Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

LinkerGen: Guided Autoregressive Generation of Organic Linkers for Reticular Chemistry

Patrick W.V. Butler[1,2,#], Marcello Costamagna[1,#], Felix-Cosmin Mocanu[1], Simon D. Rihm[3], Feroz Farazi[1], Sebastian Mosbach[1,2], Jethro Akroyd[1,2], Markus Kraft[1,2,3]

released: November 2, 2026

[1] Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[2] CARES
Cambridge Centre for Advanced Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

[3] CMPG
GRIPS – Gründerinnenzentrum Pirmasens
Delaware Avenue 1–3
66953 Pirmasens
Germany

[#] These authors contributed equally

Preprint No. 349



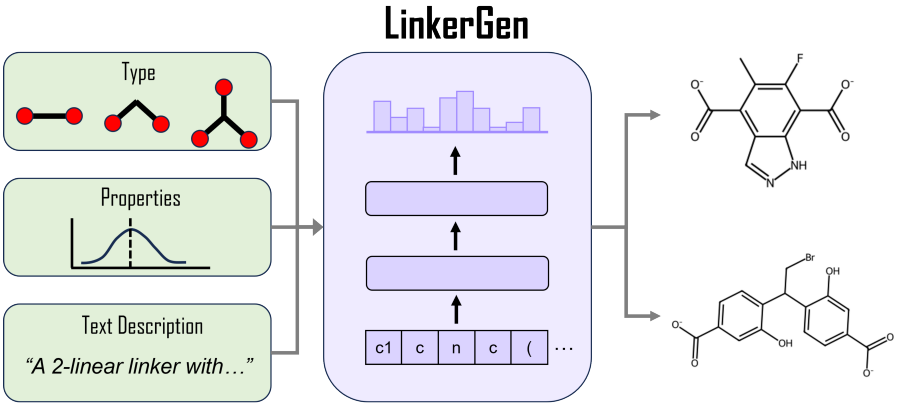
Keywords: Reticular Chemistry, Inverse Design, Metal–Organic Polyhedra, Computational Materials Discovery, Multi-Modal Transformer

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

E-Mail: mk306@cam.ac.uk
World Wide Web: https://como.ceb.cam.ac.uk/

Abstract

Organic linker design for reticular materials requires balancing molecular diversity with structural and chemical constraints. Here, we introduce LinkerGen, a compact SMILES-based autoregressive model guided by chemical building unit (CBU) types, numerical property targets, or text descriptions. We train and evaluate two model sizes on approximately 22 million fragment-enumerated carboxylate linkers spanning 2-linear, 2-bent, and 3-planar CBU types. Model size and sampling temperature adjust the balance between validity and novelty, while conditioning increases the yield of distinct, valid, novel candidates satisfying selected requests compared with unconditioned generation followed by filtering. Text conditioning combines molecular requirements in a single natural-language request, although its effectiveness depends on training coverage and prompt formulation. A prospective application connects literature-informed linker generation to metal–organic polyhedron (MOP) assembly and calculated binding energies for 5-fluorouracil, illustrating how chemical design hypotheses can be explored through computational screening. LinkerGen provides a steerable source of candidate building units and a basis for adaptation to other linker families and design objectives through suitable training data and conditioning information.



Highlights

- Generative model for reticular organic chemical building units (CBUs) widen the chemical linker space available for the design of metal-organic polyhedra.
- Conditioning enriches generated populations in linkers with desired properties.
- Text conditioning allows complex chemical requirements to be effectively expressed in natural language.

Contents

1 Introduction **4**

2 Methods **5**

2.1 Training Data 5

2.2 Model Architecture and Training 6

2.3 Linker Generation and Evaluation 6

2.4 Prospective 5-Fluorouracil Application 7

3 Results **8**

3.1 Unconditioned Generation 8

3.2 Conditioning on Linker Type 8

3.3 Conditioning on Linker Properties 10

3.4 Text Description Conditioning 11

3.5 Conditioning Efficiency 13

3.6 Prospective Application to MOP Hosts for 5-Fluorouracil 14

4 Discussion **16**

5 Conclusions **18**

A Appendix **20**

A.1 Dataset Construction and Characterisation 20

A.1.1 Starting Dataset 20

A.1.2 Dataset Splits 20

A.1.3 CBU-Type Assignment 21

A.1.4 Molecular Descriptors 21

A.1.5 Property Coverage 21

A.2 Model Architecture and Training 22

A.2.1 Representation and Architecture 22

A.2.2 Training Procedure 23

A.2.3 Expanded-Corpus Checkpoint Selection 24

A.3 Training and Generation Times 25

A.3.1 Hardware and Training Times 25

A.3.2 Generation and Assessment Times 25
A.4 Unconditioned Generation and Novelty 26
A.5 Conditioned Generation 27
A.5.1 Numerical-Property Conditioning 27
A.5.2 Additional Property Targets 28
A.5.3 Conditioning versus Filtering 28
A.6 Text Conditioning and Robustness 30
A.6.1 Training Descriptions 30
A.6.2 Sensitivity to Wording 31
A.7 Synthetic Accessibility Proxies 32
A.8 Prospective 5-Fluorouracil Application 34
A.8.1 Literature Context and Generation Prompts 34
A.8.2 Candidate Selection and Workflow Outcomes 34
A.8.3 Assembly and Host–Guest Calculations 35
A.8.4 Cage Windows and Final Guest Location 36
A.8.5 Host Reorganisation and Reference Energies 38
A.8.6 Lowest-Energy LinkerGen Candidate 39
A.9 Data and Code Availability 40

References **41**

1 Introduction

Reticular chemistry offers a modular approach to constructing materials by connecting chemical building units (CBUs) into defined architectures. [1] This approach has enabled the development of metal–organic frameworks (MOFs), [2] covalent organic frameworks (COFs), [3] and discrete metal–organic polyhedra (MOPs), [4] with applications in gas capture and separation, catalysis, water harvesting, and drug delivery. [5–9] Their structural and chemical diversity offers opportunities to tailor materials to particular applications, but identifying suitable building units and their combinations remains a substantial design challenge. [10] In metal–organic systems, inorganic nodes and organic linkers jointly determine the assembled architecture and its properties. Inorganic components contribute coordination geometry, connectivity, and chemical interaction sites, while organic linkers contribute the arrangement of binding groups, molecular dimensions, flexibility, and functional groups. These contributions are interdependent: changing the metal can alter adsorption within an otherwise retained framework family, [11] while modifying the organic ligand can change cage geometry and nuclearity. [12] Within suitable structural families, linker extension and functionalisation also allow pore dimensions and chemical environments to be varied systematically. [13, 14] Generating linkers with specified structural and chemical characteristics therefore provides one means of exploring material designs, alongside the selection of inorganic components and assembly architectures.

Several computational approaches provide access to linker candidates. Libraries of experimentally characterised building units supply established components that can be recombined with selected nodes and topologies to construct hypothetical materials for screening. [15–18] Broader molecular databases extend this search to compounds not previously used in reticular materials, selected according to their coordinating groups, connectivity, and structural characteristics.[19] Fragment-based methods construct additional linkers by recombining molecular units according to defined templates, using enumeration or evolutionary optimisation to explore the resulting space. [20, 21] These approaches provide both candidate structures and chemically structured datasets, creating opportunities to learn from available linker chemistry and direct further exploration towards particular design requirements.

Generative machine learning (ML) provides a complementary means of exploring molecular space. Molecular sequence models learn patterns in structures represented as SMILES [22] or SELFIES [23] strings and sample additional molecules from the learned distributions. [24, 25] In drug discovery, fragment-linking methods such as DeLinker, [26] SyntaLinker, [27] 3DLinker, [28] and DiffLinker [29] generate connections between specified substructures. These methods incorporate different forms of structural guidance, including spatial arrangement in three-dimensional approaches, while Link-INVENT [30] uses reinforcement learning to optimise specified objectives. Although reticular design imposes different requirements, related methods have been applied directly to this domain. A pretrained DiffLinker model has been used to generate MOF linkers for subsequent assembly and CO_2 adsorption screening. [31] More recently, Nexerra^{R1} combined chemical language modelling with seed-guided and scaffold-constrained MOF linker design and a flow-based extension for property-directed generation. The workflow includes framework

assembly, with the scaffold-constrained approach leading to the experimental synthesis of a new MOF. [32]

Conditioning guides molecular generation through molecular classes, numerical targets, or textual descriptions of desired features. [33, 34] Descriptor-conditioned models enable steering towards requested property ranges, [34] while text–molecule models such as nach0 [35] and ChemLML [36] allow requirements to be expressed through language. The effectiveness of this control depends on training coverage and how requirements are represented: rare property combinations may be difficult to reproduce, and text-guided generation may be sensitive to wording. [32, 37, 38] For linker design, this motivates assessing both the ability to generate requested chemistry and the limits imposed by the available training examples.

Here, we introduce LinkerGen, a SMILES-based autoregressive language model for generating organic CBUs under specified molecular requirements. We train and evaluate two model sizes using approximately 22 million unique carboxylate linkers reconstructed from the fragments and assembly templates of a previously reported library [20] (Supporting Information, Section A.1), covering the 2-linear, 2-bent, and 3-planar CBU types. We assess how model size and sampling temperature affect validity, uniqueness, and novelty, defining novelty against the complete reference dataset. We then evaluate conditioning on CBU type, numerical descriptors, and textual descriptions, comparing the yield of candidates satisfying each request with that obtained from unconditioned generation followed by identical filtering. We also examine requests with limited training support, sensitivity to prompt formulation, and the incorporation of additional numerical descriptors.

To connect molecular generation with a reticular application, we use descriptions informed by an LLM-assisted literature review to generate linkers for MOP hosts for 5-fluorouracil. Generated candidates undergo molecular selection, three-dimensional construction, cage assembly, and computational binding-energy screening. Comparisons with generic text-guided generation and direct LLM generation allow us to examine how the selected molecular requirements translate into host–guest performance after assembly. LinkerGen source code is freely available on GitHub: github.com/pwvbutler/linkergen.

2 Methods

2.1 Training Data

The starting dataset was reconstructed from the molecular fragment collection and 12 CBU templates introduced in our previous work, spanning the 2-linear, 2-bent, and 3-planar construction classes. [20, 21] The source fragments were originally extracted from ChEMBL36 [39] using BRICS fragmentation. [40] Assembly was restricted to carboxylate binding groups and used a revised RDKit-based procedure connecting fragment attachment sites through single bonds. [41] The assembled structures were validated using RDKit, retaining only connected molecules with the required number of carboxylate binding groups and no residual attachment markers or radical centres. Stereochemical annotations were removed, and canonical SMILES were deduplicated, yielding 22 074 197

unique CBUs.

Molecules were assigned reproducibly to training, validation, and test partitions in approximately 90:5:5 proportions. These assignments were preserved during conditioned training. Numerical annotations were calculated using RDKit, while separate text corpora were constructed from annotated molecular subsets using description templates, property dropout, and varied property ordering. Dataset construction and descriptor definitions are detailed in Section A.1 of the Supporting Information; text-corpus preparation is described in Section A.6.

2.2 Model Architecture and Training

LinkerGen is an autoregressive transformer decoder operating on SMILES strings. [22, 42] Self-attention processes the growing molecular sequence, while optional cross-attention incorporates conditioning information. CBU types and numerical properties are mapped to conditioning vectors through a feed-forward network. Text descriptions are embedded using the frozen nach0 encoder [35] and projected into the decoder embedding space. Cross-attention is bypassed during unconditioned generation.

Two model sizes were evaluated, with 279 865 and 1 621 305 stored parameters in the small and large unconditioned models, respectively, including inactive conditioning components. Models were trained using AdamW [43] to minimise cross-entropy loss for next-token prediction. Unconditioned training comprised ten epochs, with checkpoints selected by minimum validation loss. These checkpoints initialised separate conditioned-training runs: ten additional epochs for CBU-type and numerical conditioning, and ten or twenty epochs for text conditioning, depending on the corpus. For the expanded-corpus text models, checkpoints within 5% of the minimum validation loss were compared using 1 000 generation attempts under the research prompt. Selecting the checkpoint producing the most distinct valid molecules gave epoch 15 for both model sizes. Architecture, parameter counts for each conditioning mode, and training details are provided in Section A.2 of the Supporting Information.

2.3 Linker Generation and Evaluation

Unless otherwise stated, each condition was evaluated at a temperature of 1 through ten independent sampling runs of 10 000 attempts using the same trained checkpoint. Generation proceeded from a start-of-sequence token until an end-of-sequence token or the maximum length of 512 tokens was reached. Results are reported as means and standard deviations across runs.

Molecules were considered valid if they could be parsed by RDKit, were closed shell, contained at least two anionic carboxylate groups, and received a supported CBU-type assignment using the procedure in Section A.1.3 of the Supporting Information. Valid outputs were deduplicated within each run using canonical SMILES. Novel molecules were distinct valid outputs absent from the complete reference dataset, including all three partitions. Validity, uniqueness, and novelty were expressed as percentages of all genera-

tion attempts.

Conditioned generation was additionally evaluated against the requested properties or structural features, using criteria and tolerances specified alongside each experiment. To assess sampling efficiency, we compared the yield of distinct valid novel molecules satisfying each request with that obtained from unconditioned generation followed by identical filtering. Yields were expressed as percentages of all generation attempts, measuring the number of qualifying candidates obtained from a fixed sampling budget. Synthetic-accessibility proxy calculations and reference-dataset comparisons are described separately in Section A.7 of the Supporting Information.

2.4 Prospective 5-Fluorouracil Application

An LLM-assisted literature review informed a research prompt for generating linkers for tetrahedral M_4L_6 MOP hosts for 5-fluorouracil. Both the small and large text-conditioned models trained on the expanded corpus were run with this prompt. For comparison, the small model was also run with the generic description *"A 2-linear linker"*. Each LinkerGen condition used three independent runs of 5 000 generation attempts at $T = 1$, with 100 distinct accepted candidates selected per run. Research-prompt sets prioritised candidates satisfying all measurable requirements, followed by random selection from the remaining accepted outputs; generic sets were selected randomly from accepted 2-linear outputs. As an additional comparison, GPT-5.6-Sol and GPT-6-Astra were each given the literature-review report and asked directly to propose 100 carboxylate linkers for the same host–guest application. Selection preceded energy calculations, and failed candidates were not replaced. The review, generation prompts, and selection procedures are provided in Sections A.8.1 and A.8.2 of the Supporting Information.

A previously reported MOP assembly workflow combined six organic linkers with four $[Zr_3O(OH)_3(C_5H_5)_3]$ metal CBUs to construct each cage. [44, 45] Calculations used `mace-matpes-r2scan-0` with D3(BJ) dispersion using r2SCAN parameters. [46–48] Four chloride counterions neutralised each cage; solvent was not included. Following separate host and guest relaxation, up to 720 guest placements were screened per cage and up to three distinct low-energy poses were relaxed.

Binding energies were calculated for the lowest-energy accepted complex relative to the relaxed empty host, including counterions, and free guest. The host reference used the lower accepted energy obtained from initial host relaxation or further relaxation of a host extracted from a relaxed complex. Comparisons included only candidates satisfying the structural and guest-location criteria. Calculation settings, acceptance criteria, and statistical procedures are detailed in Sections A.8.3 and A.8.2 of the Supporting Information.

3 Results

3.1 Unconditioned Generation

We first evaluated unconditioned generation to establish how model size and sampling temperature affect the production of valid and novel CBUs. Both models were trained on the same training partition and evaluated through ten independent sampling runs of 10 000 attempts per setting. Table 1 reports validity, uniqueness, and novelty as percentages of all generation attempts; unique and novel yields count distinct valid molecules, with novelty assessed against the complete reference dataset.

At $T = 1.0$, validity reached 95.58% for the small model and 99.68% for the large model, whereas distinct valid novel yields were 28.31% and 1.03%, respectively. Uniqueness closely followed validity, indicating that repeated molecules within individual runs were uncommon. The large model therefore generated valid CBUs more consistently, while the small model produced substantially more molecules absent from the enumerated reference collection.

Increasing the sampling temperature raised novel yield at the expense of validity. Between $T = 0.8$ and $T = 1.5$, novel yield increased from 18.76% to 41.37% for the small model and from 0.31% to 6.72% for the large model. Over the same interval, validity decreased from 98.64% to 73.52% and from 99.95% to 93.81%, respectively. Figure 1 illustrates this balance for the small model and distinguishes novelty relative to the training partition from novelty relative to the complete dataset. The latter excludes recovery of validation and test molecules and therefore measures expansion beyond the enumerated collection. The trade-off between validity and novel yield for both models is shown in the Supporting Information (Figure A.3).

At $T = 1.0$, novel molecules broadly retained the synthetic-accessibility and complexity profiles of the complete reference dataset, as assessed using SA score, [49] RAscore, [50] SCScore, [51] and SYBA [52] (Supporting Information, Section A.7 and Figure A.7). These similarities concern the proxy distributions and do not establish synthetic feasibility.

We used $T = 1.0$ for subsequent conditioning experiments because both models retained high validity, while the small model also produced a substantial yield of novel candidates.

3.2 Conditioning on Linker Type

CBU-type conditioning allows generation to target a construction class required by a selected assembly model. This is particularly relevant for the present dataset, in which 2-bent and 2-linear CBUs account for 63.75% and 36.19% of molecules, respectively, while only 0.054% are 3-planar. The complete dataset contains 11 850 3-planar molecules, of which 10 630 belong to the training partition.

At epoch 10 and $T = 1.0$, both models followed requests for the common classes with high accuracy (Figure 2). Among valid outputs, 99.54% and 99.49% matched the requested 2-linear and 2-bent types for the small model, compared with 99.94% and 99.82% for the

Table 1: *Unconditioned generation at sampling temperature T. Values are percentages of all attempts, reported as mean ± standard deviation across ten runs of 10 000 attempts. Unique and novel yields count distinct valid molecules; novelty is assessed against the complete reference dataset.*

Model	T	Valid (%)	Unique (%)	Novel (%)
Small	0.8	98.64 ± 0.14	98.60 ± 0.14	18.76 ± 0.73
	1.0	95.58 ± 0.16	95.56 ± 0.17	28.31 ± 0.59
	1.5	73.52 ± 0.37	73.51 ± 0.36	41.37 ± 0.70
Large	0.8	99.95 ± 0.02	99.89 ± 0.02	0.31 ± 0.05
	1.0	99.68 ± 0.05	99.65 ± 0.05	1.03 ± 0.10
	1.5	93.81 ± 0.29	93.72 ± 0.29	6.72 ± 0.27

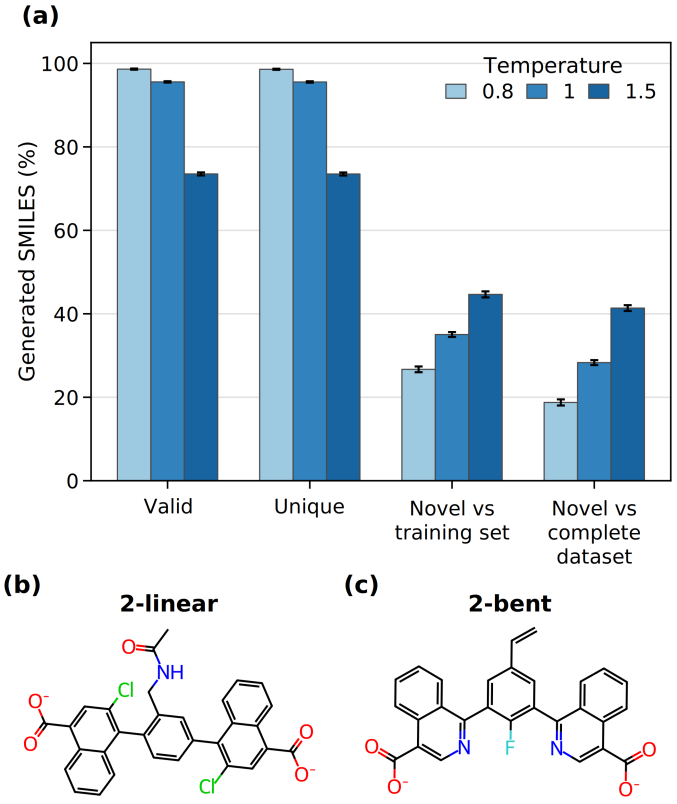


Figure 1: *Unconditioned generation. (a) Small-model validity, uniqueness, and novelty relative to the training partition and complete reference dataset. Bars show mean percentages of all attempts; error bars indicate standard deviations across ten runs of 10 000 attempts per temperature. (b,c) Example novel 2-linear and 2-bent CBUs, respectively.*

large model. The rare 3-planar class was more challenging: matching accuracy among valid outputs was 80.61% for the small model and 98.14% for the large model, while validity was 62.39% and 95.95%, respectively.

The distinction between accuracy and candidate yield was especially important for the 3-

planar request. Per 10 000 attempts, the small model generated an average of 4 317 distinct valid molecules of the requested type, including 2 292 absent from the complete reference dataset. The large model generated approximately 6 355 distinct valid matches, but only 715 were novel. Thus, the large model followed the requested class more reliably and produced more distinct matches, whereas the small model produced more novel matches.

CBU-type conditioning therefore provided access to the rare construction class despite its low representation in training.

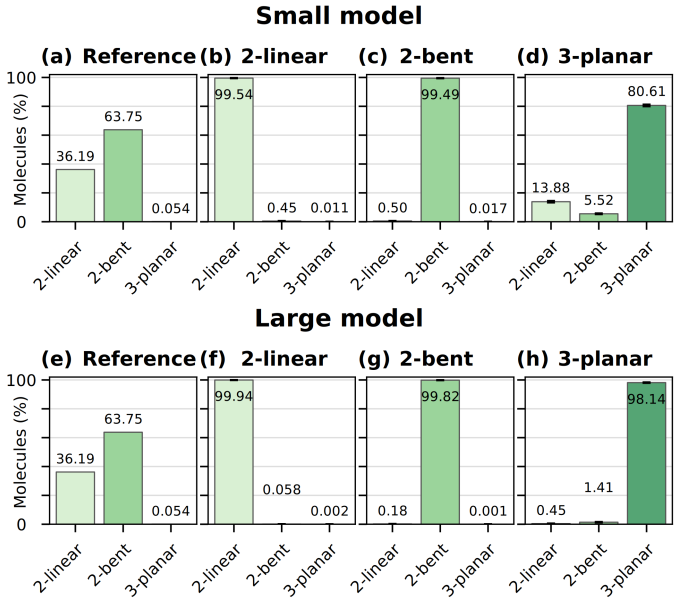


Figure 2: *CBU-type conditioning for the small (top, a–d) and large (bottom, e–h) models. Panels show the reference dataset composition (a, e) and valid-output compositions for requests specifying 2-linear (b, f), 2-bent (c, g), and 3-planar (d, h) CBUs. Generated percentages are means across ten independent runs of 10 000 attempts at $T = 1$; error bars indicate standard deviations. Numbers label the corresponding percentages.*

3.3 Conditioning on Linker Properties

We next examined whether generation could satisfy jointly specified structural and compositional requirements. Both models were conditioned on CBU type, binding-site graph distance, and the heteroatom fraction of the molecular core, excluding the carboxylate binding groups. All evaluated requests specified 2-linear CBUs. We tested binding-site distances of 5, 9, and 13 bonds at a fixed heteroatom fraction of 0.25, and heteroatom fractions of 0.25, 0.50, and 0.75 at a fixed distance of 13 bonds. A generated molecule was counted as a match if it was valid, belonged to the requested CBU type, had the exact requested binding-site distance, and had a heteroatom fraction within ± 0.05 of the target. Matching rates were calculated over all generation attempts, counting each occurrence of a matching molecule.

Both models shifted the generated distributions towards the requested values (Figure 3). All requests jointly specified CBU type, binding-site distance, and heteroatom fraction; the figure displays the two numerical-property distributions separately, whereas the matching rates below require each molecule to satisfy all three criteria simultaneously. Distance 13 accounts for 98.50% of 2-linear molecules in the training partition. For a request specifying this distance and a heteroatom fraction of 0.25, 96.35% of attempts satisfied all criteria for the small model, compared with 99.43% for the large model. Keeping the requested heteroatom fraction at 0.25 but reducing the distance to 9 or 5 bonds lowered these rates to 42.23% and 13.91% for the small model and 74.82% and 37.57% for the large model, respectively. These shorter-distance combinations have substantially less training support, with only 1 214 and 840 training molecules satisfying the respective joint criteria.

At distance 13, both models responded to heteroatom-fraction targets of 0.25 and 0.50, but neither reliably reached 0.75. For the latter request, joint-matching rates were only 0.112% and 0.016% of attempts for the small and large models, respectively. No 2-linear training molecules simultaneously satisfy distance 13 and a heteroatom fraction within ± 0.05 of 0.75. The poor performance therefore coincided with an unrepresented combination of requested properties.

The yield of distinct valid novel molecules satisfying all requested criteria also varied substantially between targets. For the small model, it reached 38.88% for distance 13 and heteroatom fraction 0.50, and 9.96% and 5.05% for distances 9 and 5 at heteroatom fraction 0.25. The extreme heteroatom-fraction request yielded only 0.079%, equivalent to approximately eight novel matching molecules per 10 000 attempts (Supporting Information, Table A.9).

Conditioning was also successfully extended to core molecular weight and cLogP (Supporting Information, Figure A.4 and Table A.9).

3.4 Text Description Conditioning

Text conditioning allows molecular requirements to be supplied through a single description. We evaluated models trained separately on an initial corpus of 60 000 molecule–description pairs and an expanded corpus with greater coverage. Descriptions were assembled from CBU types, molecular properties, elements, and functional groups.

The initial small model showed strong control over fluorination. Changing the prompt from *"A 2-linear linker"* to *"A 2-linear linker with fluoride functional groups"* increased the proportion of fluorine-containing molecules among valid 2-linear outputs from 15.12% to 98.88%. Across ten sampling runs of 10 000 attempts, validity also increased, from 85.59% to 95.18%. Enrichment occurred across several fluorinated motifs (Figure 4), showing that a simple textual request could substantially alter the generated chemistry.

More demanding requests revealed weaker control over property combinations and sensitivity to wording (Supporting Information, Section A.6). A description requesting strong binding of 5-fluorouracil predominantly promoted fluorinated structures. This response was consistent with an association between the wording and fluorinated molecular features, but did not establish control over host–guest binding.

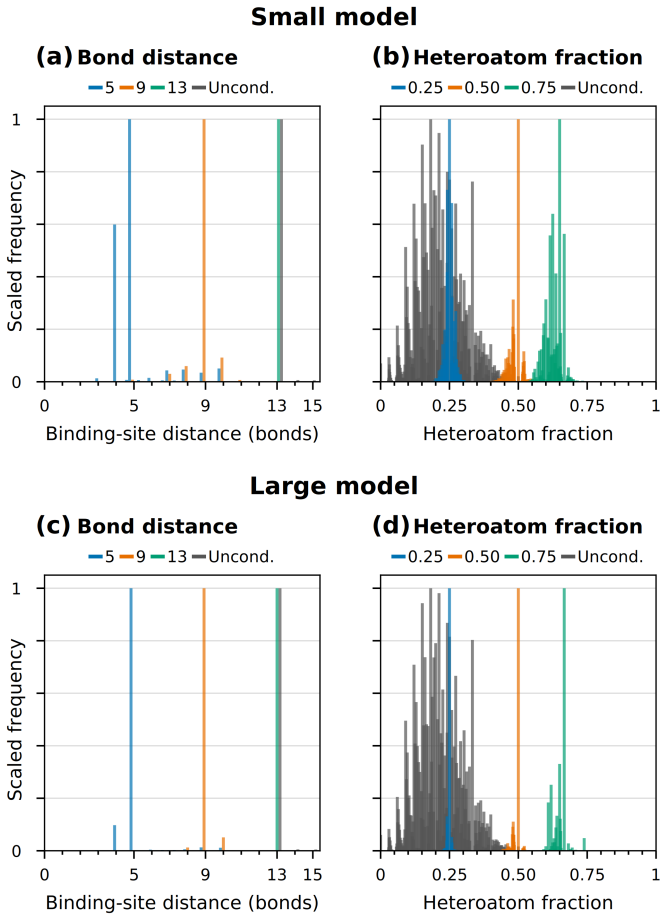


Figure 3: *Numerical conditioning for the small (top, a–b) and large (bottom, c–d) models. All requests jointly specify 2-linear CBU type, binding-site distance, and heteroatom fraction; panels show the resulting numerical-property distributions separately, including outputs that do not satisfy the complete request. Panels (a, c) vary the requested distance (5, 9, or 13 bonds) at a heteroatom fraction of 0.25; panels (b, d) vary the requested heteroatom fraction (0.25, 0.50, or 0.75) at a distance of 13 bonds. Grey distributions show unconditioned generation. Each distribution is independently scaled to a maximum of one; overlapping distance stems are offset for visibility.*

The expanded corpus contained approximately 1.21 million molecule–description pairs and increased coverage of five-bond and urea-containing 2-linear linkers. For both model sizes, epoch 15 was selected using validation loss and the number of distinct valid molecules generated under the research prompt introduced in Section 2.4. Checkpoint selection is detailed in the Supporting Information (Section A.2 and Table A.4). With these models, an explicit request for five bonds between binding sites was satisfied by 71.19% of valid 2-linear outputs for the small model and 97.66% for the large model.

Combining requirements remained sensitive to prompt formulation. For the research prompt, 35.11% and 21.31% of valid 2-linear outputs from the small and large models

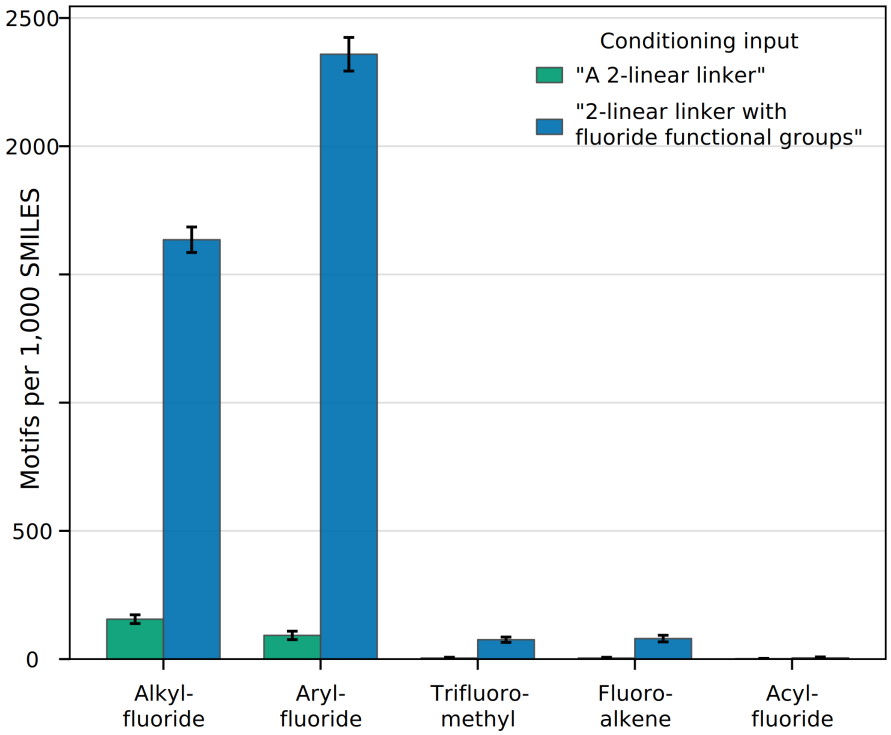


Figure 4: *Fluorinated-motif enrichment for the small model under generic and fluorination-specific prompts. Bars show mean motif counts in the first 1 000 attempts of each of ten independent runs; error bars indicate standard deviations. Molecules may contain multiple motif occurrences and contribute to several categories.*

satisfied both the five-bond distance and urea requirements. These percentages assess those two features rather than compliance with the complete description. A reformulation that changed the wording and order of the requirements reduced these proportions to 5.04% and 7.54%, respectively. Other variants also changed or omitted requirements, so the comparisons assess sensitivity to prompt formulation rather than word order alone (Supporting Information, Section A.6 and Figure A.5).

3.5 Conditioning Efficiency

To compare the practical benefit of the conditioning interfaces, we measured the yield of distinct valid novel molecules satisfying each request, expressed as a percentage of all generation attempts. Each qualifying molecule was counted once per run; invalid outputs, additional occurrences, molecules present in the complete reference dataset, and outputs failing the request did not contribute. We compared conditioned generation with unconditioned generation followed by identical request filters, using the small model at $T = 1.0$ and epoch-10 conditioned checkpoints.

The benefit varied substantially with the requested population (Figure 5). For CBU-type

conditioning, novel matching yield increased from 0.021% to 22.92% for 3-planar CBUs, corresponding to approximately two versus 2 292 candidates per 10 000 attempts. Improvements were smaller for the common classes: yields increased from 8.19% to 14.70% for 2-linear and from 20.09% to 24.17% for 2-bent CBUs. Numerical conditioning likewise increased access to rarely sampled combinations. For example, the distance-13 and heteroatom-fraction-0.50 request yielded 38.88% novel matches, compared with 0.091% after unconditioned generation and filtering.

Text conditioning with the initial-corpus model increased novel matching yield from 8.19% to 41.29% for the generic 2-linear request. Requests for amine-containing and fluorine-containing 2-linear linkers increased the corresponding yields from 2.13% to 23.13% and from 1.45% to 26.22%, respectively. These represent approximately 5.0-, 10.9-, and 18.1-fold improvements over the corresponding unconditioned controls.

Across the evaluated requests, conditioning increased the yield of distinct valid novel candidates satisfying the specified requirements. This comparison measures efficiency per generation attempt; it does not account for training or computational cost.

3.6 Prospective Application to MOP Hosts for 5-Fluorouracil

We examined whether literature-informed text conditioning could guide the generation of linkers for M_4L_6 MOP hosts for 5-fluorouracil. The research prompt requested a 2-linear linker with five bonds between binding sites, high polarity, medium functionalisation, a very low fraction of sp^3 carbons, and urea groups. We compared the small and large text-conditioned models receiving this prompt with the small model receiving the generic description *"A 2-linear linker"*. Additional comparisons used linkers proposed directly by GPT-5.6-Sol and GPT-6-Astra, each given the literature-review report (Supporting Information, Section A.8.1).

The three sampling runs for each LinkerGen condition supplied 900 selected candidates in total, while the two ChatGPT models supplied 100 candidates each. Each research-prompt run yielded at most two distinct accepted molecules satisfying all measurable requirements among 5 000 generation attempts. Nevertheless, among successfully assembled cages, the requested five-bond distance occurred in 93–97% of the small-model research sets and 82–85% of the large-model research sets, compared with 4–10% of the generic sets. The research prompt therefore strongly enriched short linkers despite limited fulfilment of the complete description (Supporting Information, Section A.8.2).

Assembly produced 822 cages from the LinkerGen candidates. After applying the common evaluation criteria, 626 were retained: 229 from generic generation, 221 from the small research-conditioned model, and 176 from the large research-conditioned model. These correspond to 76.3%, 73.7%, and 58.7% of the 300 selected candidates per condition, respectively. The non-LinkerGen GPT-5.6-Sol and GPT-6-Astra sets yielded 72 and 65 retained cages. Binding-energy comparisons refer to these retained populations; acceptance criteria and per-run outcomes are provided in the Supporting Information (Section A.8.3 and Table A.11).

Across the three runs, mean calculated binding energies ranged from -1.081 to -1.071 eV for generic generation, from -0.965 to -0.938 eV for the small research-conditioned

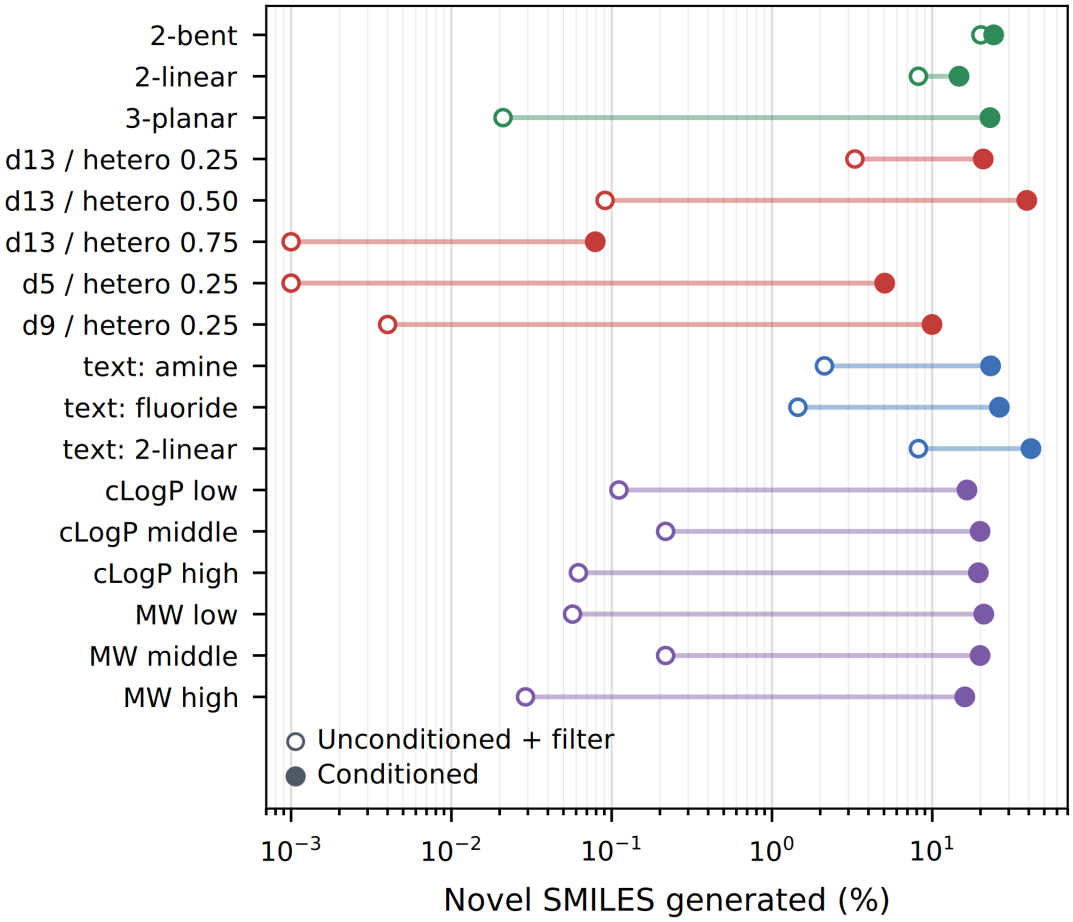


Figure 5: *Distinct valid novel matching yields for conditioned generation (filled markers) and unconditioned generation followed by identical filtering (open markers). Values are percentages of all attempts for the small model at $T = 1$ (epoch 10). Zero yields are placed at 0.001% on the logarithmic axis solely for display.*

model, and from −1.090 to −1.057 eV for the large research-conditioned model. The generic and large-model means remained close, while the small research-conditioned model gave less favourable means in every run. The corresponding means for the single GPT-5.6-Sol and GPT-6-Astra sets were −1.032 and −0.921 eV. Figure 6 compares the pooled binding-energy distributions; per-run means and confidence intervals are reported in the Supporting Information (Table A.11).

Despite its less favourable mean, the small research-conditioned model supplied the candidate with the lowest calculated binding energy, −2.139 eV, compared with minima of −2.095 and −1.993 eV for generic generation and the large research-conditioned model, respectively. This candidate had the requested five-bond distance but contained no urea group and had low core polarity under the conditioning definitions. Its optimised host–guest structure showed several short, directional N–H···O contacts between linker groups and the carbonyl oxygens of 5-fluorouracil (Figure 7).

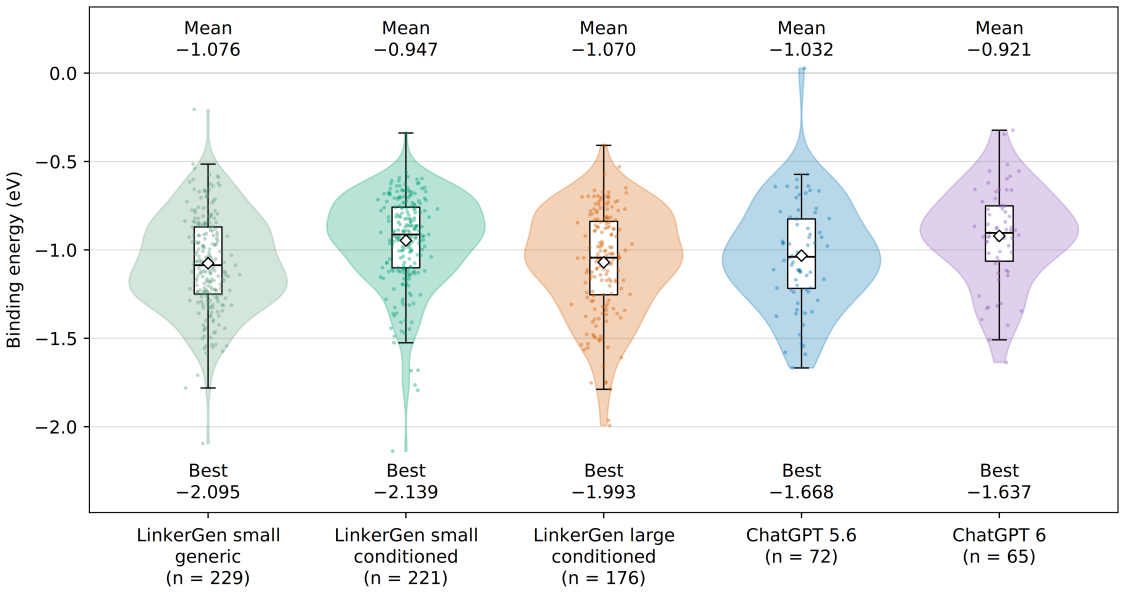


Figure 6: *Calculated 5-fluorouracil binding energies for accepted MOP hosts. LinkerGen distributions pool three sampling runs per condition; each ChatGPT distribution represents one set. Dots show individual evaluations, boxes the interquartile ranges, horizontal lines the medians, and diamonds the means. Counts, means, and minima are annotated.*

4 Discussion

LinkerGen makes selected regions of linker chemistry more accessible for computational screening. Even a large molecular collection may contain relatively few candidates satisfying a particular design request, making unrestricted generation followed by filtering inefficient. The enrichment observed for rare CBU types and selected property combinations shows how conditioning can direct sampling towards relevant candidates, including molecules absent from the reference collection. Enumeration and generation thus serve complementary roles: enumeration supplies a chemically structured foundation, while the trained model allows that chemistry to be explored under changing requirements. Within the tested settings, the small model's greater novel yield also shows that increasing model size does not necessarily improve the supply of new candidates.

Text conditioning offers an accessible way to express these requirements, but its usefulness depends on the connection between the language of a request and the chemistry represented in training. A description that appears reasonable to a chemist may combine features that rarely occur together in the available examples. The expanded corpus supported generation of short and urea-containing linkers, while the prompt comparisons showed that combining such requirements remained sensitive to their expression. Preparing molecular annotations and checking the generated structures are therefore part of using the interface effectively. Text provides a convenient means of communicating a design hypothesis; its fulfilment still has to be established from the outputs.

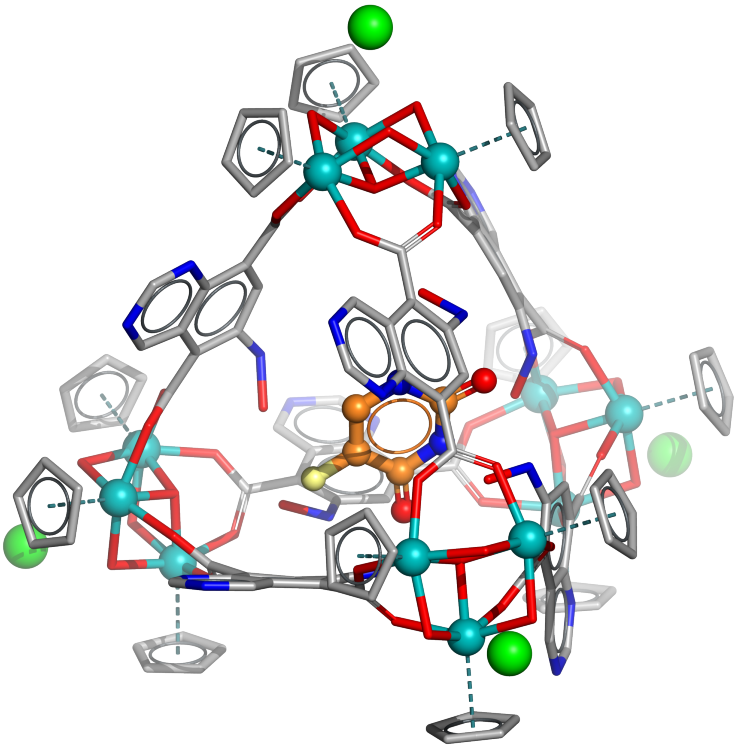


Figure 7: *Optimised M_4L_6 host containing 5-fluorouracil for the lowest-energy candidate identified ($E_{\text{bind}} = -2.139$ eV), generated by the small research-conditioned model. Guest carbon atoms are orange, zirconium atoms cyan, and chloride counterions green.*

The 5-fluorouracil application extends this distinction from molecular compliance to performance after assembly. LinkerGen supplied candidates with calculated binding energies competitive with those obtained from direct ChatGPT generation within the shared screening protocol. The small research-conditioned model produced the candidate with the lowest calculated energy despite having the least favourable mean among the LinkerGen conditions. Population averages and individual candidates therefore answer different screening questions: the former describe the typical accepted output, while the latter identify structures worth examining more closely. The lowest-energy candidate lacked the requested urea group, and its optimised structure instead showed short, directional contacts between other linker groups and the guest (Supporting Information, Section A.8.6). This provides a concrete example of a promising arrangement outside the complete set of features envisaged in the request.

Literature-informed conditioning can guide exploration, but it also favours some possibilities at the expense of others. Here, the research prompt strongly enriched short linkers, whereas generic generation explored a broader range of lengths and achieved mean binding energies similar to those of the large research-conditioned model. The selected requirements therefore directed the generated chemistry without providing a consistent advantage in the chosen performance measure. Few candidates satisfied the complete

description, however, so this comparison cannot separate the suitability of the chemical hypothesis from the model's ability to realise it. Broader exploration remains valuable alongside targeted generation because it can reveal favourable candidates outside the proposed requirements. These comparisons concern selected populations and accepted calculations, rather than a general ranking of generators; differences in workflow completion also matter when assessing how many candidates yield usable screening results.

The application also highlights that conditioning acts on individual linkers, whereas binding depends on the assembled host. Inorganic nodes contribute both the coordination geometry that organises the linkers and a chemical environment of their own, including metal centres and associated ligands. Together with the linkers and counterions, these components determine which interaction sites are accessible to the guest. All hosts in the present comparison shared the same node, allowing linker variation to be examined within one assembly family while leaving the contribution of node design unexplored. Considering alternative nodes alongside organic linkers could therefore reveal favourable combinations beyond those accessible through linker modification alone.

LinkerGen provides a component of such a design process, with its present capabilities established within a defined carboxylate-linker domain. Extension to other binding groups and linker families will require suitable training data and criteria for evaluating the resulting chemistry. Feedback from assembly and property calculations [44, 53] could inform subsequent requests or support optimisation through reinforcement learning.[24] This would allow downstream evaluation to improve both the candidates being generated and the chemical assumptions used to guide their generation.

5 Conclusions

In this work, we introduced LinkerGen, a compact SMILES-based autoregressive model for generating organic building units guided by CBU type, numerical properties, or text descriptions. Trained on approximately 22 million carboxylate linkers, LinkerGen generated molecules beyond the complete reference collection and increased the yield of distinct valid novel candidates meeting selected requirements compared with unconditioned generation followed by filtering. Model size and sampling temperature offered control over the balance between validity and novelty, while training coverage and prompt formulation influenced how reliably requests were fulfilled.

The prospective 5-fluorouracil application connected text-guided generation to MOP assembly and computational binding-energy screening. It illustrated the value of combining literature-informed guidance with broader exploration, allowing chemical design hypotheses to be tested while retaining access to candidates outside their assumptions. LinkerGen therefore provides a practical means of translating chemical design ideas into candidate structures whose suitability can be assessed in the assembled material.

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

M.K. gratefully acknowledges funding from the European Union's Horizon Europe research and innovation programme under grant agreements 101188248 (CLIMATE-ADAPT4EOSC), 101226137 (TOGETHER), and 101309693 (RESILIACT), as well as support from the Alexander von Humboldt Foundation and the Massachusetts Institute of Technology. P.W.V.B. acknowledges the support of the CARES office, where part of the research was conducted.

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data and code availability

The supplementary information contains additional details on the computational methods, example text descriptions, LLM prompts used, learning curves and example novel CBUs and fragments. The code for LinkerGen is available on GitHub under MIT license: https://github.com/pwvbutler/linkergen.The data and model checkpoints supporting this work are available on Zenodo at doi:10.5281/zenodo.23669416.

A Appendix

A.1 Dataset Construction and Characterisation

A.1.1 Starting Dataset

The starting dataset was reconstructed from the molecular fragments and 12 CBU templates introduced in our previous work.[20, 21] The source fragments were originally extracted from ChEMBL36 [39] using BRICS fragmentation. [40] Fragment preparation converted attachment markers to dummy atoms and applied the replacements and exclusions recorded in the accompanying fragment-selection rules.

Assembly was restricted to carboxylate binding groups and used a revised RDKit-based graph procedure. [41] Attachment markers were removed and their neighbouring atoms connected by single bonds, followed by molecular sanitisation. Products were checked for connectivity, the expected number of binding groups, and the absence of residual attachment markers. Stereochemical annotations were removed, and canonical SMILES were deduplicated across all templates. Independent validation confirmed that the resulting molecules were closed shell.

The final collection contained 22 074 197 unique molecules: 22 062 347 with two carboxylate groups and 11 850 with three. Reference CBU labels retained the construction classes associated with the source collection, with the cyclic-core and acyclic-core 2-bent subclasses combined into a single 2-bent class. Fragment preparation, assembly rules, validation scripts, and dataset checksums are included in the accompanying data and code described in Section A.9.

A.1.2 Dataset Splits

Canonical SMILES were assigned deterministically to training, validation, and test partitions using a SHA-256 hash with a fixed salt, a constant string combined with each SMILES before hashing. The resulting values were used to allocate molecules in approximately 90:5:5 proportions. Assignment therefore depends on molecular identity rather than file order.

The resulting partitions contained no shared canonical SMILES and were preserved during conditioned training, including text-corpus preparation. Table A.1 reports their composition.

Table A.1: *Molecule counts by reference CBU type and dataset partition.*

Partition	2-linear	2-bent	3-planar	Total
Training	7 189 937	12 665 738	10 630	19 866 305
Validation	399 793	704 121	610	1 104 524
Test	399 626	703 132	610	1 103 368
Complete	7 989 356	14 072 991	11 850	22 074 197

A.1.3 CBU-Type Assignment

Generated molecules were assigned CBU types using a fragment-based classification procedure. Anionic carboxylate groups were replaced by dummy atoms. Eligible single, non-ring bonds along shortest paths from the first binding-site dummy atom to the others were cut to identify constituent fragments. Eligible bonds had at least one ring-atom endpoint and did not involve hydrogen or dummy atoms.

Molecules containing a fragment with more than three attachment sites were rejected. A maximum of three attachment sites gave a 3-planar assignment, without an independent planarity test. Otherwise, attachment-vector angles were calculated from ETKDGv3 fragment conformers generated using a fixed random seed. [54, 55] Dummy atoms were temporarily replaced by isotopically labelled hydrogens during embedding. Mean pairwise angles above 160° identified linear fragments, while angles between 90° and 130° identified bent fragments. A recognised bent fragment took precedence over a linear fragment; molecules without a supported assignment were rejected.

These operational labels describe the construction classes used in evaluation. They do not independently establish the geometry or assembly compatibility of the complete linker.

A.1.4 Molecular Descriptors

For descriptor calculation, carboxylate or carboxylic-acid groups matching `C(=O)[O-,OH]` were replaced by dummy attachment atoms, and the resulting core was sanitised. Binding-site graph distance was the mean shortest-path distance over all pairs of attachment atoms. For molecules with two binding sites, this is a single distance measured in bonds and is equivalent to the distance between the corresponding carboxylate carbon atoms in the original graph.

Heteroatom fraction was the proportion of non-hydrogen core atoms that are not carbon, excluding dummy attachment atoms. Additional core descriptors were calculated using RDKit: molecular weight, cLogP, topological polar surface area (TPSA), sp^3 carbon fraction, rotatable-bond count, and ring count. Dummy attachment atoms were retained in these calculations. The sp^3 fraction denotes the proportion of core carbon atoms assigned sp^3 hybridisation.

For functional-group detection, a separate copy of the core was prepared with the dummy attachment atoms replaced by hydrogens. Functional groups were identified using the archived SMARTS definitions. Functionalisation denotes the number of detected functional-group labels, rather than the total number of individual group occurrences. Their use in text descriptions and the associated property categories are described in Section A.6.

A.1.5 Property Coverage

Table A.2 reports the training support for the numerical requests examined in the main text. Among 2-linear training molecules, 7 081 817 (98.50%) have a binding-site distance of 13 bonds. The tested five- and nine-bond combinations are much less represented, and no training molecules satisfy the distance-13, heteroatom-fraction-0.75 request within

either reported tolerance. The main generation comparisons use an absolute heteroatom-fraction tolerance of ± 0.05; the narrower ± 0.025 interval provides additional context.

Table A.2: *Training support among 2-linear CBUs. Distance matching is exact; joint counts also require heteroatom fraction (HF) within the indicated tolerance.*

Distance (bonds)	HF target	Distance only	Joint (± 0.025 HF)	Joint (± 0.05 HF)
5	0.25	2 390	403	840
9	0.25	3 571	593	1 214
13	0.25	7 081 817	1 405 269	2 765 581
13	0.50	7 081 817	13 538	32 422
13	0.75	7 081 817	0	0

A.2 Model Architecture and Training

A.2.1 Representation and Architecture

LinkerGen is an autoregressive transformer decoder that combines causal self-attention over the growing SMILES sequence with optional cross-attention to conditioning inputs (Figure A.1). [42] The vocabulary contains 57 tokens representing molecular symbols and special tokens for sequence boundaries, padding, and unknown inputs. The maximum sequence length is 512 tokens.

For CBU-type conditioning, the requested class is encoded as a one-hot vector representing 2-linear, 2-bent, or 3-planar. Numerical conditioning adds binding-site distance and core heteroatom fraction. Numerical inputs are standardised using training-set means and standard deviations, with the same transformation applied during validation, testing, and generation. A small feed-forward network maps these inputs to eight conditioning vectors accessed through cross-attention.

Text descriptions are embedded using the pretrained nach0 encoder, which produces 768-dimensional representations. [35] A trainable projection maps these representations into the decoder embedding space. The nach0 encoder remains frozen during training. Cross-attention is bypassed when generation is unconditioned.

The small and large configurations are summarised in Table A.3. Stored unconditioned-model parameter counts include inactive conditioning components; text-conditioned counts exclude the frozen nach0 encoder.

The CBU-type conditioned models contain 279 545 and 1 620 665 parameters for the small and large configurations, respectively. The corresponding numerical-property models contain 279 673 and 1 620 921 parameters. These differences reflect the conditioning networks used for each input type.

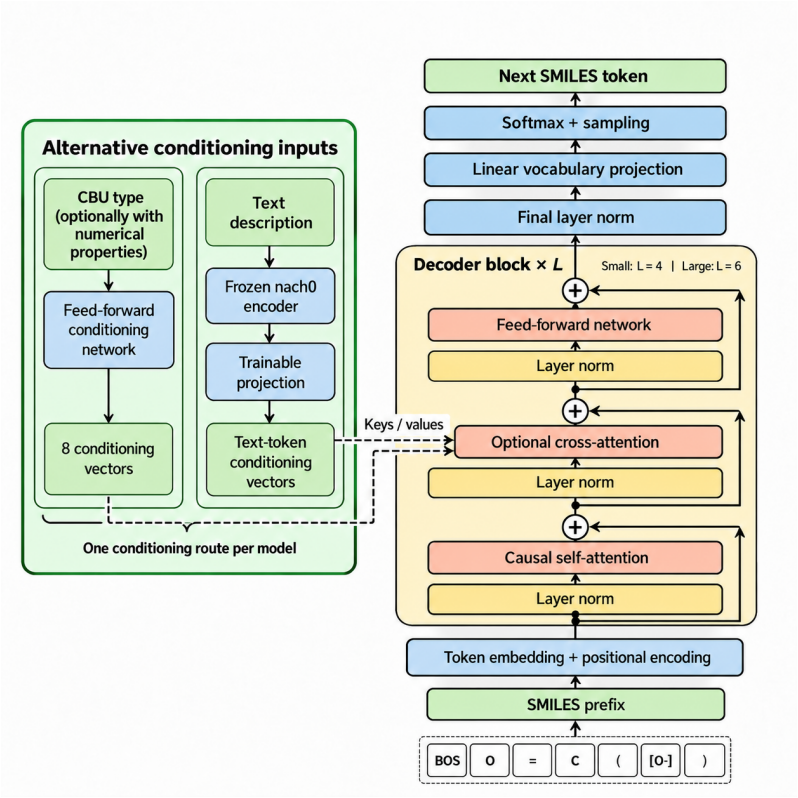


Figure A.1: *LinkerGen architecture. Each decoder block combines causal self-attention, optional cross-attention, and a feed-forward network. Conditioning uses either CBU-type and numerical inputs or text descriptions embedded by the frozen nach0 encoder. Cross-attention is bypassed during unconditioned generation.*

Table A.3: *Model configurations. d: embedding dimension; d_{ff}: feed-forward dimension; N_{base}: stored unconditioned parameters; N_{text}: trainable text-conditioned parameters.*

Model	d	d_{ff}	Heads	Layers	N_{base}	N_{text}
Small	64	256	4	4	279 865	327 993
Large	128	512	8	6	1 621 305	1 717 561

A.2.2 Training Procedure

Models were trained to minimise cross-entropy loss for next-token prediction, excluding padding tokens. Optimisation used AdamW [43] with a learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay of 0.01, and gradient clipping at 1.0. Training used a fixed random seed and mixed precision, with bfloat16 computations and float32 model weights and optimiser states. Validation loss was averaged over predicted tokens.

Both unconditioned models were trained for ten epochs, using batch sizes of 1 024 for the small model and 512 for the large model. Dropout rates were 0.05 for the embeddings and

0.1 within the decoder layers. The checkpoint with the lowest validation loss was selected, giving epoch 10 for the small model and epoch 9 for the large model (Figure A.2).

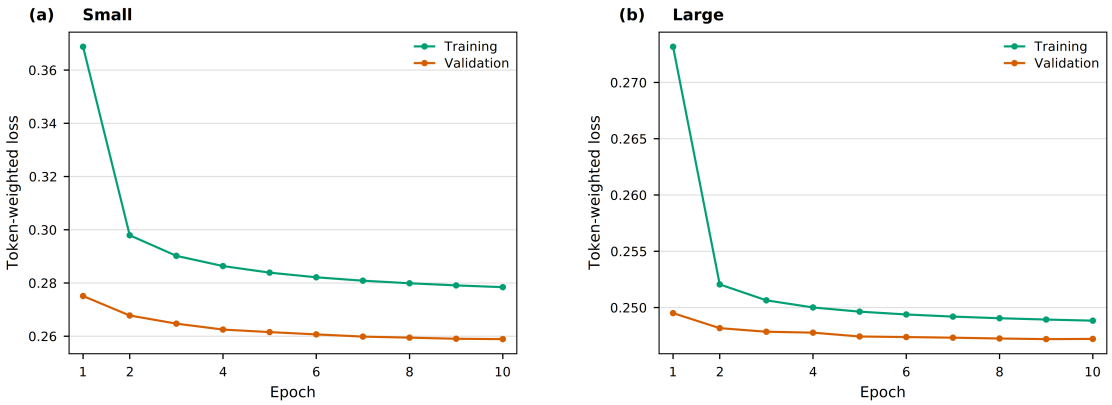


Figure A.2: *Training and validation losses for the small (left) and large (right) unconditioned models.*

The selected checkpoints initialised separate conditioned-training runs. For CBU-type and numerical conditioning, the conditioning networks, cross-attention layers, and associated layer normalisations were newly initialised. Both models were trained for ten additional epochs without class reweighting or oversampling. The epoch-10 checkpoints were used for evaluation.

Text-conditioned models also started from the selected unconditioned checkpoints, using a newly initialised text projection and the frozen nach0 encoder. Training on the initial description corpus used ten epochs and a batch size of 32; training on the expanded corpus used twenty epochs and a batch size of 64. These were separate training runs, rather than successive stages of one text-training procedure. Corpus construction is described in Section A.6.

A.2.3 Expanded-Corpus Checkpoint Selection

For the expanded-corpus text models, checkpoint selection considered both validation loss and the yield of distinct valid molecules. At each epoch, 1 000 generation attempts were made at $T = 1$ using the research prompt reported in Section A.8.1 and a fixed sampling seed.

Among checkpoints with validation loss no more than 5% above the minimum, the checkpoint producing the most distinct valid molecules was selected. This criterion selected epoch 15 for both model sizes (Table A.4). Selection did not use binding energies or complete request compliance. The selected checkpoints were used for the final prompt-formulation experiments and the prospective MOP application.

Table A.4: *Selected expanded-corpus text checkpoints. Generation counts refer to 1 000 attempts under the research prompt at $T = 1$.*

Model	Epoch	Validation loss	Valid outputs	Distinct valid molecules
Small	15	0.0973	628	360
Large	15	0.0690	667	131

A.3 Training and Generation Times

A.3.1 Hardware and Training Times

Training and generation used an NVIDIA GeForce RTX 5080 GPU with PyTorch 2.13.0 and CUDA 13.0. Training used bfloat16 mixed precision; generation and molecular evaluation used Python 3.12.3 and RDKit 2026.03.4.

Approximate training times are reported in Table A.5. Each conditioned model was initialised separately from the corresponding unconditioned checkpoint, so these durations represent additional training time. Text-training times include encoding descriptions with the frozen nach0 encoder.

Table A.5: *Approximate training times on an NVIDIA GeForce RTX 5080. Conditioned-training times exclude prior unconditioned training.*

Training configuration	Epochs	Batch size	Time (h)
Unconditioned, small	10	1 024	1.5
Unconditioned, large	10	512	3.9[a]
CBU type, small	10	1 024	2.0
CBU type, large	10	512	5.2
Numerical, small	10	1 024	2.0
Numerical, large	10	512	5.2
Initial text, small	10	32	0.2
Initial text, large	10	32	0.2
Expanded text, small	20	64	5.1
Expanded text, large	20	64	5.8

[a]Estimated duration.

A.3.2 Generation and Assessment Times

Unconditioned generation was timed using a sampling batch size of 64. Timings include SMILES generation, transfer to the CPU, and molecular assessment: RDKit parsing, canonicalisation, CBU-type assignment, and synthetic-accessibility (SA) scoring. Model loading and novelty assessment against the reference dataset are excluded.

Generating and assessing 10 000 attempts took approximately 82–87 seconds for the small model and 94–102 seconds for the large model (Table A.6). These measurements describe the combined generation and assessment workflow; conditioned-generation timings were not included in this comparison.

Table A.6: *Unconditioned generation and assessment times for 10 000 attempts, reported as mean ± standard deviation across ten runs. Throughput is calculated from the mean elapsed time.*

Model	T	Time (s)	Attempts s^{-1}
Small	0.8	82.3 ± 8.5	121.5
	1.0	87.3 ± 4.3	114.6
	1.5	85.5 ± 13.4	117.0
Large	0.8	101.7 ± 3.0	98.3
	1.0	99.9 ± 1.8	100.1
	1.5	93.7 ± 7.1	106.7

A.4 Unconditioned Generation and Novelty

Table A.7 extends the main-text results by including greedy generation ($T = 0$) and novelty relative to the training partition. At $T = 0$, each model repeatedly generated a single valid molecule present in the training partition, giving a distinct valid yield of 0.01% and zero novelty.

Novelty relative to the training partition includes molecules present in the validation and test partitions. Novelty against the complete reference dataset excludes these molecules and therefore measures generation beyond the enumerated collection.

Increasing sampling temperature raised distinct valid novel yield at the expense of validity. Across the tested non-zero temperatures, the small model generated more molecules absent from the complete reference dataset, whereas the large model retained higher validity (Figure A.3).

Table A.7: *Unconditioned generation across ten runs of 10 000 attempts: mean ± standard deviation. All values are percentages of attempts. Unique and novel yields count distinct valid molecules within each run; novelty uses the indicated reference set.*

Model	T	Valid	Unique	Novel (training)	Novel (complete)
Small	0.0	100.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	0.8	98.64 ± 0.14	98.60 ± 0.14	26.69 ± 0.69	18.76 ± 0.73
	1.0	95.58 ± 0.16	95.56 ± 0.17	35.04 ± 0.59	28.31 ± 0.59
	1.5	73.52 ± 0.37	73.51 ± 0.36	44.65 ± 0.72	41.37 ± 0.70
Large	0.0	100.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	0.8	99.95 ± 0.02	99.89 ± 0.02	10.22 ± 0.25	0.31 ± 0.05
	1.0	99.68 ± 0.05	99.65 ± 0.05	10.92 ± 0.30	1.03 ± 0.10
	1.5	93.81 ± 0.29	93.72 ± 0.29	15.42 ± 0.33	6.72 ± 0.27

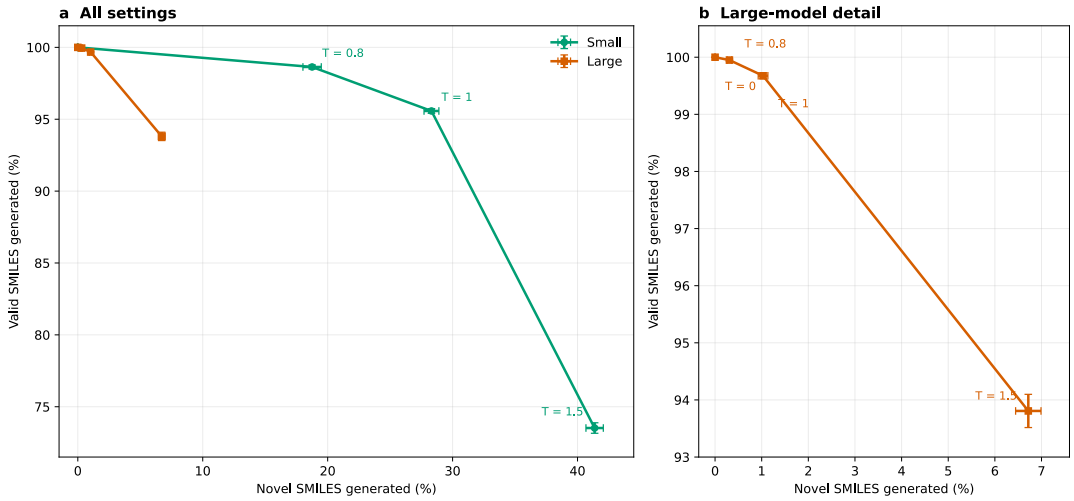


Figure A.3: *Validity–novelty trade-off across sampling temperatures for both models. Validity and distinct valid novel yield are percentages of all attempts, with novelty assessed against the complete reference dataset.*

A.5 Conditioned Generation

A.5.1 Numerical-Property Conditioning

Table A.8 reports generation conditioned on CBU type, binding-site graph distance, and core heteroatom fraction (HF). All requests specified 2-linear CBUs. A matching output was a valid molecule of the requested type with the exact requested distance and HF within ±0.05 of the target. Matching rates count every occurrence of a matching molecule, whereas unique yields count distinct valid molecules within each run. Both are expressed as percentages of all generation attempts.

Table A.8: *Numerical conditioning at epoch 10 and $T = 1$: mean $\pm$ standard deviation across ten runs of 10 000 attempts. All values are percentages of attempts; joint matching includes repeated outputs.*

Model	Distance (bonds)	HF	Valid	Unique	Joint matching
Small	5	0.25	45.89 ± 0.56	30.09 ± 0.31	13.91 ± 0.41
	9	0.25	80.03 ± 0.44	45.48 ± 0.27	42.23 ± 0.61
	13	0.25	97.24 ± 0.14	96.89 ± 0.13	96.35 ± 0.18
	13	0.50	91.45 ± 0.23	81.12 ± 0.36	85.07 ± 0.30
	13	0.75	48.37 ± 0.55	17.52 ± 0.20	0.11 ± 0.03
Large	5	0.25	56.44 ± 0.35	26.93 ± 0.15	37.57 ± 0.44
	9	0.25	92.33 ± 0.35	31.36 ± 0.27	74.82 ± 0.38
	13	0.25	99.61 ± 0.09	97.33 ± 0.16	99.43 ± 0.12
	13	0.50	99.13 ± 0.08	55.47 ± 0.33	96.64 ± 0.16
	13	0.75	75.19 ± 0.38	4.93 ± 0.23	0.02 ± 0.02

A.5.2 Additional Property Targets

The numerical conditioning inputs were extended to include core molecular weight (MW) and cLogP, and the small model was trained for ten conditioned epochs. All evaluated requests specified 2-linear CBUs, a binding-site distance of 13 bonds, and HF of 0.25. MW targets were varied at a fixed cLogP of 4.411, and cLogP targets at a fixed MW of 467.216 Da (Figure A.4). The two series therefore share the central request of MW 467.216 Da and cLogP 4.411.

Each condition was evaluated through three sampling runs of 10 000 attempts at $T = 1$. Matching required the requested type and exact distance, with tolerances of ± 0.05 for HF, ± 25 Da for MW, and ± 0.5 for cLogP.

A.5.3 Conditioning versus Filtering

Table A.9 compares the yield of distinct valid novel molecules satisfying each request with that obtained from unconditioned generation followed by identical filtering. All comparisons use the small model at $T = 1$ and epoch-10 conditioned checkpoints. Text-conditioned results use the initial description corpus.

Molecules were deduplicated within each run, and novelty was assessed against the complete reference dataset. Yields are percentages of all generation attempts. For the MW/cLogP requests, all fixed requirements and tolerances specified above were applied jointly. Their shared central request is reported once.

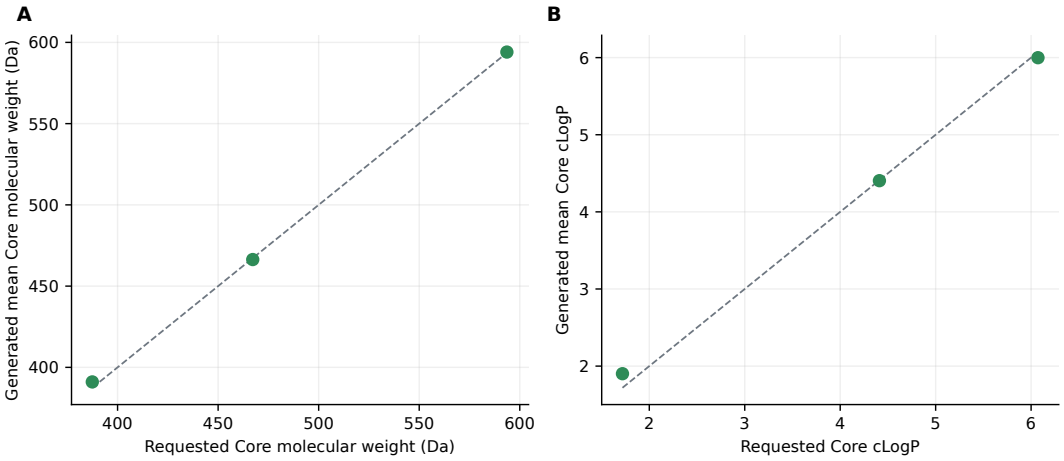


Figure A.4: *Core MW and cLogP conditioning for the small model at epoch 10 and $T = 1$. MW targets were 387.432, 467.216, and 593.592 Da; cLogP targets were 1.719, 4.411, and 6.072. Results pool three runs of 10 000 attempts per condition.*

Table A.9: *Distinct valid novel matching yields (% of attempts), reported as mean ± standard deviation. Each run contains 10 000 attempts. The final column gives the number of conditioned runs; all unconditioned controls use ten runs. Zero denotes no observed qualifying molecules.*

Request	Conditioned	Unconditioned + filtering	Conditioned runs
2-bent	24.166 ± 0.749	20.093 ± 0.381	10
2-linear	14.696 ± 0.399	8.194 ± 0.354	10
3-planar	22.923 ± 0.272	0.021 ± 0.017	10
2-linear, $d = 5$, HF = 0.25	5.053 ± 0.207	0.000 ± 0.000	10
2-linear, $d = 9$, HF = 0.25	9.963 ± 0.196	0.004 ± 0.005	10
2-linear, $d = 13$, HF = 0.25	20.810 ± 0.334	3.291 ± 0.239	10
2-linear, $d = 13$, HF = 0.50	38.876 ± 0.418	0.091 ± 0.036	10
2-linear, $d = 13$, HF = 0.75	0.079 ± 0.019	0.000 ± 0.000	10
MW = 387.432 Da	20.967 ± 0.424	0.057 ± 0.026	3
MW = 467.216 Da, cLogP = 4.411	19.880 ± 0.328	0.217 ± 0.037	3
MW = 593.592 Da	15.963 ± 0.038	0.029 ± 0.014	3
cLogP = 1.719	16.470 ± 0.403	0.111 ± 0.036	3
cLogP = 6.072	19.390 ± 0.170	0.062 ± 0.026	3
Text: 2-linear	41.291 ± 0.562	8.194 ± 0.354	10
Text: 2-linear + amine	23.125 ± 0.464	2.125 ± 0.127	10
Text: 2-linear + fluorine	26.222 ± 0.299	1.451 ± 0.096	10

A.6 Text Conditioning and Robustness

A.6.1 Training Descriptions

Descriptions combined CBU type, core molecular properties, elements, and functional groups. Properties were expressed numerically or through qualitative categories ranging from very low to very high. Polarity refers to core topological polar surface area (TPSA), flexibility to the rotatable-bond count, and lipophilicity to cLogP. Functional groups were identified using SMARTS patterns; functionalisation denotes the number of detected functional-group labels rather than the number of individual group occurrences.

The functional-group and text-conditioning-versus-filtering experiments used models trained on an initial corpus of 60 000 molecule–description pairs. The final prompt-formulation and application experiments used an expanded corpus containing 1 209 591 training pairs from 581 051 distinct molecules, with 1 703 validation and 1 701 test pairs. Molecular training, validation, and test partitions remained separate, including when several descriptions were generated for the same molecule.

The expanded corpus contained 100 000 training pairs from each of the 12 construction templates, supplemented with five-bond and urea-containing 2-linear linkers. Molecules from smaller template populations were reused with varied descriptions. Only three molecules in the complete reference dataset jointly satisfied the 2-linear, five-bond, and urea criteria; these were included without repetition. Thus, improved coverage of the individual features did not provide extensive training support for their combination.

Expanded-corpus descriptions used four sentence templates, with varied property selection, numerical or categorical expression, and ordering. Qualitative categories were defined using the 20th, 40th, 60th, and 80th percentiles of the training candidate pool (Table A.10). These are dataset-specific categories rather than universal chemical thresholds. Description templates and the corpus-specific annotation settings are included with the accompanying data.

Table A.10: *Category boundaries for expanded-corpus descriptions. The four boundaries separate very low, low, medium, high, and very high values; values equal to a boundary enter the higher category.*

Descriptor	b_1	b_2	b_3	b_4
Binding-site distance (bonds)	11	12	13	13
Heteroatom fraction	0.1515	0.2069	0.2581	0.3200
cLogP	1.357	2.771	4.030	5.595
TPSA (Å^2)	38.38	65.34	91.50	124.38
sp^3 carbon fraction	0.0800	0.1500	0.2273	0.3333
Ring count	3	4	4	5
Rotatable-bond count	3	4	6	8
Molecular weight (Da)	368.385	406.0496	445.059	560.025
Functional-group label count	3	3	4	5

Repeated boundaries arise from discrete descriptor values and leave some categories

empty. For example, low flexibility corresponds to three rotatable bonds, whereas medium polarity corresponds to 65.34 Å$^2 \leq$ TPSA $<$ 91.50 Å^2.

A.6.2 Sensitivity to Wording

The initial models were tested with four requests: 2-linear type, a five-bond distance combined with heteroatom fraction 0.25, 3-planar type combined with low molecular weight and high lipophilicity, and fluorinated 2-linear linkers. Each request was expressed in five styles: training-style, paraphrased, technical, terse, and reordered. At epoch 10, each formulation was evaluated through three runs of 10 000 attempts at $T = 1$. Type and fluorination requests were followed more consistently than requests combining several properties.

Separate diagnostic tests used one run of 1 000 attempts per prompt and model at epoch 10. For the request *"Generate a carboxylate CBU that is not 2-linear."*, 97.78% and 97.95% of valid outputs remained 2-linear for the small and large models, respectively. This test showed that the negative instruction was not reliably followed.

The final expanded-corpus models were evaluated at epoch 15 using ten runs of 10 000 attempts per prompt at $T = 1$ (Figure A.5). The research prompt is reproduced in Section A.8.1. Its reordered reformulation was:

> *"With inward-facing urea groups, very low sp3 fraction, medium functionalisation, and high polarity, generate a 2-linear linker with 5 bonds between its binding sites."*

Among valid 2-linear outputs, the proportion satisfying both the five-bond distance and urea requirements fell from 35.11% to 5.04% for the small model and from 21.31% to 7.54% for the large model under this reformulation. This comparison changes both clause order and wording, so it does not isolate an effect of order alone. Other reformulations also omitted or softened requirements and therefore assess sensitivity to the formulation of the request.

Matching percentages here count repeated outputs and use valid 2-linear molecules as the denominator. They assess only distance and urea presence; full application-request compliance is evaluated separately in Section A.8.2. Exact prompts and results for all formulations are included with the accompanying data.

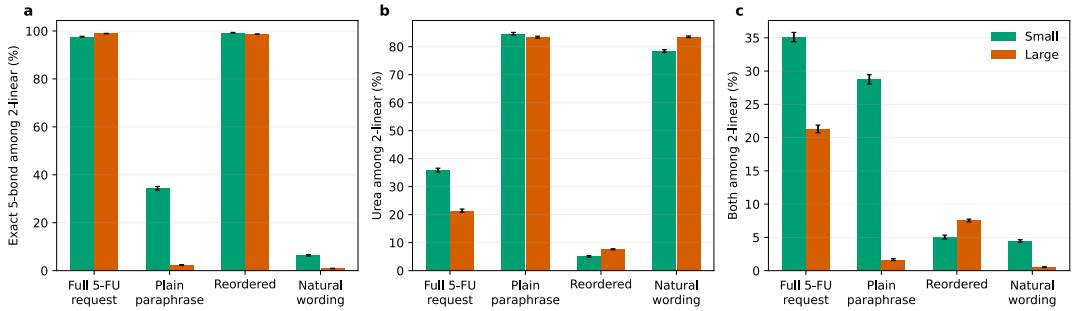


Figure A.5: *Sensitivity of the expanded-corpus models to research-prompt reformulations at epoch 15. Joint matching denotes a five-bond distance and urea group among valid 2-linear outputs. Each prompt uses ten runs of 10 000 attempts at* $T = 1$.

A.7 Synthetic Accessibility Proxies

Four proxies were used to assess generated candidates. SA score combines fragment contributions and molecular complexity, while SCScore estimates synthetic complexity from reaction data; lower values are favourable. [49, 51] RAscore predicts the likelihood of a retrosynthetic planner finding a route, while SYBA evaluates fragment-based synthetic accessibility; higher values are favourable. [50, 52]

Scores were calculated for distinct valid molecules absent from the complete reference dataset, pooled and deduplicated across ten sampling runs per model and temperature. SA scores used the generated carboxylate structures; RAscore, SCScore, and SYBA used structures processed with RDKit's Uncharger. These scores were applied after generation and did not determine validity or novelty.

Figure A.6 summarises the scores across sampling temperatures. For the small model, increasing temperature slightly raised median SA scores, whereas median RAscore and SYBA increased and median SCScore decreased. The proxies therefore did not indicate a uniform change in predicted accessibility. SYBA medians remained negative for all settings.

All 22 074 197 molecules in the complete reference dataset were scored using the same molecular preparation and scoring procedures, without subsampling. Figure A.7 compares this baseline with distinct valid novel molecules generated at $T = 1$: 28 297 for the small model and 1 029 for the large model. Each molecule contributed once to its respective population, irrespective of its generation frequency.

The generated populations broadly followed the reference distributions across all four proxies, with modest shifts towards higher SA scores and lower RAscores. SCScore shifted slightly towards higher values for the small model and lower values for the large model, while the SYBA distributions closely overlapped the reference distribution. Generation beyond the reference collection therefore largely preserved these proxy profiles, although the scores do not establish feasible synthesis routes.

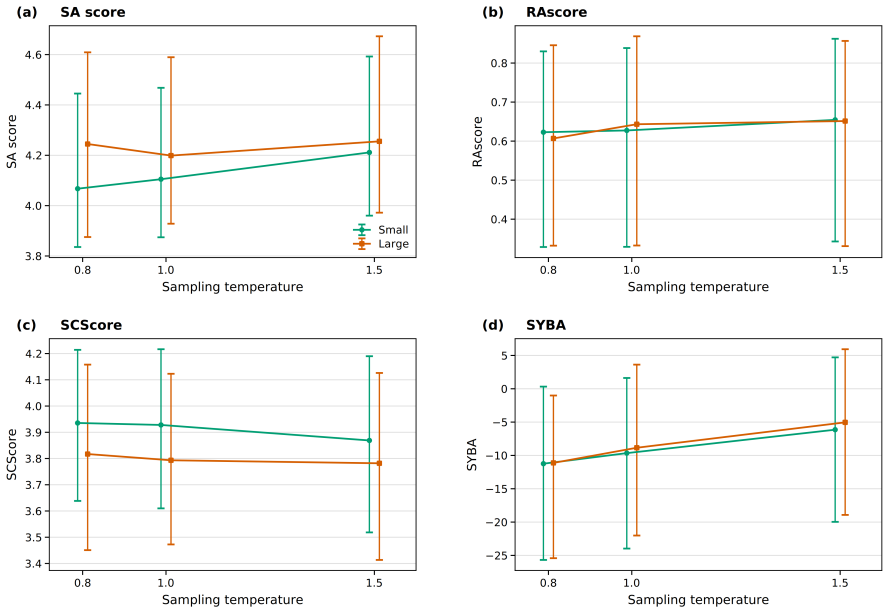


Figure A.6: *Synthetic-accessibility proxies across sampling temperatures for distinct valid novel molecules. Points show medians and bars span the interquartile range. Molecules were pooled and deduplicated across ten runs per setting.*

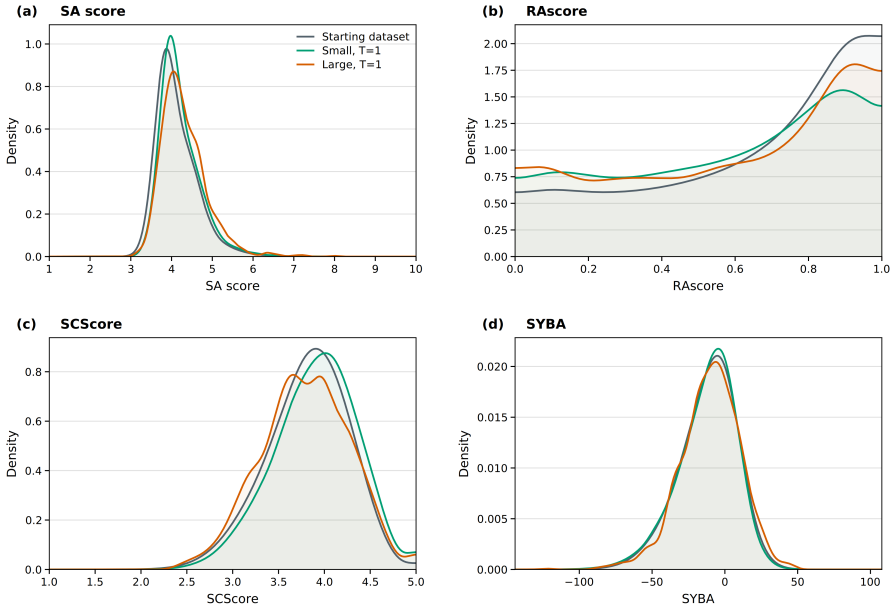


Figure A.7: *Synthetic-accessibility proxies for the complete reference dataset ($n = 22\,074\,197$) and distinct valid novel molecules generated at $T = 1$ by the small ($n = 28\,297$) and large ($n = 1\,029$) models. Generated populations pool ten runs of 10 000 attempts each. Densities integrate to one and share a smoothing bandwidth within each panel.*

A.8 Prospective 5-Fluorouracil Application

A.8.1 Literature Context and Generation Prompts

An LLM-assisted literature review informed LinkerGen conditioning and direct generation with GPT-5.6-Sol and GPT-6-Astra. It supplied a chemical design hypothesis; neither LinkerGen nor its text encoder was trained to predict host–guest binding energies. The complete review and generation requests are included with the accompanying data.

Both LinkerGen models received:

> *A 2-linear linker with 5 bonds between binding sites, high polarity, medium functionalisation, and very low sp^3 carbon fraction, incorporating inward-facing urea functional groups.*

The small-model generic control received *"A 2-linear linker"*.

The direct ChatGPT request asked for 100 SMILES strings for 2-linear carboxylate linkers expected to favour non-reactive binding of 5-fluorouracil in an M_4L_6 tetrahedral cage. It included reference cavity volumes of 477.216, 1188.62, and 2172.616 Å^3 and inner-sphere diameters of 9.657, 12.475, and 14.417 Å for unfunctionalised linkers containing one, two, and three benzene rings, respectively. The corresponding binding-site distances were 5, 9, and 13 bonds. The supplied guest van der Waals volume was 98.52 Å^3. Thus, both generation approaches used the same literature context, but their requests were not identical.

A.8.2 Candidate Selection and Workflow Outcomes

The small generic, small research-conditioned, and large research-conditioned models each supplied three independent runs of 5 000 generation attempts at $T = 1$, using the expanded-corpus epoch-15 checkpoints. Each run supplied 100 selected candidates, giving 900 LinkerGen candidate occurrences. Each ChatGPT model supplied one set of 100 candidates.

Structures were converted to carboxylates and deduplicated within each pool after removing explicit hydrogens and stereochemical annotations. Accepted molecules were closed shell, contained exactly two carboxylate sites, had net charge -2, received a supported CBU-type assignment, and contained no anionic pyrazolate motif. Research-prompt sets included all candidates meeting the operational compliance criteria, followed by fixed-seed random selection from the remaining accepted molecules. Generic sets were selected randomly from accepted 2-linear molecules. Selection preceded energy calculations, and failed candidates were not replaced. Occurrences in different sampling runs were retained separately.

Operational compliance required a 2-linear type, a binding-site distance rounding to five bonds, a urea motif, core TPSA satisfying $91.5 \leq \text{TPSA} < 124.38$ Å^2, medium functionalisation, and an sp^3 carbon fraction below 0.08. Functional-group orientation after assembly was not assessed. Each model supplied one, zero, and two compliant candidates in runs 1, 2, and 3, respectively.

Assembly produced 822 LinkerGen cages, of which 687 passed the computational workflow checks. Subsequent linker-integrity and small-ring filters excluded 22 and 39 evaluations, respectively, leaving 626. The ChatGPT sets yielded 72 and 65 retained evaluations (Table A.11).

Binding-energy means and 95% confidence intervals were calculated for retained evaluations only. Intervals use 10 000 bootstrap resamples within each set and the 2.5th and 97.5th percentiles of the resampled means. They describe within-set variation, excluding uncertainty from the energy model and conformational search. Pooled LinkerGen results retain occurrences across runs and therefore do not represent globally unique candidate sets.

Table A.11: *Application outcomes. Each row starts from 100 selected candidates. The $d = 5$ column counts five-bond linkers among assembled cages. Energies and bootstrap intervals use retained evaluations.*

Condition	Run	Assembled	$d = 5$	Retained	Mean E_{bind} (eV)	95% CI (eV)
Small generic	1	100	4	78	-1.081	$[-1.139, -1.022]$
Small generic	2	97	10	73	-1.075	$[-1.134, -1.017]$
Small generic	3	98	6	78	-1.071	$[-1.135, -1.013]$
Small conditioned	1	96	91	73	-0.938	$[-0.996, -0.882]$
Small conditioned	2	95	88	75	-0.938	$[-1.002, -0.879]$
Small conditioned	3	96	93	73	-0.965	$[-1.025, -0.909]$
Large conditioned	1	82	69	58	-1.057	$[-1.134, -0.982]$
Large conditioned	2	80	66	58	-1.063	$[-1.143, -0.988]$
Large conditioned	3	78	66	60	-1.090	$[-1.160, -1.021]$
GPT-5.6-Sol	Single	82	—	72	-1.032	$[-1.099, -0.963]$
GPT-6-Astra	Single	88	—	65	-0.921	$[-0.988, -0.857]$

A.8.3 Assembly and Host–Guest Calculations

A previously reported assembly workflow constructed tetrahedral cages from six organic linkers and four $[Zr_3O(OH)_3(C_5H_5)_3]$ metal CBUs, using assembly-model and metal-CBU data from The World Avatar. [44, 45] Linker conformers were generated using ETKDGv3 and optimised with MMFF94s, or UFF when MMFF parameters were unavailable. [54, 55] Embedding used random-coordinate retries when necessary. Binding-group coplanarisation was attempted before alignment with the prescribed cage geometry and overlap checks. Failed assemblies were retried with alternative conformers, up to ten times.

Calculations used `mace-matpes-r2scan-0` with D3(BJ) dispersion using r2SCAN parameters, float64 arithmetic, and no solvent. [46–48] Four chloride counterions neutralised each cage. Each chloride was initially placed outside a Zr_3 node towards its bridging hydroxyl groups, with a mean Cl$\cdots$O distance of 3.1 Å. Optimisation used ASE PreconLBFGS with an exponential preconditioner ($A = 3$), Armijo line search, a force threshold of 0.01 eV Å^{-1}, and a maximum of 5 000 steps, with FIRE as a fallback after optimiser exceptions. [56, 57]

The counterion-containing host and free guest were first relaxed separately. Guest placement sampled the cage centre, four node-directed sites, four window-directed sites, and

six linker-directed sites. Off-centre sites lay halfway along the available cavity radius. Sampling 48 super-Fibonacci orientations per site gave up to 720 placements per cage. Placements with intermolecular contacts below 1.8 Å were discarded. Up to three lowest-energy placements separated by at least 1.0 Å guest-coordinate RMSD were then relaxed.

Accepted complexes had to converge, preserve the checked internal host and guest connectivity, and place the guest centre of mass inside the tetrahedron defined by the four node centroids. Each chloride had to remain within 6.0 Å of a node. Cyclopentadienyl slippage was excluded using a maximum associated Zr–C distance of 3.15 Å. Other internal bonds were checked using 1.15 times the sum of covalent radii. These checks did not constitute an exhaustive search for new host–guest bonds. Empty-host references also had to converge and pass the host-integrity and counterion checks.

To improve the empty-host reference, hosts were extracted from relaxed complexes meeting the host-integrity and counterion criteria. The extracted host with the lowest single-point energy was relaxed without the guest. Its energy replaced the initial empty-host reference if lower and if the required integrity and convergence checks were satisfied. This reference could originate from a different pose from the accepted complex used for binding.

Binding energies were calculated as

$$E_{\mathrm{bind}} = E^{*}_{HG} - E^{\mathrm{ref}}_{H} - E^{0}_{G}, \tag{A.1}$$

where E^{*}_{HG} is the lowest-energy accepted complex, E^{ref}_{H} is the selected empty-host energy including counterions, and E^{0}_{G} is the relaxed free-guest energy. These are gas-phase screening energies rather than solution binding free energies.

The final comparison additionally required each linker's heavy-atom skeleton to match its input structure, allowing proton transfer and tautomeric changes. Linkers containing three- or four-membered rings involving N or O were excluded by an explicit selection rule.

A.8.4 Cage Windows and Final Guest Location

Geometry analyses used the 822 assembled LinkerGen cages, including cases excluded from the final binding-energy comparison. The four Zr_3 node centroids defined a tetrahedron whose faces supplied the window planes. Window dimensions were measured on the relaxed empty-host reference geometries.

For a point $\mathbf{q}$ in a window plane, the available radius was

$$r_{\mathrm{free}}(\mathbf{q}) = \min_{i} \left(\|\mathbf{q} - \mathbf{x}_i\| - r_i^{\mathrm{vdW}} \right), \tag{A.2}$$

where $\mathbf{x}_i$ and r_i^{vdW} denote host-atom positions and assigned van der Waals radii. All host atoms, including hydrogens and counterions, contributed. A face grid with nominal spacing 0.25 Å was followed by local refinement in the face plane. The reported diameter was twice the largest available radius found, and each cage was represented by its largest window. These are static clearance estimates, not passage barriers. The radius assignments and search settings are included with the accompanying analysis.

Guest location was assessed on each relaxed complex. A pose was classified as inside only if its guest centre of mass lay strictly inside all four node-centroid face planes. The outside fraction uses all relaxed poses, including those failing other acceptance criteria.

There were 2 416 relaxed poses across 806 cages; 16 cages supplied no relaxed pose. Of these poses, 220 failed the location criterion. Window dimensions and outside fractions differed with linker distance (Table A.12; Figure A.8). The Spearman correlation between largest-window diameter and per-cage outside fraction was 0.50 across the 806 cages with relaxed poses. These descriptive associations do not measure guest-escape probabilities or retention kinetics.

Table A.12: *Window dimensions and guest location for assembled LinkerGen cages. Window diameters are medians [first quartile, third quartile]. Counts include all relaxed poses.*

Distance (bonds)	Cages	Largest window (Å)	Poses	Outside	Outside (%)
5	493	5.34 [4.97, 5.69]	1470	8	0.54
9	106	7.02 [5.91, 7.67]	315	103	32.70
13	195	8.19 [6.86, 8.69]	550	92	16.73
Other	28	6.90 [5.78, 8.46]	81	17	20.99

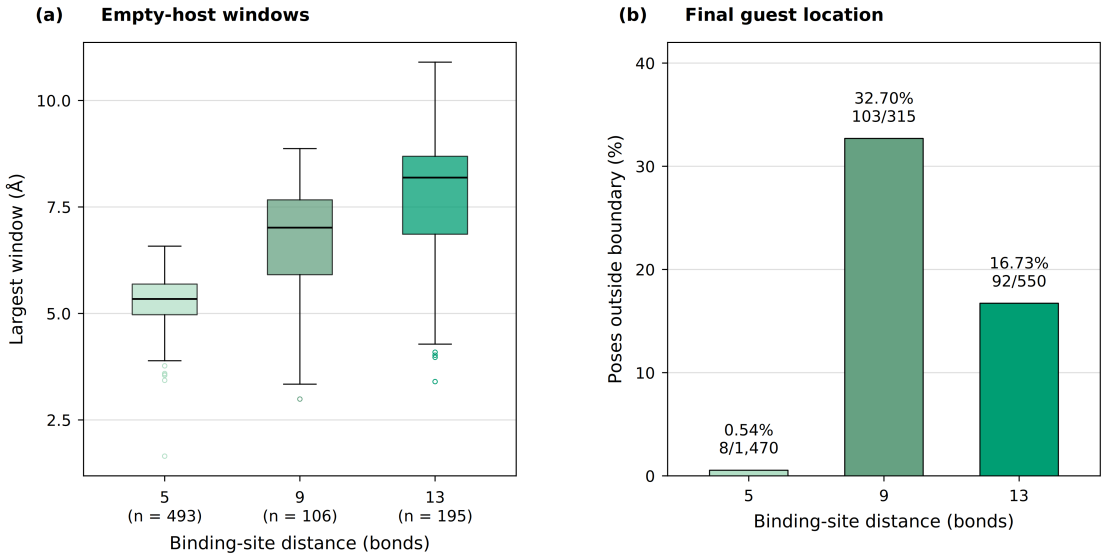


Figure A.8: *Linker distance, cage windows, and guest location. (a) Largest-window diameters; boxes show interquartile ranges, lines medians, and whiskers extend to 1.5 interquartile ranges. (b) Fractions and counts of relaxed poses outside the cavity boundary. The 28 cages with other distances are omitted.*

A.8.5 Host Reorganisation and Reference Energies

This analysis used the 626 retained LinkerGen evaluations. Host coordinates from each lowest-energy accepted complex were compared with the corresponding empty-host reference after rigid least-squares alignment without reflection. Atom correspondence was preserved without symmetry remapping. Host RMSDs included all heavy atoms and counterions; Zr-only RMSDs used a separate alignment of the 12 Zr atoms.

The generic population had the largest median host RMSD, followed by the large and small conditioned populations (Table A.13). This ordering held in all three runs. Grouping instead by five-, nine-, and thirteen-bond distances gave median host RMSDs of 0.129, 0.184, and 0.323 Å, respectively. These compare optimised geometries rather than thermal flexibility, and the distance groups also differ in chemistry.

Empty-host reference lowering was defined as

$$\Delta E_H = E_H^{\text{first}} - E_H^{\text{ref}} \geq 0. \tag{A.3}$$

For a fixed complex and free-guest energy, adopting the lower reference makes the binding energy less negative by ΔE_H. This prevents additional empty-host stabilisation discovered through complex optimisation from being attributed to guest binding.

Reference lowering exceeded 0.1 eV in 24.0% of generic, 14.5% of small conditioned, and 17.6% of large conditioned evaluations (Figure A.9). The largest correction, 3.104 eV, occurred for a five-bond linker from the large conditioned model. All reported binding energies include these corrections; the selected references are not guaranteed global minima.

Table A.13: *Host reorganisation and reference lowering across retained LinkerGen evaluations. RMSDs are medians; energy changes include evaluations with zero correction.*

Population	n	Host RMSD (Å)	Zr RMSD (Å)	$\Delta E_H > 0.1$ eV Count (%)	Mean ΔE_H (eV)	Max. ΔE_H (eV)
Small generic	229	0.263	0.090	55 (24.0)	0.093	1.152
Small conditioned	221	0.129	0.046	32 (14.5)	0.051	1.287
Large conditioned	176	0.165	0.050	31 (17.6)	0.080	3.104

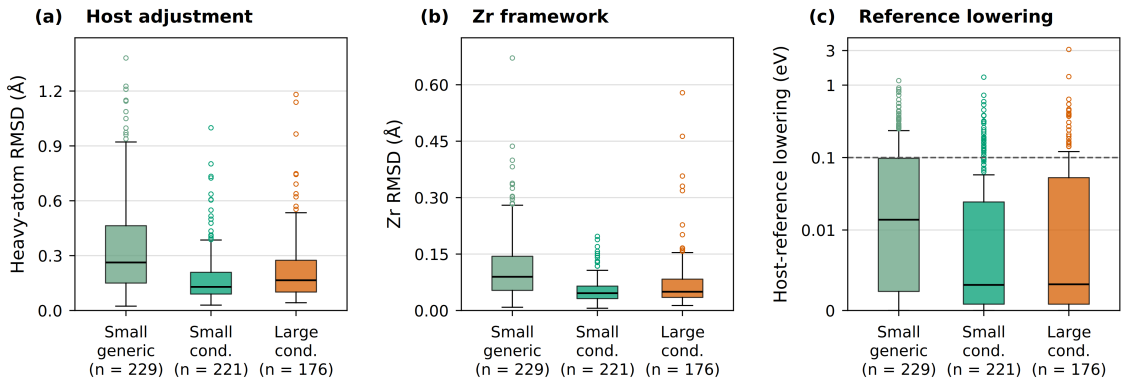


Figure A.9: *Host reorganisation and reference lowering: (a) host heavy-atom RMSD, (b) separately aligned Zr-only RMSD, and (c) ΔE_H. Boxes show interquartile ranges, lines medians, and whiskers extend to 1.5 interquartile ranges. Panel (c) is linear below 0.01 eV and logarithmic above; the dashed line marks 0.1 eV.*

A.8.6 Lowest-Energy LinkerGen Candidate

The lowest calculated binding energy was −2.139 eV for `small_research_r02_041`, from the small conditioned model's second run. The linker had five bonds between binding sites, no urea motif, core TPSA of 58.04 Å^2, and an sp^3 carbon fraction of zero:

```
O=C([O-])c1c(NO)cc(C(=O)[O-])c2ncncc12
```

It therefore did not fulfil the complete research request.

Intermolecular N/O–H···N/O contacts were identified using donor–H distances of at most 1.20 Å, H···acceptor distances of at most 2.50 Å, and donor–H···acceptor angles of at least 140°. Six contacts met these criteria: five host N–H contacts to the guest's carbonyl oxygens and one guest N–H contact to a host oxygen (Table A.14). All participating host atoms belonged to organic linkers. The initial placement label `node_3` does not identify the final interaction environment.

At the complex geometry, the interaction energy was −2.903 eV. Host and guest deformation energies were +0.454 and +0.309 eV, respectively, where

$$E_{\mathrm{int}} = E^{*}_{HG} - E^{HG}_{H} - E^{HG}_{G}, \tag{A.4}$$

$$E_{\mathrm{bind}} = E_{\mathrm{int}} + \left(E^{HG}_{H} - E^{\mathrm{ref}}_{H}\right) + \left(E^{HG}_{G} - E^{0}_{G}\right). \tag{A.5}$$

Here, E^{HG}_{H} and E^{HG}_{G} are single-point energies at the complex geometry. These deformation terms differ from the empty-host reference lowering, which was 0.011 eV for this candidate. The contact analysis describes the optimised structure without assigning energies to individual contacts.

Table A.14: *Intermolecular contacts in the lowest-energy LinkerGen complex. D and A denote donor and acceptor atoms. Indices are one-based atom positions in the supplied XYZ file; host atoms are 1–302 and guest atoms 303–314.*

Donor–H	Acceptor	Donor in	H···A (Å)	D···A (Å)	Angle (°)
N6–H19	O303	Host	1.85	2.80	152.8
N29–H42	O303	Host	1.76	2.76	162.8
N52–H65	O310	Host	1.77	2.78	164.3
N75–H88	O303	Host	1.80	2.77	156.2
N98–H111	O310	Host	1.90	2.85	154.4
N305–H312	O76	Guest	1.76	2.77	164.3

A.9 Data and Code Availability

Dataset preparation scripts, model checkpoints, generated molecules, analysis code, numerical results, and application structures are available in the LinkerGen results archive on Zenodo at doi:10.5281/zenodo.23669416. Software versions and model identifiers are documented with the corresponding workflows.

References

[1] Omar M. Yaghi. Reticular Chemistry—Construction, Properties, and Precision Reactions of Frameworks. *Journal of the American Chemical Society*, 138(48):15507–15509, 2016. doi:10.1021/jacs.6b11821.

[2] Omar M Yaghi, Michael O'Keeffe, Nathan W Ockwig, Hee K Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423(6941):705–714, 2003. doi:10.1038/nature01650.

[3] Stefan Wuttke. Toward the Nobel Prize: Dissecting Fundamental Principles and Applications of MOF and COF Materials. *Advanced Materials*, 37(52):e71859, 2025. doi:10.1002/adma.71859.

[4] Aleksandar Kondinski, Angiras Menon, Daniel Nurkowski, Feroz Farazi, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Automated Rational Design of Metal–Organic Polyhedra. *Journal of the American Chemical Society*, 144(26):11713–11728, 2022. doi:10.1021/jacs.2c03402.

[5] Meili Ding, Robinson W Flaig, Hai-Long Jiang, and Omar M Yaghi. Carbon capture and conversion using metal–organic frameworks and MOF-based materials. *Chemical Society Reviews*, 48(10):2783–2828, 2019. doi:10.1039/C8CS00829A.

[6] Bingbing Chen, Dong Fan, Rosana V Pinto, Iurii Dovgaliuk, Shyamapada Nandi, Debanjan Chakraborty, Nuria García-Moncada, Alexandre Vimont, Charles J McMonagle, Marta Bordonhos, et al. A scalable robust microporous Al-MOF for post-combustion carbon capture. *Advanced Science*, 11(21):2401070, 2024. doi:10.1002/advs.202401070.

[7] Frederik Vermoortele, Rob Ameloot, Alexandre Vimont, Christian Serre, and Dirk De Vos. An amino-modified Zr-terephthalate metal–organic framework as an acid–base catalyst for cross-aldol condensation. *Chemical Communications*, 47(5):1521–1523, 2011. doi:10.1039/C0CC03038D.

[8] Wentao Xu and Omar M Yaghi. Metal–organic frameworks for water harvesting from air, anywhere, anytime. *ACS central science*, 6(8):1348–1354, 2020. doi:10.1021/acscentsci.0c00678.

[9] Ming-Xue Wu and Ying-Wei Yang. Metal–organic framework (MOF)-based drug/cargo delivery and cancer therapy. *Advanced Materials*, 29(23):1606134, 2017. doi:10.1002/adma.201606134.

[10] Austin M. Mroz, Victor Posligua, Andrew Tarzia, Emma H. Wolpert, and Kim E. Jelfs. Into the Unknown: How Computation Can Help Explore Uncharted Material Space. *Journal of the American Chemical Society*, 144(41):18730–18743, 2022. doi:10.1021/jacs.2c06833.

[11] Wendy L. Queen, Matthew R. Hudson, Eric D. Bloch, Jarad A. Mason, Miguel I. Gonzalez, Jason S. Lee, David Gygi, Joshua D. Howe, Kyuho Lee, Tamim A. Darwish, Michael James, Vanessa K. Peterson, Simon J. Teat, Berend Smit, Jeffrey B.

Neaton, Jeffrey R. Long, and Craig M. Brown. Comprehensive study of carbon dioxide adsorption in the metal–organic frameworks M_2(dobdc) (M = Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chemical Science*, 5(12):4569–4581, 2014. doi:10.1039/C4SC02064B. URL https://doi.org/10.1039/C4SC02064B.

[12] Aeri J Gosselin, Gerald E Decker, Brett W McNichols, James E Baumann, Glenn PA Yap, Alan Sellinger, and Eric D Bloch. Ligand-based phase control in porous zirconium coordination cages. *Chemistry of Materials*, 32(13):5872–5878, 2020. doi:10.1021/acs.chemmater.0c01965.

[13] Nikita Hanikel, Daria Kurandina, Saumil Chheda, Zhiling Zheng, Zichao Rong, S Ephraim Neumann, Joachim Sauer, J Ilja Siepmann, Laura Gagliardi, and Omar M Yaghi. MOF linker extension strategy for enhanced atmospheric water harvesting. *ACS Central Science*, 9(3):551–557, 2023. doi:10.1021/acscentsci.3c00018.

[14] Mohamed Eddaoudi, Jaheon Kim, Nathaniel Rosi, David Vodak, Joseph Wachter, Michael O'Keeffe, and Omar M. Yaghi. Systematic design of pore size and functionality in isoreticular MOFs and their application in methane storage. *Science*, 295 (5554):469–472, 2002. doi:10.1126/science.1067208. URL https://doi.org/10.1126/science.1067208.

[15] Peter G Boyd and Tom K Woo. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*, 18 (21):3777–3792, 2016. doi:10.1039/C6CE00407E.

[16] Yamil J Colón, Diego A Gómez-Gualdrón, and Randall Q Snurr. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Crystal Growth & Design*, 17(11):5801–5810, 2017. doi:10.1021/acs.cgd.7b00848.

[17] Marco Gibaldi, Ohmin Kwon, Andrew White, Jake Burner, and Tom K. Woo. The HEALED SBU Library of Chemically Realistic Building Blocks for Construction of Hypothetical Metal–Organic Frameworks. *ACS Applied Materials & Interfaces*, 14(38):43372–43386, 2022. doi:10.1021/acsami.2c13100.

[18] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry*, 4(2):83–89, 2012. doi:10.1038/nchem.1192.

[19] Gianmarco G. Terrones, Shih-Peng Huang, Akash K. Ball, Changhwan Oh, Matthew P. Rivera, and Heather J. Kulik. What makes a plausible metal–organic framework linker? discovery from a general chemical database. *Digital Discovery*, 5(9):3766–3779, 2026. doi:10.1039/D6DD00200E. URL https://doi.org/10.1039/D6DD00200E.

[20] Patrick W. V. Butler, Simon D. Rihm, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Leveraging general molecular fragments to expand the design space of metal–organic polyhedra. *Physical Chemistry Chemical Physics*, 28(29):17787–17791, 2026. doi:10.1039/d6cp02069k.

[21] Patrick W.V. Butler, Simon D. Rihm, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Inverse Design of Metal–Organic Polyhedra through Molecular Fragmentation and Evolutionary Optimisation. *Journal of Chemical Information and Modeling*, 66(7):3933–3943, 2026. doi:10.1021/acs.jcim.5c02956.

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

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

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

[25] Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018. doi:10.1021/acscentsci.7b00512. URL https://doi.org/10.1021/acscentsci.7b00512.

[26] Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3D linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995, 2020. doi:10.1021/acs.jcim.9b01120. URL https://doi.org/10.1021/acs.jcim.9b01120.

[27] Yuyao Yang, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen. SyntaLinker: Automatic fragment linking with deep conditional transformer neural networks. *Chemical Science*, 11(31):8312–8322, 2020. doi:10.1039/D0SC03126G.

[28] Yinan Huang, Xingang Peng, Jianzhu Ma, and Muhan Zhang. 3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design. arXiv preprint arXiv:2205.07309, 2022. URL https://arxiv.org/abs/2205.07309.

[29] Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024. doi:10.1038/s42256-024-00815-9.

[30] Jeff Guo, Franziska Knuth, Christian Margreitter, Jon Paul Janet, Kostas Papadopoulos, Ola Engkvist, and Atanas Patronov. Link-INVENT: generative linker design with reinforcement learning. *Digital Discovery*, 2(2):392–408, 2023. doi:10.1039/D2DD00115B. URL https://doi.org/10.1039/D2DD00115B.

[31] Hyun Park, Xiaoli Yan, Ruijie Zhu, Eliu A. Huerta, Santanu Chaudhuri, Donny Cooper, Ian Foster, and Emad Tajkhorshid. A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture. *Communications Chemistry*, 7(1):21, 2024. doi:10.1038/s42004-023-01090-2.

[32] Dhruv Menon, Vivek Singh, Xu Chen, Mohammad Reza Alizadeh Kiapi, Ivan Zyuzin, Hamish W. MacLeod, Nakul Rampal, William Shepard, Omar M. Yaghi, and David Fairen-Jimenez. Flow-guided chemical language modeling for linker design in reticular chemistry. *Journal of the American Chemical Society*, 148(37): 39546–39555, 2026. doi:10.1021/jacs.6c06920. URL https://doi.org/10.1021/jacs.6c06920.

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014. URL https://arxiv.org/abs/1411.1784.

[34] Panagiotis-Christos Kotsias, Josep Arús-Pous, Hongming Chen, Ola Engkvist, Christian Tyrchan, and Esben Jannik Bjerrum. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence*, 2(5):254–265, 2020. doi:10.1038/s42256-020-0174-5. URL https://doi.org/10.1038/s42256-020-0174-5.

[35] Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, and Alex Zhavoronkov. Nach0: Multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389, 2024. doi:10.1039/d4sc00966e.

[36] Yifan Deng, Spencer S. Ericksen, and Anthony Gitter. Chemical language model linker: Blending text and molecules with modular adapters. *Journal of Chemical Information and Modeling*, 65(17):8944–8956, 2025. doi:10.1021/acs.jcim.5c00853. URL https://doi.org/10.1021/acs.jcim.5c00853.

[37] Xiang Fu, Tian Xie, Andrew S. Rosen, Tommi Jaakkola, and Jake Smith. MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design. arXiv preprint arXiv:2310.10732, 2023. URL https://arxiv.org/abs/2310.10732.

[38] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar K Farha, Randall Q Snurr, and Álan Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1): 76–86, 2021. doi:10.1038/s42256-020-00271-1.

[39] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024. doi:10.1093/nar/gkad1004.

[40] 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. doi:10.1002/cmdc.200800178.

[41] RDKit. RDKit: Open-source cheminformatics. https://www.rdkit.org. URL https://www.rdkit.org.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[44] Aleksandar Kondinski, Andrea M Oyarzún, Simon D Rihm, Jiaru Bai, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Automated Assembly Modelling of Metal-Organic Polyhedra. *European Journal of Inorganic Chemistry*, page e202500115, 2025. doi:10.1002/ejic.202500115.

[45] Jiaru Bai, Simon D Rihm, Aleksandar Kondinski, Fabio Saluz, Xinhong Deng, George Brownbridge, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Twa: The World Avatar Python package for dynamic knowledge graphs and its application in reticular chemistry. *Digital Discovery*, 4(8):2123–2135, 2025. doi:10.1039/D5DD00069F.

[46] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. arXiv preprint arXiv:2206.07697, 2023. URL https://arxiv.org/abs/2206.07697.

[47] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf,

Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163 (18):184110, 2025. doi:10.1063/5.0297006.

[48] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7):1456–1465, 2011. doi:10.1002/jcc.21759.

[49] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009. doi:10.1186/1758-2946-1-8.

[50] Amol Thakkar, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science*, 12(9):3339–3349, 2021. doi:10.1039/D0SC05401A.

[51] Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. SCScore: Synthetic complexity learned from a reaction corpus. *Journal of Chemical Information and Modeling*, 58(2):252–261, 2018. doi:10.1021/acs.jcim.7b00622.

[52] Milan Voršilák, Michal Kolář, Ivan Čmelo, and Daniel Svozil. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics*, 12:35, 2020. doi:10.1186/s13321-020-00439-2.

[53] Patrick W. V. Butler, Arravind Subramanian, Simon D. Rihm, Ari F. Fischer, Tej S. Choksi, Aleksandar Kondinski, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Computational modelling workflows for metal–organic polyhedra in The World Avatar. *Physical Chemistry Chemical Physics*, 28(24):14828–14838, 2026. doi:10.1039/d5cp04871k.

[54] Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015. doi:10.1021/acs.jcim.5b00654.

[55] Shuzhe Wang, Jagna Witek, Gregory A. Landrum, and Sereina Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of Chemical Information and Modeling*, 60(4):2044–2058, 2020. doi:10.1021/acs.jcim.0c00025.

[56] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel

Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27): 273002, 2017. doi:10.1088/1361-648X/aa680e.

[57] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural Relaxation Made Simple. *Physical Review Letters*, 97(17):170201, 2006. doi:10.1103/PhysRevLett.97.170201.