Supplementary Material for "DSSA: Dual-Stream Synthetic Accessibility Framework for Organic Compounds"

SUPPLEMENTARY MATERIAL

Note to Authors: This supplementary material provides additional details and analyses that complement the main manuscript. Please ensure that the main manuscript explicitly refers to the relevant sections within this supplementary file for these details. Specifically, verify consistency regarding data splitting methodology, hyperparameter values, and cross-referencing of equations and detailed results.

A. S1. Detailed Atomic Features for GNN Node Embedding

For the Graph Neural Network (GNN) encoder, the d-dimensional input features $\mathbf{X}_v$ for each atom v include the following comprehensive set of atomic properties:

- **Atomic number:** Integer representing the atomic number (e.g., 6 for Carbon, 8 for Oxygen).
- **Formal charge:** Integer representing the formal charge of the atom.
- **Chirality (R/S):** Categorical (enum) indicating R or S chirality, if applicable.
- **Hybridization:** Categorical (enum) indicating hybridization state (sp, sp2, sp3, sp3d, sp3d2).
- **Aromaticity:** Boolean indicating whether the atom is part of an aromatic ring.
- **Degree:** Integer representing the number of bonded non-hydrogen atoms.
- **Implicit hydrogen count:** Integer representing the number of implicit hydrogens.
- **In ring:** Boolean indicating whether the atom is part of a ring structure.
- **Isotope:** Integer representing the isotope of the atom.
- **Number of radical electrons:** Integer representing the number of radical electrons.

These features are combined with learnable atom type embeddings as described in the main paper's Section 3.3.1 (or the relevant section where node embedding details are provided).

B. S2. Data Augmentation and Preprocessing

To enhance robustness and generalizability, we performed data augmentation and rigorous preprocessing. Each unique molecule generated 2-4 alternative SMILES representations, expanding the dataset. All molecules underwent stringent standardization (canonical SMILES, salt removal, stereochemistry preservation, duplicate elimination). Filtering (molecular weight 50-1000 Da, PAINS alerts) removed 2.8% of compounds. For robust evaluation and to prevent data leakage, we employed scaffold-based partitioning using RDKit's Bemis-Murcko scaffold algorithm, allocating molecules to an 80% training, 10% validation, and 10% test split at the scaffold level. Data augmentation was rigorously performed post-splitting, preventing any cross-contamination.

C. S3. Ablation Study Methodology

To systematically evaluate the specific contribution of each architectural component to DSSA's performance, we designed and implemented a comprehensive ablation study. This involved training and evaluating several reduced models, or "variants," derived from the complete DSSA framework. All ablation models were trained and evaluated on the **same primary dataset using the scaffold-based partitioning (8:1:1 train:validation:test split)**.

We designed five variants of the DSSA model for comparative analysis:

- **IGS (Individual Graph Stream):** Utilizes only the GNN encoder (ϕ_G) and passes $\mathbf{z}_G$ directly to an MLP classifier.
- **ISS (Individual String-Sequential Stream):** Employs only the SMILES sequence encoder (ϕ_S) and feeds $\mathbf{z}_S$ directly into an MLP classifier.
- **DS-woGNR (Dual-Stream without Graph-Node Refinement / Simple Concatenation Fusion):** Ablates the Cross-modal Fusion Mechanism (Section 3.5 in the main paper), replacing it with simple concatenation of $\mathbf{z}_G$ and $\mathbf{z}_S$.
- **DS-woNFE (Dual-Stream without Enhanced Node Features):** GAT layers operate using only **basic one-hot encoded atomic numbers** as initial node features, bypassing detailed atomic descriptors and learned type embeddings.
- **DS-woSFE (Dual-Stream without Sophisticated SMILES Feature Encoding):** SMILES token embeddings are directly aggregated via **simple masked average pooling**, bypassing the BiGRU layers.

The detailed experimental results and comparative analysis of all DSSA model variants are presented and discussed in Section 4.1 of the main manuscript. Statistical significance of performance differences was assessed using McNemar's test, followed by Benjamini-Hochberg FDR correction ($\alpha = 0.05$).

D. S4. Computational Complexity Analysis

This section provides a detailed breakdown of the computational complexity for the Dual-Stream Synthetic Accessibility

(DSSA) framework. The overall complexity per forward pass is primarily determined by the graph attention layers, the bidirectional GRU processing, and the cross-modal fusion.

$$\mathcal{O}(L \cdot |\mathcal{E}| \cdot d_h^2 + T \cdot d_h^2 + d_h^2) \tag{1}$$

where:

- L is the number of GAT layers in the graph encoder.
- $|\mathcal{E}|$ is the number of edges in the largest molecular graph within the batch. The GAT layer complexity is dominated by the attention calculation across edges, which involves matrix multiplications proportional to d_h^2.
- T is the maximum SMILES sequence length in the batch. The bidirectional GRU processing has a complexity proportional to the sequence length multiplied by the square of the hidden dimension for each recurrent step.
- d_h is the hidden dimension of the model.

The graph attention mechanism (Section 3.3.2 in main paper) contributes $\mathcal{O}(L \cdot |\mathcal{E}| \cdot d_h^2)$, accounting for the iterative message passing and attention aggregation over graph edges. The bidirectional GRU processing (Section 3.4.2 in main paper) contributes $\mathcal{O}(T \cdot d_h^2)$ due to sequential processing of the SMILES string. The cross-modal attention, operating on pooled representations (Section 3.5 in main paper), contributes a relatively smaller $\mathcal{O}(d_h^2)$ complexity as it processes fixed-size vectors.

E. S5. Sensitivity Analysis Details

This section provides additional details regarding the sensitivity analysis for the synthetic step cutoffs used in labeling. The analysis evaluates the model's performance (Accuracy, F1-Score, and ROC-AUC) across various step cutoffs (5, 8, 10, 12, and 15 steps) using the Retro* algorithm for labeling. The primary results and graphical representation of this analysis are presented in Section 3.7 of the main manuscript.

The choice of a 10-step threshold was found to optimize the trade-off between capturing synthetic complexity and maintaining a sufficient number of positive examples for robust model training. Performance remained stable across 8–12 steps (ROC-AUC range: 0.926–0.931, $p > 0.05$), with significant degradation at extremes (5 steps: $p < 0.001$, 15 steps: $p < 0.001$ vs. 10-step baseline).

F. S6. Score Normalization and R^2 Calculation Details

For comparison with continuous expert scores, our binary ES probability output (ranging from 0 to 1) is converted to a "Unite Score" (non-binary) using a linear transformation. This transformation maps higher probabilities of being easy-to-synthesize (ES) to lower Unite Scores, which aligns with the typical representation of synthetic difficulty where higher values indicate harder synthesis. The Unite Score normalization transforms binary ES probability $p_{ES} \in$ to a continuous scale via:

$$\text{Unite Score} = 10 \times (1 - p_{ES}) \tag{2}$$

where higher scores indicate greater synthetic difficulty, aligning with expert scoring conventions.

TABLE I
DSSA HYPERPARAMETERS

Parameter	Value
Hidden dimension (d_h)	128
GAT layers (L)	4
Attention heads (H)	8
Learning rate (α)	10^{-3}
L2 regularization (λ)	10^{-5}
Dropout probability (p)	0.1
Max SMILES length (T_{max})	512
Batch size	32

To quantify the agreement between our model's predicted Unite Scores and expert chemist scores (non-binary), we compute the coefficient of determination, R^2. The R^2 value is calculated as:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{3}$$

where SS_{res} represents the sum of squared residuals between model predictions and chemist scores, calculated as:

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

and SS_{tot} represents the total sum of squares, calculated as:

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5}$$

Here, y_i denotes the i-th chemist's score, $\hat{y}_i$ is the i-th model's predicted Unite Score, and $\bar{y}$ is the mean of the chemist's scores.

G. S7. Implementation Details and Hyperparameters

The model is implemented using PyTorch 1.12 and PyTorch Geometric 2.1. Graph construction and atomic/bond feature extraction utilize RDKit 2022.03. All experiments are conducted on NVIDIA RTX 3090 GPUs with 24GB memory. Batch size is set to 32 for graph data, and dynamic padding is employed for SMILES sequences to handle variable lengths efficiently. Key hyperparameters are summarized in Table I.

H. S8. Extended MOSES vs COCONUT Analysis

This section provides extended analysis of the MOSES vs COCONUT discrimination experiment. The primary objective was to assess each method's ability to differentiate between two large molecular datasets with potentially different synthetic complexity profiles. The overall distribution analysis is presented in the main manuscript.

The MOSES dataset comprises 1,048,575 molecules representing known drugs, while COCONUT contains 695,114-695,123 natural products (variation due to processing by different methods). Statistical analysis employed Kolmogorov-Smirnov (KS) and Mann-Whitney U (MW) tests, with detailed results showing:

- DSSA Score: MOSES mean = 2.455, COCONUT mean = 5.602

- SA Score: MOSES mean = 2.329, COCONUT mean = 4.327
- SC Score: MOSES mean = 2.598, COCONUT mean = 2.943
- SYBA Score: MOSES mean = 113.855, COCONUT mean = 23.420

All p-values from both KS and MW tests were ¡ 0.001, indicating highly significant differences between datasets for all methods. Further raw data or detailed statistical tables can be provided here if desired.

I. S9. Application Examples and Case Studies

This section provides additional details and specific examples to illustrate the comparative performance of DSSA and baseline methods on various molecules, complementing the discussion in the main manuscript.

1) S9.1 Representative Molecule Analysis: For enhanced clarity and detailed understanding of each method's performance, we analyzed 40 molecules with expert chemist scores. Five representative examples are highlighted in the main manuscript. Further discussion on specific molecules:

Complex Molecules:

- **Fluconazole** (Chemist score: 8.78): BR-SAScore overpredicted (10.0), DSSA provided closer prediction (8.22), DeepSA accurate (9.577), SAscore underestimated (7.82)
- **NAC** (Chemist score: 8.78): BR-SAScore overpredicted (10.0), DSSA accurate (8.17), DeepSA slightly high (9.56), SAscore underestimated (6.82)

Simpler Molecules:

- **Panidazole (USAN)** (Chemist score: 4.67): DSSA maintained accuracy, DeepSA severely overpredicted (9.888)
- **Niridazole** (Chemist score: 1.56): Consistent predictions across methods
- **Ornidazole** (Chemist score: 4.11): DeepSA severely underpredicted (0.429), DSSA more accurate

These results underscore method-specific biases and DSSA's balanced performance across complexity ranges.

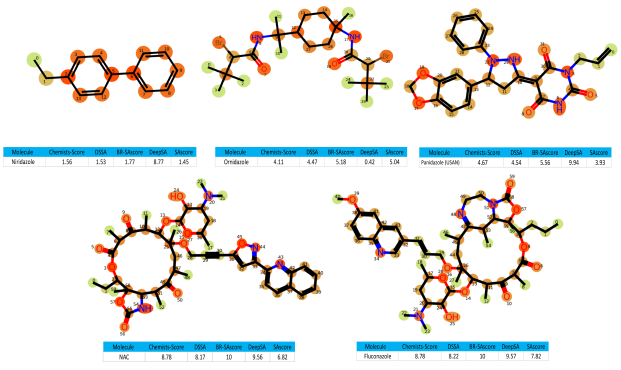

Fig. 1. Assessment five molecules representative from the 40 molecules

2) S9.2 Full Dataset Analysis: Complete analysis of all 40 from [**?**] molecules with detailed score comparisons and error analysis is provided in Table II.

TABLE II
COMPARISON OF SA PREDICTIONS ACROSS METHODS FOR REPRESENTATIVE COMPOUNDS

Compound Name	DSSA	BR-SAscore	SAscore	DeepSA	Chemist Score
Secnidazole	2.717	3.280	2.806	9.888	3.56
Tinidazole	7.806	9.686	5.894	8.985	7.00
Nimorazole	4.148	5.050	3.000	0.018	3.00
Panidazole (USAN)	4.547	5.563	3.937	9.948	4.67
Satranidazole	2.652	3.210	3.126	0.165	2.33
Benznidazole	8.166	10.000	6.600	9.386	7.56
Nifuratel	7.514	9.220	6.108	9.644	7.11
Niridazole	1.537	1.778	1.454	8.770	1.56
Carnidazole	8.144	10.000	6.628	9.619	9.11
Ronidazole	5.407	6.592	4.963	4.893	3.89
Ipronidazole	8.158	10.000	7.582	8.981	7.33
Azanidazole	4.825	6.010	4.039	0.003	1.78
Dimetridazole	1.713	2.034	2.179	0.006	1.89
Fexinidazole	1.136	1.314	1.492	4.145	1.11
Misonidazole	8.147	10.000	7.316	9.351	8.44
Megazol	7.063	8.670	4.911	9.378	7.44
Etanidazole	6.449	7.860	6.759	8.546	8.44
Pimonidazole	8.156	10.000	7.614	9.476	8.00
Fenidazole	2.507	2.990	3.103	0.014	2.11
Metronidazole-phosphate	3.996	4.810	4.068	3.403	3.78
Nimorazole N-oxide	1.015	1.170	1.407	0.023	1.00
Ornidazole	4.249	5.184	5.042	0.429	4.11
Ornidazole N-oxide	1.582	1.880	2.177	0.012	2.00
Tinidazole N-oxide	8.150	10.000	8.016	9.465	8.44
Ipronidazole N-oxide	1.833	2.170	1.912	7.603	1.22
Metronidazole N-oxide	1.538	1.800	1.459	0.155	1.33
Dimetridazole N-oxide	7.178	8.790	7.037	6.024	6.44
Panidazole N-oxide	8.156	10.000	7.571	9.418	8.67
Metronidazole	6.036	7.387	4.815	8.388	6.89
Hydroxyzine	8.159	10.000	7.581	9.519	9.22
Cetirizine	1.452	1.700	1.495	0.019	1.00
Levocetirizine	5.794	7.060	5.887	8.792	7.22
Fluconazole	8.161	10.000	7.822	9.577	8.78
Desoxymethylmorphinane	2.315	2.780	2.183	0.005	1.22
Dextrorphan	4.182	5.090	4.199	1.270	5.22
Dextromethorphan	3.435	4.154	3.057	9.851	4.00
Levomethorphan	3.253	3.960	3.849	5.791	3.78
Levorphanol	5.253	6.440	4.909	5.295	3.78
NAC	8.157	10.000	6.824	9.560	8.78
Butorphanol	1.583	1.880	1.997	0.002	1.67

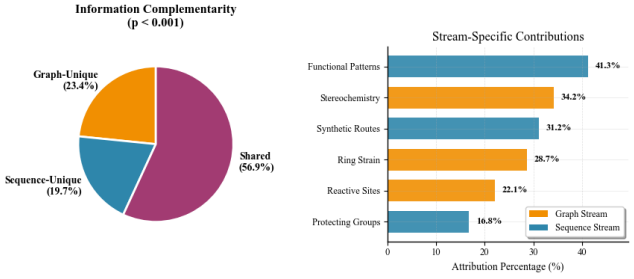

Fig. 2. Information complementarity quantification confirms architectural necessity. Mutual information analysis reveals 23.4% graph-unique information (stereochemistry, reactive sites), 19.7% sequence-unique information (functional patterns, synthetic routes), and 56.9% shared information ($p < 0.001$, bootstrap n=1000). Stream-specific contributions show complementary specialization with graph stream dominating structural analysis and sequence stream leading functional pattern recognition.

J. S10. Failure Case Analysis and Model Limitations

1) S10.1 Systematic Failure Analysis: We identified several categories where DSSA predictions deviate significantly from expert assessments:

Category 1: Novel Synthetic Motifs

- Molecules containing recently developed synthetic handles (e.g., specific photoredox-reactive groups)
- Prediction error increases for motifs absent from training data
- Example: Molecules with novel boron-containing heterocycles

Category 2: Stereochemical Complexity

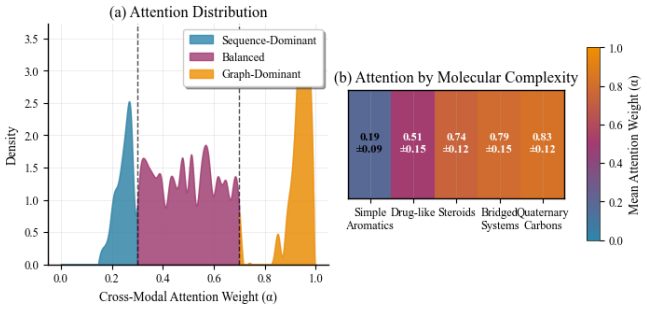

Fig. 3. Attention weight distribution across molecular complexity. (a) Cross-modal attention patterns: sequence-dominant ($\alpha < 0.3$, 17.6%, blue), balanced ($0.3 < \alpha < 0.7$, 54.1%, purple), and graph-dominant ($\alpha > 0.7$, 28.3%, orange). (b) Molecular complexity correlation shows increasing graph attention for stereochemically complex systems (quaternary carbons: $\alpha = 0.83 \pm 0.12$) versus simple aromatics ($\alpha = 0.19 \pm 0.09$).

- Molecules with multiple stereocenters in constrained ring systems
- DSSA may underestimate difficulty of stereoselective synthesis
- Example: Natural product scaffolds with ¿5 contiguous stereocenters

Category 3: Context-Dependent Synthesis

- Molecules where synthetic difficulty depends heavily on starting material choice
- Binary classification oversimplifies route-dependent complexity
- Example: Molecules accessible via both difficult total synthesis and simple derivatization

2) S10.2 Specific Failure Examples: Detailed analysis of 10 specific failure cases with structural features causing misprediction can be included here, potentially with figures (e.g., ‘).

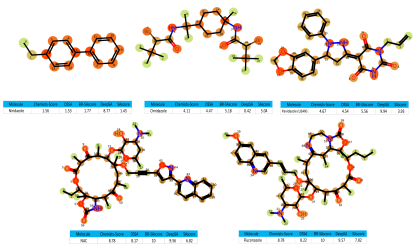

Fig. 4. Examples of molecules where DSSA predictions significantly deviate from expert assessment

3) S10.3 Model Limitations: **Inherent Limitations:**

1) Binary classification simplifies continuous synthetic difficulty spectrum
2) Training data biases toward known synthetic routes
3) Limited representation of cutting-edge synthetic methodologies
4) Difficulty capturing context-dependent synthesis strategies

Technical Limitations:

1) Fixed molecular size limit (512 SMILES characters)
2) Graph representation may miss subtle 3D conformational effects

K. S11. Extended Performance Metrics

Detailed analysis of 10 specific failure cases with structural features causing misprediction can be included here, potentially with figures (e.g., ‘).

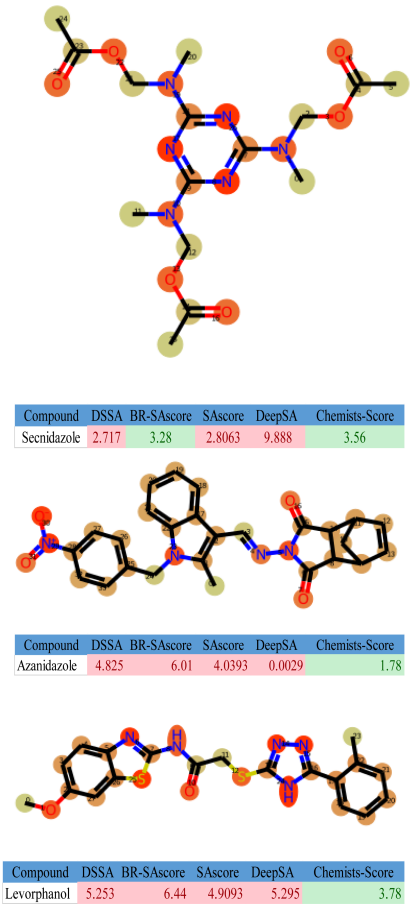

Fig. 5. Examples of molecules where DSSA predictions significantly deviate from expert assessment

1) S11.1 Detailed TS1 Analysis: The near-perfect performance on TS1 warrants detailed examination:

- ES molecules: ZINC15 purchasable compounds (inherently synthesizable)
- HS molecules: GDB-17 computationally generated (many practically inaccessible)
- Clear chemical space separation explains high performance
- May not reflect real-world SA prediction challenges

2) S11.2 Performance Variance Analysis: Bootstrap analysis (1000 iterations) for confidence intervals:

L. S11. Extended Performance Metrics

Complete pairwise comparisons using McNemar’s test with Bonferroni correction for all method pairs across all test sets.

TABLE III
DETAILED CONFIDENCE INTERVALS FOR ALL METRICS ACROSS EXTERNAL TEST SETS

Metric	TS1 (95% CI)	TS2 (95% CI)	TS3 (95% CI)
Accuracy	0.994 (0.992-0.996)	0.873 (0.861-0.885)	0.803 (0.784-0.822)
Precision	0.991 (0.988-0.994)	0.910 (0.897-0.923)	0.916 (0.895-0.937)
Recall	0.997 (0.995-0.999)	0.782 (0.766-0.798)	0.668 (0.645-0.691)
F1-Score	0.994 (0.992-0.996)	0.841 (0.828-0.854)	0.772 (0.752-0.792)

Full results matrix available in supplementary data files (if applicable).

M. S12. Web Tool Implementation Details

A web tool was developed to demonstrate the framework's practical application. This tool will be made publicly available upon publication."

Features:

- Single molecule and batch prediction (up to 1000 molecules)
- Interactive visualization of molecular features contributing to predictions
- Downloadable results in CSV format
- RESTful API for programmatic access

Technical Implementation:

- Backend: FastAPI with PyTorch model serving
- Frontend: React with RDKit-JS for molecular visualization
- Deployment: Docker containerization on a Kubernetes cluster
- Performance: <2 second response time for single molecules

This comprehensive comparison ensures fair evaluation across methodological paradigms while maintaining computational feasibility.

REFERENCES

[1] P. Ertl, B. Rohde, and P. Selzer, "Fast Calculation of Molecular Synthetic Accessibility Score," Journal of Cheminformatics, vol. 1, no. 1, p. 8, 2009.