Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Computational Assessment of Molecular Synthetic Accessibility using Economic Indicators

Friedrich Hastedt[a], Klaus Hellgardt[b], Sophia Yaliraki[a], Antonio del Rio Chanona[a*], Dongda Zhang[b*]

[a] Imperial College London, Exhibition Rd, South Kensington, London SW7 2BX, United Kingdom
[b] University of Manchester, Oxford Rd, Manchester M13 9PL, United Kingdom
* Corresponding Author: a.del-rio-chanona@imperial.ac.uk, dongda.zhang@manchester.ac.uk

ABSTRACT

The rapid advancement of computational drug discovery has enabled the generation of vast virtual libraries of promising drug candidates. However, evaluating the synthetic accessibility (SA) of these compounds remains a critical bottleneck. While computer-aided synthesis planning (CASP) tools can provide synthesis routes to the candidate, their computational demands make them impractical for large-scale screening. Existing rapid SA scoring methods, struggle to generalize to out-of-distribution molecules and do not account for economic viability. To address these challenges, we present MolPrice, an accurate and reliable price prediction tool. By introducing a novel self-supervised learning approach, MolPrice achieves robust generalization to diverse molecular structures of various complexities. Our comprehensive analysis of model architectures and molecular representations reveals that substructure-based features strongly correlate with market prices, supporting the relationship between synthetic complexity and economic value. MolPrice performs well on the standard literature SA benchmark, showcasing its ability for SA estimation. MolPrice thus serves as both an accurate molecular price predictor and a rapid synthetic accessibility assessment tool, enhancing the efficiency of modern drug discovery pipelines.

Keywords: Molecular Complexity, Machine Learning, Synthetic Accessibility, Virtual Screening, Retrosynthesis

INTRODUCTION

Generating promising compounds for drug discovery has become increasingly efficient. With the emergence of several open-source tools [1], one can design a whole library of thousands/millions of molecules with desired physical properties in hours.

However, taking these molecules from the virtual library to laboratory testing is a major challenge. This is because the virtual compound needs to be synthesized first, which is costly and time-consuming. Finding synthesis pathways from known building blocks to virtual molecules is a laborious process known as retrosynthesis. Within the past decade, researchers have developed computer-aided synthesis planning tools (CASP) [2] that automate retrosynthetic planning. Whilst these tools are powerful, they are still too slow (~0.5 minutes per molecule). As molecular design routines require a large number of molecular evaluations, CASP tools do not scale [3]. For example, to screen a library of 1 billion molecules within hours, the time scale should be in ms per molecule.

To overcome this issue, researchers have developed efficient scoring systems that can evaluate molecules for *synthetic accessibility* (SA) within milliseconds. The scoring systems classify molecules as either *easy-to-synthesize* (ES) or *hard-to-synthesize* (HS). ES molecules are proceeded for further analysis and HS are usually discarded. Calculated SA scores are based on i) complexity-based indicators [4,5] or ii) retrosynthetic analysis [3,6]. Complexity-based models calculate a score depending on structural indicators such as the presence/absence of functional groups and the number of stereocentres in the molecule [4]. Retrosynthetic-based models predict the output of the slower CASP tool. For example, *Thakkar et al.* [6] attempt to predict whether the CASP tool can successfully find a synthesis route to the molecule. Whilst both methodologies have their merits, they face a significant limitation: the inability to generalize to out-of-distribution molecules. Furthermore, the scoring system is either based on binary classification or

a user-defined scale, lacking physical interpretation.

Inspired by *Sanchez-Garcia et. al.* [7], we propose to assess the synthetic accessibility of a molecule based on its market price. The authors were the first to draw a connection between SA and molecular price. The price prediction model (CoPriNet) presented in [7] accurately recalls the price of purchasable, ES molecules. Similar to other scoring systems, their model fails to generalize to out-of-distribution molecules. In particular, the model is prone to assign a similar price range to both ES and HS molecules. In short, their model cannot distinguish molecular complexity and cannot be used for SA scoring.

This work aims to overcome these challenges by introducing MolPrice, a fast predictive model for molecular price. Using a new self-supervised training methodology, MolPrice learns to distinguish purchasable (ES) molecules from hard-to-synthesize (HS), out-of-distribution molecules. Through systematic analysis of different model architectures and molecular representations, we demonstrate that features capturing substructures strongly correlate with market prices, supporting the fact that synthetic complexity is reflected in economic value. MolPrice can thus be used as an initial starting point for economic analysis (market price of product – reactants).

Finally, MolPrice can be used as an accurate tool for molecular price prediction and as an SA estimation for virtual screening or molecular generation.

METHODOLOGY

Model Development

Data Source

The data was provided by Molport as a snapshot of their purchasable database from March 2024. The database consists of over 5M molecules with their corresponding price in USD per unit weight. Molport provides several prices (by different suppliers) for a range of weights. The first step to obtain a robust dataset is to find a single price value per molecule. We follow the steps below to obtain this database:

1. Molecules that cannot be read by RDKit [8] are filtered out.
2. For each molecule, all corresponding prices are converted from USD per weight to USD per mmol.
3. For each molecule, the lowest price is obtained in USD per mmol.
4. The prices are converted to a logarithmic scale for model training [7].
5. Finally, molecules with a price < 2 USD/mmol on the logarithmic scale are removed (these are mostly stabilizing salts in solution or metals)

The final dataset is plotted in Figure 1.

We acknowledge that over time new molecules may be added to the database. In that case, one can simply retrain the model in a matter of minutes (as outlined in Results and Discussion).

For self-supervised learning, the HS molecules dataset is obtained from *Yu et. al.* [9] consisting of 400K molecules.

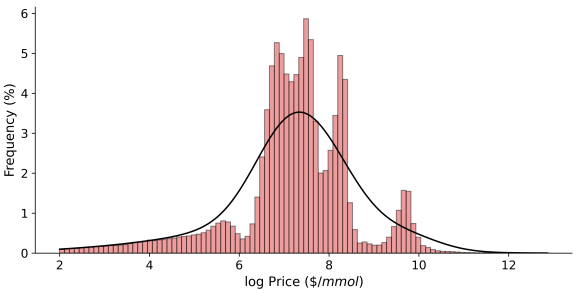


Figure 1: Distribution of price dataset

Model Details

One can use a variety of models for molecular property prediction, the best depending on the specific use case. We screened different molecular featurization schemes and model architectures to select the most accurate and robust model. Particularly, in case of molecular featurization, it is interesting to explore the features that capture the relationship between price and molecule best. This way, we are posing the question whether functional groups or larger molecular moieties determine the economic value of a compound. Below we outline the tested model featurization:

- **Molecular Fingerprints** encode the 2-dimensional molecule in a bit vector. Each bit in the vector indicates the presence or absence of a molecular substructure. Various fingerprint generation algorithms exist. We select four popular algorithms: Morgan (ECFP), Atom-pair, Topological torsion (all three provided by RDKit) and MinHash (Mhfp) [10]. Morgan and Mhfp focus on the circular environment around atoms. Atom-pair considers pairs of atoms and the shortest path between them. Topological torsion considers sequential patterns. One could argue that Morgan/Mhfp best represent functional groups (FGs), atom-pair captures the overall shape of the molecule and topological torsion captures the backbone patterns. The reader is referred to the sources for more detailed information [8,10]. The various fingerprints are then used as input to a multilayer perceptron (MLP).
- **2D Graphs** include the connectivity between atoms in our modelling approach, we explore the direct use of the 2D molecular graph. The graph

features include descriptors for nodes (atom) and edge (bond) features. We use the graph model presented in [6] as the graph neural network (GNN) and as the literature baseline within the results.

- **SMILES** is a text-based representation of a molecule. We explore both the vanilla Transformer and a pre-trained RoBertA model by DeepChem [11]. As a note, the SMILES is tokenized per atom basis.
- **Functional Groups** are features that are constructed through the extended functional groups (EFG) algorithm [12]. Intuitively, EFGs are substructures in the molecule characterizing functionalization such as alcohols, ketones, amides. As molecules have different number of EFGs (different input length), a long-short term memory (LSTM) is a natural modelling choice. As an alternative, one could also construct an EFG fingerprint and use an MLP. This will be investigated in the future.

All models are trained on the MSE loss:

$$L = \frac{1}{M}\sum(y_i - \hat{y}_i)^2, \quad (1)$$

where y_i and $\hat{y}$ are the ground-truth labels and model output, respectively and M is the dataset size. The regression models are compared using mean squared error (MSE), coefficient of determination (R^2), Spearman's correlation coefficient (r_s) and most importantly, inference time. The selection of time as an evaluation criterion follows naturally from previous arguments; the purpose of *SA* scoring systems is to perform large numbers of evaluations. The Molport dataset is split into 90/5/5 portions for training/validation/testing. This leaves ~500k molecules for validation and testing.

Self-Supervised Learning

Following model training, the model is still unlikely to generalize to unseen, complex molecules. This is because all molecules in our database are purchasable or in other words, easy to synthesize. Molecules which are not synthesizable (HS) generally will not have a price label since they are not offered by suppliers.

To allow our model to learn price labels for the HS molecules, we use a contrastive learning approach. Particularly, the approach is based on two principles:

First, assuming the model has learnt a good correlation between the *latent* molecule representation and the price, we only need to focus on learning the *latent* representation of the HS molecules. Herein, we define the *latent* representation as the final vector in the model, prior to the readout function. The readout function *f* is defined as $f(\mathcal{R}^N) \rightarrow \mathcal{R}$, that is the function that converts the *latent* representation into the price output. A visual explanation is shown in Figure 2. Following on with this assumption, we freeze the parameters (weights) in the readout function during self-supervised learning. In other words, the only parameters that can be updated during the training, are the parameters that compute the *latent* representation, prior to the readout function.

Second, we assume that each dimension of the *latent* representation follows a Gaussian distribution. The assumption is reasonable as the price distribution of the model output (Figure 3a) is approximate Gaussian (by visual inspection). Using these two principles, we define the loss function for self-supervised learning as:

$$L = \frac{1}{M}\sum_{i\in ES}(y_i - \hat{y}_i)^2 + \lambda \cdot \frac{1}{N}\sum_{j=0}^{N} H_j, \quad (2)$$

$$H_j = \sqrt{\frac{2\sigma_{ES,j}\sigma_{HS,j}}{{\sigma_{ES,j}}^2 + {\sigma_{HS,j}}^2}}\exp\left(-1/4\frac{(\mu_{ES,j}-\mu_{HS,j})^2}{{\sigma_{ES,j}}^2 + {\sigma_{HS,j}}^2}\right), \quad (3)$$

where H_j is the analytical Hellinger distance computed $\forall j \in \{1, \ldots, N\}$ with *N* being the *latent* vector size for Gaussian distributions. In short, the Hellinger distance enforces a separation in the *latent* representation of ES molecules versus HS molecules. The Hellinger distance tends to a minimum of zero when the ES distribution $\mathcal{N}(\mu_{ES,j}, \sigma_{ES,j})$ does not overlap with the HS distribution $\mathcal{N}(\mu_{HS,j}, \sigma_{HS,j})$. Since the parameters of the readout function are not updated during training, a different *latent* vector representation for the HS molecules should lead to a different price. The first term in Eq. 2 is the MSE loss computed over the set of ES molecules and therefore acts as a regularisation term. λ is a hyperparameter and was set to a low value of 0.05 by default. By scaling the loss of the Hellinger distance, we do not permit the model to perform poorly on the initial price prediction task. Intuitively, if the two loss terms in Eq. 2 are not conflicting, we should obtain an accurate model for price prediction as well as a model that can distinguish ES from HS molecules.

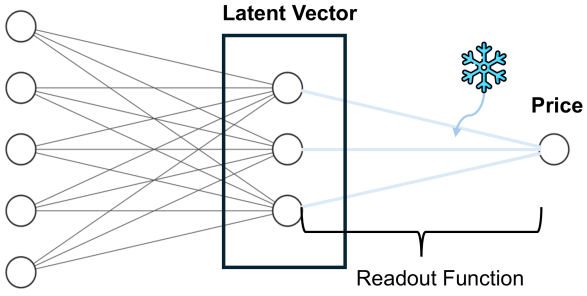


Figure 2: Toy example with a latent vector of 3D (N=3). Parameters of the readout functions are frozen.

It is to be noted that there is no theoretical guarantee that the predicted price for the HS molecules will be larger than the predicted price of the ES molecules. This

Table 1: Results for different model architectures and molecular featurization schemes on the hold-out test set. The metrics include mean squared error (MSE), coefficient of determination (R^2), Spearman's correlation coefficient (r_s) and inference time in ms (per single cpu core per molecule)

Architecture	Feature/Model Type	MSE (↓)	R^2 (↑)	r_s (↑)	Inference Time (ms/mol)
MLP	Atom pair	0.58	0.72	0.84	1.4
	Topological	0.61	0.71	0.84	1.4
	Morgan	0.51	**0.76**	**0.88**	**1.4**
	Mhfp	**0.50**	**0.76**	**0.88**	**1.4**
GNN (CoPriNet)	2D Graph	0.52	0.74	0.85	21.5
Transformer	SMILES (Vanilla)	0.88	0.58	0.72	41.1
	SMILES (RoBERTa)	0.58	0.72	0.85	29.1
LSTM	EFGs	0.62	0.70	0.82	3.6

is because we do not incorporate price labels for the HS molecules and solely rely on the contrastive loss (Eq. 2)

To verify that MolPrice has effectively learnt the difference between ES and HS molecules, we compare its performance to classical SA scoring systems on the test dataset TS1-3 provided in [9], the standard for comparing SA scoring tools.

RESULTS AND DISCUSSION

Price Prediction

The results for price prediction on the test set are shown in Table 1. Analysing the performance metrics, the neural network (MLP) with Morgan/Mhfp featurization performs best in terms of MSE, R^2 and r_s. Particularly, the Morgan/Mhfp featurization is preferred over the two other fingerprints tested (Atom pair and Topological). This can be explained by considering the fingerprint construction algorithm: Both Morgan/Mhfp fingerprints are constructed through circular environments around atoms in the molecule. This featurization resembles functional group (substructure) encoding. On the other hand, atom-pair and topological torsion fingerprints consider the sequential order of atoms in a molecule and fail to fully capture fine-grained details such as functional groups.

The GNN model (CoPriNet) is seen to nearly match the fingerprint MLP models. Possibly this can be explained by comparing the internal model workings: Particularly, the message-passing operation commonly used in GNNs is similar to the circular encoding algorithm (for fingerprint construction). Thus, both model architectures share a certain degree of knowledge.

As expected, the pre-trained Transformer (RoBERTa) model performs better than its vanilla counterpart. The pre-training allows the model to learn the SMILES language prior to the downstream prediction task. Nevertheless, both SMILES-based models fail to outperform the GNN or fingerprint MLP. As SMILES are tokenized per atom basis, the model is not aware of molecular substructures, a piece of information available to both the GNN and MLP.

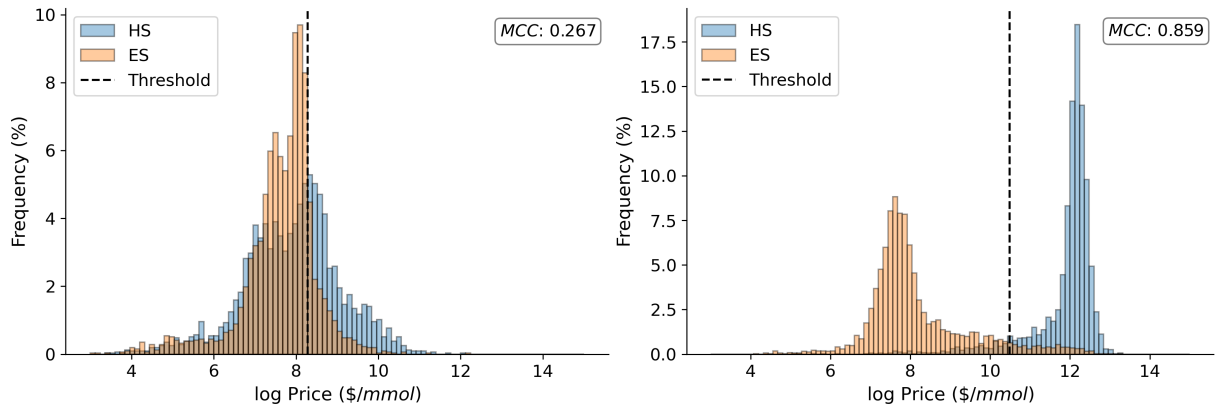
Finally, the extended functional group (EFGs) LSTM is seen to perform acceptably. We believe that the LSTM fails to construct a good latent representation of the molecule due to the sequential input of the EFGs. Instead of using an LSTM, it may be worth constructing a new fingerprint based on EFGs and using an MLP as the model.

Last, a very important factor not yet discussed is the model's inference time. For example, if one wants to screen a database of 1 billion molecules and the scoring takes 1 second per molecule, it would take ~30 years to screen the entire database. From Table 1, we see that the MLP is favourable with a quick inference time of 1.4 ms per molecule (tested on an Intel vPro i9 laptop cpu). The same database could now be screened within 12 days. Both the literature GNN (CoPriNet, literature model) and Transformer models take 20-40x longer for the same task. This is because both models perform a large number of matrix multiplication operations, whereas the MLP only performs 3 overall. The training time of the MLP lies within 30 minutes, whereas both the GNN and Transformer model takes several days, rendering training on a new dataset computationally intensive.

The detailed analysis of different molecular featurization highlights that features capturing substructures correlate well with market prices (Morgan, Mhfp, GNN). This finding underscores the relationship between molecular complexity and market price.

Learning Synthetic Accessibility

The top two performing models (Mhfp and Morgan) provide the starting point for self-supervised learning. Figure 3 shows the predicted price for both ES and HS molecules before and after self-supervised learning for the Morgan fingerprint. For both Morgan and Mhfp fingerprints, the algorithms learn to assign prices to HS molecules that are different to the ES, purchasable molecule. Empirically, we observe that the model prefers to assign a higher price range to the HS molecules compared to the ES molecules. This is desired, but as outlined previously,



a: Before self-supervised learning b: After self-supervised learning

Figure 3: Distributions of ES/HS molecules predicted by MolPrice for Morgan fingerprint. The degree of distribution separation is measured through the Matthews correlation coefficient (MCC).

there is no mathematical guarantee for this observation.

Statistically, the separation between the ES and HS distribution is confirmed in the table below:

Table 2: Statistics on HS/ES classification and Price

Fingerprint	**MCC** (↑)	**AUC** (↑)	$\mathbf{R^2}$ (↑)
Morgan	0.86	0.97	0.54
Mhfp	0.89	0.98	0.73

In Table 2, the distribution separation is measured using Matthew's correlation coefficient (MCC) and area under receiver operating characteristic curve (AUC). These two metrics indicate how well MolPrice can distinguish between ES and HS molecules (a binary classification problem). The Mhfp fingerprint slightly outperforms the Morgan fingerprint in both metrics.

More importantly, for the Morgan fingerprint, the MSE loss and Hellinger distance separation were seen to have conflicting objectives. To obtain a higher AUC, the hyperparameter λ in Eq. 2 was set to a higher value. This came with a decrease in the accuracy of price prediction for the ES dataset as shown by a lower R^2 compared to Table 1. For the Mhfp fingerprint, the R^2 only slightly decreased from 0.76 to 0.73. From Table 2, we can deduct that the MLP along with the Mhfp fingerprint results in an accurate price prediction as well as a SA scoring tool.

Comparison to other SA scoring tools

To confirm that MolPrice is indeed a competitive SA scoring tool, we benchmarked and compared MolPrice to other state-of-the-art SA tools. This analysis is shown in Table 3. From the table, it is seen that MolPrice performs competitively compared to other SA screening tools. Only GASA outperforms MolPrice on the test datasets apart from TS3 – accuracy. While MolPrice is not the top performing model, two things should be kept in mind when analysing the results:

1. GASA and other models were exclusively trained for SA scoring, *i.e.*, the only objective given to the model is the binary classification between ES and HS molecules. MolPrice has to balance between predicting accurate labels for the price as well as learning the difference between the two molecule groups
2. The test sets TS2 and TS3 were curated by using a CASP tool. In short, the HS molecules are labelled as hard-to-synthesize if the CASP tool fails to return a synthesis route for the molecule. It is to note that CASP tools are not perfect and highly depend on the database that it was trained on. Given this, it could be that some molecules that were labelled as ES are

Table 3: Comparison of our model MolPrice to other, commonly used SA tools. The test datasets are obtained from [9]. For TS1-3, the area under receiver operating characteristic curve (AUC) is reported as well as the accuracy for TS3. The best performance is highlighted in bold, the second best in italics.

Model	**TS1 (AUC)**	**TS2 (AUC)**	**TS3 (AUC)**	**TS3 (Acc)**
GASA [9]	**0.999**	*0.881*	**0.850**	*0.759*
SCScore [5]	0.881	0.487	0.575	0.511
SAScore [4]	0.989	**0.919**	0.772	0.664
RAScore [6]	0.919	0.865	0.790	0.701
MolPrice (ours)	*0.991*	0.871	*0.833*	**0.771**

actually HS and vice versa.

Given above, it is fair to conclude that MolPrice performs well on these benchmarks. We plan to extend SA testing in future works on case studies as outlined in the next section to solidify the use case of MolPrice of molecular generation and screening.

CONCLUSION AND OUTLOOK

In this work, we present MolPrice, an accurate tool for both price prediction and synthetic accessibility measurement. MolPrice is based on the Mhfp fingerprint and a fast-to-evaluate MLP, which were carefully selected after exploring a range of model architectures and molecule featurization schemes. Through this exploration, we discovered the importance of molecular substructures for predicting prices, supporting the notion that molecular complexity is reflected in economic value.

By introducing a new contrastive learning approach, we steer MolPrice towards molecules that are out-of-distribution and/or complex.

In the future, we will test MolPrice on different case studies. These case studies will range from multi-objective virtual screening [13] to molecular generation such as in [3]. Furthermore, we want to include MolPrice directly as guidance in retrosynthesis planning. In short, we are interested to see whether MolPrice can steer the synthesis search towards economically viable routes and/or speed up the convergence of the CASP tool.

DIGITAL SUPPLEMENTARY MATERIAL

MolPrice is hosted on GitHub and usable through https://github.com/fredhastedt/MolPrice

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) funding grant EP/S023232/1.

We thank Molport (https://www.molport.com/) for sharing their private data used for model training.

REFERENCES

1. Du Y, Jamasb AR, Guo J, et al. Machine learning-aided generative molecular design. *Nat Mach Intell*. 2024;6(6):589-604. doi:10.1038/s42256-024-00843-5
2. Coley CW, Green WH, Jensen KF. Machine Learning in Computer-Aided Synthesis Planning. *Acc Chem Res*. 2018;51(5):1281-1289. doi:10.1021/acs.accounts.8b00087
3. Liu CH, Korablyov M, Jastrzębski S, Włodarczyk-Pruszyński P, Bengio Y, Segler M. RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software. *J Chem Inf Model*. 2022;62(10):2293-2300. doi:10.1021/acs.jcim.1c01476
4. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminformatics*. 2009;1(1):8. doi:10.1186/1758-2946-1-8
5. Coley CW, Rogers L, Green WH, Jensen KF. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J Chem Inf Model*. 2018;58(2):252-261. doi:10.1021/acs.jcim.7b00622
6. Thakkar A, Chadimová V, Bjerrum EJ, Engkvist O, Reymond JL. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem Sci*. 2021;12(9):3339-3349. doi:10.1039/D0SC05401A
7. Sanchez-Garcia R, Havasi D, Takács G, et al. CoPriNet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation. *Digit Discov*. 2023;2(1):103-111. doi:10.1039/D2DD00071G
8. Landrum G. RDKit: Open-source chemoinformatics. https://www.rdkit.org/docs
9. Yu J, Wang J, Zhao H, et al. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *J Chem Inf Model*. 2022;62(12):2973-2986. doi:10.1021/acs.jcim.2c00038
10. Probst D, Reymond JL. A probabilistic molecular fingerprint for big data settings. *J Cheminformatics*. 2018;10(1):66. doi:10.1186/s13321-018-0321-8
11. Ramsundar B, Eastman, Walters P, Pande V, Leswing K, Wu Z. *Deep Learning for the Life Sciences*. O'Reilly Media
12. Lu J, Xia S, Lu J, Zhang Y. Dataset Construction to Explore Chemical Space with 3D Geometry and Deep Learning. *J Chem Inf Model*. 2021;61(3):1095-1104. doi:10.1021/acs.jcim.1c00007
13. Fromer JC, Graff DE, Coley CW. Pareto optimization to accelerate multi-objective virtual screening. *Digit Discov*. 2024;3(3):467-481. doi:10.1039/D3DD00227F