.

# Conclusions

Availability and ease of synthesises are crucial properties that all drug-like compounds should exhibit to be progressed in the drug discovery pipeline. Due to its importance, several approximations for these properties have been developed. In this manuscript we have illustrated some of the limitations of current synthetic accessibility (SA) estimations for use in estimating availability, including the poor correlation between SA estimations and compound price. The practical implications of this lack of correlation are potentially far ranging since SA estimations are commonly employed for compound prioritization and price is an important variable when deciding which compounds should be assayed. More importantly, most *de novo* generated molecules are biased or optimized against simple SA measurements such as SAscore, thus they may well suggest feasible but prohibitively expensive designs that hardly ever will be selected for progression.

With the aim of alleviating this problem, we propose CoPriNet, a deep learning-based method designed to obtain predictions for compound prices using only their molecular 2D graph. Our approach, evaluated on an independent test set, exhibits far better performance than existing alternatives and an excellent throughput, being able to process up to one thousand molecules per second. This speed means that CoPriNet could be employed in high-throughput settings such as virtual screening or de novo compound generation/optimization, for which retrosynthesis-based approaches are too computationally demanding.

# Methods

## Datasets

Two main sources of compounds were employed in this work. The first is the Mcule catalogue (Kiss *et al.*, 2012), that contains more than 40 million compounds and their up-to-date prices compiled from more than a hundred vendors. In order to avoid common errors that may arise from the integration of different catalogues (misdrawn and incorrect structures), the catalogue is curated using the Mcule Advanced Curation process (MAC) that involves a rigorous molecule registration system with various structural checks, and various steps of standardization, preparation and correction, ensuring that the information contained in the catalogue is highly reliable. From this catalogue we extracted the subset of in-stock compounds (~7M), that was divided into train, validation, and test partitions randomly. The price of each compound was extracted on March 2021 from the Mcule database as the price for 1 g of compound. Prices were then converted from $/g to $/mmol because, as suggested by Fukunishi et al. (Fukunishi *et al.*, 2014), correlations with SA measurements were stronger. All statistics and Figures included in this work are derived from the compounds in the test set except when explicitly stated. The test set is also referred to as the purchasable compounds dataset (PC) throughout this manuscript as it only contains purchasable compounds. For the generalizability study, a random subset of 100K virtual compounds was also extracted from the Mcule catalogue as a separate independent testing set.

The second source of compounds was the ZINC database (Sterling and Irwin, 2015) from which we extracted a subset comprising only non-commercially available natural products, that we refer to as the NPNP (Non-Purchasable Natural products) dataset. We use this dataset as an approximate set of non-synthesizable compounds.

We also employed two of the datasets compiled by Cao and Coley (Gao and Coley, 2020). Particularly, their dataset of molecules compiled from five different sources: MOSES (Polykovskiy *et al.*, 2020), ZINC (Sterling and Irwin, 2015), ChEMBL (Gaulton *et al.*, 2012), Sheridan et al. (Sheridan *et al.*, 2014), and GBD-17 (Ruddigkeit *et al.*, 2012); and their dataset of *de novo* generated molecules comprising of two subsets of molecules optimized against multiple properties including or not the SAscore.

## Synthetic accessibility calculations

Four distinct SA metrics were employed in this work: SAscore (Ertl and Schuffenhauer, 2009), SCScore (Coley *et al.*, 2018), the AstraZeneca RAscore (Amol Thakkar et al., 2021) and SYBA (Voršilák *et al.*, 2020). All of them were executed using default parameters. Additionally, the retrosynthesis-based score ManifoldSA was computed using the Postera Manifold API v1 (https://api.postera.ai/api/v1/docs/). ManifoldSA summarizes retrosynthesis results into a number between 0 (easy) and 1 (difficult) that estimates the synthetic accessibility of a compound. For comparison ease, we used 1 – ManifoldSA instead. Discretization was carried out considering that compounds with ManifoldSA < 0.5 are synthesizable and compounds with ManifoldSA < 0.2 are

easily synthesizable. In summary, Manifold first performs a tree-search to compute possible retrosynthetic routes from the target molecule to purchasable starting materials, using Molecular Transformer to predict the probability of success of each step. The ManifoldSA is then computed by considering the feasibility and robustness of multiple routes to the molecule, taking into account probability of success at each step of a route. Manifold algorithm has been reported to be used in synthesis-driven de novo design (Morris *et al.*, 2021).

## Retrosynthesis calculations

Retrosynthesis prediction was carried out using the Postera Manifold API, that implements the molecular transformer approach (Schwaller *et al.*, 2019b; Lee *et al.*, 2019). We employed the v1 retrosynthesis endpoint using a depth search of four and the Mcule catalogue as the source of building blocks.

For price estimation from retrosynthesis predictions we employed a simple heuristic that only considers the cost of the building blocks neglecting any additional cost. Thus, taking into account that the retrosynthesis results obtained for each compound tend to include several pathways, potentially involving multiple building blocks, we employed two simple strategies. The first one assumes ideal route ranking, thus overestimates the performance (ignoring non-reactant costs), by selecting the route that better matched our price records. The second strategy just reports the minimum price route. Both approaches could be easily improved considering aspects such as reaction types involved, yield prediction, etc, but this data is not generally available in the public domain and the usage of predictors for such properties is outside the scope of this work. Other approaches such as the method proposed by Bodowski et al. (Badowski *et al.*, 2019) could not be employed as they are not publicly available.

## CoPriNet Graph Neural Network

To create our price prediction GNN we represented compounds as 2D graphs with atoms corresponding to nodes and bonds to edges. Nodes are encoded using five features: atomic number, valence, formal charge, number of neighbours, and aromaticity. Edges are represented with four features: bond type, aromaticity, conjugation, and ring membership.

Our GNN first projects the node and edges features using a learnable linear transformation from dimension five and four to 75 and 50 respectively. After that, ten blocks consisting of a PNA layer (Corso *et al.*, 2020), batch normalization (Ioffe and Szegedy, 2015) and ReLU (Nair and Hinton, 2010) activation are applied one after another. Then, an embedding for the graph is obtained applying a Set2set layer (Vinyals *et al.*, 2015). Finally, two dense layers with batch normalization and ReLU activation and one last linear layer with one single unit are applied to the graph embedding. A schematic of our GNN architecture is depicted in Figure 4. The hyperparameters were selected by cross-validation over the validation dataset, exhibiting a robust behaviour. See Supplementary Material Section 7 for more details.

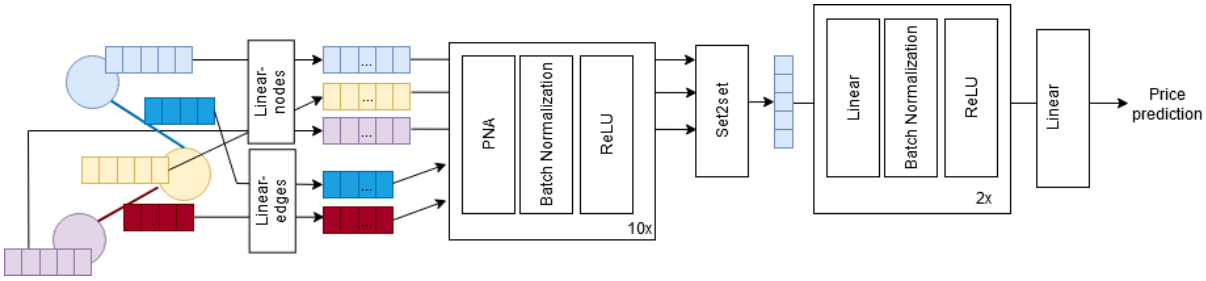

Figure 4. Price prediction Graph Neural Network architecture. The graph of a molecule consisting of nodes (circles) and edges (dark blue and red lines) is encoded as node vectors (dimension five, pale blue, green and purple rectangles) and edge vectors (dimension four, dark blue and red rectangles). The node and edge vectors are embedded into higher dimensionality embeddings using independent learnable weights for the nodes (Linear-nodes) and for the edges (Linear-edges). After that, the node and edge embeddings are processed by ten blocks of PNA layer, batch normalization and ReLu activation, updating the state of the embeddings after each block. Then, the processed embeddings of all the nodes are combined into one single graph embedding using a Set2Set layer. Finally, the graph embedding is processed by two blocks of linear layer, batch normalization and ReLu activation from which the price prediction is obtained using a linear layer with one single unit.

Our network was trained using the Adam optimizer (Kingma and Ba, 2014) with a batch size of 512 graphs. Initial learning rate was set to $10^{-5}$ and decreased by a factor of 0.1 when the validation loss did not improve during 25 epochs. The mean squared error was used as the loss function.

### Evaluation metrics

The correlation between continuous variables was measured using the absolute value of the Pearson Correlation Coefficient (PCC, Eq. 1) and the Spearman's Rank Correlation Coefficient (SRCC, Eq. 2).

$$PCC = abs\left(\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})}\sqrt{\sum(Y_i-\bar{Y})}}\right) \quad \text{Eq. 1}$$

$$SRCC = abs\left(1-\frac{6\sum\left(R(X_i)-R(Y_i)\right)^2}{n(n^2-1)}\right) \quad \text{Eq. 2}$$

where $X_i$ and $Y_i$ are the $i$-th observations of the variable $X$ and $Y$, $\bar{X}$ is the average of variable $X$, $R(X_i)$ is the ranking of the $i$-th observation of the variable $X$ and n is the number of observations.

Binary classification performance was evaluated using the Matthews Correlation Coefficient (MCC, Eq. 3) at the threshold that maximizes its value.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{Eq. 3}$$

where $TP$ is the number of true positive predictions, $TN$ is the number of true negative predictions, $FP$ is the number of false positive predictions and $FN$ is the number of false negative predictions.

## Code and data availability

CoPriNet source code, trained models and test dataset are available at https://github.com/oxpig/CoPriNet.


## Acknowledgments

We thank Mcule.com for sharing their private data and their support and assistance. We thank PosteraAI for their support and assistance. This work has been economically supported by the Rosetrees Trust (Ref M940).

# References


Allu,T.K. and Oprea,T.I. (2005) Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *Journal of Chemical Information and Modeling*, **45**, 1237–1243.

Badowski,T. *et al.* (2019) Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chemical Science*, **10**, 4640–4651.

Barone,R. and Chanon,M. (2001) A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *Journal of Chemical Information and Computer Sciences*, **41**, 269–272.

Basu,S. *et al.* (2020) Novel cyclohexanone compound as a potential ligand against SARS-CoV-2 main-protease. *Microbial pathogenesis*, **149**, 104546.

Boda,K. *et al.* (2007) Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design 2007 21:6*, **21**, 311–325.

Coley,C.W. *et al.* (2019) A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, **365**.

Coley,C.W. *et al.* (2018) SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling*, **58**, 252–261.

Corso,G. *et al.* (2020) Principal Neighbourhood Aggregation for Graph Nets. *Advances in Neural Information Processing Systems*, **2020-December**.

Dai,H. *et al.* (2020) Retrosynthesis Prediction with Conditional Graph Logic Network. *Advances in Neural Information Processing Systems*, **32**.

Ertl,P. and Schuffenhauer,A. (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics 2009 1:1*, **1**, 1–11.

Fukunishi,Y. *et al.* (2014) Prediction of synthetic accessibility based on commercially available compound databases. *Journal of Chemical Information and Modeling*, **54**, 3259–3267.

Gao,W. and Coley,C.W. (2020) The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling*, **60**.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, **40**, D1100.

Genheden,S. *et al.* (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics 2020 12:1*, **12**, 1–9.

Gillet,V.J. *et al.* (1995) SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspectives in Drug Discovery and Design 1995 3:1*, **3**, 34–50.

Green,D.V.S. *et al.* (2020) BRADSHAW: a system for automated molecular design. *Journal of Computer-Aided Molecular Design*, **34**, 747–765.

Hendrickson,J.B. *et al.* (1987) Molecular Complexity: A Simplified Formula Adapted to Individual Atoms. *Journal of Chemical Information and Computer Sciences*, **27**, 63–67.

Huang,Q. *et al.* (2011) RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. *Journal of Chemical Information and Modeling*, **51**, 2768–2777.

Huang,Y. *et al.* (2019) Discovery of novel allosteric site and covalent inhibitors of FBPase with potent hypoglycemic effects. *European Journal of Medicinal Chemistry*, **184**, 111749.

Humbeck,L. *et al.* (2018) CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein–Protein Interaction Modulators. *ChemMedChem*, **13**, 532–539.

Ihlenfeldt,W.-D. and Gasteiger,J. (1996) Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angewandte Chemie International Edition in English*, **34**, 2613–2633.

Imrie,F. *et al.* (2020) Deep Generative Models for 3D Linker Design. *Journal of chemical information and modeling*, **60**, 1983–1995.

Imrie,F. *et al.* (2021a) Generating property-matched decoy molecules using deep learning. *Bioinformatics*, **37**, 2134.

Imrie,F. *et al.* (2021b) Generating property-matched decoy molecules using deep learning. *Bioinformatics*, **37**, 2134.

Ioffe,S. and Szegedy,C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning, ICML 2015*, **1**, 448–456.

Khemchandani,Y. *et al.* (2020a) DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *Journal of Cheminformatics 2020 12:1*, **12**, 1–17.

Khemchandani,Y. *et al.* (2020b) DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: A graph convolution and reinforcement learning approach. *Journal of Cheminformatics*, **12**.

Kim,S. *et al.* (2016) PubChem Substance and Compound databases. *Nucleic Acids Research*, **44**, D1202.

Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kiss,R. *et al.* (2012) http://Mcule.com: a public web service for drug discovery. *Journal of Cheminformatics 2012 4:1*, **4**, 1–1.

Lee,A.A. *et al.* (2019) Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, **55**, 12152–12155.

Leguy,J. *et al.* (2020) EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *Journal of Cheminformatics 2020 12:1*, **12**, 1–19.

Lu,Y. and Li,M. (2021) A New Computer Model for Evaluating the Selective Binding Affinity of Phenylalkylamines to T-Type Ca(2+) Channels. *Pharmaceuticals (Basel, Switzerland)*, **14**.

Makara,G.M. *et al.* (2021) Derivatization Design of Synthetically Accessible Space for Optimization: In Silico Synthesis vs Deep Generative Design. *ACS Medicinal Chemistry Letters*, **12**, 185–194.

Morris,A. *et al.* (2021) Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model. *Chemical Communications*, **57**, 5909–5912.

Nair,V. and Hinton,G.E. (2010) Rectified linear units improve Restricted Boltzmann machines. In, *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*., pp. 807–814.

Nicolaou,C.A. and Brown,N. (2013) Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies*, **10**, e427–e435.

Omolabi,K.F. *et al.* (2021) A probable means to an end: exploring P131 pharmacophoric scaffold to identify potential inhibitors of Cryptosporidium parvum inosine monophosphate dehydrogenase. *Journal of molecular modeling*, **27**, 35.

Podolyan,Y. *et al.* (2010) Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods. *Journal of Chemical Information and Modeling*, **50**, 979–991.

Polykovskiy,D. *et al.* (2020) Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, **0**, 1931.

Prykhodko,O. *et al.* (2019) A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, **11**.

Ruddigkeit,L. *et al.* (2012) Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, **52**, 2864–2875.

Schwaller,P. *et al.* (2019a) Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, **5**, 1572–1583.

Schwaller,P. *et al.* (2019b) Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, **5**, 1572–1583.

Sheridan,R.P. *et al.* (2014) Modeling a Crowdsourced Definition of Molecular Complexity. *Journal of Chemical Information and Modeling*, **54**, 1604–1616.

Sterling,T. and Irwin,J.J. (2015) ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, **55**, 2324–2337.

Thakkar,A. *et al.* (2021) Retrosynthetic accessibility score (RAscore)-rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science*, **12**, 3339–3349.

Vinyals,O. *et al.* (2015) Order Matters: Sequence to sequence for sets. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Voršilák,M. *et al.* (2020) SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics 2020 12:1*, **12**, 1–13.

Voršilák,M. and Svozil,D. (2017) Nonpher: computational method for design of hard-to-synthesize structures. *Journal of Cheminformatics 2017 9:1*, **9**, 1–7.

Yassine,R. *et al.* (2021) Active Learning and the Potential of Neural Networks Accelerate Molecular Screening for the Design of a New Molecule Effective against SARS-CoV-2. *BioMed Research International*, **2021**.

Yiming Mo *et al.* (2021) Evaluating and clustering retrosynthesis pathways with learned strategy. *Chemical Science*, **12**, 1469–1478.

Zhou,Z. *et al.* (2019) Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports 2019 9:1*, **9**, 1–10.