Supplementary Information (SI) for Chemical Science.
This journal is © The Royal Society of Chemistry 2026

Supplementary Notes

Implementations of the competitive baselines

In this work, we adapted several widely used and powerful molecule generators for comparison, including JTVAE, VJTNN, MoLeR, and a control model (random generation). For these methods, we used the officially released pre-trained models and adapted them to perform few-shot optimization tasks. Specifically, for a generative model with latent space $\mathcal{Z}$, we performed K-shot optimization by first training a small surrogate regressor on the latent representations of the K modification cases. For a query molecule, we encoded it into the latent space and then applied gradient-based optimization to its latent vector to match the property change direction (increase or decrease) indicated by the few-shot examples. The resulting optimized latent vector was then decoded to generate the new molecule. This procedure was applied consistently across JTVAE, VJTNN, and MoLeR. The random generation model was implemented by random sampling based on the latent space of the pre-trained JTVAE.

We further incorporated advanced large language models as benchmarks, including ChatGLM, ChatGPT3.5, and GPT-4. These models require carefully constructed prompts to generate relevant output data. In crafting these prompts, we integrated the characteristic descriptions of the molecules to be optimized and instances of a few-shot optimization to ensure that the model can accurately understand the task requirements. For ChatGLM, we used the officially released code and pre-trained models, deployed and executed on a local server environment, to obtain the model prediction results. In contrast, for ChatGPT3.5 and GPT-4, we utilized their online service capabilities by transmitting the processed input data through their official API interfaces, thereby eliciting the corresponding model outputs from remote servers.

Implementations of the biological activity predictors

To evaluate the generated molecules on biological activities, we employed ChemProp, a message-passing neural network model that represents molecules as graphs with atoms as nodes and bonds as edges. For each target biological activity, assays with a sufficient number of molecules were extracted from the ChEMBL database. Given the diversity of the bioassay targets, we implemented an automated, multi-round training pipeline to develop a robust and optimized predictor for each specific assay, rather than using a single, fixed architecture.

Prior to training, data underwent specific preprocessing: activity values exhibiting a wide dynamic range were $\log_{10}$-transformed, followed by StandardScaler normalization. The fitted scaler was saved for subsequent inverse transformation of predictions. This automated pipeline then iteratively trained models with progressively increasing complexity. Initial rounds employed a baseline architecture (e.g., hidden size 600, 50 epochs), while subsequent rounds automatically scaled up the model's capacity (e.g., hidden size up to 1800, epochs up to 200, and ensemble sizes up to 5) and incorporated 2D RDKit features. Models were trained using the Adam optimizer (ChemProp defaults) with an adaptive early stopping patience (increasing from 5 to 20 epochs). For each training round, the assay data were split into a 90.00% training/validation set and a 10.00% held-out test set. The 90.00% portion was then split internally by ChemProp using scaffold-balanced splitting for model training and validation.

The pipeline automatically tracked the performance (Pearson correlation) on the 10.00% held-out test set. This iterative training process continued until the predictor's performance on this set met or exceeded our target threshold of 0.75. As shown in Table S5, this process yielded predictors with strong correlations, enabling reliable evaluation of the generated molecules. During DrugLLM evaluation, the twenty target biological activities used for testing were excluded from DrugLLM's training set, allowing an unbiased assessment of few-shot optimization on previously unseen targets.

Implementation details of Functional Group Tokenization (FGT)

To facilitate a precise understanding of the Functional Group Tokenization (FGT) framework, we provide the formal definitions of the variables and functions utilized in Algorithm S1 and Algorithm S2.

Notation and Variables: Let $G = (V, E)$ denote the molecular graph, where V represents the set of atoms (nodes) and E represents the set of chemical bonds (edges). The output FGT sequence S is a canonical string representation consisting of a core token followed by a series of fragment tokens and their attachment descriptors, separated by "/". The set of structural groups $\mathcal{G}$ consists of chemically meaningful fragments (e.g., fused rings, functional motifs) identified during decomposition. During the encoding stage, we maintain a connection stack $\mathcal{C}$, which is an ordered list of tuples (g_i, a_m, a_g); the order of $\mathcal{C}$ follows the sequence of fragment removal and is reversed during decoding to ensure correct reconstruction. The central scaffold of the molecule, g_0 (Core Group), is the largest fused ring system that remains after all peripheral groups have been pruned. To handle connectivity, we define a_m and a_g as topological attachment descriptors. These are multi-dimensional signatures based on Breadth-First Search (BFS) atom environments: a_m identifies the attachment point on the main structure, and a_g identifies the point within the isolated fragment. Unlike volatile atom indices, these descriptors are invariant to SMILES canonicalization order. Finally, v_m and v_g represent the resolved atom nodes, where the function LocateAtom maps the descriptors a_m and a_g back to specific nodes in the graph objects during decoding.

Algorithmic Logic and Robustness: To ensure that one molecule always maps to exactly one FGT string, we implement two layers of canonicalization: initial SMILES canonicalization to standardize input, and lexicographical sorting of fragments during the removal check. If multiple fragments are eligible for removal, the one with the lowest lexicographical rank is processed first. A significant technical contribution of FGT is its resilience to index shifting. By using the BFS-based descriptors (a_m, a_g) instead of raw integer indices, the algorithm effectively captures the "local chemical environment." This allows the decoding process to reliably re-identify the correct splicing site even if the underlying software re-indexes the atoms during fragment instantiation.

Comparative evaluation of tokenization schemes

To validate the effectiveness of our proposed Functional Group Tokenization (FGT) scheme, particularly its suitability for generative language modeling, we conducted a comparative analysis against two established fragmentation algorithms: RECAP and BRICS. The evaluation focused on quantifying the efficiency (trade-offs between coverage and vocabulary size) and expressiveness (coverage) of FGT relative to these methods.

A subset of the ZINC database, consisting of approximately 400,000 drug-like molecules, was used for the analysis. Each molecule in the dataset was tokenized using FGT, RECAP, and BRICS. Identical preprocessing steps were applied to all methods to ensure a fair comparison. Performance was measured by:

- **Molecular coverage:** The fraction of molecules successfully decomposed by the algorithm.
- **Fragment coverage:** The fraction of all generated fragments included in the algorithm's vocabulary.
- **Vocabulary size:** The total number of unique fragment tokens required to represent the dataset.

Table S2 summarizes the decomposition performance. FGT achieved near-total molecular coverage (99.95%) and fragment coverage (99.99%) with a remarkably compact vocabulary of only 4,796 unique tokens. In contrast, BRICS achieved 98.34% molecular coverage but required a much larger vocabulary of 32,874 tokens. RECAP performed the poorest, covering only 64.75% of molecules and generating an impractical vocabulary exceeding 500,000 tokens. These results demonstrate that FGT provides a highly efficient and expressive molecular representation. It achieves comprehensive coverage with a minimal and dense vocabulary. This compactness is critical for the effective training of a generative language model. The vocabularies produced by retrosynthetic-focused methods (BRICS, RECAP) are significantly larger and sparser, making them less suitable for this purpose.

Evaluation of generative quality in molecular optimization

We evaluated the generative quality of DrugLLM in a few-shot molecular optimization setting, generating chemically valid, unique, and novel molecules relative to the provided leads, rather than under a purely de novo generation scenario. This evaluation employed the K-pair few-shot learning framework previously used for molecular optimization (detailed in Fig. 1a), where the model was conditioned on K molecular modification pairs and a target molecule. All molecules were encoded using our Functional Group Tokenization (FGT) scheme, which systematically decomposes structures into a hierarchical vocabulary of interpretable fragments. This representation mitigates the cyclic dependencies inherent in SMILES and facilitates consistent, structured reasoning.

Molecular generation was conducted in an autoregressive manner using the Hugging Face Transformers library on a single GPU with half-precision (FP16) computation. The key hyperparameters were set as follows: a sampling temperature of 0.90, a nucleus sampling threshold (Top-p) of 0.80, and a maximum of 128 newly generated tokens. For comparative analysis, several representative baseline models were evaluated under the same few-shot optimization protocol and identical generation settings. We quantitatively assessed the generated molecules using three standard metrics:

- **Validity:** The proportion of generated FGT sequences that are verified as chemically valid by RDKit.
- **Uniqueness:** The percentage of unique molecules among all validly generated molecules.
- **Novelty:** Fraction of validly generated molecules that are not present in the input K-pair examples, reflecting conditional novelty relative to the few-shot context.

As detailed in Table S3, DrugLLM achieves 100.00% validity, 99.49% uniqueness, and 99.02% novelty, meeting or exceeding the performance of all baseline models under this setting. These results indicate that DrugLLM maintains high fidelity and diversity while performing targeted molecular optimization, supporting reliable few-shot property improvement and demonstrating the practical utility of the FGT representation in structured molecular design.

Scalability analysis of FGT on macrocyclic structures

To evaluate the scalability and robustness of the FGT scheme, we conducted a systematic analysis of vocabulary growth using macrocyclic molecules. This evaluation is designed to determine whether the fragment-based decomposition leads to a vocabulary explosion when applied to complex structural classes.

To represent diverse chemical spaces, we selected two specialized macrocyclic datasets: the Macformer dataset ($n = 5,551$) and the Macrocycle-DB ($n = 50,653$). All molecular structures were represented as canonical SMILES strings. For each dataset, molecules were processed sequentially, and FGT decomposed each molecule into its constituent functional groups and structural cores. The cumulative number of unique FGT tokens was recorded as a function of the number of molecules processed, allowing the plotting of vocabulary growth curves. A linear growth reference ($y = x$) was included to represent the theoretical worst-case scenario in which every new molecule introduces a completely unique set of tokens.

As illustrated in Fig. S5, the growth of unique FGT tokens follows a distinct sub-linear trend. Specifically, for the Macformer dataset, only 146 unique tokens were generated for 5,551 molecules. Even as the sample size increased to over 50,000 in the Macrocycle-DB, the vocabulary size remained manageable at 5,799 tokens.

These results demonstrate that FGT effectively captures the structural redundancy inherent in macrocyclic chemistry, where diverse molecules often share recurring fragments. Compared to atom-level modeling, FGT's modest vocabulary increase is offset by a 53.27% reduction in sequence length, providing a superior balance between semantic density and computational efficiency for LLMs.

Robustness and fidelity of FGT round-trip reconstruction

To validate the reliability of the FGT scheme as a reversible molecular representation, we conducted a large-scale round-trip reconstruction experiment. The objective was to ensure that FGT strings can be faithfully decoded back into their original chemical structures without ambiguity. We performed this "stress test" across three distinct datasets: the ZINC database (a subset of 10,000,000 molecules), the Macrocycle-DB, and the Macformer dataset.

The experimental procedure involved: (1) converting original SMILES strings into FGT sequences using the hierarchical decomposition algorithm; (2) decoding the FGT sequences back into SMILES format using the connectivity-based splicing logic; and (3) performing a canonical SMILES comparison and RDKit-based isomorphism check to verify if the reconstructed molecule is identical to the source. To quantitatively evaluate the reconstruction performance, we defined the following two metrics:

- **Success rate:** The percentage of molecules perfectly reconstructed. A reconstruction is considered successful only if the canonicalized SMILES matches the original input and passes the RDKit isomorphism check.
- **Ambiguity rate:** The percentage of molecules for which FGT-to-SMILES conversion fails to recover the exact original structure. Such cases are rare, typically arising from unresolved stereochemistry or complex connectivity in highly strained ring systems.

As summarized in Table S6, the FGT scheme demonstrated exceptional reconstruction fidelity. Specifically, the success rate for the ZINC dataset reached 99.97%, indicating near-perfect reversibility for standard drug-like molecules. Even for challenging macrocyclic structures with complex ring systems, the reconstruction success rate remained consistently high at 98.41% (Macrocycle-DB) and 99.89% (Macformer dataset). These results underscore the robustness of FGT across diverse and complex chemical architectures, confirming that the framework provides a stable and reliable foundation for the generative tasks performed by DrugLLM.

Patch clamp experiments

Before electrophysiological recordings, human embryonic kidney (HEK293) cells transfected with hHCN2 cDNA were incubated in normal Tyrode's solution for 2-3 hours at room temperature. The measurement of f-current (I_f) was performed via the patch-clamp technique in the whole-cell configuration, following previously established protocols. Patch-clamp pipettes had a resistance of 3-5 MΩ when filled with the internal solution. Cells were superfused with a gravity-controlled microsuperfusion system, allowing rapid changes in the perfusing solution. I_f currents were elicited from a holding potential of -40 mV to more negative voltages at -120 mV. Cells were superfused with modified Tyrode's solution. After recording I_f properties in control conditions, cells were superfused with solutions containing the test compounds. Dose-response curves were fitted using the Hill equation:

$$Y = \text{Bottom} + \frac{X^k \times (\text{Top} - \text{Bottom})}{X^k + \text{IC}_{50}^k}$$

where "Top" is the maximum effect, "Bottom" is the minimum effect, and k corresponds to the slope factor. All results were expressed as mean ± s.e.m.

Chemical synthesis

All reagents were obtained from commercial suppliers and used without further purification. During synthesis, reaction mixtures were purified by silica gel flash chromatography on 200 − 400 mesh silica gel 60 from Qingdao Haiyang Chemical with UV detection at 254 nm. Reported yields are isolated yields after purification of each intermediate. Final clean (purity ≥ 95.00%, LC-MS Agilent 1100 Series LC/MSD) compounds were used for the study. ^{1}H NMR and ^{13}C NMR spectra were recorded on a Bruker AVANCE NEO spectrometer (Bruker Company, Germany), using TMS as an internal standard and CD_3OD, DMSO-d_6 or $CDCl_3$ as solvent. Chemical shift is given in ppm (δ). High-resolution mass spectra (HRMS) were obtained using Agilent P/N G1969-90010. High-resolution mass spectra were reported for the molecular ion $[M + Na]^+$ or $[M + H]^+$. Thin-layer chromatography (TLC) analysis was carried out on silica gel plates GF254 (Qingdao Haiyang Chemical, China). Detailed experiments and characterization of the new compounds are included below. Detailed synthetic schemes and steps are provided in Figure S3, with compound characterization data, including HRMS and NMR, listed below.

(Z)-3-(4-chlorobut-2-en-1-yl)-7,8-dimethoxy-1,3-dihydro-2H-benzo[d]azepin-2-one (2)

Under nitrogen flow, potassium tert-butoxide (7.75 mmol, 0.87 g) was added to a suspension of 1 (4.56 mmol, 1.0 g) in anhydrous DMSO (4 mL). This solution was added dropwise, under nitrogen flow, to a solution of cis-1,4-dichlorobut-2-ene (10.49 mmol, 1.31 g) in anhydrous DMSO (15 mL), and the mixture was left stirring at room temperature for 1 h. Ice was then added to the reaction mixture, and the product was extracted with diethyl ether. The organic layers were collected and dried with Na_2SO_4, and the solvent was removed under vacuum. The residue was then purified by flash chromatography (cyclohexane/ethyl acetate: 5/5). Compound 2 was obtained in 60.00% yield as a white solid.

^{1}H NMR (400 MHz, $CDCl_3$): δ (ppm) 6.78 (s, 1H), 6.74 (s, 1H), 6.38 (d, J = 9.12 Hz, 1H), 6.20 (d, J = 9.16 Hz, 1H), 5.80 (ddt, J = 11.0, 7.9, 1.6 Hz, 1H), 5.53 (dt, J = 10.7, 7.1 Hz, 1H), 4.27 (d, J = 8.8 Hz, 2H), 4.16 (d, J = 7.9 Hz, 2H), 3.90 (s, 3H), 3.88 (s, 3H), 3.46 (s, 2H).

^{13}C NMR (101 MHz, $CDCl_3$) δ (ppm) 167.6, 150.0, 148.1, 129.2, 128.9, 127.4, 126.3, 124.6, 117.6, 111.2, 109.6, 56.0, 43.9, 43.1, 38.6.

HRMS (ESI-TOF) m/z: $[M + Na]^+$ Calcd for $C_{16}H_{18}NNaO_3^+$ 330.0868, found 330.0873.

(Z)-7,8-dimethoxy-3-(4-((3-methoxyphenethyl)(methyl)amino)but-2-en-1-yl)-1,3-dihydro-2H-benzo[d]azepin-2-one (HCN2-M1)

A mixture of 3 (1.62 mmol, 0.50 g) and 2-(3-methoxyphenyl)-N-methylethan-1-amine (3.25 mmol, 0.54 g) in anhydrous triethylamine (4 mL) was heated at 60 °C for 7 h under nitrogen flow. After cooling, $CHCl_3$ was added to the mixture, which was then washed with water. The organic phase was collected, dried with Na_2SO_4, and the solvent was removed under vacuum. The residue was purified by flash chromatography (CH_2Cl_2/MeOH 95:5 as eluent). HCN2-M1 was obtained as an oil in 76.00% yield.

^{1}H NMR (400 MHz, DMSO-d_6) δ (ppm) 7.24 (t, J = 7.8 Hz, 1H), 6.92 – 6.80 (m, 5H), 6.52 (dd, J = 9.1, 3.3 Hz, 1H), 6.40 (d, J = 9.0 Hz, 1H), 5.86 – 5.77 (m, 1H), 5.73 – 5.64 (m, 1H), 4.29 (d, J = 7.1 Hz, 2H), 4.01 (d, J = 7.5 Hz, 1H), 3.92 (d, J = 7.9 Hz, 1H), 3.81 (s, 6H), 3.78 (s, 3H), 3.42 (s, 3H), 3.07 – 2.97 (m, 2H), 2.86 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d_6) δ (ppm) 166.5, 159.4, 149.4, 147.6, 138.4, 133.5, 129.6, 128.1, 126.1, 124.0, 121.3, 120.8, 116.5, 114.3, 112.2, 111.6, 110.0, 55.5, 55.4, 55.1, 54.9, 51.5, 44.2, 42.1, 38.6, 29.5.

HRMS (ESI-TOF) m/z: $[M + H]^+$ Calcd for $C_{26}H_{33}N_2O_4^+$ 437.2435, found 437.2495.

(Z)-7,8-dimethoxy-3-(4-(methyl(phenethyl)amino)but-2-en-1-yl)-1,3-dihydro-2H-benzo[d]azepin-2-one (HCN2-M2)

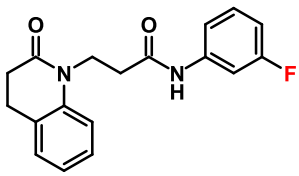
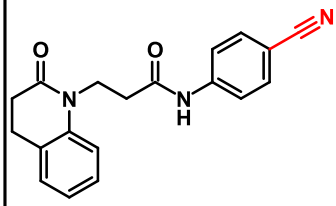
A mixture of 3 (1.62 mmol, 0.50 g) and N-methyl-2-phenylethan-1-amine (0.78 mmol, 0.44 g) in anhydrous triethylamine (4 mL) was heated at 60 °C for 7 h under nitrogen flow. After cooling, $CHCl_3$ was added to the mixture, which was then washed with water. The organic phase was collected, dried with Na_2SO_4, and the solvent was removed under vacuum. The residue was purified by flash chromatography (CH_2Cl_2/MeOH 95:5 as eluent). HCN2-M2 was obtained as an oil in 72.00% yield.

^{1}H NMR (400 MHz, MeOD) δ (ppm) 7.26 (t, J = 7.8 Hz, 1H), 6.92 (s, 1H), 6.89 – 6.82 (m, 4H), 6.42 (d, J = 9.2 Hz, 1H), 6.36 (d, J = 9.2 Hz, 1H), 5.78 – 5.70 (m, 1H), 5.69 – 5.59 (m, 1H), 4.26 (t, J = 5.9 Hz, 2H), 3.97 – 3.85 (m, 2H), 3.76 (d, J = 3.4 Hz, 6H), 3.75 (s, 3H), 3.37 (s, 2H), 3.29 – 3.23 (m, 2H), 3.05 – 3.00 (m, 2H), 2.74 (s, 3H).

^{13}C NMR (101 MHz, MeOD) δ (ppm) 169.4, 151.6, 149.7, 141.0, 130.2, 129.7, 129.5, 129.2, 128.8, 128.0, 127.3, 126.0, 119.4, 112.5, 111.4, 60.0, 56.6, 56.5, 54.8, 45.7, 43.5, 42.2, 34.1.

HRMS (ESI-TOF) m/z: $[M + H]^+$ Calcd for $C_{25}H_{31}N_2O_3^+$ 407.2330, found 407.2398.

Supplementary Figures

	Molecule A	Molecule B	Similarity	Token Length
Structure			Similar	-
SMILES	O=C(CCN1C(=O)CCc2ccccc21)Nc1cccc(F)c1	N#Cc1ccc(NC(=O)CCN2C(=O)CCc3ccccc32)cc1	Different	39/37
FGT	Iso 0/0 Ben 5/0 Flu 1/0 Eth 0/10 IndOne	Iso 0/0 Ben 6/0 Cya 7/0 Eth 0/10 IndOne	Similar	17/17 **(54% reduction)**

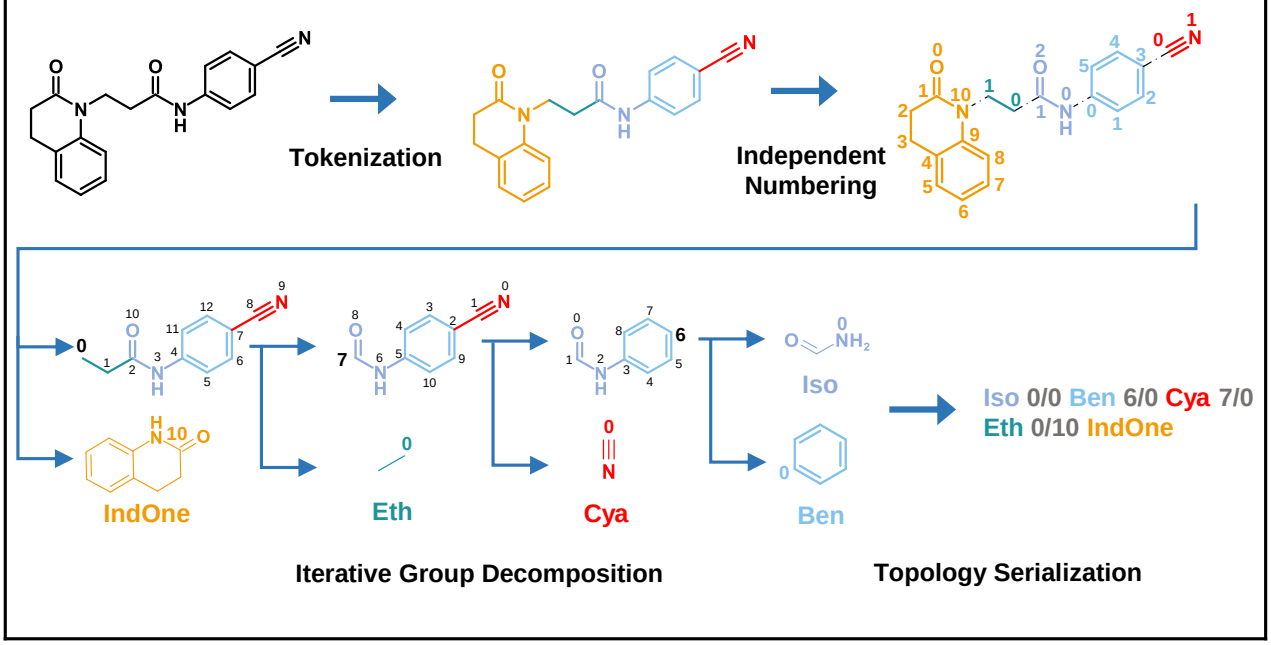


Fig. S1: **Comparison between FGT and SMILES, along with the assembly process of FGT.**

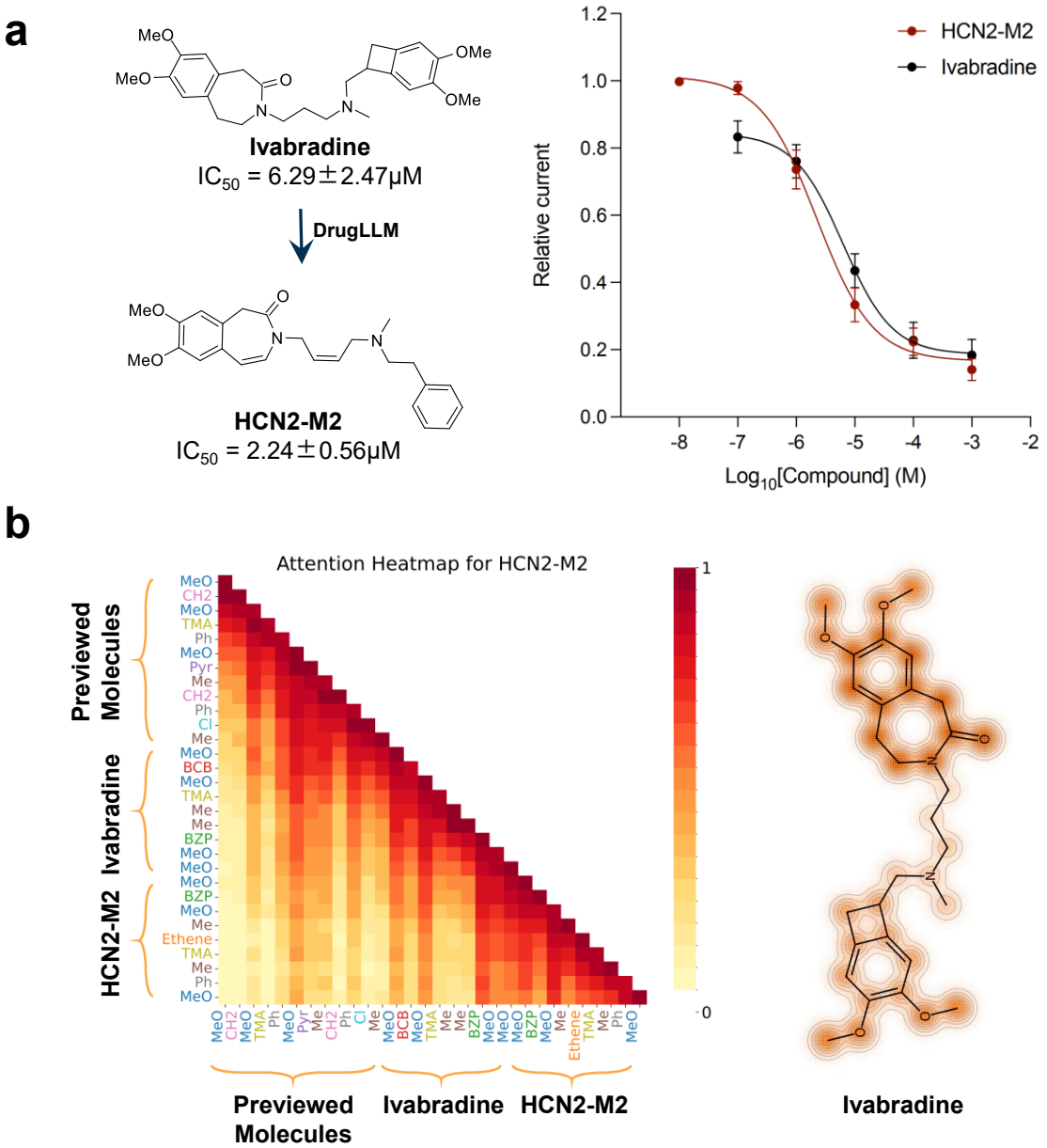


Fig. S2: **Applications of DrugLLM to optimization of HCN2 inhibitors.** (**a**) The inputs and generation (i.e., HCN2-M2) of DrugLLM, along with the dose-response curves of HCN2-M2 on the human HCN2 isoform heterologously expressed in HEK293 cells. (**b**) The attention maps of DrugLLM when generating HCN2-M2, and the visualization of the attention activations when generating HCN2-M2.

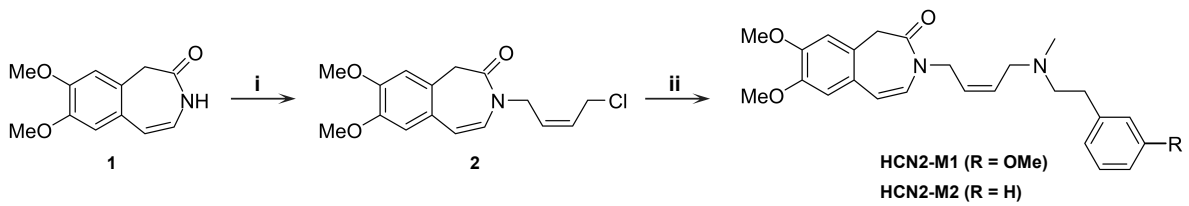


Fig. S3: **Reagents and conditions.** (i) t-BuOK, cis-1,4-dichlorobut-2-ene; (ii) 2-(3-methoxyphenyl)-N-methylethan-1-amine or N-methyl-2-phenylethan-1-amine.

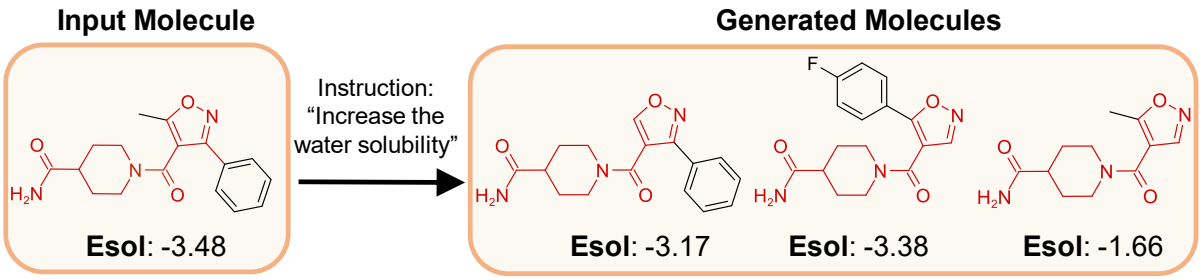


Fig. S4: **Examples of region-constrained molecular modification guided by natural language instruction.** The red-marked region indicates the fixed scaffold.

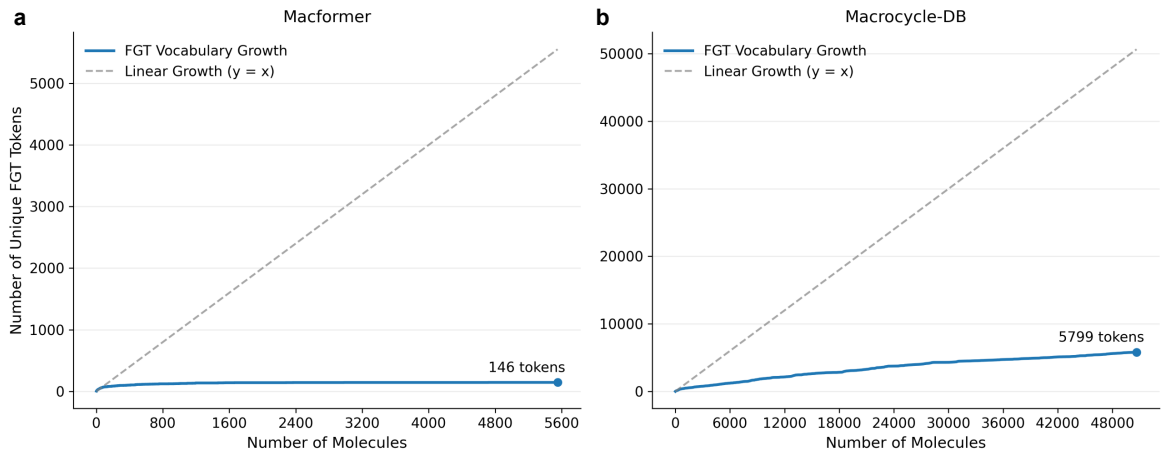


Fig. S5: **Scalability analysis of the FGT token vocabulary across macrocyclic datasets.** (**a**) On the Macformer dataset of 5,551 macrocycles, FGT produces only 146 unique tokens. (**b**) On the Macrocycle-DB of 50,653 molecules, the number of unique tokens remains moderate at 5,799. Both growth curves exhibit a distinct sub-linear trend and significantly deviate from the worst-case linear growth reference (gray dashed line, y=x), demonstrating FGT's efficiency and robustness against vocabulary explosion when representing diverse and complex chemical structures.

Supplementary Tables

Table S1. Pre-training dataset statistics for DrugLLM. Molecules from ZINC and ChEMBL are grouped into modification paragraphs, where each paragraph contains multiple molecule pairs aiming to optimize a common property.

Dataset	# Molecules	# Paragraphs	# Tasks	# Tokens
ZINC	4.50M	0.60M	770	330M
ChEMBL	180.20M	24.00M	10,100	12.58B
Total	184.70M	24.60M	10,870	12.91B

Table S2. Comparison of vocabulary efficiency and fragment coverage among different molecular decomposition methods. FGT achieves high molecular and fragment coverage with a minimal vocabulary size, demonstrating its compactness and generalization capability.

Method	Vocabulary size ↓	Molecular coverage (%) ↑	Fragment coverage (%) ↑	Uncovered fragments ↓
FGT	4,796	99.95	99.99	30
RECAP	503,037	64.75	90.28	33,918
BRICS	32,874	98.34	99.68	1,347

Table S3. Evaluation of generative quality under molecular optimization settings. DrugLLM achieves high validity, uniqueness, and novelty while enabling effective property optimization.

Model	Validity (%) ↑	Uniqueness (%) ↑	Novelty (%) ↑	ESOL suc. (%) ↑	LogP suc. (%) ↑	SA suc. (%) ↑	TPSA suc. (%) ↑
DrugLLM	100.00	99.49	99.02	76.31	72.44	59.11	63.51
VJTNN	100.00	99.28	89.14	44.49	46.35	49.88	42.50
JTVAE	100.00	94.12	99.47	53.13	50.75	50.88	51.25
MoLeR	100.00	99.78	97.99	51.63	48.50	50.75	46.63
Random	100.00	99.59	99.53	54.50	51.13	51.75	49.88

Table S4. Description of bioassay targets used in Table 1. Each Bioassay target corresponds to a ChEMBL assay ID representing a specific experimental measurement of biological activity.

Bioassay target	Description
CHEMBL1794496	Cell-based assay to confirm inhibitors of beta cell apoptosis
CHEMBL2354301	HTS using PAX8 luciferase reporter in RMG-I cells
CHEMBL1613983	Microorganism-based assay to identify compounds cytotoxic to SK(-)GAS
CHEMBL1738500	Cell-based assay to identify potentiators of HSF1
CHEMBL1614183	High-throughput screen for inhibitors of Mycobacterium tuberculosis H37Rv
CHEMBL1963888	Dose-response confirmation for antagonists of CRF-binding protein / CRF-R2 complex
CHEMBL4296185	Antibacterial activity against *Escherichia coli* ATCC 25922
CHEMBL4296190	Antifungal activity against *Cryptococcus neoformans* H99
CHEMBL1613886	qHTS for inhibitors/substrates of Cytochrome P450 3A4
CHEMBL1614481	qHTS for selective cytotoxicity in p53 null cancer cells
CHEMBL1963722	Kinome panel assay for ROCK2
CHEMBL1963723	Kinome panel assay for CDK2
CHEMBL1963727	Kinome panel assay for STK6
CHEMBL1963788	Kinome panel assay for KDR
CHEMBL1963790	Kinome panel assay for CDK5
CHEMBL1963807	Kinome panel assay for STK12
CHEMBL1963814	Kinome panel assay for ROCK1
CHEMBL1963835	Kinome panel assay for PRKACA
CHEMBL1964107	Kinome panel assay for DYRK1A
CHEMBL1964119	Kinome panel assay for STK3

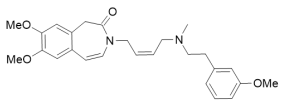
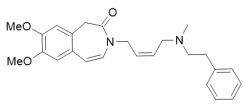
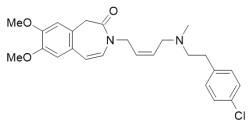
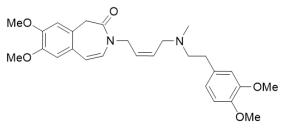
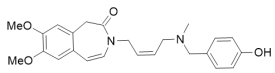
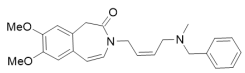
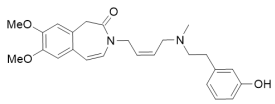
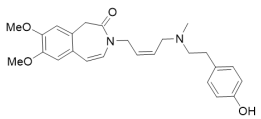
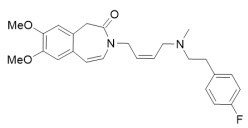
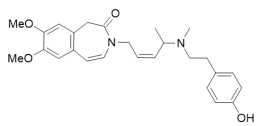
Table S5. Predictive performance of ChemProp-based biological activity predictors across different bioassay targets. The models exhibit strong correlations with experimental measurements, ensuring reliable evaluation of generated molecules.

Bioassay target	Pearson correlation
CHEMBL1794496	0.79
CHEMBL2354301	0.83
CHEMBL1613983	0.82
CHEMBL1738500	0.77
CHEMBL1614183	0.76
CHEMBL1963888	0.80
CHEMBL4296185	0.78
CHEMBL4296190	0.79
CHEMBL1613886	0.76
CHEMBL1614481	0.75
CHEMBL1963722	0.83
CHEMBL1963723	0.92
CHEMBL1963727	0.78
CHEMBL1963788	0.80
CHEMBL1963790	0.79
CHEMBL1963807	0.75
CHEMBL1963814	0.91
CHEMBL1963835	0.78
CHEMBL1964107	0.81
CHEMBL1964119	0.75

Table S6. Round-trip reconstruction success rates and ambiguity rates of FGT across different datasets.

Dataset	# Molecules	# Reconstructed	Success rate (%) ↑	Ambiguity rate (%) ↓
ZINC	10,000,000	9,997,412	99.97	0.03
Macrocycle-DB	50,653	49,849	98.41	1.59
Macformer	5,551	5,545	99.89	0.11

Table S7. DrugLLM-generated ivabradine derivatives targeting HCN2. Ten examples of SMILES and their 2D molecular structures generated by DrugLLM to optimize ivabradine for enhanced activity on HCN2.

SMILES	2D structure
COc1cccc(CCN(C)CC=CCN2C=Cc3cc(OC)c(OC)cc3CC2=O)c1	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)CCc1ccccc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)CCc1ccc(Cl)cc1)C=C2	
COc1ccc(CCN(C)CC=CCN2C=Cc3cc(OC)c(OC)cc3CC2=O)cc1OC	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)Cc1ccc(O)cc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)Cc1ccccc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)CCc1ccc(O)cc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)CCc1ccc(Cl)cc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CCN(C)CCc1ccc(F)cc1)C=C2	
COc1cc2c(cc1OC)CC(=O)N(CC=CC(C)N(C)CCc1cccc(O)c1)C=C2	

Supplementary Algorithm

Algorithm S1 Deterministic FGT encoding with canonical decomposition

Require: Molecular graph $G = (V, E)$
Ensure: Unique FGT string S

$G \leftarrow \text{CanonicalizeSMILES}(G)$ ▷ Remove atom order, stereo, and SMILES randomness
$\mathcal{G} \leftarrow \text{ExtractStructuralGroups}(G)$ ▷ Fused rings + unsaturated motifs, rule-based
$\mathcal{C} \leftarrow \emptyset$ ▷ Connection records
while $|\mathcal{G}| > 1$ **do**
  for all $g \in \mathcal{G}$ in canonical order **do**
    if $\text{IsPeripheral}(g)$ **and** $\text{IsRemovable}(g)$ **then**
      $(g', a_m, a_g) \leftarrow \text{IdentifyConnection}(G, g)$ ▷ Using atom signatures
      $\mathcal{C} \leftarrow \mathcal{C} \cup (g', a_m, a_g)$
      $G \leftarrow G \setminus g'$
      $\text{UpdateStructuralGroups}(G, \mathcal{G})$
      break
    end if
  end for
end while
$g_0 \leftarrow \text{RemainingCoreGroup}(\mathcal{G})$
$S \leftarrow \text{DictID}(g_0)$
for all $(g_i, a_m, a_g) \in \mathcal{C}$ in removal order **do**
  $S \leftarrow S \oplus \text{Encode}(a_m, a_g, \text{DictID}(g_i))$
end for
return S

Algorithm S2 Deterministic FGT decoding

Require: FGT string S
Ensure: Reconstructed molecular graph G

Parse S into core token g_0 and connection list $\mathcal{C}$
$G \leftarrow \text{InstantiateGroup}(g_0)$
for all (g_i, a_m, a_g) in $\mathcal{C}$ sequentially **do**
  $G_i \leftarrow \text{InstantiateGroup}(g_i)$
  $v_m \leftarrow \text{LocateAtom}(G, a_m)$
  $v_g \leftarrow \text{LocateAtom}(G_i, a_g)$
  $G \leftarrow \text{Attach}(G, G_i, v_m, v_g)$
end for
return G