A LARGE SCALE INVESTIGATION OF SCALING LIMITS IN CHEMICAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chemical Language Models (CLMs) are increasingly used in *de novo* drug design, driven by recent growth in model scale, compute, and dataset size. However, the relationship between design choices, training dynamics, and downstream generation quality remains poorly understood. We present a compute-controlled scaling study of CLMs comprising more than 30,000 experiments across molecular representations (SMILES, SELFIES, SAFE), tokenizations (atom-level and byte-pair encoding), model scales (0.5M–1B parameters), leakage-controlled datasets (MOSES, ChEMBL, PubChem, ZINC-22), and architectures (decoder-only and encoder–decoder). By fitting IsoFLOP profiles, we establish clear scaling trends in pretraining loss, but find that these improvements do not translate into comparable gains in goal-directed molecular design. Layer-wise probing and sparse autoencoder analysis reveal continued development of chemical representations: chemical syntax saturates early, while semantic properties emerge more slowly and become increasingly accessible with further training and model scale. These representational gains coexist with diminishing improvements in goal-directed generation under the evaluated protocols and oracle budgets. Our resulting suite of models, NovoMolGen, achieves state-of-the-art results, outperforming prior CLMs and specialized generative models in goal-directed molecular generation across drug discovery tasks. These findings expose a disconnect between chemical representation learning and downstream molecular design, motivating the development of pretraining paradigms that more directly learn chemical semantics.

1 INTRODUCTION

Chemical Language Models (CLMs) have largely adopted the scaling paradigm of Natural Language Processing (NLP), pretraining increasingly larger autoregressive transformers on massive molecular corpora. The implicit assumption is that minimising next-token cross-entropy translates directly into better molecular design (Olivecrona et al., 2017; Bagal et al., 2022; Ross et al., 2024). In NLP, scaling laws describe predictable improvements in language modeling as model size, training data, and compute increase (Kaplan et al., 2020; Hoffmann et al., 2022). However, *de novo* molecular generation optimises candidates using downstream iteration loops against biological and physicochemical property oracles (Gao et al., 2022). Understanding molecular scaling therefore requires tracing its effects from sequence modeling and learned chemical representations to goal-directed generation. Whether continued improvements in the former translate into comparable gains in molecular design remains poorly understood. Recent studies have explored scaling behaviour in CLMs, revealing different outcomes across evaluation settings. While Frey et al. (2023) observed scaling in CLMs, their analysis did not establish how these gains translate to goal-directed molecular generation. More recently, Xu et al. (2026) studied compute-controlled scaling across representations and property prediction, while Medina et al. (2026) observed that hit-generation rates saturate with model scale in reinforcement-learning-based molecular optimisation. However, these studies do not examine how goal-directed generation changes along the pretraining scaling curve, nor do they account for the cost of downstream optimisation. These findings motivate our central hypothesis: *improvements along the compute-optimal pretraining frontier do not necessarily translate into better goal-directed design.*

To establish how scaling translates into molecular design performance, we study CLMs using IsoFLOP profiling across model scales spanning 0.5M to 1B parameters. Our experimental design includes

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.

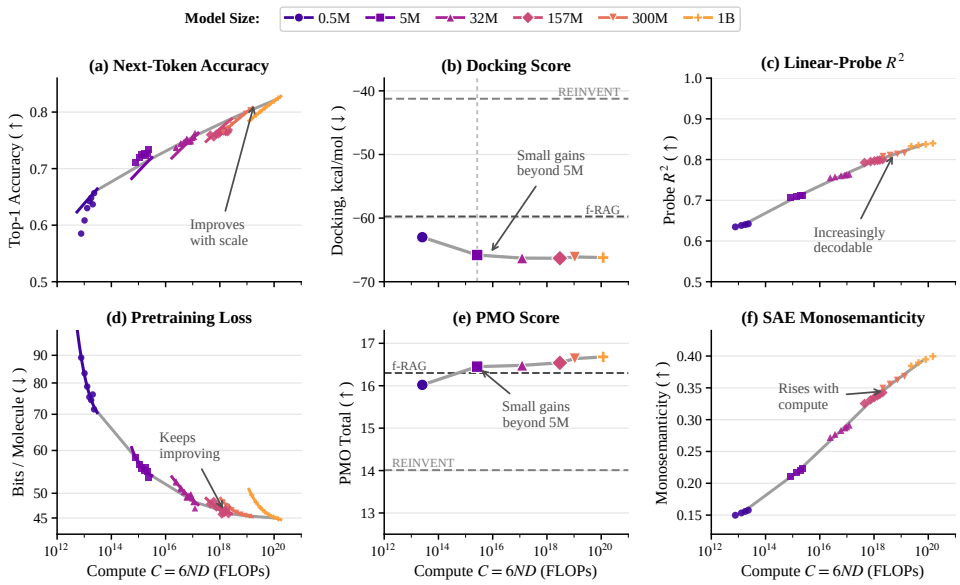


Figure 1: Pretraining and internal representations improve with compute, while goal-directed generation shows diminishing returns. Metrics against pretraining compute $C = 6ND$ for models from 0.5M to 1B parameters. **(a)** Next-token accuracy rises with scale. **(b, e)** After RL fine-tuning, the total top 5% docking score over five targets (↓) and the total PMO score over 22 tasks show only small gains beyond ∼5M. **(c)** Middle-layer linear-probe R^2, averaged over 36 property targets, keeps rising. **(d)** Validation set bits per molecule decrease steadily as compute increases. **(f)** SAE monosemanticity, the mean over features of the largest absolute property correlation, keeps rising.

controlled comparisons of transformer architectures (decoder-only and encoder–decoder), string representations (SMILES, SELFIES, SAFE) (Weininger, 1988; Krenn et al., 2022; Noutahi et al., 2024), tokenisation strategies (atom-level and byte-pair encoding) (Wadell et al., 2026), and datasets (MOSES (Polykovskiy et al., 2020), ChEMBL (Gaulton et al., 2012), PubChem (Kim et al., 2016), and ZINC-22 (Tingle et al., 2023)). To support comparisons across tokenisations and representations, we report validation set negative log-likelihood in bits per molecule, normalising loss per molecule rather than per token. Checkpoints are evaluated on the Practical Molecular Optimisation (PMO) benchmark (Gao et al., 2022) and structure-based docking (Alhossary et al., 2015) under fixed oracle budgets, with representations analysed using layer-wise probing (Belinkov, 2022) and BatchTopK Sparse Autoencoders (SAEs) (Bussmann et al., 2024).

Our large-scale empirical investigation in this work reveals a paradox between pretraining efficiency, internal representations, and practical generative performance (Figure 1):

- Pretraining cross-entropy loss, measured in bits per molecule, decreases smoothly with increasing compute. IsoFLOP profiles reveal power-law relationships between compute-optimal model size and training tokens ($N_{\text{opt}} \propto C^a$, $D_{\text{opt}} \propto C^b$), demonstrating that chemical sequence modelling follows compute-optimal scaling trends within the studied regime.
- Latent chemical representations become richer with increasing compute, as revealed by layer-wise linear probes and SAE feature dictionaries. Models acquire basic chemical syntax in early checkpoints, while molecular properties and functional-group features become more accessible during subsequent training.
- Downstream goal-directed generation largely saturates beyond approximately 5M parameters, despite improvements in bits per molecule and linear accessibility of chemical features. Performance on PMO and structure-based docking flattens under the evaluated optimisation protocols and oracle budgets. A 5M-parameter model achieves performance comparable to models scaling up to 1B parameters across multi-objective optimisation tasks.

These findings demonstrate that scaling parameter counts under current self-supervised string objectives yields diminishing returns in downstream molecular design, despite continued improvements in

pretraining loss and learned chemical representations. Because our probing shows that additional chemical features remain accessible during training, this divergence suggests that the evaluated objectives may limit their downstream utility. Practically, compact CLMs ($\sim$5M parameters) preserve competitive downstream performance while substantially reducing the computational cost of generation and policy updates. To motivate better evaluation standards, we open-source the NovoMolGen family of CLMs, releasing our pretraining checkpoints and downstream task pipelines.

2 Experimental Setup

We study how pretraining scale and design choices affect *de novo* molecular generation. After describing the datasets and design axes (§ 2.1), we evaluate every model at three levels: validation set likelihood (§ 2.2), the linear accessibility of chemical information in its representations (§ 2.3), and goal-directed optimisation under fixed oracle budgets (§ 2.4).

2.1 Datasets and Design Axes

We pretrain our models on datasets of increasing size and diversity: MOSES (Polykovskiy et al., 2020), ChEMBL (Gaulton et al., 2012), PubChem (Kim et al., 2016), and ZINC-22 (Tingle et al., 2023). For ZINC-22, the largest publicly available molecular library ($\approx$96B synthesisable compounds), we sample 1.5B molecules proportionally across its heavy-atom-count and logP tranches. To prevent data leakage, we remove from every corpus all molecules with a Tanimoto similarity above 0.5 (2,048-bit ECFP, radius 2) to any molecule in a set of $\sim$21,000 known actives for the PMO and inhibitors for docking targets and top 1% molecules from the Tartarus benchmarks (Nigam et al., 2023). This active (molecule) set also serves as an unseen set for probing and interpretability. We create four evaluation splits with increasing levels of distributional shift. These include a random split, a scaffold split of held-out Bemis–Murcko scaffolds, a molecular-weight split of the heaviest molecules, and a maximum-dissimilarity split of the fingerprint clusters farthest from the corpus. The random split is the in-distribution reference, and the other three measure generalisation. We vary the model size (0.5M–1B parameters), molecular representation (SMILES, SELFIES, SAFE), tokenisation (atom-level, byte-pair encoding), pretraining corpus, and architecture (decoder-only, encoder–decoder). The reference configuration, a decoder-only model trained on PubChem SMILES with atom-level tokenisation, spans the full size range, and each remaining axis is varied one at a time at matched pretraining compute. Preprocessing and the full experimental grid are detailed in Appx. B.

2.2 Pretraining and Scaling Analysis

We introduce NovoMolGen, a family of 42 CLMs pretrained from scratch across 0.5M–1B parameters, using Qwen3 (Yang et al., 2025) (decoder-only) backbone with next-token prediction and T5Gemma (Zhang et al., 2025) (encoder–decoder) backbone with prefix language modelling. All models use AdamW with a warmup–stable–decay (WSD) schedule and token-aware batching of approximately 25,000 non-padding tokens per optimiser step. Because different molecular representations and tokenisers produce different numbers of tokens, per-token losses are not directly comparable. We therefore report validation set negative log-likelihood in bits per molecule as $L_{\mathrm{mol}} = -\frac{1}{M \ln 2} \sum_{i=1}^{M} \sum_{t=1}^{T_i} \ln p_\theta(x_{i,t} \mid x_{i,<t})$, where M is the number of molecules, T_i is the sequence length including EOS, $x_{i,t}$ is the t-th token, and p_θ is the model probability. For T5Gemma, the encoder receives only a beginning-of-sequence token, so the decoder scores the whole molecule autoregressively and L_{mol} stays comparable to decoder-only models. We estimate non-embedding compute as $C \approx 6ND$, with N the non-embedding parameter count and D the cumulative non-padding training tokens, and use $C \approx 3ND$ for T5Gemma, whose scored tokens pass through the decoder, which holds half the non-embedding parameters. To locate the compute-optimal frontier, we build IsoFLOP profiles that compare model sizes at a fixed compute budget. The loss-minimising size at each budget gives $N_{\mathrm{opt}}(C)$ and $D_{\mathrm{opt}}(C)$. Model configurations and fitting procedures are detailed in Appx. C.

2.3 Probing Internal Chemical Representations

For each frozen checkpoint, we mean-pool residual-stream activations of the active set molecules over non-padding positions at early, middle, and late layers. Linear ridge-regression probes (Belinkov,

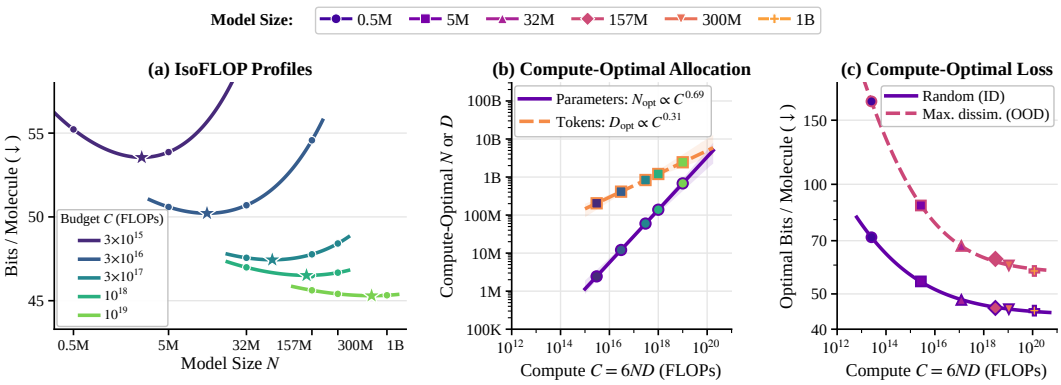


Figure 2: Chemical sequence modelling follows compute-optimal scaling. Validation set bits per molecule (↓) of the PubChem–SMILES–atomwise decoder-only models. **(a)** IsoFLOP profiles at the five compute budgets of our IsoFLOP grid, evaluated at each budget's bracketing model sizes; curves and points come from a parametric loss surface fit to all training trajectories, and stars mark the loss-minimising size. **(b)** Compute-optimal parameters and training tokens follow power laws in compute, with shaded bootstrap intervals. **(c)** Compute-optimal loss on the random (in-distribution) and maximum-dissimilarity (out-of-distribution) splits, with colour and marker denoting model size.

2022) predict 36 properties, grouped into physicochemical descriptors (e.g. molecular weight, logP), PMO objectives and docking scores, and a logistic probe measures chemical syntax by separating valid from corrupted sequences. Probes use a fixed 80/20 split with regularisation from five-fold cross-validation, and we report validation set R^2 and accuracy. We compute each probe's improvement over the randomly initialised network as a fraction of its final improvement relative to training tokens to compare the model's learning pace. To characterise individual features, we train BatchTopK sparse autoencoders (SAEs) on the token-level activations of the same molecules (Bussmann et al., 2024; Cunningham et al., 2023). We measure reconstruction quality, sparsity and monosemanticity, the mean over live features of the largest absolute property correlation, and compare linear probes on SAE features with probes on residual-stream neurons for detecting the ten most common functional groups, each trained on the training split and scored on the validation split (F1). Further details on the probing setup, SAE training and extended analyses are provided in Appx. D.

2.4 Goal-Directed Fine-Tuning and Evaluation

Finally, we test whether pretraining gains translate into goal-directed design under limited oracle calls. We fine-tune each checkpoint with the REINVENT framework (Olivecrona et al., 2017) using Augmented Hill-Climb (Thomas et al., 2022), a well-established reinforcement learning (RL) approach shown to perform competitively against more complex policy optimisers like PPO (Bou et al., 2024; Schulman et al., 2017). The checkpoint initialises a fixed prior $P_{\text{prior}}(x)$ and a trainable agent $P_{\text{agent}}(x)$ that learns to generate molecules x with high task-specific reward $s(x)$. For the highest-scoring molecules in each batch, the agent's log-likelihood is regressed towards the augmented target $\log P_{\text{prior}}(x) + \sigma s(x)$, and an experience-replay buffer reintroduces the best molecules found. We evaluate on 22 PMO tasks (excluding Valsartan SMARTS) with 10,000 oracle calls per task and five protein–ligand docking targets with 3,000 calls per target. PMO performance is the budget-normalised area under the top-10 reward curve (AUC top-10), capturing optimisation quality and sample efficiency. For docking, we report the mean docking score of the top 5% of novel candidates. Reward definitions, hyperparameter selection and oracle scoring are provided in Appx. E.

3 Results

We first establish how validation set loss scales with model size, and compute and locate the compute-optimal allocation of parameters and tokens (§ 3.1). We then ask which design choices change the loss, comparing representations, tokenisers, datasets and architectures at matched compute (§ 3.2), and whether lower loss transfers to out-of-distribution splits and unconstrained sampling (§ 3.3).

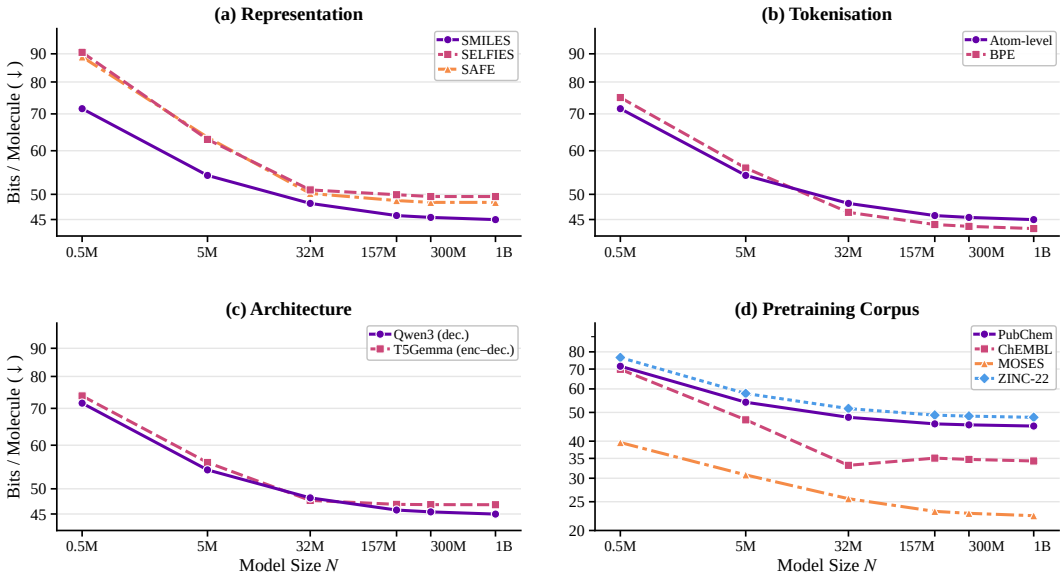


Figure 3: Representation and corpus change the loss far more than tokenisation or architecture. Validation set bits per molecule (↓) at $D = 20N$, varying one design axis at a time around the PubChem–SMILES–atomwise decoder-only reference. **(a)** SMILES is lowest at every scale. **(b)** Byte-pair tokenisation is slightly worse than atom-level tokenisation at small scale and lower from 32M onwards. **(c)** T5Gemma matches the decoder-only model at 32M and trails it at larger scales. **(d)** Corpora are scored on their own validation splits, so their ordering reflects corpus diversity.

We next show how chemical information becomes linearly accessible inside the model as training proceeds (§ 3.4). Finally, we test whether any of these pretraining gains transfer to goal-directed design under fixed oracle budgets (§ 3.5).

3.1 Chemical Sequence Modelling Follows Compute-Optimal Scaling

For the PubChem–SMILES–atomwise reference configuration, validation set loss decreases steadily with pretraining compute across model sizes from 0.5M to 1B parameters (Figure 1d). Trajectories of different sizes overlap, so a larger model does not necessarily reach a lower loss at the same budget. IsoFLOP profiles capture the trade-off between model capacity and the number of training tokens. At a fixed budget, a larger model processes fewer training tokens, while a smaller model sees more tokens but has less capacity. Balancing these competing effects yields a U-shaped profile, whose minimum marks the compute-optimal model size (Figure 2a). We fit a parametric loss surface to the reference losses and summarise the loss-minimising allocation as $N_{\text{opt}}(C) \propto C^{a}$ and $D_{\text{opt}}(C) = \frac{C}{6N_{\text{opt}}(C)} \propto C^{1-a}$, where N_{opt} is the optimal model size, D_{opt} is the corresponding data size, and a is the fitted scaling exponent. The fitted allocations follow an approximate power law, with $N_{\text{opt}} \propto C^{0.69}$ and $D_{\text{opt}} \propto C^{0.31}$ for models from 0.5M to 1B (Figure 2b), so parameters grow faster than tokens, close to the natural-language allocation ($N_{\text{opt}} \propto C^{0.73}$, $D_{\text{opt}} \propto C^{0.27}$) of Kaplan et al. (2020). Our grid spans budgets from 3×10^{15} to 10^{19} FLOPs, where the compute-optimal size ranges from ∼2.5M to ∼700M parameters. Training to 20 tokens per parameter reaches within 1% of the fitted optimal loss at every budget, so we train and compare all sizes at $D = 20N$ and call these checkpoints compute-optimal. The compute-optimal loss then follows a saturating power law on both the random and maximum-dissimilarity splits (Figure 2c). On the random split, it falls from 71.5 bits per molecule at 0.5M to 48.1 at 32M and 45.0 at 1B. Returns diminish at larger budgets but never vanish, and the out-of-distribution loss follows the same trend, its ratio to the in-distribution loss is narrowing from 2.4 at 0.5M to 1.3 at 1B. The IsoFLOP grid and the loss fit are detailed in Appx. F.

Findings: Validation set loss follows power laws in compute for chemical language with $N_{\text{opt}} \propto C^{0.69}$ and $D_{\text{opt}} \propto C^{0.31}$, similar to the allocation for natural language reported by Kaplan et al. (2020).

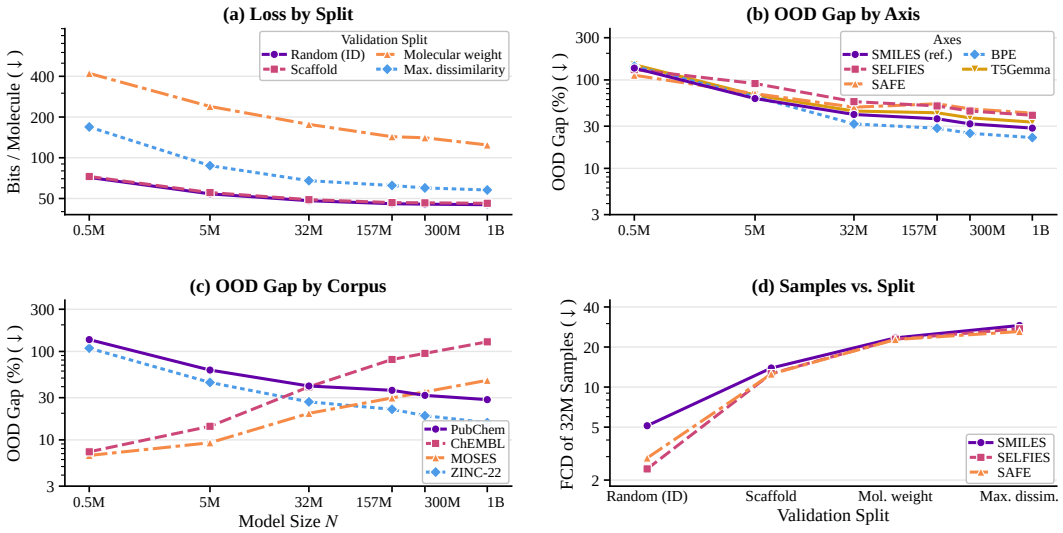


Figure 4: Out-Of-Distribution (OOD) generalisation improves with scale but depends on representation and corpus. The OOD gap is the relative increase in bits per molecule from the random to the maximum-dissimilarity split (↓). **(a)** Validation set loss of the reference configuration on each split. **(b)** OOD gap for each design axis. **(c)** The OOD gap for each corpus, measured on its own splits, narrows with scale on PubChem and ZINC-22 but widens on MOSES and ChEMBL. **(d)** FCD between 30,000 molecules sampled from each 32M model and each split.

3.2 Molecular Representation and Corpus Change the Pretraining Loss

We vary representation, tokenisation, architecture and corpus at matched model size and training tokens (Figure 3). Representation matters most. SMILES reaches lower bits per molecule than SAFE and SELFIES at every scale from 0.5M to 1B, with gaps of 2.1 and 2.8 bits per molecule at 32M that persist at larger scales (Figure 3a). SELFIES tokens are individually easier to predict and give the lowest bits per token, but a SELFIES molecule needs about 155 tokens against 47 for SMILES, so its per-molecule loss stays higher. Tokenisation has a smaller, scale-dependent effect. Byte-pair encoding is slightly worse than atom-level at 5M (55.8 versus 54.1) and better from 32M onwards (46.4 versus 48.1), as larger models exploit the larger vocabulary (Figure 3b). Qwen3 and T5Gemma reach 48.1 and 47.6 bits per molecule at 32M, and T5Gemma then trails the decoder-only model by 1.1–1.8 bits from 157M to 1B (Figure 3c), showing that model architecture does not affect pretraining loss. The corpus changes the loss most, from 25.6 bits per molecule on MOSES to 33.2 on ChEMBL, 48.1 on PubChem and 51.5 on ZINC-22 at 32M (Figure 3d). The ordering follows corpus diversity, from narrow, drug-like MOSES to the broad make-on-demand space of ZINC-22, and not sequence length, since ChEMBL molecules are longer than PubChem molecules yet score lower. Appx. G breaks these losses into bits per token and tokens per molecule.

Findings: Molecular representation and corpus have the largest effect improving pretraining loss, whereas tokenisation and architecture change it by only a few per cent. Per-molecule loss is a better metric than per-token loss for CLM model ranking.

3.3 Studies on Scaling Models and OOD Generalization

The four evaluation splits analyse increasingly strong OOD shifts. The scaffold split behaves almost as in-distribution, at ∼2% higher bits per molecule than the random split at every scale, while the molecular-weight and maximum-dissimilarity splits are genuinely out of distribution, at 5.9× and 2.4× the random-split loss at 0.5M and 2.8× and 1.3× at 1B (Figure 4a). The gap also depends on the design axes and corpus. At 32M it is the smallest for byte-pair SMILES (+32%), then atom-level SMILES (+41%), and largest for SELFIES (+57%) (Figure 4b). The gap reduces with scale on PubChem and ZINC-22 corpus, but widens on the smaller MOSES and ChEMBL, which suggests a closer alignment to their distributions (Figure 4c). For unconstrained generation, we sample 30,000

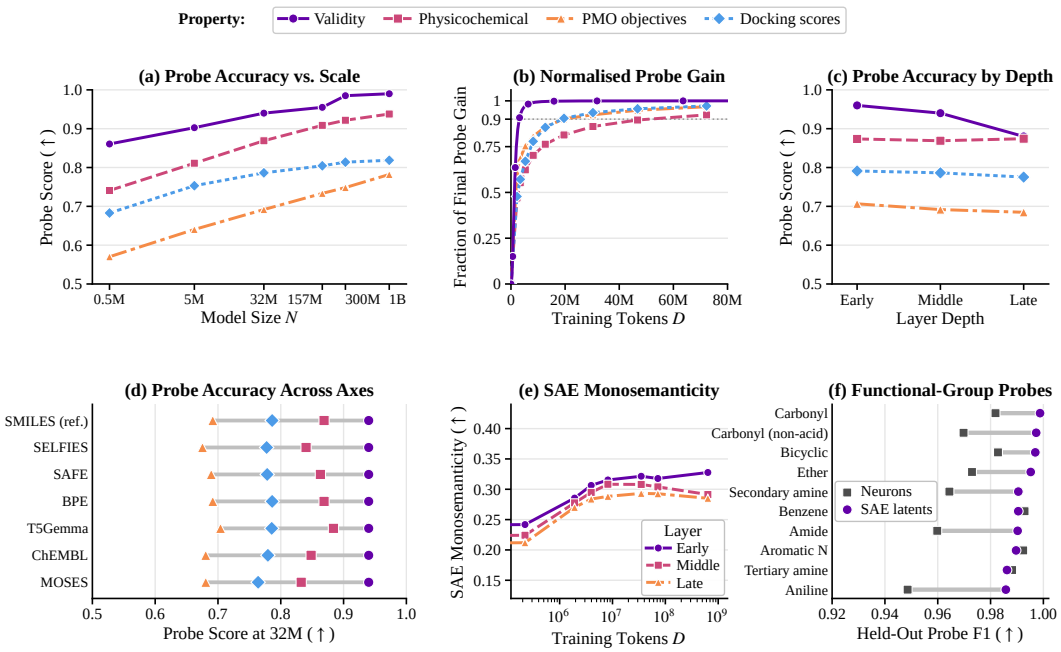


Figure 5: Chemical syntax is learned rapidly, while semantic properties emerge later and more slowly. Linear probes and SAEs on the active set activations. Validity is the accuracy of separating valid molecules from corrupted sequences, and the other properties report probe R^2. **(a)** Middle-layer probe score from 0.5M to 1B parameters. **(b)** Probe gain over the randomly initialised network as a fraction of the gain at the final checkpoint (mean over layers), against training tokens expressed for the 32M model. **(c)** Probe score by depth at 32M. **(d)** Probe score at 32M across design axes. **(e)** SAE monosemanticity of the 32M model during training by depth. **(f)** Validation set F1 of linear probes on SAE latents and on residual-stream neurons for ten common functional groups at 32M.

molecules per model at temperature $T = 1.0$ without top-k or top-p filtering and measuring their Fréchet ChemNet Distance (FCD) (Preuer et al., 2018) to each split, averaging over pretraining seeds at 32M. On the smaller MOSES and ChEMBL corpora, FCD falls with loss, from 1.82 to 0.14 and from 3.13 to 0.22 between 0.5M and 32M. On PubChem, 32M SMILES has the lowest bits per molecule but an FCD of 5.1, against 2.4 for SELFIES and 2.9 for SAFE, and within PubChem SMILES, the FCD rises from 2.8 to 5.1 between 5M and 32M, even as the loss falls from 54.1 to 48.1. Sampled molecules show alignment to the training distribution, their FCD rising from 2.4–5.1 against the random split to 26–29 against the maximum-dissimilarity split for every representation (Figure 4d). Better loss on OOD molecules does not influence the sampling process, so split loss alone cannot effectively select a better CLM. This motivates testing whether scaling pretraining supports goal-directed optimisation. Full unconstrained-generation metrics are reported in Appx. H.

Findings: Scaling CLMs improve the OOD split losses, but does not lead to better generalization in unconstrained molecular generation.

3.4 Learning of Internal Chemical Representations at Scale

We ask whether chemical information becomes more accessible as pretraining progresses. Using frozen checkpoints from 0.5M–1B models, we probe 36 molecular properties and validity on the active set at early, middle and late layers (Figure 5). Every property probe improves with scale, rising between 0.5M and 1B from 0.74 to 0.94 R^2 for physicochemical descriptors, 0.68 to 0.82 for docking scores and 0.57 to 0.78 for PMO objectives, while validity is already 0.86 at 0.5M and reaches 0.99 at 1B (Figure 5a). The validity probe reaches 90% of its final gain after 0.1 tokens per parameter, whereas docking and PMO objectives need a median of 0.6 and physicochemical properties need 1.5 (Figure 5b). Linear accessibility also varies with depth. Docking and PMO R^2 fall slightly towards late layers (0.79 to 0.78 and 0.71 to 0.68), while physicochemical descriptors

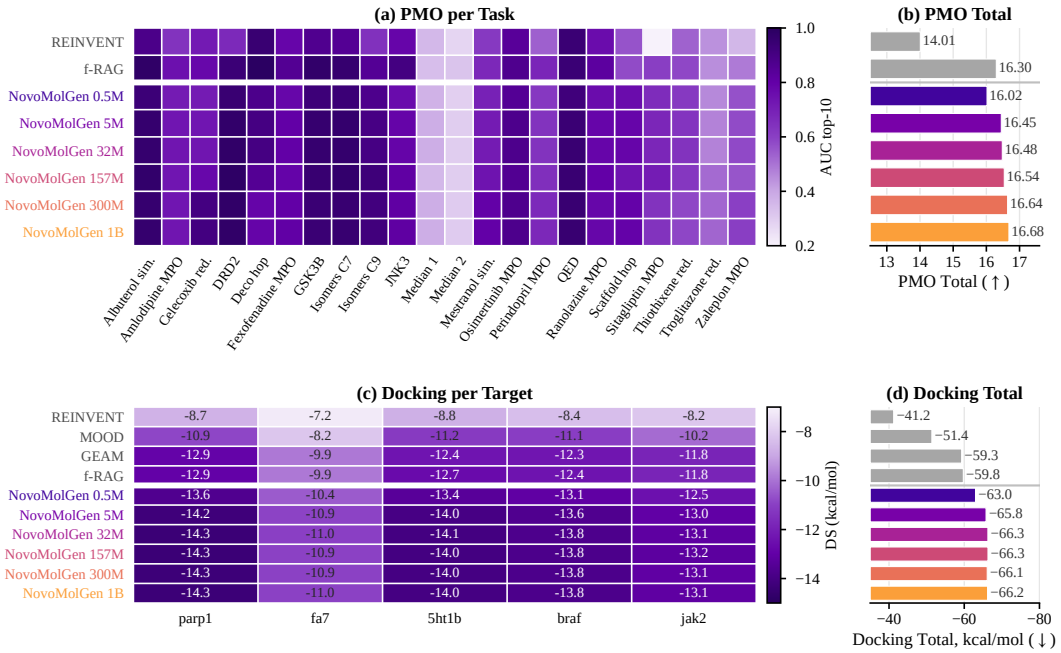


Figure 6: NovoMolGen outperforms strong baselines on goal-directed generation from 5M parameters, with little gain from further scale. Rows compare baselines (grey) with NovoMolGen-SMILES models from 0.5M to 1B parameters. **(a)** PMO AUC top-10 per task with 10,000 oracle calls (↑). **(b)** Total PMO score over the 22 tasks. **(c)** Top-5% docking score of novel molecules per target with 3,000 oracle calls, in kcal/mol. **(d)** Total top 5% docking score over the five targets in kcal/mol (↓). Baseline results are taken from Lee et al. (2024a).

stay similarly decodable (Figure 5c). This ordering of higher validity and lower PMO objectives holds for every design axis at 32M (Figure 5d). BatchTopK SAEs show stronger feature–property associations as training progresses. For the 32M model, early-layer monosemanticity rises from 0.24 to 0.33, while middle- and late-layer values peak mid-training and then decline slightly (Figure 5e), and the fraction of features above the property-association threshold rises from 0.10 to 0.21. These measures indicate alignment with the evaluated properties rather than single-concept encoding. For 7 of the 10 common functional groups, a linear probe on SAE latents beats a probe on residual-stream neurons by up to 0.04 F1 (Figure 5f). Probe configurations, SAE curves, and functional-group F1 scores are in Appx. D.

Findings: CLMs learn what is a valid molecule (syntax) early and learn the molecular properties (semantics) much later during pretraining and keep improving with scale.

3.5 Impact of Scaling on Goal-Directed Molecular Generation

We evaluate goal-directed generation on two complementary benchmarks, the Practical Molecular Optimisation (PMO) benchmark (Gao et al., 2022) and a multi-objective protein–ligand docking benchmark following Lee et al. (2023). For PMO we follow the original protocol of 10,000 oracle calls per task and fine-tune each pretrained model with hyperparameters shared across tasks (Appx. E). NovoMolGen surpasses REINVENT (Olivecrona et al., 2017) and f-RAG (Lee et al., 2024a) from 5M parameters onwards, scoring 16.45 against 16.30 for f-RAG (Figure 6a,b). Scaling to 1B raises the total only to 16.68, a diminishing and often negligible gain (Appx. I). For each docking target (*parp1*, *fa7*, *5ht1b*, *braf*, *jak2*), we generate 3,000 molecules and keep unique candidates with Tanimoto similarity below 0.4 to the training set. A novel hit scores below the median of known actives and satisfies QED > 0.5 and SA < 5, the mean docking score of the top 5% of novel candidates. Against fragment-based, 3D and diffusion baselines, including GEAM (Lee et al., 2024b), MOOD (Lee et al., 2023) and f-RAG, NovoMolGen attains better novel docking scores on every

target (Figure 6c,d). Gains beyond 5M are marginal, the total docking score moving only from -65.8 at 5M to -66.2 at 1B. Docking setup, reward curves and top molecules are in Appx. J.

Findings: NovoMolGen model with 5M parameters outperforms REINVENT and f-RAG on PMO and docking targets. Scaling NovoMolGen models upto 1B parameters yield marginal improvement in goal-directed molecular generation.

4 Discussion and Recommendations

Our analysis, over 30,000 experiments and more than one billion evaluated molecules, reveals a divergence between pretraining and downstream performance in CLMs. Validation set loss keeps improving from 0.5M to 1B (Figure 2), and the same holds across representations, tokenisers and corpora (Figure 3) and even on the out-of-distribution splits (Figure 4). Probe R^2 for PMO objectives and SAE monosemanticity rises between 0.5M and 1B parameter models (Figure 5). PMO and docking scores saturate at $\sim$5M for every representation, tokeniser and architecture, where a compute-optimally trained 5M model already surpasses REINVENT and f-RAG and matches a 1B model. Despite a significant decrease in pretraining loss (9 bits per molecule) between 5M and 1B parameter models, the total PMO score increases only from 16.45 to 16.68 (Figure 6). Similarly, docking scores lie within 0.2 kcal/mol across model sizes on every target (Figure 6). Six of the 22 PMO tasks already reach AUC top-10 ≥ 0.9 at 32M. Median 1 and Median 2 scores stay below 0.4 for every model and baseline. At 10,000 and 3,000 oracle calls, these benchmarks mainly measure how efficiently fine-tuning exploits the oracle and only needs a small pretrained prior proposing valid and diverse molecules. These results indicate that this saturation is not a limitation of the transformer architecture, but a consequence of the current pretraining data, modelling objectives and benchmarks. Unlike genomic or proteomic data, where evolutionary pressure has embedded a rich, functional learning signal, chemical libraries such as PubChem and ZINC-22 are curated for coverage and synthesisability, so their dominant signal teaches chemical syntax rather than functional semantics. Similar limits of scale-driven approaches have been reported for proteins (Fournier et al., 2024) and DNA (Boshar et al., 2025). Further progress will require objectives beyond chemical validity, for example, signals from protein–ligand interactions, physicochemical properties or experimental bioactivity combined with self-supervision, or reinforcement learning introduced earlier in training as a "fitness" signal analogous to natural selection. Our results suggest several practical guidelines for building and evaluating next-generation CLMs:

- **Train better sample-efficient algorithms, not larger models.** A compute-optimal 5M model uses $\sim 4 \times 10^4$ fewer pretraining FLOPs than a 1B model, and makes each policy update $\sim$200 times cheaper, at comparable PMO and docking performance.
- **SMILES remains the best representation.** SMILES with a decoder-only model gives the lowest loss and the highest downstream scores, while tokenisation and architecture change the loss by only a few per cent.
- **Report bits per molecule and test on OOD splits.** Per-token loss favours long representations such as SELFIES, and the scaffold split is effectively in-distribution, so molecular-weight and maximum-dissimilarity splits are more informative.
- **Better benchmarks.** Harder multi-objective optimisation tasks, targets absent from public training data and stricter oracle budgets would provide a better test of whether larger models help.

5 Conclusions

We presented a compute-controlled study of CLMs from 0.5M to 1B parameters across representation, tokenisation, corpus and architecture. Pretraining loss follows compute-optimal scaling, and internal representations are enriched in two phases, chemical syntax first and chemical semantics next. Yet, goal-directed generation saturates at $\sim$5M parameters, where NovoMolGen already surpasses REINVENT and f-RAG and matches a model two orders of magnitude larger. Although NLP-inspired assumptions that larger models and longer training hold for pretraining and chemical representations, current benchmarks cannot capture the complexity of chemistry. This motivates pretraining modelling objectives that enable CLMs to learn chemical semantics directly. NovoMolGen can also be extended to other tasks, such as text-instructed design and fragment-constrained generation.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing, improve clarity and readability and assist with the retrieval and discovery of related literature. We have reviewed all AI-assisted work and revised the text to reflect our intended meaning and scientific claims. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appx. B describes the datasets, standardisation, leakage filtering and evaluation splits. Appx. C specifies the model architectures, optimiser, learning-rate schedule and token-aware batching. Appx. F reports the IsoFLOP grid, the fitted loss surface and the bootstrap procedure, and Appx. D describes the probing and sparse-autoencoder protocols. Appx. E details the Augmented Hill-Climb objective, reward definitions, oracle budgets and hyperparameter search. We will release the NovoMolGen model suite, intermediate pretraining checkpoints, sparse-autoencoder dictionaries, and the code for pretraining, probing and downstream optimisation to reproduce every reported result.

REFERENCES

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*, 31(13):2214–2216, July 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv082. URL https://doi.org/10.1093/bioinformatics/btv082.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling*, 62(9): 2064–2076, May 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00600. URL https://doi.org/10.1021/acs.jcim.1c00600. Publisher: American Chemical Society.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, pp. 2025–12, 2025.

Albert Bou, Morgan Thomas, Sebastian Dittert, Carles Navarro, Maciej Majewski, Ye Wang, Shivam Patel, Gary Tresadern, Mazen Ahmad, Vincent Moens, et al. Acegen: Reinforcement learning of generative chemical agents for drug discovery. *Journal of Chemical Information and Modeling*, 64 (15):5900–5911, 2024.

Marc Brockschmidt. GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1144–1152. PMLR, November 2020. URL https://proceedings.mlr.press/v119/brockschmidt20a.html. ISSN: 2640-3498.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Nicola De Cao and Thomas Kipf. MolGAN: An implicit generative model for small molecular graphs, September 2022. URL http://arxiv.org/abs/1805.11973. arXiv:1805.11973 [stat].

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 3(10): 1503–1507, 2008. ISSN 1860-7187. doi: 10.1002/cmdc.200800178. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200800178. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.200800178.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, November 2002. ISSN 0095-2338. doi: 10.1021/ci010132r. URL https://doi.org/10.1021/ci010132r. Publisher: American Chemical Society.

Peter Eckmann, Kunyang Sun, Bo Zhao, Mudong Feng, Michael Gilson, and Rose Yu. LIMO: Latent Inceptionism for Targeted Molecule Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5777–5792. PMLR, June 2022. URL https://proceedings.mlr.press/v162/eckmann22a.html. ISSN: 2640-3498.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, February 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00438-4. URL https://www.nature.com/articles/s42256-021-00438-4. Publisher: Nature Publishing Group.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. Domain-agnostic molecular generation with chemical feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9rPyHyjfwP.

Quentin Fournier, Robert M Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, pp. 2024–09, 2024.

Nathan C Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor W Coley, and Vijay Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W. Coley. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization, October 2022. URL http://arxiv.org/abs/2206.12411. arXiv:2206.12411.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and others. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. *Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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 International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272, Sydney, NSW, Australia, August 2017. JMLR.org.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk,

Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models, June 2024. URL http://arxiv.org/abs/2402.00838. arXiv:2402.00838 [cs].

Jeff Guo and Philippe Schwaller. Saturn: Sample-efficient generative molecular design using memory manipulation. *arXiv preprint arXiv:2405.17066*, 2024.

Zhichun Guo, Kehan Guo, Bozhao Nan, Yijun Tian, Roshni G. Iyer, Yihong Ma, Olaf Wiest, Xiangliang Zhang, Wei Wang, Chuxu Zhang, and Nitesh V. Chawla. Graph-based molecular representation learning. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 6638–6646. International Joint Conferences on Artificial Intelligence Organization, August 2023. doi: 10.24963/ijcai.2023/744.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572. Publisher: American Chemical Society.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pp. 6840–6851, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 30016–30030, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8867–8887. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html. ISSN: 2640-3498.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7): 1757–1768, 2012.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, January 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac3ffb. URL https://dx.doi.org/10.1088/2632-2153/ac3ffb. Publisher: IOP Publishing.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2323–2332. PMLR, July 2018. URL https://proceedings.mlr.press/v80/jin18a.html. ISSN: 2640-3498.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pp. 4839–4848. JMLR.org, July 2020.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10362–10383. PMLR, June 2022. URL https://proceedings.mlr.press/v162/jo22a.html. ISSN: 2640-3498.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, and others. PubChem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Lingkai Kong, Jiaming Cui, Haotian Sun, Yuchen Zhuang, B. Aditya Prakash, and Chao Zhang. Autoregressive Diffusion Model for Graph Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17391–17408. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kong23b.html. ISSN: 2640-3498.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588, October 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100588. URL https://www.sciencedirect.com/science/article/pii/S2666389922002069.

Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007/.

Maksim Kuznetsov and Daniil Polykovskiy. MolGrow: A Graph Normalizing Flow for Hierarchical Molecular Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 8226–8234, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i9.17001. URL https://ojs.aaai.org/index.php/AAAI/article/view/17001. Number: 9.

Tuan Le, Robin Winter, Frank Noé, and Djork-Arné Clevert. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chemical Science*, 11(38): 10378–10389, October 2020. ISSN 2041-6539. doi: 10.1039/D0SC03115A. URL https://pubs.rsc.org/en/content/articlelanding/2020/sc/d0sc03115a. Publisher: The Royal Society of Chemistry.

Seul Lee, Jaehyeong Jo, and Sung Ju Hwang. Exploring Chemical Space with Score-based Out-of-distribution Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 18872–18892. PMLR, July 2023. URL https://proceedings.mlr.press/v202/lee23f.html. ISSN: 2640-3498.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Arash Vahdat, and Weili Nie. Molecule generation with fragment retrieval augmentation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9798331314385.

Seul Lee, Seanie Lee, Kenji Kawaguchi, and Sung Ju Hwang. Drug discovery with dynamic goal-aware fragments. In *International Conference on Machine Learning*, pp. 26731–26751. PMLR, 2024b.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. Limits to depth-efficiencies of self-attention. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Anchen Li, Elena Casiraghi, and Juho Rousu. Chemical reaction enhanced graph learning for molecule representation. *Bioinformatics*, 40(10):btae558, October 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae558. URL https://doi.org/10.1093/bioinformatics/btae558.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. GeomGCL: Geometric Graph Contrastive Learning for Molecular Property Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4541–4549, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i4.20377. URL https://ojs.aaai.org/index.php/AAAI/article/view/20377. Number: 4.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. *Provably Powerful Graph Networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Eyal Mazuz, Guy Shtar, Bracha Shapira, and Lior Rokach. Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports*, 13(1):8799, May 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-35648-w. URL https://www.nature.com/articles/s41598-023-35648-w. Number: 1 Publisher: Nature Publishing Group.

Borja Medina, Alessandro Tibo, Jiazhen He, Jon Paul Janet, and Nicklas Österbacka. Diversity Beats Size Scaling for Chemical Language Models. *npj Drug Discovery*, 3(1):26, August 2026. ISSN 3005-1452. doi: 10.1038/s44386-026-00056-3.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33014602. URL https://ojs.aaai.org/index.php/AAAI/article/view/4384. Number: 01.

AkshatKumar Nigam, Robert Pollice, Gary Tom, Kjell Jorner, John Willes, Luca Thiede, Anshul Kundaje, and Alan Aspuru-Guzik. Tartarus: A benchmarking platform for realistic and practical inverse molecular design. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3263–3306. Curran Associates, Inc., 2023. doi: 10.52202/075280-0145.

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(4):796–804, 2024. doi: 10.1039/D4DD00019F. URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00019f. Publisher: Royal Society of Chemistry.

Noel O'Boyle and Andrew Dalke. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures, September 2018. URL https://chemrxiv.org/engage/chemrxiv/article-details/60c73ed6567dfe7e5fec388d.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, September 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0235-x. URL https://doi.org/10.1186/s13321-017-0235-x.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 11, December 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.565644. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.565644/full. Publisher: Frontiers.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter 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.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. Publisher: American Chemical Society.

Jerret Ross, Brian Belgodere, Samuel C. Hoffman, Vijil Chenthamarakshan, Youssef Mroueh, and Payel Das. GP-MoLFormer: A Foundation Model For Molecular Generation, April 2024. URL http://arxiv.org/abs/2405.04912. arXiv:2405.04912 [q-bio].

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, September 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. Publisher: American Chemical Society.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Martin Simonovsky and Nikos Komodakis. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, February 2018. URL http://arxiv.org/abs/1802.03480. arXiv:1802.03480 [cs] version: 1.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265. PMLR, June 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. ISSN: 1938-7228.

Ruslan N. Tazhigulov, Joshua Schiller, Jacob Oppenheim, and Max Winston. Molecular Fingerprints for Robust and Efficient ML-Driven Molecular Generation, November 2022. URL http://arxiv.org/abs/2211.09086. arXiv:2211.09086 [cs].

Morgan Thomas, Noel M O'Boyle, Andreas Bender, and Chris De Graaf. Augmented hill-climb increases reinforcement learning efficiency for language-based de novo molecule generation. *Journal of cheminformatics*, 14(1):68, 2022.

Benjamin I. Tingle, Khanh G. Tang, Mar Castanon, John J. Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S. Moroz, and John J. Irwin. ZINC-22-A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *Journal of Chemical Information and Modeling*, 63(4):1166–1176, February 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.2c01253. URL https://doi.org/10.1021/acs.jcim.2c01253. Publisher: American Chemical Society.

Alexius Wadell, Anoushka Bhutani, and Venkatasubramanian Viswanathan. Tokenization for Molecular Foundation Models. *Journal of Chemical Information and Modeling*, 66(3):1384–1393, February 2026. ISSN 1549-9596. doi: 10.1021/acs.jcim.5c01856.

Ye Wang, Honggang Zhao, Simone Sciabola, and Wenlu Wang. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules*, 28 (11):4430, January 2023. ISSN 1420-3049. doi: 10.3390/molecules28114430. URL https://www.mdpi.com/1420-3049/28/11/4430. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

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, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society.

Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science*, 10(34):8016–8024, August 2019. ISSN 2041-6539. doi: 10.1039/C9SC01928F. URL https://pubs.rsc.org/en/content/articlelanding/2019/sc/c9sc01928f. Publisher: The Royal Society of Chemistry.

Jiacheng Xiong, Zhaoping Xiong, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Graph neural networks for automated *de novo* drug design. *Drug Discovery Today*, 26(6):1382–1393, June 2021. ISSN 1359-6446. doi: 10.1016/j.drudis.2021.02.011. URL https://www.sciencedirect.com/science/article/pii/S1359644621000787.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, August 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b00959. URL https://doi.org/10.1021/acs.jmedchem.9b00959. Publisher: American Chemical Society.

Dong Xu, Qihua Pan, Sisi Yuan, Jianqiang Li, Zexuan Zhu, and Junkai Ji. Unveiling scaling behaviors in molecular language models: Effects of model size, data, and representation. *arXiv preprint arXiv:2601.22757*, 2026.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025.

Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, and Sung Ju Hwang. Hit and lead discovery with explorative rl and fragment-based molecule generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Chengxi Zang and Fei Wang. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 617–626, New York, NY, USA, August 2020. Association for Computing Machinery. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403104. URL https://dl.acm.org/doi/10.1145/3394486.3403104.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. Encoder-Decoder Gemma: Improving the Quality-Efficiency Trade-Off via Adaptation, April 2025.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 793–803, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330961. URL https://dl.acm.org/doi/10.1145/3292500.3330961.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Related Work

A.1 Representation of Molecules

Molecules are commonly depicted using structural diagrams, traditionally drawn with pen and paper, to represent bonds and atoms visually. However, in chem-informatics, more advanced representations are needed for the computational processing of molecular structures. In this context, "molecular representations" encompass any encoding of a chemical compound that can be employed for computational exploration of the chemical space. The current approaches to representing molecules can be broadly classified into four types: (i) Vector-based representations, (ii) Graph-based representations, (iii) 3D-based representations, and (iv) String-based representations.

Vector-based: Topological fingerprints, such as Extended Connectivity FingerPrints (ECFP) (Rogers & Hahn, 2010) and Molecular ACCess System (MACCS) (Durant et al., 2002), have traditionally been employed for substructure and molecule similarity searches. These fingerprints encode molecules as a sequence of bits in an identifier list, each denoting the presence or absence of a specific substructure. Although each molecular structure can be deterministically mapped to a fingerprint, the fingerprints are only partially invertible (Le et al., 2020), which limits their applicability in *de novo* molecule generation (Gómez-Bombarelli et al., 2018; Tazhigulov et al., 2022). Additionally, the fingerprints can be augmented with 2D molecular descriptors such as Molecular Weight, QED score, and Number of Aromatic Rings to help impose specific constraints on the generated molecules and make them more aligned with desired chemical properties or biological activity.

Graph-based: A 2D molecular graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes (atoms) and $\mathcal{E}$ is the set of edges (bonds). The type of atoms and edges can be represented using a feature matrix $\boldsymbol{X}$. Graph Neural Networks (GNNs) have been used to learn the representations of molecules (Kipf & Welling, 2017; Xu et al., 2019; Xiong et al., 2021) for tasks such as reaction prediction, property prediction and drug discovery. The initial frameworks for learning molecule representation used Message Passing Neural Networks (MPNNs) to compute the atom embedding based on neighbourhood information capturing local interaction effects (Gilmer et al., 2017). Although many variants of GNNs have been proposed (Morris et al., 2019; Maron et al., 2019; Zhang et al., 2019; Brockschmidt, 2020; Xiong et al., 2020; Li et al., 2024), challenges remain in terms of higher-order expressivity, scalability, and computational cost.

3D-based: While using Graph Neural Networks (GNNs) on 2D molecular graphs is convenient and seem to be the obvious choice, the resulting representations often overlook crucial spatial information, such as the spatial direction and torsion angles between atoms (Guo et al., 2023). Recent advancements in molecule representation learning have focused on integrating 3D coordinate information into 2D molecular graphs (Luo et al., 2021; Fang et al., 2022; Li et al., 2022). Uni-Mol (Zhou et al., 2023) introduced a pretraining framework capable of directly utilizing 3D positions as inputs and outputs. However, a significant challenge in this approach is the existence of multiple low-energy conformations for a given molecule. These conformations are not easily accessible and are particularly difficult to compute, especially for large molecules and the vast chemical space, which spans billions of possible molecules.

String-based: 2D molecular structures can also be encoded as linear notations, which use specialized languages to represent molecular structures and compositions in chemistry. The earliest example of such a molecular language was developed in the 1980s by Weininger (1988). The SMILES (Simplified Molecular Input Line Entry System) notation encodes atoms, bonds, and connectivity patterns using ASCII strings, where atoms are represented by characters (e.g., 'C' for carbon, 'N' for nitrogen) and bonds by special characters (e.g., '-' for a single bond, '=' for a double bond, '#' for a triple bond). However, the syntax rules and restrictive grammar of SMILES can result in many invalid molecules during parsing, even when the string appears to represent a plausible molecular structure. To address some of these limitations, DeepSMILES (O'Boyle & Dalke, 2018) was introduced, which avoids the issue of unbalanced parentheses by using only closing parentheses, where the number of parentheses indicates the branch length. More recently, SELFIES (Self-Referencing Embedded Strings) (Krenn et al., 2022) was developed as a linear notation that is 100% robust; every SELFIES string corresponds to a valid molecule, even for entirely random strings. Additionally, SAFE (Sequential Attachment-based Fragment Embedding) (Noutahi et al., 2024) introduced a framework for fragment-constrained molecule generation tasks while maintaining compatibility with existing SMILES parsers.

String-based molecular representations offer a computationally efficient and scalable approach to exploring the vast chemical space using unlabelled data without relying on additional information such as 3D geometry or complex optimization techniques. Despite their simplicity, these methods capture essential chemical information and structural features, making them valuable tools in computational chemistry and drug discovery.

A.2 Deep Generative Models for *De Novo* Molecule Generation

Deep generative models have become a key approach for *de novo* molecule generation, facilitating the discovery of novel compounds by capturing complex patterns within the vast chemical space. Numerous methods have emerged, each focused on different molecular representations and assembly strategies (Olivecrona et al., 2017; Jin et al., 2018; Polykovskiy et al., 2020; Jin et al., 2020; Bagal et al., 2022; Eckmann et al., 2022; Jo et al., 2022; Irwin et al., 2022; Fang et al., 2024; Lee et al., 2023; Yang et al., 2021; Ross et al., 2024).

Assembly Methods: Early approaches (Gómez-Bombarelli et al., 2018; Jin et al., 2018; Winter et al., 2019; Tazhigulov et al., 2022) primarily utilized Variational Autoencoders (VAEs) to transform SMILES representations into a continuous latent space, followed by sampling and decoding to generate discrete molecular structures. Graph-based generative models have emerged as a natural extension, directly leveraging the molecular graph structure where atoms are represented as nodes and bonds as edges (Cao & Kipf, 2022). GraphVAE (Graph Variational Autoencoder) (Simonovsky & Komodakis, 2018) encodes and decodes molecules using edge-conditioned graph convolutions. MoFlow (Zang & Wang, 2020), a flow-based graph generative model learns invertible mappings between molecular graphs and their latent representations. Graph generation approaches utilize small molecular building blocks such as atoms and their performance degrades significantly for larger molecules. To tackle this problem recent works employ significantly larger and more flexible graph motifs as basic building blocks (Kuznetsov & Polykovskiy, 2021). In parallel, 3D molecule generation has gained substantial attention, particularly through the use of diffusion models. G-Schnet (Gebauer et al., 2019), for instance, utilizes an autoregressive process to iteratively sample atoms and bonds in 3D space. Similarly, inspired by the success of diffusion models in other domains (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kong et al., 2023), Hoogeboom et al. (2022) proposed an equivariant diffusion model for generating novel 3D molecular structures.

Optimization Methods: Molecule optimization involves navigating an immense and complex chemical space, which requires sophisticated algorithms capable of efficiently searching for and generating molecules with optimal characteristics. Several computational approaches have been developed to tackle this challenge, each with distinct mechanisms for exploring the design space and optimizing molecular properties. Reinforcement Learning treats molecule optimization as a sequential decision-making problem. In this context, the state typically represents a partially generated molecule, and actions correspond to modifications at the graph or string level. The reward function is based on the properties of the generated molecules, guiding the model toward desirable outcomes. Bayesian Optimization operates by learning a continuous latent space of molecular representations, optimizing target properties by navigating through this latent space. Genetic algorithms, inspired by natural evolutionary processes, explore the chemical space through operations such as mutation and crossover applied to a pool of candidate molecules, promoting diversity and exploration. Gradient ascent methods, on the other hand, estimate the gradient of a molecular property across the chemical space and use backpropagation to optimize molecular structures. Hill Climbing is an iterative optimization method with high-performing molecules from previous rounds incorporated into the training data to refine the generative model progressively.

While graph-based and 3D deep generative models have made significant strides in generating molecular structures, recent advances in natural language processing have opened new possibilities for *de novo* molecule generation and optimization. Large Language Models (LLMs) offer a novel approach for navigating the large chemical space, presenting new opportunities for optimizing molecular properties in a scalable and computationally feasible manner. Combining these models with traditional optimization techniques can further enhance the search for *de novo* molecules with desired properties, marking a significant step forward in drug discovery.

A.3 Language Models in Molecule Generation

LLMs can effectively model sequential data and have shown remarkable proficiency in understanding and generating human language (Dubey et al., 2024; Jiang et al., 2023; Groeneveld et al., 2024). These architectures are now being repurposed to explore and generate molecular structures. When treated as sequences of tokens, 1D molecular representations inherently encode chemical information, including 2D bonding topology patterns, while LLMs further enhance this by learning to generate diverse molecular structures. These models leverage unlabelled molecular data from across the chemical space, offering a novel approach to *de novo* molecule generation and broadening the potential for chemical discovery. MolGPT (Bagal et al., 2022) employs a decoder-only transformer architecture, inspired by GPT, to predict SMILES token sequences for molecular generation. Building on these advancements, models like cMolGPT (Wang et al., 2023) have been developed to generate target-specific compounds by incorporating conditional training for property optimization. Taiga (Mazuz et al., 2023) extends this approach by employing a two-stage framework: first, the model treats molecular generation as a language modeling task by predicting the next token in SMILES strings. Subsequently, reinforcement learning (RL) is applied to optimize simple chemical properties such as QED (Quantitative Estimate of Drug-likeness) and logP (Octanol-Water Partition Coefficient). MolGen (Fang et al., 2024), in contrast, utilizes an encoder-decoder BART architecture, focusing on generating chemically valid molecules through the SELFIES notation. It incorporates a chemical feedback mechanism to align generative probabilities with real-world chemical preferences. SAFE-GPT (Noutahi et al., 2024) introduces a new line notation and trains a GPT-like model on 1.1 billion SAFE notations, demonstrating versatile and robust performance in both *de novo* and fragment-constrained molecule generation tasks.

Tokenizers, which convert raw text sequences into tokens, are a critical component of modern language models. In CLMs, token vocabularies are often constructed using a predefined regular expression developed by Schwaller et al. (2019), which splits SMILES strings into relevant tokens representing each atom (e.g., 'C' for carbon, 'N' for nitrogen). Less commonly, subword tokenization algorithms such as Byte Pair Encoding (BPE) (Sennrich et al., 2016) or Unigram (Kudo, 2018) are employed, sometimes combined with atomwise pretokenization and BPE. Given that tokenizer design impacts every stage of the modeling pipeline, this study explores the effects of using learned tokenization methods versus hand-crafted approaches.

B Dataset Preparation and Design Axes Details

We pretrain on four public molecular corpora of increasing size and diversity. MOSES (Polykovskiy et al., 2020) contains about 1.9M drug-like molecules with a restricted atom set. ChEMBL (Gaulton et al., 2012) contains about 1.9M bioactive molecules drawn from the medicinal-chemistry literature. PubChem (Kim et al., 2016) contributes about 100M molecules after filtering, and forms the reference corpus for our scaling sweep. ZINC-22 (Tingle et al., 2023) is the largest library, and we sample 1.5B molecules proportionally across its heavy-atom-count and logP tranches so that the sample preserves the coverage of the full make-on-demand space. Tab. B1 summarises the corpora.

B.1 Standardisation

Each corpus passes through the same offline pipeline. We canonicalise every SMILES string with RDKit, drop entries that RDKit cannot parse, strip salts and normalise charges, and deduplicate on the canonical string in a single streaming pass so that no molecule appears twice. We then convert the surviving molecules to SELFIES and SAFE, and build atom-level and byte-pair tokenisers for each representation from the same underlying molecules. This keeps every representation and tokeniser aligned to one shared set of compounds.

B.2 Leakage control

To keep the evaluation targets unseen during pretraining, we assemble an active set of about 21,000 known actives for the PMO and docking targets together with the Tartarus benchmark molecules (Nigam et al., 2023). We compute 2,048-bit ECFP fingerprints of radius 2 and remove from every corpus any molecule whose Tanimoto similarity to an active set member exceeds 0.5. The active set is never used for training and later serves as the validation probe set in Appx. D.

B.3 EVALUATION SPLITS

From the filtered molecules we hold out four splits of increasing distribution shift. The random split is a uniform sample and serves as the in-distribution reference. The scaffold split holds out entire Bemis-Murcko scaffolds. The molecular-weight split holds out the heaviest molecules. The maximum-dissimilarity split holds out the fingerprint clusters that sit farthest from the rest of the corpus. Each split contains 175,000 molecules and is scored on its own, so a single model is measured against progressively harder generalisation rather than a single validation set.

B.4 DESIGN GRID

The reference configuration is a Qwen3 decoder-only model trained on PubChem SMILES with atom-level tokenisation, and it spans the full range from 0.5M to 1B parameters. Each remaining axis is varied one at a time around this reference at matched model size and training tokens, as summarised in Tab. B2. This one-axis-at-a-time design keeps the number of pretraining runs tractable while still isolating the effect of each design choice on validation set loss, learned representations and goal-directed generation.

Table B1: Pretraining corpora. Token counts use the atom-level SMILES tokeniser. Sizes are reported after standardisation and leakage filtering.

Corpus	Molecules	Median heavy atoms	Median SMILES tokens	Character
MOSES	1.9M	21	28	Narrow, drug-like, restricted atoms
ChEMBL	1.9M	27	48	Bioactive, literature-derived
PubChem	100M	24	47	Broad drug-like (reference corpus)
ZINC-22	1.5B (sampled)	29	52	Make-on-demand, largest coverage

Table B2: Experimental design grid. Each axis is varied around the PubChem–SMILES–atomwise Qwen3 reference at matched model size and training tokens.

Axis	Variants	Model sizes
Reference	PubChem, SMILES, atom-level, Qwen3	0.5M, 5M, 32M, 157M, 300M, 1B
Representation	SELFIES, SAFE	0.5M, 5M, 32M
Tokenisation	Byte-pair encoding	0.5M, 5M, 32M, 157M, 300M, 1B
Architecture	T5Gemma (encoder–decoder)	0.5M, 5M, 32M, 157M, 300M, 1B
Pretraining corpus	MOSES, ChEMBL, ZINC-22	0.5M, 5M, 32M, 157M, 300M, 1B

B.5 DATASET DIVERSITY

To ensure the diversity of the dataset, we confirmed that it contains a broad range of molecular structures, which is crucial for generating a wide variety of valid, novel, and unique molecules as illustrated in Figure B1. The left plot shows the distribution of molecular lengths, tokenized using an atomwise tokenizer, demonstrating variation in molecule size across batches. The right plot displays the Tanimoto similarity between molecule pairs, showcasing the structural diversity in each batch an essential factor for robust molecular generation.

The four pretraining corpora also differ in their molecular properties, which helps explain why they are not equally hard to model. Figure B2 compares six properties across MOSES, ChEMBL, PubChem and ZINC-22, each estimated from 6,000 sampled molecules. MOSES, ChEMBL and PubChem are drug-like and overlap closely across most properties, with QED concentrated above 0.7 and logP around 2 to 3. ZINC-22 is the outlier. It is less drug-like, with QED peaking near 0.5 rather than above 0.7, and its molecules are heavier and more tightly concentrated in molecular weight, with more rings and more rotatable bonds. This broader make-on-demand chemistry matches the corpus ordering of validation set loss, where the narrow, drug-like MOSES distribution is the easiest to model and the broad ZINC-22 space is the hardest.

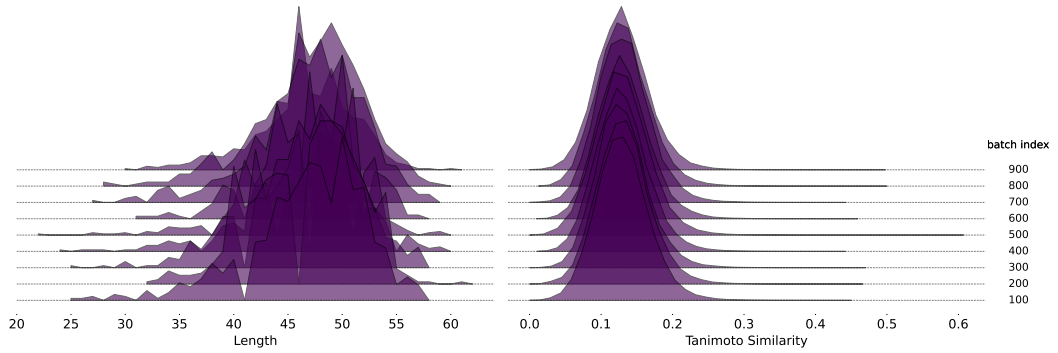


Figure B1: Diversity of the data across batches. The left plot shows the distribution of molecular lengths, tokenized using an atomwise tokenizer, indicating variation in molecule length within each batch. The right plot shows Tanimoto similarity between molecule pairs, demonstrating structural diversity in each batch.

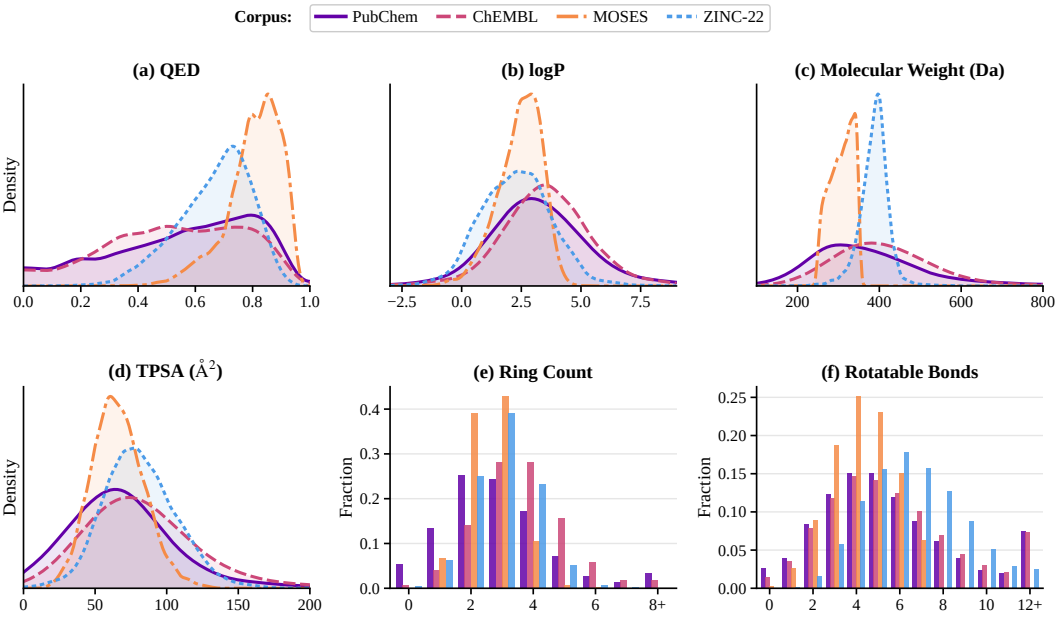


Figure B2: Molecular-property distributions of the four pretraining corpora. Each corpus is represented by 6,000 molecules sampled uniformly across all of its shards, with properties computed by RDKit on the largest fragment. **(a–d)** Kernel density estimates of QED, logP, molecular weight and TPSA. **(e, f)** Fraction of molecules by ring count and number of rotatable bonds, with the last bin including all larger values.

B.6 Property Distributions

The distribution of properties serves as a valuable tool for visually evaluating the generated structures. We present a kernel density estimation of these distributions and calculate the Wasserstein-1 distance to compare the distributions of the generated and reference datasets. We use the following properties:

1. **Quantitative Estimation of Drug-likeness (QED)**: A metric derived from medicinal chemistry principles that quantifies the drug-likeness of molecules on a scale from 0 to 1.
2. **Synthetic Accessibility (SA)**: A measure of a molecule's synthesizability, calculated based on the contributions of molecular fragments. The metric ranges from 10 (difficult to synthesize) to 2 (easily synthesizable).

3. **Octanol-Water Partition Coefficient (logP)**: Represents the ratio of a compound's concentration in the octanol phase to its concentration in the aqueous phase in a two-phase octanol/water system, serving as an indicator of solubility.
4. **Molecular Weight (MW)**: Evaluates whether the generated set is biased toward heavier or lighter molecules, computed as the sum of atomic weights.
5. **Topological Polar Surface Area (TPSA)**: Estimated based on functional group contributions from a database of substructures, this metric reflects lipid solubility and molecular polarity. Higher TPSA values indicate reduced absorption and distribution within the body.
6. **Bertz Complexity**: A graph-theoretical measure that quantifies molecular complexity using structural invariants and information-theoretic principles.
7. **Number of Rings (NumRings) and Rotatable Bonds**: Represents the number of independent closed-ring structures and rotatable bonds within a molecule, which are essential for analyzing molecular topology and are commonly used in cheminformatics for compound classification and comparison.

The kernel density estimation plots in Figure B3 indicate that NovoMolGen successfully generates molecules whose property distributions align closely with those of both the random split and the scaffold split. Additionally, a lower Wasserstein-1 distance is observed across all properties, further demonstrating the model's ability to replicate the reference distributions. The training dataset contains a higher proportion of molecules with favorable drug-like properties, including higher drug-likeness scores (QED > 0.6), greater synthetic accessibility (SA < 4), optimal solubility ($1 <$ logP < 4), and an ideal molecular weight range ($300 <$ MW < 500). By closely matching this distribution, NovoMolGen generates molecules with an increased likelihood of exhibiting drug-like characteristics. Furthermore, the model effectively captures molecular topology, as reflected in the distributions of NumRings, Rotatable Bonds, and Bertz Complexity. These results highlight the potential of NovoMolGen in generating chemically relevant and synthetically accessible molecules suitable for drug discovery applications.

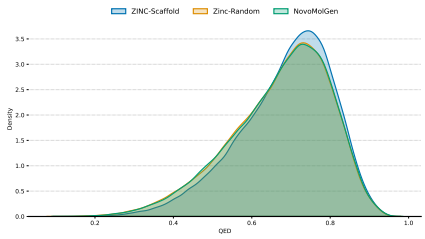
C Pretraining Configuration

We implement the Qwen3 decoder-only and T5Gemma encoder–decoder models in the HuggingFace Trainer framework with FlashAttention kernels (Dao et al., 2022) for efficiency, and train in mixed precision with bfloat16. We use the AdamW optimiser with a peak learning rate of 6×10^{-4} and a warmup–stable–decay (WSD) schedule. The learning rate warms up over the first 2% of steps, is held constant through the stable phase, and each reported checkpoint is taken after a short linear cooldown to 6×10^{-5} branched from the stable trajectory. This matches the schedule used for the IsoFLOP profiles in Appx. F.

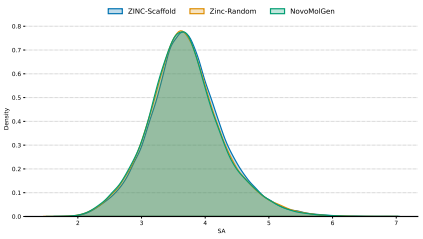
We use token-aware batching of approximately 25,000 non-padding tokens per optimiser step, with gradient accumulation and per-device batch size chosen to fit 4 NVIDIA A100 GPUs (80 GB each). Weight decay is 0.01, gradient clipping uses a maximum norm of 1.0, and AdamW uses $\beta_1 = 0.9$ and $\beta_2 = 0.95$. Design-axis comparisons match model size and training tokens at $D = 20N$ rather than raw compute. For T5Gemma we estimate compute as $C \approx 3ND$ because scored tokens pass only through the decoder, so a T5Gemma model matched in size and tokens to a Qwen3 model uses roughly half the training compute. The architectural configurations are summarised in Tab. C1, following the width-to-depth ratio of Levine et al. (2020).

Table C1: NovoMolGen Configurations

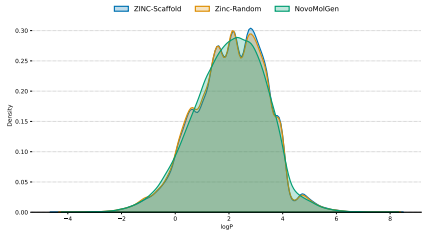
Components	NovoMolGen-32M	NovoMolGen-157M	NovoMolGen-300M
Attention Heads	8	10	12
Hidden Layers	12	24	32
Hidden Size	512	640	768
Intermediate Size	1024	2560	3072



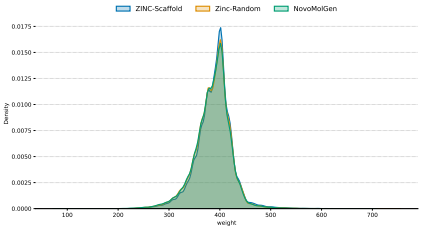
(a) QED Distribution



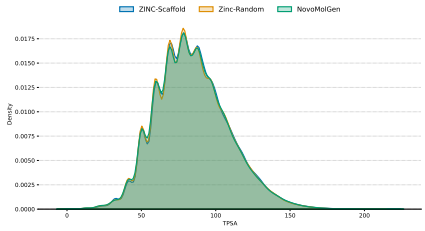
(b) SA Distribution



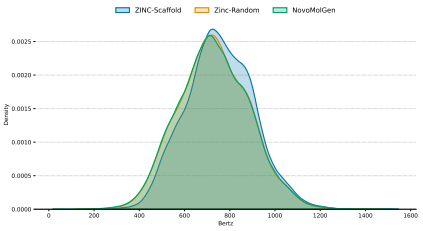
(c) logP Distribution



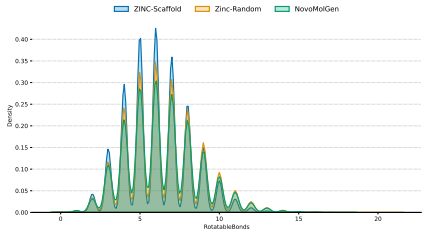
(d) MW Distribution



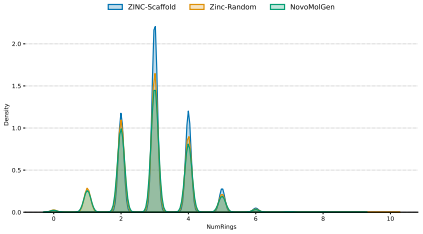
(e) TPSA Distribution



(f) Bertz Complexity Distribution



(g) Rotatable Bonds Distribution



(h) NumRings Distribution

Figure B3: Molecular-property distributions for the random and scaffold reference splits (175,000 molecules each) and a generated set from **NovoMolGen-SMILES-BPE-32M** (100,000 molecules). The properties include QED, SA, logP, MW, TPSA, Bertz complexity, number of rotatable bonds, and number of rings.

D Mechanistic Probing Methodology

D.1 Layer-wise Probing

To quantify the explicit encoding of chemical properties within NovoMolGen's internal representations, we employed the linear probing method. This technique involves training simple linear classifiers or regressors on the model's frozen activation vectors to predict specific external properties, where high performance indicates that the information is linearly accessible at that specific depth. Let $h_\ell^{(i)} \in \mathbb{R}^d$ denote the activation vector of the i-th molecule at layer ℓ. We trained a separate linear probe for each layer and property.

For continuous regression tasks (e.g., LogP, QED, Molecular Weight), we utilised **Ridge Regression**. The probe learns weights $W \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$ to minimise the regularised squared error:

$$\mathcal{L}_{reg} = \sum_i ||y^{(i)} - (W^T h_\ell^{(i)} + b)||^2 + \alpha||W||_2^2 \tag{1}$$

where activations were standardised (zero mean, unit variance) prior to training, and the regularisation strength α was selected by five-fold cross-validation on the training split. For categorical classification tasks, such as validity checks, we utilised **Logistic Regression**, minimising the cross-entropy loss with L_2 regularisation (using the multinomial formulation for multiclass problems).

We evaluated the model on two distinct task categories. First, for **Property Prediction**, we curated a set of over 30 chemical descriptors, ranging from simple physicochemical properties (e.g., Molecular Weight, NumRings) to complex bioactivity scores (e.g., Docking Scores, QED), with ground-truth values computed using RDKit and standard docking simulations. Second, for **Validity Detection**, we assessed the model's ability to distinguish valid chemical syntax from noise via a binary classification task. The positive class consisted of activations from valid SMILES sequences, while the negative class consisted of "corrupted" activations extracted from structurally invalid or randomised sequences.

We use a fixed 80/20 train/test split and report performance on the validation set 20%, with the regularisation strength selected by five-fold cross-validation on the training portion. For regression we report the coefficient of determination (R^2), mean absolute error (MAE) and Spearman rank correlation (ρ), and for classification we report accuracy and F_1. All probes use the `scikit-learn` library (Pedregosa et al., 2011), with input activations frozen and detached from the computational graph so that probing does not update the language model parameters.

D.2 Sparse Autoencoders Training

To investigate the internal representations of NovoMolGen, we trained Sparse Autoencoders (SAEs) on the residual stream activations of specific transformer layers. We use the **BatchTopK** sparse autoencoder (Bussmann et al., 2024), which enforces sparsity across a whole batch rather than per input. For a batch X of B activation vectors it retains the kB largest activations over the entire batch, so the number of active features can vary between molecules while averaging k per input. This improves the trade-off between reconstruction fidelity and sparsity compared with a fixed L_1 penalty. The SAE has an encoder $W_{enc} \in \mathbb{R}^{d_{model} \times d_{dict}}$ and a decoder $W_{dec} \in \mathbb{R}^{d_{dict} \times d_{model}}$, and the forward pass is

$$a(x) = \text{ReLU}(xW_{enc} + b_{enc}) \tag{2}$$

$$f(X) = \text{BatchTopK}_k\big(a(X)\big) \tag{3}$$

$$\hat{x} = f(x)W_{dec} + b_{dec} \tag{4}$$

where BatchTopK_k keeps the kB largest entries across the batch X and zeroes the rest, and $d_{dict} = M \times d_{model}$ is the dictionary size.

Activations were extracted from different layers of (early, middle and late) **NovoMolGen**. To ensure diverse, uncorrelated training batches, we employed an activation-buffering strategy. The source data consisted of the active set molecules, which were tokenised using the model's native tokeniser (BPE or AtomWise) with a fixed context length of 64 tokens. We utilised a shuffled `ActivationBuffer` with a capacity of 3×10^4 contexts, continuously refreshed with batches of 512 sequences from the language model to prevent distributional shift during training.

All SAEs were trained using the settings detailed in Table D1. We employed an auxiliary loss ($\mathcal{L}_{aux}$) to mitigate "dead features" (neurons that never activate).

To assess the quality of the learned dictionary, we evaluated each SAE on a validation split of the active set using four metrics: (1) **Fraction of Variance Explained (FVE)**, $1 - \|\hat{x} - x\|^2/\|x\|^2$; (2) L_0 **norm**, the average number of active features per token (target $\approx k$); (3) **Fraction Alive**, the fraction of dictionary features that fire at least once; and (4) **CE Loss Recovered**, the fraction of the language-modelling loss recovered when the residual stream is replaced by the SAE reconstruction (Cunningham et al., 2023). To relate features to chemistry, we correlate each live feature with

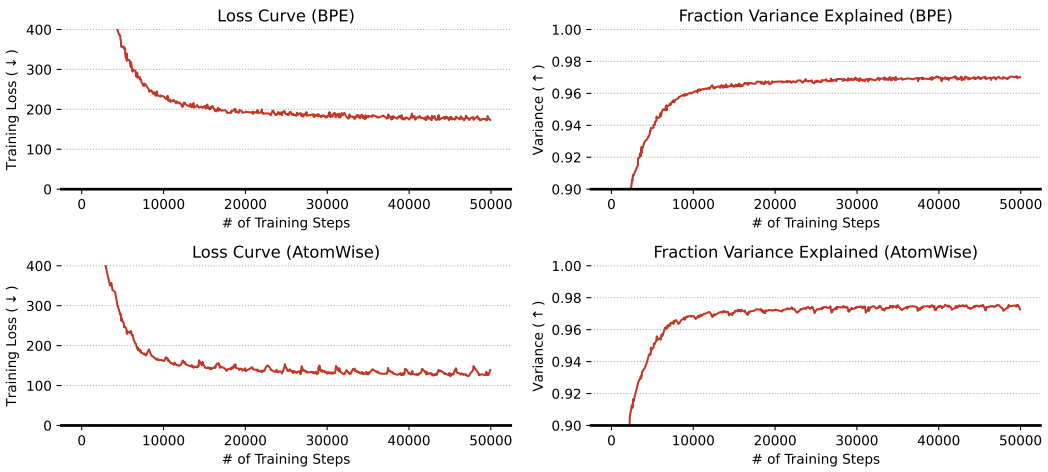


Figure D1: Training curves for loss and fraction of variance explained for NovoMolGen-SMILES-BPE-32M at checkpoint 75000 and layer 6.

Table D1: Hyperparameters for Sparse Autoencoder Training.

Hyperparameter	**Value**	**Description**
Dictionary Multiplier	$16\times$	Expansion factor ($16 \times 512 = 8192$ features)
Sparsity (k)	$\approx 2\%$	Top-k active features allowed
Optimizer	Adam	Standard configuration
Learning Rate	3×10^{-4}	Linear warmup over first 2,500 steps
Batch Size (SAE)	4,096	Activation vectors per optimization step
Training Steps	50,000	Total updates ($\approx 2 \times 10^8$ tokens)
Dead Feature Loss (α)	$1/32$	Coefficient for auxiliary loss
Threshold Adaption (β)	0.999	EMA decay for threshold adaptation

the 36 property targets used for probing, namely the physicochemical descriptors, PMO objectives and docking scores. For each molecule we max-pool a feature's activations over its tokens to obtain one value per molecule, and take its Pearson correlation with each property. **Monosemanticity** is the mean over live features of the largest absolute property correlation, and we call a feature property-associated when this value exceeds 0.4. The training curves in Figure D1 show that the fraction of variance explained rises quickly and saturates early in training.

To empirically validate the hypothesis that Sparse Autoencoder (SAE) features explicitly disentangle abstract chemical concepts better than the model's raw basis directions, we conducted a controlled substructure-detection experiment. We curated a set of 10 common chemical functional groups defined by SMARTS patterns and generated binary ground-truth labels $y \in \{0,1\}^N$ for a random subset of $N = 10,000$ valid molecules from the active set using RDKit. For each molecule, we extracted the internal activations from the target transformer layer and mean-pooled them over non-padding positions, yielding one activation vector per molecule. We then trained a Logistic Regression classifier with class balancing and standard scaling, once on the SAE latents and once on the residual-stream neurons, to predict the presence of each substructure. Performance was evaluated by validation set F_1-score with 5-fold stratified cross-validation to guard against class imbalance. We report the F_1 of the probe on SAE latents against the probe on neurons in Tab. D2, and read a positive difference as evidence that the concept is more linearly accessible in the SAE dictionary. The SAE latents win for seven of the ten groups, and where the neurons win the margin is negligible, so the wider dictionary offers a modest but consistent gain rather than a decisive one.

Table D2: Validation set F_1 for detecting the ten most common functional groups from a linear probe on SAE latents against a probe on residual-stream neurons (middle layer of the 32M PubChem–SMILES–atomwise reference at the final checkpoint, averaged over folds). SAE latents give the higher F_1 for seven of the ten groups, with gains of up to about 0.04. Where neurons win, the margin is within 0.003. **Bold** indicates the higher score.

Functional Group	SAE latents	Neurons	Δ
Carbonyl	**0.999**	0.982	+0.017
Carbonyl (non-acid)	**0.997**	0.965	+0.032
Bicyclic	**0.997**	0.984	+0.013
Ether	**0.995**	0.968	+0.027
Secondary amine	**0.993**	0.963	+0.030
Amide	**0.993**	0.960	+0.033
Aniline	**0.986**	0.952	+0.034
Benzene	0.991	**0.992**	−0.001
Aromatic N	0.990	**0.993**	−0.003
Tertiary amine	0.988	**0.989**	−0.001

E FINE-TUNING METHODOLOGY

The agent is trained to minimize the cost function $J(X_k)$ over the subset of top-k molecules, X_k, defined as:

$$J(X_k) = \frac{1}{k} \sum_{x \in X_k} \left[\log P_{\text{prior}}(x) - \log P_{\text{agent}}(x) + \sigma \cdot s(x)\right]^2 \tag{5}$$

where σ is a scaling factor that controls the influence of the reward $s(x)$. The reward is the task score mapped to $[0, 1]$. For PMO tasks it is the benchmark oracle score, and for docking it is the QuickVina2 docking score min-max scaled to $[0, 1]$ across the batch. This loss balances the generation of high-reward molecules against adherence to the chemically valid distribution of the prior. We add a penalty term,

$$J_p(X_k) = \frac{-1}{\frac{1}{k}\sum_{x \in X_k} \log P_{\text{agent}}(x)},$$

which grows as the agent becomes overconfident and its mean log-likelihood approaches zero, and shrinks as that likelihood decreases. Minimising it therefore discourages the agent from collapsing onto a narrow set of very high likelihood molecules, which preserves exploration and stabilises training. The final objective combines these components,

$$J_\theta = J + \lambda \cdot J_p, \tag{6}$$

where λ weights the penalty term. To improve sample efficiency we use an experience replay buffer (Guo & Schwaller, 2024) that stores the 100 top-performing molecules found during fine-tuning. It is pre-initialised with the 100 highest-reward ZINC-250K molecules per task (Irwin et al., 2012). These seed molecules are scored once before optimisation and are not counted against the per-task oracle budget, and they are excluded from the reported novel candidates so they cannot inflate the downstream metrics.

To optimize our models for goal-directed design, we performed a systematic hyperparameter search using a REINVENT-inspired Hill Climbing framework. We ran a separate search for each of the two task categories. For multi-property optimization we tuned on the **Perindopril_MPO** and **Zaleplon_MPO** tasks from the PMO benchmark, and for protein-ligand docking we tuned on the **fa7** and **jak2** targets. Within each category, settings were selected by the aggregated score over its two tasks, and each configuration was evaluated over three random seeds.

The hyperparameter space explored includes the following, with the exact grid varying slightly across representation and size variants:

- **Penalty Coefficient** (λ): $[10, 100, 500, 2000]$
- **Batch Size**: $[32, 64, 128]$
- **Sigma** (σ): $[500, 1000, 2000, 2500]$

- **Learning Rate** (lr): $[1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}]$
- **Fraction of Top-k Molecules**: $[0.1, 0.5]$

The default value for penalty coefficient from the REINVENT implementation (5000) proved unsuitable for our models, likely due to differences in the log-likelihood scale of generated molecules, leading to unstable training and a high proportion of invalid solutions.

We evaluated 96 configurations, each on both task categories over three seeds, for **576 runs** in total, where a run is one configuration evaluated on one task category for one seed ($96 \times 2 \times 3 = 576$). Our model could not generate any molecules for the **Valsartan_SMARTS** task, necessitating its exclusion from our analysis. This limitation stems directly from our choice of pre-training data, which lacks the specific chemical substructures required by the task's SMARTS pattern. We note that this is a data-dependent constraint, and other models may succeed where ours could not. For instance, a method like f**-RAG** , which builds its fragment vocabulary from the `UniChem` database, would likely be able to generate valid solutions provided that this database contains the relevant chemical motifs. The parallel coordinates plot are shown in Figures E1 to E10.

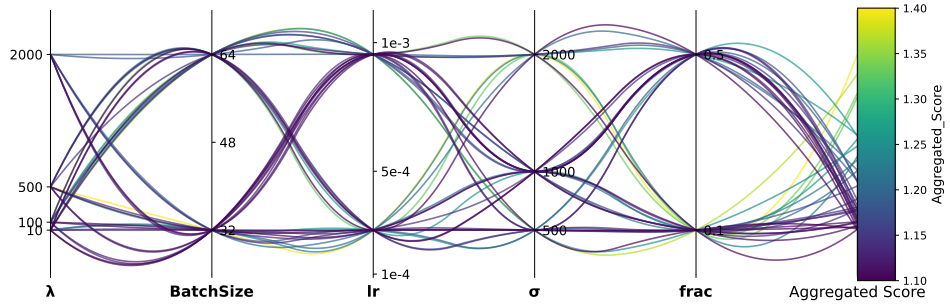


Figure E1: Parallel Coordinates Plot for the **SMILES (AtomWise)** model (32M), showing the importance and effects of hyperparameters on the Aggregated Score.

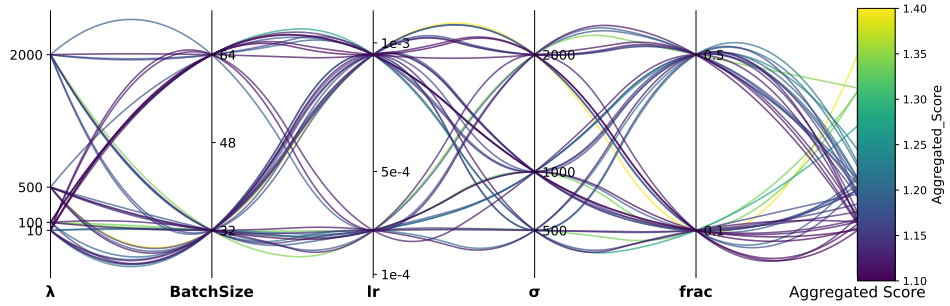


Figure E2: Parallel Coordinates Plot for the **SMILES (BPE)** model (32M).

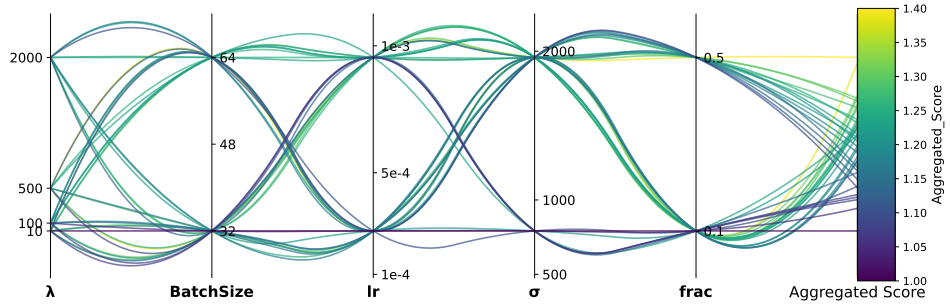


Figure E3: Parallel Coordinates Plot for the **SAFE (AtomWise)** model.

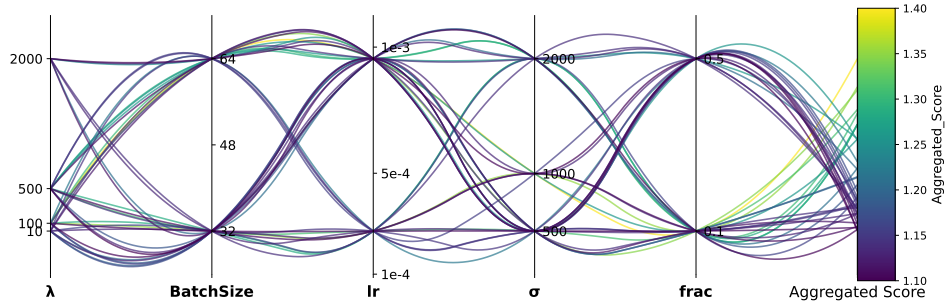


Figure E4: Parallel Coordinates Plot for the **SAFE (BPE)** model.

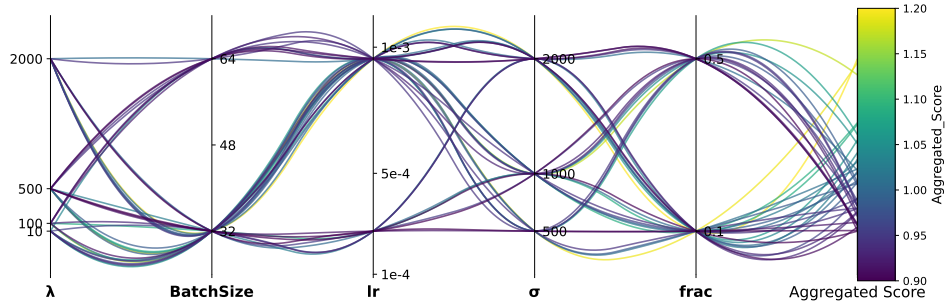


Figure E5: Parallel Coordinates Plot for the **SELFIES (AtomWise)** model.

F ISOFLOP EXPERIMENTS

We locate the compute-optimal frontier with IsoFLOP profiles, which compare model sizes at a fixed compute budget $C \approx 6ND$. Training a separate model to convergence at every (N, C) pair would be prohibitively expensive, so we use a single warmup–stable–decay trajectory per model size. Each model runs one long warmup–stable phase, and every IsoFLOP point is measured after a short linear cooldown branched from that phase at the target token count. This gives a well-converged checkpoint at each budget without retraining from scratch and keeps the grid affordable.

Tab. F1 lists the five compute budgets and the model sizes evaluated near each budget's optimum. Following Hoffmann et al. (2022), we fit a parametric loss surface

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{7}$$

to the validation set losses of the reference models, then minimise it at fixed $C = 6ND$ to read the loss-minimising size $N_{\text{opt}}(C)$ at each budget. The minima follow an approximate power law, $N_{\text{opt}} \propto C^{0.69}$ and $D_{\text{opt}} \propto C^{0.31}$, with a 68% bootstrap interval on the parameter exponent of 0.59

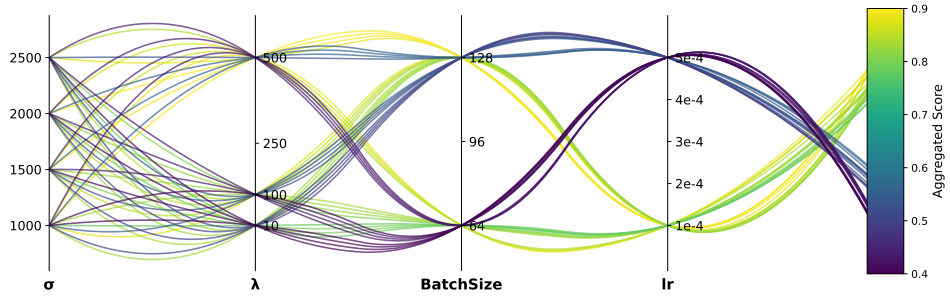


Figure E6: Parallel Coordinates Plot for the **SELFIES (BPE)** model.

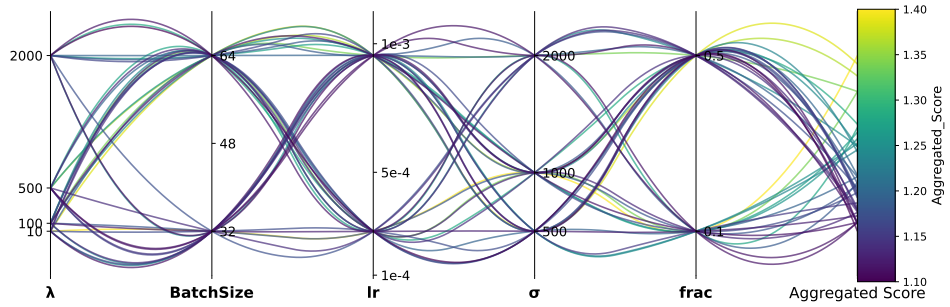


Figure E7: Parallel Coordinates Plot for the **SMILES (AtomWise)** model (157M).

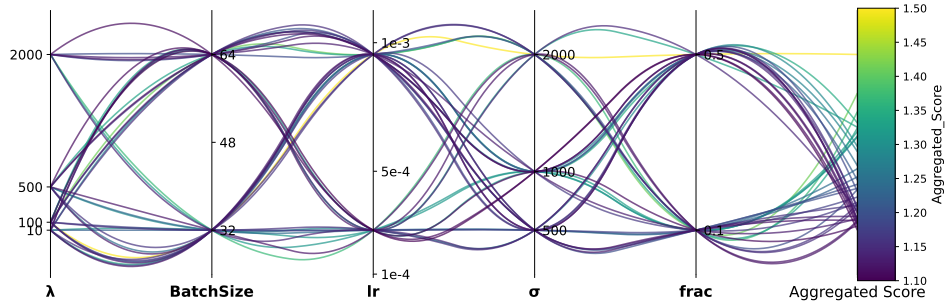


Figure E8: Parallel Coordinates Plot for the **SMILES (BPE)** model (157M).

to 0.76 obtained by resampling the training trajectories 1,000 times. Across the grid the compute-optimal size grows from about 2.5M parameters at 3×10^{15} FLOPs to about 700M at 10^{19} FLOPs, while the optimal token count grows more slowly, consistent with an exponent below one-half.

The optimal token-to-parameter ratio drifts along the frontier, so we adopt a single practical rule for all reported comparisons. Training every model to $D = 20N$ stays within 1% of the fitted optimal loss across the budget range of the grid, 3×10^{15} to 10^{19} FLOPs, so we train and compare all sizes at $D = 20N$ and refer to these checkpoints as compute-optimal. This is a near-optimal practical rule rather than the exact optimum, since $D = 20N$ corresponds to $N \propto C^{0.5}$ whereas the fitted allocation is $N_{\text{opt}} \propto C^{0.69}$. For the smallest and largest sizes in the reported sweep (0.5M and 1B), which fall just outside the fitted budget range, the near-optimality is inferred from the fitted loss surface rather than from bracketing IsoFLOP points.

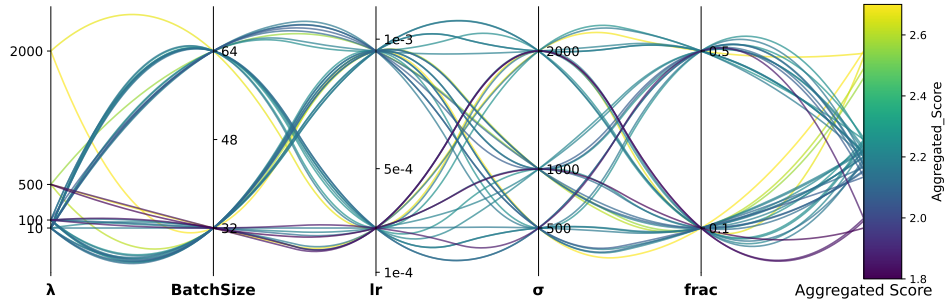


Figure E9: Parallel Coordinates Plot for the **SMILES (AtomWise)** model (300M).

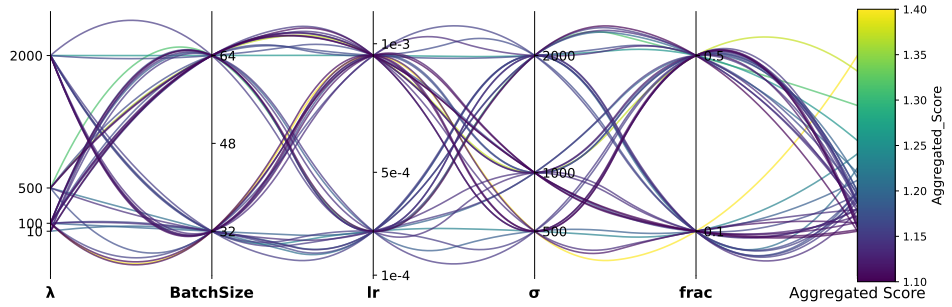


Figure E10: Parallel Coordinates Plot for the **SMILES (BPE)** model (300M).

Table F1: IsoFLOP grid. For each compute budget we evaluate the model sizes that bracket the optimum and read the compute-optimal parameter count N_{opt} and token count D_{opt} from the fitted loss surface.

Budget C (FLOPs)	Evaluated sizes	N_{opt}	D_{opt} (tokens)
3×10^{15}	0.5M, 5M, 32M	2.5M	2.0×10^{8}
3×10^{16}	5M, 32M, 157M	12M	4.2×10^{8}
3×10^{17}	5M, 32M, 157M	55M	9.1×10^{8}
1×10^{18}	32M, 157M, 300M	130M	1.3×10^{9}
1×10^{19}	157M, 300M, 1B	700M	2.4×10^{9}

G Loss Metrics

We compare models in bits per molecule because per-token loss is not comparable across representations and tokenisers that encode the same molecule with different sequence lengths. Tab. G1 decomposes the validation set bits per molecule at 32M into bits per token and tokens per molecule for each encoding and corpus, all on the random split. The decomposition explains why the per-molecule ranking differs from the per-token ranking. SELFIES has the lowest bits per token, yet it needs about 155 tokens to encode a molecule against 47 for SMILES, so its bits per molecule are higher. Byte-pair encoding raises bits per token but shortens the sequence, which is why it overtakes atom-level tokenisation from 32M onwards. Across corpora the per-molecule loss tracks corpus diversity rather than sequence length, since ChEMBL molecules are longer than PubChem molecules yet reach a lower loss.

Tab. G2 reports the validation set bits per molecule of the PubChem–SMILES–atomwise reference across the full size range on the random split.

Table G1: Validation set loss at 32M decomposed into bits per token and tokens per molecule on the random split. Bits per molecule equals bits per token times tokens per molecule.

Axis	Setting	Bits / molecule (↓)	Bits / token	Tokens / molecule
Representation	SMILES (atomwise)	48.1	1.02	47
	SELFIES (atomwise)	50.9	0.33	155
	SAFE (atomwise)	50.2	0.95	53
Tokenisation	SMILES (atom-level)	48.1	1.02	47
	SMILES (byte-pair)	46.4	1.86	25
Corpus (SMILES, atomwise)	MOSES	25.6	0.91	28
	ChEMBL	33.2	0.69	48
	PubChem	48.1	1.02	47
	ZINC-22	51.5	0.99	52

Table G2: Validation set bits per molecule of the PubChem–SMILES–atomwise reference on the random split, trained at $D = 20N$.

Model size	0.5M	5M	32M	157M	300M	1B
Bits / molecule (↓)	71.5	54.1	48.1	46.0	45.5	45.0

G.1 BPE Substructures

We provide visualization of different substructures extracted from the random 100M subset from ZINC-22 in Figure G1.

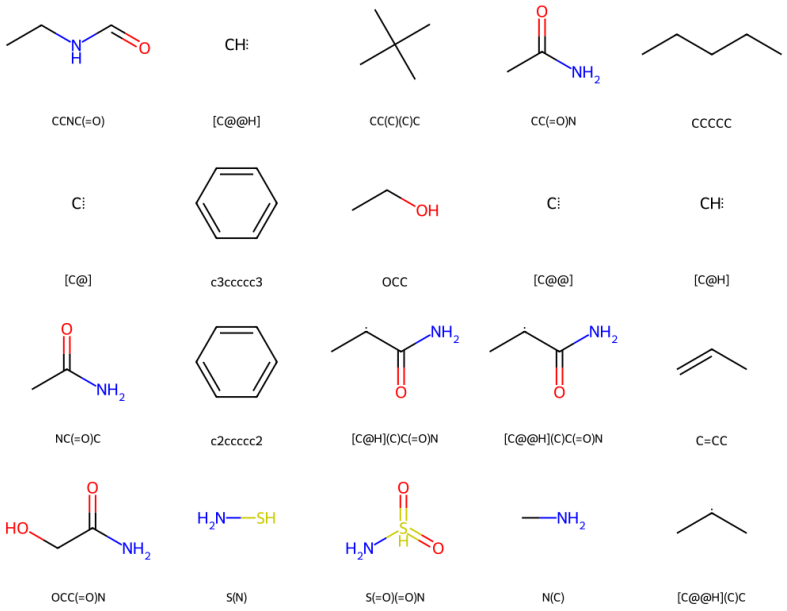


Figure G1: Substructures present in BPE tokenization

H OOD Generalization

This section reports the per-split behaviour behind the out-of-distribution analysis and the full unconstrained-generation metrics. All values use the same $D = 20N$ checkpoints and the same 175,000-molecule validation splits.

H.1 Loss across splits

Tab. H1 reports the validation set bits per molecule of the PubChem–SMILES–atomwise reference on all four splits. The scaffold split stays close to the random split at every scale, so it behaves almost as in-distribution. The molecular-weight and maximum-dissimilarity splits are far harder, and their loss ratio to the random split narrows as the model grows, from 5.9× and 2.4× at 0.5M to 2.8× and 1.3× at 1B. Larger models therefore generalise better in likelihood, even though the shifted splits remain much harder than the in-distribution reference.

H.2 OOD gap across design axes

The relative increase in bits per molecule from the random to the maximum-dissimilarity split depends on the encoding. At 32M this gap is smallest for byte-pair SMILES and largest for SELFIES, as shown in Tab. H2. Measured on their own splits, the gap narrows with scale on the broad PubChem

and ZINC-22 corpora but widens on the narrower MOSES and ChEMBL corpora, which points to tighter fitting of the smaller training distributions.

H.3 Unconstrained generation

We sample 30,000 molecules per model at temperature $T = 1.0$ without top-k or top-p filtering and score them against the validation splits. Tab. H3 reports validity, internal diversity, novelty, and the distribution metrics on the random split across the design axes and the reference size sweep. Lower loss does not translate into a lower Fréchet ChemNet Distance. On PubChem the FCD of the SMILES reference rises from 2.8 at 5M to 5.1 at 32M even though the loss falls, and SELFIES and SAFE reach a lower FCD than SMILES at 32M despite higher bits per molecule. On the narrower MOSES and ChEMBL corpora the FCD is far lower and does fall with loss, since those distributions are easier to match. Tab. H4 shows that sampled molecules stay anchored to the training distribution, with FCD rising from a few units against the random split to 26 to 29 against the maximum-dissimilarity split for every representation.

Table H1: Validation set bits per molecule of the PubChem–SMILES–atomwise reference on all four evaluation splits (↓). The final row is the maximum-dissimilarity loss as a multiple of the random-split loss.

Split	0.5M	5M	32M	157M	300M	1B
Random	71.5	54.1	48.1	46.0	45.5	45.0
Scaffold	72.9	55.2	49.1	46.9	46.4	45.9
Molecular weight	422	243	168	143	132	126
Maximum dissimilarity	172	92	68	62	60	59
Max-dissim. / random	2.4	1.7	1.4	1.35	1.32	1.3

Table H2: Out-of-distribution gap at 32M, the relative increase in bits per molecule from the random to the maximum-dissimilarity split (↓).

Axis	Setting	OOD gap (%)
Representation	SMILES (atomwise)	41
	SELFIES (atomwise)	57
	SAFE (atomwise)	50
Tokenisation	Atom-level	41
	Byte-pair	32
Architecture	Qwen3 (decoder)	41
	T5Gemma (enc–dec)	46

H.4 Generation Metric Definitions

We define the distribution-learning metrics used to assess unconstrained generation, following the MOSES benchmark (Polykovskiy et al., 2020). These metrics are computed based on the generated set of molecules from the model, denoted as G, and two reference sets, the random split (Test) and the scaffold split (TestSF), each holding out molecules from the same corpus the model was trained on. All reported metrics are calculated using the subset of G consisting of valid molecules identified through a post-generation filtering process except for validity. Novelty is computed against the full training corpus rather than a small reference set, so it is not directly comparable to values that prior work reports against smaller datasets such as MOSES. Similar care applies to other distribution-based metrics, such as SNN and scaffold similarity.

Table H3: Unconstrained-generation metrics on the random split for NovoMolGen across design axes. Values are averaged over pretraining seeds where available. Each corpus is scored on its own random split, so the FCD scale differs across corpora. Arrows show the preferred direction.

Axis	Setting	Size	Valid (↑)	IntDiv (↑)	Novelty (↑)	FCD (↓)	SNN (↑)	Frag (↑)	Scaf (↑)
Ref. size	PubChem, SMILES	0.5M	0.310	0.892	0.989	3.92	0.386	0.948	0.469
		5M	0.901	0.894	0.970	2.84	0.402	0.892	0.722
		32M	0.961	0.884	0.979	5.12	0.395	0.901	0.314
		157M	0.968	0.883	0.980	5.05	0.398	0.905	0.330
		300M	0.972	0.882	0.980	4.95	0.399	0.908	0.335
		1B	0.978	0.882	0.981	4.82	0.401	0.910	0.340
Representation	SELFIES (atomwise)	32M	0.994	0.894	0.956	2.43	0.400	0.954	0.662
	SAFE (atomwise)	32M	0.947	0.895	0.986	2.93	0.383	0.963	0.595
Tokenisation	SMILES (byte-pair)	32M	0.965	0.885	0.982	4.60	0.398	0.910	0.340
Corpus	MOSES	32M	0.989	0.854	0.716	0.14	0.564	0.999	0.890
	ChEMBL	32M	0.978	0.876	0.896	0.22	0.482	0.999	0.544
	ZINC-22	32M	0.960	0.900	0.990	6.50	0.380	0.885	0.300

Table H4: Fréchet ChemNet Distance (↓) between 30,000 molecules sampled from each 32M PubChem model and each evaluation split. Samples stay anchored to the training distribution, so the FCD grows sharply on the shifted splits.

Representation (32M)	Random	Scaffold	Molecular weight	Maximum dissimilarity
SMILES (atomwise)	5.12	13.9	23.4	29.0
SELFIES (atomwise)	2.43	12.6	22.9	27.4
SAFE (atomwise)	2.93	12.5	22.8	26.2

1. **Validity**: Validity is determined using RDKit's molecular structure parser, which verifies atomic valency and the consistency of bonds within aromatic rings. The metric ensures that the model adheres to relevant chemical constraints and measures the proportion of valid molecules generated within G. For molecular representations other than SMILES, we convert the generated set to SMILES and assess whether the corresponding decoder successfully decodes the molecule string to SMILES format. This step is essential, as representations that adhere to their respective syntactic rules may still produce chemically invalid molecules.
2. **Novelty**: This metric quantifies the proportion of molecules in G that do not appear in the training dataset. The molecules in G are canonicalized and compared against the full training corpus.
3. **Internal Diversity (IntDiv)**: This metric quantifies the chemical diversity within a set of generated molecules (G). It is calculated by taking 1 minus the average pairwise Tanimoto similarity of the Morgan fingerprints for all molecules in the set. Higher IntDiv scores, which range from 0 (no diversity) to 1 (maximum diversity), indicate a more structurally varied set of generated molecules.
4. **Fréchet ChemNet Distance (FCD)**: Derived from the activations of the penultimate layer of ChemNet, a deep neural network trained to predict the biological activities of drugs, this measure captures the chemical and biological properties of molecules. Activations for canonical SMILES representations of molecules are compared between the generated and reference sets. Lower values indicate better overlap, and the metric is non-negative.
5. **Fragment Similarity (Frag)**: The distribution of BRICS fragments (Degen et al., 2008) is compared between the generated and reference sets. Higher values signify a closer match in fragment distributions, ensuring no fragment is disproportionately overrepresented or underrepresented.
6. **Scaffold Similarity (Scaff)**: Similar to Fragment Similarity, this comparison uses Bemis-Murcko scaffolds instead of BRICS fragments to evaluate the resemblance between scaffolds in the generated and reference datasets.
7. **Similarity to Nearest Neighbor (SNN)**: The average Tanimoto similarity between each molecule in the generated set and its nearest counterpart in the reference set is calculated.

Lower values suggest the generated molecules are farther from the reference set's manifold, while higher values indicate closer alignment.

H.5 Distribution Learning on the Scaffold Split

We sample 30,000 molecules per model and score them against the random split (Test) and the scaffold split (TestSF), each holding out 175,000 molecules, averaging over three seeds. Tab. H5 compares the two splits for the reference size sweep and the representation axis at 32M. The scaffold split is harder in sample space than in likelihood. Validation set loss on the scaffold split is only about 2% above the random split, yet the Fréchet ChemNet Distance and scaffold similarity degrade sharply, because the held-out scaffolds are almost absent from the generated distribution. Within the SMILES reference the metrics change little from 32M onwards. The full set of random-split metrics across every design axis is given in Tab. H3.

Table H5: Random split (Test) versus scaffold split (TestSF) for NovoMolGen. FCD is lower-is-better, SNN and scaffold similarity are higher-is-better. The scaffold split degrades sample-space metrics even though its validation loss stays close to the random split.

Representation	Size	FCD (↓)		SNN (↑)		Scaf (↑)	
		Test	TestSF	Test	TestSF	Test	TestSF
SMILES (atomwise)	0.5M	3.92	15.03	0.386	0.333	0.469	0.064
	5M	2.84	13.17	0.402	0.347	0.722	0.063
	32M	5.12	13.91	0.395	0.345	0.314	0.028
SELFIES (atomwise)	32M	2.43	12.61	0.400	0.345	0.662	0.041
SAFE (atomwise)	32M	2.93	12.53	0.383	0.336	0.595	0.092

I PMO Benchmark

Before the per-task tables, Figure I1 summarises the central downstream finding across design axes and model sizes. Lower pretraining loss does not translate into better goal-directed design, and every axis plateaus in both total PMO score and total docking score beyond ∼5M parameters.

This appendix provides a comprehensive analysis of the PMO benchmark results for **NovoMolGen**, assessing its performance across varying model sizes, tokenization strategies, and intermediate training checkpoints. In Figure I2, the results are presented, highlighting the influence of model size and tokenization approaches on the performance of NovoMolGen. Additionally, Figure I3 tracks the performance progression of NovoMolGen-32M-Atomwise across intermediate checkpoints, demonstrating the evolution of model performance with the optimal hyperparameter configuration and atomwise tokenization. The evaluation includes comparisons with the REINVENT and f-RAG baselines, with the mean and standard deviation of 3 independent runs presented in Tabs. I1 and I2. The reward curve for MPO tasks is shown in Figure I5.

J Protein-Ligand Docking

The docking score reflects the strength of interaction between a ligand and its target protein. In this study, we focus on optimizing binding affinity for five human proteins, each implicated in various diseases:

1. ***parp1***: Poly(ADP-ribose) polymerase 1 (PARP1) is a nuclear enzyme involved in DNA repair, transcriptional regulation, and cell survival. PARP1 inhibitors are widely studied for cancer therapy, particularly in tumors with defective DNA repair mechanisms.
2. ***fa7***: Factor VII (FA7), also known as proconvertin, is a coagulation protein essential for blood clotting. Deficiencies in FA7 are associated with bleeding disorders, including hemophilia-like conditions.

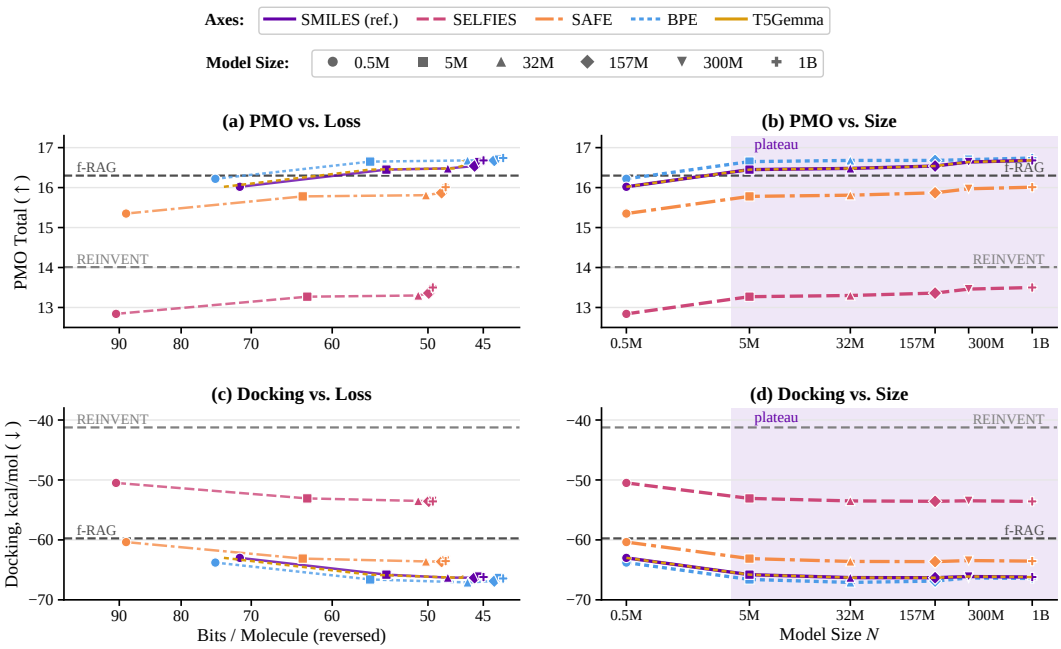


Figure I1: Lower pretraining loss does not buy better goal-directed design. Downstream performance after RL fine-tuning of decoder-only PubChem models across design axes (colour) and model sizes from 0.5M to 1B (marker). **(a)** Total PMO score over 22 tasks (higher is better) against validation set bits per molecule, with the loss axis reversed. **(b)** Every axis plateaus in total PMO score beyond ∼5M (shaded). **(c, d)** The same comparisons for the total top-5% novel docking score over five targets (lower is better). Dashed lines mark f-RAG and REINVENT (Lee et al., 2024a).

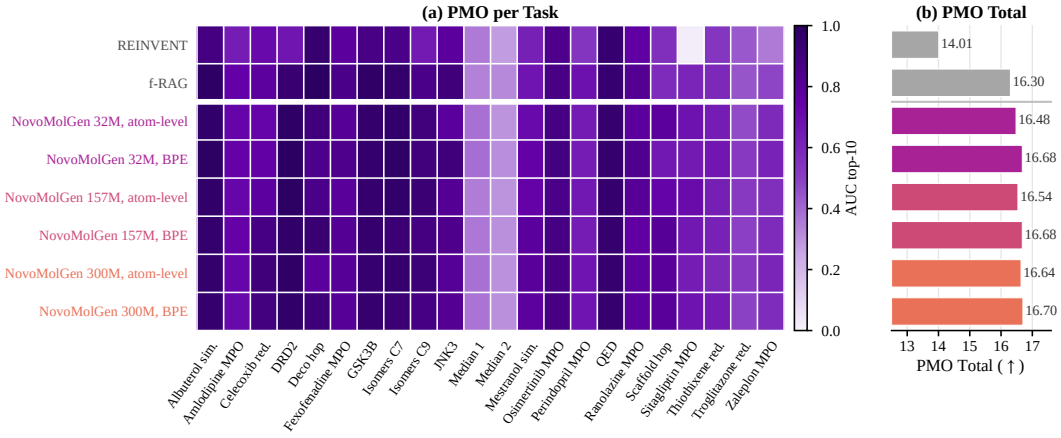


Figure I2: PMO benchmark results for **NovoMolGen-SMILES** across model sizes (32M, 157M, 300M) and tokenization strategies. The **heatmap** (left) shows normalized scores per task, while the **bar chart** (right) presents total scores (higher is better). Baselines are REINVENT and f-RAG .

3. ***5ht1b***: The 5-hydroxytryptamine receptor 1B (5-HT1B) is a serotonin receptor involved in neurotransmitter regulation. It plays a role in mood disorders, anxiety, and migraine, making it a therapeutic target for psychiatric and neurological conditions.

4. ***braf***: B-Raf (serine/threonine-protein kinase B-Raf) is a key regulator of the MAP kinase/ERK signaling pathway, which controls cell proliferation and differentiation. Mutations in BRAF are commonly associated with various cancers, including melanoma and colorectal cancer.

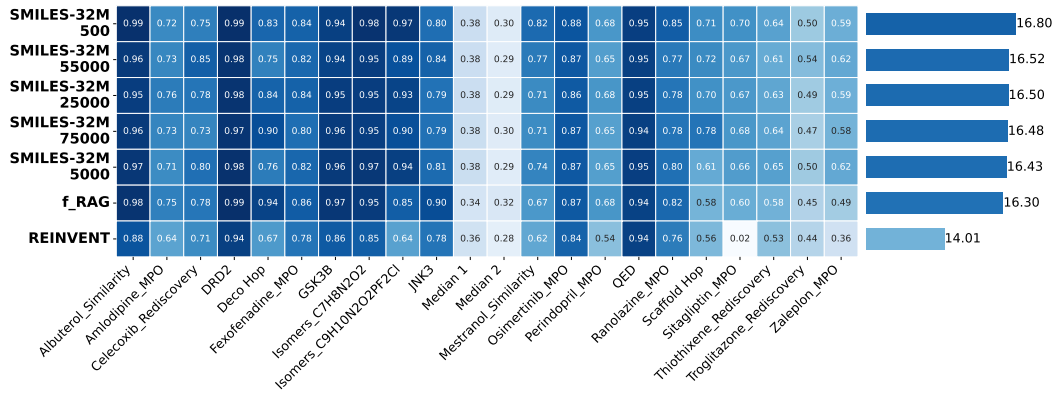


Figure I3: PMO benchmark results for **NovoMolGen-SMILES-32M (AtomWise)** across intermediate training checkpoints. The **heatmap** (left) shows normalized scores per task, while the **bar chart** (right) presents total scores (higher is better). Baselines are REINVENT and f-RAG .

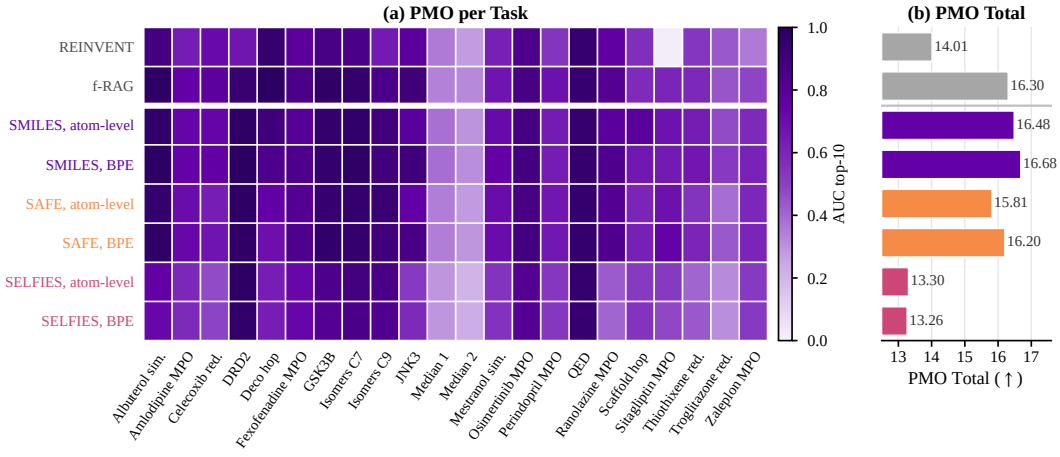


Figure I4: PMO benchmark results for all molecule types **NovoMolGen-32M**. The **heatmap** (left) shows normalized scores per task, while the **bar chart** (right) presents total scores (higher is better). Baselines are REINVENT and f-RAG .

5. ***jak2***: Janus kinase 2 (JAK2) is a tyrosine kinase involved in the JAK/STAT signaling pathway, which regulates hematopoiesis and immune function. Mutations in JAK2 are linked to myeloproliferative disorders such as polycythemia vera and essential thrombocythemia.

We employ the docking program QuickVina 2 (Alhossary et al., 2015) to compute docking scores, setting the exhaustiveness parameter to 1. The timeout for 3D conformer generation is set to 20 seconds, while the docking score calculation timeout is set to 50 seconds. PDB files, box sizes and center coordinates for the target proteins are taken from Lee et al. (2023). To compute the maximum Tanimoto similarity with the training molecules, we use the ZINC-250K molecules to identify the unique and novel compounds. This approach is adopted because searching against the full training corpus is computationally infeasible. The novelty results are presented in Tab. J2 and the visualization of generated novel hits are presented in Figure J2. The reward curve is shown in Figure J1.

Tab. J1 reports the top-5% novel docking score per target for NovoMolGen and the fragment-based, 3D and diffusion baselines. NovoMolGen reaches a better score than every baseline on all five targets from 5M parameters, and the total score improves only marginally from -65.8 at 5M to -66.2 at 1B, so the docking benefit of scale is small once a compact prior is in place.

Table I1: PMO **AUC Top-10** results. Results are reported as mean$_{(\text{std})}$ over three independent model initializations for **SMILES** molecule type. **Blue** denotes the best performing model, while **Pink** represents the second-best model by mean score.

Oracle	f-RAG	REINVENT	NovoMolGen-32M (AtomWise)	NovoMolGen-157M (AtomWise)	NovoMolGen-300M (AtomWise)
albuterol_similarity	**0.977**$_{(0.002)}$	0.882$_{(0.006)}$	0.958$_{(0.003)}$	**0.968**$_{(0.002)}$	0.954$_{(0.002)}$
amlodipine_mpo	**0.749**$_{(0.019)}$	0.635$_{(0.035)}$	**0.732**$_{(0.024)}$	0.730$_{(0.037)}$	0.730$_{(0.007)}$
celecoxib_rediscovery	**0.778**$_{(0.007)}$	0.713$_{(0.067)}$	0.732$_{(0.009)}$	0.777$_{(0.132)}$	**0.904**$_{(0.022)}$
drd2	0.936$_{(0.011)}$	0.666$_{(0.044)}$	**0.973**$_{(0.001)}$	**0.983**$_{(0.002)}$	0.972$_{(0.001)}$
deco_hop	**0.992**$_{(0.000)}$	**0.945**$_{(0.007)}$	0.902$_{(0.007)}$	0.853$_{(0.115)}$	0.785$_{(0.130)}$
fexofenadine_mpo	**0.856**$_{(0.016)}$	0.784$_{(0.006)}$	0.802$_{(0.008)}$	0.790$_{(0.028)}$	**0.803**$_{(0.018)}$
gsk3b	**0.969**$_{(0.003)}$	0.865$_{(0.043)}$	**0.955**$_{(0.006)}$	0.949$_{(0.005)}$	0.941$_{(0.012)}$
isomers_c7h8n2o2	0.955$_{(0.008)}$	0.852$_{(0.036)}$	**0.961**$_{(0.001)}$	**0.965**$_{(0.002)}$	0.949$_{(0.005)}$
isomers_c9h10n2o2pf2cl	0.850$_{(0.005)}$	0.642$_{(0.054)}$	0.898$_{(0.032)}$	**0.926**$_{(0.024)}$	**0.915**$_{(0.022)}$
jnk3	**0.904**$_{(0.004)}$	0.783$_{(0.023)}$	0.787$_{(0.014)}$	**0.809**$_{(0.066)}$	0.808$_{(0.045)}$
median1	0.340$_{(0.007)}$	0.356$_{(0.009)}$	**0.380**$_{(0.003)}$	0.354$_{(0.027)}$	**0.379**$_{(0.002)}$
median2	**0.323**$_{(0.005)}$	0.276$_{(0.008)}$	0.300$_{(0.013)}$	0.299$_{(0.001)}$	**0.304**$_{(0.019)}$
mestranol_similarity	0.671$_{(0.021)}$	0.618$_{(0.048)}$	0.712$_{(0.047)}$	**0.739**$_{(0.068)}$	**0.796**$_{(0.046)}$
osimertinib_mpo	0.866$_{(0.009)}$	0.837$_{(0.009)}$	**0.874**$_{(0.013)}$	0.858$_{(0.001)}$	**0.871**$_{(0.011)}$
perindopril_mpo	**0.681**$_{(0.017)}$	0.537$_{(0.016)}$	0.637$_{(0.050)}$	0.649$_{(0.037)}$	**0.668**$_{(0.011)}$
qed	0.939$_{(0.001)}$	0.941$_{(0.000)}$	0.945$_{(0.000)}$	**0.946**$_{(0.000)}$	**0.945**$_{(0.000)}$
ranolazine_mpo	**0.820**$_{(0.016)}$	0.760$_{(0.009)}$	0.784$_{(0.007)}$	**0.797**$_{(0.023)}$	0.779$_{(0.024)}$
scaffold_hop	0.576$_{(0.014)}$	0.560$_{(0.019)}$	**0.783**$_{(0.123)}$	0.735$_{(0.140)}$	**0.791**$_{(0.160)}$
sitagliptin_mpo	0.601$_{(0.011)}$	0.021$_{(0.009)}$	**0.677**$_{(0.018)}$	**0.708**$_{(0.038)}$	0.635$_{(0.043)}$
thiothixene_rediscovery	0.584$_{(0.009)}$	0.534$_{(0.013)}$	**0.635**$_{(0.018)}$	**0.624**$_{(0.050)}$	0.584$_{(0.062)}$
troglitazone_rediscovery	0.448$_{(0.017)}$	0.441$_{(0.032)}$	0.470$_{(0.042)}$	**0.516**$_{(0.010)}$	**0.524**$_{(0.058)}$
zaleplon_mpo	0.486$_{(0.004)}$	0.358$_{(0.062)}$	**0.579**$_{(0.023)}$	0.559$_{(0.013)}$	**0.599**$_{(0.000)}$
Sum	16.30	14.01	16.48	**16.54**	**16.64**

Table I2: PMO **AUC Top-10** results. Results are reported as mean$_{(\text{std})}$ over three independent model initializations for **SMILES** molecule type. **Blue** denotes the best performing model, while **Pink** represents the second-best model by mean score.

Oracle	f-RAG	REINVENT	NovoMolGen-32M (BPE)	NovoMolGen-157M (BPE)	NovoMolGen-300M (BPE)
albuterol_similarity	**0.977**$_{(0.002)}$	0.882$_{(0.006)}$	**0.981**$_{(0.001)}$	0.951$_{(0.000)}$	0.949$_{(0.001)}$
amlodipine_mpo	**0.749**$_{(0.019)}$	0.635$_{(0.035)}$	**0.739**$_{(0.021)}$	0.739$_{(0.026)}$	0.713$_{(0.010)}$
celecoxib_rediscovery	0.778$_{(0.007)}$	0.713$_{(0.067)}$	0.751$_{(0.190)}$	**0.872**$_{(0.061)}$	**0.890**$_{(0.018)}$
drd2	0.936$_{(0.011)}$	0.666$_{(0.044)}$	**0.987**$_{(0.001)}$	0.967$_{(0.001)}$	**0.967**$_{(0.000)}$
deco_hop	**0.992**$_{(0.000)}$	**0.945**$_{(0.007)}$	0.844$_{(0.121)}$	0.937$_{(0.008)}$	0.914$_{(0.051)}$
fexofenadine_mpo	**0.856**$_{(0.016)}$	0.784$_{(0.006)}$	**0.837**$_{(0.022)}$	0.810$_{(0.012)}$	0.804$_{(0.002)}$
gsk3b	**0.969**$_{(0.003)}$	0.865$_{(0.043)}$	**0.953**$_{(0.013)}$	0.947$_{(0.000)}$	0.944$_{(0.006)}$
isomers_c7h8n2o2	**0.955**$_{(0.008)}$	0.852$_{(0.036)}$	**0.975**$_{(0.003)}$	0.929$_{(0.006)}$	0.928$_{(0.006)}$
isomers_c9h10n2o2pf2cl	0.850$_{(0.005)}$	0.642$_{(0.054)}$	**0.897**$_{(0.027)}$	**0.876**$_{(0.005)}$	0.874$_{(0.011)}$
jnk3	**0.904**$_{(0.004)}$	0.783$_{(0.023)}$	**0.899**$_{(0.049)}$	0.839$_{(0.044)}$	0.818$_{(0.025)}$
median1	0.340$_{(0.007)}$	0.356$_{(0.009)}$	**0.388**$_{(0.003)}$	0.364$_{(0.013)}$	**0.374**$_{(0.002)}$
median2	**0.323**$_{(0.005)}$	0.276$_{(0.008)}$	**0.310**$_{(0.031)}$	0.305$_{(0.016)}$	0.307$_{(0.021)}$
mestranol_similarity	0.671$_{(0.021)}$	0.618$_{(0.048)}$	0.745$_{(0.028)}$	**0.789**$_{(0.018)}$	**0.780**$_{(0.040)}$
osimertinib_mpo	0.866$_{(0.009)}$	0.837$_{(0.009)}$	**0.885**$_{(0.012)}$	**0.873**$_{(0.008)}$	0.870$_{(0.008)}$
perindopril_mpo	**0.681**$_{(0.017)}$	0.537$_{(0.016)}$	0.635$_{(0.018)}$	0.636$_{(0.036)}$	**0.669**$_{(0.049)}$
qed	0.939$_{(0.001)}$	0.941$_{(0.000)}$	**0.947**$_{(0.000)}$	0.945$_{(0.000)}$	**0.945**$_{(0.000)}$
ranolazine_mpo	**0.820**$_{(0.016)}$	0.760$_{(0.009)}$	**0.830**$_{(0.010)}$	0.760$_{(0.018)}$	0.780$_{(0.012)}$
scaffold_hop	0.576$_{(0.014)}$	0.560$_{(0.019)}$	0.650$_{(0.002)}$	**0.803**$_{(0.063)}$	**0.790**$_{(0.134)}$
sitagliptin_mpo	0.601$_{(0.011)}$	0.021$_{(0.009)}$	0.643$_{(0.034)}$	**0.650**$_{(0.040)}$	**0.673**$_{(0.044)}$
thiothixene_rediscovery	0.584$_{(0.009)}$	0.534$_{(0.013)}$	**0.657**$_{(0.012)}$	**0.615**$_{(0.010)}$	0.638$_{(0.003)}$
troglitazone_rediscovery	0.448$_{(0.017)}$	0.441$_{(0.032)}$	**0.517**$_{(0.032)}$	**0.497**$_{(0.016)}$	0.496$_{(0.003)}$
zaleplon_mpo	0.486$_{(0.004)}$	0.358$_{(0.062)}$	**0.611**$_{(0.027)}$	**0.578**$_{(0.041)}$	0.574$_{(0.010)}$
Sum	16.30	14.01	**16.68**	**16.68**	**16.70**

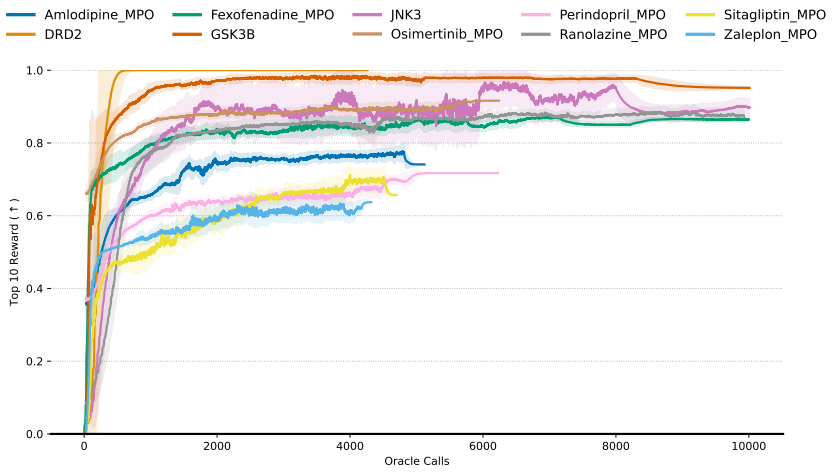


Figure I5: Sample efficiency of **NovoMolGen-SMILES-BPE-32M** on goal-directed generation tasks from the PMO benchmark. The plot tracks the top-10 reward as a function of oracle calls, showing how quickly the model discovers high-reward molecules.

Table J1: Top-5% novel docking score per target (kcal/mol, ↓). Baselines are taken from Lee et al. (2024a). **Blue** denotes the best score per target. Totals are the reported sums, so rounded per-target entries may not add up exactly.

Method	parp1	fa7	5ht1b	braf	jak2	**Sum**
REINVENT	−8.7	−7.2	−8.8	−8.4	−8.2	−41.2
MOOD	−10.9	−8.2	−11.2	−11.1	−10.2	−51.4
GEAM	−12.9	−9.9	−12.4	−12.3	−11.8	−59.3
f-RAG	−12.9	−9.9	−12.7	−12.4	−11.8	−59.8
NovoMolGen-0.5M	−13.6	−10.4	−13.4	−13.1	−12.5	−63.0
NovoMolGen-5M	−14.2	−10.9	−14.0	−13.6	−13.0	−65.8
NovoMolGen-32M	**−14.3**	**−11.0**	**−14.1**	**−13.8**	−13.1	**−66.3**
NovoMolGen-157M	**−14.3**	−10.9	−14.0	**−13.8**	**−13.2**	**−66.3**
NovoMolGen-300M	**−14.3**	−10.9	−14.0	**−13.8**	−13.1	−66.1
NovoMolGen-1B	**−14.3**	**−11.0**	−14.0	**−13.8**	−13.1	−66.2

Table J2: Novelty comparison on protein-ligand docking tasks. We report Novelty (%), calculated with respect to the ZINC-250k dataset. Results show the mean and standard deviation from 3 independent runs. Baseline data is sourced from Lee et al. (2024b). **Blue** denotes the best model by mean score, while **Pink** represents the second-best model by mean score.

Method	parp1	fa7	5ht1b	braf	jak2
REINVENT	$9.894_{(2.178)}$	$10.731_{(1.516)}$	$11.605_{(3.688)}$	$8.715_{(2.712)}$	$11.456_{(1.793)}$
MORLD	$\mathbf{98.433_{(1.189)}}$	$\mathbf{97.967_{(1.764)}}$	$\mathbf{98.787_{(0.743)}}$	$\mathbf{96.993_{(2.787)}}$	$\mathbf{97.720_{(0.995)}}$
HierVAE	$60.453_{(17.165)}$	$24.853_{(15.416)}$	$48.107_{(1.988)}$	$59.747_{(16.403)}$	$85.200_{(14.262)}$
FREED	$74.640_{(2.953)}$	$78.787_{(2.132)}$	$75.027_{(5.194)}$	$73.653_{(4.312)}$	$75.907_{(5.916)}$
MOOD	$84.180_{(2.123)}$	$83.180_{(1.519)}$	$84.613_{(0.822)}$	$87.413_{(0.830)}$	$83.273_{(1.455)}$
GEAM	$\mathbf{88.611_{(3.107)}}$	$\mathbf{89.378_{(2.619)}}$	$\mathbf{84.222_{(2.968)}}$	$\mathbf{90.322_{(3.467)}}$	$\mathbf{89.222_{(1.824)}}$
SMILES-32M-ATOMWISE	$30.34_{(7.0)}$	$45.84_{(26.0)}$	$15.38_{(2.0)}$	$23.14_{(3.0)}$	$41.48_{(8.0)}$
SMILES-157M-ATOMWISE	$54.31_{(5.0)}$	$46.26_{(8.0)}$	$27.28_{(9.0)}$	$29.01_{(16.0)}$	$44.68_{(4.0)}$
SMILES-300M-ATOMWISE	$37.78_{(16.0)}$	$66.12_{(10.0)}$	$42.69_{(17.0)}$	$33.23_{(18.0)}$	$37.83_{(6.0)}$
SMILES-32M-BPE	$56.28_{(6.0)}$	$62.02_{(6.0)}$	$33.69_{(8.0)}$	$52.73_{(8.0)}$	$48.42_{(7.0)}$
SMILES-157M-BPE	$45.98_{(5.0)}$	$62.06_{(2.0)}$	$43.62_{(4.0)}$	$63.26_{(5.0)}$	$51.64_{(4.0)}$
SMILES-300M-BPE	$50.12_{(9.0)}$	$31.37_{(4.0)}$	$40.44_{(3.0)}$	$53.24_{(3.0)}$	$45.87_{(1.0)}$

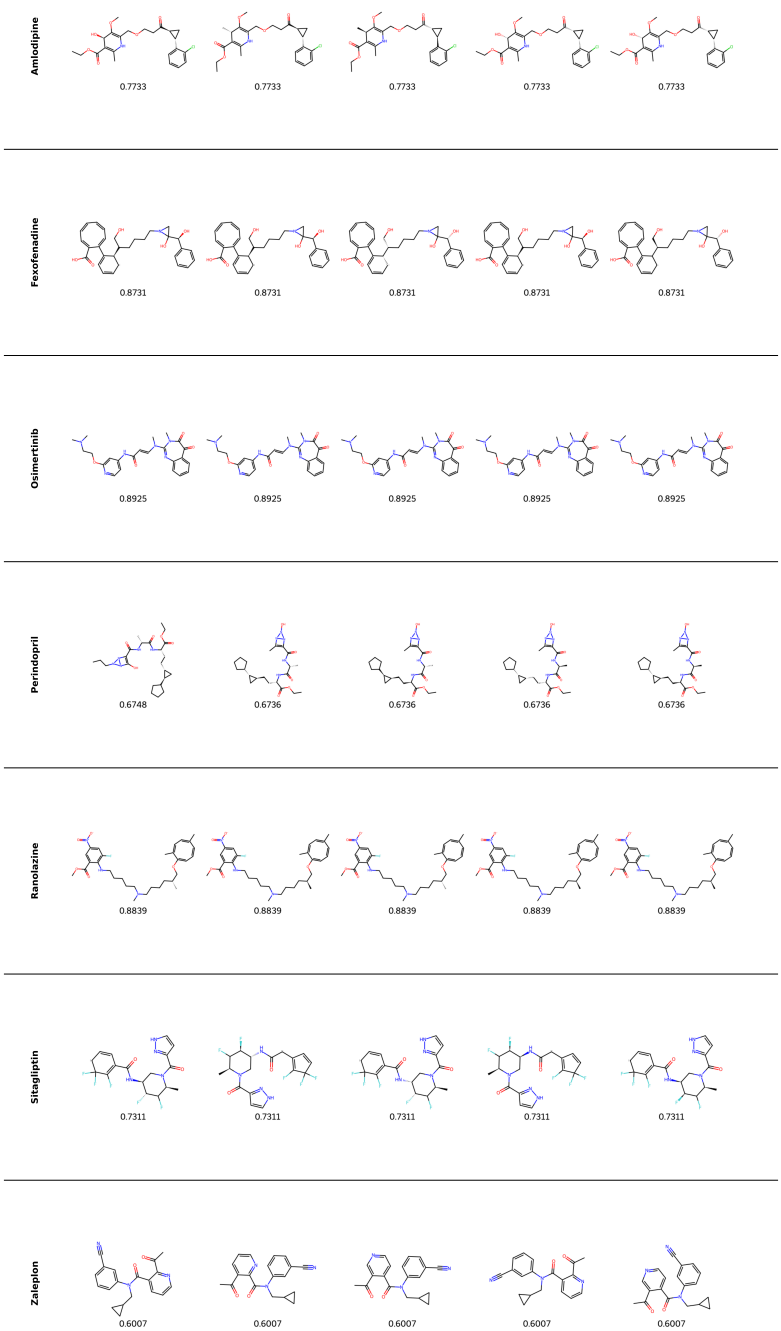


Figure I6: Examples of the generated top-5 molecules from a single run of **NovoMolGen-SMILES-BPE-32M**. The scores are provided at the bottom of each generated molecule.

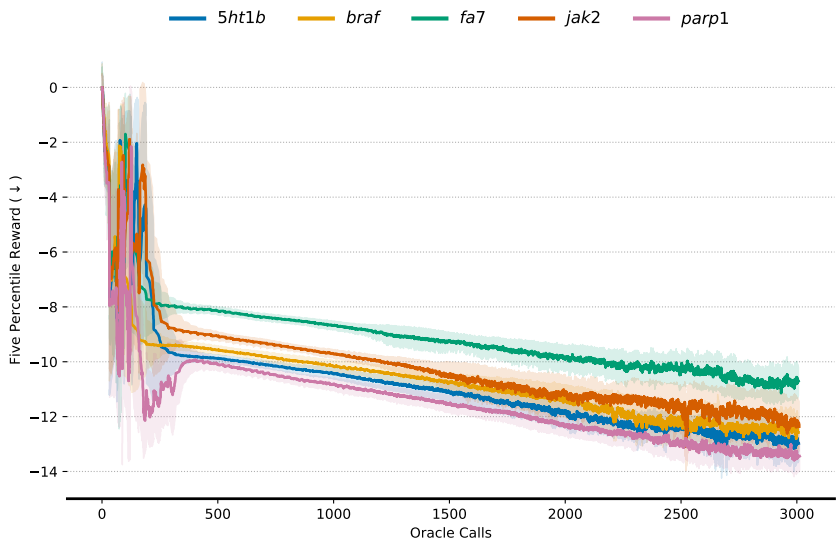


Figure J1: Sample efficiency of **NovoMolGen-SMILES-BPE-32M** on Protein-ligand docking goal-directed generation tasks. The plot tracks the top 5th percentile of the reward distribution as a function of oracle calls, demonstrating our model's ability to rapidly discover high-reward molecules.

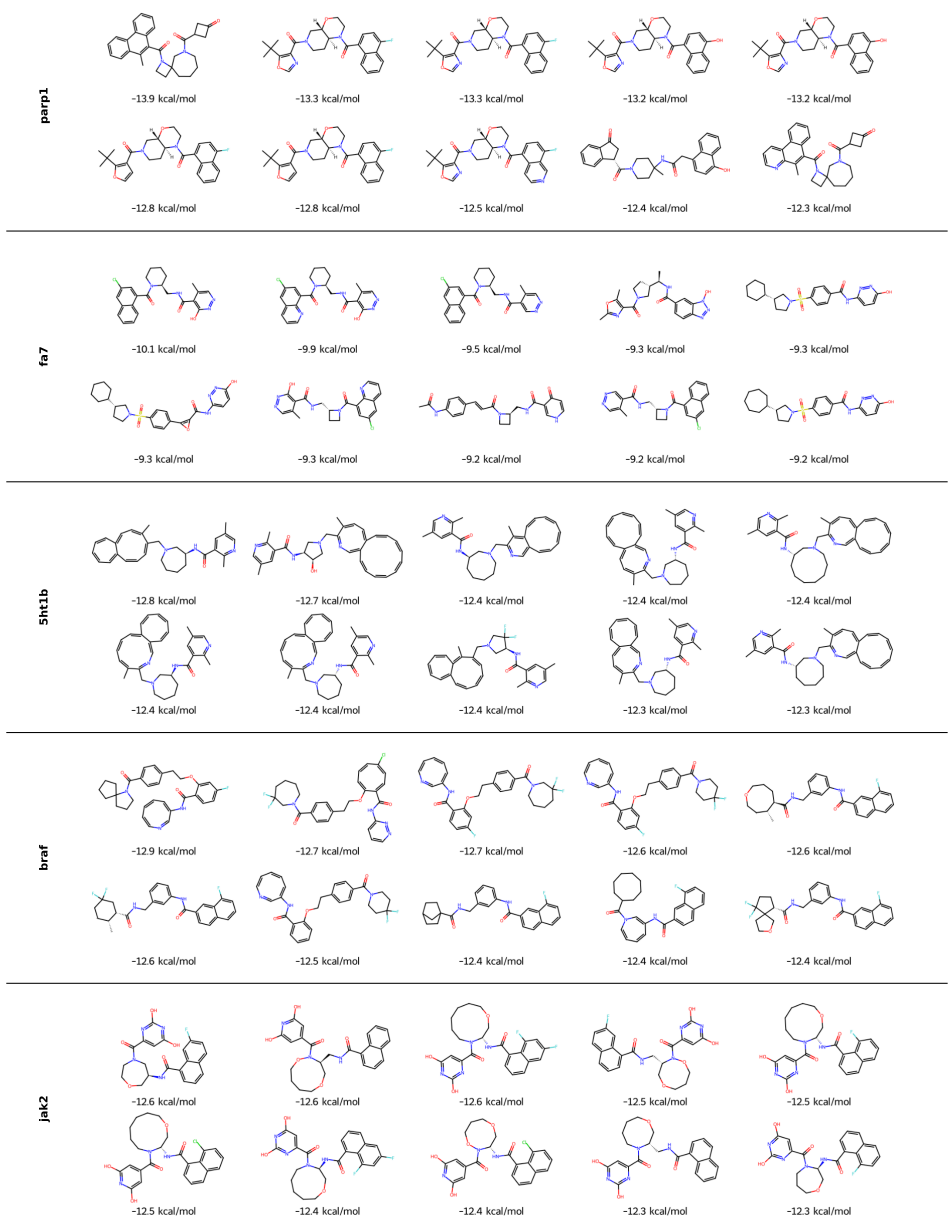


Figure J2: Examples of the generated top-10 molecules from a single run of **NovoMolGen-SMILES-BPE-32M**, shown for the stricter showcase filter QED > 0.5 and SA < 4 rather than the SA < 5 used in the novel-hit definition. The docking scores are given at the bottom of each molecule.