JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

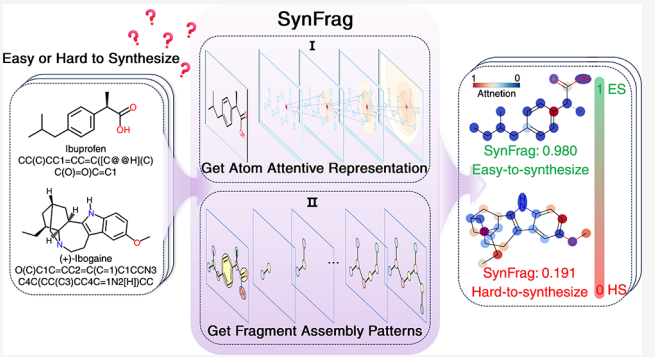
SynFrag: Synthetic Accessibility Predictor Based on Fragment Assembly Generation in Drug Discovery

Xiang Zhang, Jia Liu, Bufan Xu, Zihan Zhang, Zifu Huang, Kaixian Chen, Dingyan Wang,* and Xutong Li*

Cite This: https://doi.org/10.1021/acs.jcim.5c02450 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: AI-driven molecular generation encounters a "generation-synthesis gap": most computationally designed molecules cannot be synthesized in laboratories, limiting AI-assisted drug design (AIDD) applications. Current approaches to assess synthetic accessibility (SA) include computer-aided synthesis planning (CASP) tools that perform retrosynthetic searches and machine learning-based SA prediction models that provide rapid scoring. CASP tools are computationally expensive for high-throughput screening, while existing SA prediction models may lack chemical synthesis logic or exhibit variable performance across different chemical spaces. We developed SynFrag, an SA prediction model using fragment assembly autoregressive generation to learn stepwise molecular construction patterns. Self-supervised pretraining on millions of unlabeled molecules enables the learning of dynamic fragment assembly patterns beyond fragment occurrence statistics or reaction step annotations. This approach captures connectivity relationships relevant to synthesis difficulty cliffs, where minor structural changes substantially alter SA. Evaluation across public benchmarks, clinical drugs with intermediates, and AI-generated molecules shows consistent performance across diverse chemical spaces. The model produces subsecond predictions with attention mechanisms corresponding to key reactive sites. SynFrag provides computational efficiency suitable for large-scale screening while maintaining interpretability for detailed SA assessment in drug discovery workflows. Online platform: https://synfrag.simm.ac.cn. Code and data available: https://github.com/simmzx/SynFrag.



1. INTRODUCTION

AI-driven molecular generation has advanced in exploring chemical spaces and designing compounds with desired properties.[1−3] However, a "generation-synthesis gap"[4] limits practical impact: computationally designed molecules often cannot be synthesized in laboratories. Generative models produce varying proportions of molecules amenable to synthesis using existing methodologies,[5] constraining practical implementation in AI-assisted drug design (AIDD). Synthetic accessibility (SA) prediction serves as a bridging approach to address this bottleneck in AIDD.

SA is a multidimensional concept encompassing structural complexity, reaction feasibility, yield, and reagent availability. Current SA prediction approaches fall into two categories. Computer-aided synthesis planning (CASP) methods, including ASKCOS[6] and AiZynthFinder,[7] perform retrosynthetic searches based on MCTS or A* algorithm but require substantial computational time (2−6 min per molecule) and depend on reaction template databases.[8] Machine learning approaches employ various strategies: SAscore[9] uses fragment frequency statistics from drug databases without considering interfragment connectivity. SCScore[10] learns relative complexity using neural networks trained on schemes (sequences of reactions). SYBA[11] applies naive Bayes classification based on fragment occurrence. RAscore[8] and GASA[12] train neural network and graph attention-based model using binary labels derived from CASP results. DeepSA[13] treats SMILES as chemical language using pretrained BERT. BR-SAscore[14] incorporates the fingerprints of reaction centers and building blocks to enhance fragment information.

However, existing approaches mostly face three limitations in drug discovery workflows. First, fragment-level statistical methods, such as SAscore and SYBA, rely on fragment occurrence frequencies without capturing assembly patterns. This limits the ability to identify synthesis difficulty cliffs, as patterns may be implicit across fragments. Second, supervised methods depend on binary labels from CASP tools, leaving

Received: October 9, 2025
Revised: February 1, 2026
Accepted: February 25, 2026

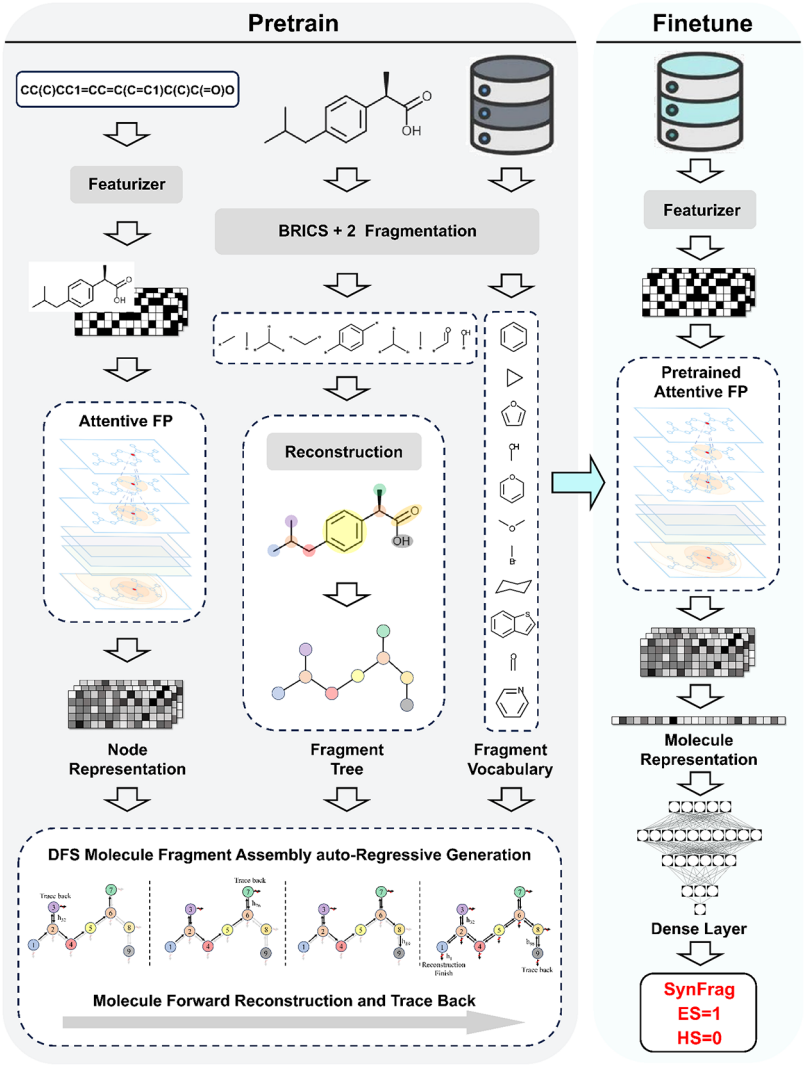


Figure 1. SynFrag model architecture Left (Pretraining): Molecules are converted to graphs and processed through AttentiveFP to obtain atom and fragment representations. BRICS+2 fragmentation decomposes molecules into fragments, constructing fragment trees with fragments as nodes and broken bonds as edges. DFS-based autoregressive generation trains dual predictors (topology and label predictors) to reconstruct molecules by sequential prediction of fragment assembly. Right (Finetuning): Pretrained AttentiveFP is fine-tuned on labeled SA data for easy-to-synthesize (ES) and hard-to-synthesize (HS) classification. Output range [0, 1]: values near 1 indicate ES; values near 0 indicate HS.

large chemical databases underutilized for learning chemical principles. Third, current methods show variable performance[15] on AI-generated molecules or cannot discriminate between intermediates in multistep syntheses[16]—both relevant for evaluating molecular design outputs and optimizing synthetic routes.

We developed SynFrag, a model using fragment assembly autoregressive generation pretraining (Figure 1). SynFrag decomposes molecules into fragments and reconstructs them step-by-step through autoregressive generation, learning stepwise construction patterns, and is expected to simulate how chemists think in assembling building blocks in synthesis. The pretrain-finetune framework uses 9.18 million unlabeled molecules with depth-first-search (DFS) assembly sequences to capture fragment connectivity patterns in generation processes. This approach provides subsecond predictions suitable for high-throughput screening, maintains performance across diverse chemical spaces, including drug molecules and AI-generated structures, and serves interpretable visualizations.

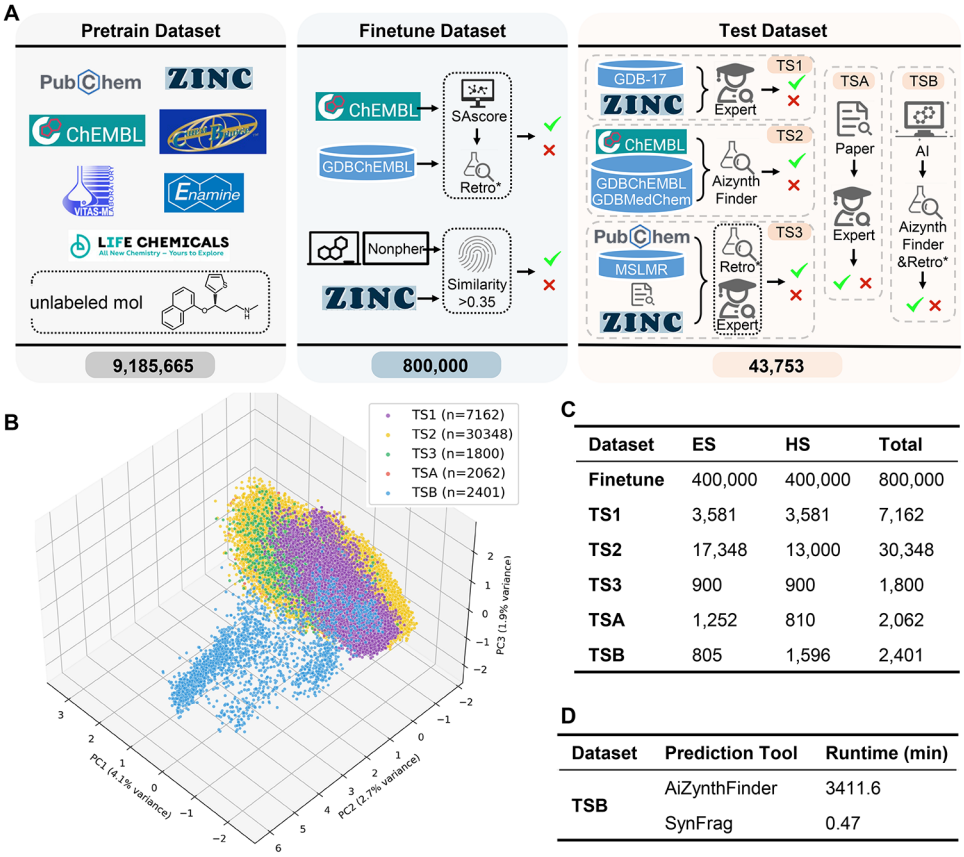
The main contributions include: (1) developing SynFrag with fragment assembly strategy that captures synthesis logic beyond traditional fragment patterns; (2) constructing test sets with clinical drugs (TSA) and AI-generated molecules (TSB) for practical benchmarks; (3) evaluating models' performance in identifying synthesis difficulty cliffs and discriminating relative synthesis difficulty of intermediates in scheme; (4) providing an open platform combining computational efficiency with interpretability. SynFrag addresses SA prediction requirements in AIDD workflows.

2. METHODS AND MATERIALS

2.1. Data Set

2.1.1. Pretraining Data Set. We integrated commercially available molecules from seven databases to construct the pretraining data set. Public databases included ZINC,[17] PubChem,[18] and ChEMBL.[19] Commercial databases included Enamine,[20] VitasM, Chembridge,[21] and Lifechemical.[22] Data preprocessing employed three-stage filtering: (i) Molecular validity verification using RDKit[23] cheminformatics toolkit to validate chemical validity of SMILES

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX



C

Dataset	ES	HS	Total
Finetune	400,000	400,000	800,000
TS1	3,581	3,581	7,162
TS2	17,348	13,000	30,348
TS3	900	900	1,800
TSA	1,252	810	2,062
TSB	805	1,596	2,401

D

Dataset	Prediction Tool	Runtime (min)
TSB	AiZynthFinder	3411.6
	SynFrag	0.47

Figure 2. Data set construction and characterization.(A) Overview of data sets sources and construction pipeline. Pretraining data set (9.18 M molecules) from public databases (ZINC, PubChem, ChEMBL) and commercial suppliers (Enamine, VitasM, Chembridge, Lifechemical). Finetuning data set (800 K molecules) from boundary molecules (SAscore [3.5, 6]) annotated by Retro* and difficulty cliff pairs (Tanimoto similarity >0.35) from ZINC15 (ES) and Nonpher-generated molecules (HS). Test data sets (43,753 total): TS1−TS3 are public benchmarks; TSA contains clinical drugs and intermediates; TSB contains AI-generated molecules. (B) PCA visualization of test sets in chemical space using molecular fingerprints. (C) Distribution of ES/HS across data sets. (D) Runtime comparison: SynFrag processes 2401 molecules in 0.47 min; AiZynthFinder requires 3411.6 min.

strings, excluding structures violating valence rules; (ii) Fragmentation compatibility assessment to confirm molecules could be decomposed into fragments according to BRICS+2 rules and reconstructed as molecular tree structures, eliminating molecules with fragmentation or reconstruction failures; (iii) Data quality control through duplicate removal and integrity validation. After filtering, 9,185,665 molecules were obtained as the pretraining data set (Figure 2), covering chemical space from small compounds to complex natural products. Notably, beyond the above preprocessing steps, we did not implement additional standardization procedures to explicitly canonicalize tautomers and protonation states (e.g., −COOH vs $-COO^-$). Beyond the comparison of data sets pre- and poststandardization, posthoc analysis—including evaluation on standardized test sets and assessment of prediction consistency across representative tautomer pairs—indicated that such standardization was of low importance for SynFrag development (Tables S1 and S2). This is likely because the pretraining task primarily learns fragment assembly patterns rather than ionization-state-specific features. Nevertheless, we recommend incorporating complete standardization procedures into data processing workflows as best practice.

2.1.2. Finetuning Data Set. The fine-tuning stage was specifically designed for two scenarios in SA prediction: ambiguous boundary molecules and synthesis difficulty cliffs. We adopted an established multisource training data set previously used in SA prediction studies (GASA, DeepSA, BR-SAscore),[12−14] consisting of two components. (i) Decision-boundary molecules: molecules with SAscore values in [3.5, 6] were collected from ChEMBL and CDBChEMBL, representing ambiguous cases, then the retrosynthesis tool Retro*[24] provided annotations: compounds with feasible synthetic routes (≤10 steps) were labeled easy-to-synthesize (ES); others were labeled hard-to-synthesize (HS).[25] (ii)Synthesis difficulty cliff set: commercially available compounds from ZINC15 served as ES molecules; structurally complex molecules generated by the Nonpher[26] model served as HS molecules; molecular pairs with fingerprint similarity >0.35 were retained to capture cases where similar structures diverge in SA.

2.1.3. Test Data Set. Five test sets were constructed covering different application scenarios, totaling 43,753 molecules.

Public benchmark test sets (TS1−TS3): TS1, constructed by Voršilák et al.,[11] contains 7162 molecules with controlled design: HS molecules (3581) from complexity analysis of the GDB17 database and ES molecules (3581) from expert curation of the ZINC database.[27] TS2, developed by Thakkar et al.,[8] comprises 30,348 molecules sampled from ChEMBL, CDBChEMBL, and other databases, with annotations based on retrosynthesis planning from AiZynthFinder. TS3 contains 1800 molecules (900 ES, 900 HS) from expert-curated ZINC molecules, literature-reported molecules with expert assessment, and molecules from MLSMR and PubChem annotated using Retro*. The three test sets represent increasing difficulty gradients (Figure S1).[12] Notably, label consistency between the test and fine-tuning data sets is crucial for model development and evaluation. Since Retro* serves as the primary labeling method for fine-tuning data, we relabeled TS1 and TS2 using identical Retro* configurations and compared them with the original benchmark labels. Results showed substantial agreement between different labeling approaches (Tables S3 and S4), and the small proportion

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

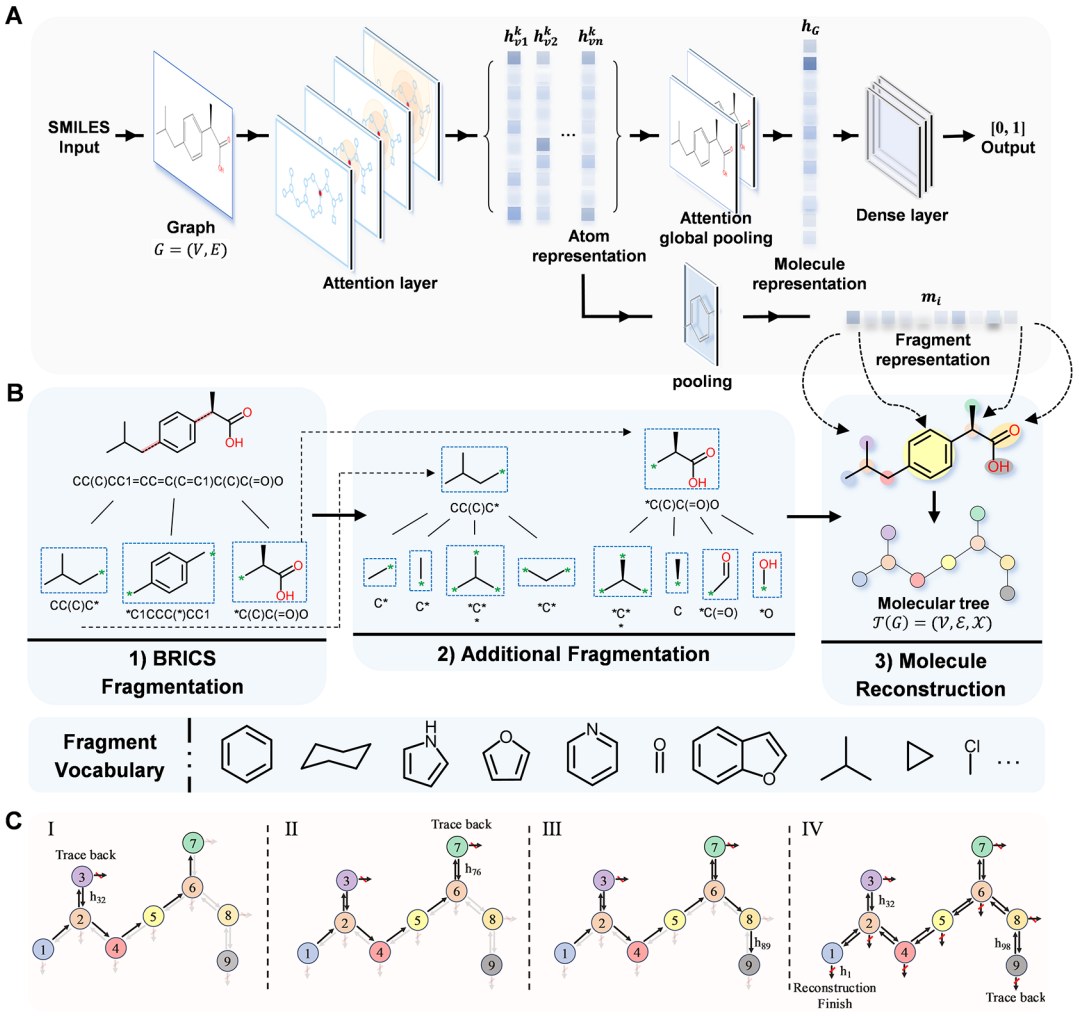


Figure 3. SynFrag technical components. (A) AttentiveFP architecture for molecular representation learning. SMILES input is converted to molecular graph and processed through attention layers to update atomic representations, attention-weighted global pooling for molecular representation, and dense layers for output [0,1]. Pooling operations aggregate atomic features into fragment representations. (B) used the BRICS +2 fragmentation strategy. Step 1: BRICS rules decompose molecules into initial fragments. Step 2: Additional rules refine fragmentation. Step 3: Fragments are organized into tree structure. Fragment vocabulary defines the label space for assembly prediction (C) Example of ibuprofen for the DFS fragment assembly generation process: The assembly initiates from root node 1, then the label predictor and topology predictor alternate in this process, conducting exploration along the sequence (1 → 2 → 3). When the topological predictor at node 3 correctly determines no subsequent connections exist, backtracking operations are executed to branching node 2, continuing exploration of another side chain (2 → 4 → ··· → 7 → 6 → 8 → 9). Upon completion of tracing back for each individual node with no remaining node to connect with, the fragment autoregressive assembly task finishes with successful generation (reconstruction) of the original molecule. This process simulates stepwise synthesis from building blocks.

of inconsistent labels was predominantly concentrated in the SA decision boundary region (Table S5). This suggests that current labeling strategies achieve a reasonable balance between consistency and methodological diversity.

Real-world application test sets (TSA, TSB): TSA (2062 molecules) covers drug discovery scenarios with four sources: (i) FDA-approved drugs from the past 8 years and clinical trial candidates,[28−35] (ii) their key synthetic intermediates, (iii) total synthesis of organic compounds reported in the past 5 years, and (iv) in-house lead compounds synthesized in our laboratory. Molecules with publicly available synthetic routes of ≤10 steps were labeled ES. Those requiring >10 steps or lacking disclosed routes were labeled HS. In-house leads were annotated by senior medicinal chemists. TSB (2401 molecules) assesses model performance on novel chemical entities. Molecules were generated by the Graph GA algorithm,[36] producing structurally novel derivatives while preserving pharmacophoric groups. Dual-validation annotation was used: Retro* and AiZynthFinder were applied with identical settings (maximum depth = 10 steps, time limit = 360 s, full template library). Molecules were labeled ES if both tools identified ≤10 steps, HS if both required >10 steps or failed, and excluded if results were inconsistent. All five test sets were deduplicated using canonical SMILES to prevent molecular overlap across data sets.

2.2. SynFrag Model

SynFrag uses fragment assembly autoregressive generation. In organic synthesis, chemists start with commercially available building blocks and perform stepwise assembly through systematic reactions to obtain the target molecules. Each reaction step builds upon the products of the preceding steps. SynFrag converts this assembly process into machine learning tasks through pretraining that reconstructs molecules by sequentially predicting fragment assembly using a depth-first search (DFS). The model starts from an initial fragment and progressively builds complete molecular structures. This training approach is expected to capture interfragment connectivity patterns, assembly sequences, and structural relationships relevant to synthetic accessibility. The autoregressive framework learns the implicit

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

knowledge of individual fragments and structural relationships between fragment assembly steps, reducing reliance on labeled data compared to traditional fragment frequency-based or solely supervised methods.

2.2.1. AttentiveFP Module. Message-passing networks provide molecular representations. AttentiveFP is a graph neural network for molecular property prediction.[37] Molecular graph construction: Molecules are converted to molecular graphs $G = (V, E)$ using RDKit, where V represents atomic nodes, and E represents chemical bonds. Each atom is initialized as a feature vector containing an atomic type, formal charge, hybridization state, aromaticity, and chirality. Each bond is initialized with a bond type, aromaticity, and ring membership features. Graph convolution featurizers (DeepChem implementation) enrich node and edge representations (Table S6).[38] Message-passing mechanism: During K-round updates, each atom aggregates information from neighboring atoms through attention mechanisms. Attention scores compute feature compatibility between central and neighboring atoms (eq 1):

$$e_{ij} = a\left(\left[Wh_i \middle\| Wh_j\right]\right), \quad j \in N_{(i)} \tag{1}$$

where $a(\cdot)$ represents the attention network, W denotes learnable parameters, $\|$ indicates feature concatenation, and $N_{(i)}$ represents neighbors of atom i. Attention weights undergo softmax normalization (eq 2). Neighbor information is aggregated through attention-weighted summation to form context vectors (eq 3), and atomic representations (Figure 3A) are updated through gated recurrent units (GRU) (eq 4):[39]

$$a_{ij} = \mathrm{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{l\in N(i)} \exp(e_{il})} \tag{2}$$

$$C_i^{k-1} = \mathrm{elu}\left(\sum_{j\in N(i)} a_{ij} \cdot W_2 h_j^{k-1}\right) \tag{3}$$

$$h_i^k = \mathrm{GRU}^{k-1}(C_i^{k-1}, h_i^{k-1}) \tag{4}$$

where elu denotes exponential linear unit activation. The context vector C_i^{k-1} aggregates neighbor features h_j^{k-1} weighted by attention coefficients a_{ij} from eq 2. The GRU gating mechanism then integrates this aggregated neighborhood information with the previous hidden state, h_i^{k-1}, to produce updated atomic representations.

In parallel, BRICS+2 fragmentation decomposes molecules into fragments (Figure 3B), constructing molecular trees $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ with fragments as nodes and broken bonds as edges. Fragment representations m_i are obtained through pooling operations from constituent atomic representations (eq 5), where M_i represents the set of atoms in the i-th fragment. Molecular-level representations h_G are obtained through attention-weighted global pooling (eq 6),[40] where f represents the attention score network, and α_v indicates each atom's contribution weight to the molecular representation.

$$m_i = \mathrm{Pool}(\{h_v^K | v \in M_i\}) \tag{5}$$

$$h_G = \sum_{v\in V} \alpha_v \cdot h_v^K, \quad \alpha_v = \frac{\exp(f(h_v^K))}{\sum_{u\in V} \exp(f(h_u^K))} \tag{6}$$

2.2.2. BRICS + 2 Fragmentation. We improved BRICS+2 fragmentation based on the BRICS algorithm by Degen et al.[41] BRICS defines 16 retrosynthetic bond-breaking rules that decompose molecules into fragments (e.g., aromatic C−C bonds, aromatic C−O bonds, and N-aliphatic C bonds). The original BRICS strategy has two limitations: (i) Incomplete fragmentation: complex molecules may retain large fragments, increasing learning complexity; (ii) Uneven fragment distribution: oversized fragments exhibit low occurrence frequencies, affecting pretraining. We introduced two additional rules to form BRICS+2: (i) Ring−chain bond-breaking: bonds connecting intraring and extra-ring atoms are broken to form independent ring and chain fragments; (ii) Branching point bond-breaking: nonring atoms connecting three or more bonds are broken to form independent atom fragments. These rules reduce fragment combination complexity and increase common functional group frequency in the fragment vocabulary (Figure 3B).[42]

2.2.3. DFS-Based Fragment Assembly Autoregressive Generation Pretraining. SynFrag converts molecular construction processes into machine learning tasks. Chemists assemble complex molecules from building blocks through stepwise reactions, and our model learns this process through pretraining:

Initially, molecular trees $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ serve as representations for the autoregressive task.[43−46] $\mathcal{V}$ represents the set of fragment node m_i, $\mathcal{E}$ represents broken bond edge sets, and $\mathcal{X}$ represents fragment vocabulary. The pretraining objective maximizes molecular tree generation probability[47] $p(\mathcal{T}(G); \theta)$ (eq 7), seeking optimal parameters $\theta^* = \mathrm{argmax}_\theta p(\mathcal{T}(G); \theta)$. Two predictors are used: (i) topology predictors determine whether next nodes connect to current nodes; (ii) label predictors determine specific labels of the next nodes. Autoregressive generation is modeled as conditional probability products (eq 8):

$$p(\mathcal{T}(G); \theta) = \mathbb{E}_\pi[p_\theta(\mathcal{V}^\pi, \mathcal{E}^\pi)] \tag{7}$$

$$\log p_\theta(\mathcal{V}, \mathcal{E}) = \sum_{i=1}^{|\mathcal{V}|} \log p_\theta\left(\mathcal{V}_i, \mathcal{E}_i \middle| \mathcal{V}_{<i}, \mathcal{E}_{<i}\right) \tag{8}$$

where π denotes fragment assembly sequences, $\mathbb{E}$ denotes expectation over all possible sequences, implemented by DFS recursion,[48] and $\mathcal{V}^\pi$ and $\mathcal{E}^\pi$ denote node labels and edge connections arranged by sequence π.

Autoregressive fragment assembly starts from root nodes (1st atom located).[49] Topology predictors determine whether connected next fragments exist. Label predictors determine specific fragment types when they exist (Figure 3C). Each prediction uses historical node labels $\mathcal{V}_{<i}$ and edge connections $\mathcal{E}_{<i}$, embodying autoregressive characteristics. The DFS algorithm determines the fragment assembly order, representing linear growth in organic synthesis that completing main chain before branch modifications.

Fragment-level message-passing and dual predictors are implemented. Fragment nodes exchange information through the molecular tree structure. Fragment-level hidden state updates as eq 9, where $h_{i,j}$ represents the hidden state of fragment i at iteration j, m_i denotes the fragment representation from eq 5.

$$h_{i,j} = \mathrm{GRU}\left(m_i, \sum_{(k,i)\in\mathcal{E}_i} h_{k,i}\right) \tag{9}$$

Topology predictors use sigmoid functions to estimate connection probabilities (eq 10):

$$p_t = \sigma\left(U^d \tau\left(W_1^d m_i + W_2^d \sum_{(k,i)\in\mathcal{E}_i} h_{k,i}\right)\right) \tag{10}$$

Label predictors select fragment types in vocabulary through softmax functions (eq 11):

$$q_j = \mathrm{softmax}\left(U^l \tau\left(W^l h_{i,j}\right)\right) \tag{11}$$

where σ represents sigmoid functions, τ represents relu functions, and U^d, W_1^d represent learnable parameters. Pretraining loss $\mathcal{L}_{pre}$ integrates cross-entropy from both predictors (eq 12),[50−52] where $\hat{p}_t$ and $\hat{q}_j$ denote true connections and labels:

$$\mathcal{L}_{\mathrm{pre}} = \sum_t \mathcal{L}_{\mathrm{topo}}(p_t, \hat{p}_t) + \sum_j \mathcal{L}_{\mathrm{label}}(q_j, \hat{q}_j) \tag{12}$$

Through end-to-end training, the model learns implicit knowledge about fragments and their assembly logic.

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

2.2.4. Finetuning for SA Prediction. AttentiveFP is initialized using pretrained parameters. (Figure 1) During forward propagation, AttentiveFP processes molecular graphs to obtain atomic representations, aggregates molecular representations through attention global pooling, and outputs SA prediction values through dense layers and sigmoid activation. Loss function $\mathcal{L}_{fine}$ employ binary cross-entropy (eq 13):

$$\mathcal{L}_{\text{fine}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \quad (13)$$

2.2.5. Model Training and Hyperparameter Optimization. Pretraining used 9.18 million molecules with Adam optimizer (learning rate 1e−4, batch size 64).[53] The gradient clipping threshold was 1.0.[54] Cosine annealing learning rate scheduling decreased rates to 10% of the initial values after 100 epochs.[55] Pretraining used 4 TESLA A100 GPUs. Finetuning employed random search with early stopping.[56,57] Search spaces included the following: message-passing layers [2, 3, 4, 5], aggregation layers [1, 2, 3], hidden units [100, 200, 300, 400], dropout [0.1, 0.3, 0.5], learning rates [0.01, 0.001, 0.0001, 0.00001], batch sizes [16, 32, 64], and weight decay [0.001, 0.0001, 0.00001] (Table S7 and Figure S2).

2.3. Evaluation Metrics

SA prediction is defined as a binary classification task distinguishing hard-to-synthesize (HS) and easy-to-synthesize (ES) molecules. Based on confusion matrix elements (TP, FP, TN, FN), evaluation metrics include:

Accuracy measures the overall classification performance. Precision evaluates the reliability of ES predictions. Sensitivity (Recall) measures the completeness of ES identification. High sensitivity reduces false negatives of the ES candidates. Specificity measures the HS identification accuracy. High specificity reduces resources spent on difficult candidates (Table 1). F-score is the harmonic mean of precision and recall. AUROC calculates the area under the ROC curve, plotting the true positive rate against the false positive rate. Values range from 0 to 1; higher values indicate better discrimination. PR-AUC is used for imbalanced data sets.[58] Cohen's d quantifies separation between ES and HS prediction distributions: $|d|$ = 0.2 (small), $|d|$ = 0.5 (medium), $|d|$ = 0.8 (large).[59]

Table 1. Evaluation Metrics for the Synthetic Accessibility Prediction[a]

evaluation metric	equation
accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
precision	$\frac{TP}{TP + FP}$
recall	$\frac{TP}{TP + FN}$
specificity	$\frac{TN}{TN + FP}$
F-score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Cohen's d	$\frac{(\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{(n_1 + n_2 - 2)}}}$

[a]TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative. μ_1, μ_2 are prediction means for ES and HS; σ_1, σ_2 are standard deviations; n_1, n_2 are sample sizes.

2.3.1. Comparative Experiments and Ablation Studies. Seven SA prediction methods across three categories were compared with SynFrag: (i) Fragment-based statistical methods: SAscore (fragment frequency with complexity penalty), BR-SAscore (SAscore with reaction center information), SYBA (naive Bayes classification on fragments); (ii) Reaction routes methods: SCScore (learning from reaction step complexity), RAscore (using AiZynthFinder retrosynthesis results); (iii) Deep learning methods: GASA (graph attention network) and DeepSA (BERT on SMILES sequences). For models with predefined thresholds (e.g., SYBA: 0.0, GASA: 0.5, DeepSA: 0.5), official parameters were used. Otherwise, optimal thresholds were determined by grid search on each test set.[60]

Ablation experiments quantify pretraining contributions. SynFrag with and without pretraining, AttrMasking and MolCLR were adopted to compare the contribution of pretraining strategies to SA prediction. These models used a consistent AttentiveFP backbone and were trained for identical epochs, and tested by TS2 and TS3. AUROC curves throughout the training were recorded.

2.3.2. Chemical Interpretability Analysis. Four analyses were performed: (i) Molecular descriptor analysis: SynFrag prediction distributions across chiral center counts (0, 1, 2, 3, ≥4) and molecular weight ranges (<500 Da, ≥ 500 Da). Four contrasting cases were examined: molecules with minimal/multiple chiral centers and low/high molecular weights but opposite SA label. (ii) Using steroid derivatives as a case study, we investigate whether SynFrag misclassifies these compounds as ES owing to the full preservation of their synthetically challenging fragments in pretraining. (iii) Synthesis difficulty cliff analysis: Three in-house molecular pairs with high structural similarity (Tanimoto > 0.83) but different SA were selected, representing N−N bond instability, distance-dependent electronic effects, and steric hindrance. (iv) Multistep synthetic route analysis and attention mechanism analysis: Predictions for atogepant and its 9 intermediates were normalized across models (SynFrag, SCScore, BR-SAscore, DeepSA) for comparison. Attention weights for the final product atogepant were examined for correspondence with key reactive sites.

3. RESULTS AND DISCUSSION

3.1. Systematic Performance Evaluation on Public Test Sets and Real-world Application Scenarios

To comprehensively validate SynFrag's performance, systematic comparisons with seven mainstream models were conducted across five test sets. The evaluation framework encompassed three public benchmark test sets (TS1−TS3) and two specially constructed real-world application scenario test sets (TSA clinical drug data set, TSB AI-generated molecule data set), with comparison models spanning three technical paradigms: traditional fragment frequency, reaction routes-based learning, and deep learning.

SynFrag demonstrated a superior performance across all test sets. In public benchmark testing (Table 2), TS1 achieved AUROC values of 1.000, TS2 obtained 0.940 (2.9% improvement over the SOTA DeepSA's 0.913), and TS3 reached 0.894 (comparable to DeepSA's 0.896 while maintaining advantages in Accuracy and Specificity); SynFrag also obtained high Specificity values across all test sets (0.891−0.995), ensuring accurate identification of synthesis barriers. More importantly, SynFrag exhibited outstanding performance in real-world application scenarios (Figure 4): TSA achieved AUROC 0.945, and TSB reached 0.889, both representing the highest values among all models.

To rigorously assess the statistical significance of these performance differences, we employed DeLong's test for comparing AUROC values, which accounts for the correlation structure when evaluating different models on the same test set. With the Bonferroni correction applied for multiple comparisons (28 pairwise comparisons, α = 0.00179), a comprehensive pairwise comparison was conducted across all eight methods and five test sets (Figure S3). Notably, the statistical comparison between SynFrag and SYBA (a representative fragment-frequency-only method) revealed significant improvements across all five test sets (Table S8). The minimal difference on TS1 (ΔAUROC = 0.002) reflects near-ceiling performance, whereas on TSB, SynFrag achieved a

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Performance Comparison on Public Benchmark Test Sets[a]

data set	model	accuracy	AUROC	specificity	PRAUC	recall	*F*-score	threshold
TS1	SAscore	0.989	*0.999*	0.986	0.998	0.992	0.989	4.50
	SCScore	0.608	0.641	0.518	0.585	0.698	0.641	3.10
	SYBA	0.962	0.998	0.925	0.693	**1.000**	**1.000**	0.00
	RAscore	0.919	0.981	0.970	0.692	0.867	0.874	0.50
	GASA	0.987	*0.999*	0.975	**1.000**	*0.999*	0.987	0.50
	DeepSA	*0.995*	**1.000**	*0.990*	**1.000**	**1.000**	0.995	0.50
	BR-SAscore	0.831	*0.999*	0.662	*0.999*	**1.000**	0.855	5.00
	SynFrag	**0.997**	**1.000**	**0.995**	**1.000**	**1.000**	*0.997*	0.50
TS2	SAscore	*0.841*	0.919	0.869	*0.909*	0.804	0.813	3.40
	SCScore	0.449	0.373	0.071	0.341	**0.954**	0.597	2.30
	SYBA	0.787	0.862	0.905	0.730	0.627	0.711	0.00
	RAscore	0.751	0.865	**0.950**	0.726	0.485	0.630	0.50
	GASA	0.796	0.876	0.885	0.840	0.677	0.740	0.50
	DeepSA	0.838	0.913	0.911	0.904	0.730	0.795	0.50
	BR-SAscore	0.827	*0.925*	0.774	0.902	*0.898*	*0.816*	5.00
	SynFrag	**0.869**	**0.940**	*0.912*	**0.929**	0.817	**0.839**	0.50
TS3	SAscore	0.707	0.772	0.641	0.721	0.772	0.725	3.10
	SCScore	0.511	0.425	0.046	0.443	**0.977**	0.666	2.20
	SYBA	0.647	0.790	**0.907**	0.654	0.387	0.513	0.00
	RAscore	0.701	0.790	0.831	0.654	0.571	0.620	0.50
	GASA	0.760	0.849	0.874	0.812	0.645	0.729	0.50
	DeepSA	*0.817*	**0.896**	0.881	**0.912**	0.753	0.804	0.50
	BR-SAscore	0.804	0.869	0.702	0.819	*0.905*	**0.822**	5.00
	SynFrag	**0.820**	*0.894*	*0.891*	*0.869*	0.762	*0.806*	0.50

[a]**Bold**: best performance; *italic*: second-best performance. Threshold represents the cutoff value for binary classification. For SYBA, RAscore, GASA, DeepSA, and SynFrag, molecules above the threshold are classified as ES; for SAscore, SCScore, and BR-SAscore, molecules above the threshold are classified as HS.

substantial improvement of ΔAUROC = 0.221 (33.1%), highlighting its superior performance on AI-generated molecules.

Performance on TSA and TSB data sets reveals differences across technical approaches (Figure 4). SynFrag achieved Cohen's *d* of 2.814 (TSA) and 1.783 (TSB), while BR-SAscore and SAscore achieved *d* = 2.488 and 2.247 on TSA. Box plots (Figure 4A,B) illustrate the distributions underlying these effect sizes. On TSA, SynFrag shows clear separation between ES and HS distributions, with ES molecules concentrated at high prediction values (median ≈ 0.95) and HS molecules at low values (median ≈ 0.05). On TSB, while separation decreased (*d* = 1.783), SynFrag maintained distinct distributions with minimal overlap. Traditional fragment-based methods showed marked changes: BR-SAscore's Cohen's *d* decreased from 2.488 to 1.622, and SAscore from 2.247 to 1.485 on TSB, with increasing distribution overlap visible in box plots. This pattern suggests a reduced generalization to novel chemical structures. DeepSA showed highly variable performance, with overlapping distributions on TSA (*d* = 0.257) but improved separation on TSB (*d* = 1.066).

AUROC analysis confirms these patterns (Figure 4C,D). BR-SAscore and SAscore achieved 0.942 and 0.937 on TSA, but decreased to 0.866 and 0.848 on TSB. DeepSA achieved 0.362 on TSA and 0.797 on TSB, indicating variable performance across chemical spaces. This discrepancy may relate to the limitations of the pure SMILES sequence representation or insufficient model generalization. SynFrag maintained consistent performance as the fragment assembly pretraining approach captures generalizable synthesis pattern knowledge across different molecular distributions. These results set the stage for the scenario-specific analyses that follow.

To examine whether the observed improvements stem specifically from fragment assembly pretraining rather than generic pretraining benefits, we conducted ablation experiments comparing four conditions under documented pretraining and identical finetuning settings: (1) SynFrag with fragment assembly pretraining, (2) SynFrag w/o (without pretraining, random Xavier initialization as baseline), (3) AttrMasking (Hu et al., attribute masking pretraining),[61] and (4) MolCLR (Wang et al., molecular contrastive learning pretraining).[62] All methods used the same AttentiveFP backbone architecture and finetuning protocol.

SynFrag achieved the highest performance (Figure 4E,F), with an AUROC of 0.934 on TS2 and 0.887 on TS3, representing relative improvements of 2.50 and 3.13% over the baseline, respectively. In contrast, MolCLR showed minimal improvements (0.03% on TS2, 0.66% on TS3), and AttrMasking exhibited slight negative transfer (−0.86% on TS2, −0.97% on TS3). The learning curves revealed distinct training dynamics: while MolCLR and AttrMasking showed early improvements followed by plateaus around 50K iterations, SynFrag demonstrated sustained improvement throughout training, with AUROC continuing to increase in later stages. This sustained learning trajectory suggests that fragment assembly pretraining captures task-relevant structural features that align more closely with the downstream objective of SA prediction. The minimal or negative effects of AttrMasking and MolCLR, methods validated on multiple molecular property prediction tasks, highlight the task-specific nature of pretraining benefits and underscore the advantage of

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

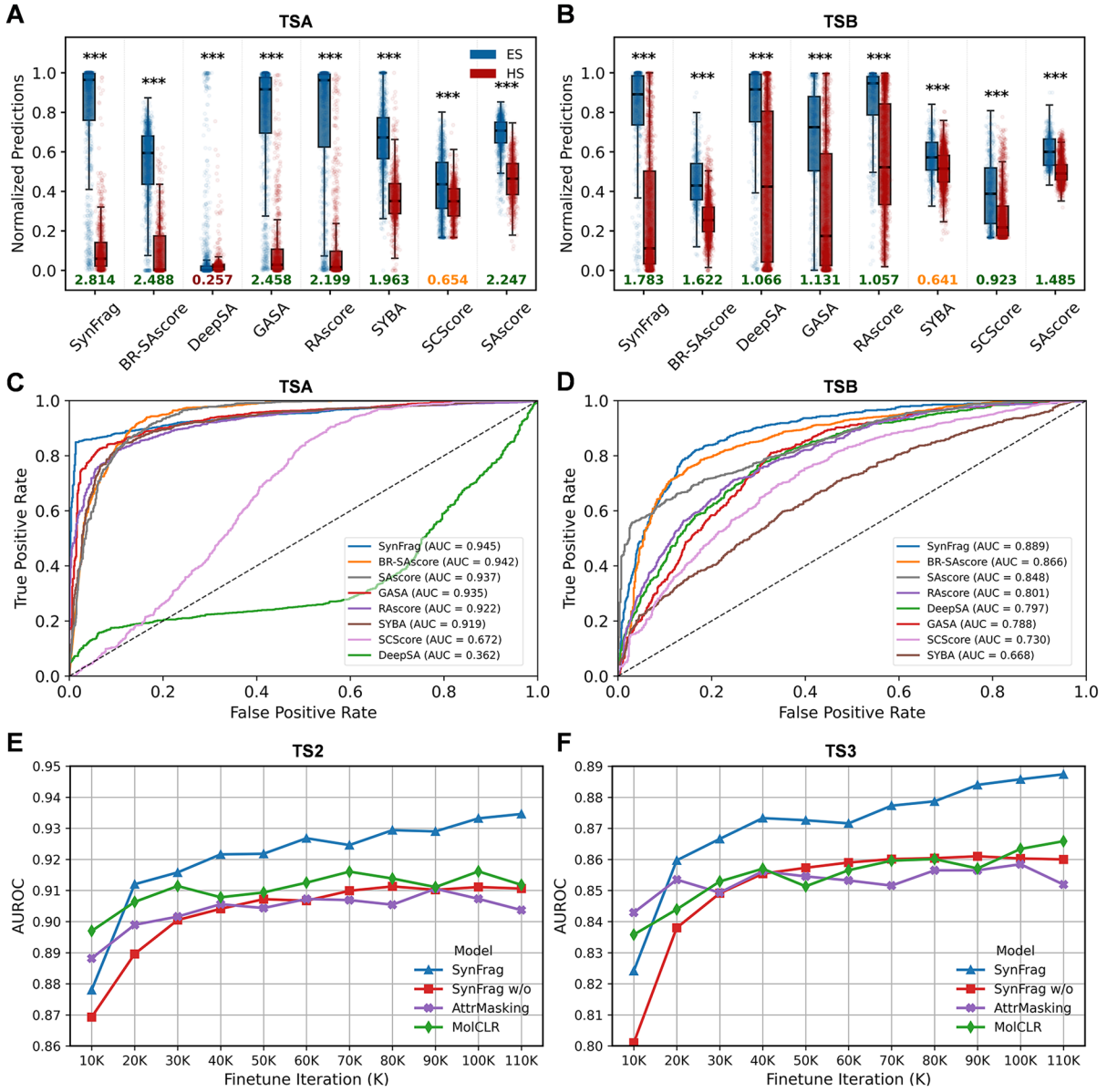


Figure 4. Performance evaluation on real-world scenarios and ablation studies. (A, B) Box plots comparing normalized prediction distributions across eight methods on TSA (clinical drugs/intermediates, n = 2,062) and TSB (AI-generated molecules, n = 2,401). Blue: ES molecules; red: HS molecules. Center line: median. Numbers below indicate Cohen's d (orange values < 0.8 indicate relative weak separation). Statistical significance by the Mann−Whitney U test;[63] ***p < 0.001. (C, D) ROC curves for TSA and TSB. Dashed diagonal: random classifier (AUROC = 0.5) (E, F) Finetuning dynamics in ablation study on TS2 and TS3 comparing four pretraining conditions. Blue: SynFrag (ours); red: SynFrag w/o (no pretraining, baseline); green: MolCLR (contrastive learning pretraining); purple: AttrMasking (attribute masking pretraining). x-axis: finetuning iterations.

SynFrag's synthesis-oriented pretraining strategy for SA prediction.

3.2. Chemical Principle Validation of Model Predictions

To assess whether SynFrag predictions are grounded in chemical reasoning rather than statistical correlations, we examined relationships between SynFrag predictions and molecular descriptors through distribution analysis and case studies.[64]

Chiral center dependency showed hierarchical patterns (Figure 5A). SynFrag prediction values decreased progressively as chiral center counts increased, with all consecutive groups showing significant differences (p < 0.001). This stepwise decline aligns with synthetic chemistry principles: each additional stereocenter typically doubles the number of possible stereoisomers and increases stereoselective synthesis complexity.[65] However, the model assigned high scores to Gwanakoside A (8 chiral centers, SynFrag = 0.5934), consistent with established glycosylation methodologies[66] that address stereochemical complexity in carbohydrate synthesis[67] (Figure 5C).

Molecular weight (MW) showed a statistically significant but modest influence on predictions[68] (Figure 5B). The small effect size (Cohen's d = 0.37) and substantial distribution overlap indicate limited reliance on the molecular size. This pattern reflects chemical reality: large biomolecules like peptides and oligonucleotides can be synthesized through building block assembly strategies,[69] while some small

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

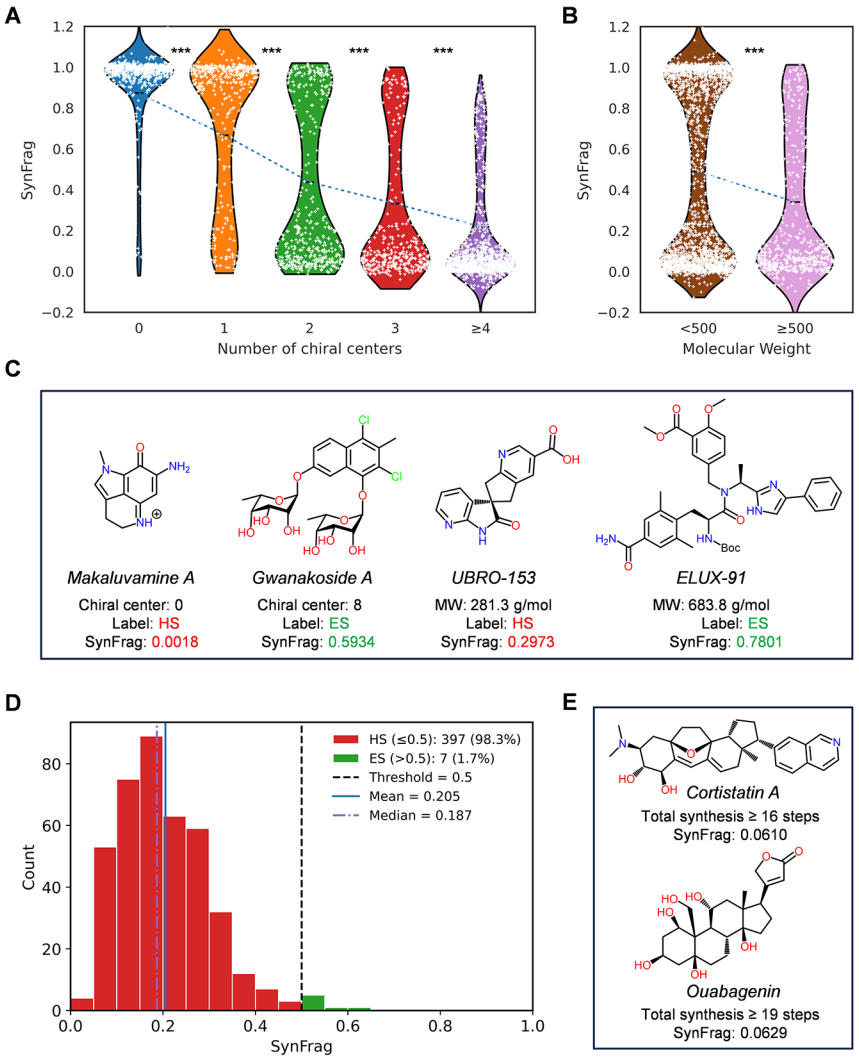


Figure 5. Chemical rationality validation and analysis. (A, B) Violin plots showing SynFrag prediction distributions across descriptors (TSA & TSB, n = 4,463). White dots: SynFrag predictions; dashed lines: mean trends. (A) by chiral center counts. Kruskal–Wallis test followed by Dunn's posthoc test with Bonferroni correction;[70] ***p < 0.001 for all consecutive comparisons. (B) Distribution by MW. Independent samples t test, ***p < 0.001, Cohen's d = 0.37. (C) Cases about SynFrag predictions are based on chemical principles rather than simple descriptor correlations. (D) Distribution of SynFrag predictions for 404 steroid compounds. Red bars: HS (≤0.5); green bars: ES (>0.5). Dashed lines indicate threshold (0.5), mean (0.205), and median (0.187). (E) Representative cases of synthetically challenging steroids with documented total syntheses: Cortistatin A (SynFrag = 0.0610; ≥ 16 synthetic steps) and Ouabagenin (SynFrag = 0.0629; ≥ 19 synthetic steps).

molecules with strained rings or reactive functional groups present synthesis challenges.

Case validation examined four molecules with a ground-truth label that challenge descriptor-based predictions (Figure 5C). Makaluvamine A[71] (0 chiral centers, SynFrag: 0.0018) is a minimal chiral centers molecule with HS prediction due to its fused heterocyclic system.[72] Gwanakoside A (8 chiral centers, SynFrag: 0.5934) is a complex glycoside, but with ES prediction, consistent with established synthetic methodologies. UBRO-153 (MW: 281.3 g/mol; SynFrag: 0.2973) demonstrates that low MW does not ensure synthesis ease when molecules contain densely functionalized heterocycles. ELUX-91 (MW: 683.8 g/mol, SynFrag: 0.7801) shows that large molecules composed of modular building blocks can have high SynFrag predictions. These results indicate that predictions emerge from learned synthesis patterns rather than simple descriptor correlations.

3.3. Validation on Complex Polycyclic Scaffolds

A potential issue with SynFrag's fragment-based pretraining is whether complex ring systems, although preserved as intact fragments during pretraining, pose substantial total synthesis difficulty and would be misclassified as ES. To address this, we validated SynFrag on steroid compounds, whose tetracyclic steroid cores exist as intact fragments in the fragment vocabulary.

SynFrag's prediction results on 404 steroid compounds (Figure 5D, Table S9) showed consistently low scores (mean = 0.205; median = 0.187), with 98.3% of compounds classified as HS. This indicates that SynFrag does not systematically misclassify such compounds as ES merely because their complex polycyclic scaffolds exist as intact fragments in the pretraining fragment vocabulary. Two representative cases supported by a well-documented total synthesis were selected for further illustration (Figure 5E). Cortistatin A, a marine steroidal alkaloid featuring a unique 9(10–19)-abeo-androstane skeleton and an oxabicyclo [3.2.1]octene motif, required

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

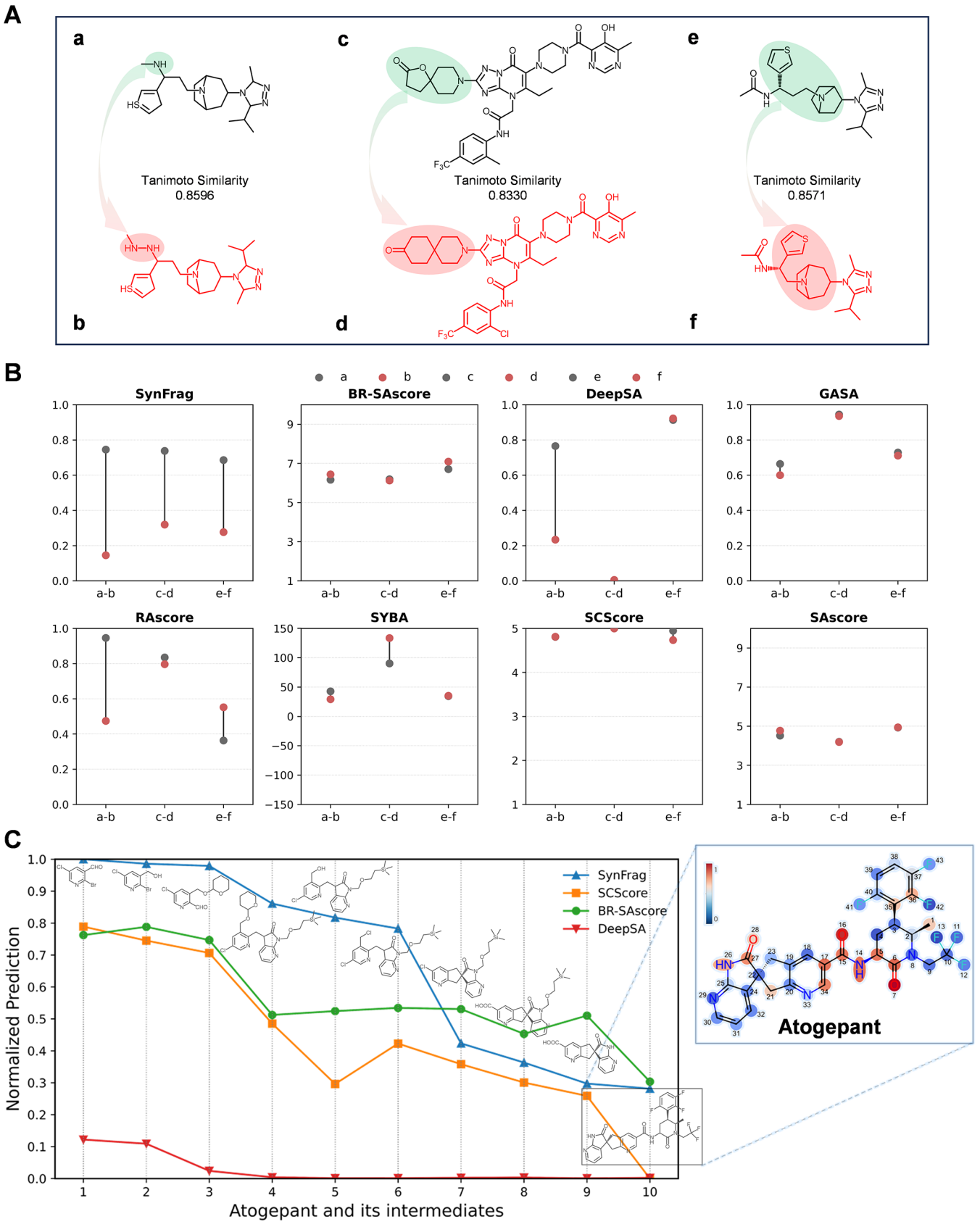


Figure 6. Synthesis difficulty cliff identification and multistep routes assessment. (A) Three pairs of molecules with high similarity but opposite SA. Red: HS, green/red shadow: structures responsible for SA divergence. (B) Predictions comparisons for three pairs across eight methods, *y*-axis: each model's prediction (threshold see Table 2) (C) Normalized predictions (1 = ES, 0 = HS) for atogepant and its 9 synthetic intermediates. Inset: SynFrag attention weight heatmap for atogepant; shadow colors indicate contribution to prediction. High-attention area corresponds to key reactive sites in atogepant's synthesis route.

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Figure 7. SynFrag web platform interface. (A) CSV file upload interface accepting SMILES strings (up to 250,000 molecules per submission). (B) Results page with download options. Generated links can be shared for accessing results and attention heatmaps. Email results is optional. Results table shows Index, SMILES, SynFrag scores, and Status. Summary statistics show total and successful predictions and average scores. (C) Interactive attention heatmap showing atom-level contributions, alerting users to potential challenges in specific structural synthesis. (D) CASP tools integration for retrosynthesis route planning.

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

≥16 synthetic steps in multiple groundbreaking total synthesis studies;[73] SynFrag scores it 0.061, correctly classifying it as HS. Ouabagenin, one of the highly oxygenated cardenolides with six hydroxyl groups and one β-oriented butenolide, is a highly challenging synthetic target that requires ≥19 steps via different synthetic routes.[74] SynFrag assigned it a score of 0.063 and correctly classified it as HS.

We speculate that this robustness originates from SynFrag's multilevel learning strategy: AttentiveFP captures atom-level features including stereochemical characteristics and functional group contexts, while the autoregressive assembly task is responsible for learning fragment connectivity patterns. These components work synergistically, enabling SynFrag to evaluate SA based on the complete molecular context rather than fragment occurrence frequency.

3.4. Identification of Synthesis Difficulty Cliffs in Lead Optimization

AI models for lead optimization enable automated generation of structural variants, typically through R-group substitutions.[75] Such modifications play a key role in medicinal chemistry, yet many AI-generated proposals do not consider SA. Small functional changes can create"synthesis difficulty cliffs" that are not typically detected by current generative frameworks.[76] The practical utility of AI-designed analogs depends on whether these cliffs can be identified, requiring SA models to capture accessibility changes arising from minor substituent variations.

We analyzed cases from our in-house compound collection where molecules with high structural similarity exhibit divergent SA profiles. Three pairs were selected (Tanimoto similarity:[77] 0.8596, 0.833, 0.8571), each representing distinct chemical principles underlying synthesis difficulty (Figure 6A): (1) pair a-b: N−N bond instability arising from lone pair repulsion;[78] (2) pair c-d: distance-dependent carbonyl induction effects,[79] where increased separation in d weakens nucleophilic site activation; (3) pair e-f: steric hindrance from the larger bridged ring in f combined with a shorter chain. Notably, these molecular pairs have comparable reagent availability profiles, allowing the practical synthesis difficulty differences to be primarily attributed to intrinsic structural factors.

Results (Figure 6B) show that SynFrag correctly distinguished all three pairs: a-b (0.7520 vs 0.1450), c-d (0.7384 vs 0.3194), and e-f (0.6856 vs 0.2837). In contrast, other methods showed limited performance: RAscore and DeepSA succeeded only on pairs a-b, while BR-SAscore, GASA, SYBA, SCScore, and SAscore failed across all cases. These results demonstrate that SynFrag captures how subtle structural variations influence SA, complementing reagent availability considerations in practical synthesis planning.

This capability partially addresses a key challenge in AI-driven lead optimization. Existing SA models often lack sensitivity to detect fine-grained synthesis cliffs, while CASP tools require substantial computational time that limits large-scale application. SynFrag provides subsecond predictions with the ability to distinguish structurally similar molecules differing in synthetic difficulty. This ability enables efficient filtering of synthetically problematic variants during high-throughput screening, potentially increasing chemists' willingness to synthesize AI-generated candidates. For screening AI-generated libraries, SynFrag thus represents an intermediate approach between rapid but less sensitive traditional SA models and accurate, but time-consuming, CASP tools.

3.5. Synthesis Complexity Assessment in Multi-Step Routes and Attention Mechanism Analysis

Assessment of relative SA among intermediates in multistep synthetic routes is a challenge in medicinal chemistry, influencing synthesis strategy optimization.[80] SynFrag showed a performance on the TSA test set, which contains drugs and their synthetic intermediates. This supports applicability to drug discovery scenarios and the ability to discriminate SA differences among intermediates. We analyzed the synthesis route of atogepant[33,81] as a case, encompassing reaction types including Knoevenagel condensations, nucleophilic substitution-based heterocycle construction, and amide bond formation.

Normalized predictions showed different model behaviors (Figure 6C). SynFrag exhibited a monotonic decline from 1.0 (initial reactant) to 0.28 (atogepant), capturing progressive complexity accumulation through ring formations, stereocenters, and functional modifications. The transition from intermediate 6 to 7 (6th step, intramolecular nucleophilic substitution forming a cyclic structure) showed a notable prediction decrease in SynFrag, consistent with increased synthetic complexity at this cyclization step. SCScore showed similar trends (0.79→0) with a decrease at the 6th step but a fluctuation at intermediates 5−6. In contrast, BR-SAscore displayed reversals, dropping at intermediate 4 and then recovering, indicating limitations of the traditional fragment frequency method. DeepSA maintained near-zero predictions (0.12→0.002) with limited discrimination between intermediates, suggesting that the SMILES sequence representation-based model has limitations for intermediates analysis in multistep routes.

Attention weight analysis examined correspondence with key reactive sites.[82] Attention heatmap for atogepant (Figure 6C inset) shows high attention weights at atoms 27 and 28 (3rd step, Knoevenagel reaction),[83] atom 21 (6th step, intramolecular nucleophilic substitution),[84] atoms 17 and 34 (7th step, carbonylation), atom 26 (8th step, hydrolysis deprotection), and atoms 5, 14, 15, 16 (9th step, amide condensation).[85] This correspondence between attention patterns and key reactions[86,87] suggests SynFrag predictions incorporate information about chemical reactivity and structural sensitivity.

The fragment assembly pretraining approach enabled SynFrag to capture incremental changes in synthetic complexity similar to synthetic-route-based SCScore and CASP tools, supporting SA prediction based on structural assembly patterns.

3.6. SynFrag Online Prediction Platform

To facilitate practical applications in drug discovery, we developed an open-access web platform for batch SA prediction. The platform accepts CSV files containing SMILES strings and returns SynFrag prediction scores with attention weight visualizations (Figure 7). Enable high-throughput screening, given that it is tens of thousands of times faster than CASP tools. For users requiring detailed retrosynthetic analysis, the platform provides integration with AiZynthFinder and SYNTHIA,[7,88] allowing transition from rapid SA screening to comprehensive route planning.

5. CONCLUSIONS

This study presents SynFrag, a molecular synthetic accessibility (SA) prediction model using fragment assembly autoregressive

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

generation pretraining. Through self-supervised learning on 9.18 million unlabeled molecules, SynFrag learns fragment connectivity patterns relevant to SA.

Evaluation across multiple test sets showed consistent performance. The model achieved great performance on the public test set (TS1−TS3), AUROC 0.945 on clinical drugs and intermediates (TSA) and 0.889 on AI-generated molecules (TSB). SynFrag provides subsecond predictions for high-throughput screening. The model distinguishes synthesis difficulty cliffs, where minor structural changes alter SA, which is relevant for lead optimization. Attention mechanisms show correspondence with key reactive sites, suggesting that predictions incorporate chemical reactivity intuition.

The web platform provides batch prediction with attention visualizations and integration with the CASP tools. This combination of prediction speed and interpretability addresses two scenarios in drug discovery: screening large AI-generated libraries for SA and evaluating complexity progression in multistep synthetic routes.

Considering the limitations of current SA assessment, future research can be carried out in the following directions: integrating commercial databases to incorporate reagent availability as a decisive factor; improving ES/HS annotations beyond synthetic step counts, such as yields and costs; constructing models for multiple SA-influencing metrics via multitask learning frameworks; and performing fine-grained SA prediction based on first-principles, such as predictions of reaction energy changes and reaction energy barriers.

As AI-driven molecular generation continues to develop, SynFrag represents an advanced attempt to bridge the gap between computational design and laboratory synthesis in drug discovery. SynFrag provides predictions at speeds suitable for large-scale screening while maintaining interpretability via attention mechanisms. Integrating this model with generative models enables SA-aware molecular design; coupling it with automated synthesis platforms can accelerate the experimental validation process.

■ ASSOCIATED CONTENT

Data Availability Statement

The pretraining &finetuning data set, public benchmark test sets, steroid compounds and scenario application test sets are available at https://github.com/simmzx/SynFrag/tree/main/data. All the codes of SynFrag are available at https://github.com/simmzx/SynFrag. The open-access SynFrag platform at https://synfrag.simm.ac.cn.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02450.

> Posthoc analysis of the impact of standardization on SynFrag; consistency of Retro* in relabeling TS1 and TS2; atom and bond features used in SynFrag; hyperparameter search configuration for SynFrag fine-tuning; statistical tests between SynFrag and other models; heatmap of the fingerprint similarities between the ES and HS compounds in 3 public test sets; and heatmap of AUROC on TS3 for different hyperparameter search configurations (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Dingyan Wang − *Lingang Laboratory, Shanghai 200031, China*; Email: wangdy@lglab.ac.cn

Xutong Li − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; University of Chinese Academy of Sciences, Beijing 100049, China*; orcid.org/0000-0001-9547-0643; Email: lixutong@simm.ac.cn

Authors

Xiang Zhang − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Jia Liu − *Nanjing University of Chinese Medicine, Nanjing 210023, China*

Bufan Xu − *Nanjing University of Chinese Medicine, Nanjing 210023, China; Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Zihan Zhang − *Information Management Office, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Zifu Huang − *Information Management Office, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Kaixian Chen − *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; Nanjing University of Chinese Medicine, Nanjing 210023, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02450

Author Contributions

X.Z.: Conceptualization; Data curation; Investigation; Software; Writing—original draft; Writing—review and editing. J.L.: Data curation; Validation. B.X.: Validation. Z.Z.: Resources. Z.H.: Resources. K.C.: Supervision. D.W.: Conceptualization; Supervision. X.L.: Conceptualization; Supervision; Writing—review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0830000, China), National Natural Science Foundation of China (82204278, China), National Key Research and Development Program of China (2023YFC2305904, China, and 2022YFC3400504, China), and the Lingang Laboratory (LGL-8888). We also acknowledge the Shanghai Supercomputer Center for providing computing resources.

REFERENCES

(1) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.

(2) Wang, M.; Hsieh, C.-Y.; Wang, J.; Wang, D.; Weng, G.; Shen, C.; Yao, X.; Bing, Z.; Li, H.; Cao, D.; Hou, T. Relation: A Deep Generative Model for Structure-based De Novo Drug Design. *J. Med. Chem.* **2022**, *65*, 9478−9492.

(3) Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular Design in Drug Discovery: A Comprehensive Review of Deep Generative Models. *Brief. Bioinform.* **2021**, *22*, No. bbab344.

(4) Schneider, P.; et al. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discovery* **2020**, *19*, 353−364.

(5) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chem. Sci.* **2020**, *11*, 154−168.

(6) Tu, Z.; Choure, S. J.; Fong, M. H.; Roh, J.; Levin, I.; Yu, K.; Joung, J. F.; Morgan, N.; Li, S.; Sun, X.; Morgan, M.; Liles, J. P.; Jensen, K. F.; Coley, C. W.; et al. ASKCOS: Open-source, Data-driven Synthesis Planning. *Acc. Chem. Res.* **2025**, *58*, 1764−1775.

(7) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-source Software for Retrosynthetic Planning. *J. Cheminform.* **2020**, *12*, 70.

(8) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) − Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. *Chem. Sci.* **2021**, *12*, 3339−3349.

(9) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.

(10) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252−261.

(11) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminform.* **2020**, *12*, 35.

(12) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *J. Chem. Inf. Model.* **2022**, *62*, 2973−2986.

(13) Wang, S.; Wang, L.; Li, F.; Bai, F. DeepSA: a deep-learning driven predictor of compound synthesis accessibility. *J. Cheminform.* **2023**, *15*, 103.

(14) Chen, S.; Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *J. Cheminform.* **2024**, *16*, 83.

(15) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60*, 5714−5723.

(16) Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. Critical Assessment of Synthetic Accessibility Scores in Computer-assisted Synthesis Planning. *J. Cheminform.* **2023**, *15*, 6.

(17) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(18) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.

(19) Gaulton, A.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(20) Shivanyuk, A. N.; Ryabukhin, S. V.; Tolmachev, A.; Stetsenko, S. Enamine Real Database: Making Chemical Diversity Real. *Chem. Today* **2007**, *25*, 58−59.

(21) Desai, P. V.; Patny, A.; Sabnis, Y.; Tekwani, B.; Gut, J.; Rosenthal, P.; Srivastava, A.; Avery, M. Identification of Novel Parasitic Cysteine Protease Inhibitors Using Virtual Screening. 1. The ChemBridge Database. *J. Med. Chem.* **2004**, *47*, 6609−6615.

(22) Yadav, R.; Imran, M.; Dhamija, P.; Suchal, K.; Handu, S. Virtual Screening, ADMET Prediction and Dynamics Simulation of Potential Compounds Targeting the Main Protease of SARS-CoV-2. *J. Biomol. Struct. Dyn.* **2021**, *39*, 6617−6632.

(23) Landrum, G. RDKit: Open-source Cheminformatics from Machine Learning to Chemical Registration. In *Abstr. Pap. Am. Chem. Soc.*, **2019**; p *258*.

(24) Chen, B.; Li, C.; Dai, H.; Song, L. Retro∗: Learning Retrosynthetic Planning with Neural Guided A∗ Search. In *Proceedings of the 37th International Conference on Machine Learning*, 2020; pp 1608−1616.

(25) Klucznik, T.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4*, 522−532.

(26) Voršilák, M.; Svozil, D. Nonpher: Computational Method for Design of Hard-to-synthesize Structures. *J. Cheminform.* **2017**, *9*, 20.

(27) Kutchukian, P. S.; Shakhnovich, E. I. De Novo Design: Balancing Novelty and Confined Chemical Space. *Expert Opin. Drug Discovery* **2010**, *5*, 789−812.

(28) Flick, A. C.; Ding, H. X.; Leverett, C. A.; Fink, S. J.; O'Donnell, C. J. Synthetic Approaches to New Drugs Approved During 2016. *J. Med. Chem.* **2018**, *61*, 7004−7031.

(29) Flick, A. C.; Leverett, C. A.; Ding, H. X.; McInturff, E.; Fink, S. J.; Helal, C. J.; O'Donnell, C. J. Synthetic Approaches to the New Drugs Approved During 2017. *J. Med. Chem.* **2019**, *62*, 7340−7382.

(30) Flick, A. C.; Leverett, C. A.; Ding, H. X.; McInturff, E.; Fink, S. J.; Helal, C. J.; DeForest, J. C.; Morse, P. D.; Mahapatra, S.; O'Donnell, C. J. Synthetic Approaches to New Drugs Approved during 2018. *J. Med. Chem.* **2020**, *63*, 10652−10704.

(31) Yuan, S.; Yu, B.; Liu, H.-M. New Drug Approvals for 2019: Synthesis and Clinical Applications. *Eur. J. Med. Chem.* **2020**, *205*, No. 112667.

(32) Yuan, S.; Luo, Y.-Q.; Zuo, J.-H.; Liu, H.; Li, F.; Yu, B. New Drug Approvals for 2020: Synthesis and Clinical Applications. *Eur. J. Med. Chem.* **2021**, *215*, No. 113284.

(33) Yuan, S.; Wang, D.-S.; Liu, H.; Zhang, S.-N.; Yang, W.-G.; Lv, M.; Zhou, Y.-X.; Zhang, S.-Y.; Song, J.; Liu, H.-M. New Drug Approvals for 2021: Synthesis and Clinical Applications. *Eur. J. Med. Chem.* **2023**, *245*, No. 114898.

(34) Yuan, S.; Shen, D.-D.; Jia, R.; Sun, J.-S.; Song, J.; Liu, H.-M. New Drug Approvals for 2022: Synthesis and Clinical Applications. *Med. Res. Rev.* **2023**, *43*, 2352−2391.

(35) Wang, Y.-T.; Yang, P.-C.; Zhang, Y.-F.; Sun, J.-F. Synthesis and Clinical Application of New Drugs Approved by FDA in 2023. *Eur. J. Med. Chem.* **2024**, *265*, No. 116124.

(36) Jensen, J. H. A Graph-based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10*, 3567−3572.

(37) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.

(38) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2224−2232.

(39) Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014; pp 1724−1735.

(40) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* 2015.

(41) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-like' Chemical Fragment Spaces. *ChemMedChem.* **2008**, *3*, 1503−1507.

(42) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.
(43) Wang, Y.; Pang, C.; Wang, Y.; Jin, J.; Zhang, J.; Zeng, X.; Su, R.; Zou, Q.; Wei, L. Retrosynthesis Prediction with an Interpretable Deep-learning Framework Based on Molecular Assembly Tasks. *Nat. Commun.* **2023**, *14*, 6155.
(44) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018; pp 2323−2332.
(45) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4*, 1256−1264.
(46) Hafidi, H.; Ghogho, M.; Ciblat, P.; Swami, A. GraphCL: Contrastive Self-supervised Learning of Graph Representations. *arXiv preprint arXiv:2007.08025* 2020.
(47) Bengio, Y.; Léonard, N.; Courville, A. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432* 2013.
(48) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-directed Molecular Graph Generation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6410−6421.
(49) Kong, X.; Huang, W.; Tan, Z.; Liu, Y. Molecule Generation by Principal Subgraph Mining and Assembling. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2550−2563.
(50) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.
(51) Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; Lee, C. K. Motif-based Graph Self-supervised Learning for Molecular Property Prediction. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15870−15882.
(52) Wang, X.; et al. Generic Interpretable Reaction Condition Predictions with Open Reaction Condition Datasets and Unsupervised Learning of Reaction Center. *Research* **2023**, *6*, 0231.
(53) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* 2015.
(54) Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983* 2017.
(55) Pascanu, R.; Mikolov, T.; Bengio, Y. On the Difficulty of Training Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Machine Learning*. 2013; pp 1310−1318.
(56) Bergstra, J.; Bengio, Y. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281−305.
(57) Prechelt, L. In *Neural Networks: Tricks of the Trade*. Montavon, G.; Orr, G. B., Müller, K.-R., Eds.; Springer: Berlin, 2012; pp 53−67.
(58) Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006; pp 233−240.
(59) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.
(60) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.
(61) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* 2019.
(62) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **2022**, *4*, 279−287.
(63) Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50−60.
(64) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(65) Calcaterra, A.; D'Acquarica, I. The Market of Chiral Drugs: Chiral Switches versus De Novo Enantiomerically Pure Compounds. *J. Pharm. Biomed. Anal.* **2018**, *147*, 323−340.
(66) Demchenko, A. V. *Handbook of Chemical Glycosylation: Advances in Stereoselectivity and Therapeutic Relevance*; Wiley-VCH: Weinheim, 2008.
(67) Zhu, X.; Schmidt, R. R. New Principles for Glycoside-bond Formation. *Angew. Chem., Int. Ed.* **2009**, *48*, 1900−1934.
(68) Gaich, T.; Baran, P. S. Aiming for the Ideal Synthesis. *J. Org. Chem.* **2010**, *75*, 4657−4673.
(69) Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J. Am. Chem. Soc.* **1963**, *85*, 2149−2154.
(70) Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **1964**, *6*, 241−252.
(71) An, J.; Jackson, I. R. K.; Tuccinardi, J. P.; Wood, J. L. Pyrroloiminoquinone Alkaloids: Total Synthesis of Makaluvamines A and K. *Org. Lett.* **2023**, *25*, 1868−1871.
(72) Sharma, V.; Kumar, P.; Pathak, D. Biological Importance of the Indole Nucleus in Recent Years: A Comprehensive Review. *J. Heterocycl. Chem.* **2010**, *47*, 491−502.
(73) Hatcher, J. M.; Wang, E. S.; Johannessen, L.; Kwiatkowski, N.; Sim, T.; Gray, N. S. Development of highly potent and selective steroidal inhibitors and degraders of CDK8. *ACS medicinal chemistry letters* **2018**, *9*, 540−545.
(74) Sun, J.; Chen, Y.; Ragab, S. S.; Gu, W.; Tang, Z.; Tang, Y.; Tang, W. Total Syntheses of Polyhydroxylated Steroids by an Unsaturation-Functionalization Strategy. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202303639.
(75) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443−4458.
(76) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* **2014**, *19*, 1069−1080.
(77) Dimova, D.; Stumpfe, D.; Bajorath, J. Quantifying the Fingerprint Descriptor Dependence of Structure-activity Relationship Information on a Large Scale. *J. Chem. Inf. Model.* **2013**, *53*, 2275−2281.
(78) Engel, P. S. Mechanism of the Thermal and Photochemical Decomposition of Azoalkanes. *Chem. Rev.* **1980**, *80*, 99−150.
(79) Clayden, J.; Greeves, N.; Warren, S. *Organic Chemistry*, 2nd ed.; Oxford University Press: Oxford, 2012.
(80) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.
(81) Lipton, R. B.; Goadsby, P. J.; Smith, J.; Schaeffler, B. A.; Biondi, D. M.; Hirman, J.; Pederson, S.; Allan, B.; Cady, R. Efficacy and Safety of Eptinezumab in Patients with Chronic Migraine: PROMISE-2. *Neurology* **2020**, *94*, e1365−e1377.
(82) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-based Neural Networks. *Nat. Mach. Intell.* **2021**, *3*, 144−152.
(83) van Beurden, K.; de Koning, S.; Molendijk, D.; van Schijndel, J. The Knoevenagel Reaction: A Review of the Unfinished Treasure Map to Forming Carbon−carbon Bonds. *Green Chem. Lett. Rev.* **2020**, *13*, 349−364.
(84) Tomilov, Y. V.; Menchikov, L. G.; Novikov, R. A.; Ivanova, O. A.; Trushkov, I. V. Methods for the Synthesis of Donor-acceptor Cyclopropanes. *Russ. Chem. Rev.* **2018**, *87*, 201−250.
(85) Montalbetti, C. A. G. N.; Falque, V. Amide Bond Formation and Peptide Coupling. *Tetrahedron* **2005**, *61*, 10827−10852.
(86) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2*, No. 015016.
(87) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-convolutional

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370−377.
(88) Sinha, S.; Obkircher, M.; Yaragani, M.; Deokar, R.; Kumar, S.; Gardener, E. Computer-assisted Synthetic Route Optimization Using SYNTHIA™Retrosynthesis Software. *Org. Process Res. Dev.* **2024**, *28*, 1196−1205.

https://doi.org/10.1021/acs.jcim.5c02450
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX