Reaction-first de novo molecular design for geroscience targets: synthesizability by construction, quantum-derived search, and calibrated confidence

GeroQubit Research

GeroQubit, Bengaluru, India

Correspondence: hello@geroqubit.com · Version 1.1 · 26 August 2026

ABSTRACT

Motivation. Generative models for small-molecule design routinely propose structures that no chemist can make, and report activity predictions whose accuracy has been established only on molecules resembling the training set. Both problems are acute in geroscience, where target sets are small, published binders are sparse, and the therapeutic hypothesis usually requires a genuinely novel chemotype.

Approach. We describe a design system whose genome is a synthesis recipe rather than a structure. A candidate is five integers naming a scaffold, two building blocks and two reaction templates; a molecule exists only if those templates fire under RDKit, so a route accompanies every candidate by construction. Search is conditioned in-loop on a ten-component tissue axis derived from compiled expression-clock weights, and every candidate is filtered by hard ADMET, structural-alert and route-integrity gates before selection. A twelve-component hallmark axis is computed per molecule and reported rather than optimised against, and a directional ageing reference assembled for 15,705 genes across four independent sources informs target selection and the applicability analyses, not the search loop.

Results. A quantum-derived, classically computed search layer of nine operators raises usable products per reaction evaluation by 22.9% (95% CI 16.2–33.1%; 24 of 24 paired runs), predicting whether a template fires at AUC 0.946 on 9,173 unseen reactant pairs against 0.637 for the rule table it replaced. A low-rank decomposition over the recipe tensor lifts mean batch quality by 19% and yield from 7.6 to 9.3 candidates per run. Adding an aryl C–N coupling won 12 of 12 paired comparisons. Co-synthesizable set selection raises shared-reagent fraction across a delivered batch by 0.0717 (95% CI 0.0367–0.1099), so one order supplies the set. Runs are byte-reproducible from a job identifier, and the whole pipeline is CPU-only.

Application. Across 21 runnable targets in two programmes, lifespan extension and cellular rejuvenation, the system delivers candidates with a two-step route, named reagents, a genotoxicity screen applied to the reagents as well as the product, and a calibrated statement of how far each molecule sits from the region where its predicted scores are reliable. We give a worked senescence programme in which one designed candidate is available from a make-on-demand supplier at 10 mg scale, which reduces the cost of a first biological test from a custom synthesis campaign to a catalogue order.

Keywords: de novo drug design; synthetic accessibility; reaction templates; geroscience; cellular senescence; applicability domain; evolutionary algorithms

1. Introduction

A molecular design system becomes useful at the moment a chemist can act on its output. That requires two things beyond a predicted score: a route by which the proposed structure can actually be assembled,

and a statement of how far the prediction can be trusted for that particular molecule. This manuscript describes a system built to supply both, and reports what it delivers across 21 geroscience targets.

Both requirements respond to well-documented properties of the field. Post-hoc synthetic accessibility scores[1] rank difficulty without supplying a route, and an audit of generative output found a substantial fraction that expert chemists judged impractical[3]. Separately, activity and ADMET models validated on random or scaffold splits are tested on compounds resembling their training data[10], so their reported accuracy describes a regime different from the one generative design operates in[11,12]. Designing the route rather than the structure addresses the first directly; reporting a per-molecule applicability figure addresses the second.

Geroscience sharpens both. Ageing carries no approved regulatory indication, target sets are small and unevenly characterised, and for several targets of interest the published binder count is in the tens. A design system for this domain must therefore produce candidates a chemist can actually attempt, and must be explicit about the point at which its own predictions stop being informative.

Section 2 places the approach in the synthesis-aware generation literature. Sections 3 to 5 describe the representation, the biological conditioning and the quantum-derived search layer. Section 6 states the evaluation protocol and defines every reported quantity, which matters because the headline result is a ratio whose denominator is not self-evident. Section 7 gives measured results, Section 8 the calibration work, and Section 9 the controls that govern which mechanisms were kept.

2. Related work

Making generated molecules synthesisable is an active area, and the approaches divide by where the constraint is applied. Post-hoc scoring attaches a difficulty estimate to a finished structure[1]; it ranks but cannot guarantee. Retrosynthesis-in-the-loop filters generated structures through a planner, which is faithful but expensive and still permits the generator to spend its search on unreachable regions.

Closest to this work is bottom-up generation, in which the object being searched is an assembly tree rather than a graph. SynNet represents a molecule as a synthetic route over purchasable building blocks and reaction templates, decoding tree to structure[4]. Related systems constrain generation to in-house building-block inventories so that proposals are makeable with reagents a specific laboratory already holds[6], and a recent review sets out why "generate then check" underperforms "generate only what is reachable"[5].

Our representation is in that family and differs in three respects. First, the genome is a fixed-length integer vector rather than a variable-depth tree, which makes crossover and mutation ordinary operations and permits an explicit edit distance between two candidates' routes. Second, the reaction set is audited rather than mined-and-trusted: templates enter only with a named bond transformation and a matched reagent profile. Third, conditioning is biological rather than purely chemical, over a tissue and hallmark state described in Section 4.

For evaluation we report against established generative benchmarks[7,8,9] while noting in Section 7 why saturated metrics on those suites are uninformative for this system.

3. Representation and assembly

3.1 *Recipes as the unit of search*

A candidate is a vector of five integers: a scaffold index, a first building-block index, a first reaction-template index, a second building-block index and a second template index. Decoding applies the named templates through RDKit's reaction machinery[23]. Where the chosen template does not fire, a fallback chain attempts other compatible templates and then a direct attachment; where no step succeeds, the candidate is discarded rather than repaired.

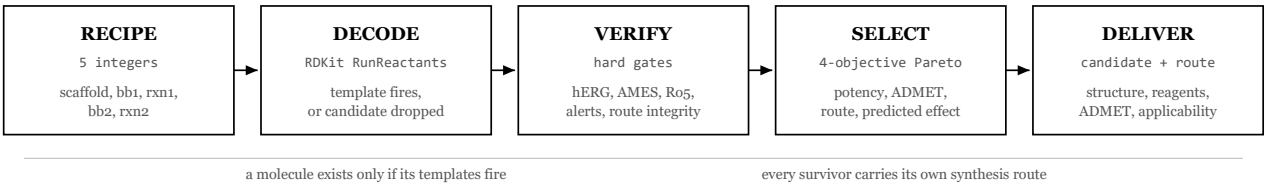


Figure 1. The pipeline end to end. The genome is a synthesis recipe, so a molecule exists only if its named templates fire, and every survivor carries the route that produced it. Candidates that fall through to direct attachment at both steps are discarded, because the recorded route would otherwise be fictitious.

The consequence is structural rather than statistical. The search cannot express a molecule for which no assembly route exists, so synthesizability stops competing with potency as an objective. It also makes route-level quantities computable that a structure-first generator cannot define: the weighted edit distance between two candidates' recipes, and the set of analogues reachable by changing exactly one building block while holding the scaffold and both reactions fixed. The latter yields an orderable series for structure-activity work rather than a set of unrelated structures.

3.2 *The reaction library*

The engine carries 73 reaction templates. Curated templates were written by hand against a named bond transformation; mined templates were admitted from a patent-derived corpus only after firing on a substrate grid drawn from both target programmes and only after being given a reagent profile matching the bond they form. Templates that fire but cannot be classified are refused. That refusal is deliberate: a template with an unverified reagent profile prescribes chemistry that may not correspond to the bond it makes.

3.3 *Search over the recipe space*

The recipe space is large but not smooth: a single-index change can move from a product to no product at all, because a template either fires or does not. The search therefore combines a population-based method with several mechanisms aimed at that discontinuity.

The population is initialised by enumerating validated assemblies rather than by sampling uniformly, so a substantial fraction of the starting set already decodes. Variation uses per-slot mutation with an adaptive rate governed by a success-ratio rule, so the rate rises while the search is finding improvements and falls when it is not. An archive over behavioural niches preserves solutions that are locally suboptimal but occupy distinct regions, which matters because the objective is not a single peak. Scaffold choice is treated as a bandit problem and re-evaluated periodically, since which skeletons decode successfully is target-dependent and not known in advance. Restarts follow a cosine schedule, and a predictive stopping rule ends a run when the projected remaining improvement falls below a threshold.

A separate mechanism fits two low-rank decompositions over the five-slot tensor: one for the objective and one for the probability that a candidate survives the gates. The second is trained on rejected candidates, which are otherwise discarded. This addresses a measured pathology in which increasing the compute budget produced fewer and worse molecules, because the search was optimising a smoothed objective while selection enforced a hard one, and the population converged into a region where every candidate was rejected. Rank matters: at rank one the decomposition reduces to independent per-slot preferences and reproduces the earlier failure, which makes the claim testable rather than rhetorical.

An important negative constraint governs where any of this may act. Every mechanism we have tested that folded an additional signal into the in-loop objective degraded results, while the same signal applied at selection did not. Search-layer mechanisms therefore act on sampling or on selection, and the objective and gates are left untouched.

4. Biological conditioning

4.1 *Targets*

Twenty-one targets are runnable, in two programmes maintained separately. The lifespan set was drawn from the geroprotector literature[20]; the rejuvenation set was admitted by rule, requiring ageing perturbation evidence and at least 50 human single-protein binders at a defined potency threshold. SIRT6 is the only member of both.

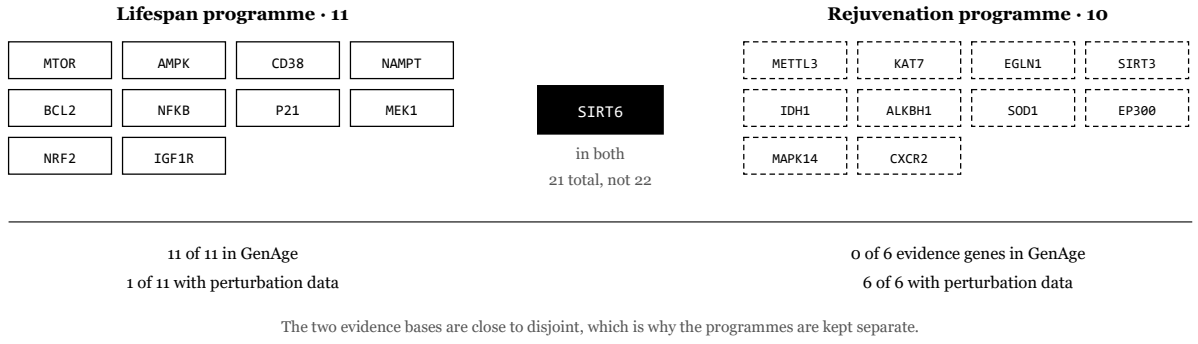


Figure 8. The two target programmes. Every lifespan target appears in the curated longevity-gene database and one of eleven carries ageing perturbation data; none of the six genes that qualified the rejuvenation set on evidence grounds appears in that database and all six carry perturbation data. The 11 of 11 figure is circular, since the lifespan set was selected from that literature; the 0 of 6 and the 1 of 11 versus 6 of 6 comparison are not, and they are what motivate keeping the programmes apart.

4.2 *The rejuvenation programme*

The second programme addresses a different question from lifespan extension: not whether an intervention lengthens survival, but whether it moves cells toward a younger state. Those are separable both conceptually and, as Figure 8 shows, in the evidence. Reprogramming work has demonstrated that cellular age can be reduced without altering the nutrient-sensing pathways that dominate the lifespan literature[26], and target-level rejuvenation evidence lands most heavily on epigenetic enzymes[27].

Targets enter this programme by rule rather than by judgement. A candidate target requires documented ageing perturbation evidence from at least two independent sources, and at least fifty human single-protein binders above a defined potency threshold, so that reference chemistry exists to design against. An audit stage then checks the annotated direction of intervention, the chemotype spread of the available binders, and whether those binders are in fact small molecules rather than peptides.

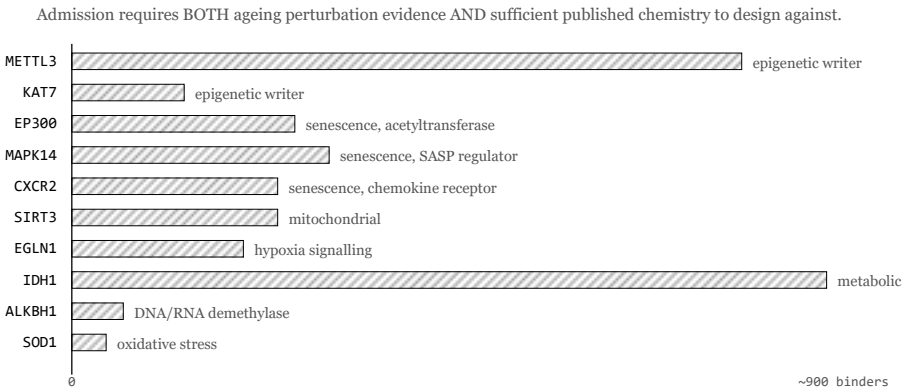


Figure 10. Rejuvenation targets by available reference chemistry, with the biological axis each addresses. Admission requires both evidence and chemistry; a target with strong perturbation evidence and no published small-molecule binders is recorded as an opportunity rather than admitted, because the activity model would have no reference set to score against.

That audit is not decorative. Applying it to candidate targets rejected one whose entire active set proved to be peptidic on inspection, and another whose available binders collapsed into a single chemotype accounting for most of the set. Both would have produced candidates carrying an activity score computed against a reference that could not support it. Three further targets were withdrawn from the runnable list altogether after direct checks in two independent compound databases showed no usable small-molecule chemistry; they remain listed as opportunities, and the absence is itself a finding, because de novo design is the one approach that does not require prior chemistry.

The senescence axis is the most developed of these. Senescent cells persist without dividing and secrete an inflammatory programme; interrupting that programme without killing the cell is a distinct therapeutic strategy from clearance, and it is where three of our rejuvenation targets sit[16,17,18].

4.3 *The ageing state*

Two ageing axes are computed[13,14], and they enter the system at different points. Reporting them as one object would overstate what the search optimises, so they are separated here.

The **in-loop** axis has ten tissue components. Each is an expression-clock reversal score computed from compiled per-tissue clock weights, and a single-tissue score combines the target tissue with a specified inter-tissue coupling matrix in the ratio 0.60 to 0.40. This term carries 0.45 of the in-loop fitness and is the only ageing axis the search optimises against.

The **twelve hallmark components** are computed per molecule as compartment reachability: whether a molecule's physicochemistry is consistent with reaching the subcellular compartment in which a given hallmark process is executed. They are reported with each candidate and are deliberately excluded from the objective. That exclusion is measured rather than stylistic — conditioning generation on this axis regressed 21 of 32 paired runs and reduced yield to zero on one target, while the same axis is informative when read alongside a candidate.

Separately, a directional ageing reference assembled for **15,705 genes across four independent sources** informs target selection and is one input to direction assignment, alongside an audit against annotated target direction. It is also the basis for the applicability analyses reported below. The search loop does not read it, and no claim here depends on it doing so.

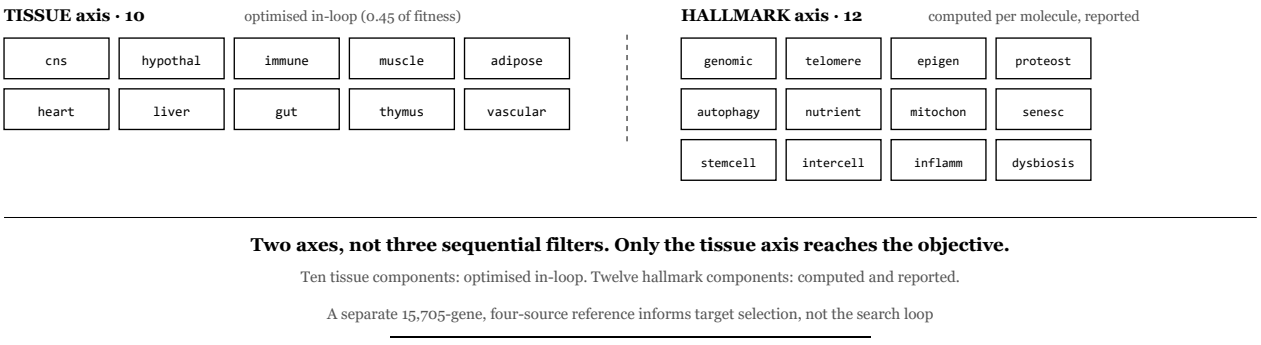


Figure 2. The two ageing axes and where each acts. The ten tissue components are optimised in-loop; the twelve hallmark components are computed per molecule and reported, not optimised against. Keeping them separate matters in two ways: read as prose, "target by tissue by hallmark" is usually taken to mean three filters applied in sequence, which it is not; and only one of the two axes reaches the objective.

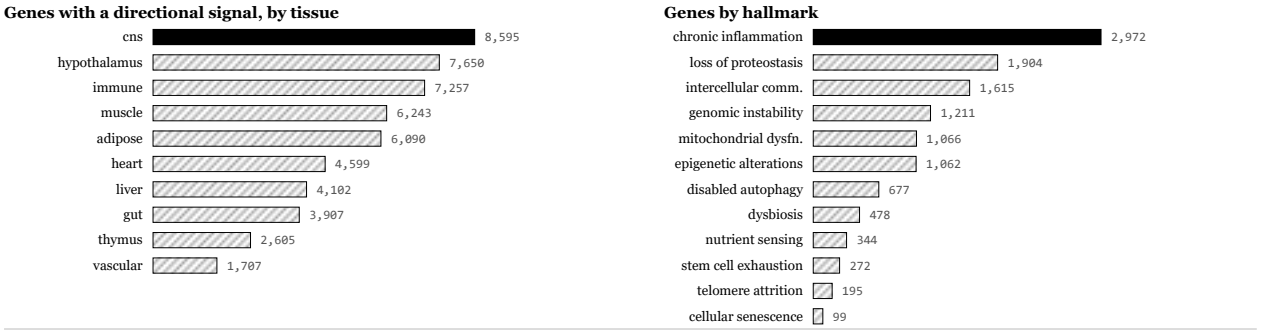


Figure 3. Coverage of the 15,705-gene directional reference, per tissue (left, n = 10) and per hallmark (right, n = 12). This describes the reference used for target selection and applicability analysis; it is not the coverage of the in-loop tissue axis, which is computed from compiled clock weights. The spread is reported because it is uneven by a factor of five across tissues and thirty across hallmarks, and a uniform presentation of twelve hallmark axes would imply they are observed equally well. Cellular senescence, the hallmark with the most active therapeutic literature, has the thinnest directional coverage here.

4.4 *Quantum-derived search and similarity*

Three components of this system are formulated in the language of quantum mechanics and evaluated classically. In an optional search mode, disabled by default, the two ageing axes above are carried together as a single 22-component state vector; the reference set for a target is treated as a density operator rather than as a list of neighbours; and similarity is computed with a quantum kernel over amplitude-encoded count features[24]. Following the convention in that literature, these are computed on conventional hardware; we note it because the phrase is used loosely elsewhere and a reader is entitled to know which claim is being made.

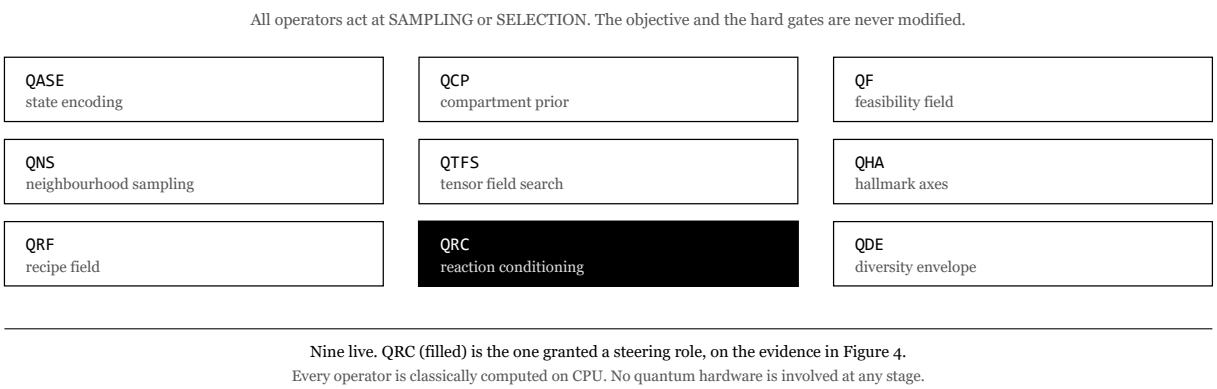


Figure 9. The nine live operators in the search layer, with the function each performs. Every one acts at sampling or at selection; none modifies the objective or the hard gates, which is a constraint we adopted after repeated observation that signals folded into the objective degraded results while the same signals applied at selection did not.

The density-operator formulation earns its place by supplying two quantities that no nearest-neighbour measure can express. The first is spectral leakage: the fraction of a query state lying outside the subspace the reference set spans. A distance to the closest known active is blind to direction; leakage is not, and it distinguishes a molecule that is merely far from the reference from one that points somewhere the reference has no mass at all. The second is effective rank. A 200-compound reference set for one of our targets spans an effective rank of 8.45, so it is not 200 independent directions, and a measure built on counting neighbours cannot report that.

Within the kernel, the term that carries genuinely non-classical structure is the two-body reduced density matrix. Feature co-occurrence is invisible to any function of single-feature marginals, so this term is not reproducible by a kernel on one-body quantities however it is tuned; it improves a ranking metric by +0.0112 (95% CI +0.0083 to +0.0145) and is computed through an identity that never forms the full density matrix, at a cost of two matrix multiplications. The one-body part of the same kernel does admit a classical reduction, as the wider literature on bandwidth-tuned quantum kernels reports[25]; we separate the two because a method description that cannot be checked component by component is not a method description.

5. Scoring, gates and safety screening

5.1 *Aggregation*

In-loop fitness combines structural quality, similarity to a target's reference binders, similarity to a geroprotector reference set and a pharmacophore term, modulated by an ADMET proxy and physicochemical penalties. Delivered candidates are re-scored with a geometric mean over activity, drug-likeness, synthetic accessibility, pharmacophore fit and shape.

The geometric mean is deliberate. Under a weighted arithmetic mean a strong potency estimate offsets a poor safety profile; under a geometric mean a single weak axis collapses the aggregate, which is the behaviour a screening decision requires. Two earlier scalarisations were replaced after they permitted exactly that trade. The property holds only while every axis carries signal, so one term was removed from the aggregate after audit showed it was a per-target near-constant inflating every candidate by several percent without discriminating between them.

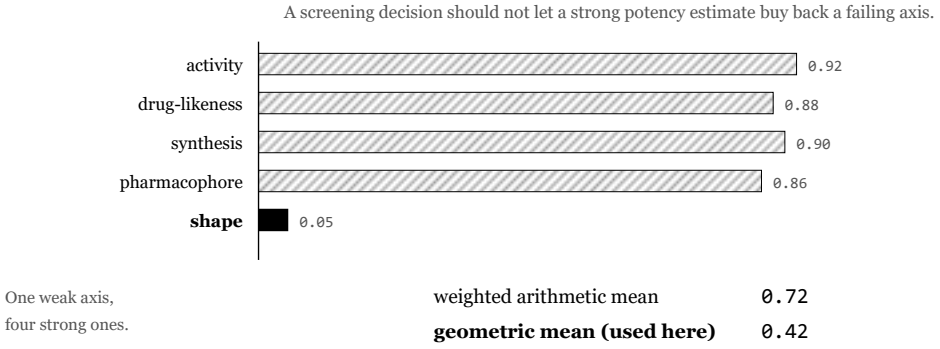


Figure 14. Why the delivered aggregate is a geometric mean. The candidate shown is strong on four axes and fails one. A weighted arithmetic mean returns a respectable 0.72 and the failure disappears into the average; the geometric mean returns 0.42 and it does not. The property holds only while every axis carries signal, which is why an axis found to be a per-target near-constant was removed from the aggregate rather than down-weighted.

5.2 *Hard gates*

Verification drops rather than penalises. A candidate is discarded on predicted cardiac or mutagenicity liability above threshold, two or more violations of standard oral physicochemical limits, excessive synthetic complexity or rotatable-bond count, the presence of a radical, or a route in which both assembly steps fell through to direct attachment.

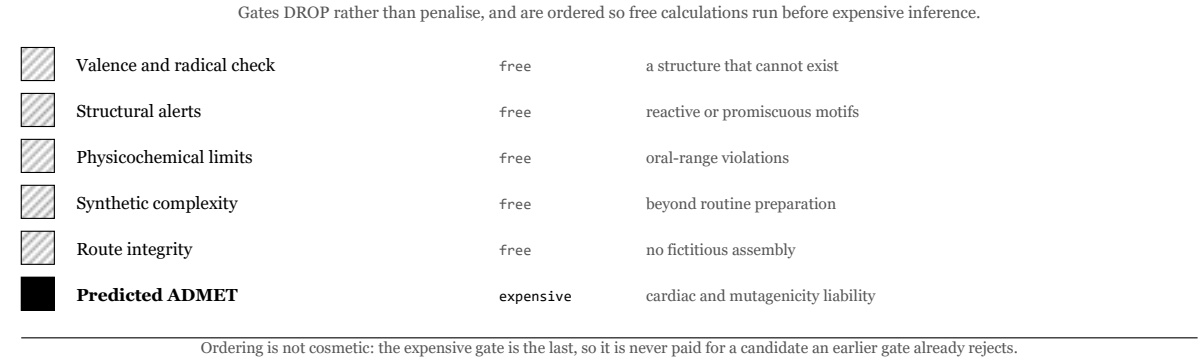


Figure 13. The gate cascade. Each gate removes a candidate outright rather than reducing its score, so a candidate failing any one of them cannot be recovered by strength elsewhere. Ordering matters for cost: the only expensive step is last, and is never paid for a candidate an earlier gate already rejects.

5.3 *Genotoxicity screening over the route*

Because the route is known, structural alerts can be evaluated on the reagents as well as on the product. A DNA-reactive building block is a handling and regulatory consideration even when it is consumed during the synthesis and absent from the final structure, and a product-only screen cannot see it. We apply a standard alert set to every reactant in the recorded route as well as to the product. On a validation set of known genotoxic and known benign structures the implementation flagged all positives and passed all negatives, which establishes that the alerts are wired correctly rather than that they predict genotoxicity well.

5.4 *Developability*

Candidates also receive a second aggregate computed over only those axes that do not depend on similarity to known actives: drug-likeness, synthetic accessibility, predicted ADMET and physical-form readiness. This separates two questions that a single score conflates. A candidate can be a well-behaved, makeable molecule while its predicted activity is low precisely because it is unlike anything the activity model has seen, and reporting both figures makes that distinction visible instead of averaging it away.

6. Evaluation protocol

6.1 *Definitions*

Reported quantities are defined as follows, since two of them are ratios whose denominators are not self-evident.

- **Aggregate score.** A geometric mean over activity, drug-likeness[2], synthetic accessibility[1], pharmacophore fit and shape, then multiplied by structural alert, developability and route-quality factors, each in [0,1]. Bounded in [0,1].
- **Peak.** The maximum aggregate score across the candidates a single run delivers. It is a batch-best, not a mean, and is therefore sensitive to the size of the delivered set, which is held fixed within any comparison.
- **Yield.** The count of candidates surviving all gates and delivered by a run.
- **Products per reaction evaluation.** The number of distinct valid product molecules obtained, divided by the number of calls to the reaction engine. This is the headline efficiency metric; the denominator is attempted template applications, including those that fail.
- **Batch route sharing.** The mean fraction of reagents shared across the delivered set, which determines whether one order supplies a whole batch.

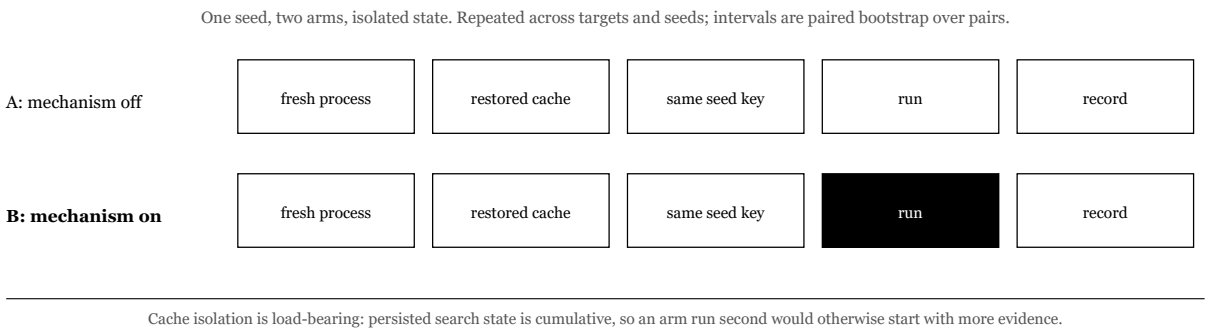


Figure 15. The paired measurement design used for every comparison in Table 1. Both arms of a pair share a seed key and differ only in the mechanism under test. Cache isolation is not a formality: persisted search state accumulates across runs, so without it the arm executed second begins with strictly more evidence than the first.

6.2 *Experimental design and compute*

All comparisons use paired runs: a separate operating-system process per arm, a distinct random key per seed, and snapshot isolation of every persisted cache around each arm. Intervals are 95% paired bootstrap over run pairs. That protocol is not incidental. Persisted search state in this system is cumulative rather than converging, so an arm executed second begins with strictly more evidence than the first, and three separate false regressions were traced to that alone before isolation was enforced.

The engine is CPU-only and uses no GPU at any stage. Experiments reported here were executed on commodity cloud instances of 2 to 16 vCPUs; a single run at the population and generation counts used for these comparisons completes in tens of seconds to a few minutes depending on target. The absence of a training step is a property of the design: all scorers are closed-form or pre-trained, so there is no model fitting in the loop.

7. Results

7.1 *Reaction feasibility prediction*

Decoding is the stage at which most candidates are lost, so the order in which templates are attempted determines how much of the compute budget produces molecules. We replaced a hand-written rule table with a model that predicts, from the functional groups of a reactant, whether a given template will fire, trained on the full record of past decode outcomes, including the rejected attempts that are ordinarily discarded.

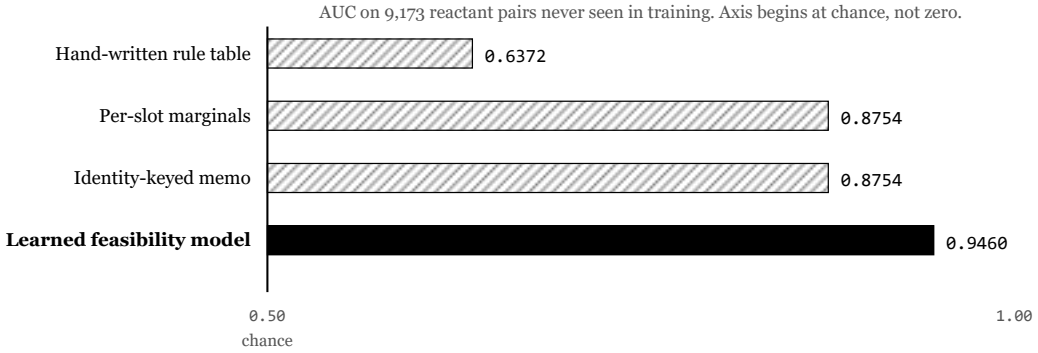


Figure 4. Feasibility prediction, evaluated on n = 9,173 reactant-template pairs held out of training. The axis begins at 0.50 because that is the score a coin flip earns; drawn from zero, a coin flip would appear half full. AUC is reported rather than log-loss because the rule table emits only two distinct probabilities and cannot compete on a calibration-sensitive metric, which would flatter the learned model for the wrong reason.

Ordering the fallback chain by that prediction changes which template is attempted first and removes none from consideration, so the set of reachable products is identical with the mechanism on or off. The gain is scheduling, not chemistry.

7.2 *Measured effects of shipped mechanisms*

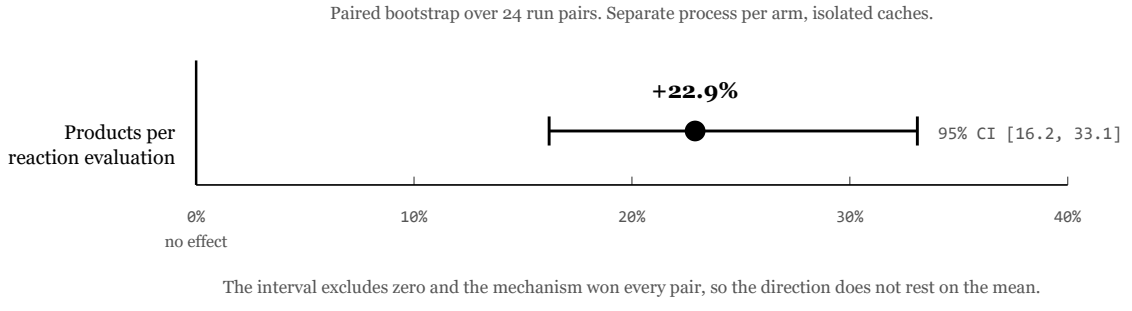


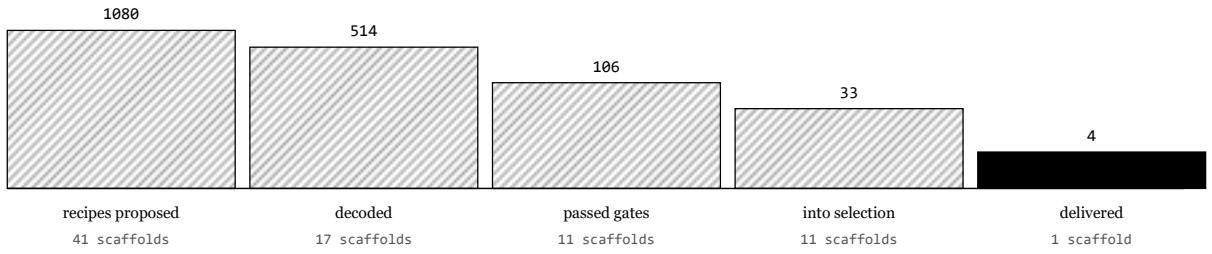
Figure 5. The headline efficiency effect with its 95% paired-bootstrap interval, over 8 targets × 3 seeds (n = 24 pairs). The mechanism won every pair, so the direction does not depend on the mean. Other measured effects are reported in Table 1 rather than plotted here, because they carry different units and a shared axis would invite a comparison that is not meaningful.

Table 1. Measured effects of mechanisms enabled by default. Intervals are 95% paired bootstrap over run pairs. W/T/L counts wins, ties and losses across paired runs.

Mechanism	Effect	95% CI	n pairs	W/T/L
Feasibility-ordered decoding	+22.9% products per reaction evaluation	+16.2 to +33.1%	24	24/0/0
Tensor recipe search	+19% mean peak; yield 7.64 → 9.32	—	33	8/2/1
Aryl C–N coupling template	peak improvement, no regression	—	12	12/0/0
Co-synthesizable set selection	+0.0717 batch route sharing	+0.0367 to +0.1099	33	12/14/0
Champion-preserving selection	delivered peak never below best available	—	24	5/19/0

7.3 *Where diversity is lost*

Because decoding dominates attrition, we instrumented one run to count distinct scaffolds entering and leaving every stage.



Decoding is the dominant loss, and it is scaffold-selective: most of the skeleton diversity is gone before any gate has run.

Figure 7. Attrition through the pipeline for one representative target, with distinct scaffold counts beneath each stage. Decoding is the dominant loss and it is scaffold-selective: skeleton diversity falls furthest before any gate has run. This localises a limitation rather than demonstrating a capability, and n = 1, so no quantity here is a performance figure.

The mechanism is the compatibility table: whether a scaffold and template are compatible is predicted from functional groups rather than verified by substructure matching, and that prediction over-promises unevenly across targets. Tightening it costs decode yield rather than correctness, and a strict-matching variant collapsed scaffold diversity by an order of magnitude. We report the trade-off rather than a resolution.

7.4 *What a delivered candidate carries*

The output of a run is not a list of structures. Each candidate is delivered with the information required to decide whether to make it, and four of those items exist only because the route rather than the structure is the object of search.

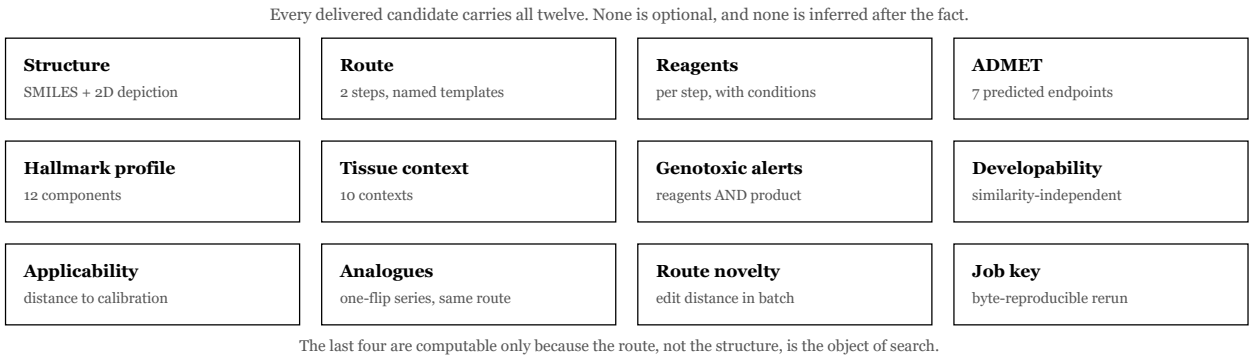


Figure 12. Contents of a delivered candidate. Analogue series, route novelty, reagent-level alerts and reproducible rerun are computable only from a representation that holds the synthesis route; a generator emitting structures alone cannot produce any of the four.

7.5 *Reproducibility from an identifier*

A run is identified by a job key, and that key determines the output. Supplying the same key returns byte-identical molecules in the default environment, with no special configuration, so any candidate we have ever delivered can be regenerated and audited from its identifier alone. This is a requirement rather than a convenience: a design that cannot be replayed cannot be defended in a regulatory filing or reproduced by a collaborator.

7.6 *Analogue series from the representation*

Holding the scaffold and both reaction templates fixed while varying one building block enumerates a set of molecules that share a synthesis route by construction. Every member is reachable from the same reagents and the same two steps, so the series can be ordered and prepared as a block rather than as a set of unrelated preparations.

On a delivered candidate this produced 119 analogues spanning fingerprint similarity from 0.89 down to 0.79 to the parent. That gradient is the useful property: a flat series tests nothing and a disconnected one is not a series. A structure-first generator cannot compute this, because holding a route fixed requires knowing the route, and it has none.

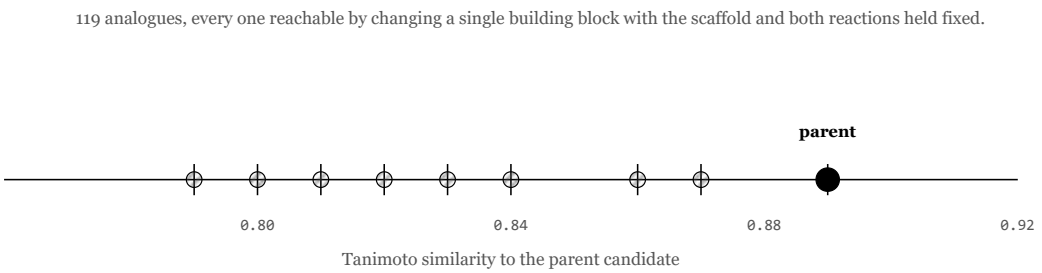


Figure 16. An analogue series generated by one-flip enumeration around a delivered candidate: 119 molecules spanning Tanimoto 0.89 to 0.79 to the parent, every one sharing the scaffold and both reaction steps. The gradient is what makes it usable, since a series with no spread tests nothing.

The same representation yields a route-novelty measure

The same representation yields a route-novelty measure, defined as weighted edit distance over the five slots. It describes how much of the assembly differs between two candidates and is therefore a statement about the produced batch rather than about the chemical universe. We report it as a diversity property of a delivered set, not as a claim of novelty against known chemistry.

8. The applicability boundary

The central quantitative finding of this work is a constraint rather than a capability, and it governs how every predicted score here should be read.

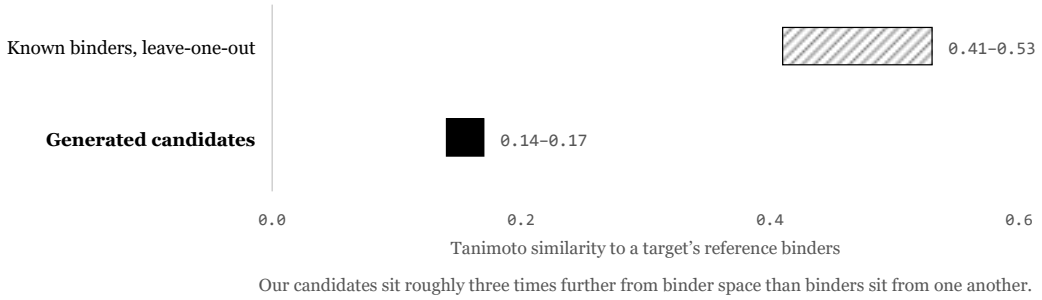


Figure 6. Chemical distance to a target's reference binders, Tanimoto on Morgan fingerprints. Known binders are evaluated leave-one-out against the remainder of their own set. The separation is the definition of a novel chemotype and simultaneously the reason predictions on our own output are less reliable than any benchmark figure would suggest.

Recovery accuracy on our own output falls to AUC 0.62 from 0.945 on a conventional held-out split. Degradation of ADMET models beyond the structural frontier has been reported independently on six public tasks, where the training penalties tested there did not resolve it[12]; we adopt that wording deliberately, since a single preprint on six tasks does not establish that the effect is universal or that no method can address it.

We tested whether richer physics recovers the loss. A combination of shape, electrostatic and desolvation descriptors scored 0.572 against 0.610 for plain two-dimensional fingerprint similarity on novel chemotypes. It did not. We therefore treat the boundary as something to measure and disclose rather than to overcome, and every delivered candidate carries an applicability figure alongside its score.

8.1 *What is shipped instead of accuracy*

Since the boundary cannot be removed, the design decision is what to do at it. Every delivered candidate carries a distance figure describing how far it sits from the region where the scoring models were calibrated, presented alongside the score rather than folded into it. Folding it in would produce a single number that is neither a prediction nor a confidence, and would hide precisely the case a medicinal chemist needs to see: a high score computed far outside the domain that produced it.

Four separate applicability thresholds operate in the system, on different quantities and with different calibrations, and they are deliberately not combined into one. They are reported for detection and never used as gates, because a candidate outside the domain is not thereby a bad candidate; it is a candidate whose predicted score carries little information. Removing it would select against exactly the novelty the system exists to produce.

9. Validation, controls and scope

Every mechanism in this system is required to beat a control specified before the measurement runs. That discipline is the reason the results in Section 7 can be relied on, and it is worth showing what it costs: three candidate operators were built, measured against their declared controls, and withdrawn. We describe them because a method that reports only its successes gives a reader no way to calibrate the ones it keeps.

- **State multiplicity.** Exceeded its permutation control only at grid resolutions coarse enough to collapse 25 molecules into as few as 2 bins, and returned exactly the null value wherever molecules re-

mained distinguishable. A grid-free reformulation with bandwidth read from the data also scored below control.

- **Identity-keyed reaction memoisation.** Improved prediction on previously seen reactant pairs by 0.045 and on unseen pairs by 0.000. All of its apparent skill was recall. Retained as a cache; withdrawn as a model.
- **Pairwise slot co-occurrence.** Raw pointwise mutual information appeared large, up to 2.32 nats. A permutation control that shuffled each slot independently, preserving marginal frequencies and table sparsity while destroying genuine coupling, scored equally on 21 targets of 21. The signal was a sparsity artifact throughout. A mutual-information value on a sparse contingency table will look substantial without such a control.

Table 2. Pre-registered retrospective evaluations of the structure-to-lifespan predictor. Specifications were committed before results in both cases. Neither interval excludes chance.

Evaluation	n	AUC	95% CI	Verdict
Mouse intervention programme	35	0.645	0.447–0.826	Null; underpowered by design
Invertebrate screening programme	133	0.532	0.391–0.671	Null; adequately powered

A post-hoc power analysis shows the first design could only have reached significance at AUC 0.695 or above, so it was underpowered by construction and cannot be read as evidence against the model. The second was not: at n = 133 the rank correlation with the continuous outcome was −0.023 (p = 0.80), where the model's internal cross-validated correlation would have been detectable. We conclude the predictor has no demonstrated external ranking ability. It is used only for post-hoc presentation and never inside the search loop.

SCOPE OF THESE NULLS

Synthesizability by construction, route novelty, design-time genotoxicity screening over reagents, developability scoring and reproducibility do not depend on the lifespan model and are unaffected. What is affected is any claim that a delivered candidate is likely to extend lifespan. We make no such claim anywhere in this manuscript.

9.1 *Benchmark position*

On established generative benchmarks[7,8] this system reports validity, uniqueness and novelty at ceiling. We consider those figures uninformative here rather than favourable: for a generator that assembles molecules through verified reactions over a deduplicated block library, all three are close to tautological. On a multi-property optimisation suite[9] we match published performance on a drug-likeness objective while every candidate carries a route, and score zero on a dopamine-receptor objective because the building-block library is geroscience-specialised. We report the second alongside the first.

10. Worked example

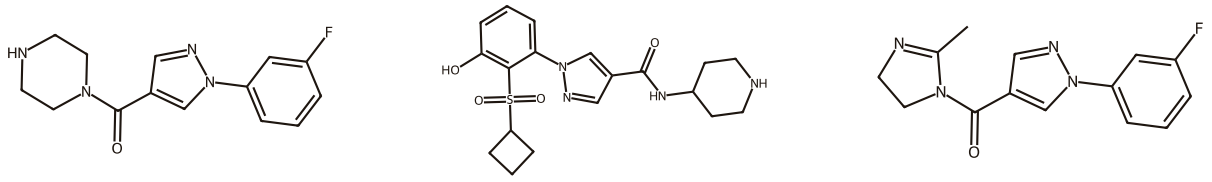
CXCR2 illustrates the intended use. Chemokine signalling through CXCR2 reinforces cellular senescence: receptor knockdown relieves both replicative and oncogene-induced senescence, and forced expression drives premature senescence[16]. The receptor sits inside a feedback loop sustaining the senescence-associated secretory phenotype, alongside other senescence-associated targets in our set[17,18].

Every CXCR2 antagonist that reached patients was developed for respiratory disease, and the class did not succeed there; one phase 2b programme reported no improvement on symptom or health-status endpoints together with more frequent exacerbations in treated groups[19]. Those trials tested whether reducing neutrophilic inflammation improves lung function, which is a different hypothesis from

whether interrupting the senescence loop benefits ageing tissue, and no clinical programme has tested the second.

10.1 *The three candidates*

The system produced three candidates against CXCR2. All are small, all clear the standard oral physicochemical limits without violation, and all are assembled by two named reactions from commercially available reagents.



GQ-SM-06
MW 274.3 · QED 0.89 · SA 2.20

GQ-SM-07
MW 404.5 · QED 0.69 · SA 2.80

GQ-SM-08
MW 272.3 · QED 0.84 · SA 2.60

Figure 17. The three delivered CXCR2 candidates as drawn by RDKit from the SMILES in Table 3. All three are built on an N-aryl pyrazole carboxamide framework, which is the motif the engine converged on for this target rather than one supplied to it.

Table 3. The three CXCR2 candidates in full. Ro5 counts Lipinski violations; hERG and AMES are predicted liabilities on [0,1] where lower is better; developability is the similarity-independent aggregate defined in Section 5.4; maxTan is the highest Tanimoto similarity to any compound in the reference binder set for this target.

ID	MW	logP	QED	SA	Ro5	hERG	AMES	Dev.	maxTan
GQ-SM-06	274.3	1.06	0.890	2.20	0	0.299	0.215	0.917	0.185
GQ-SM-07	404.5	1.39	0.694	2.80	0	0.297	0.139	0.822	0.250
GQ-SM-08	272.3	1.89	0.838	2.60	0	0.140	0.081	0.883	0.207

Table 4. Structures and routes as machine-readable SMILES, so that every claim in Table 3 can be recomputed independently. Both steps are templated for all three candidates; none relies on a fallback attachment.

ID	Product SMILES	Step 1	Step 2
GQ-SM-06	`O=C(c1cnn(-c2cccc(F)c2)c1)N1CCNCC1`	N-arylation	amide coupling
GQ-SM-07	`O=C(NC1CCNCC1)c1cnn(-c2cccc(O)c2S(=O)(=O)C2CCC2)c1`	N-arylation	amide coupling
GQ-SM-08	`CC1=NCCN1C(=O)c1cnn(-c2cccc(F)c2)c1`	N-arylation	amide coupling

10.2 *The shared route*

The three routes converge on a single starting material. Each begins with 1H-pyrazole-4-carboxylic acid, arylates the pyrazole nitrogen under copper catalysis, and couples the resulting acid to an amine. The candidates differ only in the aryl halide and the amine partner, and two of the three share the aryl halide as well, so their routes differ in exactly one reagent.

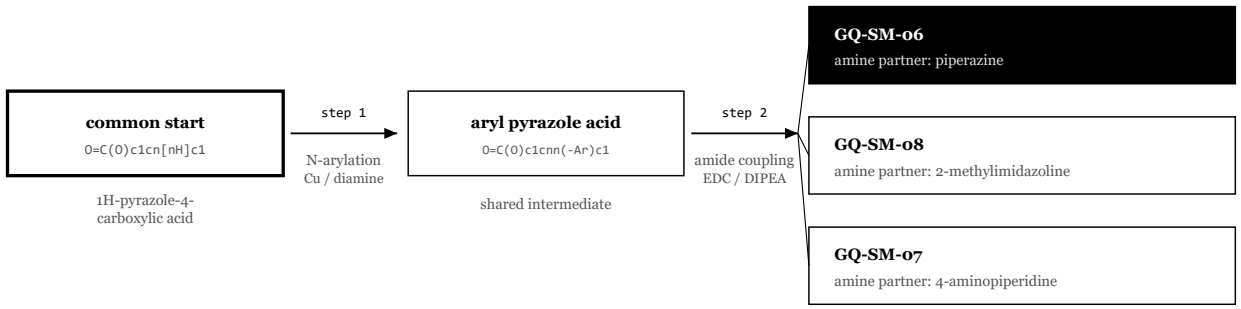


Figure 18. The common synthetic route. A shared starting material and a shared intermediate mean the three candidates can be prepared as one campaign with a single first step, rather than as three independent preparations. This is a direct consequence of co-synthesizable set selection (Table 1), which selects a delivered batch for shared reagents rather than choosing each member on its own merit.

That structure is what makes the set orderable. A single N-arylation supplies the intermediate for all three, and three amide couplings then diverge to the products. The reagent list for the whole batch is one heteroaromatic acid, two aryl halides and three amines.

Table 5. Chemical distance for the CXCR2 series, Tanimoto on Morgan fingerprints. Clinical compounds are compared pairwise among themselves (n = 4 compounds, 6 pairs); candidates are compared to the nearest clinical compound (n = 3). The two ranges nearly meet, which is the point of reporting both.

Comparison	n	Range	Reading
Clinical compounds, pairwise	6	0.106–0.167	The class is already structurally diverse
Our candidates, to nearest clinical compound	3	0.185–0.250	Outside the class, but not by a wide margin

Each candidate carries a route with named reagents, a predicted ADMET profile, a structural alert screen applied to the reagents as well as the product, and its distance from the region where those predictions are calibrated. One is listed by a make-on-demand supplier at 10 mg scale. That is a fact about a catalogue entry and not a statement about the molecule.

10.3 *Reaching a first biological test*

The practical obstacle to testing a designed molecule is rarely the assay. Worm and cell healthspan screening is available commercially at low thousands of dollars per compound. The cost driver is obtaining the compound: first-time synthesis of a novel structure is the expensive class of custom chemistry, quotes are slow, and cost is close to flat between milligram and hundred-milligram scale because labour dominates, so asking for less material saves little.

Designing over a fixed library of commercially available building blocks changes what is available at the end. Because every candidate is assembled from catalogue reagents through two named reactions, its product frequently falls inside the enumerated space that make-on-demand suppliers already offer. One of the three CXCR2 candidates is listed by such a supplier at 10 mg scale with a quoted lead time of 30 days to India, at a price in the low hundreds of dollars. That is a catalogue order rather than a synthesis campaign.

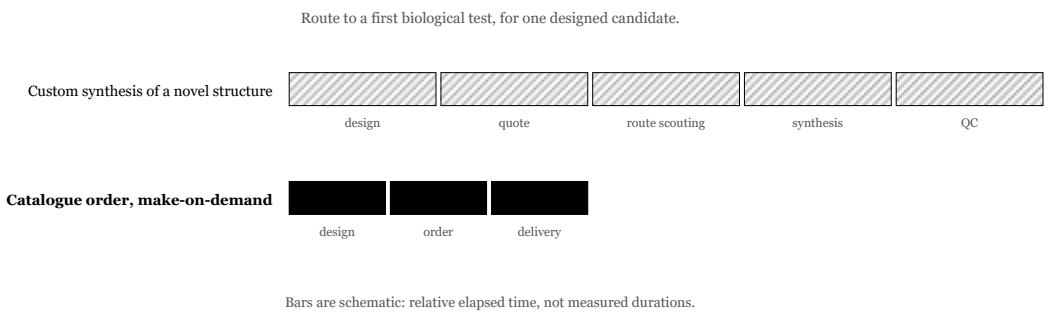


Figure 11. Route to a first biological test. Bars are schematic and show relative elapsed time rather than measured durations. The upper lane is the customary path for a novel structure; the lower lane is available when the designed product is already enumerated by a make-on-demand supplier, which is a consequence of designing over catalogue building blocks rather than an additional service.

Where a candidate is not catalogue-available, the recorded route still shortens the quoting step, because a supplier receives named reagents and two named transformations rather than a structure to be analysed retrosynthetically. We make no claim about a specific price in that case.

What the system does not provide is any evidence that these compounds engage CXCR2 or affect senescence. They are hypotheses with executable routes attached, which is the entire claim. The value is that the hypothesis can be tested in weeks by ordering starting materials, rather than debated in terms of whether the proposed structure could be made at all.

11. Discussion

Three positions follow from the results above, and each is narrower than the result might be taken to support.

Constraining the representation is cheaper than scoring the constraint. Because a molecule cannot exist here without a route, no part of the objective spends effort on synthesizability, and no delivered candidate requires retrospective triage for makeability. The cost is expressive: chemistry outside the template and building-block library is unreachable, and Figure 7 shows the reachable set is narrower than the compatibility table predicts. That trade favours this representation for programmes where candidates must be prepared quickly, and disfavours it where the goal is to explore chemical space broadly.

Efficiency gains in this setting came from scheduling, not from a better objective. The largest measured improvement here changes which reaction is attempted first and leaves the reachable product set identical. Similar reasoning applies to the tensor decomposition, which supplies feasibility information the objective never contained rather than re-weighting terms the objective already had. Every mechanism we tested that instead pressed harder on the existing objective helped some targets and degraded others, converging on no global benefit.

Calibrated disclosure is a more defensible aim than a higher score. Our candidates sit far outside the region where any published accuracy figure applies, and adding physics did not recover the loss. Given that, the useful engineering target is knowing when a prediction is uninformative, which is measurable, rather than claiming accuracy that no method in this regime currently demonstrates. This position is falsifiable: a method retaining benchmark accuracy on molecules at Tanimoto 0.15 from its training distribution would refute it.

We make no claim of superior predictive accuracy over any published system, and Section 8 explains why no method should currently claim it in this regime. What this system offers instead is a different deliverable: every proposal arrives with an executable route, a safety screen applied to the reagents as

well as the product, a calibrated statement of predictive reliability, and exact reproducibility from an identifier. For a programme deciding what to make next, those are the properties that determine whether a computational result can be acted on.

12. Limitations

No compound described here has been synthesised or assayed, and no efficacy claim is made for any candidate. Predicted ADMET values come from models trained on approved and clinical compounds and are least reliable precisely on the novel chemotypes this system produces, as Section 7 quantifies. Building-block availability is inferred from public compound records, which establishes that a reagent is known and not that it is purchasable. Scaffold-template compatibility is predicted rather than verified, with the consequences shown in Figure 7. Tissue coupling weights and hallmark priors are hand-assigned; they are used as ordinal inputs and their numerical values are not reported here as measurements. The hallmark framework itself is a taxonomy rather than a causal theory[15], and our own analysis finds the twelve axes carry roughly three independent directions rather than twelve. Finally, every model here is static: it describes differences between age groups, not rates, and a dynamical account of ageing requires longitudinal data we do not hold[22,21].

Data and code availability

Measurement harnesses reside alongside the components they measure, and each experiment reported here is reproducible from a pinned configuration. Source is available for audit on request.

Appendix A. Reaction classes

Templates are grouped by the bond they form. Each carries a reagent profile matched to that bond; a template whose transformation cannot be classified is refused rather than admitted with a default profile, because a default would prescribe reagents for chemistry it does not perform.

Table A1. Reaction classes in the shipped library (73 templates in total). Curated templates were written against a named transformation; mined templates were admitted from a patent-derived corpus only after firing on a substrate grid covering both target programmes and receiving a matched reagent profile.

Class	Bond formed	Reagent family	Note
Amide coupling	C(=O)–N	Carbodiimide or uronium activator, tertiary amine base	Most frequent class
Aryl amination	aryl C–N (amine)	Palladium, phosphine ligand, alkoxide base	Forms the aryl–morpholine motif
Amide N-arylation	aryl C–N (amide)	Copper, diamine ligand, carbonate base	Not a palladium substrate
Azole N-arylation	aryl C–N (azole)	Copper, phenanthroline ligand	Aromatic nitrogen pattern required
Nucleophilic aromatic substitution	aryl C–X displacement	Base only, polar aprotic solvent	Gated on ring activation
Reductive amination	C–N from carbonyl	Borohydride reducing agent	—
Sulfonamide formation	S(=O)2–N	Sulfonyl chloride, base	—
Ether and ester formation	C–O	Alkylation or acylation conditions	—

Appendix B. Reporting conventions

Two conventions are applied throughout, stated here so that figures in this manuscript can be compared with those reported elsewhere.

- Discrimination axes are drawn from the chance level rather than from zero. An AUC bar drawn from zero renders a coin flip as half full, which flatters every weak result.
- Point estimates from small-sample evaluations are never reported without their interval. The two retrospective evaluations in Table 2 are meaningful only as intervals.

References

1. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform*. 2009;1:8.
2. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat Chem*. 2012;4(2):90–98.
3. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. *J Chem Inf Model*. 2020;60(12):5714–5723.
4. Gao W, Mercado R, Coley CW. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. arXiv:2110.06389. 2021.
5. Generative de novo design and virtual screening of synthesizable molecules. *Curr Opin Struct Biol*. 2023;83:102715.
6. Generate what you can make: achieving in-house synthesizability with readily available resources in de novo drug design. *J Cheminform*. 2024;16:106.
7. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. *J Chem Inf Model*. 2019;59(3):1096–1108.
8. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, et al. Molecular Sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol*. 2020;11:565644.
9. Gao W, Fu T, Sun J, Coley CW. Sample efficiency matters: a benchmark for practical molecular optimization. *Adv Neural Inf Process Syst*. 2022;35.
10. Step forward cross validation for bioactivity prediction: out of distribution validation in drug discovery. bioRxiv 2024.07.02.601740. 2024.
11. ADMEOOD: out-of-distribution benchmark for drug property prediction. arXiv:2310.07253. 2023.
12. Zheng, Guo, Wang, Liu, Chen. Beyond scaffold splits: structural-frontier evaluation reveals hidden failures in ADMET models. arXiv:2607.10729. 2026. Preprint; not peer reviewed.
13. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*. 2013;153(6):1194–1217.
14. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: an expanding universe. *Cell*. 2023;186(2):243–278.
15. Gems D, de Magalhães JP. The hoverfly and the wasp: a critique of the hallmarks of aging as a paradigm. *Ageing Res Rev*. 2021;70:101407.
16. Acosta JC, O'Loghlen A, Banito A, et al. Chemokine signaling via the CXCR2 receptor reinforces senescence. *Cell*. 2008;133(6):1006–1018. PMID 18555777.
17. Sen P, Lan Y, Li CY, et al. Histone acetyltransferase p300 induces de novo super-enhancers to drive cellular senescence. *Mol Cell*. 2019;73(4):684–698. PMID 30773298.
18. Freund A, Patil CK, Campisi J. p38MAPK is a novel DNA damage response-independent regulator of the senescence-associated secretory phenotype. *EMBO J*. 2011;30(8):1536–1548. PMID 21399611.
19. Lazaar AL, Miller BE, Donald AC, et al. CXCR2 antagonist for patients with chronic obstructive pulmonary disease with chronic mucus hypersecretion: a phase 2b trial. *Respir Res*. 2020;21:149.
20. Harrison DE, Strong R, Sharp ZD, et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*. 2009;460(7253):392–395.
21. Oh HS-H, Rutledge J, Nachun D, et al. Organ aging signatures in the plasma proteome track health and disease. *Nature*. 2023;624(7990):164–172.
22. Pyrkov TV, Avchaciov K, Tarkhov AE, et al. Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nat Commun*. 2021;12:2765.

23. Landrum G. RDKit: open-source cheminformatics. https://www.rdkit.org
24. Huang HY, Broughton M, Mohseni M, et al. Power of data in quantum machine learning. *Nat Commun*. 2021;12:2631.
25. Flórez-Ablan M, Roth M, Schnabel J. On the similarity of bandwidth-tuned quantum kernels and classical kernels. arXiv:2503.05602. 2025.
26. Sarkar TJ, Quarta M, Mukherjee S, et al. Transient non-integrative expression of nuclear reprogramming factors promotes multifaceted amelioration of aging in human cells. *Nat Commun*. 2020;11:1545.
27. Wang W, Zheng Y, Sun S, et al. A genome-wide CRISPR-based screen identifies KAT7 as a driver of cellular senescence. *Sci Transl Med*. 2021;13(575):eabd2655.