

RESEARCH ARTICLE

Predicting the Price of Molecules Using Their Predicted Synthetic Pathways

Massina Abderrahmane | Hamza Tajmouati | Vinicius Barros Ribeiro da Silva | Quentin Perron

Iktos, Paris, France

Correspondence: Quentin Perron (quentin.perron@iktos.com)**Received:** 30 January 2024 | **Revised:** 11 November 2024 | **Accepted:** 3 January 2025**Funding:** IKTOS**Keywords:** deep learning | molecular price prediction | retrosynthesis | synthetic accessibility

ABSTRACT

Currently, numerous metrics allow chemists and computational chemists to refine and filter libraries of virtual molecules in order to prioritize their synthesis. Some of the most commonly used metrics and models are QSAR models, docking scores, diverse druggability metrics, and synthetic feasibility scores to name only a few. To our knowledge, among the known metrics, a function which estimates the price of a novel virtual molecule and which takes into account the availability and price of starting materials has not been considered before in literature. Being able to make such a prediction could improve and accelerate the decision-making process related to the cost-of-goods. Taking advantage of recent advances in the field of Computer Aided Synthetic Planning (CASP), we decided to investigate if the predicted retrosynthetic pathways of a given molecule and the prices of its associated starting materials could be good features to predict the price of that compound. In this work, we present a deep learning model, RetroPriceNet, that predicts the price of molecules using their predicted synthetic pathways. On a holdout test set, the model achieves better performance than the state-of-the-art model. The developed approach takes into account the synthetic feasibility of molecules and the availability and prices of the starting materials.

1 | Introduction

During the research and development (R&D) process of a new drug, chemists use different metrics to refine and filter libraries of virtual molecules in order to prioritize their synthesis, such as QSAR models' predictions of biological or physchem properties [1], docking scores [2], druggability metrics [3], and synthetic feasibility scores [4, 5], to name only a few. On top of those well known metrics, to our knowledge, no function which estimates the price of a novel virtual molecule taking into account its synthetic pathway and the associated starting

materials (price, provider, and availability) has been previously considered in literature.

Indeed, there are only few reports in the literature of works related to the prediction of a molecule's price. In 2019, Badowski & all [6] proposed an expert-rule based system to compute the cost of an existing synthetic pathway. They represented the cost of a molecule by the cost of the least expensive synthetic route. Each route's cost was computed as the sum of the cost of the initial reactants, plus the cost of the reactions required to synthesize the molecule weighted by the

Abbreviations: CASP, Computer Aided Synthetic Planning; CMP, Commercial Molecules with Prices; MAE, Mean Absolute Error; MCTS, Monte Carlo Tree Search; MSR, Multi-Step Retrosynthesis; PRp, Product fingerprint and Reactants prices; PRRp, Product fingerprint, Reaction fingerprint and Reactants prices; Rp, Reactants prices; RRRp, Reaction fingerprint and Reactants prices; SaaS, Software as a Service; SSR, Single-Step Retrosynthesis; VPC, Virtual Private Cloud

yield of the reaction. This method has a limitation, as it relies on the cost and yield of a reaction which are often not readily available and can be difficult to predict, especially for molecules that have not been previously described. Consequently, this limitation affects the practical usefulness of the method since the primary objective of a price prediction solution is to predict the price of entirely new molecules.

In that direction, in 2022, CoPriNet [7], a deep learning model, was introduced to predict the price of a molecule using only its chemical structure as input. It is a graph neural network that was trained on providers catalog prices, to predict the price of a molecule in dollars per millimole, directly from its molecular 2D graph, achieving good performance on an external test set. To the best of our knowledge, it is the state-of-the-art method for compound price prediction. However, it is well known that the price of a given molecule depends on many parameters such as the number of synthetic steps, the sequence of reactions, the price of the reagents and reactants, the yields of reactions, the human resources, and so on. Therefore, a model that solely predicts price based on the molecular graph will likely incorrectly estimate prices for molecules that are not feasible. Furthermore, such a model is unable to account for practical constraints in real-life scenarios such as the availability of the building blocks, key intermediates, or specific chemistry.

To the best of our knowledge, no method that uses the predicted retrosynthetic route of a given molecule to estimate its price has been published so far. Indeed, having rapid access to an accurately predicted retrosynthetic pathway was a very difficult task until recently.

Significant advances have recently been made in retrosynthesis technology, driven by the integration of artificial intelligence (AI) and machine learning algorithms that can help predict reaction outcomes and retrosynthetic disconnections [8]. These advances have enabled the development of powerful retrosynthesis tools, which can generate high-quality synthetic routes for complex molecules more efficiently and effectively than traditional expert methods [9]. Many tools are available today for molecular retrosynthesis. Some of them are public platforms such as ASKCOS [10] and AiZynthFinder [11], and others are commercial platforms such as Synthia [12], ChemAIRS [13] and Spaya [14].

We hypothesized that having access to the full retrosynthetic pathway of a given molecule and the price of the starting materials should allow progress toward the estimation of the price of that route. Following this intuition, we decided to investigate in that direction, using Spaya, our in-house Computer Aided Synthetic Planning (CASP) technology. Consequently, we have developed RetroPriceNet, a deep learning model which predicts the price of a new molecule in dollars per gram using its predicted synthetic pathway and the price of the predicted commercial starting materials.

Spaya is a template-based retrosynthesis AI software that computes synthetic routes of molecules and ranks them based on a synthesizability score. Spaya relies on neural networks and data driven procedures that are combined using Monte

Carlo Tree Search (MCTS) to perform chemical synthesis planning [15]. The first important block consists of a neural network, referred to as the expansion network, that serves as a retrosynthesis template predictor. It guides the search towards promising pathways by suggesting a limited set of viable retrosynthesis templates for molecular decomposition. Spaya's expansion network was trained on Pistachio reactions [16]. The second block consists of quality filters that estimate whether the proposed reactions are actually feasible. For this purpose, in-house filters were developed, allowing us to assess the quality of reactions against those documented in the literature. Monte Carlo Tree Search is then performed in 4 main steps: selection, expansion, simulation and rollout. Final recovered routes are scored using a proprietary score named the Rscore [4]. The set of molecule building blocks used in Spaya is supplied by multiple providers and is updated every three months.

RetroPriceNet takes the predicted routes found by Spaya as input, with the possibility of using search options such as a selection of providers, a maximal number of steps, a maximal price of starting materials, a set of name reactions to pass by or to avoid, the presence of key intermediates, and so on. For each molecule, out of the top routes found by Spaya, the best route is selected to feed RetroPriceNet, based on the Rscore [4], Iktos proprietary synthetic feasibility score, which is used to compare synthetic routes based on their probability of success. RetroPriceNet was trained on a large database of 6.2 million commercial molecules along with their prices and associated predicted synthetic routes. It demonstrated good performance surpassing the current state-of-the-art methods.

2 | Materials and Methods

2.1 | Price Prediction Pipeline

The proposed method is composed of two main steps as shown in Figure 1:

- The first step of the method consists of recovering a predicted synthetic pathway from the input molecule. Spaya is used for this purpose.
- The second step of the method consists of predicting the price of a molecule using the previously identified synthetic route. For this purpose, we built RetroPriceNet, a deep neural network that predicts the price of a molecule by taking as input its retrosynthetic tree predicted by Spaya and the price of the associated starting materials.

2.2 | Datasets

In order to train RetroPriceNet in a supervised manner, it is necessary to build a dataset composed of molecules with a synthetic route and a price. The price variable is the output to be predicted. To take into account the potential price variations among different providers for the same molecule, we established a price standardization pipeline (as detailed in Section 3.2.1) to define the price of each molecule in the dataset.

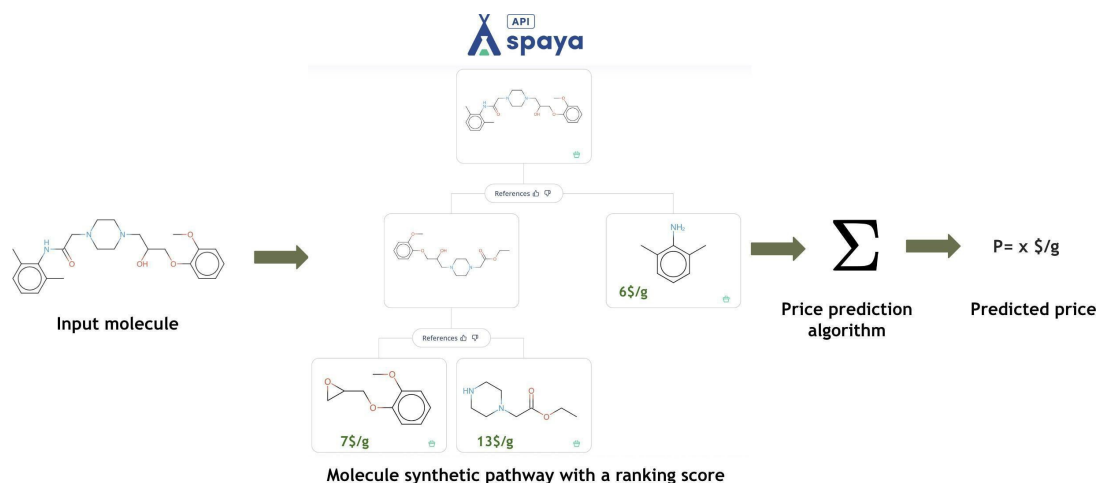


FIGURE 1 | Method pipeline for predicting the price of a molecule.

To come up with predicted synthetic routes, retrosynthetic analysis using Spaya was run on the molecules.

Two datasets were prepared to train the model: a training dataset of molecules with multi-step synthetic routes and a pretraining dataset of molecules with only single-step synthetic routes. The pretraining is optional and aims to improve the performance of the model. The datasets are presented in Sections 3.2.2 and 3.2.3.

2.2.1 | Price Standardization Pipeline

Different sources of compounds were employed in this work. The catalogs of commercial starting materials available in Spaya in March 2022 were recovered from the following providers: MolPort, eMolecules, ChemSpace, ChemBridge, Life-Chemicals, Apollo, 1plusChem, TciChemicals, Key-Organics, ChemTellec, AdvancedChemBlocks and A2bChem.

In order to merge the different catalogs, a process of data curation was performed on each catalog taking as input building blocks SMILES string [17], and using the following steps to generate standardized data:

1. Check the length of the SMILES string < 512;
2. Read the input SMILES string as an RDKit [18] molecule object;
3. Convert isotopes to their normal form except deuteriums;
4. Neutralize functional groups;
5. Split the molecule into multiple structures if necessary, filter salts, solvents and fragments with non organic elements and select the largest fragment.

After data curation, the price standardization pipeline presented in Figure 2 was implemented:

- Step 1 (selection of packaging): as the same molecule can have different prices depending on the packaging, we chose to use only the price of the 1 gram packaging.
- Step 2 (average price computation): the average price of each molecule was computed from the different providers after having removed molecules for which the ratio between the maximum price and the minimum price was higher than 10.

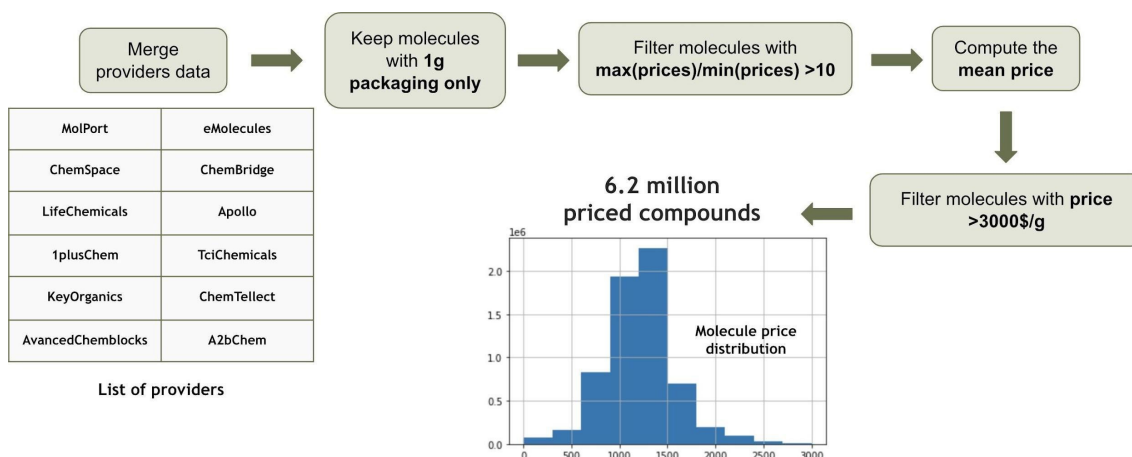


FIGURE 2 | Price standardization pipeline.

- Step 3 (data filtering): compounds with a price greater than 3000 dollars per gram were filtered out, as they represented a very small subset of the dataset.

The output of the price standardization pipeline is a dataset of commercial molecules with a price: *CMP dataset*. It contains a list of molecules represented by their SMILES string with their average price in dollars/gram. The final CMP dataset is composed of 6.2 million priced compounds.

2.2.2 | Training Dataset

RetroPriceNet is trained on predicted synthetic routes identified by Spaya for molecules of the CMP dataset.

Since a retrosynthesis calculation is time consuming (around 10 s/molecule), only a subset of the dataset was processed. A sample of 400,000 random molecules was selected to form the training dataset and their synthetic routes were computed using Spaya API, a cloud-based platform providing Spaya retrosynthesis services through a scalable RESTful API [14]. The retrosynthesis was constrained to the set of building blocks recovered from the CMP dataset, to ensure that all predicted synthetic routes have reactants with prices. Moreover, to reduce the computational cost, the retrosynthesis running time was constrained, with a timeout period set to one minute.

The retrosynthetic routes where the price of the final molecule was higher than the price of the initial reactants were filtered out, in order to be closer to real life use cases. For each molecule, only the best route was selected, i.e., the route with the highest Rscore. Finally, only molecules having a route with a maximum of 6 synthesis steps were retained, resulting in the removal of a small fraction of the molecules from the training dataset (1441 molecules).

The full data preparation pipeline is illustrated in Figure 3. The final dataset is named “multi-step synthesis route” dataset (*MSR dataset*) and consists of 357,172 priced molecules associated with their corresponding multi-step synthetic routes.

The difference between the size of the dataset before and after retrosynthesis originates from two main reasons:

1. Molecules for which Spaya finds no route. This could be due to missing starting materials or new reactions that are not yet included in Spaya, or due to the defined retrosynthesis conditions (the number of steps, the prices of starting materials, and the timeout of retrosynthesis).
2. The filter which is applied on the number of steps (≤ 6) of the obtained synthetic routes after Spaya’s retrosynthesis.

2.2.3 | Pretraining Dataset

RetroPriceNet can be pretrained on a dataset of molecules with only single step synthetic routes. To generate the pretraining dataset, retrosynthesis was performed on the full CMP dataset (6.2 million molecules) using Spaya API with the same configuration described in the previous section, adding a condition on the routes to be single step. For each molecule, the best scored route was selected as described above. The pretraining dataset is named “single-step synthesis route” dataset (*SSR dataset*). The final SSR dataset is composed of 5.1 million of priced molecules associated with a single-step synthetic route.

2.3 | RetroPriceNet

2.3.1 | Model Architecture

RetroPriceNet is a feed forward neural network (M_θ) [19] that predicts the price of a molecule from its predicted single step synthetic route and the price of its reactants (Figure 4). From the single step of synthesis, two vectors are computed:

- The vector of the counted Morgan Fingerprint [20] of the product molecule. In the counted Morgan Fingerprint, each bit represents the appearance count of the motif in the molecule. Radius 3 and size 8192 were used to compute the fingerprint vector.

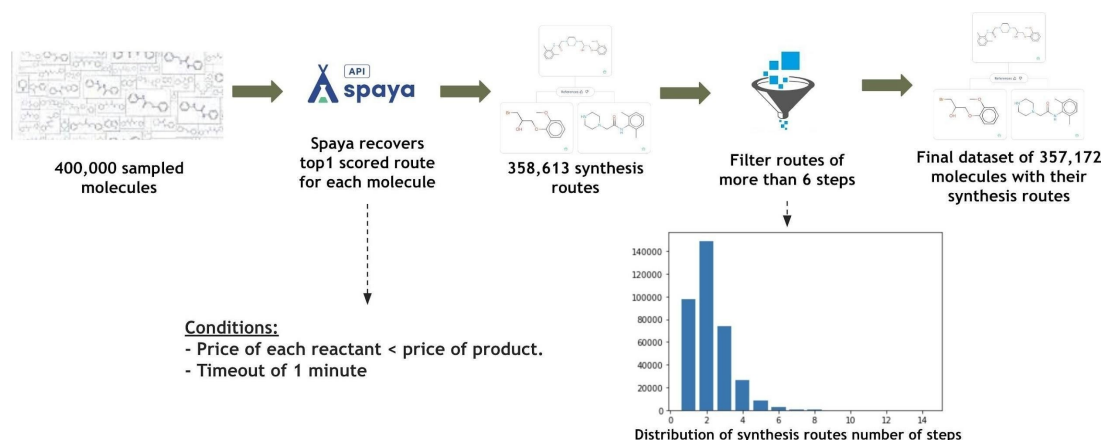


FIGURE 3 | Training dataset computation pipeline.

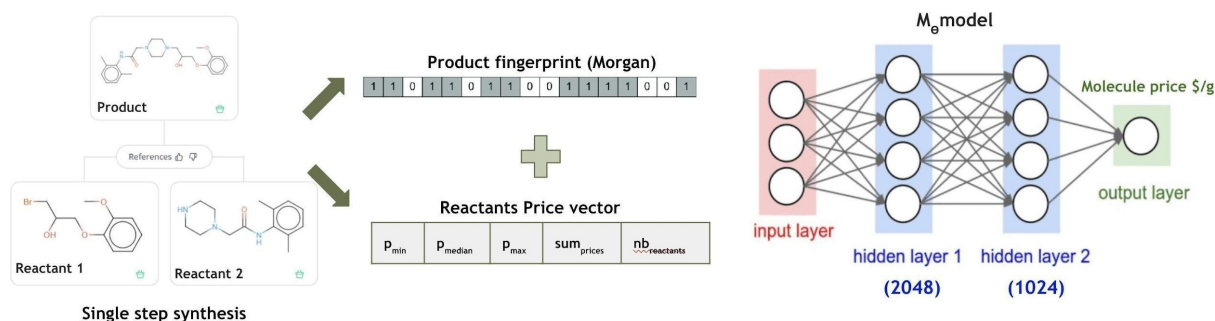


FIGURE 4 | RetroPriceNet prediction pipeline.

- A vector representing reactants prices. As the number of reactants vary from one reaction to another, the vector of reactants prices was standardized by computing 5 values: the minimum price of reactants, the median price of reactants, the maximum price of reactants, the sum of reactant prices and the number of reactants.

These two vectors are concatenated and given to the Model M_θ that outputs the predicted price of the product.

2.3.2 | Price Inference

To predict the price of molecules with multi-step synthetic routes, RetroPriceNet is used along the synthetic route to predict the price of intermediate components step-by-step until reaching the final molecule, as shown in Figure 5. The inference procedure to compute the predicted price of a given molecule p_θ is described in Algorithm 1.

2.3.3 | Loss Definition

The loss of the model is defined in Equation (1) as a mean squared error loss between the predicted price (p_θ) of the root molecule and its true price (y) weighted by the score (s) representing the ranking score (Rscore [4]) of the route. The Rscore was used as a weighting factor in the training loss in order to prioritize the information coming from well scored routes against the one coming from badly scored routes during the training.

$$L = E[s(p_\theta - y)^2] \quad (1)$$

2.4 | Model Training Experiments

Several model training experiments were performed to assess the impact of different parameters/settings.

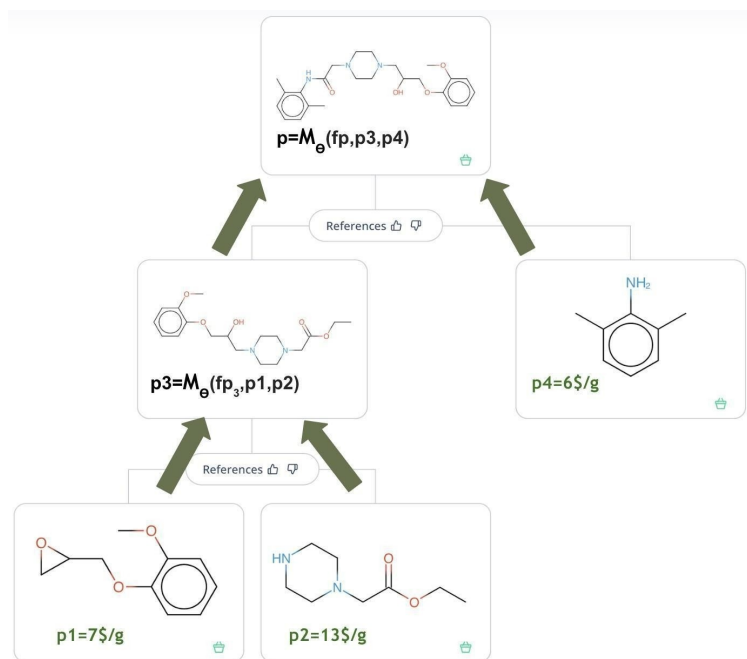


FIGURE 5 | Prediction along synthetic tree using RetroPriceNet (M_θ) model.

Algorithm 1 Price inference.

```

1:  $x \leftarrow$  Synthetic tree
2:  $M_\theta \leftarrow$  RetroPriceNet model with  $\theta$  weights
3:  $steps \leftarrow x[steps]$ 
4:  $S \leftarrow len(steps)$ 
5: for  $i = 1..S$  do
6:    $v \leftarrow compute\_features(Product(steps^{(i)}), price[Reactants(steps^{(i)})])$ 
7:    $p_\theta \leftarrow M_\theta(v)$ 
8:    $price[Product(steps^{(i)})] \leftarrow p_\theta$ 
9: return  $p_\theta$ 

```

▷ Synthesis steps ordered from bottom to top

▷ Figure 4

▷ The predicted price of intermediate product

▷ The predicted price of the final molecule

2.4.1 | Impact of Pretraining

As RetroPriceNet performs its predictions using single step synthetic routes, and recovering single-step synthetic routes is less time consuming than recovering deep routes through Spaya, we hypothesized that it could be interesting to pretrain RetroPriceNet on a large dataset of single-step synthetic routes. To assess the impact of pretraining on model performance, we conducted the following experiments:

1. Training RetroPriceNet from scratch on MSR dataset, the multi-step synthetic routes of molecules recovered from Spaya [14].
2. Pretraining RetroPriceNet on SSR dataset, single-step synthetic routes of molecules first, before training the model on MSR dataset in a second step.

2.4.2 | Experiments on Model Input Representations

Experiments were also conducted to assess the impact of the input representations on the performance of the model. Different representations were tested:

- Product fingerprint + Reactants prices (PRp)
- Reaction fingerprint + Reactants prices (RRp)
- Product fingerprint + Reaction fingerprint + Reactants prices (PRRp)
- Reactants prices (Rp)

Product fingerprint was computed as the counted Morgan fingerprint of the product molecule with radius 3 and size 8192. The reaction fingerprint was computed as the difference between the product and reactants fingerprint. Reactants fingerprint is computed as the fingerprint of the joined molecule composed of all reactants. Product and reactants fingerprints were also computed using the counted version of the Morgan Fingerprint with radius 3 and size 8192.

2.5 | Model Evaluation

Evaluation Procedure. Both MSR (multi-step retrosynthetic routes) and SSR (single-step retrosynthetic routes) datasets

were randomly split into train (80%) and test (20%) sets. RetroPriceNet was evaluated on the test set of the MSR dataset. To fairly evaluate the effect of pretraining, MSR test set and SSR dataset have to be disjoint. Hence, the molecules from the MSR test set were filtered out of the SSR dataset. The mean absolute error (MAE) in Equation (2) and the coefficient of determination (R2 score) in Equation (3) were used as evaluation metrics to compare regression models. (y_i , \hat{y}_i and \bar{y} represent the true price, the predicted price, and the empirical mean of the observed prices, respectively.)

$$MAE = \sum_{i=1}^D |y_i - \hat{y}_i| \quad (2)$$

$$R2 \text{ score} = 1 - \sum_{i=1}^D \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2} \quad (3)$$

The models were also evaluated in a classification context, using the accuracy metric defined in Equation (4) 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.

Accuracy was computed on defined price ranges, because when considering the potential application of the model, we hypothesized that chemists don't need the precise price of the molecule, rather they need a metric that indicates whether the molecule is cheap, has an average price or is expensive. Four ranges (in \$/g) were defined: <100, [100, 500], [500, 1000] and >1000.

Accuracy was also computed according to a percentage of tolerated error (α) on the price of the molecule. A molecule was considered well predicted if the predicted error did not exceed the tolerated error as shown in Equations (5) and (6) where $Price_{pred}$ represents the predicted price of the molecule, $Price$ represents the real price of the molecule and α represents the tolerated error in percentage.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Price_{pred} \subseteq [Price \pm Error] \quad (5)$$

$$\text{Error} = \alpha \cdot \text{Price} \quad (6)$$

Baseline: RetroPriceNet was compared to two baselines:

- Similarity baseline: assigns the price of the most similar molecule in the MSR (multi-step retrosynthetic routes) training set. Tanimoto similarity on Morgan Fingerprints [20] was used for this use case.
- CoPriNet [7]: a state-of-the-art compound price prediction model, was trained on pairs of molecular 2D graphs and their corresponding prices in dollars/millimole (\$/mmol). We used our CMP (commercial molecules with price) dataset to train CoPriNet, splitting it into 80% for training and 20% for testing, following the RetroPriceNet splits for a fair comparison. To align with RetroPriceNet, CoPriNet's predictions in \$/mmol were converted to \$/g. Similarly, RetroPriceNet's predictions in \$/g were converted to \$/mmol to ensure a fair evaluation.

3 | Results and Discussion

Our model, RetroPriceNet, predicts the price of molecules as 1 gram packaging in \$/g, using their predicted synthetic routes. The price of the intermediate compounds found in the retrosynthetic tree are predicted step-by-step, starting from the commercial building blocks until reaching the target molecule. By providing such information, the model is trained to learn a function which links the price of a molecule to the price of its predicted “ingredients” (reactants) and the associated recipe (reactions). We believe that this function implicitly incorporates hidden variables, such as reaction yield, reaction cost and labor-related expenses, which account for the actual price of a given molecule on top of the reactant prices.

Two experiments were performed:

- Training RetroPriceNet from scratch on the MSR (multi-step retrosynthetic routes) dataset.
- Pre-training RetroPriceNet on the SSR (single-step retrosynthetic routes) dataset before launching the training on the MSR dataset.

Table 1 shows a comparative analysis of the similarity baseline performance metrics against RetroPriceNet both with and without the pretraining step. The results indicate that RetroPriceNet outperforms the baseline. The baseline achieves an R2 score of only 0.18 and an MAE (mean absolute error) of 212.9 \$/g. The accuracy of the baseline on price ranges is 74.8%, which, although lower than the performance of RetroPriceNet, remains comparable. This can be attributed to the broad categorization of the price ranges. A significant performance drop would likely occur if a more precise definition of price ranges was used.

Results also show the importance of pretraining, improving MAE from 168.6 to 107.2 and R2 score from 0.54 to 0.75. Pretraining the model on a large amount of data (5.1 million) helped improve its performance. Based on this result, we decided to build the model using the pretraining step in all the following experiments.

Table 2 shows the performance of model input representations (PRp, RRp, PRRp, and Rp) on the test set. The representation giving the better results on our test set was the PRp representation (Using product fingerprint and reactants prices). RetroPriceNet_PRp achieved an R2 score of 0.75 and a mean absolute error of 107.2 \$/g.

The input representation with the lowest performance was the Rp representation, that uses reactants prices only, which appears not sufficient to predict the price accurately. However, surprisingly, the use of the reaction fingerprint, in the RRp and PRRp representations, did not help to improve the perform-

TABLE 1 | Comparison of RetroPriceNet performance on the test set (with similarity baseline, pretraining step and without pretraining step).

Metric	Similarity baseline	RetroPriceNet	RetroPriceNet pretrained
R2 score	0.18	0.54	0.75
Mean absolute error	212.9 \$/g	168.6 \$/g	107.2 \$/g
Accuracy on price ranges:	74.8%	78.4%	87.7%
Cheap [0 \$, 100 \$]			
Medium [100 \$, 500 \$]			
Expensive [500 \$, 1000 \$]			
Very expensive > 1000 \$			

TABLE 2 | RetroPriceNet performance on test set by input representation.

Model	R2 score	Mean absolute error
RetroPriceNet_Rp	0.37	205.9 \$/g
RetroPriceNet_RRp	0.57	162.1 \$/g
RetroPriceNet_PRp	0.75	107.2 \$/g
RetroPriceNet_PRRp	0.7	125.9 \$/g

ance, which we suppose is due to the lack of precision when it comes to the synthetic route. Synthetic routes are calculated using the CASP software Spaya, and even if the generated routes are well scored, there is no guarantee that these are the actual routes used to build the molecules. Using reaction information can therefore lead to additional noise that reduces performance.

We finally compared our best model RetroPriceNet_PRp to the state-of-the-art model CoPriNet on the same test set. As shown in Tables 3 and 4, CoPriNet has relatively good performance, but it does not surpass RetroPriceNet. When considering predictions in \$/gram (Table 3), RetroPriceNet demonstrated superior performance with a 0.07 point improvement in R2_score and an 18.7 \$/g advantage in terms of MAE compared to CoPriNet. Similarly, for predictions in \$/mmol (Table 4), RetroPriceNet outperformed CoPriNet with a 0.06 increase in R2_score and a 5.1 \$/mmol reduction in MAE. The subsequent analysis specifically focuses on the predictions in \$/gram.

To further examine the mean absolute error (MAE) of the predictors, we plotted the MAE as a function of molecule prices, categorized into 100 \$/g bins, as shown in Figure 6. The figure illustrates that for molecules priced between 0 \$/g and 600 \$/g, CopriNet demonstrates slightly better performance compared to RetroPriceNet. For molecules priced between 600 \$/g and 1200 \$/g, both models exhibit equivalent performance in MAE. However, for molecules with prices exceeding 1200 \$/g, RetroPriceNet significantly outperforms CopriNet.

The accuracy within a 10% margin of tolerated error for predicted prices was also evaluated across molecular price categories, divided into 100 \$/g increments, as shown in Figure 7. The figure demonstrates that both RetroPriceNet and CopriNet achieve comparable accuracy for molecules priced below 1200 \$/g. However, for molecules priced above 1200 \$/g, RetroPriceNet demonstrates superior accuracy.

From a broader perspective, it is evident that both models exhibit poor performance for both very inexpensive and very

TABLE 3 | CoPriNet results on test set in \$/gram.

Metrics	RetroPriceNetPre	CoPriNet
R2 score	0.75	0.68
Mean absolute error	107.2 \$/g	125.9 \$/g
Accuracy on price ranges:	87.7%	85.4%
Cheap [0 \$, 100 \$]		
Medium [100 \$, 500 \$]		
Expensive [500 \$,1000 \$]		
Very expensive > 1000 \$		
Inference time	5 ms + ~10 s for retrosynthesis	40 ms

TABLE 4 | CoPriNet results on test set in \$/mmol.

Metrics	RetroPriceNetPre	CoPriNet
R2 score	0.84	0.78
Mean absolute error	25.7 \$/mmol	30.8 \$/mmol

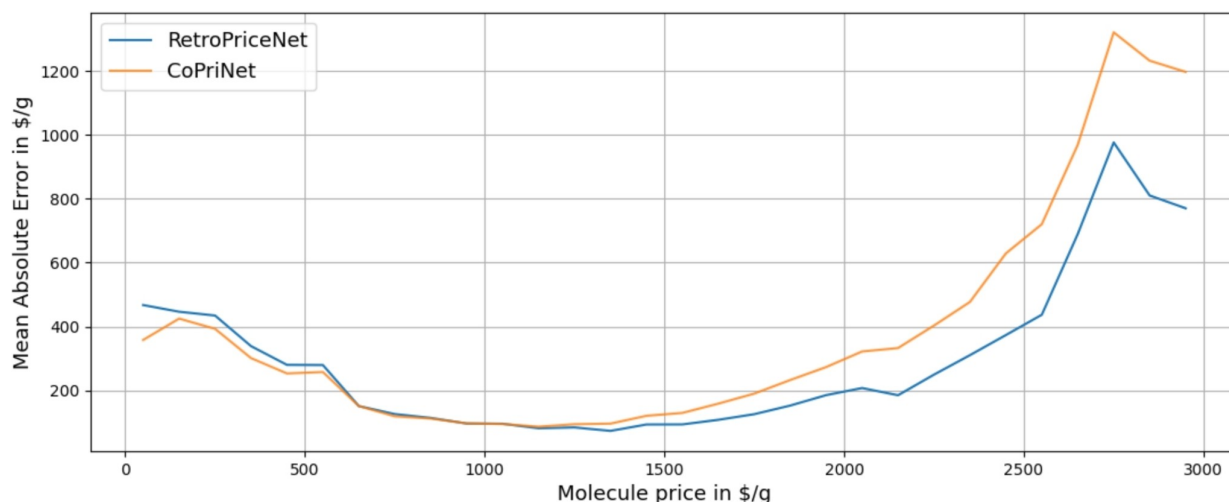


FIGURE 6 | Mean absolute error, by molecule prices categorized into \$100/g bins.

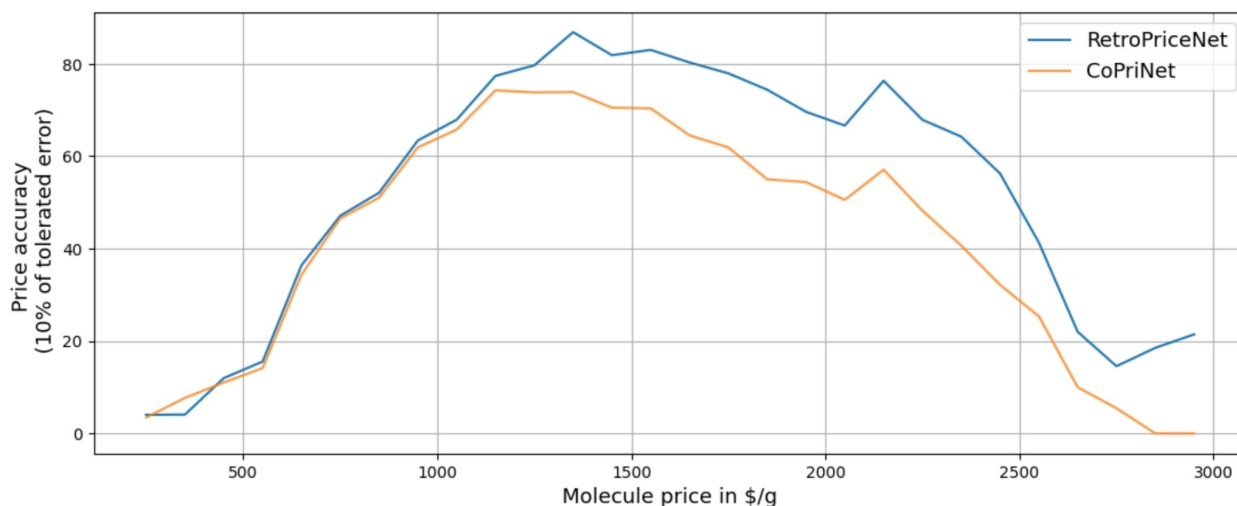


FIGURE 7 | Accuracy of predicted price with 10% of tolerated error, by molecule prices categorized into \$ 100/g bins.

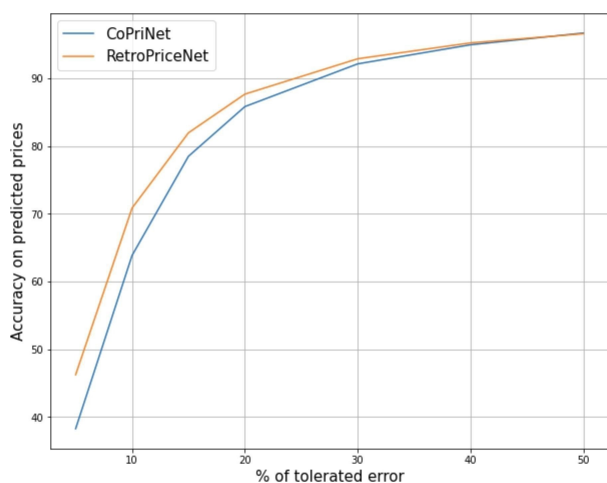


FIGURE 8 | Accuracy on predicted prices by percentage of tolerated errors.

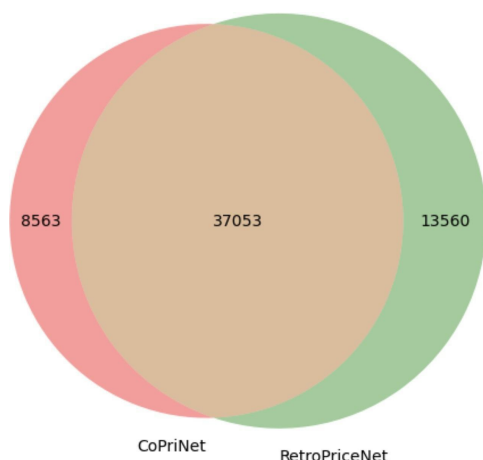


FIGURE 9 | Venn diagram on correctly predicted prices with RetroPriceNet and CoPriNet on the MSR test set. (We consider that the prediction is correct if the model price error is inferior to 10% of the price of the molecule.).

expensive molecules, with an accuracy falling below 20%. This phenomenon can be attributed to the under-representation of these molecules in the training dataset. Enhancing the model's performance on these molecules represents a potential area for future improvement.

We also computed the accuracy of the predicted prices by percentage of tolerated error. Figure 8 shows that RetroPriceNet achieves better accuracy when the tolerated error is low. With 5% of tolerated error, RetroPriceNet achieved 46% accuracy compared to 38% for CoPriNet and with 10% of tolerated error, RetroPriceNet achieved 70% accuracy compared to 63% for CoPriNet.

Figure 9 shows a Venn diagram presenting the number of molecules that are well predicted by both models with a tolerated error of 10% on the MSR test set (71 K molecules). The graph shows that both models predict an important percentage of the molecules well, however, there are some molecules that are well predicted by CoPriNet only and some molecules that are well predicted by RetroPriceNet only. For illustrative purposes, Figures 10 and 11 present some examples of those molecules. Figure 12 shows examples of badly predicted molecules by both models. Exploring the conditions under which each model demonstrates superior prediction capabilities is a potential avenue for further study, but it falls outside the scope of this paper.

We also analyzed RetroPriceNet predictions according to the depth of the molecule's synthetic pathway. Figure 13 shows that RetroPriceNet predicts better the price of molecules with short routes rather than deep routes. This can be explained by the accumulation of prediction errors along the route.

The main difference between CoPriNet and RetroPriceNet is that even if CoPriNet does not require a retrosynthesis step, which makes the inference time shorter (40 ms) compared to the average time of 10s required for retrieving a retrosynthetic route using Spaya API, not considering the synthetic route when predicting the price of molecules implies that the

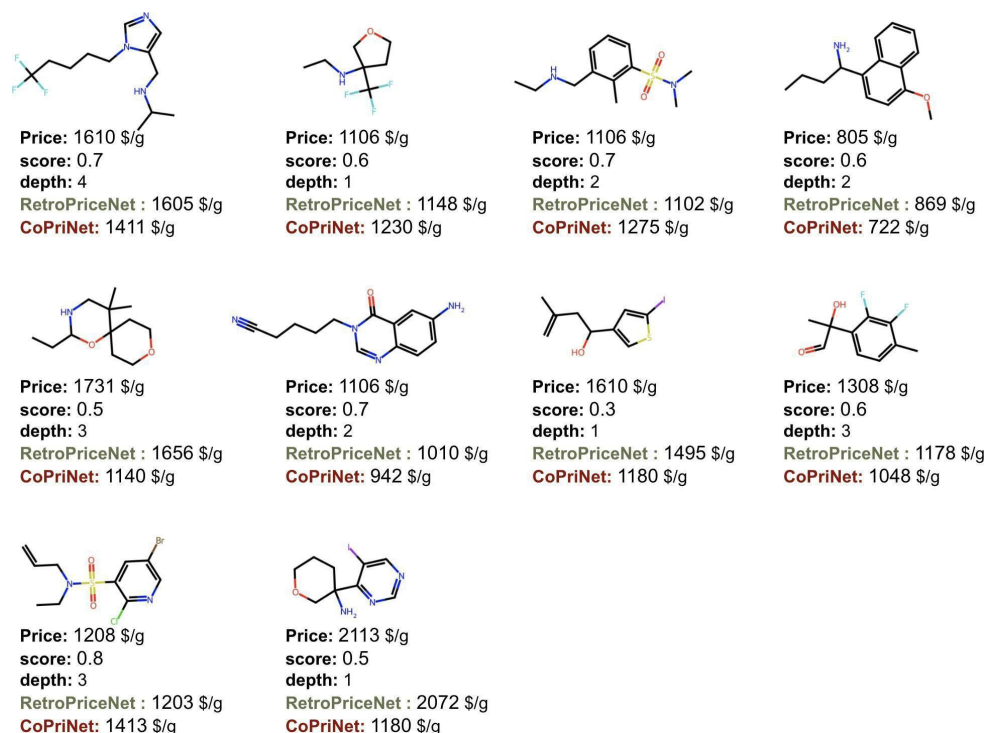


FIGURE 10 | Examples of well-predicted molecules with RetroPriceNet (<10% error).

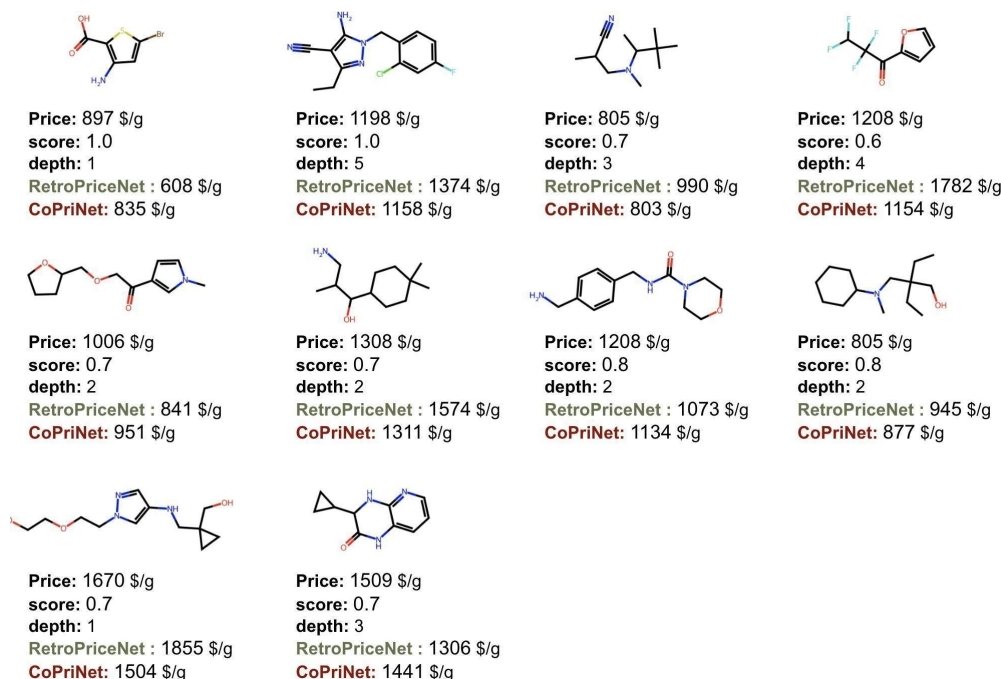


FIGURE 11 | Examples of well-predicted molecules with CoPriNet (<10% error).

prediction should be false for unfeasible molecules. In our analysis, we consider unfeasible molecules as molecules which fail retrosynthesis through Spaya without any specified conditions. Because price prediction is directly linked to molecule feasibility and synthesis complexity, we expect to predict huge prices for unfeasible molecules from a model based on molecular features only [7].

We performed tests on CoPriNet model by making small modifications on the Ranolazine molecule to make it not feasible through Spaya, and tried to predict its price using CoPriNet. Results in Table 5 and Figure 14 show that the predicted price for the unfeasible molecule is completely in the range of possible prices. In conclusion, one important drawback of building a price prediction model without factoring in the retrosynthesis route is that the computed models do not

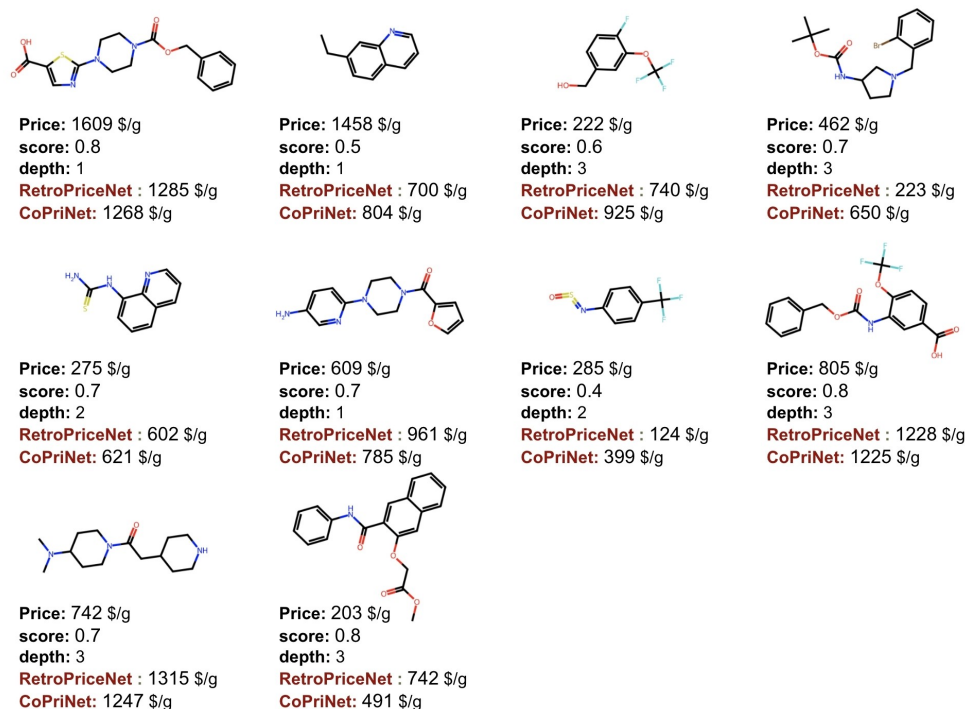


FIGURE 12 | Examples of badly-predicted molecules with CoPriNet and RetroPriceNet (<10% error).

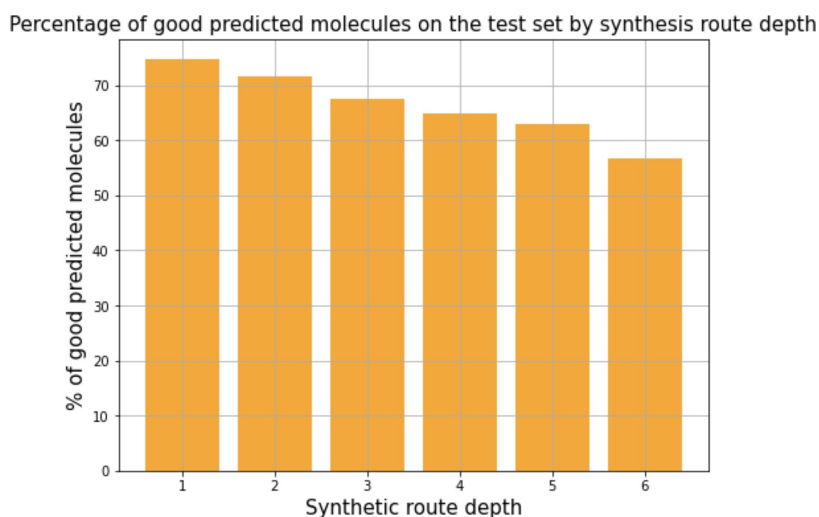


FIGURE 13 | Percentage of well-predicted molecules by tree depth (tolerated error of 10%).

distinguish between feasible and unfeasible molecules. However, it is important to note that relying on the retrosynthetic route provided by Spaya may have limitations for molecules that are in practice feasible but for which our predictive retrosynthesis tool Spaya cannot find any route. In such cases, RetroPriceNet cannot determine a price due to the absence of a synthetic pathway. However, Spaya is able to find synthetic routes for over 90% of randomly selected molecules from the ChEMBL database [21] (data not shown), therefore the likelihood of encountering such an issue in practice remains low.

In a broader context, future research could explore the phenomenon of price cliffs and examine how both models

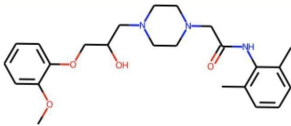
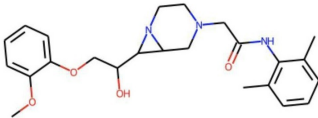
perform with very similar molecules that exhibit significantly different prices.

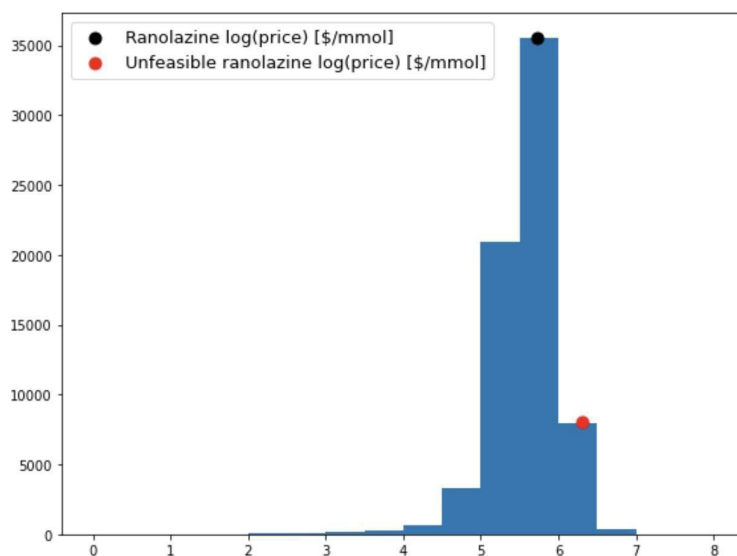
4 | Conclusion

This work presents the development of RetroPriceNet, a model that predicts the price of molecules in \$/g from their predicted synthetic routes. RetroPriceNet predicts step-by-step along the synthetic route until predicting the price of the input molecule.

In order to train RetroPriceNet, a database of commercial compounds with prices was built using data from chemicals suppliers and predicted synthetic routes obtained using Spaya

TABLE 5 | Prediction of unfeasible molecule using CoPriNet and RetroPriceNet.

	Molecule	CoPriNet Price (\$/g)	RetroPriceNet Price (\$/g)
Ranolazine		728	957
Unfeasible Ranolazine derivative		1284	No retrosynthesis route found

**FIGURE 14** | Distribution of prices of molecules in log(\$/mmol) in the MSR test set.

API retrosynthesis on those molecules. RetroPriceNet was trained on the database of commercial molecules with their price and predicted synthetic route. Two different training setups were tested: training directly RetroPriceNet on multi-step synthetic routes and pretraining RetroPriceNet on single-step synthetic routes before training on multi-step synthetic routes. The results showed the importance of the pretraining step to improve model performance. Different representations of the retrosynthesis step were also tested and the representation giving the best performances was PRp representation (product fingerprint and reactants prices). Our model RetroPriceNet was finally compared to CoPriNet, a state-of-the-art model in compound price prediction that predicts the price directly from the molecular graph in \$/mmol.

The best configuration of RetroPriceNet achieves superior performances on the MSR test set and is able to predict molecular prices more accurately than CoPriNet. In addition, RetroPriceNet, in contrast to CoPriNet, takes into account synthetic accessibility, starting materials availability and all other constraints that the user may decide to include in the retrosynthesis search (price, number of steps, key intermedi-

ates, and so on), which justifies the use of RetroPriceNet model in more demanding contexts. On the other hand, achieving this quality in the prediction requires more computational resources compared to CoPriNet. Those two approaches can be complementary to sort molecules by price.

Author Contributions

The study was designed by M. A., H. T. V.B. and Q.P. provided chemical expertise in the interpretation of model outputs. M. A., H.T, V. B. and Q.P. wrote the manuscript, which was critically reviewed by all authors. The author(s) read and approved the final manuscript.

Acknowledgments

The authors would like to thank IKTOS for having supported this study.

Conflicts of Interest

The authors are employees at IKTOS. The authors declare no competing interests in relationship with this manuscript.

Data Availability Statement

Spaya (<https://spaya.ai/>) is a Software as a Service (SaaS) platform freely accessible on the web and running on Iktos's secure Virtual Private Cloud (VPC) on Amazon Web Services (AWS). Iktos will integrate RetroPriceNet into Spaya, enabling users to access the technology via a graphical interface.

References

1. P. Gramatica, E. Papa, and A. Sangion, "QSAR Modeling of Cumulative Environmental End-Points for the Prioritization of Hazardous Chemicals," *Environmental Science: Processes & Impacts* 20.1 (2018): 38–47.
2. A. Anighoro and J. Bajorath, "Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes," *Journal of Chemical Information and Modeling* 56, no. 3 (2016): 580–587, <https://doi.org/10.1021/acs.jcim.5b00745>.
3. X. Barril, "Druggability Predictions: Methods, Limitations, and Applications," *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3, no. 4 (2013): 327–338.
4. M. Parrot, et al., "Integrating Synthetic Accessibility with AI-Based Generative Drug Design," *Journal of Cheminformatics* 15 (2023): 83, <https://doi.org/10.1186/s13321-023-00742-8>.
5. P. Bonnet, "Is Chemical Synthetic Accessibility Computationally Predictable for Drug and Lead-Like Molecules? A Comparative Assessment Between Medicinal and Computational Chemists," *European Journal of Medicinal Chemistry* 54 (2012): 679–689, <https://doi.org/10.1016/j.ejmech.2012.06.024>.
6. T. Badowski, K. Molga, and B. A. Grzybowski, "Selection of Cost-Effective Yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans," *Chemical Science* 10 (2019): 4640–4651, <https://doi.org/10.1039/C8SC05611K>.
7. R. Sanchez-Garcia, et al., "CoPriNet: Graph Neural Networks Provide Accurate and Rapid Compound Price Prediction for Molecule Prioritization," *Digital Discovery* 2 (2023): 103–111, <https://doi.org/10.1039/D2DD00071G>.
8. T. J. Struble, et al., "Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis," *Journal of Medicinal Chemistry* 63, no. 16 (2020): 8667–8682, <https://doi.org/10.1021/acs.jmedchem.9b02120>.
9. J. Dong, et al., "Deep Learning in Retrosynthesis Planning: Datasets, Models and Tools," *Briefings in Bioinformatics* 23, no. 1 (2022): bbab391, <https://doi.org/10.1093/bib/bbab391>.
10. C. W. Coley, et al., "A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning," *Science* 365, no. 6453 (2019): eaax1566, <https://doi.org/10.1126/science.aax1566>.
11. S. Genheden, et al., "AiZynthFinder: a Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning," *Journal of Cheminformatics* 12 (2020): 70, <https://doi.org/10.1186/s13321-020-00472-1>.
12. S. Szymkuć, et al., "Computer-Assisted Synthetic Planning: the End of the Beginning," *Angewandte Chemie International Edition* 55, no. 20 (2016): 5904–5937, <https://doi.org/10.1002/anie.201506101>.
13. ChemAIRS: Advanced Reaction Informatics Platform in the Industry. 2024, URL: <https://www.chemical.ai/>.
14. Iktos CASP Freely Available at Spaya: AI-Driven Retrosynthesis Platform. 2024: <https://spaya.ai/> and SpayaAPI <https://iktos.ai/spaya-api/>.
15. Segler, M. H. S., M. Preuss, and M. P. Waller. "Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI." in 555.7698 (2018): 604–610, <https://doi.org/10.1038/nature25978>.
16. Pistachio: Reaction data, Querying and Analytics. 2024. URL: <https://www.nextmovesoftware.com/pistachio.html>.
17. D. Weininger, "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences* 28, no. 1 (1988): 31–36, <https://doi.org/10.1021/ci00057a005>.
18. A. P. Bento, et al., "An Open Source Chemical Structure Curation Pipeline Using RDKit," *Journal of Cheminformatics* 12 (2020): 1–16, <https://doi.org/10.1186/s13321-020-00456-1>.
19. D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to Multi-Layer Feed-Forward Neural Networks," *Chemometrics and Intelligent Laboratory Systems* 39, no. 1 (1997): 43–62, [https://doi.org/10.1016/S0169-7439\(97\)00061-0](https://doi.org/10.1016/S0169-7439(97)00061-0).
20. D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling* 50, no. 5 (2010): 742–754, <https://doi.org/10.1021/ci100050t>.
21. A. Gaulton, et al., "The ChEMBL Database in 2017," *Nucleic Acids Research* 45, no. D1 (2017): D945–D954, <https://doi.org/10.1093/nar/gkw1074>.