STAMP as an addressable molecular language for fragment-level learning and editing

Supplementary Information

Yuxuan Ren[1], Jiatong Li[1,2], Ziqiao Meng[1], Yatao Bian[1,†]

[1]National University of Singapore
[2]Hong Kong Polytechnic University
[†]Correspondence to: ybian@nus.edu.sg

Contents

Supplementary Note 1: Algorithms **2**

Supplementary Note 2: Supporting construction metrics **3**

Supplementary Note 3: Subtree-crossover optimization loop **5**

Supplementary Note 4: Lead optimization: per-cell results and protocol **6**

Supplementary Note 5: Vocabulary and training configurations **7**

Supplementary Note 6: MoleculeNet: full per-task results **7**

Supplementary Note 7: MoleculeACE: comparison across 30 targets **8**

Supplementary Note 8: Token-space locality of matched molecular edits **9**

Supplementary Note 9: Structure-matched frozen activity-cliff control **9**

Supplementary Note 10: RePRA comparison of nine pretrained models **10**

Supplementary Note 11: Language-model interfaces and indexed controls **11**

Supplementary Note 1: Algorithms

Supplementary Algorithm 1 Fragment Tree Construction

Require: Molecule G, fragmentation scheme $\mathcal{F}$
Ensure: Rooted fragment tree T, or fragmentation failure
Compute fragments $F_1, \ldots, F_k$ and reduced graph G_R under $\mathcal{F}$
root $\leftarrow \arg\max_{F_i} \deg_{G_R}(F_i)$, ties broken by heavy-atom count
Require G_R to be a tree; root it at root to obtain T
Rank children by descending subtree size; select the first as the main continuation
Traverse the remaining children as branches, then the main continuation, each depth-first
return T with canonical DFS traversal

Supplementary Algorithm 2 STAMP Stack-Machine Decoder

Require: STAMP sequence x with motif groups and branch prefixes
Ensure: Molecular graph G, or decoding failure
$G \leftarrow \emptyset$; port stack $S \leftarrow \emptyset$; branch-frame stack $B \leftarrow \emptyset$
for all motif groups or branch tokens in x, from left to right **do**
    if next item is $[B_n]$ **then**
        $p \leftarrow \textsc{Pop}(S)$; push frame (n, S) onto B; $S \leftarrow [p]$
    else
        read motif and its joints; place motif in G
        if root motif **then**
            push all output ports onto S in reverse traversal order
        else
            pop parent port from S; connect it to the motif's first joint
            push remaining output ports onto S in reverse traversal order
        end if
        decrement the remaining-group counter of every frame in B
        while B is nonempty and its top frame has zero remaining groups **do**
            pop the completed frame and restore its saved port stack as S
        end while
    end if
end for
Generated-sequence fallbacks: out-of-range $[J_i] \rightarrow$ wrap modulo atom count; saturated atom $\rightarrow$ scan to next free-valence atom; overlong branch $\rightarrow$ truncate to remaining groups; empty port stack $\rightarrow$ start a new component; leftover open ports $\rightarrow$ cap with implicit H.
sanitize the assembled graph; **return** G or decoding failure

Supplementary Note 2: Supporting construction metrics

Eight-representation control. We compare eight molecular representations on the same 20,000 molecules sampled from the ZINC-250K training split. Exact reconstruction compares canonical isomeric SMILES, with InChIKey equality as a fallback for equivalent encodings. Valid decoding is reported separately. Encoding, decoding and tokenization are described in Methods.

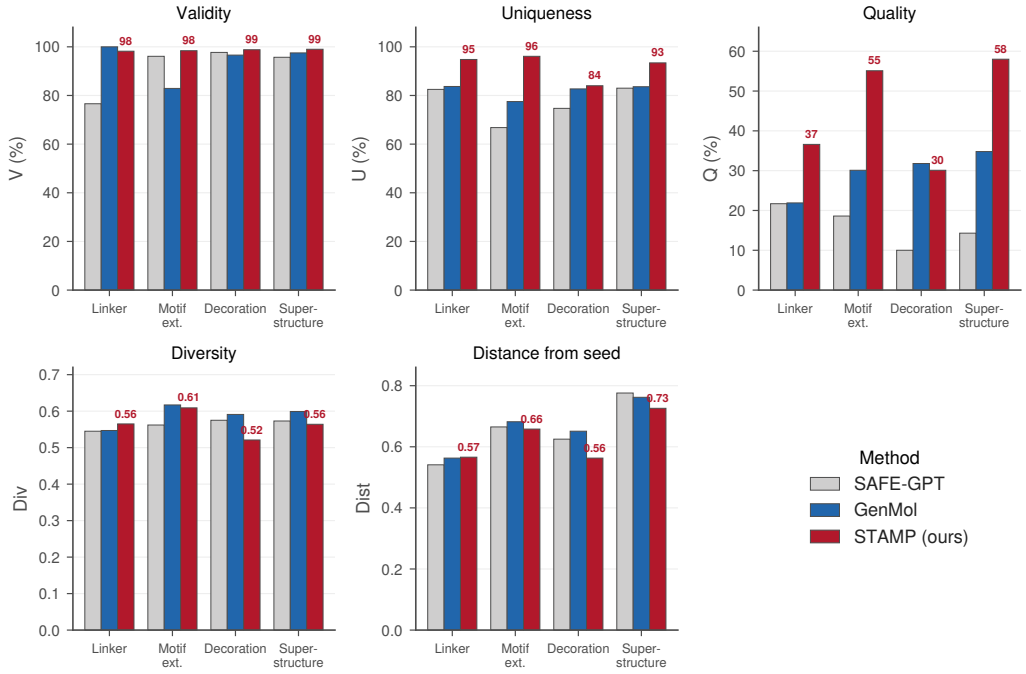
Supplementary Table 1: Representation statistics on a common 20,000-molecule ZINC-250K subset. All eight methods encoded every input. Mean and p_{95} denote mean and 95th-percentile sequence lengths in tokens; $|V|$ is the number of distinct tokens observed on this subset. Decode validity and exact molecular identity are percentages of all inputs.

Representation	Mean	p_{95}	$\|V\|$	Valid (%)	Exact (%)
SMILES	44.25	60	32	100.000	100.000
SELFIES	37.42	52	71	100.000	100.000
Group-SELFIES	40.63	56	67	100.000	100.000
SAFE	47.11	63	54	99.965	96.570
fragSMILES	16.66	25	1,849	99.040	93.345
t-SMILES (TSDY)	66.59	88	39	99.970	15.925
Molcano	47.14	66	286	100.000	57.085
STAMP	15.31	24	6,783	100.000	100.000

Supplementary Table 2: Controlled de novo generation for eight representations with the same 4-layer causal Transformer backbone ($d = 256$), optimization schedule and 10,000 generation requests per method. V is validity over all requests, U is uniqueness among valid outputs, N is novelty among unique valid outputs, Div is internal fingerprint diversity, and Q is the fraction of all requests yielding a unique valid molecule with QED≥ 0.6 and SA≤ 4. V, U, N and Q are percentages. LM vocab includes control tokens and is fitted on the training corpus. Values are from one training and sampling run per representation.

Representation	V	U	N	Div	Q	LM vocab
SMILES	79.66	99.99	99.99	0.8776	55.21	37
SELFIES	100.00	100.00	99.98	0.8932	43.21	95
Group-SELFIES	100.00	100.00	99.98	0.8881	46.40	90
SAFE	47.81	99.98	100.00	0.8802	32.77	65
fragSMILES	91.87	99.99	99.91	0.8797	66.76	4,303
t-SMILES (TSDY)	87.71	100.00	100.00	0.8801	58.42	50
Molcano	40.21	100.00	99.95	0.8687	31.37	349
STAMP	97.38	99.97	99.79	0.8663	75.54	6,783

The shared input sets contain 100,000 training and 10,000 validation molecules. Molcano encoded 99,998 training inputs and all validation inputs; the other representations encoded every input. Training uses seed 42 for all representations; decoding settings are specified in Methods.



Supplementary Figure 1: Fragment-constrained generation metrics. Validity, uniqueness, constraint-aware quality, diversity and Tanimoto distance from the supplied seed for SAFE-GPT, GenMol and STAMP. The four settings are linker design, motif extension, scaffold decoration and superstructure generation. Linker design also covers the shared scaffold-morphing formulation.

Supplementary Table 3: De novo generation under the GenMol evaluation protocol. SAFE-GPT and GenMol values are from[5], using the SAFE corpus of ZINC and UniChem molecules. STAMP uses a ZINC-mined corpus and reports mean ± s.d. across five seeds. V, U and Q denote validity, uniqueness and quality yield; Div is internal fingerprint diversity. Metrics are defined in Methods. Bold marks the best value in each column.

Method	V (%)	U (%)	Q (%)	Div
SAFE-GPT	94.0 ± 0.4	**100.0 ± 0.0**	54.7 ± 0.3	0.879 ± 0.001
GenMol w/o conf. sampling	96.7 ± 0.3	99.3 ± 0.2	53.8 ± 1.7	**0.896 ± 0.001**
GenMol $N = 1$ ($\tau = 0.5$, $r = 0.5$)	**100.0 ± 0.0**	99.7 ± 0.1	84.6 ± 0.8	0.818 ± 0.001
GenMol $N = 2$	97.6 ± 0.7	99.5 ± 0.2	76.2 ± 1.3	0.843 ± 0.002
GenMol $N = 3$	95.6 ± 0.5	99.0 ± 0.1	67.1 ± 0.7	0.861 ± 0.001
STAMP (ours)	99.7 ± 0.2	99.4 ± 0.3	**88.0 ± 1.1**	0.861 ± 0.001

Supplementary Note 3: Subtree-crossover optimization loop

Supplementary Algorithm 3 applies TreeCrossover within a fixed-budget optimization loop. The donor pool groups subtrees by port signature and retains at most 50 per signature and 5,000 overall, ranked by oracle reward. Molecules contribute donor subtrees when their reward exceeds $\theta = \max(0.7r^*, 0.3)$, where r^* is the highest reward observed.

Supplementary Algorithm 3 Port-signature subtree crossover in the optimization loop

Require: oracle $\mathcal{O}$ with budget B; population size K; operator TreeCrossover

initialize population $\mathcal{P}$; donor pool $\mathcal{D} \leftarrow \emptyset$
while oracle calls $< B$ **do**
  sample parents $s_A, s_B \sim \mathcal{P}$
  $s' \leftarrow \textsc{TreeCrossover}(s_A, s_B)$
  if $s' = \varnothing$ **or** s' fails size filter **then continue**
  end if
  $r \leftarrow \mathcal{O}(s')$ ▷ one oracle call, cache-deduplicated
  $\mathcal{P} \leftarrow \textsc{TopK}(\mathcal{P} \cup \{(s', r)\}, K)$
  if $r > \theta$ **then**
    $\mathcal{D} \leftarrow \mathcal{D} \cup \textsc{HarvestSubtrees}(s', r)$ ▷ keyed by port signature
  end if
end while
return best molecule in $\mathcal{P}$

Supplementary Note 4: Lead optimization: per-cell results and protocol

Supplementary Table 4 adds RetMol to the STAMP, GenMol and Graph GA comparison in Fig. 4c. Rows identify the protein target and starting molecule; the two column groups use molecular-similarity thresholds of 0.4 and 0.6.

Supplementary Table 4: Lead optimization on GenMol's docking benchmark. Rows identify the target and starting molecule, and δ is the molecular-similarity threshold. Each cell gives the best QuickVina2 score among constraint-satisfying candidates that improve on the starting molecule, averaged over three optimization seeds. Lower scores are better; bold marks the best method within each threshold group. A dash indicates no qualifying candidate and contributes zero to the reported total.

		$\delta = 0.4$				$\delta = 0.6$			
Starting molecule	Initial score	**STAMP**	GenMol	RetMol	Graph GA	**STAMP**	GenMol	RetMol	Graph GA
parp1/0	-7.3	**-11.9**	-10.6	-9.0	-8.3	**-11.3**	-10.4	–	-8.6
parp1/1	-7.8	**-11.9**	-11.0	-10.7	-8.9	**-11.1**	-9.7	–	-8.1
parp1/2	-8.2	-11.2	**-11.3**	-10.9	–	**-10.9**	-9.2	–	–
fa7/0	-6.4	**-8.6**	-8.4	-8.0	-7.8	**-8.3**	-7.3	-7.6	-7.6
fa7/1	-6.7	**-8.7**	-8.4	–	-8.2	**-8.1**	-7.6	–	-7.6
fa7/2	-8.5	**-8.9**	–	–	–	–	–	–	–
5ht1b/0	-4.5	**-14.4**	-12.9	-12.1	-11.7	**-13.1**	-12.1	–	-11.3
5ht1b/1	-7.6	**-12.5**	-12.3	-9.0	-12.1	**-12.1**	-12.0	-10.0	-12.0
5ht1b/2	-9.8	-11.5	**-11.6**	–	–	**-11.3**	-10.5	–	–
braf/0	-9.3	-10.4	**-10.8**	–	-9.8	–	–	–	–
braf/1	-9.4	-10.9	-10.8	**-11.6**	–	**-10.9**	-9.7	–	–
braf/2	-9.8	-11.2	-10.6	–	**-11.6**	**-10.9**	-10.5	–	-10.4
jak2/0	-7.7	**-10.4**	-10.2	-8.2	-8.7	**-10.3**	-9.3	–	-8.1
jak2/1	-8.0	**-11.0**	-10.0	-9.0	-9.2	**-10.3**	-9.4	–	-9.2
jak2/2	-8.6	**-10.6**	-9.8	–	–	–	–	–	–
Total		**-164.1**	-148.7	-88.5	-96.3	**-128.6**	-117.7	-17.6	-82.9

Docking protocol. STAMP applies the QED, SA and similarity constraints before docking and allocates 1,000 docking calls to candidates satisfying these constraints.

Supplementary Note 5: Vocabulary and training configurations

The cleaned PubChem corpus contains 9.69M molecules and 138M token occurrences. The selected PS-VAE-mined vocabulary contains 4,666 motifs, with 99% cumulative token-occurrence coverage and 96.85% motif-only coverage. Structural markers dominate the highest-frequency entries; remaining motif occurrences use atom-level fallback.

Supplementary Table 5: Training configurations by experiment. Larger-scale generation and fragment infilling use separate AO-GPT checkpoints[6]. Vocabulary size includes motif, structural and control tokens; the controlled row gives the STAMP vocabulary, with the other seven listed in Supplementary Table 2. Batch size is global. Training duration is the evaluated checkpoint step for de novo AO-GPT and the fine-tuning duration for infill models. The lower panel gives final encoder validation metrics and training times on one H200 GPU.

Configuration	Architecture	Vocab	Training	Batch	LR
Controlled de novo (Fig. 2a)	4-layer, d=256	6,783	10 epochs	256	3×10^{-4}
Large-scale AO-GPT (Fig. 2b)	12-layer, d=768	12,659	300,000 steps	1,024	3×10^{-4}
Fragment infill (three tasks)	12-layer, d=768	12,659	3 epochs	256	10^{-4}
Scaffold decoration	12-layer, d=768	12,659	10 epochs	256	5×10^{-5}
Foundation MLM	6-layer enc., d=768	4,761	10 epochs	512	5×10^{-4}
Foundation MTR	6-layer enc., d=768	4,761	5 epochs	512	3×10^{-4}
Fine-tuning (MoleculeNet)	encoder + task head	—	20 epochs	32	2×10^{-5}

Configuration	Validation metric	Value	Training time
Foundation MLM	Perplexity	2.45	2 h 45 min
Foundation MTR	MSE (z-scored descriptors)	0.034	3 h

Supplementary Note 6: MoleculeNet: full per-task results

Supplementary Table 6 compares STAMP with published ChemBERTa-2 10M results under the corresponding pretraining objectives.

Supplementary Table 6: MoleculeNet evaluation at 10M-molecule pretraining scale. STAMP fine-tuning results are compared with published ChemBERTa-2 10M values under matched MLM and MTR objectives[7]. Regression uses RMSE and classification uses ROC-AUC. The Tox21 task is SR-p53. Bold marks the better value in each pair.

	MLM		MTR			MLM		MTR	
	STAMP	BPE	STAMP	BPE		STAMP	BPE	STAMP	BPE
	Regression (RMSE ↓)					Classification (ROC-AUC ↑)			
BACE-reg	**1.526**	1.611	1.489	**1.417**	BACE-clf	**0.786**	0.729	**0.797**	0.783
Clearance	**49.61**	53.86	**47.17**	48.93	BBBP	**0.702**	0.696	**0.749**	0.733
ESOL	1.018	**0.961**	0.978	**0.858**	ClinTox	**0.897**	0.349	**0.890**	0.601
Lipo	**0.848**	1.009	**0.737**	0.744	Tox21 SR-p53	0.721	**0.748**	0.741	**0.827**

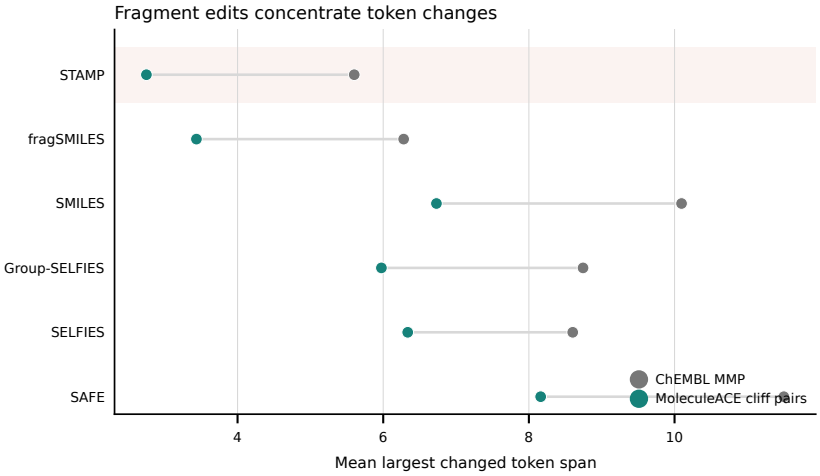
Supplementary Note 7: MoleculeACE: comparison across 30 targets

Supplementary Table 7 compares STAMP-MTR and 19 baselines across the same 30 MoleculeACE targets. STAMP and the released ChemBERTa-2 10M encoder were fine-tuned using the matched protocol in Methods. Other baseline results are from the MoleculeACE benchmark authors [1].

Supplementary Table 7: MoleculeACE results averaged over 30 targets. The 20 methods are sorted by full-set RMSE. Columns report full-set RMSE, activity-cliff RMSE and their difference. Bold marks the minimum in each error column.

Method	RMSE_{all}	$\text{RMSE}_{\text{cliff}}$	Δ_{cliff}
SVM + ECFP4	**0.671**	**0.751**	0.080
GBM + ECFP4	0.695	0.783	0.089
RF + ECFP4	0.700	0.786	0.086
STAMP-MTR (ours)	0.719	0.825	0.106
KNN + ECFP4	0.729	0.839	0.110
LSTM + SMILES (aug 10)	0.742	0.871	0.129
GBM + MACCS	0.750	0.842	0.092
SVM + MACCS	0.752	0.850	0.098
ChemBERTa-2 BPE	0.753	0.865	0.112
RF + MACCS	0.756	0.847	0.091
MLP + ECFP4	0.763	0.833	0.070
KNN + MACCS	0.811	0.900	0.089
RF + PHYSCHEM	0.881	0.930	0.049
GBM + PHYSCHEM	0.891	0.946	0.055
Transformer + TOKENS (aug 10)	0.896	0.970	0.074
GCN + GRAPH	0.914	0.974	0.060
CNN + SMILES (aug 10)	0.930	0.968	0.038
GAT + GRAPH	0.947	1.005	0.059
MPNN + GRAPH	1.094	1.120	0.026
AFP + GRAPH	1.186	1.192	**0.006**

Supplementary Note 8: Token-space locality of matched molecular edits

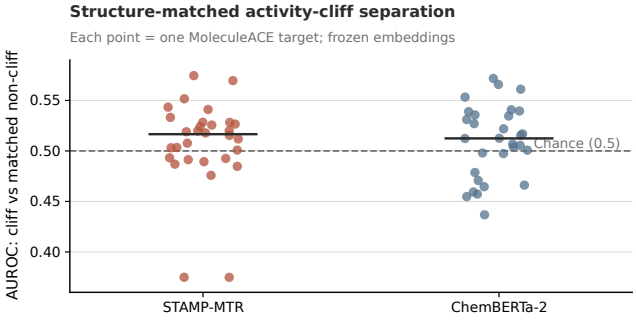


Supplementary Figure 2: Token-space locality of molecular edits. Mean largest changed token span for ChEMBL matched molecular pairs and MoleculeACE activity-cliff pairs under the same sequence-alignment definition. Each changed span is a maximal contiguous region returned by the alignment.

Supplementary Note 9: Structure-matched frozen activity-cliff control

Within each MoleculeACE target, matched molecular pairs are labelled as cliffs when the activity difference $\Delta y \geq 1.0$ and non-cliffs when $\Delta y \leq 0.25$. Pairs are divided into 20 ECFP4-distance quantile bins. Equal numbers of cliffs and non-cliffs are sampled within each target–distance-bin stratum, capped at 100 pairs per class. Target-level ROC-AUC measures whether frozen-embedding distance separates the classes. Matching is repeated with 20 prespecified seeds.

Mean target-level ROC-AUC across the repeated samples was 0.508 for STAMP-MTR and 0.509 for ChemBERTa-2. Raw ECFP4 distance and bin index gave 0.501 and 0.500, respectively (Supplementary Fig. 3). Frozen distance scores were close to chance after matching structural distance within targets.



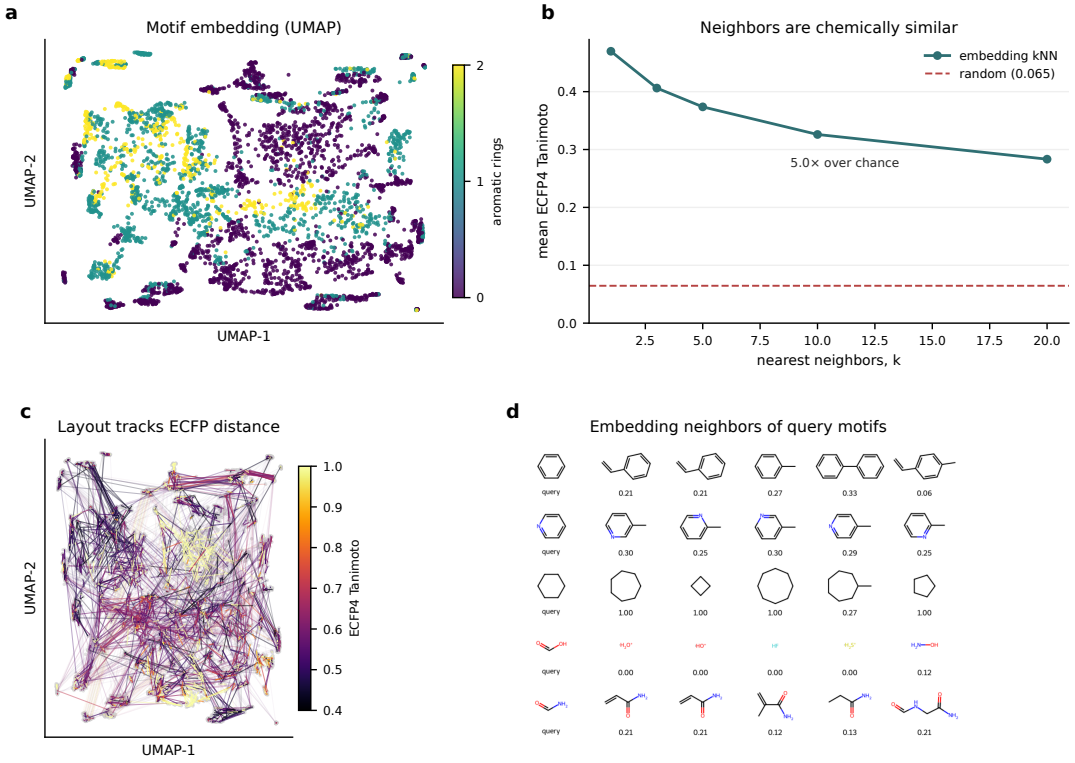
Supplementary Figure 3: Frozen activity-cliff discrimination after structural matching. Each point represents one MoleculeACE target. ROC-AUC measures separation of activity-cliff pairs from non-cliff pairs matched within ECFP4-distance strata, using frozen-embedding distance as the score. Horizontal segments show medians; the dashed line indicates ROC-AUC 0.5.

Supplementary Note 10: RePRA comparison of nine pretrained models

RePRA[2] compares threshold-defined discordance between representation and property distances. For molecule pairs (i, j), representation distance $d_z(\mathbf{z}_i, \mathbf{z}_j)$ and property distance $d_y(\mathbf{y}_i, \mathbf{y}_j)$ are MinMax-normalized to $[0, 1]$. Thresholds identify scaffold-hop pairs with large d_z and small d_y, and activity-cliff pairs with small d_z and large d_y. For their respective counts C_1 and C_4, the diagnostic is

$$s_{IR} = \frac{C_1^{\mathrm{PTM}}}{C_1^{\mathrm{ECFP}}} + \frac{C_4^{\mathrm{PTM}}}{C_4^{\mathrm{ECFP}}}. \tag{1}$$

The ECFP4 reference has $s_{IR} = 2$; lower values indicate a smaller sum of normalized discordant-pair counts. Supplementary Table 8 reports the eight task-level scores for all nine pretrained models.



Supplementary Figure 4: Contextual motif-embedding analysis for STAMP-MTR. UMAP visualization of motif representations, ECFP4 similarity of embedding neighbours and representative motif neighbourhoods. Neighbouring motifs share ring and functional-group features. Figure 3b summarizes neighbour enrichment.

Supplementary Table 8: RePRA s_{IR} on eight MoleculeNet tasks. Lower values indicate fewer threshold-defined discordant pairs relative to ECFP4. STAMP-MLM and STAMP-MTR are compared with the seven pretrained models reported in[2]. Bold marks the lowest value for each task.

Task	STAMP-MLM	STAMP-MTR	GROVER	ChemBERT	GraLoG	MAT	S.T.	PreGNNs	Pre8
BACE	**1.606**	1.668	2.393	2.341	2.426	2.452	2.302	2.495	2.301
BBBP	2.027	**1.997**	3.054	2.767	3.595	3.172	2.045	3.177	2.035
ClinTox	2.532	2.897	3.354	2.546	3.383	3.163	**1.951**	3.069	3.081
SIDER	1.405	1.739	1.248	1.526	1.252	1.595	1.547	1.437	**1.247**
Tox21	1.883	**1.840**	3.731	2.657	3.938	3.287	2.750	3.226	3.458
ESOL	1.921	**1.514**	1.923	1.784	2.047	1.930	1.819	1.960	1.813
FreeSolv	1.960	**1.642**	2.106	2.040	2.126	2.018	1.896	2.058	1.849
Lipo	1.913	**1.831**	1.975	2.119	1.993	2.367	1.988	1.983	1.935

Supplementary Note 11: Language-model interfaces and indexed controls

Evaluation scope. Qwen and DeepSeek denote Qwen-2.5-7B-Instruct[3] and DeepSeek-V4-Pro[4], respectively. All six S^2-Bench files contribute 500 tasks each. Family-level success is the equal-weight mean of the three corresponding subtasks. Each STAMP interface is reported separately.

Independent atom-indexed SMILES interface. RDKit removes existing atom maps, produces canonical isomeric SMILES and reparses the string before assigning 1-based atom IDs. Input includes canonical and mapped SMILES, element labels and a bond-typed adjacency list. IDs are local to each molecule, with no correspondence between molecules. The executor does not depend on the STAMP codec, fragment tree, motif bank or node templates. The model selects `delete_atoms`, `replace_atoms` or `attach_fragment`, explicitly specifying atom sets, retained anchors and a generated fragment. Source IDs are 1-based; fragment connection indices are 0-based.

Selected atom sets must be nonempty, unique and connected, and cannot remove the entire molecule. Replacement fragments must be valid, nonempty, connected SMILES without dummy atoms; functional-group names are not converted to structures. Single-boundary replacements require one retained anchor. Multiple boundaries require an explicit connection list covering every original boundary exactly once and preserving its bond type. Ambiguous anchors also require the removed boundary atom. Addition allows single, double or triple bonds and checks available valence. The executor applies the declared selection, sanitizes the product and checks connectivity; it neither relocates an invalid target nor supplies a missing connection.

STAMP execution rules. The primary node interface labels every motif intersecting a fixed SMARTS match. A label may span multiple motifs. Subtree growth selects the C/N atom with the most available hydrogens within the chosen node, breaking ties by atom order. If none is available, it probes other hydrogen-bearing heavy atoms within that node, including O–H and S–H sites. Known group names map to fixed structures; generated fragment SMILES attach through the first hydrogen-bearing heavy atom in parse order. If none exists, only a radical atom passing a single-bond valence probe is accepted. These fixed rules define the STAMP execution interface. The selected subtree is converted into an atom set for the shared RDKit graph editor.

Supplementary Table 9: Native-input molecular editing and optimization. Entries give success/validity in percent for all 500 cases per row and interface. STAMP denotes the primary subtree-action interface. Indexed SMILES denotes independent atom actions; the remaining columns rewrite complete strings. RDKit parsability defines validity, with additional connectivity checks in the action and STAMP paths.

Model	Task	STAMP	Indexed SMILES	SMILES rewrite	STAMP rewrite
Qwen	Add	73.40/91.20	24.40/33.80	22.00/62.00	41.80/83.80
Qwen	Delete	68.00/77.80	18.00/21.40	49.40/83.80	47.20/90.20
Qwen	Substitute	44.60/77.00	1.00/1.80	16.40/77.40	15.20/71.80
Qwen	LogP	65.60/96.20	1.00/1.60	35.60/61.00	24.60/39.60
Qwen	MR	85.40/96.80	4.20/5.40	32.80/60.40	32.20/57.00
Qwen	QED	40.80/93.80	2.40/5.40	33.00/74.00	21.40/46.20
DeepSeek	Add	96.60/100.00	86.60/95.80	73.20/90.60	60.80/94.60
DeepSeek	Delete	76.60/81.00	75.80/80.60	68.80/99.00	65.20/100.00
DeepSeek	Substitute	83.00/97.80	87.40/98.80	86.60/99.80	71.60/99.20
DeepSeek	LogP	84.60/99.20	83.00/92.00	61.80/98.40	64.60/97.20
DeepSeek	MR	85.00/99.00	80.60/94.20	65.40/96.80	56.80/96.40
DeepSeek	QED	56.40/99.00	52.80/93.60	33.60/97.40	32.00/98.00

DeepSeek's indexed baseline reached 83.27% mean success on MolEdit and 72.13% on MolOpt, compared with 85.40% and 75.33% for STAMP subtree actions. Indexed substitution exceeded STAMP (87.40% versus 83.00%). Qwen's indexed baseline reached 14.47% and 2.53%; only 9/500 substitution outputs were valid.

Transfer construction and execution. Demonstrations are constructed from input molecules in the S^2-TOMG-Bench substitution subset (`MolEdit_SubComponent.csv`) and accepted only when the removed group decreases by one and the added group increases by one. A different eligible B is selected, with at most 12 tasks per transformation; transformation types are interleaved before taking 120 examples. Nine transformations contribute 12 cases each, thiol-to-hydroxyl contributes one and halo-to-nitrile contributes 11. Aldehyde-to-carboxyl contributes no transfer examples. The 120 distinct triples reuse 54 A, 120 A' and 56 B strings, comprising 180 distinct public molecular strings overall.

Each condition independently infers a removed/added plan from A/A' in one call. Names are normalized and checked only for vocabulary membership and inequality. Valid but incorrect plans pass unchanged to an independent second-stage conversation containing B and the predicted plan. Stage two uses one model call. STAMP transfer replaces a single node, preserving descendants, parent and all labelled boundary ports through RDKit `molzip`; the model writes the full new fragment. The node-port executor uses the fragment written by the model. Raw-STAMP transfer presents STAMP strings. Success is scored on the final product against the true group-count change, without additionally requiring the predicted plan to be correct.

The shared-label transfer conditions add the same atom graphs and all atom-level labels to A, A' and B. STAMP also maps node membership onto the common atom IDs. The conditions share chemical facts and retain their native grouping, action syntax and connection rules. Each condition infers its own plan.

Supplementary Table 10: Two-stage transfer, including indexed-SMILES controls. Every row uses the same 120 examples. SR, validity and plan Exact are percentages over all examples; Exact requires both group names. Invalid plans terminate at stage one. Invalid edits count failures only among examples entering stage two. These counts are distinct from incorrect but valid plans.

Input	Model	Interface	SR	Validity	Plan Exact	Invalid plan	Invalid edit
Native	Qwen	STAMP node-port	30.83	60.00	77.50	6	42
Native	Qwen	Indexed SMILES	4.17	9.17	25.00	21	88
Native	Qwen	SMILES rewrite	13.33	62.50	16.67	18	27
Native	Qwen	STAMP rewrite	10.00	56.67	8.33	30	22
Native	DeepSeek	STAMP node-port	87.50	95.83	93.33	1	4
Native	DeepSeek	Indexed SMILES	61.67	75.83	73.33	22	7
Native	DeepSeek	SMILES rewrite	55.83	75.83	66.67	26	3
Native	DeepSeek	STAMP rewrite	48.33	77.50	61.67	25	2
Shared	Qwen	STAMP node-port	32.50	53.33	75.83	4	52
Shared	Qwen	Indexed SMILES	5.83	8.33	74.17	2	108
Shared	DeepSeek	STAMP node-port	85.00	94.17	98.33	0	7
Shared	DeepSeek	Indexed SMILES	80.00	93.33	97.50	0	8

Shared-label transfer narrowed the DeepSeek difference to 85.00% versus 80.00% success. For Qwen, plan accuracy was similar (75.83% versus 74.17%), but indexed actions failed to produce valid products in 108/118 second-stage cases, versus 52/116 with STAMP. These failures occur during execution of the predicted edit.

Independent identification dataset. Input molecules from the S^2-TOMG-Bench addition, deletion, substitution and LogP subsets (`MolEdit_AddComponent.csv`, `MolEdit_DelComponent.csv`, `MolEdit_SubComponent.csv` and `MolOpt_LogP.csv`) are merged and deduplicated before constructing seed/edited pairs. The constructor verifies the specified group-count changes, interleaves transformation types and selects 100 pairs. Reciprocal halo/hydroxyl changes contribute ten cases each; thiol-to-hydroxyl and aldehyde-to-carboxyl contribute four each; the remaining eight types contribute nine each. The set contains 52 distinct seed and 100 edited strings and differs from the transfer set. All molecule views are independently generated, without aligned nodes or atoms. All four views provide matched chemical information and use one call, the same output vocabulary and no repair.

Supplementary Table 11: Independent change identification with matched information. Accuracies are percentages over the same 100 molecule pairs per row. Exact requires both labels; removed and added report the individual fields. Invalid responses remain in the denominator. No molecular edit is executed.

Model	Input view	Exact	Removed	Added
Qwen	STAMP nodes	71.00	79.00	87.00
Qwen	Indexed SMILES	19.00	43.00	46.00
Qwen	SMILES string	15.00	24.00	34.00
Qwen	STAMP string	5.00	12.00	38.00
DeepSeek	STAMP nodes	97.00	98.00	99.00
DeepSeek	Indexed SMILES	71.00	81.00	90.00
DeepSeek	SMILES string	71.00	81.00	90.00
DeepSeek	STAMP string	64.00	74.00	88.00

Implementation. Evaluation used Python 3.12.12 and RDKit 2025.09.4. Deterministic compiler replay reused first-round responses; changed repair prompts triggered the corresponding new calls.

Supplementary References

1. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
2. Zhang, Z., Bian, Y., Xie, A., Han, P. & Zhou, S. Can pretrained models really learn better molecular representations for AI-aided drug discovery? *J. Chem. Inf. Model.* **64**, 2921–2930 (2024).
3. Yang, A. *et al.* Qwen2.5 technical report. Preprint at `https://arxiv.org/abs/2412.15115` (2024).
4. DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. Preprint at `https://arxiv.org/abs/2606.19348` (2026).
5. Lee, S. *et al.* GenMol: A drug discovery generalist with discrete diffusion. Preprint at `https://arxiv.org/abs/2501.06158` (2025).
6. Xue, S. *et al.* Any-order GPT as masked diffusion model: Decoupling formulation and architecture. Preprint at `https://arxiv.org/abs/2506.19935` (2025).
7. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: Towards chemical foundation models. Preprint at `https://arxiv.org/abs/2209.01712` (2022).