SIGMA: SEMANTIC IDENTIFIER GROUPING FOR MOLECULAR AUTOREGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive molecular models assign probability to molecular serializations even though chemical identity is invariant to serialization. Equivalent serializations can induce different internal states along a shared molecular continuation. Randomized strings broaden exposure, but do not explicitly supervise correspondence between those states. We introduce SIGMA, a dense suffix-position objective built from chemically certified same-suffix triplets: two equivalent histories, one non-equivalent history, and a shared suffix. SIGMA aligns corresponding pre-token hidden states along the continuation and enforces a finite similarity margin over the non-equivalent history, leaving the language-model objective, decoder, and inference procedure unchanged. We compare SIGMA with canonical training, randomized-serialization training, and last-token alignment across four datasets under SMILES and SELFIES. SIGMA lowers test-reference Fréchet ChemNet Distance against every control in all three training seeds on all four SELFIES datasets and on QM9 and ZINC under SMILES. Position-wise analyses show improved state correspondence, chemical discrimination, and top-1 next-token agreement while preserving between-molecule information. Full-corpus ZINC ablations under both representations identify correct state correspondence, rather than extra computation alone, as an effective ingredient. Beyond generation, a whole-molecule SIGMA extension improves mean predictive performance on all six molecular property benchmarks and reduces serialization sensitivity on every task relative to a matched randomized-serialization control.

1 INTRODUCTION

The combinatorial scale of drug-like chemical space precludes exhaustive experimental enumeration, motivating generative models that propose candidates beyond observed libraries (Reymond et al., 2012; Ruddigkeit et al., 2012; Polykovskiy et al., 2020). Autoregressive models factorize molecular generation into next-token decisions over serialized graphs, introducing a structural mismatch: probability is assigned in string space, whereas molecular identity is invariant to serialization. SMILES admits different graph traversals (Weininger, 1988), SELFIES retains alternative valid orders (Krenn et al., 2020), and SAFE uses fragments and attachment identifiers (Noutahi et al., 2024). For example, `CCO` and `C(C)O` both represent ethanol. Maximum likelihood remains correct for each string but does not explicitly couple equivalent histories, even when chemical checks establish their correspondence along an identical continuation.

The mismatch matters at the level where an autoregressive model acts: the probability distribution for its next token. Chemically equivalent molecular descriptions can elicit inconsistent internal representations or downstream predictions (Ganeeva et al., 2024; Yan et al., 2025). Randomized molecular strings broaden training exposure and form a strong baseline (Bjerrum, 2017; Arús-Pous et al., 2019), but observing two complete strings for one molecule does not identify which intermediate states correspond. Aligning only their final states likewise leaves intermediate prediction events without direct auxiliary supervision. We call the ordered pre-token hidden states along a continuation its *suffix prediction path*. Chemical verification of a shared continuation lets us align these states across equivalent histories while distinguishing a chemically different history.

We introduce SIGMA—**S**emantic **I**dentifier **G**rouping for **M**olecular **A**utoregression—to make intermediate-state correspondence an explicit training signal. The construction holds the continu-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

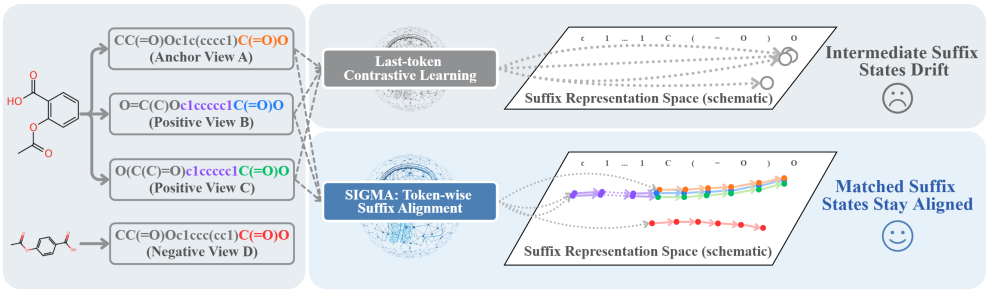


Figure 1: Last-token alignment directly supervises only the endpoint, whereas SIGMA aligns every position of an exact shared suffix and retains a margin-gated non-equivalent history. The artwork uses positive B for the upper example and positive C for the lower example; these are alternative illustrations, not simultaneous inputs. Each implemented triplet contains one anchor, one equivalent positive, and one non-equivalent negative. Artwork D therefore corresponds to the negative denoted C in the Method notation.

ation fixed while varying the serialization history: an anchor A, an equivalent positive B, and a valid but non-equivalent negative C share exactly the same suffix. Equivalence and non-equivalence are verified wherever the intermediate strings are jointly decodable, using the molecular-identity rule defined in the Method section. At each suffix position, the aligned states are taken immediately before predicting the same token, holding the prediction target fixed. SIGMA trains the anchor to identify the corresponding positive state and enforces a finite similarity margin over the negative. The negative term becomes inactive once that margin is satisfied.

Figure 1 contrasts this dense supervision with endpoint-only alignment. SIGMA leaves the LM head, decoder, and inference cost unchanged. We test whether improved correspondence translates into consistent next-token predictions without erasing distinctions between molecules, separating directly supervised state metrics from predictive agreement and information preservation. Generation and property prediction assess utility beyond the alignment objective.

We make four contributions:

- We formulate *serialization-induced prediction inconsistency* along a common continuation, separating explicit state correspondence from string enumeration.
- We introduce chemically certified same-suffix triplets and a dense suffix-position objective with a margin-gated non-equivalent history.
- We obtain lower FCD than canonical training, randomized-serialization training, and last-token alignment in all three seeds on six of eight SMILES/SELFIES representation–dataset blocks under matched training and decoding; SAFE exposes a fragment-assembly limitation.
- Mechanism analyses show improved next-token agreement with retained molecular information; compute-matched ZINC ablations support correct correspondence as an effective component. A whole-molecule extension improves mean performance and serialization stability on six property benchmarks over a matched randomized-serialization control.

2 Related Work

Chemical language models and molecular strings. Transformer backbones (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Jiang et al., 2023) underpin chemical models for representation learning and property prediction (Chithrananda et al., 2020; Wang et al., 2019; Honda et al., 2019; Fabian et al., 2020), autoregressive design (Bagal et al., 2022; Li et al., 2023), and tool use (Bran et al., 2024). Training resources range from PubChem, ChEMBL, and ZINC corpora (Wang et al., 2009; Gaulton et al., 2012; Irwin & Shoichet, 2005; Sterling & Irwin, 2015) to graph and multimodal biological views (Xu et al., 2024; Seal et al., 2022; 2023). SMILES admits multiple valid traversals (Weininger, 1988); DeepSMILES revises ring and branch

syntax (O'Boyle & Dalke, 2018), SELFIES constrains the grammar (Krenn et al., 2020), and t-SMILES/SAFE use fragments (Wu et al., 2024; Noutahi et al., 2024). These representations define molecular syntax; SIGMA adds supervision for corresponding hidden states along an identical token continuation.

Order augmentation and molecular language models. Randomized SMILES can improve molecular generative models by exposing more traversals (Bjerrum, 2017; Arús-Pous et al., 2019); related augmentation is also effective in reaction modeling (Tetko et al., 2020). All-SMILES aggregates several complete serializations (Alperstein et al., 2019). Equivalent descriptions can nevertheless elicit inconsistent representations or task predictions (Ganeeva et al., 2024; Yan et al., 2025), while strong point predictions need not imply reliable uncertainty estimates (Scalia et al., 2020). Enumeration supplies additional likelihood targets without explicitly matching the internal states that face a shared continuation. Our comparisons test the value of that correspondence signal beyond randomized-serialization training.

Molecular generation. Graph generators construct molecules through graph-space actions (You et al., 2018; Shi et al., 2020), supported by message-passing representations (Gilmer et al., 2017; Hu et al., 2020); reinforcement learning optimizes task rewards under sample-efficiency constraints (Olivecrona et al., 2017; Gao et al., 2022), and fragment retrieval augments generation (Lee et al., 2024). Learning-order models learn a state-dependent graph-generation order (Wang et al., 2025b). SIGMA addresses a complementary question: how the hidden states of a fixed string model should correspond when equivalent histories face a common continuation. It changes the training objective without learning a new generation order or altering the decoder.

Contrastive sequence learning. Contrastive learning commonly makes two augmented views agree at a pooled representation (Chen et al., 2020); molecular approaches similarly compare complete strings, graphs, or other global views (Wang et al., 2022; Pinheiro et al., 2022; Qian et al., 2024; Zhu et al., 2023). Token-aware contrastive objectives have also been used to shape individual language states rather than only pooled embeddings (Su et al., 2022; Zhou et al., 2024). At the same token-step granularity, PSV-PPO uses a partial-SMILES validity table to reshape entropy and distribution-matching losses at every autoregressive step, providing feedback before completion (Wang et al., 2025a). SIGMA instead aligns pre-token states from chemically equivalent histories and imposes a relative margin over a non-equivalent history, all under the same continuation. This construction fixes both the token target and its suffix offset. Last-token alignment tests sparse endpoint supervision, while matched ablations isolate supervision density from loss form.

Molecular generation evaluation. Benchmarks such as MOSES and GuacaMol emphasize that validity alone is insufficient (Polykovskiy et al., 2020; Brown et al., 2019). We therefore report Fréchet ChemNet Distance together with validity and likelihood. Because distribution fidelity does not verify the proposed mechanism, we separately measure hidden-state correspondence, next-token predictive consistency, and retention of molecular information.

3 METHOD

EQUIVALENT SERIALIZATION HISTORIES

Let $x = (x_1, \ldots, x_T)$ be a tokenized molecular string. The prefix $x_{<t}$ is its *history*, and $(x_{<t}, x_t)$ defines a *prediction event*. For optimizer-batch example i, let T_i count non-padding target tokens. SIGMA retains the token-normalized language-model (LM) loss,

$$\mathcal{L}_{\mathrm{LM}} = \frac{\sum_i \sum_{t=1}^{T_i} -\log p_\theta(x_{i,t} \mid x_{i,<t})}{\sum_i T_i}. \tag{1}$$

SIGMA supplements this likelihood with correspondence supervision between chemically verified histories along a shared continuation.

Illustrative construction. Consider equivalent SMILES histories $A =$ `CCO` and $B =$ `C(C)O`, the non-equivalent history $C =$ `CCC`, and shared suffix $S =$ `CC`. After each appended token, A and B

remain equivalent and C remains distinct; their completions are `CCOCC`, `C(C)OCC`, and `CCCCC`. Initial equivalence alone is insufficient: appending `C` to equivalent prefixes `CCO` and `OCC` instead yields different molecules, `CCOC` and `OCCC`. SIGMA rejects this pair by checking identity along the continuation, not only at its start.

Formally, prefixes A, B, C share the nonempty token sequence $S = (s_1, \ldots, s_m)$; $\oplus$ denotes concatenation. At every jointly decodable checkpoint, histories extending A and B must have the same identity and those extending C a different one. Each comparison therefore fixes the next-token target and suffix offset while varying the history.

For a decoded string x, let $\mathrm{id}_{\mathrm{std}}(x)$ denote its complete Standard InChIKey, computed with RDKit after a canonical isomeric-SMILES round trip. Equality of these keys is the molecular-identity criterion used throughout the paper, including all reported equivalence checks and identity collisions. Standard InChI normalizes mobile hydrogen and some protonation/charge arrangements and omits the Fixed-H layer; this criterion is therefore coarser than tautomer-specific graph identity (Heller et al., 2015; InChI Trust, 2012). For the fixed continuation S,

$$A \equiv_S B \quad \Longleftrightarrow \quad \mathrm{id}_{\mathrm{std}}(A \oplus S) = \mathrm{id}_{\mathrm{std}}(B \oplus S),$$

subject to the position-wise chemical verification below. To measure whether equivalent histories actually make similar decisions, let $\pi_j^X = p_\theta(\cdot \mid X \oplus s_{<j})$:

$$\mathcal{D}_\theta(A, B; S) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{JS}(\pi_j^A, \pi_j^B). \tag{2}$$

Here JS denotes Jensen–Shannon divergence, so $\mathcal{D}_\theta(A, B; S) = 0$ exactly when the two histories yield identical next-token distributions at every suffix position. The chemical checks establish identity correspondence along the selected continuation, not equality of the distributions over alternative next tokens. SIGMA regularizes the pre-token hidden states rather than directly minimizing Eq. (2); we use this divergence to test whether state-level supervision also improves predictive consistency.

Chemically Certified Same-Suffix Triplets

Retained triplets require lossless tokenization, exact suffix reuse, matching decodability at each offset, valid context lengths, and the required identities; equivalent completions must recover the recorded source identity. Negatives are searched in length buckets satisfying $||C| - |A|| \leq \max(3, \mathrm{round}(.25|A|))$, with a hash-seeded start and fixed search limit. Length matching does not replace chemical verification. Algorithm A1 in Appendix D.1 gives the construction; Appendix E.7 verifies the example's intermediate and completed identities.

Suffix-Relative Pre-Token Alignment

A *state* is the final-layer hidden vector supplied to the LM head. Let g_t^X denote that vector at absolute position t in $X \oplus S$, and $r_X(j)$ the position immediately before suffix token s_j. We compare

$$h_j^X = g^X_{r_X(j)} = H_\theta(X \oplus s_{<j}), \qquad X \in \{A, B, C\}. \tag{3}$$

If the first suffix target has index b_X, then $r_X(j) = b_X + j - 2$. With zero-based indices and an explicit `GO` token, the example yields first-suffix state positions $(3, 5, 3)$, not one shared absolute index. Comparing pre-token states avoids conditioning on the target itself. Supervision covers real suffix tokens and excludes `EOS`.

Dense Suffix-Position Matching with Margin-Gated Negatives

At each suffix offset j, the positive term uses the anchor state h_j^A as a query to retrieve the corresponding pre-token position $r_B(j)$ in the positive sequence $B \oplus S$. The candidate set $\mathcal{V}_B$ contains all non-padding positions in that sequence, from the start-of-sequence (`GO`) through the end-of-sequence (`EOS`) state, while supervised targets are restricted to pre-token positions of real suffix

tokens. Other positions compete for retrieval probability but are not individually verified as chemically non-equivalent to the anchor. With cosine similarity and temperature τ,

$$q_\tau(k \mid h_j^A; B) = \frac{\exp(\cos(h_j^A, g_k^B)/\tau)}{\sum_{\ell \in \mathcal{V}_B} \exp(\cos(h_j^A, g_\ell^B)/\tau)},$$
$$\ell_j^+ = -\log q_\tau(r_B(j) \mid h_j^A; B). \tag{4}$$

Padding logits are set to $-\infty$ before softmax, excluding padding from the normalizer. The number of real candidates $|\mathcal{V}_B|$ still varies with length, so raw matching losses are not compared across datasets.

For $s_j^+ = \cos(h_j^A, h_j^B)$ and $s_j^- = \cos(h_j^A, h_j^C)$, the negative term requires a relative similarity margin γ:

$$\ell_j^{\text{gate}} = \max(0, s_j^- - s_j^+ + \gamma). \tag{5}$$

It vanishes when $s_j^+ \geq s_j^- + \gamma$, requesting finite separation rather than continued repulsion after the margin is met.

Molecular equivalence is symmetric, but the implemented position-matching objective is directed: states from A query positions in B. The loss is not averaged with the reverse $B \to A$ direction.

Let $\mathcal{I}$ index the retained example–suffix-offset pairs in an optimizer batch. Averaging the position-matching and margin losses over $\mathcal{I}$ gives $\mathcal{L}_+$ and $\mathcal{L}_{\text{gate}}$, respectively. The full training objective combines these terms with the ordinary next-token loss:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{aux}} \left(\mathcal{L}_+ + \rho_{\text{neg}} \mathcal{L}_{\text{gate}} \right). \tag{6}$$

Here λ_{aux} scales auxiliary supervision, while ρ_{neg} sets the margin term's weight relative to position matching. The effective negative coefficient is $\lambda_{\text{neg}}^{\text{eff}} = \lambda_{\text{aux}} \rho_{\text{neg}}$. We use the shared validation-selected setting $\lambda_{\text{aux}} = .075$, $\rho_{\text{neg}} = 1$, $\gamma = .10$, and $\tau = .07$.

Controlled Optimization

Coverage and initialization are fixed within each comparison. SELFIES uses the same verified training-split subset for all methods; its LM targets come from $A \oplus S$, or $B \oplus S$ for randomized training. SMILES uses the complete LM corpus and a separate triplet stream. Micro-batching preserves objective normalization. Auxiliary training adds no inference parameters or decoding cost; Appendices F and J.1 give execution details and training costs.

Why the Objective Has This Form

The following properties connect the design choices to their mathematical consequences. Appendix E.5 gives the complete statements and proofs.

Well-defined supervision and finite separation. Suffix-relative indexing identifies the unique pre-token state for each suffix prediction event. Masking padding before the position-matching softmax makes the loss and real-state gradients invariant to additional right padding. High matching probability implies a lower bound on the target–competitor cosine gap, while $\ell_j^{\text{gate}} = 0$ if and only if $s_j^+ - s_j^- \geq \gamma$. Consequently, the negative term vanishes once the required ordering is achieved, whereas collapsing aligned positive and negative states to the same direction incurs a nonzero margin loss when $\gamma > 0$.

Why every continuation position is aligned. Let the shared LM head be $z = Wh + b$ and $M_W = \max_u \|w_u\|_2$, where w_u is the row associated with token u. Evaluating the same continuation S under two histories gives

$$|\log p_\theta(S \mid A) - \log p_\theta(S \mid B)| \leq 2M_W \sum_{j=1}^{m} \|h_j^A - h_j^B\|_2. \tag{7}$$

Every suffix position contributes to the bound on teacher-forced continuation likelihood. Last-token alignment directly supervises only one of these states: shared parameters may affect earlier states, but an endpoint constraint alone does not bound their distances. The bound involves Euclidean distances, which cosine matching alone cannot control without control of state norms. We therefore evaluate next-token agreement and divergence, suffix-token likelihood, and retained molecular information alongside the auxiliary objective.

Local versus global scope. Verified triplets constrain only a subset of the histories reachable during generation. Two models can agree on that subset yet differ at an unconstrained history reached with probability r. If their outgoing distributions there differ by total variation δ, the construction in Appendix E.5 yields sequence distributions differing by $r\delta$. Local agreement therefore does not determine the complete generation distribution, motivating FCD as a separate empirical evaluation.

Applying the Verification to SMILES, SELFIES, and SAFE

SMILES uses atomwise tokens and parseable prefixes; SELFIES uses bracket tokens and decodable checkpoints. SAFE retains dots and attachment identifiers and additionally checks fragment and connection boundaries. Appendix D gives the verification rules and triplet counts.

4 Experimental Setup

Data, representations, and coverage. We evaluate QM9 (Ramakrishnan et al., 2014), ZINC (Irwin & Shoichet, 2005), ChEMBL (Gaulton et al., 2012), and MOSES (Polykovskiy et al., 2020), spanning small organic, purchasable drug-like, bioactive, and benchmark drug-like molecules. SMILES and SELFIES form the primary three-seed study; SAFE provides a single-seed fragment-assembly analysis. All methods share initialization and LM training rows within each block. SMILES uses the full training split plus auxiliary triplets; SELFIES uses a certified training-split subset. Validation data select settings and checkpoints. Separately, SELFIES FCD uses a method-independent certifiable test subset solely as an evaluation reference, never for optimization or training-triplet construction. Absolute SMILES and SELFIES FCD values are therefore not corpus-matched comparisons. Appendices D and F give split, triplet, and reference counts and the independent RDKit/InChI verification (Landrum et al., 2026; Heller et al., 2015).

Controlled methods and training. Within each representation–dataset comparison, all methods share a six-layer, eight-head, 256-dimensional GPT-2, tokenizer, optimizer schedule, three training epochs, and seeds 101, 202, and 303. They differ in the serialization used for LM training and the auxiliary supervision:

- **Canonical** applies ordinary next-token likelihood to the designated canonical molecular strings.
- **Randomized-serialization training** applies the same likelihood to their paired randomized valid serializations, with no alignment loss.
- **Last-token alignment** adds one positive-versus-negative comparison at the pre-token state for the final real suffix token.
- **SIGMA** applies positive suffix-position matching and a margin-gated negative at every real suffix position.

The randomized control replaces each canonical training view with its fixed paired serialization; it does not resample views online. Last-token alignment tests endpoint-only supervision. The shared validation-selected SIGMA setting uses $\lambda_{\mathrm{aux}} = .075$, $\rho_{\mathrm{neg}} = 1$, and $\gamma = .10$, giving effective positive and gated-negative coefficients of .075. Appendix F lists the full hyperparameters and filtering procedure.

Evaluation and reporting. Fréchet ChemNet Distance (FCD; lower is better) measures generation fidelity (Preuer et al., 2018). Validity, diversity, connectedness, likelihood, and representation probes provide complementary checks. The primary SMILES/SELFIES comparison uses untruncated categorical sampling ($p = 1$) at temperature 1, targeting 5,000 valid molecules per model and

Table 1: Test-reference FCD under SMILES and SELFIES (mean ± standard deviation across three seeds; lower is better; bold is best).

	SMILES				SELFIES			
Dataset	Canon.	Random.	Last-token	SIGMA	Canon.	Random.	Last-token	SIGMA
QM9	.665±.061	.801±.029	.647±.057	**.583±.044**	.856±.064	1.051±.108	.849±.042	**.685±.012**
ZINC	1.966±.246	2.349±.207	1.996±.208	**1.835±.198**	6.805±.112	8.176±.116	6.986±.201	**6.253±.124**
ChEMBL	**1.591±.085**	1.733±.096	1.603±.100	1.609±.088	3.163±.068	3.763±.076	3.264±.106	**3.110±.064**
MOSES-300k	5.797±.150	6.136±.039	5.805±.069	**5.750±.085**	9.661±.270	11.765±.104	9.940±.052	**9.200±.152**

generation repetition. For each of three training seeds, we average FCD over five paired generations; tables report the mean ± standard deviation of these seed-level averages. Loss weights are selected using validation FCD and NLL. The primary generation results in Table 1 use frozen formal-test references; the separately reported objective ablations use validation references. Appendices A, F, and G give metric definitions, reference construction, per-seed results, and the complete statistical analysis.

Molecular property prediction. For whole-molecule prediction on six MoleculeNet benchmarks (Wu et al., 2018), we apply the equivalence principle at the EOS state, which represents the complete molecule for an affine property head. Two serializations of the same molecule form the positive pair, and a label-aware negative is constrained by a finite margin. Validation-selected checkpoints are compared with a matched randomized-serialization model without alignment. Appendix F.7 specifies the property head, task-specific settings, and training protocol.

5 Results

We evaluate generation fidelity, predictive consistency, and retained molecular information, then use controlled ablations to distinguish correct correspondence from additional computation.

Generation Fidelity Across Representations

Under SMILES (Table 1), SIGMA lowers FCD from the strongest control's $.647 \pm .057$ to $.583 \pm .044$ on QM9 and from $1.966 \pm .246$ to $1.835 \pm .198$ on ZINC, reductions of 9.9% and 6.7%. SIGMA outperforms every control in all three seeds on both datasets. On MOSES-300k, its mean FCD of $5.750 \pm .085$ is lower than canonical training ($5.797 \pm .150$) and last-token alignment ($5.805 \pm .069$), although these smaller gains vary across seeds. ChEMBL remains the exception: canonical training has the lowest mean ($1.591 \pm .085$ versus SIGMA's $1.609 \pm .088$). SIGMA has lower mean FCD than randomized training but slightly higher mean FCD than last-token alignment. Appendices A.4–A.5 report per-seed results and paired comparisons.

Validity also improves over canonical training on QM9, ZINC, and MOSES-300k (.7601→.8055, .7382→.7610, and .8183→.8315), while decreasing slightly on ChEMBL (.8723→.8671). The QM9 and ZINC FCD gains thus accompany more valid outputs, rather than increased rejection. ChEMBL illustrates a different outcome: local correspondence improves without a lower mean FCD over canonical training. Appendix I.2 examines repeated-token position competition, ring closure, and stopping calibration in this case.

Under SELFIES (Table 1), SIGMA has the lowest mean FCD on every dataset and in every training seed. On QM9, FCD falls from the strongest control's $.849 \pm .042$ to $.685 \pm .012$; on ZINC, it falls from $6.805 \pm .112$ to $6.253 \pm .124$. Relative reductions over the strongest control are 19.4%, 8.1%, 1.7%, and 4.8% on QM9, ZINC, ChEMBL, and MOSES-300k, respectively. Appendices A.4–A.5 give per-seed results and the complete comparisons.

CONSISTENCY ACROSS SUFFIX POSITIONS

An independent implementation verifies the QM9 supervision under the declared identity rule: retained positive pairs share a Standard InChIKey, retained negative histories have a different key, and all three strings reuse the exact suffix. It finds zero anchor–negative Standard-InChIKey collisions in that corpus.

The mechanism analysis tests whether state alignment translates into more consistent predictions while retaining molecular information. At each suffix position, we measure next-token distribution agreement and shared-suffix likelihood alongside positive-position retrieval and negative ordering. Retrieval and ordering are directly constrained by the auxiliary loss; next-token agreement is not explicitly optimized by it. Likelihood checks whether greater agreement accompanies better token prediction, and profiles over suffix positions locate these effects along the continuation.

With sim denoting cosine similarity, the two primary state-level indicators at offset j are

$$R_j^+ = \mathbf{1}\left[\arg\max_{t \in \mathcal{V}_B} \mathrm{sim}(h_j^A, g_t^B) = r_B(j)\right],$$
$$V_j^- = \mathbf{1}\left[\mathrm{sim}(h_j^A, h_j^C) \geq \mathrm{sim}(h_j^A, h_j^B)\right]. \tag{8}$$

$R_j^+ = 1$ denotes correct retrieval of the suffix-aligned position. $V_j^- = 1$ denotes an ordering violation: the negative is at least as similar to the anchor as the positive. We summarize these indicators in ten normalized suffix-position bins to locate the effect. A separate density ablation tests the benefit of supervising multiple positions rather than one.

Top-1 next-token agreement between equivalent histories improves across all four SMILES datasets and three seeds, rising from .7234 to .8014 on QM9. The changes therefore extend beyond the directly supervised state quantities to model predictions. Figure 2 shows that retrieval and ordering improve throughout the continuation; on QM9, the ordering-violation rate falls from .5405 to .2383 and correct-position retrieval rises from .8875 to .9912. All six suffix-level diagnostics improve in every dataset and seed. On ChEMBL, these local improvements coexist with higher mean FCD than canonical training, showing that local predictive consistency and global distribution fidelity measure distinct aspects of model behavior.

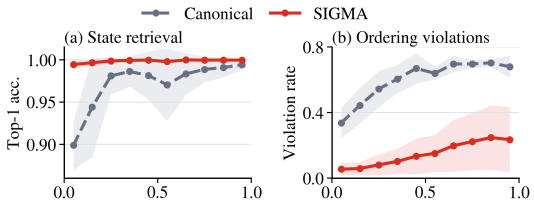


Figure 2: Position-wise SMILES mechanism profiles. Curves average four datasets and three seeds; ribbons show one standard deviation over the 12 dataset–seed conditions. Suffix progress runs from 0 (start) to 1 (end). SIGMA improves state retrieval and reduces ordering violations throughout the continuation.

SELFIES shows the same pattern across all four datasets and three seeds: all six diagnostics improve, including lower negative log-likelihood (NLL) for the shared suffix. Appendix C defines each diagnostic and reports the complete cross-domain results.

Chemical information is preserved. Covariance effective rank measures how broadly embedding variation is spread across directions; descriptor probes measure how well molecular properties can be recovered from those embeddings. Effective rank and descriptor-probe R^2 increase on all four SMILES domains, while fingerprint macro-AUROC remains within .004 of canonical training. Scaffold-conditioned similarities also retain between-molecule separation. Together, these checks indicate that greater agreement between equivalent histories does not come at the expense of collapsing molecular representations. Appendix C reports the probe definitions and complete values.

SAFE boundary. In the single-seed SAFE analysis, local consistency improves, but last-token alignment retains lower FCD and higher connectedness on all four domains. Among connected outputs, SIGMA instead has lower FCD on three domains. This contrast points to fragment assembly as a remaining bottleneck: local correspondence does not ensure connected molecules. Appendix I.1 reports the full results and connectivity analysis.

Molecular property prediction. The whole-molecule extension improves all six task means over matched randomized training (Table 2); ESOL, FreeSolv, Lipophilicity, and BACE improve in every seed. ESOL RMSE falls from $1.013 \pm .081$ to $.970 \pm .082$ and Lipophilicity from $.758 \pm .021$ to $.722 \pm .005$. Predictions average two serializations; their variation decreases on every task (0.2–44.9%). Appendix A.2 gives complete comparisons.

Table 2: Six-task prediction (three-seed mean $\pm$ SD; bold: better mean). RMSE ↓, AUROC ↑. Full results: Appendix A.2.

Task	Metric	Randomized serialization	SIGMA
ESOL	RMSE	$1.013 \pm .081$	$\mathbf{.970 \pm .082}$
FreeSolv	RMSE	$2.320 \pm .155$	$\mathbf{2.246 \pm .108}$
Lipophilicity	RMSE	$.758 \pm .021$	$\mathbf{.722 \pm .005}$
BACE	AUROC	$.823 \pm .012$	$\mathbf{.834 \pm .013}$
BBBP	AUROC	$.920 \pm .026$	$\mathbf{.943 \pm .005}$
ClinTox	AUROC	$.761 \pm .155$	$\mathbf{.854 \pm .061}$

What Produces the Gain?

The full-corpus ZINC ablations hold LM data, initialization, and optimizer schedule fixed. The auxiliary conditions also match the triplet stream and forward-pass budget, varying the supervision applied to those states. Table 3 compares a prespecified symmetric-KL output-distribution variant, positive-only **SIGMA-Pos**, a shuffled-correspondence control that disrupts chemical pairing while retaining the auxiliary computation, and the default **SIGMA-Margin**.

Both hidden-state variants improve over the full-corpus LM and shuffled control. SIGMA-Margin is best on SMILES ($1.043 \pm .064$), whereas SIGMA-Pos is best on SELFIES ($5.725 \pm .061$). The tested symmetric-KL variant does not reproduce these gains, supporting correct correspondence rather than computation alone.

A separate QM9 validation factorial finds a dense-position advantage over last-token supervision under both matched positive-only and positive-plus-negative objectives. Appendix B reports all three blocks, the density comparison, paired statistics, and training costs.

Table 3: Full-corpus ZINC validation FCD (mean $\pm$ seed SD; lower is better). S/F denote SMILES/SELFIES. Appendix B gives all three blocks and paired comparisons.

Training objective	S–ZINC	F–ZINC
Full-corpus LM	1.094±.044	6.330±.114
Output-distribution alignment	1.137±.043	6.401±.226
SIGMA-Pos	1.062±.069	**5.725±.061**
Shuffled-pair compute control	1.087±.023	6.333±.068
SIGMA-Margin	**1.043±.064**	5.788±.126

6 Discussion and Limitations

Verified correspondence improves predictive consistency while retaining molecular information, but ChEMBL and SAFE show that local agreement does not ensure calibrated closure, stopping, or connected assembly. Coverage differs across representations, preventing direct SMILES–SELFIES FCD comparisons. Triplet construction and auxiliary passes increase training cost, not inference cost.

Standard-InChIKey equivalence can merge mobile-H tautomers and some charge drawings; a graph-exact contract requires new triplets and re-evaluation. Broader coverage, larger backbones, and laboratory utility remain open directions.

7 Conclusion

SIGMA turns molecular serialization equivalence into dense supervision through verified same-suffix triplets. A shared setting lowers FCD against all three controls in each training seed on four SELFIES datasets and SMILES–QM9/ZINC. Position-wise diagnostics additionally show more consistent next-token predictions and retained molecular information, so the observed benefits extend beyond quantities directly optimized by the auxiliary loss. Controlled ablations support correct hidden-state correspondence as an effective ingredient, while a whole-molecule extension improves mean performance on six property benchmarks. SIGMA thus makes correspondence along the prediction path an actionable training signal, complementing token likelihood while preserving the model architecture and autoregressive inference procedure.

AI USE STATEMENT

The research direction and central ideas were developed by the authors. Generative AI tools assisted with refining hypotheses and mathematical arguments, code and analysis support, interpretation of results, manuscript editing, and figure preparation. The authors reviewed and verified AI-assisted outputs against the cited literature, retained source and result artifacts, and executable analyses used in the paper. The authors take full responsibility for the final content, results, and conclusions.

ETHICS STATEMENT

This work studies computational molecular representations using public benchmark data and does not involve human participants or clinical decision making. The method is intended to improve the reliability of molecular machine-learning models, not to establish molecular safety, synthesis feasibility, or therapeutic efficacy; those questions require appropriate experimental validation.

REPRODUCIBILITY STATEMENT

The appendices document the representation-specific triplet construction, training and sampling settings, paired statistical analysis, and mechanism checks. The anonymous code package contains the corresponding implementations, configurations, aggregate inputs, and validation tests.

REFERENCES

Zaccary Alperstein, Artem Cherkasov, and Jason Tyler Rolfe. All SMILES variational autoencoder, 2019. URL https://arxiv.org/abs/1905.13343.

Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1):71, 2019. doi: 10.1186/s13321-019-0393-0.

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2022. doi: 10.1021/acs.jcim.1c00600.

Esben Jannik Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules, 2017. URL https://arxiv.org/abs/1703.07076.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th ICML*, pp. 1597–1607. PMLR, 2020.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020. URL https://arxiv.org/abs/2011.13230.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12994–13013, 2024. doi: 10.18653/v1/2024.findings-emnlp.760.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems*, 35: 21342–21357, 2022.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012. doi: 10.1093/nar/gkr777.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th ICML*, pp. 1263–1272. PMLR, 2017.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7:23, 2015. doi: 10.1186/s13321-015-0068-4.

Shion Honda, Shoi Shi, and Hiroki R Ueda. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery, 2019. URL https://arxiv.org/abs/1911.04738.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

InChI Trust. Inchi technical faq. https://www.inchi-trust.org/technical-faq/, 2012. Last modified 2012-05-12; Sections 4.3, 4.8, and 6.1–6.2; accessed 2026-07-29.

John J Irwin and Brian K Shoichet. ZINC: a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005. doi: 10.1021/ci049714+.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947.

Greg Landrum et al. RDKit: Open-source cheminformatics. Software, https://www.rdkit.org, 2026. Version 2026.3.1; accessed 2026-07-29.

Seul Lee, Karsten Kreis, Srimukh Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Arash Vahdat, and Weili Nie. Molecule generation with fragment retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:132463–132490, 2024. doi: 10.52202/079017-4211.

Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu, and Suxia Han. DrugGPT: a GPT-based strategy for designing potential ligands targeting specific proteins. *bioRxiv*, pp. 2023.06.29.543848, 2023. doi: 10.1101/2023.06.29.543848.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pre-training approach, 2019. URL https://arxiv.org/abs/1907.11692.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C. Lim, and Prudencio Tossou. Gotta be SAFE: A new framework for molecular design. *Digital Discovery*, 3:796–804, 2024. doi: 10.1039/D4DD00019F.

Noel O'Boyle and Andrew Dalke. Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures. ChemRxiv, 2018.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017. doi: 10.1186/s13321-017-0235-x.

Gabriel A Pinheiro, Juarez LF Da Silva, and Marcos G Quiles. SmiCLR: Contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *Journal of Chemical Information and Modeling*, 62(17):3948–3960, 2022. doi: 10.1021/acs.jcim.2c00521.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020. doi: 10.3389/fphar.2020.565644.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018. doi: 10.1021/acs.jcim.8b00234.

Ying Qian, Minghua Shi, and Qian Zhang. CONSMI: contrastive learning in the simplified molecular input line entry system helps generate better molecules. *Molecules*, 29(2):495, 2024. doi: 10.3390/molecules29020495.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

Jean-Louis Reymond, Lars Ruddigkeit, Lorenz Blum, and Ruud van Deursen. The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(5):717–733, 2012. doi: 10.1002/wcms.1104.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d.

Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, 2020. doi: 10.1021/acs.jcim.9b00975.

Srijit Seal, Jordi Carreras-Puigvert, Maria-Anna Trapotsi, Hongbin Yang, Ola Spjuth, and Andreas Bender. Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. *Communications Biology*, 5(1):858, 2022. doi: 10.1038/s42003-022-03763-5.

Srijit Seal, Hongbin Yang, Maria-Anna Trapotsi, Satvik Singh, Jordi Carreras-Puigvert, Ola Spjuth, and Andreas Bender. Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. *Journal of Cheminformatics*, 15(1):56, 2023. doi: 10.1186/s13321-023-00723-x.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: a flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations*, 2020.

Teague Sterling and John J Irwin. ZINC 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2497–2507, 2022. doi: 10.18653/v1/2022.findings-naacl.191.

Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11 (1):5575, 2020. doi: 10.1038/s41467-020-19266-y.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM-BCB*, pp. 429–436, 2019. doi: 10.1145/3307339.3342186.

Xinyu Wang, Jinbo Bi, and Minghu Song. Leveraging partial SMILES validation scheme for enhanced drug design in reinforcement learning frameworks, 2025a. URL https://arxiv.org/abs/2505.00530.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(Web Server issue):W623–W633, 2009. doi: 10.1093/nar/gkp456.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022. doi: 10.1038/s42256-022-00447-x.

Zhe Wang, Jiaxin Shi, Nicolas Heess, Arthur Gretton, and Michalis K Titsias. Learning-order autoregressive models with application to molecular graph generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64282–64301. PMLR, 2025b.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Juan-Ni Wu, Tong Wang, Yue Chen, Li-Juan Tang, Hai-Long Wu, and Ru-Qin Yu. t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1):4993, 2024. doi: 10.1038/s41467-024-49388-6.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, et al. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Lei Xu, Leiming Xia, Shourun Pan, and Zhen Li. Triple generative self-supervised learning method for molecular property prediction. *International Journal of Molecular Sciences*, 25(7):3794, 2024. doi: 10.3390/ijms25073794.

Bing Yan, Angelica Chen, and Kyunghyun Cho. Inconsistency of llms in molecular representations. *Digital Discovery*, 4:2876–2892, 2025. doi: 10.1039/D5DD00176E.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in Neural Information Processing Systems*, 31, 2018.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17114–17122, 2024. doi: 10.1609/aaai.v38i15.29656.

Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Wengang Zhou, Tao Qin, Houqiang Li, and Tie-Yan Liu. Dual-view molecular pre-training. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3615–3627, 2023.

APPENDIX: ADDITIONAL RESULTS AND TECHNICAL DETAILS

The appendices begin with complete generation and property-prediction results, followed by controlled ablations and mechanism analyses. Construction rules, proofs, and experimental protocols follow the empirical evidence. Comparisons are matched within each representation–dataset block; the sensitivity studies and representation-specific boundaries retain their separate evaluation settings.

Location	Contents
Appendix A	Complete generation and property-prediction results, per-seed values, and paired comparisons.
Appendix B	Controlled full-corpus ablations and dense-versus-last-token supervision.
Appendix C	Predictive consistency, molecular-information preservation, and diagnostic definitions.
Appendices D–E	Triplet construction, identity checks, the formal objective, and proofs.
Appendices F–G	Data, optimization, generation, property-prediction protocols, and statistics.
Appendices H–I	Weight and sampling sensitivity, SAFE and ChEMBL analyses, and scope.
Appendix J	Computational cost and reproducibility resources.

A COMPLETE GENERATION AND PROPERTY-PREDICTION RESULTS

A.1 THREE-SEED SUMMARY

FCD is the mean of the three training-seed means and the reported standard deviation is across those three means. Each seed mean itself averages five paired generation repetitions.

Dataset	Canonical	Randomized serialization	Last-token alignment	SIGMA
QM9	.6655±.0610	.8013±.0286	.6474±.0568	**.5832±.0441**
ZINC	1.9655±.2463	2.3486±.2065	1.9960±.2080	**1.8348±.1975**
ChEMBL	**1.5914±.0851**	1.7327±.0958	1.6034±.0999	1.6090±.0879
MOSES-300k	5.7967±.1501	6.1355±.0387	5.8048±.0690	**5.7502±.0851**

Table A1: SMILES test-reference FCD (mean ± standard deviation across three seeds; lower is better; bold is best).

Dataset	Canonical	Randomized serialization	Last-token alignment	SIGMA
QM9	.8557±.0644	1.0510±.1080	.8491±.0418	**.6846±.0120**
ZINC	6.8046±.1125	8.1759±.1164	6.9863±.2009	**6.2527±.1241**
ChEMBL	3.1631±.0683	3.7631±.0756	3.2637±.1064	**3.1095±.0644**
MOSES-300k	9.6606±.2699	11.7648±.1037	9.9398±.0523	**9.1999±.1524**

Table A2: SELFIES test-reference FCD (mean ± standard deviation across three seeds; lower is better; bold is best).

Table A1 reports the SMILES summary and exposes the SMILES–ChEMBL boundary case alongside the clear QM9 and ZINC gains. Table A2 reports the matched SELFIES summary, for which SIGMA has the lowest mean on all four domains.

Generation results. For SELFIES, SIGMA has the lowest mean FCD among the four controlled methods on all four datasets and in each of the three seeds; all twelve paired 95% intervals are below zero. For SMILES, all three intervals are below zero on QM9 and ZINC. On MOSES, the interval is below zero against randomized-serialization training but includes zero against canonical training and last-token alignment. On ChEMBL, SIGMA is better than randomized-serialization training, has no clear difference from last-token alignment, and is slightly worse than canonical training in all three seeds. Together, the two representations show broad gains together with the SMILES–ChEMBL case in which local consistency improves but FCD does not.

Task	Metric	Randomized-serialization baseline	EOS-level SIGMA	Δ	Seeds	95% CI
ESOL	RMSE	$1.0134 \pm .0810$	$.9702 \pm .0824$	$-.0433$	3/3	$[-.0754, -.0101]$
FreeSolv	RMSE	$2.3196 \pm .1552$	$2.2463 \pm .1083$	$-.0734$	3/3	$[-.2011, .0519]$
Lipophilicity	RMSE	$.7580 \pm .0214$	$.7223 \pm .0055$	$-.0356$	3/3	$[-.0634, -.0112]$
BACE	AUROC	$.8232 \pm .0123$	$.8344 \pm .0131$	$+.0112$	3/3	$[-.0064, .0277]$
BBBP	AUROC	$.9200 \pm .0263$	$.9434 \pm .0051$	$+.0234$	2/3	$[-.0003, .0450]$
ClinTox	AUROC	$.7606 \pm .1550$	$.8542 \pm .0607$	$+.0936$	2/3	$[-.0165, .2640]$

Table A3: Complete six-task molecular-property comparison. Regression deltas are $\mathrm{RMSE}_{\mathrm{SIGMA}} - \mathrm{RMSE}_{\mathrm{baseline}}$; classification deltas are $\mathrm{AUROC}_{\mathrm{SIGMA}} - \mathrm{AUROC}_{\mathrm{baseline}}$, so the beneficial sign differs by task type. "Seeds" counts beneficial training seeds. Confidence intervals use 10,000 paired seed–molecule hierarchical bootstrap draws.

Task	Randomized serialization	SIGMA	Relative change
ESOL	.2809	.2310	-17.77%
FreeSolv	.5176	.5166	-0.19%
Lipophilicity	.2273	.1868	-17.82%
BACE	.1039	.0573	-44.89%
BBBP	.0741	.0484	-34.66%
ClinTox	.0453	.0352	-22.24%

Table A4: Mean within-molecule prediction SD across canonical and paired alternate serializations. Lower values indicate less serialization-sensitive predictions. These are descriptive mechanism measurements.

A.2 Molecular Property Prediction Results

Table A3 reports the complete six-task predictive comparison. Table A4 measures the complementary change in sensitivity to alternate serializations of the same molecule.

EOS-level SIGMA improves the mean primary metric on all six tasks and moves in the beneficial direction in all three training seeds for ESOL, FreeSolv, Lipophilicity, and BACE. The 95% intervals for ESOL and Lipophilicity are wholly on the beneficial side of zero. The other four tasks have favorable point estimates but inconclusive intervals.

The prediction gap between two serializations of the same molecule decreases on every task, with the largest relative reductions on BACE, BBBP, and ClinTox. This supports the intended serialization invariance and complements the task-level comparison.

A.3 Validation NLL, Molecular Validity, and Connectedness

Rep.	Dataset	Canonical	Randomized serialization	Last-token alignment	SIGMA
SMILES	QM9	$.911 \pm .003$	$.943 \pm .004$	$.908 \pm .001$	$.890 \pm .002$
SMILES	ZINC	$.743 \pm .004$	$.799 \pm .005$	$.746 \pm .003$	$.732 \pm .003$
SMILES	ChEMBL	$.621 \pm .001$	$.661 \pm .000$	$.620 \pm .002$	$.620 \pm .002$
SMILES	MOSES	$.610 \pm .000$	$.651 \pm .002$	$.611 \pm .001$	$.605 \pm .001$
SELFIES	QM9	$1.045 \pm .004$	$1.034 \pm .007$	$1.050 \pm .004$	$1.007 \pm .002$
SELFIES	ZINC	$.907 \pm .002$	$.992 \pm .006$	$.912 \pm .002$	$.884 \pm .005$
SELFIES	ChEMBL	$.716 \pm .002$	$.770 \pm .001$	$.718 \pm .001$	$.712 \pm .002$
SELFIES	MOSES	$.699 \pm .002$	$.767 \pm .001$	$.702 \pm .001$	$.689 \pm .000$

Table A5: Validation NLL (mean $\pm$ training-seed SD) from the same final checkpoints.

Table A5 checks whether the generation results coincide with a general loss of next-token likelihood. SIGMA lowers validation NLL on seven of eight blocks and matches canonical training to

the displayed precision on SMILES–ChEMBL. Connectedness is 1.000 for all 32 representation–dataset–method combinations. Thus the reported FCD differences are not explained by a shift toward disconnected outputs. Validity must nevertheless be read beside FCD: for example, SMILES randomized-serialization training lowers validity substantially on all four datasets. SIGMA lowers validation NLL relative to canonical training on QM9, ZINC, and MOSES-300k, while the two ChEMBL values both round to .620. Thus the weaker SMILES–ChEMBL FCD result is not a corpus-wide likelihood failure.

Novelty is not uniformly preserved. On SELFIES-QM9 it falls from .7277 for canonical training to .6944 for SIGMA, with the same direction in all three training seeds; on SMILES-QM9 it falls from .5983 to .5610. These changes make novelty an explicit secondary endpoint rather than an unqualified benefit.

Table A6 reports the complete validity, uniqueness, and novelty values underlying these guardrail statements. In particular, it contains the four SMILES validity changes quoted in the main paper; connectedness rounds to 1.000 throughout.

Representation	Dataset	Method	Validity ↑	Uniqueness ↑	Novelty ↑
SMILES	QM9	Canonical	.7601	.9823	.5983
		Randomized serialization	.6188	.9822	.6551
		Last-token alignment	.7538	.9815	.5847
		SIGMA	.8055	.9819	.5610
SMILES	ZINC	Canonical	.7382	.9999	.9996
		Randomized serialization	.6363	.9998	.9997
		Last-token alignment	.7152	.9999	.9995
		SIGMA	.7610	.9999	.9995
SMILES	ChEMBL	Canonical	.8723	.9992	.9743
		Randomized serialization	.8323	.9997	.9809
		Last-token alignment	.8703	.9994	.9744
		SIGMA	.8671	.9991	.9759
SMILES	MOSES-300k	Canonical	.8183	.9997	.9920
		Randomized serialization	.7477	.9998	.9933
		Last-token alignment	.8148	.9998	.9918
		SIGMA	.8315	.9998	.9915
SELFIES	QM9	Canonical	1.0000	.9701	.7277
		Randomized serialization	1.0000	.9687	.7838
		Last-token alignment	1.0000	.9715	.7343
		SIGMA	1.0000	.9707	.6944
SELFIES	ZINC	Canonical	1.0000	.9988	.9999
		Randomized serialization	1.0000	.9978	.9998
		Last-token alignment	1.0000	.9990	.9999
		SIGMA	1.0000	.9990	.9998
SELFIES	ChEMBL	Canonical	1.0000	.9996	.9862
		Randomized serialization	.9999	.9997	.9894
		Last-token alignment	1.0000	.9995	.9858
		SIGMA	.9999	.9996	.9866
SELFIES	MOSES-300k	Canonical	1.0000	.9996	.9961
		Randomized serialization	1.0000	.9993	.9970
		Last-token alignment	1.0000	.9994	.9967
		SIGMA	1.0000	.9996	.9958

Table A6: Secondary generation guardrails from the same three-seed, five-repetition formal-test replay as the primary FCD tables. Connectedness rounds to 1.000 for all 32 blocks. These metrics reveal the validity and novelty tradeoffs that accompany FCD rather than serving as additional selection objectives.

A.4 PER-TRAINING-SEED FCD

Table A7 reports the seed-level means underlying each mean±SD summary. Each number is the mean of five paired generation repetitions; the three directions provide a complementary stability diagnostic for the paired intervals.

Rep.	Dataset	Method	Seed 101	Seed 202	Seed 303
SMILES	QM9	Canonical	.649	.615	.733
SMILES	QM9	Randomized serialization	.797	.832	.775
SMILES	QM9	Last-token alignment	.646	.591	.705
SMILES	QM9	SIGMA	.588	.537	.625
SMILES	ZINC	Canonical	2.250	1.828	1.819
SMILES	ZINC	Randomized serialization	2.577	2.176	2.292
SMILES	ZINC	Last-token alignment	2.235	1.854	1.899
SMILES	ZINC	SIGMA	2.059	1.686	1.759
SMILES	ChEMBL	Canonical	1.675	1.505	1.595
SMILES	ChEMBL	Randomized serialization	1.814	1.627	1.757
SMILES	ChEMBL	Last-token alignment	1.707	1.508	1.595
SMILES	ChEMBL	SIGMA	1.694	1.519	1.614
SMILES	MOSES	Canonical	5.945	5.800	5.645
SMILES	MOSES	Randomized serialization	6.161	6.155	6.091
SMILES	MOSES	Last-token alignment	5.883	5.780	5.752
SMILES	MOSES	SIGMA	5.809	5.788	5.653
SELFIES	QM9	Canonical	.857	.920	.791
SELFIES	QM9	Randomized serialization	1.112	1.115	.926
SELFIES	QM9	Last-token alignment	.863	.882	.802
SELFIES	QM9	SIGMA	.682	.698	.674
SELFIES	ZINC	Canonical	6.703	6.786	6.925
SELFIES	ZINC	Randomized serialization	8.310	8.107	8.111
SELFIES	ZINC	Last-token alignment	6.755	7.092	7.112
SELFIES	ZINC	SIGMA	6.154	6.392	6.212
SELFIES	ChEMBL	Canonical	3.183	3.087	3.220
SELFIES	ChEMBL	Randomized serialization	3.850	3.720	3.719
SELFIES	ChEMBL	Last-token alignment	3.380	3.172	3.239
SELFIES	ChEMBL	SIGMA	3.162	3.038	3.129
SELFIES	MOSES	Canonical	9.373	9.701	9.908
SELFIES	MOSES	Randomized serialization	11.645	11.817	11.832
SELFIES	MOSES	Last-token alignment	9.924	9.898	9.998
SELFIES	MOSES	SIGMA	9.024	9.299	9.276

Table A7: Test-reference FCD by training seed under untruncated categorical sampling ($p = 1$, temperature 1).

A.5 Paired Contrasts

The estimand below is $\Delta = \mathrm{FCD}_{\mathrm{SIGMA}} - \mathrm{FCD}_{\mathrm{control}}$; negative values favor SIGMA. An interval wholly below zero provides clear evidence of lower FCD for SIGMA; intervals that include zero are reported as inconclusive. Seed directions are reported separately as a stability diagnostic.

Dataset	Control	Δ	95% CI	Seed directions
QM9	Canonical	-.0823	$[-.1119, -.0554]$	3/3 better
QM9	Randomized serialization	-.2181	$[-.2913, -.1543]$	3/3 better
QM9	Last-token alignment	-.0642	$[-.0858, -.0493]$	3/3 better
ZINC	Canonical	-.1307	$[-.1909, -.0627]$	3/3 better
ZINC	Randomized serialization	-.5138	$[-.5451, -.4874]$	3/3 better
ZINC	Last-token alignment	-.1612	$[-.1936, -.1282]$	3/3 better
ChEMBL	Canonical	+.0176	$[-.0080, +.0438]$	0/3 better
ChEMBL	Randomized serialization	-.1237	$[-.1599, -.0904]$	3/3 better
ChEMBL	Last-token alignment	+.0056	$[-.0193, +.0321]$	1/3 better
MOSES-300k	Canonical	-.0466	$[-.1375, +.0553]$	2/3 better
MOSES-300k	Randomized serialization	-.3853	$[-.4702, -.3100]$	3/3 better
MOSES-300k	Last-token alignment	-.0547	$[-.1307, +.0246]$	2/3 better

Table A8: Paired hierarchical-bootstrap SMILES test contrasts. The estimand is $\Delta = \mathrm{FCD}_{\mathrm{SIGMA}} - \mathrm{FCD}_{\mathrm{control}}$; negative values favor SIGMA.

Dataset	Control	Δ	95% CI	Seed directions
QM9	Canonical	-.1712	$[-.2259, -.1178]$	3/3 better
QM9	Randomized serialization	-.3664	$[-.4337, -.2552]$	3/3 better
QM9	Last-token alignment	-.1645	$[-.1967, -.1278]$	3/3 better
ZINC	Canonical	-.5519	$[-.7313, -.3676]$	3/3 better
ZINC	Randomized serialization	-1.9231	$[-2.1504, -1.7030]$	3/3 better
ZINC	Last-token alignment	-.7336	$[-.8992, -.5676]$	3/3 better
ChEMBL	Canonical	-.0536	$[-.1033, -.0085]$	3/3 better
ChEMBL	Randomized serialization	-.6536	$[-.7011, -.5848]$	3/3 better
ChEMBL	Last-token alignment	-.1542	$[-.2275, -.0875]$	3/3 better
MOSES-300k	Canonical	-.4607	$[-.6451, -.3069]$	3/3 better
MOSES-300k	Randomized serialization	-2.5648	$[-2.6604, -2.4482]$	3/3 better
MOSES-300k	Last-token alignment	-.7399	$[-.9022, -.5686]$	3/3 better

Table A9: Paired hierarchical-bootstrap SELFIES test contrasts. The estimand is $\Delta = \mathrm{FCD}_{\mathrm{SIGMA}} - \mathrm{FCD}_{\mathrm{control}}$; negative values favor SIGMA.

Table A8 gives the complete SMILES paired intervals, establishing clear reductions on QM9 and ZINC. Table A9 gives the complete SELFIES intervals; all twelve comparisons favor SIGMA with intervals wholly below zero.

B Controlled Ablations

B.1 Separating Data Coverage, Compute, and Alignment

The certified-subset study holds its triplet rows fixed across methods. To separate alignment from data filtering and additional computation, we evaluate three independently specified full-corpus validation blocks: SMILES–ZINC, SMILES–ChEMBL, and SELFIES–ZINC. Every condition uses the same full-corpus language-model stream, the same seed-level initialization, three epochs, and the same optimizer schedule. Auxiliary conditions each use three additional Transformer forwards on the same certified-triplet auxiliary stream. The five objectives are:

1. the full-corpus LM without an auxiliary objective;
2. a prespecified symmetric-KL alignment between aligned next-token distributions, treated as a SIGMA-family output-space variant;
3. positive-only certified state alignment (SIGMA-Pos);

4. the shuffled-pair compute control, which applies the SIGMA state loss after deliberately shuffling the chemical mapping and therefore preserves compute while destroying the intended correspondence; and
5. margin-gated certified state alignment (SIGMA-Margin).

For aligned pre-token states at suffix offsets $j \in \mathcal{J}$, the output-space variant is

$$\mathcal{L}_{\mathrm{out}} = \frac{1}{2|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left[D_{\mathrm{KL}}(p_j^A \| p_j^B) + D_{\mathrm{KL}}(p_j^B \| p_j^A) \right],$$

where p_j^A and p_j^B are the two full-vocabulary next-token distributions. Its total objective is $\mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{out}} \mathcal{L}_{\mathrm{out}}$ with $\lambda_{\mathrm{out}} = 1.0$, a single prespecified coefficient shared across the three blocks and seeds. This coefficient was not separately swept, so the experiment tests this particular output-space formulation rather than all possible strengths of output alignment. Gradients flow through both distributions; there is no stop-gradient teacher and no logit temperature ($T = 1$). The non-equivalent C trajectory is forwarded to match auxiliary compute but does not enter $\mathcal{L}_{\mathrm{out}}$. The variant therefore tests whether direct bidirectional decision matching is sufficient under this contract, rather than introducing an unrelated baseline. The complete grid contains 45 trained models and 225 paired FCD evaluations.

Table A10 reports the matched FCD, NLL, validity, and training-cost summaries. Table A11 gives the paired contrasts against the full-corpus LM, and Figure A1 visualizes the same three-block attribution pattern.

Block	Training objective	FCD ↓	NLL ↓	Validity ↑	Train time ×
SMILES–ChEMBL	Full-corpus LM	**1.076±.058**	.5890	0.9477	1.00
	Output-distribution alignment	1.136±.048	.5961	0.9506	3.40
	SIGMA-Pos	1.086±.071	.5882	0.9472	3.88
	Shuffled-pair compute control	1.125±.090	.5948	0.9396	3.88
	SIGMA-Margin	1.090±.050	**.5876**	0.9478	3.97
SMILES–ZINC	Full-corpus LM	1.094±.044	.6957	0.8079	1.00
	Output-distribution alignment	1.137±.043	.6966	0.8162	3.95
	SIGMA-Pos	1.062±.069	.6876	0.8206	4.29
	Shuffled-pair compute control	1.087±.023	.6940	0.7968	4.29
	SIGMA-Margin	**1.043±.064**	**.6866**	0.8219	4.50
SELFIES–ZINC	Full-corpus LM	6.330±.114	.9193	1.0000	1.00
	Output-distribution alignment	6.401±.226	.9111	0.9999	3.30
	SIGMA-Pos	**5.725±.061**	**.8905**	1.0000	3.50
	Shuffled-pair compute control	6.333±.068	.9106	0.9999	4.01
	SIGMA-Margin	5.788±.126	.8913	0.9999	4.05

Table A10: Separating data coverage, compute, and alignment on the full corpus. FCD is mean±SD over three training seeds; NLL and validity are three-seed means. Runtime is normalized to the full-corpus LM in the same representation–dataset block.

On SMILES–ZINC, both SIGMA-Pos and SIGMA-Margin improve over the full-corpus LM, with all three seed directions and intervals below zero. The tested symmetric-KL variant instead degrades FCD in all three seeds, and the shuffled-pair interval includes zero. SELFIES–ZINC gives the same identification: SIGMA-Pos and SIGMA-Margin both improve substantially, while the symmetric-KL variant and the compute control do not. SIGMA-Pos has the lower SELFIES mean, but its direct comparison with SIGMA-Margin is inconclusive ($\Delta_{\mathrm{SIGMA-positive}} = +.0625$, 95% CI $[-.0450, +.1617]$); the evidence therefore identifies certified hidden correspondence as the shared effective ingredient, not the negative term as a universally necessary component.

SMILES–ChEMBL remains a difficult case under full-corpus training. The symmetric-KL variant and the shuffled-pair compute control both have intervals wholly above zero relative to the full-corpus LM. SIGMA-Pos and SIGMA-Margin have intervals that include zero. SIGMA-Margin has a clear advantage over the symmetric-KL variant ($-.0456$, CI $[-.0693, -.0195]$) and the shuffled-pair compute control ($-.0350$, CI $[-.0792, -.0043]$), but does not beat the LM. Thus missing full-corpus coverage was a genuine earlier protocol limitation, not the full cause of the ChEMBL result.

Block	Alignment objective	ΔFCD	95% CI	Seeds improved
SMILES–ChEMBL	Output-distribution alignment	+.0593	$[+.0349, +.0792]$	0/3
	SIGMA-Pos	+.0100	$[-.0094, +.0278]$	1/3
	Shuffled-pair compute control	+.0487	$[+.0240, +.0856]$	0/3
	SIGMA-Margin	+.0137	$[-.0007, +.0265]$	0/3
SMILES–ZINC	Output-distribution alignment	+.0438	$[+.0208, +.0647]$	0/3
	SIGMA-Pos	-.0321	$[-.0571, -.0051]$	3/3
	Shuffled-pair compute control	-.0066	$[-.0324, +.0242]$	2/3
	SIGMA-Margin	-.0505	$[-.0774, -.0226]$	3/3
SELFIES–ZINC	Output-distribution alignment	+.0704	$[-.0626, +.1920]$	1/3
	SIGMA-Pos	-.6049	$[-.6890, -.5218]$	3/3
	Shuffled-pair compute control	+.0026	$[-.0713, +.0685]$	1/3
	SIGMA-Margin	-.5424	$[-.6578, -.4389]$	3/3

Table A11: Paired seed×generation-repetition hierarchical-bootstrap contrasts against the full-corpus LM. Differences are candidate minus LM, so negative values favor the auxiliary objective. No multiplicity adjustment is applied.

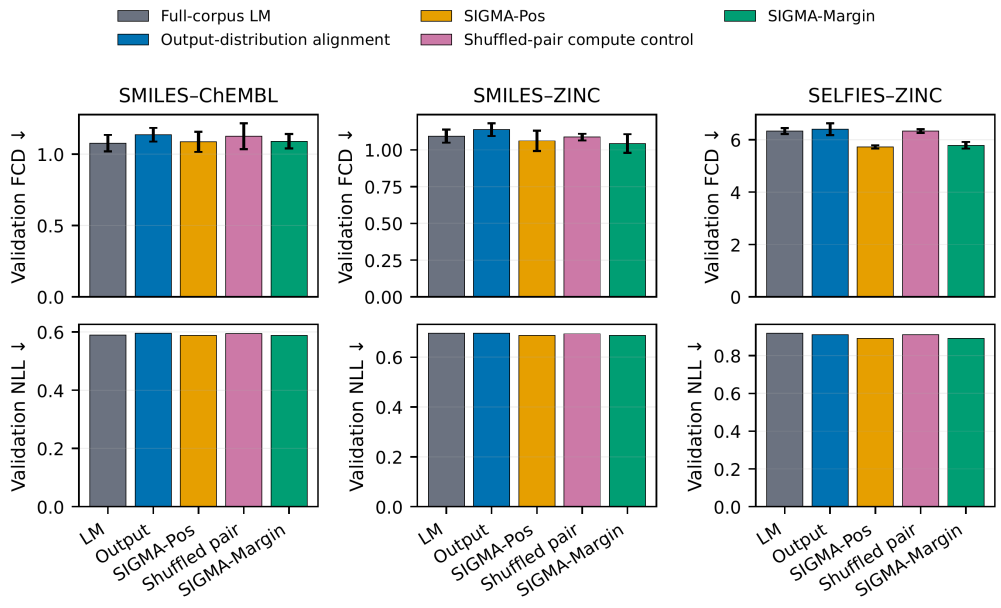


Figure A1: Separating data coverage, compute, and alignment on the full corpus. Bars are three-seed means and error bars are one standard deviation across training seeds. Correct state alignment improves FCD on both ZINC blocks, whereas the prespecified symmetric-KL variant and the shuffled-pair compute control do not. None of the auxiliary objectives exceeds the full-corpus LM on SMILES–ChEMBL.

All auxiliary objectives cost roughly 3.3–$4.5\times$ the training wall-clock of the corresponding LM because they add three Transformer forwards. The shuffled-pair compute control converts that extra computation into no gain, while the tested symmetric-KL variant often harms FCD. The comparison therefore indicates that additional computation alone does not explain the gains. Generation uses the unchanged backbone and LM head, so inference parameters, decoding operations, and proposal throughput are unchanged.

B.2 Dense Suffix-Position versus Last-Token Supervision

The last-token control directly supervises one state per certified suffix; SIGMA supervises every real suffix-token position. This full-corpus QM9 validation comparison crosses last-token versus dense suffix-position offsets with binary versus padding-safe position-matching losses. It uses the same 107,059-row full-corpus language-model stream, a separate certified-triplet auxiliary stream, identical seed-level initialization and optimizer schedule, and normalization over selected positions.

Supervised positions / loss	FCD	Validity
Last token / binary	$.4421 \pm .1033$	.7715
Every suffix position / binary	$.4335 \pm .1030$	.7882
Last token / position match, pos.	$.4200 \pm .1110$	.8052
Every suffix position / position match, pos.	$\mathbf{.3884 \pm .1191}$	**.8134**
Last token / position match, pos.+neg.	$.4710 \pm .1210$	.7302
Every suffix position / position match, pos.+neg.	$.4376 \pm .1025$	.7858

Table A12: QM9 validation-only controlled comparison of supervision density and loss formulation. FCD is mean±SD over three training seeds; lower is better. "Neg." is an always-active negative, not the final margin-gated negative.

Table A12 isolates supervision density from loss formulation. Dense suffix-position matching improves FCD over last-token matching under both the positive-only and matched positive-plus-negative comparisons. Dense suffix-position versus last-token matching improves FCD both with positive-only supervision ($\Delta = -.0317$, 95% CI $[-.0617, -.0068]$) and with the same positive-plus-negative loss ($\Delta = -.0334$, CI $[-.0567, -.0106]$). The binary contrast has the same mean direction but is inconclusive ($\Delta = -.0086$, CI $[-.0259, .0074]$). Turning on an always-active negative worsens FCD under position matching both last-token ($+.0509$) and dense suffix-position ($+.0492$) supervision. This isolates a benefit from supervising every suffix position under the tested losses and motivates the finite-margin alternative to always-active repulsion; this factorial does not itself compare negative gating rules.

C Predictive Consistency and Molecular Information

This section first reports the cross-domain mechanism and information-preservation results, then specifies the evaluation units, metrics, and probes. The controlled interventions are reported in Appendix B; pairing and metric definitions appear in Appendices C.2–C.4.

C.1 Predictive Consistency and Retained Molecular Information

Table A13 reports three-seed means and SIGMA-minus-canonical changes, with each training seed receiving equal weight. The direction of each seed-level change is checked separately. These descriptive mechanism summaries do not use the FCD bootstrap intervals; position-bin curves report means and standard deviations across dataset–seed conditions, while the controlled FCD contrasts in Section B provide the separate intervention comparisons.

Table A13 gives the complete cross-domain mechanism summary. Figure A2 resolves the same effects by normalized suffix position to test whether they persist beyond the terminal token. Across SMILES and SELFIES, semantic margin, correct-position accuracy, next-token agreement, and joint correctness rise while ordering-violation rate and suffix-token NLL fall. Every displayed direction holds in all three seeds and all four domains. The SMILES information-preservation analysis further finds increased effective rank and descriptor-probe R^2 on every domain, with Morgan fingerprint

Dataset	Method	Semantic margin	Ordering violation	Correct-position accuracy	Top-1 token agreement	Joint correctness	Suffix-token NLL
QM9	Canonical	.0511	.5405	.8875	.7234	.4900	1.0666
	SIGMA	**.1141**	**.2383**	**.9912**	**.8014**	**.5404**	**.9785**
ZINC	Canonical	.0632	.4235	.9779	.8726	.6738	.8417
	SIGMA	**.0926**	**.1902**	**.9955**	**.8970**	**.6886**	**.8155**
ChEMBL	Canonical	.0533	.4905	.9741	.8330	.6901	.7473
	SIGMA	**.1977**	**.0273**	**.9986**	**.9114**	**.7383**	**.6623**
MOSES	Canonical	.0502	.4965	.9728	.8584	.6902	.6983
	SIGMA	**.1410**	**.0880**	**.9975**	**.8961**	**.7127**	**.6660**

SELFIES change from canonical to SIGMA	QM9	ZINC	ChEMBL	MOSES
Semantic margin ↑	+.0809	+.0357	+.1229	+.0797
Ordering violation ↓	-.2540	-.2043	-.4341	-.4324
Position accuracy ↑	+.2446	+.0541	+.0665	+.0652
Top-1 token agreement ↑	+.0742	+.0320	+.0536	+.0402
Joint correctness ↑	+.0571	+.0261	+.0393	+.0313
Suffix-token NLL ↓	-.1355	-.0615	-.0795	-.0630

Table A13: Mechanism diagnostics on certified validation triplets, averaged over three training seeds. The upper block reports canonical and SIGMA values for SMILES; the lower block reports the corresponding SELFIES change from canonical training to SIGMA. Every emphasized SMILES value and every SELFIES change moves in the desired direction in all three seeds.



Figure A2: Complete shared-suffix mechanism profiles for the final SMILES model. Grey is canonical training and red is SIGMA. Each point averages the same normalized suffix bin over four datasets and three training seeds; ribbons show one standard deviation across the 12 dataset–seed conditions. Correct-position accuracy and next-token agreement are higher-is-better, while ordering-violation rate and suffix-token NLL are lower-is-better. All four effects persist across the continuation rather than appearing only at its final token.

macro-AUROC changes of only +.0025, -.0009, +.0001, and +.0038. These results reject a rank-one or chemistry-erasing collapse explanation. Extended suffix-bin profiles are descriptive; the preceding controlled ablation separately isolates position density from loss formulation.

Dataset	Training	Effective-rank fraction ↑	Descriptor R^2 ↑	Morgan AUROC ↑
QM9	Canonical	$.0500 \pm .0031$	$.7728 \pm .0059$	$.6988 \pm .0026$
	SIGMA	$\mathbf{.0571 \pm .0020}$	$\mathbf{.7846 \pm .0061}$	$\mathbf{.7013 \pm .0017}$
ZINC	Canonical	$.0770 \pm .0004$	$.6761 \pm .0095$	$\mathbf{.6773 \pm .0037}$
	SIGMA	$\mathbf{.0893 \pm .0032}$	$\mathbf{.6883 \pm .0075}$	$.6764 \pm .0045$
ChEMBL	Canonical	$.1732 \pm .0014$	$.7389 \pm .0170$	$.6743 \pm .0013$
	SIGMA	$\mathbf{.1846 \pm .0035}$	$\mathbf{.7500 \pm .0039}$	$\mathbf{.6743 \pm .0014}$
MOSES	Canonical	$.0894 \pm .0032$	$.6730 \pm .0078$	$.7028 \pm .0027$
	SIGMA	$\mathbf{.0977 \pm .0021}$	$\mathbf{.7007 \pm .0122}$	$\mathbf{.7066 \pm .0025}$

Table A14: Chemical-information preservation in SMILES hidden states (mean ± standard deviation across three training seeds). Effective rank is normalized by hidden dimension; descriptor R^2 averages the molecular property probes defined in Section C; Morgan AUROC averages the fingerprint-bit probes. Bold marks the higher mean within each dataset.
Figure A2 shows the position-resolved mechanism; Appendix I.2 gives the ChEMBL boundary diagnostics.

Table A14 gives the complete four-domain information-preservation values summarized in the main paper. SIGMA increases covariance effective rank and descriptor-probe R^2 in every domain, while Morgan-fingerprint macro-AUROC changes by at most .0038. The improvement in suffix-state correspondence therefore coexists with, rather than replaces, recoverable molecular information.

Table A15 closes the complementary scaffold-conditioned check cited in the main paper. Equivalent-view similarity rises while the same-molecule $>$ same-scaffold/different-molecule $>$ different-scaffold ordering holds in every dataset and every seed. The nearest-neighbor scaffold probe is mixed rather than uniformly improved, so the evidence supports preserved between-molecule separation, not a universal improvement of every representation probe.

Dataset	Training	Equivalent views	Same scaffold, different molecule	Different scaffold	Scaffold hit@10
QM9	Canonical	$.9240 \pm .0088$	$.7794 \pm .0161$	$.7134 \pm .0148$	$.4556 \pm .0192$
	SIGMA	$\mathbf{.9559 \pm .0050}$	$.7634 \pm .0085$	$.7033 \pm .0087$	$.4715 \pm .0079$
ZINC	Canonical	$.9800 \pm .0021$	$.7160 \pm .0147$	$.6615 \pm .0162$	$.5032 \pm .0072$
	SIGMA	$\mathbf{.9859 \pm .0002}$	$.6773 \pm .0151$	$.6243 \pm .0131$	$.5057 \pm .0034$
ChEMBL	Canonical	$.9493 \pm .0018$	$.6605 \pm .0190$	$.5529 \pm .0149$	$.5561 \pm .0505$
	SIGMA	$\mathbf{.9762 \pm .0005}$	$.5863 \pm .0133$	$.4756 \pm .0112$	$.5378 \pm .0401$
MOSES-300k	Canonical	$.9705 \pm .0017$	$.7058 \pm .0183$	$.6276 \pm .0146$	$.5000 \pm .0045$
	SIGMA	$\mathbf{.9806 \pm .0013}$	$.6723 \pm .0135$	$.6009 \pm .0050$	$.4770 \pm .0192$

Table A15: SMILES scaffold-conditioned cosine structure and scaffold retrieval (mean ± training-seed SD). In every dataset and each of the three seeds, equivalent views are more similar than different molecules sharing a scaffold, which in turn are more similar than molecules with different scaffolds. Thus the stronger equivalent-view correspondence does not erase between-molecule separation. Scaffold hit@10 is a descriptive nearest-neighbor guardrail and is not uniformly improved.

C.2 Evaluation Units and Pairing

The final cross-domain mechanism summaries compare canonical training with SIGMA on validation triplets, using at most 5,000 usable rows per checkpoint and the recorded deterministic sampling seed. SMILES uses the same row selection for each canonical–SIGMA pair within a dataset and training seed. The SELFIES runs use cell-specific sampling seeds, so their aggregate changes are matched by training seed rather than asserted to be row-paired. Metrics are averaged over offsets within each triplet, then over triplets for each molecular identity, and finally over identities and training seeds. These are checkpoint diagnostics; no retraining or free generation is required.

C.3 METRICS ALONG THE SHARED SUFFIX

For every selected suffix offset j, define cosine scores $c_j^+ = \mathrm{sim}(h_j^A, h_j^B)$ and $c_j^- = \mathrm{sim}(h_j^A, h_j^C)$. We report:

1. **correct-position accuracy**: whether h_j^B is the highest-score real position in the positive sequence;
2. **aligned-position mean reciprocal rank (MRR)**: reciprocal rank of the aligned h_j^B, averaged over evaluated suffix positions;
3. **aligned-negative position probability**: $p^-_{j,r_C(j)}$ from Section E;
4. **ordering-violation rate**: fraction with $c_j^+ \le c_j^-$;
5. **semantic margin**: $c_j^+ - c_j^-$;
6. **next-token Jensen–Shannon divergence**: JSD between the complete vocabulary distributions predicted from h_j^A and h_j^B;
7. **target-token probability gap**: absolute difference in probability assigned to the known suffix token s_j; and
8. **suffix-token NLL**: token NLL on s_j, reported separately for the two equivalent serializations and as their paired difference.

Top-1 next-token agreement is the average of $\mathbf{1}[\arg\max_u \pi_j^A(u) = \arg\max_u \pi_j^B(u)]$ under the aggregation above, not a transformation of JSD. "Both top-1 predictions correct" also requires that common prediction to equal s_j. Negative next-token divergence is not treated as a mandatory objective: $C \oplus S$ is a valid sequence whose next literal token is also s_j, so hidden-state separation and identical-token prediction can coexist.

Table A16 states the concrete question, direction, and attainable endpoint for each mechanism and information-preservation metric, preventing unlike scales from being interpreted as interchangeable. Metrics are reported overall and in ten normalized suffix-position bins. The curve therefore tests whether an effect appears only at the terminal token or persists throughout the continuation. Per-triplet aggregate rows are retained; raw position records are reduced only after the loss-mask, row-key, and denominator checks pass.

C.4 INFORMATION-PRESERVATION PROBES

Representation collapse—mapping different molecules to nearly the same hidden vector—could increase positive cosine without learning useful chemistry. We therefore evaluate final pre-EOS validation embeddings in three strata:

1. equivalent views of the same molecule;
2. different molecules with the same nonempty Bemis–Murcko scaffold; and
3. molecules with different nonempty scaffolds.

Pairs are deterministic, molecule-disjoint where required, and shared across methods. Empty scaffolds are excluded from scaffold comparisons.

Let $\lambda_1, \ldots, \lambda_d$ be eigenvalues of the centered embedding covariance and $q_i = \lambda_i / \sum_k \lambda_k$. The effective rank is

$$r_{\text{eff}} = \exp\left(-\sum_i q_i \log q_i \right).$$

We also report covariance condition number and mean pairwise cosine. Chemical information is probed with a deterministic molecule split and ridge regression for molecular weight, logP, TPSA, H-bond donors/acceptors, ring count, and fraction sp^3. A fixed-k nearest-neighbor scaffold-retrieval score tests nonlinear chemical information. SIGMA is called selective only if same-state alignment improves without broad same-scaffold/different-scaffold contraction, catastrophic effective-rank loss, or material degradation of descriptor/scaffold probes.

Metric	Concrete question	Best / full score	Worst / zero case
Correct-position accuracy	Does the A state that predicts s_j identify the matching B offset ahead of every other real B position?	1: every offset has rank 1.	0: no offset has rank 1.
Aligned-position MRR	What is the rank of the correct B offset?	1: rank 1 everywhere.	Raw MRR is never exactly 0; its per-offset minimum is $1/\lvert\mathcal{V}_B\rvert$.
Ordering-violation rate	Is the certified wrong-state offset at least as similar as the correct positive ($c_j^- \geq c_j^+$)?	0: no offset violates the ordering.	1: every offset violates it.
Semantic margin	How much larger is c_j^+ than c_j^-?	Larger is better; cosine bounds the maximum at $+2$.	Minimum -2; 0 is the decision boundary, not a universal zero score.
Next-token JSD	Do A and B predict the same next-token distribution?	0: distributions are identical.	With the implemented natural logs, $\log 2$: disjoint supports.
Target-token probability gap	Do A and B give the known s_j the same probability?	0: probabilities are equal.	1: one assigns 1 and the other 0.
Suffix-token NLL	How probable is the correct s_j?	0: target probability is 1.	No finite maximum; it diverges as target probability approaches 0.
Effective rank	How many covariance directions carry representation variance?	$\min(n-1, d)$ for equal nonzero eigenvalues.	1 under rank-one collapse; raw effective rank is not 0.
Scaffold-stratum ordering	Are same-molecule views closer than different molecules with the same scaffold, which are closer than different-scaffold molecules?	Ordering accuracy 1: every sampled comparison is correctly ordered.	0: none is correctly ordered. All distances being 0 is collapse, not success.
Property probe R^2	Can an embedding linearly recover molecular weight, logP, topological polar surface area (TPSA), etc. on held-out molecules?	1: perfect held-out prediction.	0: equal to the train-mean predictor; R^2 can be negative.
Structure probe	Do fixed-k neighbors preserve held-out scaffold labels?	Accuracy 1: all labels are correct.	Accuracy 0: none is correct; chance level is reported separately.

Table A16: Definitions, direction, and interpretation of the mechanism and information-preservation metrics. Some metrics do not have a finite worst value; those cases are stated explicitly rather than being forced onto a zero-to-one scale.

D TRIPLET CONSTRUCTION, CERTIFICATION, AND TOKENIZATION

D.1 TRIPLET CONSTRUCTION ALGORITHM

Algorithm A1: Chemical certification of same-suffix triplets

1. Read candidate (A, B, S); require a nonempty suffix, distinct prefixes, lossless tokenization, and valid vocabulary and context lengths.
2. Reconstruct $A \oplus S$ and $B \oplus S$; require the stored source identity at completion, matched parseability at each suffix offset, and equal identities whenever both partial serializations decode.
3. Search a deterministic, length-matched pool for prefix C from a different source identity.
4. Append the *same literal* S to C; require matched parseability, unequal identities at every jointly decodable offset, and unequal final identities.
5. Emit the triplet, suffix spans, parseability masks, identities, and reconstruction hashes if every verification condition holds; otherwise discard it.

Algorithm A1 makes the same-suffix construction and its rejection checks explicit. It consumes training or validation pair files. For each candidate equivalent pair it records A, B, and the exact shared suffix S, reconstructs both complete strings, and verifies the expected source identity. Candidate non-equivalent histories are searched in length buckets. If n_A is the anchor-prefix token length, the eligible lengths for C satisfy

$$|n_C - n_A| \leq \max(3, \mathrm{round}(0.25 n_A)).$$

Within each bucket, a deterministic hash-seeded start varies where the search begins. Candidates are visited with unit stride until a valid negative is found or the search limit is reached. The source row itself and candidates with the same final identity are skipped. A row is emitted only after the non-equivalence verification succeeds.

Molecular identity rule. Identity is the complete RDKit Standard InChIKey after parsing the represented molecule, emitting canonical isomeric SMILES, and reparsing that string. No application-level largest-fragment selection, salt removal, neutralization, or tautomer enumeration is performed. Standard InChI itself, however, normalizes mobile hydrogens, variable protonation, and charge and omits the Fixed-H layer. Canonical SMILES and a SHA256 graph diagnostic are stored but are not conjoined with the identity label: they diagnose a stricter tautomer/charge-specific relation than the one used in the reported certificate.

Boundary sanity check. With the RDKit identity call used in our implementation, `O=c1cccc[nH]1` (2-pyridone) and `Oc1ccccn1` (2-hydroxypyridine) retain distinct canonical isomeric SMILES but share the Standard InChIKey `UBQKCCHYAOITMY-UHFFFAOYSA-N`; non-standard InChI with `/FixedH` distinguishes them. Neutral and zwitterionic drawings can likewise share a Standard InChIKey. The accompanying code includes this tautomer pair as a unit test so that the scope of the identity contract is executable rather than implicit.

D.2 VERIFYING THE EQUIVALENT PAIR

For every retained (A, B, S):

1. tokenization is lossless and $A \neq B$ as token sequences;
2. $A \oplus S$ and $B \oplus S$ reconstruct exactly the stored full strings;
3. the two serializations have the same parseability mask at every suffix offset;
4. whenever both serializations are parseable, their complete InChIKeys are equal;
5. both full sequences decode, agree with one another, and agree with the expected source identity;
6. the suffix contains real representation tokens and the selected span excludes EOS.

SELFIES checkpoints must decode at every offset. SMILES additionally requires the two prefixes themselves to be parseable and chemically equivalent. SAFE requires equal fragment multisets in the two equivalent histories and a dot boundary at the beginning of the shared continuation; its partial fragment-assembly sequence is hashed independently.

D.3 Verifying the Non-Equivalent History

For every retained (A, C, S):

1. $C \neq A$, while the suffix token list is exactly the same S used by the positive pair;
2. the A and C prefix states are non-equivalent whenever the representation admits a decoded prefix state;
3. the anchor and negative serializations have the same parseability mask;
4. at every jointly parseable offset, the InChIKeys remain different;
5. both full strings decode and their final InChIKeys differ;
6. for SMILES and SAFE, a nonzero elemental-composition witness at the prefix is preserved after appending S.

The last condition prevents C from becoming chemically equivalent through a later same-suffix cancellation of elemental composition. The construction code stores the parseability masks, final identities, composition delta, and number of search attempts for C.

D.4 Independent Verification of the Declared Identity Relation

For QM9, certification is checked three times: during construction, by an exhaustive primary checker, and by an independently implemented reconstruction/cross-check. The independent check re-tokenizes all four fields, reconstructs $A \oplus S$, $B \oplus S$, and $C \oplus S$, and recomputes intermediate and final identities without reusing the construction code's Boolean verification columns. Other domains use the same construction checks; the zero negative-identity-collision result below refers specifically to QM9. Because both implementations intentionally recompute the same Standard InChIKey relation, implementation independence detects reconstruction and bookkeeping errors but does not turn that relation into graph-exact identity.

Rep.	Split	Source	Positive	Triplets	Neg. key collision
SMILES	train	107108	80935	80932	0
SMILES	valid	13388	10061	10061	0
SELFIES	train	45936	45936	45936	0
SELFIES	valid	5730	5730	5730	0
SAFE	train	11160	11160	11160	0
SAFE	valid	1312	1312	1312	0

Table A17: Independent QM9 verification of the declared Standard-InChIKey relation. Counts exclude the CSV header. "Neg. key collision" is the fraction of retained negative rows for which independent recomputation found equality under that relation.

Positive retention was 100% in every verified representation–split condition. Three SMILES training positives lacked an accepted negative within the length-matched search, accounting for 80,935 certified positives but 80,932 triplets.

D.5 Representation-Specific Tokenization

SMILES. The atomwise pre-tokenizer preserves bracket atoms, bond symbols, branch parentheses, and ring-closure identifiers as explicit tokens. No chemical token category is removed from dense supervision.

SELFIES. The official `split_selfies` semantics are used: each bracketed SELFIES symbol is one token. Construction and training both use that tokenizer, so the stored suffix-token count must match the runtime split exactly.

SAFE. SAFE (Noutahi et al., 2024) is tokenized into basic/atomic tokens. Dots, attachment identifiers, and connection syntax remain explicit and supervised. Treating an entire SAFE fragment as a single token would remove the token-level correspondence needed by SIGMA and is therefore not used. The four-domain data pass the same atomic-token, strict-same-suffix, and independent Standard-InChIKey checks used for the other representations. ChEMBL is the only domain in which the search could not find a certified same-suffix negative for every positive; those 659 rows were rejected rather than relabeled or weakened. All retained negatives have zero Standard-InChIKey collisions under independent reconstruction. The construction code retains atom, bond/modifier, branch, connection/ring-label, and fragment separator tokens.

Split statistic	Split	QM9	ZINC	ChEMBL	MOSES-300k
Retained triplets	train	11,160	201,399	1,139,872	290,558
Retained triplets	valid	1,312	22,415	71,193	29,057
Rejected: no certified negative	train	0	0	622	0
Rejected: no certified negative	valid	0	0	37	0

Table A18: Certified atomic-token SAFE triplets. "Unmatched" counts positive rows for which no length-matched non-equivalent prefix sharing the exact suffix was found.

Table A17 reports the independent QM9 audit. No retained non-equivalent history collides with its anchor under the declared Standard-InChIKey relation in any representation or split. Table A18 then reports the certified SAFE counts and the 659 ChEMBL rows rejected because no valid non-equivalent history was found.

D.6 Certified-Triplet Counts Used by the Controlled SMILES/SELFIES Comparison

Representation	Split	QM9	ZINC	ChEMBL	MOSES-300k
SMILES	train	80932	138750	814718	194012
SMILES	valid	10061	15413	51002	19438
SELFIES	train	45936	111041	653322	163118
SELFIES	valid	5730	12372	40863	16331

Table A19: Number of certified triplets available to the controlled common-backbone experiment. These are training examples, not counts of unique molecules.

Table A19 gives the resulting auxiliary-example counts for the controlled SMILES and SELFIES experiments. The counts vary by representation and domain, which is why comparisons are made within, rather than across, representation–dataset blocks.

E Formal Objective and Consequences

E.1 When Two Serialization Histories Are Chemically Equivalent

Let A, B, and C be tokenized prefixes and let $S = (s_1, \ldots, s_m)$ be a nonempty suffix whose token sequence is identical in all three examples. Molecular identity is represented by the complete Standard InChIKey $\mathrm{id}_{\mathrm{std}}(\cdot)$ obtained after RDKit parsing, canonical isomeric-SMILES emission, and reparsing. This operational contract does not apply an application-level salt stripper, neutralizer, parent-fragment selector, or tautomer enumerator. It nevertheless inherits Standard InChI normalization: mobile-H tautomers and some protonation/charge arrangements can share the same identifier because the Fixed-H layer is absent (Heller et al., 2015; InChI Trust, 2012). It is therefore a declared chemical equivalence relation, not atom-, bond-order-, charge-, and tautomer-specific graph equality. For this fixed continuation, define

$$A \equiv_S B \quad \Longleftrightarrow \quad \mathrm{id}_{\mathrm{std}}(A \oplus S) = \mathrm{id}_{\mathrm{std}}(B \oplus S).$$

The retained triplet satisfies

$$A \equiv_S B, \qquad C \not\equiv_S A, \tag{A1}$$

together with the position-wise checks in Section D. Canonical SMILES and graph hashes are retained as diagnostics but are not conjoined with the identity definition; doing so would define a stricter relation than the Standard-InChIKey contract used to construct and train the reported models.

For an autoregressive model, let $\pi_j^X = p_\theta(\cdot \mid X \oplus s_{<j})$. The decision inconsistency of the equivalent pair under the common continuation is

$$\mathcal{D}_\theta(A, B; S) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{JS}(\pi_j^A, \pi_j^B). \tag{A2}$$

SIGMA does not minimize this quantity directly. It constrains the hidden states that parameterize these conditionals and treats the resulting change in Eq. (A2) as a falsifiable mechanism prediction. Identity checks along S do not establish equality of the full next-token distributions over alternative continuations; state alignment is the inductive bias tested by the mechanism and generation evaluations.

E.2 Suffix-Relative Pre-Token Hidden States

Let g_t^X denote the final-layer hidden state at absolute pre-token position t in the full sequence $X \oplus S$, and let H_θ map the observed token history to that hidden vector. If the first suffix token of X occurs at target index b_X, then the state that predicts suffix token s_j is

$$r_X(j) = b_X + j - 2, \qquad h_j^X = g_{r_X(j)}^X = H_\theta(X \oplus s_{<j}). \tag{A3}$$

The subtraction by one converts a token target index into its causal pre-token state. Because b_A, b_B, and b_C may differ, equality of absolute indices does not imply equality of suffix events. Equation (A3) is therefore the unique monotone alignment that preserves both suffix offset and causal timing.

Only real suffix tokens are included. Padding and `EOS` positions are excluded, and the auxiliary mask is applied before any softmax normalization. This makes the objective invariant to the amount of right padding introduced by other examples in the minibatch.

E.3 Padding-Safe Suffix-Position Matching

Let $\mathcal{V}_B$ be the real, non-padding predictive positions in the positive sequence, including all real states from `GO` through `EOS`, and let $\mathrm{sim}(u, v)$ be cosine similarity. Only right padding is excluded from the candidate set; supervised targets remain pre-token states of real suffix symbols and never target `EOS`. The position-matching distribution from anchor state h_j^A over the positive serialization is

$$q_\tau(k \mid h_j^A; B) = \frac{\exp(\mathrm{sim}(h_j^A, g_k^B)/\tau)}{\sum_{\ell \in \mathcal{V}_B} \exp(\mathrm{sim}(h_j^A, g_\ell^B)/\tau)}, \qquad k \in \mathcal{V}_B. \tag{A4}$$

The target is not merely the positive sequence but the unique position $r_B(j)$ corresponding to the same suffix-relative prediction event:

$$\ell_j^+ = -\log q_\tau(r_B(j) \mid h_j^A; B), \qquad \mathcal{L}_+ = \frac{1}{M} \sum_{(i,j) \in \mathcal{I}} \ell_{i,j}^+. \tag{A5}$$

Here $\mathcal{I}$ contains every valid example–offset pair in the batch and $M = |\mathcal{I}|$. Replacing padded logits by $-\infty$ before Eq. (A4) guarantees that padding contributes exactly zero probability mass. The target is unique as a token position, not necessarily as a chemical state: other real positions act as retrieval competitors without individual chemical non-equivalence certificates.

The implemented position matching is directed: the designated anchor-history state h_j^A queries the alternate serialization B. Chemical equivalence is symmetric, but the optimized matching term is not averaged with a reverse $B \to A$ term. The constructor fixes these roles before training, and all methods reuse the same orientation. Because $|\mathcal{V}_B|$ varies with sequence length, raw position-matching losses are not treated as length-invariant cross-dataset metrics.

E.4 Relative Margin-Gated Negative

For the aligned positive and certified non-equivalent states, define

$$s_j^+ = \mathrm{sim}(h_j^A, h_j^B), \qquad s_j^- = \mathrm{sim}(h_j^A, h_j^C).$$

The final negative objective is

$$\ell_j^{\mathrm{gate}} = \max(0, s_j^- - s_j^+ + \gamma), \tag{A6}$$

$$\mathcal{L}_{\mathrm{gate}} = \frac{1}{M} \sum_{(i,j)\in\mathcal{I}} \ell_{i,j}^{\mathrm{gate}}. \tag{A7}$$

Here $\gamma > 0$ is the cosine margin, distinct from suffix length m. Its subgradient with respect to the two similarities is

$$\left(\frac{\partial \ell_j^{\mathrm{gate}}}{\partial s_j^+}, \frac{\partial \ell_j^{\mathrm{gate}}}{\partial s_j^-} \right) = \begin{cases} (-1, +1), & s_j^+ - s_j^- < \gamma, \\ (0, 0), & s_j^+ - s_j^- > \gamma, \end{cases}$$

with the usual set-valued subgradient at equality. Thus the negative is active precisely when the desired ordering is absent or insufficient, and it stops modifying an already separated pair.

The language-model component is the mean causal NLL over all non-padding targets in the optimizer batch,

$$\mathcal{L}_{\mathrm{LM}} = \frac{\sum_i \sum_{t=1}^{T_i} -\log p_\theta(x_{i,t} \mid x_{i,<t})}{\sum_i T_i}.$$

The complete final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{aux}} \left(\mathcal{L}_+ + \rho_{\mathrm{neg}} \mathcal{L}_{\mathrm{gate}} \right). \tag{A8}$$

The negative enters the total objective with effective coefficient $\lambda_{\mathrm{neg}}^{\mathrm{eff}} = \lambda_{\mathrm{aux}} \rho_{\mathrm{neg}}$. The validation-selected SMILES/SELFIES configuration is $\lambda_{\mathrm{aux}} = .075$, $\rho_{\mathrm{neg}} = 1$, $\gamma = .10$, and $\tau = .07$; hence both auxiliary terms have effective coefficient $.075$. This common-weight rule is shared across all primary datasets.

E.5 Four Objective Properties

These propositions establish that the objective supervises the intended pre-token prediction events, explain the finite-margin design, connect dense suffix-position alignment to a sequence-level quantity, and state the scope beyond which empirical evaluation is required.

Proposition A1 (Pre-token indexing) *Under a left-to-right token factorization with one pre-token hidden state per target token, Eq.* (A3) *is the unique mapping that pairs states which have observed $X \oplus s_{<j}$ and are about to predict s_j. Moreover, extending B with any number of arbitrary right-padding vectors leaves q_τ on $\mathcal{V}_B$, $\mathcal{L}_+$, and every gradient with respect to real states unchanged when padded logits are replaced by $-\infty$ before Eq.* (A4)*; padding-vector gradients are zero.*

Proof. The target index of s_j in sequence X is $b_X + j - 1$. Causality associates that target with the state immediately before it, at index $b_X + j - 2$. Hence $r_X(j) = b_X + j - 2$. Any smaller index has not observed all of $s_{<j}$; any larger index has already consumed s_j or a later event. Therefore no other state has both the required history and prediction target. For padding, every padded exponential is $\exp(-\infty) = 0$, so neither the numerator of a real position nor the denominator over real positions changes. The loss is algebraically independent of padded vectors, preserving all real-state gradients and giving zero padding gradient. Masking after the softmax would not have this property because padded logits would already have altered the denominator. □

Design consequence. The first clause rules out absolute-position matching when equivalent prefixes have unequal token lengths; the second rules out pairing with, or normalizing against, padding. We group them because both establish that the loss is defined on the intended next-token event rather than on batch layout.

Proposition A2 (Finite separation) *Let $r = r_B(j)$, $N = |\mathcal{V}_B| - 1$, and $c_k = \mathrm{sim}(h_j^A, g_k^B)$. If $q_\tau(r \mid h_j^A; B) \geq 1 - \varepsilon$ for $0 < \varepsilon < 1/2$, then for every $k \neq r$,*

$$c_r - c_k \geq \tau \log \frac{1-\varepsilon}{\varepsilon}.$$

Conversely, if $c_r - c_k \geq \Delta$ for every $k \neq r$, then

$$q_\tau(r \mid h_j^A; B) \geq \frac{1}{1 + N \exp(-\Delta/\tau)}.$$

In addition, $\ell_j^{\mathrm{gate}} = 0$ if and only if $s_j^+ - s_j^- \geq \gamma$. Hence a representation in which aligned positive and negative states all collapse to the same direction cannot attain zero gated loss for $\gamma > 0$.

Proof. Dividing the softmax denominator by $\exp(c_r/\tau)$ gives

$$\frac{1 - q_\tau(r)}{q_\tau(r)} = \sum_{k \neq r} \exp((c_k - c_r)/\tau).$$

Under $q_\tau(r) \geq 1 - \varepsilon$, the sum is at most $\varepsilon/(1-\varepsilon)$, so each nonnegative summand obeys the same upper bound. Taking logarithms yields the first inequality. If every score gap is at least Δ, each summand is at most $\exp(-\Delta/\tau)$; summing the N terms and inverting proves the converse. The statement for the gated term follows immediately from the zero set of the hinge. Under directional collapse, $s_j^+ = s_j^-$, so the loss equals $\gamma > 0$. □

Design consequence. The position-matching term makes the designated suffix-relative position distinguishable from competing positions in the positive serialization; this is positional discrimination, not a proof of unique chemical identity. The gated negative adds only the finite chemical ordering not supplied by position matching and then switches off; it does not demand indefinite repulsion of molecules that may share ordinary chemical features.

Proposition A3 (Stability of suffix-token likelihood) *Let the shared language-model head have logits $z(h) = Wh + b$, with row w_u for token u, and let $M_W = \max_u \|w_u\|_2$. For two histories A, B evaluated under the same teacher-forced continuation $S = (s_1, \ldots, s_m)$,*

$$|\log p_\theta(S \mid A) - \log p_\theta(S \mid B)| \leq 2M_W \sum_{j=1}^{m} \|h_j^A - h_j^B\|_2. \tag{A9}$$

If, at offset j, $\|h_j^A\|_2, \|h_j^B\|_2 \leq R_j$, $|\|h_j^A\|_2 - \|h_j^B\|_2| \leq \eta_j$, and $\mathrm{sim}(h_j^A, h_j^B) \geq 1 - \epsilon_j$, then its summand is at most $\sqrt{\eta_j^2 + 2R_j^2\epsilon_j}$.

Proof. For $f_y(h) = \log p_\theta(y \mid h)$,

$$\nabla_h f_y(h) = w_y - \sum_v p_\theta(v \mid h) w_v.$$

The gradient norm is at most $\|w_y\|_2 + \sum_v p_\theta(v \mid h)\|w_v\|_2 \leq 2M_W$. The mean-value inequality therefore gives $|f_y(h) - f_y(h')| \leq 2M_W \|h - h'\|_2$. Applying this inequality to the common target s_j at every suffix offset and using the triangle inequality proves Eq. (A9). Finally, writing $r_j = \|h_j^A\|_2$, $r'_j = \|h_j^B\|_2$, and $c_j = \mathrm{sim}(h_j^A, h_j^B)$ gives

$$\|h_j^A - h_j^B\|_2^2 = (r_j - r'_j)^2 + 2r_j r'_j(1 - c_j) \leq \eta_j^2 + 2R_j^2\epsilon_j.$$

□

Design consequence. Every suffix offset contributes one distance term to the likelihood bound. A final-position-only contrast directly supervises only its corresponding state. Earlier states can change through shared parameters, but their distances are not bounded by the endpoint constraint alone. The likelihood bound holds under the stated affine-head assumption, but cosine matching alone does not control radial differences or guarantee a small upper bound. This is why the mechanism evaluation measures next-token JSD, suffix-token NLL, state norms, effective rank, and collapse rather than treating hidden-state similarity as the endpoint.

Proposition A4 (Local–global gap) *Let $\mathcal{C}$ be the set of histories constrained by certified triplets. If a history $h^\star \notin \mathcal{C}$ and its outgoing transition remain unconstrained and are reachable with probability $r > 0$, agreement on every conditional in $\mathcal{C}$ does not identify the joint sequence distribution. In particular, two models can agree before $h^\star$ and on all of $\mathcal{C}$, while conditionals at $h^\star$ differ by total variation δ, inducing joint-sequence variation $r\delta$.*

Proof. Construct two autoregressive models with identical transitions until $h^\star$, so both reach it with probability r, and identical transitions on every constrained history. At $h^\star$, choose next-token distributions u and v with total variation distance $\mathrm{TV}(u, v) = \delta$, then terminate each resulting branch under identical continuation rules. Token sequences from different outgoing symbols are disjoint events. Their contribution to joint total variation is therefore

$$\frac{1}{2} \sum_x r|u(x) - v(x)| = r\delta.$$

Thus all certified local constraints can agree while the complete generation laws differ. Because molecular serialization is additionally many-to-one, the proposition neither predicts the sign nor the magnitude of a molecular distribution metric such as FCD. □

Design consequence. The first three propositions justify the event alignment, finite negative, and dense suffix-position supervision used by SIGMA. This fourth proposition marks their limit: local conditional improvement does not mathematically determine a free-generation metric. FCD is therefore evaluated as an independent behavioral endpoint rather than presented as a corollary.

Numerical checks. Numerical tests cover pre-token alignment, padding invariance, the position-matching inequality, finite-margin behavior, and the $r\delta$ local–global construction. Continuation-likelihood stability follows analytically from the shared affine LM head.

E.6 Controlled Comparisons and Negative-Loss Ablation

We use **SIGMA-Pos** for the certified positive-correspondence objective with $\rho_{\mathrm{neg}} = 0$ and **SIGMA-Margin** for its finite-margin extension. Unless stated otherwise, SIGMA denotes SIGMA-Margin. Both are members of the certified hidden-state SIGMA family; the ablation tests whether the negative adds value beyond their shared correspondence signal.

Within each representation–dataset block, the four methods use identical language-model row coverage, initialization, architecture, and optimization:

1. **Canonical training**: only $\mathcal{L}_{\mathrm{LM}}$ on the canonical representation.
2. **Randomized-serialization training**: only $\mathcal{L}_{\mathrm{LM}}$, with the paired randomized valid serialization as the likelihood input.
3. **Last-token alignment**: a binary positive-versus-negative contrast at the final real suffix position.
4. **SIGMA-Margin**: Eqs. (A5) and (A7) at every real suffix offset.

In the certified-subset regime, all four use the same certifiable molecule rows. In the full-corpus regime, all four use the same complete-molecule stream and auxiliary methods alone consume a separately cycled certified stream. Every block uses the same-size GPT-2 backbone. Backbone-state alignment acts on unprojected final-layer states; the language-model head and decoder remain unchanged.

An ablation uses an always-active position anti-likelihood $-\log(1 - p^-)$ as an always-active negative. That objective is retained only as an ablation because it continues to repel a negative after the relative ordering is already correct.

E.7 Worked Triplet Example

A QM9 SMILES row is

$$A = \texttt{CCO}, \qquad B = \texttt{C(C)O}, \qquad C = \texttt{CCC}, \qquad S = \texttt{CC}.$$

Their complete Standard InChIKeys are:

- Positive completions `CCOCC` and `C(C)OCC`: `RTZKZFJDLAIYFH-UHFFFAOYSA-N`.
- Negative completion `CCCCC`: `OFBQJSOFQDEBGM-UHFFFAOYSA-N`.

All three histories face the literal continuation C, C, while only (A, B) belongs to the same conditional equivalence class.

With zero-based indices and an explicit `GO` token, the first suffix C has target/state indices $4/3$ in A, $6/5$ in B, and $4/3$ in C. SIGMA therefore compares states $(3, 5, 3)$. Pairing absolute index 3 across all strings would instead compare the state after `O` in A with the state after the branch atom `C` in B, which are different predictive events. The same offset construction prevents a real suffix state from being paired with padding in a mixed-length batch.

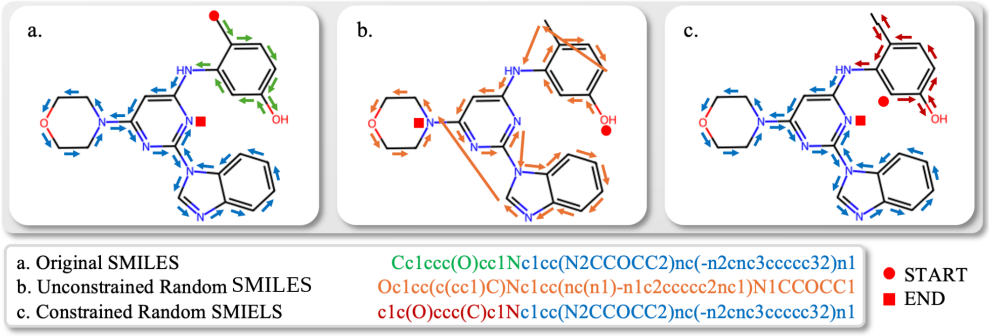


Figure A3: Illustration of constrained same-suffix randomization. (a) The anchor serialization is divided at a chemically valid cut point into a prefix and a continuation. (b) Unconstrained SMILES randomization changes the traversal of the entire molecular graph and therefore does not preserve token-level correspondence with the anchor. (c) Constrained randomization produces a distinct prefix that reaches the same intermediate molecular state, after which it reuses the anchor's exact continuation. Green and red denote the two alternative prefix traversals, while blue denotes the shared suffix with identical tokens in identical order. Circles and squares mark the start and end of each prefix traversal; arrows indicate serialization order rather than a chemical reaction or electron movement. The resulting anchor–equivalent pair can therefore be aligned by suffix-relative offset even when its prefixes have different token lengths. Length-controlled valid non-equivalent histories used as loss negatives are constructed separately and verified by molecular identity.

Figure A3 visualizes the key distinction: ordinary randomization can change the entire traversal, whereas the certified construction changes the prefix while preserving the literal suffix and its suffix-relative prediction events.

F Implementation, Data Splits, and Evaluation

F.1 Shared Backbone and Optimization

Table A20 lists the common architecture and optimization contract used by the controlled SMILES/SELFIES comparisons. The same GPT-2 class and architecture are used for all representations; only the representation vocabulary changes. There is no representation-specific model branch.

Micro-batches are an execution-only split: token-count and aligned-position weights preserve the normalization of one optimizer-batch objective; SMILES margin-gated runs fit the optimizer batch without this split. Data-loader worker count changes input prefetching only. Each dataset–seed block uses identical initial parameters across methods.

Item	Value
Backbone	`GPT2LMHeadModel`
Layers / attention heads	6 / 8
Hidden size / context	256 / 512
Attention / embedding / residual dropout	0.1 / 0.1 / 0.1
Epochs	3
Optimizer	AdamW
Learning rate / weight decay	4×10^{-4} / 0.01
Schedule	5% linear warm-up, cosine decay
Optimizer batch / micro-batch	128 / 32 where used; otherwise unsplit
Maximum batch tokens	8192
Gradient clipping	1.0
Data-loader workers	8 or 16 (execution only)
TensorFloat-32 on CUDA	enabled
Contrastive temperature τ	0.07
Training seeds	101, 202, 303

Table A20: Training configuration for the common GPT-2 backbone in the SMILES/SELFIES comparison.

Certified-subset optimizer batch. For certified-subset training, each optimizer batch contains certified triplets from the training split only. Canonical, last-token, and SIGMA training compute token-normalized causal NLL on the complete anchor serialization $A \oplus S$; randomized-serialization training computes it on the paired equivalent serialization $B \oplus S$. Auxiliary methods additionally forward $B \oplus S$ and $C \oplus S$. Their losses are averaged over selected suffix positions, combined with the token-normalized NLL, and backpropagated before one clipped optimizer update. Positive position matching is directed $A \to B$. Micro-batch weights are the fractions of non-padding LM targets and selected auxiliary positions in the original optimizer batch, so execution splitting does not change its token or position normalization.

Loss weights. The final margin-gated setting is shared by SMILES and SELFIES: $\lambda_{\text{aux}} = .075$, $\rho_{\text{neg}} = 1$, $\gamma = .10$, and $\tau = .07$. Thus $\lambda_{\text{neg}}^{\text{eff}} = \lambda_{\text{aux}} \rho_{\text{neg}} = .075$. The certified negative contributes only while it violates the positive-relative margin. The last-token binary loss is scaled by 0.075; its positive and negative similarities are the two logits of one cross-entropy, not two independently weighted terms. SAFE uses the same setting in its seed-101 fragment-assembly analysis.

Certified-subset and full-corpus regimes. The primary SELFIES comparison uses the same certified subset of the training split for all four methods: canonical, randomized, last-token alignment, and SIGMA therefore see the same row identities, but molecules without a certified triplet do not contribute ordinary LM loss. The primary SMILES comparison instead uses the full-corpus design. It decouples the two streams: every method receives the same complete-molecule language-model batch schedule (the randomized control uses its corresponding randomized serialization), while methods with alignment losses receive a separate certified-triplet auxiliary stream. Comparisons are paired only within a representation–dataset block and never treat the two coverage regimes as numerically interchangeable. The first additional full-corpus block covers 107,059 QM9 language-model rows, 24 models (eight conditions by three seeds), validation NLL, and 120 paired FCD evaluations. It confirms the dense suffix-position benefit under the same positive–negative loss formulation; the corresponding controlled ablation is reported in Appendix B. A second, independently specified controlled comparison covers SMILES–ZINC, SMILES–ChEMBL, and SELFIES–ZINC. It compares the full-corpus LM, a prespecified symmetric-KL output-space variant, positive-only certified state alignment (SIGMA-Pos), a shuffled-pair compute control, and margin-gated certified state alignment (SIGMA-Margin). The three-seed grid contains 45 trained models and 225 paired FCD evaluations (five generation repetitions per model).

F.2 Data Splits and Model Selection

The three splits have distinct roles. LM examples and auxiliary triplets used for optimization come only from the training split. Separately constructed validation triplets support validation losses and

mechanism diagnostics, not optimizer updates. Hyperparameters and checkpoint choices use validation data, including validation FCD and NLL. Formal-test molecules enter only the evaluation-reference pipeline described below, not the training-triplet pool, optimization, or model selection. After selection, the saved generated samples are scored against the fixed test references without retraining or regenerating samples. Single-seed grids are used for sensitivity analysis; the principal SMILES and SELFIES comparisons use three training seeds and five paired generation repetitions.

F.3 Datasets and Evaluation References

Representation	Dataset	LM train	LM valid	LM test	Triplet train	Triplet valid	Test reference / primary N
SMILES	QM9	107,108	13,388	13,389	80,932	10,061	13,388 / 5,000
SMILES	ZINC	220,011	24,445	5,000	138,750	15,413	5,000 / 5,000
SMILES	ChEMBL	1,273,104	79,568	238,706	814,718	51,002	20,000 / 5,000
SMILES	MOSES-300k	300,000	30,000	176,074	194,012	19,438	20,000 / 5,000
SELFIES	QM9	107,108	13,388	13,389	45,936	5,730	5,666 / 5,000
SELFIES	ZINC	220,011	24,445	5,000	111,041	12,372	2,516 / 2,516
SELFIES	ChEMBL	1,273,104	79,568	238,706	653,322	40,863	122,217 / 5,000
SELFIES	MOSES-300k	300,000	30,000	176,074	163,118	16,331	95,632 / 5,000

Table A21: Parent-split row counts, certified-triplet counts, and formal test-reference sizes. The LM columns describe the parent corpora; primary SELFIES optimization uses the training rows counted under "Triplet train," not the test-reference rows. Split and triplet counts need not be unique molecular identities. SMILES references are frozen formal-test rows; SELFIES references are the unique canonical identities in the prespecified certifiable formal-test subset described below. "Test reference" is the number of canonical reference rows read by the evaluator; the primary FCD sample size is $\min(5{,}000, \text{reference}, \text{valid generated})$.

Table A21 distinguishes language-model rows, certified-triplet rows, frozen formal-test references, and the primary FCD sample size for every representation–dataset block. QM9 uses the DeepChem `qm9.csv` snapshot with seed-0 80/10/10 splitting. The local ZINC data were inherited from the `grammarVAE` distribution, and MOSES from the `molecularsets/moses` distribution. ChEMBL uses the `chembl_guacamol_v1` split. The experiment records include processed split checksums, row counts, and preprocessing commands.

For the final evaluation, each representation–dataset block uses the frozen formal-test reference listed in Table A21. SMILES references are canonical molecular strings. SELFIES test molecules are decoded and canonicalized before ChemNet evaluation, so FCD is computed on molecular graphs rather than representation tokens.

The SELFIES evaluation references form a *certifiable formal-test subset*. Pair construction and chemical certification here determine reference eligibility only: test-derived pairs are not supplied to training or validation loaders. The filters do not use model outputs, and the resulting reference files are fixed across methods and seeds before final scoring. A test molecule is retained only when pair construction finds distinct serialization histories with an exact shared SELFIES suffix of at least five tokens, valid prefixes, and equal completed-molecule identity for the positive pair. A second strict audit decodes every suffix offset and requires the positive histories to have equal chemical identity at the initial state and after each suffix token. Retained canonical identities are then deduplicated. These filters depend only on the representation and the same chemical certificate used to define training examples; they are identical for every compared method. For ZINC, the prespecified stages reduce 5,000 formal-test molecules to 2,925 pair candidates and then to 2,516 stepwise-certified unique references. The other SELFIES reference counts in Table A21 follow the same procedure.

Hyperparameters and checkpoints were selected using validation data. The selected checkpoints and paired samples were subsequently evaluated against the prespecified formal-test references in Table A21. Reference construction is representation-only and method-independent.

F.4 Generation Protocol

The primary SMILES and SELFIES tables both use temperature 1, untruncated categorical sampling ($p = 1$), minimum length 3, and a target of 5,000 valid molecules per model and generation

repetition (up to 250,000 proposals). SMILES uses a maximum length of 64 tokens. SELFIES uses 64 tokens for QM9 and 96 for ZINC, ChEMBL, and MOSES, matching the frozen representation-specific registries. Each of three training seeds is evaluated with five paired generation seeds. Within a representation–dataset–seed–repetition group, all four methods use the same sampling seed.

F.5 FCD Implementation

The evaluator canonicalizes valid molecules, encodes them with the 512-dimensional pretrained ChemNet representation, and applies the standard Fréchet mean/covariance formula using `fcd-torch` 1.0.7 and `ChemNet_v0.13_pretrained.pt`. Pair construction records RDKit 2022.09.5 and SELFIES 2.1.1, while the current evaluation environment reports RDKit 2026.3.1 and SELFIES 2.2.0.

The formal evaluation environment is: Python 3.10.20, PyTorch 2.11.0+`cu126`, NumPy 2.2.6, SciPy 1.15.3, RDKit 2026.3.1, SELFIES 2.2.0, Hugging Face Transformers 4.24.0, SmilesPE 0.0.3, `fcd-torch` 1.0.7, and the recorded ChemNet weights.

F.6 Metrics

FCD ↓. Fréchet ChemNet Distance between generated and reference molecules in ChemNet feature space; lower is better. It is the primary distribution-fidelity metric, following the MOSES evaluation practice Polykovskiy et al. (2020).

Validity ↑. Fraction of proposals that decode and pass the chemical parser/sanitizer.

Uniqueness ↑. Fraction of canonical molecular identities that are unique among valid outputs.

Novelty ↑. Fraction of valid generated identities absent from the training identities.

Connectedness ↑. Fraction of valid molecules represented by one connected molecular graph rather than disconnected components.

Validation NLL ↓. Token-normalized causal negative log-likelihood on the validation split.

Validity, uniqueness, novelty, connectedness, and NLL complement FCD by revealing invalid-sample filtering, diversity loss, or likelihood degradation.

F.7 EOS-Level SIGMA for Molecular Property Prediction

SIGMA's shared abstraction is a task-relevant prediction state together with certified equivalent and non-equivalent contexts. In unconditional generation, the prediction states are all aligned suffix offsets. In a whole-molecule property task, the single task-relevant prediction state is EOS. The downstream block therefore reuses the pretrained six-layer GPT-2 backbone (hidden size 256, eight attention heads, context 512) and reads the EOS hidden state with an external affine property head. Language modeling and property prediction are optimized jointly for 40 epochs with batch size 64. The backbone and head learning rates are 2×10^{-5} and 5×10^{-4}, respectively, and the checkpoint with the best validation metric is selected. Baseline and SIGMA runs share the same initialization within each task and training seed.

The comparator is deliberately stronger than canonical-only training: it uses the same randomized molecular serializations and property objective, but no equivalence-alignment auxiliary. For EOS-level SIGMA, the positive pair comprises two valid serializations of the same molecule and is aligned at EOS using $1-\cos(h_{\mathrm{EOS}}, h^{+}_{\mathrm{EOS}})$. A property-aware negative is an opposite-class molecule for classification or a molecule whose target differs by at least half a training-target standard deviation for regression. Its finite-margin term is

$$\left[\mathrm{sim}(h, h^{-}) - \mathrm{sim}(h, h^{+}) + \gamma\right]_{+}.$$

Task settings are selected using validation data; the standard setting uses positive coefficient 1, negative coefficient .025, and margin .1. For FreeSolv, a validation-selected variant excludes molecules more than 1.5 training-target standard deviations from the mean from the auxiliary alignment; for BACE, the alternate serialization preserves explicit-hydrogen tokens and respects the context-length limit. These task adaptations preserve SIGMA's chemical-equivalence and finite-separation principles while applying them at the task's EOS decision rather than at every suffix decision.

The six benchmarks are ESOL, FreeSolv, Lipophilicity, BACE, BBBP, and ClinTox. The primary prediction for each molecule averages one canonical and one deterministic alternate serialization. Regression tasks use RMSE and classification tasks AUROC. Inference resamples training seeds and test molecules hierarchically for 10,000 paired bootstrap draws and reports percentile 95% intervals. Serialization stability is the within-molecule prediction standard deviation across the two paired serializations and is reported as a descriptive mechanism endpoint.

F.8 Objective-Sensitive Implementation Choices

The following seemingly minor implementation choices alter the mathematical objective or the comparison protocol:

1. **Absolute-index alignment** pairs different suffix offsets when prefix lengths differ.
2. **Unmasked padding** leaves finite zero-similarity padding columns in the softmax denominator and changes the objective with batch padding.
3. **Whole-fragment SAFE tokens** erase the basic-token prediction path and make SAFE incomparable to SMILES/SELFIES.
4. **Uncertified negatives** can accidentally represent the anchor state after the shared suffix; chemical certification removes them.
5. **Mixed sampling protocols** can reverse apparent winners, especially on QM9; each table therefore uses one stated protocol.

G Statistical Procedure

G.1 Paired Hierarchical Bootstrap

For a fixed representation, dataset, and control, let $d_{sr} = y_{sr}^{SIGMA} - y_{sr}^{\text{control}}$ be the paired FCD difference for training seed $s \in \{101, 202, 303\}$ and generation repetition $r \in \{1, \ldots, 5\}$. The point estimate is

$$\widehat{\Delta} = \frac{1}{3} \sum_{s} \frac{1}{5} \sum_{r} d_{sr}.$$

For each of 10,000 bootstrap draws, three training seeds are sampled with replacement. Within every drawn seed, five repetition indices are then sampled with replacement. The draw is the mean of the resulting three within-seed means. The 2.5% and 97.5% quantiles form the percentile interval.

G.2 Interpreting Paired Differences

For each representation and dataset, canonical training is the primary comparator; randomized-serialization training and last-token alignment localize where gains arise. We report paired hierarchical bootstrap 95% intervals. A comparison is reported as a clear reduction for SIGMA when the point estimate favors SIGMA and the upper endpoint of the 95% interval is below zero. An interval that includes zero is reported as inconclusive.

Training seed is the top-level resampling cluster; the five paired generations estimate within-model decoding variation. We additionally report the sign of each seed-level mean difference as a stability diagnostic, but do not turn it into a separate significance test. Thus the inferential object is the effect estimate and its 95% interval; seed directions help readers identify initialization sensitivity.

G.3 Multiplicity Sensitivity

The paired-effect interval above is the prespecified primary reporting rule. As a supplementary sensitivity analysis, we also apply Holm correction to the 24 primary representation–dataset–control comparisons. Nineteen of the 20 comparisons whose 95% interval lies fully below zero also have Holm-adjusted $p < .05$. The sole exception is SELFIES–ChEMBL against canonical training ($\Delta = -.0536$, 95% CI $[-.1033, -.0085]$, $p_{\text{Holm}} = .0780$); its comparisons against randomized serialization and last-token alignment remain below zero after correction. The complete unadjusted and Holm-adjusted records are included in the anonymous package.

G.4 Sampling Sensitivity

Results using $p = .95$ and $p = 1$ are reported separately so that sampling sensitivity is not conflated with changes to the training objective. Within each dataset–seed–repetition block, methods share the sampling seed and requested sample count; each $p = .95/p = 1$ pair reuses the same checkpoint and differs only in nucleus truncation. This is a paired decoding sensitivity analysis rather than method-specific tuning.

H Weight, Projection, and Sampling Sensitivity

These studies examine the sensitivity of alignment and decoding choices within their stated protocols. Their absolute FCD values are not pooled with the primary test-reference results in Appendix A.

H.1 Loss-Weight Surfaces

The one-seed validation scans study sensitivity to the positive and negative weights. The backbone-state alignment grid varied the effective positive and negative coefficients $\lambda_+ = \lambda_{\text{aux}}$ and $\lambda_- = \lambda_{\text{aux}}\rho_{\text{neg}}$ while holding the backbone, initialization, data, generation repetitions, and validation split fixed. Its selected point was $(.25, 0)$ with validation FCD .5039, compared with .6309 for the strongest control under the same protocol (20.13% lower; five of five paired generation repetitions). This setting did not transfer uniformly across domains; the shared method instead uses the margin-gated setting $(\lambda_{\text{aux}}, \rho_{\text{neg}}, \gamma) = (.075, 1, .10)$.

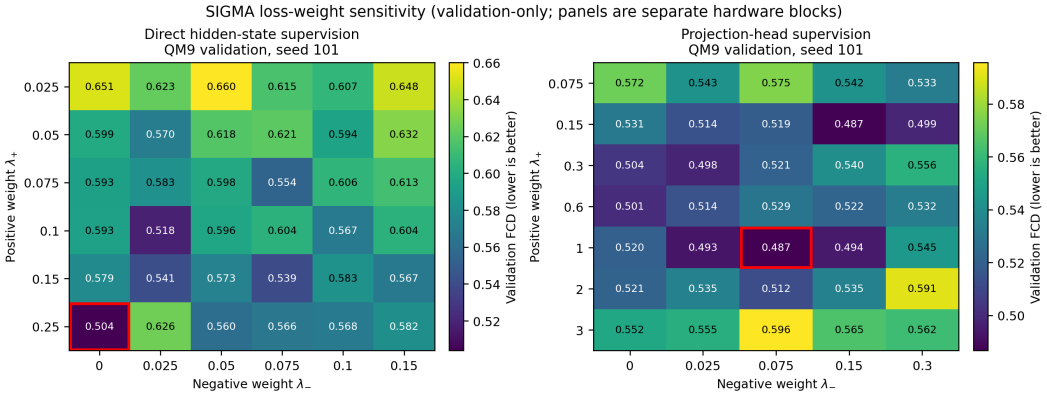


Figure A4: Validation-only loss-weight surfaces for backbone-state and projection-head alignment. The outline marks the minimum within each independently run grid.

Figure A4 shows both validation surfaces and makes their non-monotonicity explicit: increasing either coefficient does not uniformly improve FCD, and the independently selected minima differ between backbone-state and projection-head alignment.

Cross-domain transfer. A stronger positive weight that looked best in an earlier validation ablation, $(1, .075)$, does not transfer uniformly. Under untruncated categorical SMILES sampling it improves mean FCD relative to $(.075, .075)$ on QM9, ZINC, and MOSES, but worsens ChEMBL (1.637 versus 1.572). This non-monotonicity motivates selecting one shared setting across multiple validation domains rather than maximizing the weight on a single dataset.

H.2 Backbone-State versus Projection-Head Alignment

Projection-head alignment adds

$$\mathbb{R}^{256} \to \mathbb{R}^{256} \xrightarrow{\text{ReLU}} \mathbb{R}^{128}$$

only for the auxiliary loss; the LM head and generation remain attached to the unprojected backbone. In the complete 35-point QM9 validation grid, the best projection-head setting was $(1, .075)$ with FCD .4866, while the strongest backbone-state setting under the same protocol reached .4782. Thus the best projection-head point was 1.77% worse, so the unprojected final-layer state remains the default.

H.3 Sensitivity to Nucleus Truncation

Dataset	Method	FCD ↓	Validity ↑	Uniqueness ↑	Novelty ↑
QM9	Canonical	0.665 ± 0.045	0.760	0.981	0.599
QM9	Randomized serialization	0.800 ± 0.025	0.616	0.982	0.652
QM9	Last-token alignment	0.661 ± 0.059	0.758	0.981	0.583
QM9	SIGMA $(.075, .075)$	0.630 ± 0.082	0.753	0.982	0.589
QM9	SIGMA $(1, .075)$	**0.620 ± 0.051**	0.783	0.980	0.556
ZINC	Canonical	1.953 ± 0.220	0.735	1.000	1.000
ZINC	Randomized serialization	2.326 ± 0.213	0.635	1.000	1.000
ZINC	Last-token alignment	2.005 ± 0.179	0.717	1.000	0.999
ZINC	SIGMA $(.075, .075)$	1.972 ± 0.228	0.698	1.000	1.000
ZINC	SIGMA $(1, .075)$	**1.847 ± 0.219**	0.731	1.000	1.000
ChEMBL	Canonical	1.574 ± 0.095	0.873	0.999	0.976
ChEMBL	Randomized serialization	1.743 ± 0.086	0.834	1.000	0.981
ChEMBL	Last-token alignment	1.576 ± 0.082	0.869	0.999	0.977
ChEMBL	SIGMA $(.075, .075)$	**1.572 ± 0.097**	0.861	0.999	0.976
ChEMBL	SIGMA $(1, .075)$	1.637 ± 0.104	0.855	0.999	0.977
MOSES-300k	Canonical	5.840 ± 0.054	0.821	1.000	0.992
MOSES-300k	Randomized serialization	6.147 ± 0.060	0.749	1.000	0.993
MOSES-300k	Last-token alignment	5.810 ± 0.018	0.814	1.000	0.992
MOSES-300k	SIGMA $(.075, .075)$	5.793 ± 0.057	0.815	1.000	0.991
MOSES-300k	SIGMA $(1, .075)$	**5.706 ± 0.087**	0.825	1.000	0.991

Table A22: SMILES robustness under untruncated categorical sampling ($T = 1$, $p = 1$). FCD is mean ± training-seed standard deviation.

Table A22 changes only the nucleus threshold from $p = .95$ to $p = 1$ while holding the training checkpoints fixed. It isolates sampling sensitivity from training-objective effects. The untruncated-sampling table removes the extreme SMILES-QM9 degradation seen for one SIGMA seed with dense suffix-position supervision under $p = .95$, supporting the narrower interpretation that a hard top-p cutoff can amplify a modest first-token calibration shift. It does not imply that sampling is the sole cause: the model's shifted probability mass is the trigger, while top-p converts that shift into an abrupt inclusion/exclusion decision. We therefore report the sampling rule with every generation table.

I Representation-Specific Boundaries and Scope

I.1 SAFE Fragment-Assembly Limitation

Dataset	Canonical	Randomized serialization	Last-token alignment	SIGMA
QM9	.7886±.0800	.8643±.0641	**.7157±.0292**	.7691±.0476
ZINC	11.0973±.1331	11.4068±.1277	**9.5305±.1395**	9.6728±.0361
ChEMBL	4.3772±.0827	5.3912±.1347	**4.3037±.0670**	4.3252±.1048
MOSES-300k	12.1301±.1424	12.5187±.2675	**10.2188±.0911**	10.6665±.1847

Table A23: SAFE seed-101 fragment-assembly results (FCD mean ± standard deviation over five paired generation repetitions; lower is better). This single-seed analysis was not extended to additional training seeds.

Table A23 reports the single-seed SAFE boundary analysis referenced in the main paper. It distinguishes improved local correspondence from successful global fragment attachment by comparing all four training objectives under the same atomic tokenizer, common GPT-2 implementation, and margin-gated objective family. SIGMA improves over canonical and randomized-serialization training on all four domains, but loses to the last-token control on all four. Its mechanism metrics all move in the intended direction, but the fraction of connected generations falls from .916/.254/.641/.259 for last-token alignment to .915/.237/.636/.221 for SIGMA on QM9/ZINC/ChEMBL/MOSES. Conditional on connected outputs, SIGMA lowers FCD on ZINC, ChEMBL, and MOSES (4.8100, 4.7320, and 4.0255 versus 4.8956, 4.7751, and 4.3231), while QM9 moves from .6556 to .6746. Thus local suffix-prediction alignment transfers, but complete fragment assembly remains the SAFE bottleneck.

I.2 Diagnosing the SMILES–ChEMBL Boundary

SMILES–ChEMBL separates improvement in the intended local mechanism from improvement in the complete generated distribution. Aggregate validation NLL and next-token accuracy remain essentially unchanged, but ring-closure and EOS NLL worsen in all three seeds. Canonical generation already undershoots the formal test reference in molecular size and ring count; SIGMA shifts both slightly farther downward. The paired normalized Wasserstein gap consequently increases for rings ($+.0106$, 95% CI $[+.0027, +.0191]$) and aromatic rings ($+.0145$, $[+.0052, +.0250]$), while the heavy-atom interval crosses zero. Thus the boundary is localized to topological closure and stopping rather than a corpus-wide loss of token prediction.

The position-matching formulation suggests a specific source of this selective error. Every non-matching position in the positive serialization is a retrieval competitor, including positions that predict the same next token. At least one such competitor occurs at 95.9% of ChEMBL suffix positions, with 7.94 competitors on average, compared with 91.7% and 5.13 on ZINC. If a target and its K same-token competitors receive equal similarity, their combined softmax mass bounds the target probability by $1/(K+1)$, giving a loss of at least $\log(K+1)$. This is a conditional bound, not an unavoidable training penalty: repeated observed tokens need not have equivalent chemical states or next-token distributions, and even matching output distributions need not imply equal hidden-state similarities. The counts therefore motivate repeated-position competition as a hypothesis rather than establish it as the cause of the FCD result. Separate chemical-identity checks assess label consistency under the declared Standard-InChIKey relation, while auxiliary-sampling checks assess concentration on individual molecules.

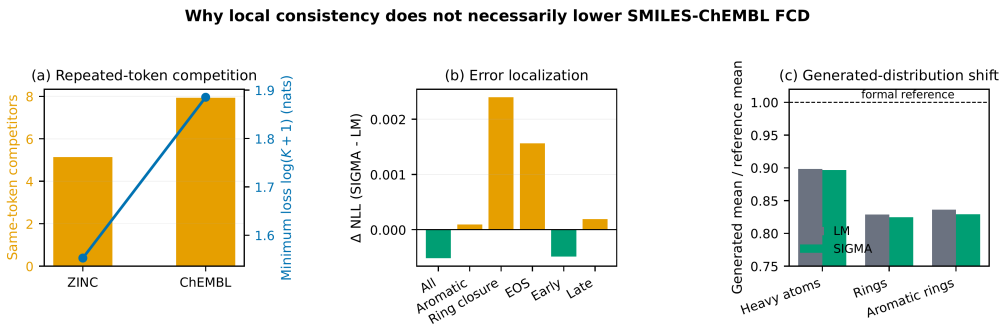


Figure A5: Diagnosing the SMILES–ChEMBL boundary. (a) ChEMBL presents more same-token retrieval competitors than ZINC. The blue curve shows the conditional $\log(K+1)$ lower bound when target and competitor similarities are equal, not an observed or unavoidable loss. (b) Aggregate token likelihood is preserved, while degradation concentrates on ring closure and EOS. (c) Both the language-model baseline and SIGMA generate molecules that are smaller and more ring-poor than the formal reference, with SIGMA shifting slightly farther in the same direction. The complete paired FCD contrasts are reported in Appendix A.5.

Figure A5 combines the three diagnostics: repeated same-token competitors, token-category calibration, and the generated-versus- reference size and ring distributions. The latter two identify a localized calibration pattern associated with the FCD boundary; the first suggests a possible training-objective explanation that these diagnostics alone do not establish causally.

I.3 Scope of the Evaluation

1. **Coverage.** SIGMA supervises examples for which a nonempty, chemically certified same-suffix triplet can be constructed. Coverage varies by representation and domain. Each primary comparison fixes language-model coverage across its four methods: SMILES uses the complete parent corpus plus a separate auxiliary stream, while SELFIES uses the certified-row corpus. Absolute values are not compared across these two coverage regimes. Full-corpus QM9, SMILES–ZINC, SMILES–ChEMBL, and SELFIES–ZINC blocks compare output alignment, hidden alignment, chemical mapping, and extra compute. The remaining representation–domain combinations have not yet received this complete controlled comparison.
2. **Chemical identity.** Standard InChIKey provides a practical, reproducible equivalence rule, not exact charged-graph identity. Its mobile-H and protonation/charge normalization can merge distinct tautomeric forms or neutral/zwitterionic drawings. The independent checker verifies reconstruction under this same declared relation; it does not establish Fixed-H or graph-exact equality. Using either stricter rule would require rebuilding the triplets and re-evaluating the models.
3. **Model scale.** The experiments use a six-layer, 256-dimensional GPT-2. Behavior under substantially larger molecular language models remains to be established.
4. **Evaluation scope.** FCD measures distributional fidelity, and the property benchmarks measure predictive transfer. Neither substitutes for prospective synthesis, biological testing, or a medicinal-chemistry campaign.
5. **Representation-specific limitations.** SMILES–ChEMBL exhibits closure and stopping errors in long, repetitive strings; repeated-position competition is a possible explanation, not a demonstrated cause. SAFE shows a different limitation: local continuation alignment can improve while global fragment attachment becomes less reliable. These observations motivate future objectives that account explicitly for repeated-token equivalence and graph-assembly constraints.

J Computational Cost and Reproducibility

J.1 Training and Inference Cost

Pair construction is an offline preprocessing cost. When the anchor also supplies the language-model example, SIGMA adds two forward passes for the positive and negative histories. With a separate full-corpus language-model stream, the auxiliary triplet batch instead adds three forward passes. Both regimes leave the model architecture and generation path unchanged. We report wall-clock time, optimizer steps, peak accelerator memory, checkpoint size, certified triplets per second, training examples per second, and generation proposals per second for identical hardware blocks.

Throughput comparisons use the same hardware. On A100-PCIE-40GB, generation reaches 365–373 proposals/s on ZINC and approximately 312 proposals/s on ChEMBL across methods, consistent with the unchanged inference graph.

J.2 Reproducibility

The anonymous code package provides the packaged scientific source for data preparation, position-wise triplet certification, all compared training objectives, paired generation, molecular evaluation, mechanism analyses, downstream evaluation, and statistical aggregation. It also contains a portable reference interface, the experiment configurations, compact aggregate results, a version-locked software environment, and unit and chemical smoke tests for alignment, masking, micro-batching, and finite-margin contracts. Machine-readable records maintain checksums, software versions, per-run metrics, and aggregation inputs without placing implementation bookkeeping in the manuscript.

Reproduction steps. Construct and independently verify the representation-specific triplets, train the controlled methods from shared seed-level initializations, generate paired repetitions with common sampling seeds, and aggregate per-seed metrics and intervals. Full-corpus, mechanism,

information-preservation, and sensitivity analyses use their recorded configurations. The package rebuilds the tables, figures, and statistical summaries from included aggregates; a provenance map links each output to its input and script. Full reruns require the public splits and ChemNet weights specified in Appendix F. Public corpora, checkpoints, and large generated-sample files are not redistributed.