# NOT EVEN ORDERED: WHAT A PUBLISHED LEADERBOARD ENTITLES A READER TO QUOTE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A leaderboard reports one total order over its methods, computed under evaluation choices its paper does not state. Those choices are a rule deciding when a prediction counts as the reference, crossed with an output budget, a post-processing step or a normalisation. We fix the items and the released predictions, re-score them in every cell of a declared grid, and ask what part of the published order survives.

We do this for 24 leaderboards in four domains, retraining nothing and reimplementing no method. Together those tables print 298 places. Read permissively, a pair stands unless some cell contradicts it, and 170 ranks stand. Read strictly, a pair stands only where every cell separates it, and 99 survive. The distance between the two readings is mostly not the choices: 159 ranks go to resolving power a benchmark never had, before any convention is varied, and 40 to the grid.

The surviving share runs from none to nine tenths, and does not replicate between two editions of one shared task. Survival is a property of a table and its grid, not of a benchmark. Of the orderings a table did establish, 6 are reversed by a declared choice and survive their own board's correction, and 5 of those survive a single correction over all 24 grids. None of the six is the matching rule's, the choice this literature discusses most. Deciding which orderings are at risk needs no pairwise statistic. We release the protocol, the audited split, the frozen predictions, the checker and the analysis code.

## 1 INTRODUCTION

This is a survey of published leaderboards and an instrument for reading them. The instrument takes a leaderboard's released predictions and a declared set of evaluation choices and returns the part of its ordering that no choice contradicts. The survey applies it to twenty-four leaderboards in four domains. One predictor in one domain is used to find the choices worth declaring. That predictor is instrumentation rather than a contribution: it is not offered as a state of the art, and nothing in the survey depends on how it performs.

The domain the choices were found in evaluates chemistry. Machine learning predicts what a molecule will undergo: the metabolites enzymes will convert a drug into, the reactants a target compound can be made from, the molecules that meet a design objective. All three are evaluated the same way. A method proposes a ranked list of structures, a curated reference records which are correct, and it is scored by how often the two agree.

Whether two structures agree is not self-evident. A single compound can be drawn in several ways a chemist reads as one substance. It may be drawn as either of two tautomers, which differ in the position of a mobile hydrogen; with or without the three-dimensional arrangement at a stereocentre; as any of several equivalent strings. An evaluation must fix a rule deciding when a predicted structure counts as the reference structure. It must also fix how many candidates a method may propose, because a method that proposes more will match more. Where predictions are generated by applying reaction templates, it must fix the convention by which a template is matched against a molecule. A template that names a hydrogen explicitly will not match a structure whose hydrogens are implicit. And it must fix which molecules the test set contains. These four choices are made in every published evaluation and stated in almost none.



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

That evaluation choices can change conclusions is established in several fields (Maziarz et al., 2025; Zhong et al., 2020; Kamalloo et al., 2023; Agarwal & Bisht, 2026). What is not established is how far any single choice moves a ranking when it is the only thing that changes. It is also not established whether a choice can *reorder* a leaderboard rather than shifting every system on it by a similar amount. The distinction matters, because a shift leaves conclusions intact and a reordering does not.

We answer that question by re-scoring predictions that other groups have released, so nothing is retrained and no method is reimplemented. Two of the four choices reorder a published leaderboard. Those reorderings are established as interactions between method and choice rather than inferred from a change in rank, and they survive a correction for the comparisons the grid creates. A third changes a reported quantity, what a rule base reaches, and reverses the ordering of the two halves of one bank without touching a leaderboard. The fourth is undeclared but produces no interaction that survives correction. Separating those two statements is possible only because the same test is applied to all four. A single condition governs when a reordering occurs: the choice must move two systems apart by more than the margin between them. That makes a benchmark more vulnerable as its leading methods converge.

**Contributions.**

- **The instrument.** What no declared choice can contradict in a published ranking is a partial order: the intersection of the per-cell total orders, computed with its longest chain from released prediction files alone. It can only fall as a grid widens, so we report a curve over the declared axes rather than one number, and count a pair's two ways of failing apart.
- **The survey.** We run it on $24$ leaderboards and $1{,}932$ pairs across retrosynthesis, metabolite prediction, protein–ligand docking and two editions of one machine translation task. They publish $298$ places, and what survives supports $170$. The share runs from none of three to nine tenths and does not replicate across two editions of the same task and language pair, so it is a property of a table and its grid, not of a benchmark.
- **What the gap is made of, which is mostly not the choices.** Requiring only a leaderboard's own cell to separate a pair takes $298$ places to $139$: $159$ ranks go to a benchmark's own power before any convention is varied and $40$ more to the grid. On none of the $24$ boards does the matching criterion certify a reversal of an ordering a published cell had separated. Verdicts are corrected by Holm over the cell-level tests each grid creates. Of the $23$ certified reversals of every kind, $15$ survive one correction over the union of all twenty-four grids. The criterion exchanges positions a benchmark never established.
- **The arithmetic.** A pair's ordering survives a change of criterion unless the change moves the two apart by more than the larger of the margins it connects. That identity is why every dominance verdict follows from $n$ scorings per cell with no pairwise statistic. A resample enters only to tell a contested pair from an unresolved one.
- **The same reversal inside a rule bank, with no leaderboard involved.** The convention a template needs is a property of whoever wrote it, so a bank assembled from several sources has no single setting that suits all of it. One bank of $7{,}581$ rules is scored on the same substrates under two conventions, both defensible and both correct arithmetic. Its curated fifth reaches $0.487$ against the mined majority's $0.800$ as deployed, and $0.668$ against $0.332$ with hydrogens expanded. The gap is $+0.313$ one way and $-0.336$ the other, and the validity floor carries $0.000$ of it. The convention does not move the number; it exchanges which half of the bank carries it. The hazard is not confined to this domain. All $42{,}554$ patterns in the standard retrosynthesis template library pin a connection count. Expanding hydrogens costs that library $82\%$ of the template–product pairs it can match, though only $11$ of $500$ products lose every applicable template.
- **The case study and the release.** The four choices were found by measuring one predictor in one domain, and that case study is reported in full: the output budget reorders a metabolite leaderboard with every ranking held fixed, while the population a dataset name refers to is ambiguous but reorders nothing that survives correction. We release the protocol, the audited split, the frozen predictions of every method compared, a checker that reports these choices from released files alone, and the analysis code.

Two claims follow, and they are different sizes. **C1**: most of what a leaderboard prints it never established: the pairs its own cell cannot separate from zero. **C2**: of the pairs it did establish, six

orderings are reversed by a declared choice and survive their own board's correction. Five of them survive the union's as well, and none is the matching rule's. Whenever both are in view, the rest of the paper says which of the two a number belongs to.

## 2 Related Work

**The domain, and what has been compared in it.** Rule-based metabolite prediction was established by SyGMa (Ridder & Wagener, 2008), which encodes phase I and phase II biotransformations as templates. It scores products by how often each template yields an observed metabolite. GLORYx (de Bruyn Kops et al., 2020) couples a comparable library to a learned model of sites of metabolism. BioTransformer (Djoumbou-Feunang et al., 2019) extends templates to microbial and environmental transformations. MetaTrans (Litsa et al., 2020) and MetaPredictor (Zhu et al., 2024) dispense with templates altogether. A further wave has appeared since: a chemical language model (Larsson et al., 2025), a mechanistically informed graph framework (Zhou et al., 2025), and a web server (Zhou et al., 2026) and a preprint (Chavan, 2026) that predict pathway, site and product jointly. None is on the board below, because that board carries the methods this work could run and freeze on this split rather than the methods the field has. The board is therefore a weak sample of the field even where the instrument reading it is not. Comparative studies recover the same coverage-precision trade-off on 85 agrochemical parents (Scholz et al., 2023). It recurs on 37 chemicals, where the broadest-coverage system also produced the most incorrect predictions (Boyce et al., 2022), and on radiolabelled ADME data (Gao et al., 2026). Each fixes one criterion for a correct prediction. None reports how sensitive its conclusion is to that choice. That one structure can be written more than one way is documented independently of this domain. Dhaked et al. (2020) catalogue tautomeric transformations the standard identifier normalises only a subset of. Hähnke et al. (2018) report roughly $60\%$ of PubChem records differing from their canonical InChI. Mansouri et al. (2024) describe a standardisation pipeline. That literature treats standardisation as preprocessing, not as a parameter of the analysis.

**Evaluation choices in adjacent fields.** None of the four choices is in doubt in general. On the matching rule, Maziarz et al. (2025) reversed a published retrosynthesis ordering under a standardised procedure, in part because one model had been scored under a relaxed stereochemistry criterion its baselines were not given. Hastedt et al. (2024) release the ranked predictions of twelve such systems. Liu et al. (2026) find divergent metric implementations in mass spectrometry. Outside chemistry, a different equivalence rule misrepresents the state of the art in text-to-SQL (Zhong et al., 2020) and changes which question-answering system leads (Kamalloo et al., 2023). On the aggregate, the metric inverts the winner in drug-response prediction (Agarwal & Bisht, 2026). The standard score in structure-based drug design rises with output atom count rather than quality (Gao et al., 2025). It motivated rebuilding that benchmark (Lin et al., 2025). Weighting by difficulty reorders natural-language leaderboards (Mishra & Arunkumar, 2021; Rodriguez et al., 2021). On the output budget the literature is older: ranked output evaluated at a declared cut-off (Järvelin & Kekäläinen, 2002), under incomplete judgements (Buckley & Voorhees, 2004), and corrected for recording propensity (Joachims et al., 2017). It also covers a cut-off metric whose winner changes with how the candidates entering it are counted (Krichene & Rendle, 2020).

**Sweeping the choices rather than fixing them.** Multiverse analysis (Steegen et al., 2016) and specification-curve analysis (Simonsohn et al., 2020) compute a result in every cell of a declared grid of defensible analytic choices. They report what survives all of them. The machine-learning counterparts account for benchmark variance (Bouthillier et al., 2021) and test a claim jointly across datasets (Dror et al., 2017). Those designs sweep an *estimate*. Section 3 sweeps an *order*, whose intersection over cells is a partial order rather than a distribution. What survives has a share and a longest chain, both computable from released prediction files. Closest in aim is work that perturbs a leaderboard and reports how far it moves. Knowles (2021) resamples the human evaluation behind the WMT ranking. Jin et al. (2026) corrects the annotations of two text-to-SQL benchmarks and reports shifts of up to nine places. Both establish that a ranking moves. Neither returns an object a reader can quote in its place, nor a rule that says in advance which pairs will move. The same fragility is reported outside chemistry. Holistic re-scoring changes which language model leads (Liang et al., 2023). A retrieval benchmark's zero-shot ordering depends on the suite it is read on (Thakur et al., 2021). An arena's ranking reflects who was allowed to submit how often (Singh et al.,

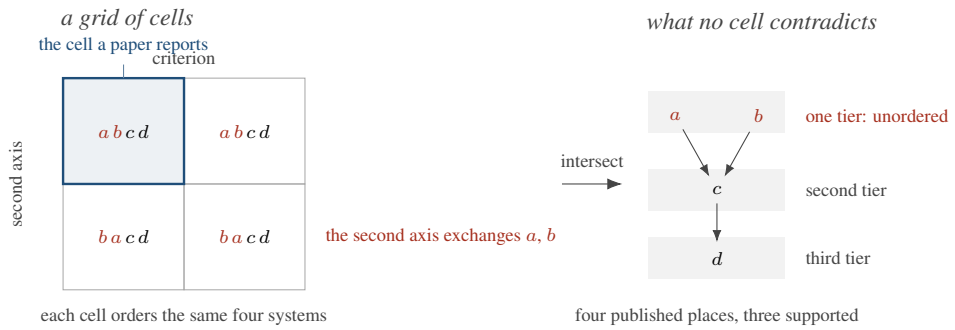



Figure 1: The instrument in miniature: four systems on a grid of four cells. Letters stand for systems, and no number in the figure is a measurement. Each cell orders the same systems, and a published table reports one cell without saying which. What survives is the intersection. Here $a$ and $b$ beat $c$ in every cell and $c$ beats $d$ in every cell, but the second axis puts $b$ above $a$ where the published cell puts $a$ above $b$. That pair survives nothing, so the four published places support three.

2025). Information retrieval has argued the statistical half of this for longer than machine learning has (Sakai, 2014; Buckley & Voorhees, 2004).

**Staged evaluation, and what is left unestablished.** Separating what a candidate generator reaches from what a later stage retains is standard (Wang et al., 2011; Covington et al., 2016; Karpukhin et al., 2020; Maziarz et al., 2025). In each case the first term is estimated from a sample or bounded by a search budget. The demonstrations above compare studies differing in several respects at once, so the effect of any one choice is not identified. Nor is it known what the first term can be when the candidate set is finite and enumerable, or whether a coverage figure obtained that way is comparable across papers. The present work addresses both.

## 3 What part of a published ranking is safe to quote

### 3.1 The instrument

The survey sweeps two of the four choices of §1, the two a released prediction file can be re-scored under. They are the matching rule, which every domain has, and whichever second choice a field reports apart from its ranking. The other two are the hydrogen convention a template match assumes and which molecules the test set contains. Neither can be varied without re-running the predictors, so both are measured in the case study instead (§3.5). Which boards enter is decided by two rules fixed before any share was read. One of them excludes a board this paper would have benefited from: a molecular-generation set whose headline quantity counts *distinct* canonical strings. Its criterion defines the metric rather than parametrising it. Its gaps are the widest relative to what the criterion moves them, so admitting it would have raised the shares reported below (Appendix D).

A leaderboard reports a total order computed in one cell of a grid its paper does not report. We fix the items and the predictions and ask of each pair whether the ordering the table asserts holds in every other cell. A cell is a criterion crossed with a second axis. The criterion is the rule deciding when a prediction is the reference. In the prediction domains the four are canonical SMILES equality, the full INCHIKEY, its stereo-blind first block, and a tautomer-aware hash. In docking they are distance to the crystal pose at two tolerances with and without a validity screen; in translation, what an expert annotation is read as. The second axis is what the field reports apart from its ranking and never jointly with it: a budget, a post-processing step, or a normalisation (Appendix D). The higher system *dominates* the lower when the margin is positive in every cell, and is *separated in every cell* when a paired interval excludes zero in each. Dominance is the intersection of the per-cell orders, so it is transitive and the finest relation no declared choice can contradict. Separation is a sub-relation, reported apart because an interval condition is not transitive.

Formally, fix a set of systems $S$, a set of items, and a grid $C$ of cells. In cell $c$ each system has a mean score, and those means induce a total preorder $\prec_c$ on $S$. A published table reports $\prec_{c_0}$ for one cell $c_0$

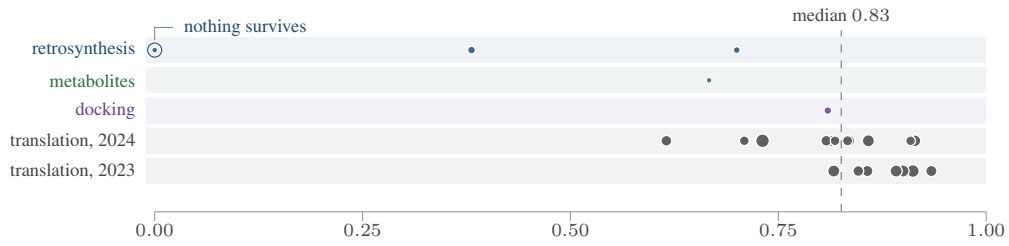



Figure 2: Each leaderboard is one dot at the share of its pairwise orderings that no cell of its own declared grid contradicts. Dot area grows with the pairs a board carries, with a floor so that a three-pair board stays visible. The two editions of the translation task are drawn apart because five of their language pairs are the same pair scored twice, a year apart, under different protocols.

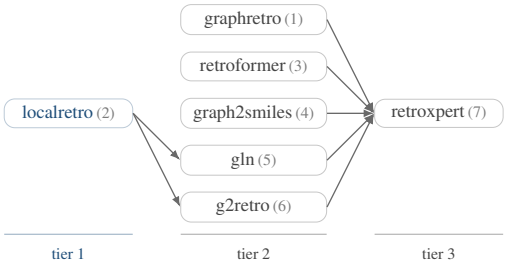



Figure 3: The dominance order of the seven-system retrosynthesis table, drawn as its transitive reduction. An arrow means the left system beats the right in all sixteen cells of the declared grid. Systems drawn together are ones the grid does not order. A tier is a system's position in the longest chain and not a block that beats the next one: the tier-1 system dominates two of the five drawn beside it, and the arrows a reader does not see are orderings the grid does not support. The number after each name is the place its published table gives it.

its paper does not name. Write $\prec_* = \bigcap_{c \in C} \prec_c$. A board ranking $n$ systems prints $n$ *places*, the rank slots a reader takes away from it. What $\prec_*$ supports is its longest chain. The number of places that survive is the length of the longest sequence of systems each of which dominates the next.

This relation is not new. It is the unanimity, or Pareto, relation on a profile of preorders, which social choice has studied since Arrow (1951) and Sen (1970). The two facts below are its standard properties. They are stated because the survey quantifies over them, and both proofs are two lines and are given in Appendix D. This paper contributes the measurement, not the relation.

> **Proposition 1.** $\prec_*$ is a strict partial order, and it is the finest relation on $S$ that no cell of $C$ can contradict.
>
> **Proposition 2.** If $C \subseteq C'$ then $\prec_*^{C'} \subseteq \prec_*^{C}$: the dominating set is monotone non-increasing in the grid, and so is the number of tiers.

Proposition 2 has a consequence for reporting. A share is not a property of a leaderboard but of the pair (leaderboard, grid), because it can only fall as a grid widens. Any paper reporting one could therefore have reported a larger one by declaring fewer choices. The remedy is to report the curve rather than the point: each declared axis alone, then their product (Appendix D).

All twenty-four are run on a grid of four criteria crossed with a second axis their field reports apart and never jointly (Appendix D). It is the budget a table is read at, whether a docked pose is minimised first, whether an error score is floored, or whether a human score is standardised within the annotator who gave it.

The last three columns are the ones to take away, because most of what a leaderboard prints was never resolved at all. Eleven of the seven-system table's twenty-one pairs carry margins its own benchmark cannot separate before any choice is varied. Across all twenty-four leaderboards the share runs from none of three to nine tenths with a median of $0.83$. It is not a constant of the field, and none

| leaderboard | places | items | dominate | contested (certified) | unresolved | places supported |
|---|---|---|---|---|---|---|
| retrosynthesis, seven systems | 7 | 4,999 | 8 of 21 | 9 (8) | 4 | 3 |
| retrosynthesis, five systems | 5 | 4,948 | 7 of 10 | 3 (2) | 0 | 3 |
| retrosynthesis, three systems | 3 | 4,982 | 0 of 3 | 2 (2) | 1 | 1 |
| metabolites, three methods | 3 | 1,170 | 2 of 3 | 1 (0) | 0 | 2 |
| docking, seven programs | 7 | 428 | 17 of 21 | 2 (0) | 2 | 4 |
| translation, nineteen systems | 19 | 486 | 125 of 171 | 17 (6) | 29 | 9 |
| translation, fifteen systems | 15 | 559 | 96 of 105 | 1 (0) | 8 | 8 |
| translation, nine further pairs | 115 | 634–1,954 | 553 of 689 | 33 (4) | 103 | 62 |
| translation, the year before | 124 | 509–1,120 | 804 of 909 | 20 (1) | 85 | 78 |
| all twenty-four | **298** | — | 1,612 of 1,932 | 88 (23) | 232 | **170** |

Table 1: What each leaderboard publishes and what survives its own grid. *Places* is the number of systems the table ranks, and *places supported* is the longest chain of the dominance order: the ranks a reader may quote without naming a cell. *Certified* counts the contested pairs surviving Holm over the family each grid creates. The last row is the survey's total.

is reported. Summed, they publish 298 places and support 170. Most of that gap is a benchmark's own power rather than a convention (Appendix D). The longest chain reports the same measurement in a readable form. Seven published places support 3 tiers, resampling to $[3, 3]$; nineteen support 9; and the same seven systems take 4 total orders across 16 cells. The system that table places first is not in the top tier of what survives. The one it places second is the only system that is (Figure 3). Docking and translation are controls. Docking's field published its reversal already (Buttenschön et al., 2024) and the grid recovers it. Translation's nineteen boards contest 71 pairs, of which 11 survive their own boards' correction and 4 the union of all twenty-four. Not one of the 11 reverses an ordering its own published cell had separated from zero. All six such reversals are the budget's, never the matching criterion's (Appendix D). A null of that shape needs a size, and the identity supplies one without a further statistic. On the two retrosynthesis boards and the metabolite one, the matching rule erases at most $0.5755$ of a margin the published cell had separated. The median comparison erases none of it. Docking is the exception and is reported as one. Translation supplies the first claim at scale, on pairs a table never established, and chemistry supplies the second. What does not survive on the docking board includes its own top pair. Because dominance is an intersection, all of this can only shrink as cells are added, so each axis is also reported alone (Appendix D).

**Three verdicts, not two.** A pair that does not dominate has failed in one of two ways that should not be added together. In the first, a cell puts the two systems the other way round with a paired interval excluding zero, so a choice decides against the published ordering. In the second, no cell does so, and the published ordering merely fails to hold everywhere on a margin the benchmark cannot resolve. The first is called *contested* and the second *unresolved*. Only the count of contested pairs is evidence that a choice reverses anything. Counting sign changes alone would let noise on a tied pair read as a reversal, the error this instrument exists to avoid.

**The five terms.** Three are verdicts and two are conditions on them. A pair *dominates* when the higher system's margin is positive in every cell, which makes dominance an intersection of the per-cell orders and therefore transitive. It is *separated in every cell* when a paired interval on that margin also excludes zero in each, reported apart because an interval condition is not transitive. A pair that does not dominate is *contested* when some cell reverses it with a paired interval excluding zero, and *unresolved* when none does and the margin is one the benchmark cannot resolve. A contested pair is *certified* when it survives Holm at $\alpha = 0.05$ over the cell-level tests its own grid creates; the word records that correction and nothing more.

### 3.2 A CONTROL WITH A PUBLISHED ANSWER

**Docking as a control with a published answer.** Three of the four boards are audited here for the first time, so the instrument's verdicts there are unvouched for. A rule that reports little surviving everywhere cannot be told apart from a rule that reports little because it is insensitive. Docking

supplies the missing arm, because its own field published a reversal of this kind before we ran anything. Buttenschön et al. (2024) argued that a pose can sit within the field's distance tolerance of the crystal ligand and still be chemically impossible. They named UNI-MOL as the program most affected, and released the per-pose table behind that argument. That table, read through the dominance order without reference to the argument, contests exactly two of the 21 pairs, and both are UNI-MOL against programs it outranks on distance alone. That is the source's own finding, recovered by an instrument that was not told it.

The verdict is read out of 21 pairs by 8 cells, so it is granted inside a family of 168 cell-level tests. Correcting for them takes it back. Neither reversal survives Holm at $\alpha = 0.05$ over the family. The board scores each item a hit or a miss, so the exact conditional test on the discordant items is available, and these $p$ are that test. The smallest reversal reaches $1.5 \times 10^{-3}$ against a threshold of $9.8 \times 10^{-4}$. Two exchanges out of 168 looks is about what noise alone would produce. This control establishes that the grid recovers the source's finding as a contested verdict, and that it certifies neither half of it.

The control does not show that this board is clean. Of its 21 pairs, 17 dominate, 2 are contested and 2 the grid leaves unresolved. One of those two is the board's own top pair, VINA against GOLD, whose margin no cell of the grid separates from zero. A benchmark can be large, dense and carefully scored and still not entitle a reader to its first place over its second. The control is that a well-run board loses four pairs where a mature retrosynthesis table loses thirteen, so the instrument distinguishes them. It is not that any board survives whole.

No pair is contested between the search-based programs and the learned ones. On these 428 items all ten such pairs dominate, because VINA and GOLD lead on distance alone before validity is asked for. The exchange the source is best known for lives on the other set it distributes, and the grid finds it there. Run unchanged on the 85 astex items, the same eight cells contest VINA against DIFFDOCK along with two other pairs, and leave 15 of 21 dominating. We report that run as a second board and not as two more cells, because a population is not a choice about how a prediction is scored. We report it at all because which reversal a grid recovers depends on which items a benchmark chose, which is the sharpest form of the paper's own argument we have found.

One axis of this board is weaker than a criterion and is labelled so. Minimising a pose before scoring it is a step the benchmark applies rather than the program. It is a choice the evaluation could have been made under. Unlike a matching rule, it does not leave the prediction alone. 99.7% of the 2,922 paired poses differ between the arms, by a median of 0.354 Å, and 6.5% of per-item verdicts change at the field's tolerance. The conservative reading is therefore the criteria alone at the published post-processing, which leaves 18 of the 21 pairs dominating rather than 17. The pair the grid loses to that axis is named in the tables below rather than absorbed into the total.

The released table forces three conventions, which are declared rather than chosen: which flatness checks a criterion may use, how a row with no scorable pose counts, and what to do with the crystal ligand. Each is stated with its effect on the published cell in Appendix D.1.

### 3.3 WHICH AXIS MOVES WHAT, AND WHAT THE GAP IS MADE OF

**Which choice the certified reversals belong to.** Of the twenty-three certified reversals, six reverse an ordering the published cell had separated. The other seventeen reverse pairs the table never established, which reaches the first claim by a second route. What matters about the six is the axis that moves them, not their domain. All six are retrosynthesis pairs. Each reverses at a budget the *published* criterion is held at, so the budget moves them and the matching rule does not. Five reverse at top-3 and top-5. The sixth, on the extrapolation board, reverses only at top-10 and is the one the union correction removes. The matching criterion therefore certifies no reversal of a separated ordering anywhere in this paper. That is a smaller second claim than a paper about matching rules might suggest. It is the one the measurements support. The criterion exchanges positions a benchmark never separated, which is the first claim. What reverses an established ordering is a budget nobody reports jointly with the ranking it produces.

**Decomposing the gap between published and supported ranks.** Proposition 2 keeps the ways a pair can fail as different quantities, and the same discipline applies to the count that rolls them up. There are three readings of "supported". Each is a longest chain in a relation on the same systems. They differ only in what an edge requires. When a leaderboard's *own* cell must separate the pair, the

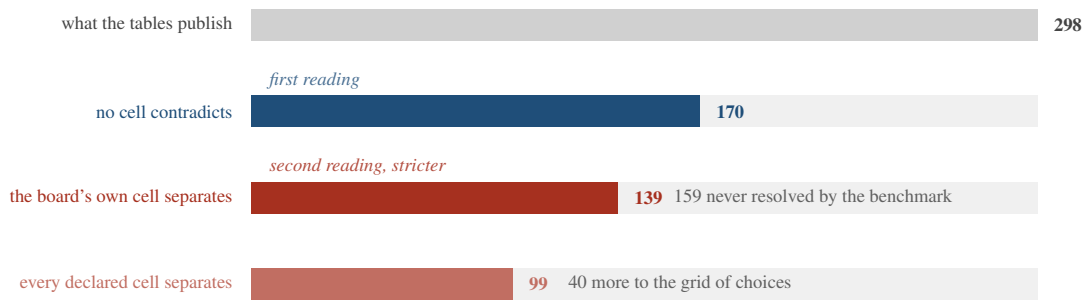



Figure 4: The two readings of what a leaderboard supports, on the same scale of published places. The upper bar asks only that no cell contradict the published order. The lower two ask each cell to separate the pair, and the places they remove are attributed.

298 published places support 139. The 159 lost go to the benchmark's power before any convention is varied. When every declared cell must separate it, they support 99, so the grid costs 40 on top. The 170 the abstract leads with is the weaker reading an intersection of orders gives, in which a cell must agree in sign and need not separate. It is quoted because dominance is what the instrument computes. A reader who wants the strictest number has 99. One who wants what the conventions cost has 40.

**The share does not replicate across two editions of one task.** Nineteen of the twenty-four boards come from one shared task, which invites the objection that they are one dataset counted nineteen times. Two editions of that task answer it. They ran a year apart under different protocols — error spans and a hundred-point score in one, direct assessments with several ratings a segment in the other. Five language pairs appear in both. The surviving share does not carry over. It moves by 0.114 on average and by 0.285 at the widest. EN-ZH stands at 0.900 in one edition and 0.615 in the next, on the same language with the same organisers. The share is a fact about a table and its grid, not about a benchmark or a language. A paper quoting one as a fingerprint, this one included, would repeat the error it is about (Appendix D).

### 3.4 Why the verdicts need no pairwise statistic

**Why the verdicts need no pairwise statistic.** Write $d$ for a pair's margin in the published cell, $d'$ for its margin in another, and $\delta_a, \delta_b$ for how far each system moves between them, so $d' = d + (\delta_a - \delta_b)$. The sign survives iff $|\delta_a - \delta_b| \le \max(|d|, |d'|)$. This is an identity, not a finding. It says only that a difference changes sign when the movement exceeds both margins. It is stated for what follows from it, not as a result. Scoring the $n$ systems once in each cell fixes every one of the $n(n-1)/2$ dominance verdicts, with no pairwise statistic anywhere. A resample enters only to tell a contested pair from an unresolved one. On the twenty-four tables that leaves 320 of 1,932 pairs failing to dominate, of which 319 are the pairs some cell reverses and one fails on an exact tie, computed from $n$ numbers per cell rather than from 1,932 paired bootstraps.

The arithmetic does not say where those pairs are, and that answer is a ratio rather than a level. The question can be asked of twenty-four leaderboards: the twenty-three plus a molecular-generation board whose second axis the grid does not cover. Their comparisons at risk run from 0% to 15.9%. The widest-gapped board is not the safest, because exposure is what happens when entries converge faster than the conventions separating them. Exposure is readable before any convention is varied (Appendix H). Substituting a typical movement for the domain in place of the per-system one is all a published table alone supports. It over-predicts fragility by a third, and is reported as a failure (Appendix G).

### 3.5 The case study the choices were found in

**How one choice breaks an ordering.** Hastedt et al. (2024) released the ranked predictions of twelve retrosynthesis systems, eleven of which parse, on three test sets (Appendix H.21). The largest group is seven systems on their 4,999 common reactions. Their predictions never move, yet they

change places at top-1. Five of twenty-one pairs exchange, and disregarding stereochemistry takes the fourth-placed system to first. An exchange describes the data. The claim is the interaction, corrected across the largest family declared in this work: the grid the two axes admit. Of its $448$ testable cells, 233 have intervals excluding zero and $110$ survive Holm at $0.05$, concentrated at top-1, the number this field leads with (Appendix H).

**How many of the margins the benchmark resolves at all.** Since the predictions are frozen, a paired bootstrap gives an interval on each margin directly. **Eleven of the twenty-one margins do not separate from zero, and five of the six adjacent margins in the published order do not.** The smallest margin this benchmark separates is $0.0146$. The largest amount any criterion moves a margin on it is $0.0154$, and the three-system group's narrowest is $0.0213$. The criterion therefore exchanges pairs the benchmark never separated. Where a margin is separated it erases the gap instead of reversing it.

**One split, two populations.** A benchmark can mix populations under one name. Of the 1,170 substrates here, $845$ are parent drugs and $325$ are annotated products of another substrate in the split. One question is about a compound a patient swallows and the other about what is left of it, averaged into one number. Methods do not treat them alike. GRAIL loses $0.152$ $[0.101, 0.202]$ of recall between the two while MetaPredictor loses $0.001$ $[-0.053, 0.057]$, an interaction of $+0.150$ $[+0.084, +0.216]$, and $+0.116$ $[+0.051, +0.181]$ against SyGMa. No paper reporting on this split reports the mixture. The other population question returns a null. A familiar name can refer to more than one released test set. Which one it refers to changes an ordering in $44$ of $380$ comparisons and $10$ of $60$ on the metabolite split. None of the $440$ interactions survives Holm at $0.05$, in a design whose blindest test would see $0.116$ and $0.202$ at $80\%$ power (Appendix H.7).

## 4 Limitations

**What the metabolite numbers bound.** The five-method table rests on $150$ shared substrates of frozen predictions, because MetaTrans's pipeline no longer reproduces and BioTransformer's configuration is unscripted. The other three are scored on the full $1{,}170$. An unannotated but real product counts as a false positive, so every precision and F1 figure is pessimistic by a bounded amount the annotation graph measures (Appendix L). Depth-2 application lifts the ceiling by $\sim 0.012$ at $8.5$ times the cost, so $0.817$ bounds this formulation and not the task. This system's rule labels are built with the substrate expanded, while its rules fire without it. Locating its shortfall in selection is therefore conditioned on a mismatch of about half the rule–substrate positives (Appendix E.5).

The cross-domain numbers come from files produced elsewhere, the reordering there is a top-1 phenomenon, and two results are bounded nulls.

## 5 Conclusion

Four choices are settled before any model is compared, and none is routinely declared. A published ranking entitles a reader to quote fewer ranks than it prints, and falls short in two ways of very different sizes: most of its pairs were never resolved, and a few that were are reversed by a choice no paper reports. Which pairs a choice reaches follows from an equivalence. Declaring is not the whole remedy, as the one bank of mixed provenance here shows.

**What to declare, and what to do with a partial order.** A leaderboard whose ordering is meant to be quoted should print five things. The rule deciding when a prediction counts as the reference, with the version of the toolkit implementing it. The number of candidates each method was allowed. Any post-processing applied before scoring. The population the test set draws from. Either a score per cell of that grid or the predictions themselves, from which any cell can be recomputed. A reader given those can compute the order reported here. Where that order is partial, a claim of the form "ours is first" is available only when no other system shares its top tier. Where one does, the statement that the data support names the tier instead. The share itself should not be compared between papers. By Proposition 2 it falls as a grid widens, so a paper declaring fewer choices reports a larger one. A quantity that rewards declaring less is not one to compete on.

### AI USE STATEMENT

Generative AI was used to implement methods: most of the analysis code, the evaluation harness and the checking scripts released with this paper were written with an AI coding assistant. It was used to clean and reformat data, in the parsers that read each leaderboard's released prediction files into the form the harness scores, and to support qualitative analysis in the eligibility screen over candidate leaderboards recorded in Appendix H.1. It was used to give feedback on methodology: several of the checks reported here, including ones that found defects in earlier drafts, were suggested that way and were then specified and run by the authors. It was not used to run the experiments: every measurement reported here was produced by a released script the authors invoked, and the artifact each wrote records the commit it ran at. It was used to develop conceptual frameworks, in that the three verdicts and the terms of §3.1 were settled in discussion with an assistant. It was used for the recommended tasks of code editing, figure creation, literature search, brainstorming and paper editing, and for drafting: most of this manuscript's prose, and the readings of measurements in it, were drafted with assistance and revised by the authors against the artifacts.

It was not used to generate research ideas or hypotheses. The four evaluation choices this paper is about were found by measuring one predictor in one domain, and the claims made about them are the authors'. It was not used to generate synthetic data: every number comes from a released prediction file or from a run of a released model, and nothing here is simulated. It was not used to develop theoretical models, to formulate mathematical claims, or to write proofs or provide ingredients for them; the propositions are the authors', and the relation they rest on is the unanimity relation of social choice, which this paper cites rather than introduces. Translation is not applicable to this work.

We have reviewed all AI-assisted work. Every quantity printed here is checked against the artifact it comes from by a released script that exits non-zero on disagreement, and code written with assistance is covered by the same gates as the rest. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work re-scores predictions that other groups released, and reports that a published ranking supports fewer of its own pairwise claims than it prints. No human subjects are involved and no new data are collected from people. Every prediction file re-scored here is one its authors chose to release, is used for the purpose it was released for, and is cited; the licences under which each was released are recorded in the archive, and no file is redistributed against its terms.

The finding is critical of published work, and two things follow. Named systems and named benchmarks are compared under choices those papers did not state, so the results say what a reader may quote and not that any author erred; where a reversal is reported it is reported with the interval that establishes it and with the cell it belongs to. And the paper's own conduct is held to the standard it asks of others: the choices this work makes are declared, the grid is swept rather than fixed at a favourable cell, and its own negative results are released.

### REPRODUCIBILITY STATEMENT

Every quantity reported here derives from a committed artifact and regenerates from the released material. The checks that establish this are released as executable scripts. One of them is intended for prediction files produced elsewhere. Given a template bank, it reports which templates depend on the hydrogen convention and whether the bank mixes conventions. Given two or more prediction files, it reports whether they were produced on the same population and how far apart their output sizes are. It requires no reference data and no trained model. Applied to the eleven released retrosynthesis files, it recovers the three test sets in seconds.

The cost of regeneration differs by analysis. Every split-level result regenerates from the committed artifacts alone, which comprise the frozen predictions, the matching-key tables and the per-substrate records. The ablations that rank or re-select candidates also require the trained checkpoints. The anonymised archive contains them together with the frozen predictions of every method compared, the audits, the evaluation harness, the split construction and the analysis code.

Three constraints bear on exact, not approximate reproduction. One result depends on a third-party distribution that is not redistributed here. The reach of BioTransformer's template set is computed

from the 994 SMIRKS shipped with that tool, which an installed copy supplies. Its weights and those of MetaPredictor are likewise third-party and are cited rather than shipped. Results measured on a subsample of a split record the cap and the seed that drew it. Without those the substrate set is not recoverable, since sampling proceeds without replacement and a different cap yields a different set of the same size, not a subset. Confidence intervals are regenerable: per-substrate rows are sorted before resampling, so a repeated run reproduces an artifact exactly.

Two standing reconciliations guard the analyses against the code they analyse. Where an analysis path re-derives a deployed output, the two are verified equal before use. The released per-candidate score dump reproduces the deployed ranking on all 1,170 substrates. Where a reimplemented loop stands in for the deployed one, it is gated against a figure the deployed loop produced, at the scale at which the claim is made. The switchable rule applier the engine arms are measured with has to reproduce the deployed loop's coverage ceiling of 0.8007 on the same 245 substrates before any of its arms is read. The released check refuses to write its artifact otherwise.

Every set treated here as external or third-party has been audited for overlap with the training and validation substrates. The GLORYx set overlaps on 24 of its 37 drugs, and §A.4 reports GRAIL on the 13 that do not overlap. The 150-substrate shared subset overlaps on none of its 150. The audit is released as a script.

## REFERENCES


Dhruv Agarwal and Riya Bisht. The metric picks the winner: Evaluation choice flips model rankings for drug-response prediction in unseen chemistry. *arXiv preprint arXiv:2606.12639*, 2026.

Kenneth J. Arrow. *Social Choice and Individual Values*. Wiley, New York, 1951.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Matthew Boyce, Brian Meyer, Chris Grulke, Lucina Lizarraga, and Grace Patlewicz. Comparing the performance and coverage of selected in silico (liver) metabolism tools relative to reported studies in the literature to inform analogue selection in read-across: A case study. *Computational Toxicology*, 21:100208, Feb 2022. ISSN 2468-1113. doi: 10.1016/j.comtox.2021.100208. URL http://dx.doi.org/10.1016/j.comtox.2021.100208.

Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32, 2004. doi: 10.1145/1008992.1009000.

Martin Buttenschön, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Swapnil Chavan. Metabolite-Gen: Joint prediction of sites of metabolism, reaction class, and metabolite structures using graph neural networks, 2026. ChemRxiv preprint, posted 6 May 2026.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pp. 191–198, 2016. doi: 10.1145/2959100.2959190.

Christina de Bruyn Kops, Martin Šícho, Angelica Mazzolari, and Johannes Kirchmair. Gloryx: Prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. *Chemical Research in Toxicology*, 34(2):286–299, Aug 2020. ISSN 1520-5010. doi: 10.1021/acs.chemrestox.0c00224. URL http://dx.doi.org/10.1021/acs.chemrestox.0c00224.

Devendra K. Dhaked, Wolf-Dietrich Ihlenfeldt, Hitesh Patel, Victorien Delannée, and Marc C. Nicklaus. Toward a comprehensive treatment of tautomerism in chemoinformatics including in

inchi v2. *Journal of Chemical Information and Modeling*, 60(3):1253–1275, Feb 2020. ISSN 1549-960X. doi: 10.1021/acs.jcim.9b01080. URL http://dx.doi.org/10.1021/acs.jcim.9b01080.

Yannick Djoumbou-Feunang, Jarlei Fiamoncini, Alberto Gil-de-la Fuente, Russell Greiner, Claudine Manach, and David S. Wishart. Biotransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics*, 11 (1), January 2019. ISSN 1758-2946. doi: 10.1186/s13321-018-0324-5. URL http://dx.doi.org/10.1186/s13321-018-0324-5.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. In *Transactions of the Association for Computational Linguistics*, volume 5, pp. 471–486, 2017.

Bowen Gao, Haichuan Tan, Yanwen Huang, Minsi Ren, Xiao Huang, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. Reframing structure-based drug design model evaluation via metrics correlated to practical needs. In *International Conference on Learning Representations (ICLR)*, 2025.

Lei Gao, Shu Yan, Kaiyuan Feng, Haonan Liu, Zihao Zhang, and Xingxing Diao. Benchmarking in silico metabolite prediction tools against human radiolabeled adme data for small-molecule drugs. *Journal of Chemical Information and Modeling*, 66(5):2918–2928, Feb 2026. ISSN 1549-960X. doi: 10.1021/acs.jcim.5c03045. URL http://dx.doi.org/10.1021/acs.jcim.5c03045.

Friedrich Hastedt, Rowan M. Bailey, Klaus Hellgardt, Sophia N. Yaliraki, Ehecatl Antonio del Rio Chanona, and Dongda Zhang. Investigating the reliability and interpretability of machine learning frameworks for chemical retrosynthesis. *Digital Discovery*, 3:1194–1212, 2024. doi: 10.1039/D4DD00007B.

Volker D. Hähnke, Sunghwan Kim, and Evan E. Bolton. Pubchem chemical structure standardization. *Journal of Cheminformatics*, 10(1), August 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0293-8. URL http://dx.doi.org/10.1186/s13321-018-0293-8.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 935–944. ACM, Aug 2016. doi: 10.1145/2939672.2939756. URL http://dx.doi.org/10.1145/2939672.2939756.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Tengjun Jin, Yoojin Choi, Yuxuan Zhu, and Daniel Kang. Pervasive annotation errors break text-to-SQL benchmarks and leaderboards. *arXiv preprint arXiv:2601.08778*, 2026. URL https://arxiv.org/abs/2601.08778.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 781–789, 2017. doi: 10.1145/3018661.3018699.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with datasail. *Nature Communications*, 16(1), April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58606-8. URL http://dx.doi.org/10.1038/s41467-025-58606-8.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.307/.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Rebecca Knowles. On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 464–477, 2021. URL https://aclanthology.org/2021.wmt-1.56/.

Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W. Klau, and Sebastian Böcker. Coverage bias in small molecule machine learning. *Nature Communications*, 16:554, 2025.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1748–1757, 2020. doi: 10.1145/3394486.3403226.

Sofia Larsson, Miranda Carlsson, Richard Beckmann, Filip Miljković, and Rocío Mercado. LAGOM: A transformer-based chemical language model for drug metabolite prediction. *Artificial Intelligence in the Life Sciences*, 8:100142, 2025. doi: 10.1016/j.ailsci.2025.100142.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. arXiv:2211.09110.

Haitao Lin, Guojiang Zhao, Odin Zhang, Yufei Huang, Lirong Wu, Cheng Tan, Zicheng Liu, Zhifeng Gao, and Stan Z. Li. CBGBench: Fill in the blank of protein-molecule complex binding graph. In *International Conference on Learning Representations (ICLR)*, 2025.

Eleni E. Litsa, Payel Das, and Lydia E. Kavraki. Prediction of drug metabolites using neural machine translation. *Chemical Science*, 11(47):12777–12788, 2020. ISSN 2041-6539. doi: 10.1039/d0sc02639e. URL http://dx.doi.org/10.1039/D0SC02639E.

Hongxuan Liu, Roman Bushuiev, Ivy Lightheart, Mrunali Manjrekar, et al. Massspecgym in the wild: Uncovering and correcting evaluation pitfalls in ai-driven molecule discovery. *arXiv preprint arXiv:2606.19624*, 2026.

Kamel Mansouri, José T. Moreira-Filho, Charles N. Lowe, Nathaniel Charest, Todd Martin, Valery Tkachenko, Richard Judson, Mike Conway, Nicole C. Kleinstreuer, and Antony J. Williams. Free and open-source qsar-ready workflow for automated standardization of chemical structures in support of qsar modeling. *Journal of Cheminformatics*, 16(1), February 2024. ISSN 1758-2946. doi: 10.1186/s13321-024-00814-3. URL http://dx.doi.org/10.1186/s13321-024-00814-3.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin H. S. Segler. Re-evaluating retrosynthesis algorithms with syntheseus. *Faraday Discussions*, 256:568–586, 2025. ISSN 1364-5498. doi: 10.1039/d4fd00093e. URL http://dx.doi.org/10.1039/D4FD00093E.

Swaroop Mishra and Anjana Arunkumar. How robust are model rankings : A leaderboard customization approach for equitable evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13561–13569, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i15.17599. URL http://dx.doi.org/10.1609/aaai.v35i15.17599.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (moses): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11, December 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.565644. URL http://dx.doi.org/10.3389/fphar.2020.565644.

Lars Ridder and Markus Wagener. Sygma: Combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem*, 3(5):821–832, May 2008. ISSN 1860-7187. doi: 10.1002/cmdc.200700312. URL http://dx.doi.org/10.1002/cmdc.200700312.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.346. URL http://dx.doi.org/10.18653/v1/2021.acl-long.346.

Tetsuya Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

Vincent-Alexander Scholz, Conrad Stork, Markus Frericks, and Johannes Kirchmair. Computational prediction of the metabolites of agrochemicals formed in rats. *Science of The Total Environment*, 895:165039, Oct 2023. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2023.165039. URL http://dx.doi.org/10.1016/j.scitotenv.2023.165039.

Amartya K. Sen. *Collective Choice and Social Welfare*. Holden-Day, San Francisco, 1970.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Shivalika Singh et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Sara Tanovic, Ewa Wieczorek, and Fernanda Duarte. An exploration of dataset bias in single-step retrosynthesis prediction. *Digital Discovery*, 2026.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Lidan Wang, Jimmy Lin, and Donald Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 105–114, 2011. doi: 10.1145/2009916.2009934.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://aclanthology.org/2020.emnlp-main.29/.

Yiling Zhou, Dejun Jiang, Xiao Wei, Jiacai Yi, Yikun Wang, Youchao Deng, and Dongsheng Cao. DeepMetab: a comprehensive and mechanistically informed graph learning framework for end-to-end drug metabolism prediction. *Chemical Science*, 16:18884–18902, 2025. doi: 10.1039/d5sc04631a.

Yiling Zhou, Sen Yang, Xiaoli Wang, Yuanhang He, Yao Tian, Jiacai Yi, Yikun Wang, Youchao Deng, Dejun Jiang, and Dongsheng Cao. DeepCYP: an integrated deep learning web server for the holistic "pathway-site product" prediction of CYP450 metabolism. *Nucleic Acids Research*, 54: W22–W32, 2026. doi: 10.1093/nar/gkag478.

Keyun Zhu, Mengting Huang, Yimeng Wang, Yaxin Gu, Weihua Li, Guixia Liu, and Yun Tang. Metapredictor: in silico prediction of drug metabolites based on deep language models with prompt engineering. *Briefings in Bioinformatics*, 25(5), Jul 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae374. URL http://dx.doi.org/10.1093/bib/bbae374.

## How to read the appendices

They divide in two, and the division matters for what rests on them. Appendices C and D carry the instrument: the definitions the propositions quantify over, the per-board tables behind every share, the three multiplicity corrections and the checks on the test itself. Appendix H carries the survey's own scope and the boards outside chemistry. Everything else is the case study — one predictor in one domain, reported in full because that is where the four choices were found and because a reader is entitled to see the measurements that produced them. No claim of the survey depends on any of it: the case study's own numbers are about that predictor, and it is not offered as a state of the art.

# A THE CASE STUDY: ONE PREDICTOR, ONE DOMAIN, FOUR CHOICES

## A.1 THE EVALUATION PROTOCOL: WHAT IS DECLARED AND WHAT IS HELD FIXED

The protocol used throughout is a matching specification, an audited split and a re-scoring harness. It is referred to below as TAME, for tautomer-aware matched evaluation, and it fixes what every number in this appendix is measured under.

**The matching criterion.** The field has never agreed what counts as a match. GLORYx uses a stereo-blind InChIKey skeleton, MetaTrans compares Morgan fingerprints, and rule-based systems that canonicalise their own output compare canonical SMILES, each silently. TAME names all five and maps both sides through one of $q \in \{\text{CANONICAL}, \text{INCHIKEY}, \text{INCHI-NO-STEREO}, \text{TANIMOTO1}, \text{INCHIKEY-TAUTOMER}\}$, a prediction counting as a hit iff $q(\hat{y}) = q(y)$ for some reference $y$, under RDKit 2022.09.5. They are canonical SMILES equality; the full InChIKey; its first block, blind to stereochemistry, charge and isotopes; Tanimoto similarity of one over Morgan fingerprints; and standardising both sides before hashing. A ceiling declares the *application* procedure too (§A.2). One caution attaches to the first: it is computed without stereochemistry on the metabolite corpus, whose references carry none, and with it in the cross-domain harnesses, whose references do, which is why the cross-domain step to INCHI-NO-STEREO moves those systems and the metabolite step cannot. The fourth is a similarity threshold rather than an identity relation and is reported but never treated as a candidate convention (Appendix D).

All five are reported for every comparison. The default is INCHIKEY-TAUTOMER: the references neither carry stereochemistry nor distinguish tautomers, and standard InChI normalises only a subset of them (Dhaked et al., 2020), so an engine emitting the keto form of an enol reference fails to match under plain INCHIKEY while producing the same substance. Standardisation of this kind is established preprocessing; adopting it as the *matching* rule is not.

**Split and leakage audit.** All splits are substrate-disjoint, and a read-only audit canonicalises every SMILES and verifies zero substrate and zero positive-pair overlap across each split pair. Leakage-aware splitting is not novel (Joeres et al., 2025); the leak that matters here is substrate-metabolite *identity* overlap, specific to this domain.

**Metrics and inference.** Results are reported as recall@$k$ together with the mean output size. Precision alone is unsuitable under incomplete annotation, where an unannotated but real metabolite counts as a false positive; recall alone is unsuitable because a method can raise it by emitting more (Appendix J).

Two words are used throughout and are not interchangeable. A difference is *separated* when its paired bootstrap interval excludes zero, which needs no family and is the only claim made about an exploratory quantity. It is *certified* when it is separated *and* survives Holm correction inside a family fixed by a rule stated before the correction was computed, the grid the declared axes admit, and not by which comparisons proved worth reporting; no family here was pre-registered and Appendix H.16 says so. Three comparisons are confirmatory: the differential criterion sensitivity of §A.4, the learned-against-prior gap of §E.10, and the cross-domain interaction, whose $448$ tests form one family. The word is used of those three and of any sentence naming its own correction and family size, which two appendix families do, and nowhere else; a released check permits exactly those forms. Two further families are corrected without being confirmatory, the $54$-cell budget sweep and the $440$ population interactions, and are called corrected rather than certified. Every remaining interval is exploratory, and every use of either word is listed against its family and adjusted $p$ in Appendix H.16. Model and configuration selection is on validation data, with two exceptions stated in Appendix B.

## A.2 WHAT A RULE BASE REACHES DEPENDS ON THE ENGINE THAT APPLIES IT

A paper that publishes a rule base usually publishes a figure for what it covers, the fraction of reference products applying every rule recovers: its *reach*. It is quoted as a property of the rules and is not one: applying a rule means deciding how the substrate is presented to the matcher, and every engine decides that silently and none reports it.

152 of SyGMa's 175 rules sit verbatim in ours, so those can be held fixed while only the procedure changes: one engine, one bank, the substrate expanded with explicit hydrogens or not. On the clean test split those two arms reach $0.200$ $[0.182, 0.218]$ and $0.390$ $[0.367, 0.413]$, a paired $+0.190$ $[+0.173, +0.208]$, and almost none of that is the convention. SyGMa's own program on the same rules reaches $0.387$, so the switch and not the reimplementation moves this. Expanding a substrate and never contracting the product leaves the drawn hydrogen on the reacting atom; contracting it lifts the expanded arm to $0.377$ $[0.354, 0.400]$, worth $+0.177$ $[+0.159, +0.195]$, and what survives against the unexpanded arm is $+0.013$ $[+0.007, +0.020]$. Thirteen fourteenths of it is an unfinished loop and the remainder the convention, both separated. The whole distance between the two configurations is $+0.339$: the engine term as the two programs stand, $+0.188$ $[+0.170, +0.205]$, $+0.145$ $[+0.128, +0.162]$ from the 23 rules our bank lacks, and $+0.0065$ $[+0.0038, +0.0096]$ from composing steps.

**The two parts fail in opposite ways.** The larger is not a matching effect: templates fire equally often expanded, and the toolkit then refuses $63\%$ of the fragments those products separate into against $0.4\%$ without the expansion, because a template written against implicit hydrogens does not name the hydrogen it replaces and the drawn one is never removed. Contract the product and that number is zero. The smaller is a matching effect and nothing else, the completed loop yielding $21{,}302$ fragments against the unexpanded arm's $21{,}106$. The convention then decides what a bank reaches without deciding which reaches further: with the loop complete BioTransformer recovers $1{,}289$ references with hydrogens drawn against $317$ with them implicit, a swing of $-0.374$, while SyGMa recovers $1{,}350$ against $1{,}390$, a swing of $+0.015$ the other way, and SyGMa leads in both. A coverage figure is the value of a bank together with a convention and is not comparable across papers that state only the bank.

The repair we registered for this works on one bank of three and on neither published one. Dispatch the convention per template rather than per run: a one-line syntactic test on each rule, whether its reactant side names a hydrogen *atom*, decides which presentation that rule sees, consulting no outcome. On SyGMa it is the identity, none of whose $175$ templates names one, which is the structural null the design needs. On BioTransformer, whose $726$ of $994$ templates make it the mixed bank by that syntax, it *loses* to the better single setting by $-0.0038$ $[-0.0067, -0.0015]$. It clears that setting only on our own bank, by $+0.0035$ $[+0.0015, +0.0058]$, whose $675$ dispatched templates were written by hand. A bank can be mixed in its syntax and single-convention in what it reaches, and the one case where dispatch pays is the bank whose provenance we control.

What is comparable without that is a bound, and it is what a reader of a coverage figure quoted without its procedure may assume: the least a bank recovers over the settings it might run under. For BioTransformer that is $0.1221$ against the $0.4963$ its best single setting reaches, four times what any one convention guarantees — and $0.4963$ is our measurement of that bank, not one its authors publish.

What a bank has to lose is readable from its rule file without running anything, and the hazard is not confined to this domain: every one of the $42{,}554$ patterns in the standard retrosynthesis template library pins a connection count, so the same step costs that library $82\%$ of the template–product pairs it can match at all (Appendix E.6).

### A.3 The decomposition and the coverage ceiling

The deep case study runs on GRAIL, a metabolite predictor built for this work: a learned selector chooses which of its 7,581 rules to apply, an engine applies them, and a learned filter ranks the result. A system of this shape is where a coverage figure can be measured exhaustively rather than sampled, because it reaches a reference only by passing three gates in sequence. With $C_{\text{full}}$ the references the whole bank recovers without a budget, $C_{\text{bud}}$ those surviving the selector and $H$ those returned,

$$\underbrace{\frac{|H|}{|U|}}_{\text{recall@}k} = \underbrace{\frac{|C_{\text{full}}|}{|U|}}_{\text{coverage}} \cdot \underbrace{\frac{|C_{\text{bud}}|}{|C_{\text{full}}|}}_{\text{selection}} \cdot \underbrace{\frac{|H|}{|C_{\text{bud}}|}}_{\text{ranking}}, \qquad |U| = \textstyle\sum_{s \in S} |Y_s|,$$

exact by construction because each set contains the next: an accounting device, not a model, in *micro* because only that form multiplies out. MetaPredictor factorises $0.612 \times 0.804 \times 1.000$, exactly its realised recall, though a beam bounds that decode rather than the model (Appendix E.3).

Table 2: Recall@15 by matching criterion on the full clean test split ($n = 1{,}170$). The final column is a paired mean difference, not the difference of the rounded columns. The same table for the external GLORYx set is Table 11.

| **method** | canon. | InChIKey (strict) | no-stereo (GLORYx) | Tanimoto=1 (fingerprint) | tautomer (ours) | gain (95% CI) |
|---|---|---|---|---|---|---|
| **GRAIL** | 0.323 | 0.331 | 0.333 | 0.324 | 0.340 | $+0.016$ $[0.010, 0.023]$ |
| MetaPredictor | 0.529 | 0.553 | 0.559 | 0.532 | 0.569 | $+0.040$ $[0.030, 0.050]$ |
| SyGMa | 0.532 | 0.558 | 0.563 | 0.533 | 0.571 | $+0.038$ $[0.029, 0.048]$ |

**The ceiling as a primitive.** Apply all 7,581 rules to every test substrate and ask what fraction of references appears anywhere in the union. That **coverage ceiling** is the best recall perfect selection and ranking would reach, here 0.817 (cluster-bootstrap 95% CI $[0.794, 0.839]$), macro $0.860$, and should be reported alongside recall for any rule-grounded method. The deployed pipeline converts $32.5\%$ of its own ceiling, and selection is where the loss sits, barely a third of what the bank reaches surviving the rule budget. Ranking is close to lossless at $0.960$ $[0.941, 0.976]$ and close to free: re-ranking the same pool by Tanimoto similarity to the substrate, nothing trained, reaches $0.342$ against the filter's $0.340$, $+0.002$ $[-0.004, +0.009]$, while reversing that order costs $0.040$ and shuffling it $0.015$. Applied to a second method the decomposition stops being self-diagnosis, since SyGMa decomposes with the inverse profile (Appendices E.7 and E.8).

### A.4 The same axis inside one benchmark, where the ceiling is measured

Published methods have their *frozen* predictions re-scored under all five criteria, so only the match rule changes across columns and never the chemistry. GRAIL, SyGMa (Ridder & Wagener, 2008) and MetaPredictor (Zhu et al., 2024) are compared on the whole clean test split, paired on identical substrates (Table 2); BioTransformer (Djoumbou-Feunang et al., 2019) and MetaTrans (Litsa et al., 2020) on the 150-substrate shared subset.

Methods are not equally sensitive to the convention. Over the full split GRAIL gains $0.016$ $[0.010, 0.023]$ from canonical to tautomer matching against MetaPredictor's $0.040$ $[0.030, 0.050]$, a differential of $+0.024$ $[0.014, 0.034]$, which survives Holm–Bonferroni over the three method pairs and over the thirty cells the rule declared for the other families, every pair by every criterion step, admits here. Unequal headroom runs against the effect: GRAIL enters lower, has more left to recover, and gains the less. Output size is a second and orthogonal confound, SyGMa emitting $81$ candidates per substrate against GRAIL's $8.4$ (Appendix J).

**Primary endpoint.** The endpoint is the interaction between criterion and method, by a paired bootstrap over substrates with $10{,}000$ resamples; every method gains under a more tolerant criterion, so only the differential can change an ordering. On the $150$-substrate subset two pairs separate, BioTransformer gaining +0.120 [+0.073, +0.171] more than GRAIL from canonical to tautomer matching and SyGMa +0.063 [+0.022, +0.106] more than MetaTrans from canonical to InChIKey, the first surviving correction and the second not. The criterion is a **method-dependent confounder** rather than a neutral scoring choice, and it moves human agreement too: two independent curations of twenty-five external drugs agree at a Jaccard of $0.145$ under strict INCHIKEY matching and $0.406$ under the tautomer-aware default, a third of what reads as disagreement about observed chemistry (Appendices H and H.16).

**External replication.** On the field's de-facto external hold-out, the 37 GLORYx drugs (de Bruyn Kops et al., 2020), the result rests on neither contaminated arm: 24 of the 37 sit in GRAIL's own train or validation split, so its row uses the 13 unseen. Under strict InChIKey matching MetaPredictor trails SyGMa by $0.130$ and one rung later the advantage is gone. The endpoint is the interaction: MetaPredictor gains $+0.135$ $[+0.072, +0.209]$ more than SyGMa across the stereo step, rejected at either family size and again by a sign-flip test. The criterion erases a significant advantage instead of reversing it, the relaxed gap covering zero while still admitting SyGMa ahead by $0.071$ (Appendix H.16).

### A.5 THE AGGREGATE AND THE BUDGET ARE ONE VARIABLE, AND THE DECOMPOSITION PREDICTS IT

Intervals here are exploratory except the budget sweep, whose family is declared below. A method emitting eighty candidates to catch a few more references is rewarded by recall and punished by F1. So on deployed output the two aggregates order the methods differently: against GRAIL the disagreement is separated both ways, SyGMa leading on recall by $0.253$ $[0.222, 0.283]$ and sitting $0.088$ $[0.076, 0.099]$ below on F1 (Appendix E).

The budget does the same work and it is what identifies the variable, because truncating a ranked list moves the emitted count with the method and its ranking held fixed, so a reordering along the sweep cannot be a property of any one system. Swept over $k$ in macro F1 the three methods take three distinct orderings: SyGMa sits above GRAIL at every budget up to 13, separated from $k = 1$ to $k = 10$ and by $0.071$ $[0.057, 0.086]$ at $k = 5$, and below from $14$ on, where the field's $15$ sits. A fourth method exhausts the field. On $291$ substrates carrying all four, MetaTox at $33.6$ candidates is last at $k \leq 10$ and first from $k = 15$; five of the six pairs change sign along that sweep and three do so between margins that each separate from zero, of which two survive Holm over the $54$-cell grid that axis admits. Its lead is volume, since on macro F1 at the same budget it is second at $0.125$ behind MetaPredictor's $0.156$ and last at $k = 1$ (Appendices H.14 and E).

Every corrected reversal along the sweep involves the one method that overproduces, and none survives correction among methods emitting alike, which is a failure to detect rather than a null: those pairs leave intervals up to $0.158$ open. The two choices act on different comparisons. The matching rule still separates methods at fixed output size, GRAIL's sensitivity sitting below both comparators by $-0.0195$ $[-0.029, -0.010]$ and $-0.025$ $[-0.035, -0.015]$ when all are matched at about eight emitted candidates (Appendix H.3); and where the budget reorders, the criterion instead erases, the GRAIL–MetaPredictor F1 gap being separated under four of the five criteria at $-0.026$ $[-0.038, -0.013]$ and indistinguishable from zero under plain INCHIKEY. The decomposition says why the aggregate splits them at all: SyGMa's selection factor is $1$ (§A.3), so a pipeline with no selection stage is scored higher by recall than by precision (Appendix E).

**A global budget is the wrong object.** It is undeclared, and it is also the wrong shape. Enzymes compete for one substrate, so its annotated products are alternatives from a shared pool: the median substrate here has one and the mean $2.22$, while the deployed policy emits $8.4$ for every substrate alike. Holding the ranking fixed and changing only how many are emitted, the deployed $k = 15$ scores $0.136$ macro F1, the best global constant $0.195$, and truncating each substrate at its own annotated count $0.214$. Emitting while a candidate scores within half the leader's score gives $0.209$, and the arm that decides the claim is not the deployed policy this paragraph has just discredited but the best constant, chosen with the same hindsight on the same split: $+0.0147$ $[+0.0059, +0.0237]$ of an available $+0.019$. Forecasting that count directly is worse than the constant it replaces, so what the pool supports is a relative rule. The advantage holds on the $845$ parent drugs, $+0.021$ $[+0.010, +0.032]$, and not on the $325$ substrates that are themselves metabolites, $-0.002$ $[-0.016, +0.013]$ (Appendices K and G).

## B ARCHITECTURE OF GRAIL

GRAIL predicts xenobiotic metabolite structures in three stages. A learned generator scores every rule in a bank of $7{,}581$ SMIRKS templates against the substrate, estimating the probability $P(r \mid s)$ that rule $r$ applies to substrate $s$. RDKit then applies each selected rule and enumerates the products, recording which rule produced which candidate. A structural filter scores each substrate–product pair, and the deployed system ranks candidates by the product of the filter and generator scores, combining a judgement about the plausibility of the pair with the generator's confidence in the rule that made it.

Both learned stages train on positive-unlabelled data. Annotated metabolites are the only positives; rule-applicable products that carry no annotation are unlabelled, not negative, because a missing annotation does not establish that a transformation does not occur. The filter exposes its raw classifier outputs so that a non-negative positive-unlabelled surrogate is never applied on top of a sigmoid when that estimator is selected, and the generator down-weights applicable but unobserved rules instead of penalising them as hard negatives.

Both stages share a featurisation: molecular graphs with 16-dimensional nodes and 18-dimensional edges, and a 1024-bit Morgan fingerprint per molecule. Unless stated otherwise, matching between predicted and reference structures uses the tautomer-aware InChIKey convention throughout this paper.

The three stages are separately inspectable: the selected rule, the enumerated product, and the filter's judgement of the pair. What the architecture contributes is interpretable rule selection paired with a positive-unlabelled filter, not recall above other metabolite predictors.

### B.1 THE GENERATOR

The generator scores the entire bank in one forward pass. A GATv2 encoder with three message-passing layers, node dimensions $16 \to 192 \to 256 \to 128$, batch normalisation, ReLU and dropout of 0.1 at each layer, and a mean-pooled readout, embeds the substrate graph into 128 dimensions. A parallel multilayer perceptron embeds its Morgan fingerprint, and a gated fusion combines the two into the substrate representation. Each rule is embedded independently by a rule-graph encoder of the same family, with hidden dimensions $128 \to 192 \to 128$, alongside a six-dimensional branch of global rule features. The full bank is re-encoded on every forward pass, which dominates the compute cost and is why inference time scales with bank size.

The logit for a rule sums four learned terms, each scaled by its own learned scalar. A cross-attention term lets the rule embedding query the substrate's atom states through single-head attention over node keys and values, producing a rule-specific local context that is concatenated with the pooled embeddings and passed through a three-layer perceptron. A global term is the cosine similarity between the substrate and rule embeddings, and a local term the cosine similarity between the attention context and the rule. A fourth term reflects whether the rule's SMARTS pattern matches. To this sum the model adds a per-rule bias and a frequency-prior term at weight 0.4, the log-odds that the rule yields a true metabolite as estimated from training statistics, then subtracts a penalty of 7.5 for rules whose SMARTS does not match, which effectively masks inapplicable rules before the sigmoid. Where a rule fires at several sites, the per-firing probabilities are combined by noisy-OR.

Training is multi-label rule classification with a margin-ranking auxiliary at weight 0.25 and margin 0.45. Rules that match the substrate but carry no annotation are down-weighted by half rather than treated as negatives, which is what makes the objective positive-unlabelled. Optimisation uses Adam at a learning rate of $10^{-4}$ with early stopping: the released preset budgets 20 epochs and the deployed generator trained for 8.

### B.2 THE FILTER

The deployed filter, which is the checkpoint behind every GRAIL number reported here, embeds the substrate and the candidate product independently. Each passes through its own three-layer GATv2 encoder over the molecule's graph, with dimensions $16 \to 192 \to 256 \to 128$ and a mean-pooled readout. The two 128-dimensional embeddings are concatenated with both Morgan fingerprints, and the resulting 2304-dimensional vector is mapped through a three-layer perceptron ($\to 128 \to 64 \to 1$, ReLU, dropout 0.1) to a single logit.

An alternative variant encodes one merged substrate–product graph, whose 18-dimensional nodes are joined by cross-edges placed according to an element-aware maximum-common-substructure correspondence between atoms, not by index order, so that chemically meaningful atom mappings are preserved across the reaction. We measured the two rather than assuming which should be the default, and report the comparison below.

The estimator is selectable: with the non-negative positive-unlabelled risk it consumes logits and with binary cross-entropy it consumes probabilities, and the released configuration trains the filter under the second. Optimisation uses Adam at a learning rate of $10^{-4}$ under the same budget, the deployed filter having trained for 8 epochs with early stopping at a patience of 7.

### B.3 CHOOSING BETWEEN THE TWO FILTERS

Both variants were trained on an identical 800-substrate subset, with the same negatives, seed and optimiser, differing only in which encoder they used, and evaluated on the full clean test split with the same generator and candidate pool.

Independent encoders win on accuracy. Recall@15 is 0.366 for the independent variant against 0.357 for the merged-graph variant, a paired difference of $-0.0090$ with a 95% paired-bootstrap interval of $[-0.017, -0.001]$ over $n = 1{,}170$ substrates; at recall@5 the margin widens, 0.282 against 0.259.

They also win on cost. Encoding a substrate and a product as one merged graph costs $3.75\times$ what encoding the two independently costs — $34.0$ ms against $9.1$ ms per pair on the released candidates — because it computes a maximum common substructure for every pair, and it pays that at every candidate at inference as well as at every training pair. Accuracy and cost point the same way. The merged-graph variant is implemented and available, and is not deployed because it loses on both counts.

### B.4 HOW THE RULE BANK WAS BUILT

The rule-based system GRAIL is compared against, SyGMa, uses hand-authored reaction rules. GRAIL's bank is mined automatically from annotated substrate–metabolite pairs, with each candidate template validated against the pair that produced it.

Mining reads only the clean, molecule-disjoint training split. For each annotated pair it first locates a ring-aware maximum common substructure, using RDKit's FMCS with ring atoms matched only to ring atoms, rings completed, any bond order allowed to match, and atoms compared by element; the match must cover at least 40% of the smaller molecule's atoms. From that common substructure it derives a reaction centre. Three kinds of atom are flagged: matched atoms whose element, formal charge, hydrogen count, aromaticity or degree changed; atom pairs whose connecting bond order changed; and leaving or entering atoms together with their matched neighbours. Several substrate–product match combinations are tried, and the one giving the smallest centre is kept. The centre is then expanded by one bond to capture its immediate reactive environment, and the expanded fragments are atom-mapped by their positional correspondence in the maximum common substructure into a candidate SMIRKS template.

Two gates follow. A template survives only if applying it to its own source substrate regenerates the annotated product after canonicalisation. The surviving, deduplicated templates then pass a selectivity filter that rejects templates producing more than 200 distinct products on average across a 50-substrate sample, and templates that are both very general and very prolific, applicable to more than 90% of sampled substrates while still averaging more than 50 products.

A second atom-mapping route is also implemented: it detects the reaction centre with RASCAL MCES, falling back to FMCS, expands by one bond as above, and emits a mapped SMIRKS using RXNMapper, a neural attention-based atom mapper. Both a rule-based and a neural mapping route are therefore available.

Because mining reads the training split alone, no test metabolite can enter a mined rule, so the coverage ceiling reported in the main text remains a held-out quantity.

### B.5 WHAT THE BANK CONTAINS

The deployed bank of $7{,}581$ templates is the deduplicated union of four earlier curated banks, of 474, 656, 500 and 1,051 rules, together with the newly mined templates.

Support across mined templates is heavily skewed. The positional miner alone extracts 5,856 unique self-tested templates from the training positives. Of these, 4,274 (73%) rest on a single training pair, and only 457 on five or more; the single best-supported template covers 1,531 pairs. The bank is therefore dominated by highly specific, near-single-example templates, not by broad general reactions.

Re-running the miner independently over the same training split re-derives templates that are all already in the bank: 5,856 of 5,856, none new. Mining is deterministic and, on this corpus, saturated

at its source. The coverage ceiling of $0.817$ is bounded by the diversity of the training reactions, not by incomplete mining, so extending coverage requires new reaction corpora or more general templates rather than further mining of the same data.

The selectivity filter is the construction-time counterpart of the breadth–precision trade measured in Appendix E. Rejecting over-general templates when the bank is built and narrowing the number of rules applied at inference act on the same quantity, one before deployment and one at the operating point.

## C Formal framework

The architecture, the coverage ceiling and the recall decomposition described in the main text follow from one compact generative model, which also gives the three propositions of Appendix E a common home. The model is a single likelihood together with an exact decomposition of recall and a map from each diagnostic to the factor it constrains.

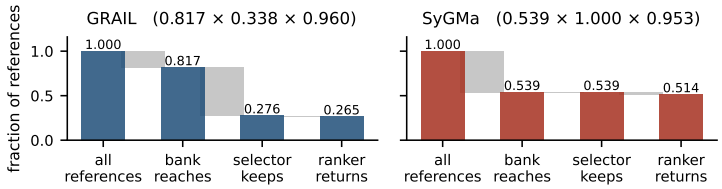



Figure 5: Coverage, selection and ranking on the clean test split ($n = 1{,}170$). Grey marks what each gate discards; each bar ends at the realised recall, in micro, which for GRAIL is $0.265$. The first bar of each waterfall is that method's reach through its *own* engine (§A.2).

**A latent-reaction mixture.** For a substrate $s$, a metabolite $m$ arises by choosing a transformation rule, choosing a site at which it fires, and applying it:

$$P(m \mid s) \;=\; \sum_{r} \sum_{\text{site} \in \text{sites}(r,s)} P(r \mid s) \cdot P(\text{site} \mid r, s) \cdot \mathbb{1}\big[\text{apply}(r, s, \text{site}) = m\big].$$

Application is deterministic, so the indicator collapses to the enumerable support of products. The latent variables are which rule fired and where; the observation, an annotated metabolite set, is positive-unlabelled, since a rule-applicable product without an annotation is unlabelled rather than a confirmed negative.

The three deployed stages approximate the marginal likelihood of this model. The generator approximates $P(r \mid s)$, and the frequency prior it stores is the marginal probability $\pi(r)$ that a rule fires. Rule application realises the deterministic support, enumerating the candidate products. The filter approximates a discriminative correction $P(\text{true} \mid s, m)$ over those candidates, and deployment ranks by the product of the filter and generator scores. Training under a positive-unlabelled objective approximates expectation-maximisation over the unobserved rule-firing indicator. This describes the model the pipeline approximates; it does not assert that the trained weights are its maximum-likelihood solution.

**An exact decomposition of recall.** Fix a cut-off $k$. For substrate $s$ let $Y_s$ be its references, and let the candidate sets nest, $R_s(k) \subseteq P_{\text{bud},s} \subseteq P_{\text{full},s}$, where $P_{\text{full},s}$ is the pool reachable by applying the whole bank once, $P_{\text{bud},s}$ the pool the generator's budget admits, and $R_s(k)$ the ranked output. Write $\text{hit}(A_s)$ for the number of references matched by a set under the tautomer-aware quotient, which is monotone under inclusion, and pool across substrates: $|U| = \sum_s |Y_s|$, $C_{\text{full}} = \sum_s \text{hit}(P_{\text{full},s})$, $C_{\text{bud}} = \sum_s \text{hit}(P_{\text{bud},s})$ and $H = \sum_s \text{hit}(R_s(k))$. Then

$$\text{recall@}k = \frac{H}{U} = \left(\frac{C_{\text{full}}}{U}\right) \cdot \left(\frac{C_{\text{bud}}}{C_{\text{full}}}\right) \cdot \left(\frac{H}{C_{\text{bud}}}\right) = \text{coverage} \cdot \text{selection} \cdot \text{ranking}.$$

Because the sets nest, every factor lies in $[0, 1]$ and the three multiply to the realised recall on any numbers. This is an accounting identity, not a theorem, and its use is to localise where recall is lost. The coverage factor coincides with the rule-bank ceiling reported in the main text.

**Which measurement constrains which factor.** Every diagnostic in Appendix E, and each of the three propositions, constrains exactly one factor.

Table 3: The factor of the recall decomposition that each diagnostic and proposition constrains.

| Factor | Diagnostics and propositions |
|---|---|
| coverage | multi-step application at depth 2, the composition of the coverage gap, the similarity-to-training stratification, the composition covariate behind the internal–external difference, Proposition 3 |
| selection | the comparison of learned selection against the frequency prior, saturation with training-set size, the propensity-scored selection loss, Proposition 2 |
| ranking | the filter and the listwise reranker, the site-of-metabolism prior, abstention on filter score, the oracle bound, Proposition 1 |

## D THE ROBUST ORDER: DEFINITIONS, PROPOSITIONS AND THE TABLES BEHIND THE SHARE

Fix a set of systems $S$, a set of items, and a grid $C$ of evaluation cells. In cell $c$ each system has a mean score, and those means induce a total preorder $\prec_c$ on $S$: exact ties occur, in four cells of the grids reported here, and a tie is not a reversal. A published table reports $\prec_{c_0}$ for one cell $c_0$ that its paper does not name. Write

$$\prec_* = \bigcap_{c \in C} \prec_c,$$

and say $a$ *dominates* $b$ when $a \prec_* b$. Say the pair is *separated in every cell* when in addition the paired bootstrap interval on the margin excludes zero in each $c$.

**Proposition 1.** $\prec_*$ is a strict partial order, and it is the finest relation on $S$ that no cell of $C$ can contradict.

*Proof.* An intersection of transitive, irreflexive relations is transitive and irreflexive. Any relation not contained in some $\prec_c$ contains a pair that cell orders the other way. □

**Proposition 2.** If $C \subseteq C'$ then $\prec_*^{C'} \subseteq \prec_*^{C}$: the dominating set is monotone non-increasing in the grid, and so is the number of tiers.

*Proof.* Immediate from the definition as an intersection; the tier count is the longest chain of a smaller relation. □

Proposition 2 is the reason a single share is not a property of a leaderboard. It is a property of the pair (leaderboard, grid), it can only fall as a grid is widened, and any paper reporting one number could have reported a larger one by declaring fewer choices. The remedy used here is to report the curve rather than a point: each declared axis alone, and their product. The tables below give that curve for every leaderboard in the paper, together with the accuracy of every system in every cell, which is the table all the derived quantities are computed from.

**A dual statistic.** The tier count says what survives; the number of distinct total orders the same systems take across the grid says how much varies. Neither is recoverable from the other: a leaderboard can take many orderings and still support several tiers if the systems that move are adjacent, and it can take few and support one if the pairs that never move are also never resolved.

**The grid is the same shape and not the same cells.** The criterion axis is four identity relations in every domain: which strings name the same compound in the prediction domains, and which poses count as the crystal one in docking. The second axis is not transferable. Retrosynthesis tables are read at top-1 through top-10 and metabolite work at fifteen, so each leaderboard is swept over the budgets its own field reports at; docking reports a single pose and has no budget, and the choice its

papers leave undeclared instead is whether that pose is energy-minimised before it is scored. Fixing one set of budgets across domains would have measured robustness to a budget nobody uses. What this costs is that the four shares are comparable in construction rather than in cells, which is why they are reported as four numbers and not averaged. One consequence is worth stating because it bounds how much the grid can find: on the seven-system board exactly one of the 120 pairs of cells gives identical hit vectors for every system, canonical and InChIKey at top-1, so the sixteen cells are sixteen and not fewer.

**Which leaderboards the grid applies to, decided before the shares were read.** A grid cell has to be a choice the evaluation could have been made under. Two conditions decide it, and both exclude cells rather than admit them. First, a matching criterion has to be an identity relation: TANIMOTO1 is reported in Table 2, because published work compares Morgan fingerprints, but it is not a grid cell, since at radius 2 over 2048 bits it returns 1 for decane against undecane and for D- against L-alanine, and a rule that identifies a homologue with its reference is not a candidate answer to whether two structures are the same compound. Second, the metric must be defined independently of the criterion. That is what excludes the molecular-generation set: its headline quantity counts *distinct* canonical strings, so the criterion is not a parameter of the evaluation but part of the definition of what is being counted, and re-scoring it under another criterion changes the metric rather than the convention. The generation results are therefore reported in Appendix H as a criterion sensitivity and not as a row here. Both rules were fixed before the shares below were read, and the second excludes the one leaderboard whose pairs stand furthest apart relative to the movement a change of criterion produces on it, which would have raised the shares. Its median gap is the smallest of any board at 0.0094 and its median movement smaller still at 0.0001, a ratio of 94 against 43 for the next board and under 20 for every other. The exclusion is therefore self-penalising in the sense that matters: the criterion has almost no purchase there, so nearly every pair would have survived.

**A leaderboard from outside chemistry, whose convention ships as a function.** The two translation rows are the same instrument on a domain sharing none of this paper's data: the ranking is human, the items are sentences, and the systems are machine-translation models. WMT's general task publishes an expert annotation of every error in every segment, its category and its severity, and turns those annotations into a number. The turning is the convention, and it is a function in the task's own repository. It makes four decisions a paper quoting an MQM score does not repeat: an error in the source text is charged nothing, a creative reinterpretation is charged nothing, a minor punctuation error is charged a tenth of an ordinary minor, and a non-translation is charged five times a major. So the criterion axis is which annotations are charged (membership, as a structure-matching rule is, with the metric fixed as the mean of the per-segment sums), and the second axis is whether that sum is floored at the non-translation weight, an MQM score being otherwise unbounded below.

The published cell is checked rather than assumed. The task does not rank by the mean over segments: it takes the mean within each domain and then the mean of those, which is what its own clustering script does by default. The cell is weighted so that its mean is that average, and both boards then reproduce the published ranking system for system. That agreement is a unit test on the weighting and not evidence about the data: the function that checks it rebuilds the same average from the same transcribed weights, and two routes to one quantity must agree. What it does establish is that the cell is the aggregation the task documents, which a reader can check against that script, as they can check the transcription of the weights, which this run does not.

Reading the published number as the mean of the column instead is a choice of exactly the kind the grid varies, and on this board it costs more than either axis in it. Under that reading the nineteen-system leaderboard has 135 dominating pairs and no certified reversal, against 125 and six under the task's own aggregation, and its published cell no longer reproduces the official ranking: nine systems change place. On the fifteen-system board the same reading changes almost nothing: 97 dominating pairs against 96, and no certified reversal either way. So the cost of getting the aggregation wrong is a fact about the board rather than a constant. Both readings are run on both boards and both are kept, so the size of the choice is measured here rather than asserted.

What the grid reports, on the aggregation the task documents, is this. On 171 pairs of 19 systems, 125 dominate, 29 are unresolved and 17 are contested of which 6 survive Holm over the 1,368 cell-level tests that grid creates, so nineteen published places support 9. On 105 pairs of 15 systems, 96 dominate, 8 are unresolved and 1 is contested, which does not survive the 840 tests of its own, so

fifteen places support 8. The two boards differ in kind and not by accident: the larger one carries a reversal past both its own 1,368 cell-level tests, and the smaller one supplies a null. The same rule fixes the cell on every board here, including the ones where it changes nothing.

**Nine more boards from the same task, and why they are nine and not one.** The same task annotated nine further language pairs under a lighter protocol, ESA, in which one expert marks the error spans in a segment and then scores it out of a hundred. The systems entered differ per pair, the annotators differ per pair, and the pairs were run independently, so these are nine leaderboards rather than one leaderboard nine times. That is the condition for reading a share as a distribution instead of as a number that happened to come out of one table. Both axes are again choices left to whoever scores: what the annotation is read as, the number out of a hundred or the marked spans as a tally, and whether a score is used as the annotator wrote it or standardised within that annotator first. The second is the field's own oldest open question in evaluation practice, and a comparison quoting human scores almost never says which it used.

The published cell is the same aggregation as on the MQM boards, the mean within a domain and then across domains, and one thing is simpler here: the number the task publishes is the annotator's score itself, so there is no weighting function to transcribe and the only convention the cell carries is the aggregation. As there, no evidence is claimed from any agreement with the published ranking.

Across the nine, 553 of 689 pairs dominate, 33 are contested of which 4 survive correction, and 115 published places support 62. The share runs from 0.62 to 0.91 with a median of 0.82: narrower than the spread across domains, which is what one would expect of nine boards from one task, and still not a constant.

**The share does not replicate across editions, and that is the strongest form of the claim.** Nine boards from one campaign, annotated under one set of instructions by one pool, cannot tell a property of a leaderboard from a property of a year. The previous edition of the same task annotated eight language pairs under a different protocol: direct assessments, no error spans, several ratings per segment, and between 82 and 174 annotators. Five of its pairs are pairs the later edition also scored. That is the one place in this paper where the same question is put twice to independent data.

The answer is that the share does not carry over. Across the five repeated pairs it moves by 0.114 on average and by as much as 0.285; EN-ZH is at 0.900 in one edition and 0.615 in the next. Nothing about the language, the domain or the difficulty changed enough to explain that, and the two campaigns were run by the same organisers on comparable data. What changed is the protocol and the pool, which is to say the leaderboard. The pools cannot be shown to be disjoint, because each campaign anonymises its annotators into its own identifier space, so the two share no identifier by construction and a rater who worked on both would be invisible. The claim made here is only that the campaigns are separate, which is what the editions are. So a share is a fact about a table and its grid, and quoting one as a benchmark's fingerprint, ours included, would be an error of the same shape as the one this paper is about. These eight boards are built on the release put through the preprocessing its own README documents: a quality-control filter that discards a rater's whole document when a planted item in it was failed, and the removal of document-level meta scores. Reading the file without it inflates every board by between a ninth and two fifths. With the filter the published cell reproduces the task's own scores to $3 \times 10^{-14}$ and orders all 909 pairs the way its published table does.

Across the eight, 804 of 909 pairs dominate, 20 are contested of which 1 survive correction, and 124 published places support 78.

**One correction per board, when the claim ranges over twenty-three.** Holm is applied within each grid, which controls the error rate of a claim about *that* leaderboard and not of the claim this paper makes. The honest family is the union of the grids: 15,752 cell-level tests. Re-applying the same procedure to the concatenated $p$-vectors, with no board re-analysed and no prediction re-scored, rejects 8,367 of them at a cutoff of $p = 6.72 \times 10^{-6}$, and the count of certified pairs falls from 23 to 15. Splitting $\alpha$ across the boards instead, at $\alpha/24$ within each grid, gives 15 as well. What the second claim is about very nearly does not move: of the 6 certified reversals of an ordering a published cell had separated, 5 survive every one of the three, and the matching rule certifies none under any of them. The sixth is the extrapolation board's, which its own grid certifies and the union does not, and it is the only pair either stricter reading removes that its board had established. That is checked under

each rather than inferred from the others, because the three are not nested: splitting $\alpha$ across the boards and pooling their grids into one family weight a board by different things, its own size against the total, and on 16 of the 24 the union is the more permissive of the two stricter readings, because its pooled cutoff is the larger one there.

What the union costs is concentrated rather than spread: the nineteen-system translation board keeps 2 of 6, two more go on EN-HI, one on ZHO-ENG and one on the extrapolation board. Those four boards are the whole fall of eight, and every pair that goes but one is the reversal of an ordering its own board had never established; the exception is named above. What removes them is the size of the family and nothing subtler: every threshold tightens by roughly the factor the family grew by, and each lost reversal misses its own rank threshold outright. A step-down procedure can also lose a small-$p$ test to an earlier failure, and on this vector it does not, since no test beyond the cutoff would have passed its own threshold. So the losses cannot be charged to the stopping rule.

**A test that assumes nothing about the shape.** Where a board scores an item a hit or a miss, the paired difference takes three values and the $p$ a correction reads is the exact conditional test on the discordant items. The translation boards are not like that: a segment score is continuous, zero-inflated and long-tailed on one side, and their $p$ is the normal approximation to a paired mean, read at thresholds between $10^{-6}$ and $10^{-3}$, which is doing work no central limit theorem promises on a few hundred items with one catastrophic output in them.

Under exchangeability there is a test that assumes nothing: flipping the sign of each paired difference leaves the null distribution alone, so the sign-flip distribution of the mean is exact by construction. Its tail is the hard part, and a resample cannot reach it: at $10^{-5}$ the draw count that would resolve it is the draw count that cannot be run. So the tail is taken from the saddlepoint instead, which is uniformly accurate far out. Computed for every certified reversal and every test within a factor of ten of its board's cutoff, 42 tests of which 25 are certified, it never returns a larger $p$ than the analytic one: the ratio runs from $0.0166$ to $0.9967$, so on these boards the approximation is the conservative of the two and not, as on a hit board, the anti-conservative one. No certified reversal is lost to it, and one further pair on the nineteen-system board would be certified under it. On the 13 tests whose tail a resample can resolve at all, the saddlepoint and $200{,}000$ draws agree to within the resample's own resolution there.

**The task's own answer to the same question.** A count no one else computes is a count no one else can check, and the objection that the instrument is idiosyncratic is answered by the one domain that computes its own version and publishes it. The translation task tests every pair with a Wilcoxon signed-rank test on the paired per-segment scores within each domain, combines the domains by Stouffer's $Z$, and merges systems into a cluster whenever that test does not separate them at $0.05$; the number of clusters is its answer to how many ranks its table supports. Their code is imported rather than reimplemented (`get_pvalues` and `get_clusters` as written) and run on all 11 boards it covers, the two MQM ones and the nine ESA pairs.

Setting the two counts side by side settles nothing, because they differ in two places at once. The test that decides whether a pair is separated is one; the construction that turns pairwise verdicts into a count is the other, and theirs walks the score-sorted table and closes a cluster only when nothing at or above the current position is tied with anything below it. A cluster is therefore an interval of the published order and can contain pairs its own test separates, while ours is the longest chain in the separation relation, which on identical verdicts is never shorter. So the two are crossed instead:

| | nineteen systems | fifteen systems |
|---|---|---|
| their test, their construction | 4 | 6 |
| our test, their construction | 3 | 5 |
| their test, our construction | 11 | 9 |
| our test, our construction | 6 | 8 |

Read down the first two rows: holding the construction fixed, our test is the stricter of the two, resolving one place fewer on each board. Read the next two: holding the test fixed, our construction is the more generous, because a longest chain counts a place whenever some pair is separated while a cluster refuses to close over an unresolved tie anywhere below it. Across all eleven boards neither direction ever reverses: our test is stricter on 10 and never more permissive, and our construction is

more generous on $10$ and never less. The number this paper reports is the last row, and the difference between it and the task's own is those two effects pulling against each other rather than either one alone.

The two rankings themselves agree system for system on $7$ of the $11$. On the other four they disagree about $6$ pairs in total, and our published cell separates none of them. So where the instruments order systems differently, they are ordering systems neither can tell apart, which is this paper's first claim arriving inside its own external check. Neither side corrects for multiplicity in any of this, so both counts are the optimistic ones, which is what makes them comparable at all.

**A null with a size.** The paper's second claim is negative: the matching rule certifies no reversal of an ordering a published cell had separated. A negative claim is worth whatever its power is. The identity supplies that without another statistic. A criterion reverses a pair exactly when it erases the whole margin between them and more, so the share of the published margin it erases, $1 - d'/d$, says how close it came, and it is read straight out of the per-cell means each board already reports. The sign matters: a criterion that moves a margin *away* from zero has not come close to reversing it, and a direction-blind ratio scores such a comparison as if it had.

Taken over the pairs the published cell separates, the only pairs with an established ordering to reverse, and with the other axis held at the board's own setting, the matching rule erases at most $0.0870$ of a margin on the seven-system retrosynthesis board, $0.5755$ on the three-system one and $0.0857$ on the metabolite board, and on all three the median comparison erases $0.0$: half the time the criterion leaves the margin exactly where the published cell put it. On the closest of the three the movement would have to be $1.74$ times larger before an ordering flipped at all. The null is a size, then, and not a silence: an axis whose movement is well inside the margins it would have to cross, rather than a test that failed to reach significance. What the axis does instead is exchange pairs the benchmark never separated, which is this paper's first claim and is measured separately.

Docking is the exception, and is reported as one. There the criterion axis does erase a margin the published cell had separated and then some: DEEPDOCK over UNIMOL, $2.8357$ of it, at the published post-processing. So on that board the matching rule is capable of the reversal the null denies elsewhere. What it is not capable of is surviving the family it was read in: that test reaches $p = 0.053$ against the board's threshold of $9.8 \times 10^{-4}$, and no reversal on that board survives, which is why its certified column reads zero. The one field that published its own reversal is where one would expect the axis to bite, and it bites hard enough to move an established margin and not hard enough to be certified.

**What the share does not measure.** It is not an error rate and no cell is privileged. The instrument takes the grid as declared and says what part of a published ordering is invariant to it; whether a given cell is one a field would defend is a question about the field, argued in the text and not settled by the instrument. Nor is the interval on a share a statement about the grid: it resamples items with the grid held fixed, so it reports how much of the share is the sample. The grid's contribution is what Proposition 2 and the curve below report.

**The share as a function of the grid.** Domination is an intersection over cells, so adding a cell can only remove pairs from it and the share is a property of the pair (leaderboard, grid). Each declared axis is therefore reported alone and against the product.

| leaderboard | sub-grid | cells | dominate | tiers | orderings |
|---|---|---|---|---|---|
| retrosynthesis, 7 systems | criteria only, at the published budget | 4 | 16 of 21 | 5 | 2 |
| | budgets only, at the published criterion | 4 | 12 of 21 | 4 | 3 |
| | the product | 16 | 8 of 21 | 3 | 4 |
| retrosynthesis, 3 systems | criteria only, at the published budget | 4 | 3 of 3 | 3 | 1 |
| | budgets only, at the published criterion | 4 | 0 of 3 | 1 | 3 |
| | the product | 16 | 0 of 3 | 1 | 3 |
| retrosynthesis, 5 systems | criteria only, at the published budget | 4 | 9 of 10 | 4 | 1 |
| | budgets only, at the published criterion | 4 | 7 of 10 | 3 | 3 |
| | the product | 16 | 7 of 10 | 3 | 3 |
| metabolites, 3 methods | criteria only, at the published budget | 4 | 3 of 3 | 3 | 1 |
| | budgets only, at the published criterion | 4 | 2 of 3 | 2 | 2 |
| | the product | 16 | 2 of 3 | 2 | 2 |
| translation, WMT24 en-de, 19 systems | criteria only, unfloored | 4 | 131 of 171 | 9 | 4 |
| | the floor only, at the published criterion | 2 | 153 of 171 | 14 | 2 |
| | the product | 8 | 125 of 171 | 9 | 6 |
| translation, WMT24 ja-zh, 15 systems | criteria only, unfloored | 4 | 96 of 105 | 8 | 3 |
| | the floor only, at the published criterion | 2 | 101 of 105 | 11 | 2 |
| | the product | 8 | 96 of 105 | 8 | 5 |
| docking, 7 programs | criteria only, at the published post-processing | 4 | 18 of 21 | 5 | 3 |
| | post-processing only, at the published criterion | 2 | 18 of 21 | 5 | 2 |
| | the product | 8 | 17 of 21 | 4 | 5 |

**retrosynthesis, 7 systems.** Published order at ('canonical', 1): graphretro > localretro > retroformer > graph2smiles > gln > g2retro > retroxpert. 4999 items. Accuracy by system and cell, from which every quantity below follows:

| system | canon@1 | canon@3 | canon@5 | canon@10 | no-st@1 | no-st@3 | no-st@5 | no-st@10 |
|---|---|---|---|---|---|---|---|---|
| graphretro | 0.538 | 0.677 | 0.716 | 0.745 | 0.550 | 0.693 | 0.733 | 0.762 |
| localretro | 0.534 | 0.769 | 0.842 | 0.910 | 0.547 | 0.788 | 0.862 | 0.930 |
| retroformer | 0.532 | 0.706 | 0.761 | 0.819 | 0.554 | 0.722 | 0.776 | 0.832 |
| graph2smiles | 0.529 | 0.661 | 0.695 | 0.724 | 0.557 | 0.679 | 0.711 | 0.738 |
| gln | 0.525 | 0.690 | 0.756 | 0.837 | 0.538 | 0.702 | 0.768 | 0.849 |
| g2retro | 0.515 | 0.722 | 0.783 | 0.837 | 0.529 | 0.741 | 0.803 | 0.858 |
| retroxpert | 0.447 | 0.598 | 0.644 | 0.694 | 0.462 | 0.610 | 0.655 | 0.703 |

| system | IK@1 | IK@3 | IK@5 | IK@10 | taut@1 | taut@3 | taut@5 | taut@10 |
|---|---|---|---|---|---|---|---|---|
| graphretro | 0.538 | 0.677 | 0.716 | 0.745 | 0.555 | 0.701 | 0.741 | 0.770 |
| localretro | 0.534 | 0.769 | 0.842 | 0.910 | 0.549 | 0.791 | 0.865 | 0.932 |
| retroformer | 0.532 | 0.706 | 0.762 | 0.819 | 0.556 | 0.724 | 0.778 | 0.834 |
| graph2smiles | 0.529 | 0.661 | 0.695 | 0.724 | 0.559 | 0.682 | 0.714 | 0.740 |
| gln | 0.525 | 0.690 | 0.756 | 0.837 | 0.540 | 0.705 | 0.771 | 0.852 |
| g2retro | 0.515 | 0.722 | 0.783 | 0.837 | 0.531 | 0.743 | 0.806 | 0.862 |
| retroxpert | 0.447 | 0.598 | 0.644 | 0.694 | 0.463 | 0.612 | 0.657 | 0.705 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| gln over g2retro | contested | no | 12 | 8 | no |
| graph2smiles over g2retro | contested | no | 12 | 12 | yes |
| graphretro over g2retro | contested | no | 12 | 12 | yes |
| localretro over g2retro | dominates | yes | 0 | 0 | yes |
| retroformer over g2retro | contested | no | 12 | 12 | yes |
| g2retro over retroxpert | dominates | yes | 0 | 0 | yes |
| graph2smiles over gln | contested | no | 12 | 12 | no |
| graphretro over gln | contested | no | 12 | 8 | no |
| localretro over gln | dominates | no | 0 | 0 | no |
| retroformer over gln | contested | no | 4 | 4 | no |
| gln over retroxpert | dominates | yes | 0 | 0 | yes |
| graphretro over graph2smiles | unresolved | no | 2 | 0 | no |
| localretro over graph2smiles | unresolved | no | 2 | 0 | no |
| retroformer over graph2smiles | unresolved | no | 2 | 0 | no |
| graph2smiles over retroxpert | dominates | yes | 0 | 0 | yes |
| graphretro over localretro | contested | no | 12 | 12 | no |
| graphretro over retroformer | contested | no | 14 | 12 | no |
| graphretro over retroxpert | dominates | yes | 0 | 0 | yes |
| localretro over retroformer | unresolved | no | 2 | 0 | no |
| localretro over retroxpert | dominates | yes | 0 | 0 | yes |
| retroformer over retroxpert | dominates | yes | 0 | 0 | yes |

**retrosynthesis, 3 systems.** Published order at ('canonical', 1): chemformer > gta > tiedtransformer. 4982 items. Accuracy by system and cell, from which every quantity below follows:

| system | canon@1 | canon@3 | canon@5 | canon@10 | no-st@1 | no-st@3 | no-st@5 | no-st@10 |
|---|---|---|---|---|---|---|---|---|
| chemformer | 0.539 | 0.609 | 0.618 | 0.624 | 0.550 | 0.621 | 0.630 | 0.636 |
| gta | 0.518 | 0.678 | 0.748 | 0.818 | 0.540 | 0.697 | 0.767 | 0.832 |
| tiedtransformer | 0.471 | 0.679 | 0.743 | 0.798 | 0.481 | 0.693 | 0.756 | 0.811 |

| system | IK@1 | IK@3 | IK@5 | IK@10 | taut@1 | taut@3 | taut@5 | taut@10 |
|---|---|---|---|---|---|---|---|---|
| chemformer | 0.539 | 0.609 | 0.618 | 0.624 | 0.553 | 0.624 | 0.634 | 0.639 |
| gta | 0.518 | 0.678 | 0.749 | 0.818 | 0.544 | 0.702 | 0.771 | 0.835 |
| tiedtransformer | 0.471 | 0.679 | 0.743 | 0.798 | 0.484 | 0.697 | 0.761 | 0.815 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| chemformer over gta | contested | no | 12 | 12 | yes |
| chemformer over tiedtransformer | contested | no | 12 | 12 | yes |
| gta over tiedtransformer | unresolved | no | 2 | 0 | yes |

**retrosynthesis, 5 systems.** Published order at ('canonical', 1): Chemformer > GraphRetro > Molecular Transformer > LocalRetro > MEGAN. 4948 items. Accuracy by system and cell, from which every quantity below follows:

| system | canon@1 | canon@3 | canon@5 | canon@10 | no-st@1 | no-st@3 | no-st@5 | no-st@10 |
|---|---|---|---|---|---|---|---|---|
| Chemformer | 0.020 | 0.038 | 0.046 | 0.051 | 0.020 | 0.038 | 0.046 | 0.052 |
| GraphRetro | 0.017 | 0.031 | 0.042 | 0.058 | 0.017 | 0.031 | 0.043 | 0.059 |
| Molecular Transformer | 0.015 | 0.029 | 0.036 | 0.037 | 0.015 | 0.029 | 0.036 | 0.037 |
| LocalRetro | 0.009 | 0.018 | 0.023 | 0.029 | 0.009 | 0.018 | 0.023 | 0.030 |
| MEGAN | 0.009 | 0.022 | 0.030 | 0.043 | 0.009 | 0.022 | 0.030 | 0.043 |

| system | IK@1 | IK@3 | IK@5 | IK@10 | taut@1 | taut@3 | taut@5 | taut@10 |
|---|---|---|---|---|---|---|---|---|
| Chemformer | 0.020 | 0.038 | 0.046 | 0.051 | 0.022 | 0.043 | 0.051 | 0.057 |
| GraphRetro | 0.017 | 0.031 | 0.042 | 0.058 | 0.019 | 0.036 | 0.050 | 0.068 |
| Molecular Transformer | 0.015 | 0.029 | 0.036 | 0.037 | 0.016 | 0.034 | 0.042 | 0.043 |
| LocalRetro | 0.009 | 0.018 | 0.023 | 0.029 | 0.011 | 0.025 | 0.031 | 0.042 |
| MEGAN | 0.009 | 0.022 | 0.030 | 0.043 | 0.011 | 0.028 | 0.039 | 0.055 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Chemformer over LocalRetro | dominates | yes | 0 | 0 | yes |
| Chemformer over MEGAN | dominates | no | 0 | 0 | yes |
| Chemformer over GraphRetro | contested | no | 4 | 2 | no |
| Chemformer over Molecular Transformer | dominates | yes | 0 | 0 | yes |
| LocalRetro over MEGAN | contested | no | 13 | 11 | no |
| GraphRetro over LocalRetro | dominates | yes | 0 | 0 | yes |
| Molecular Transformer over LocalRetro | dominates | no | 0 | 0 | yes |
| GraphRetro over MEGAN | dominates | yes | 0 | 0 | yes |
| Molecular Transformer over MEGAN | contested | no | 4 | 4 | yes |
| GraphRetro over Molecular Transformer | dominates | no | 0 | 0 | no |

**metabolites, 3 methods.** Published order at ('tautomer', 15): MetaPredictor > SyGMa > GRAIL. 1170 items. Accuracy by system and cell, from which every quantity below follows:

| system | canon@1 | canon@5 | canon@10 | canon@15 | IK@1 | IK@5 | IK@10 | IK@15 |
|---|---|---|---|---|---|---|---|---|
| MetaPredictor | 0.332 | 0.562 | 0.649 | 0.672 | 0.342 | 0.585 | 0.673 | 0.697 |
| SyGMa | 0.313 | 0.591 | 0.642 | 0.658 | 0.334 | 0.615 | 0.668 | 0.685 |
| GRAIL | 0.248 | 0.403 | 0.431 | 0.439 | 0.253 | 0.411 | 0.440 | 0.449 |

| system | no-st@1 | no-st@5 | no-st@10 | no-st@15 | taut@1 | taut@5 | taut@10 | taut@15 |
|---|---|---|---|---|---|---|---|---|
| MetaPredictor | 0.353 | 0.591 | 0.677 | 0.702 | 0.357 | 0.602 | 0.688 | 0.711 |
| SyGMa | 0.334 | 0.615 | 0.669 | 0.688 | 0.340 | 0.626 | 0.681 | 0.698 |
| GRAIL | 0.255 | 0.413 | 0.442 | 0.450 | 0.259 | 0.420 | 0.450 | 0.459 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| MetaPredictor over GRAIL | dominates | yes | 0 | 0 | yes |
| SyGMa over GRAIL | dominates | yes | 0 | 0 | yes |
| MetaPredictor over SyGMa | contested | no | 4 | 2 | no |

**translation, WMT24 en-de, 19 systems.** Published order at ('wmt', 'none'): GPT-4 > Dubformer > ONLINE-B > TranssionMT > Unbabel-Tower70B > refB > Mistral-Large > CommandR-plus > ONLINE-W > Claude-3.5 > refA > IOL-Research > Gemini-1.5-Pro > Aya23 > ONLINE-A > Llama3-70B > IKUN > IKUN-C > MSLC. 486 items. Accuracy by system and cell, from which every quantity below follows:

| system | wmt none | wmt clipped | plain none | plain clipped |
|---|---|---|---|---|
| GPT-4 | -1.641 | -1.641 | -2.503 | -2.422 |
| Dubformer | -1.814 | -1.812 | -2.812 | -2.718 |
| ONLINE-B | -1.899 | -1.888 | -2.407 | -2.279 |
| TranssionMT | -1.906 | -1.906 | -2.509 | -2.388 |
| Unbabel-Tower70B | -1.908 | -1.883 | -2.498 | -2.297 |
| refB | -1.977 | -1.870 | -2.634 | -2.342 |
| Mistral-Large | -2.110 | -2.110 | -2.794 | -2.726 |
| CommandR-plus | -2.288 | -2.265 | -3.116 | -2.965 |
| ONLINE-W | -2.321 | -2.287 | -3.200 | -3.042 |
| Claude-3.5 | -2.350 | -1.852 | -2.933 | -2.555 |
| refA | -2.461 | -2.319 | -3.172 | -2.939 |
| IOL-Research | -2.544 | -2.499 | -3.319 | -3.140 |
| Gemini-1.5-Pro | -2.775 | -1.978 | -2.268 | -2.188 |
| Aya23 | -3.214 | -3.198 | -3.852 | -3.734 |
| ONLINE-A | -3.488 | -3.419 | -4.396 | -4.147 |
| Llama3-70B | -4.258 | -4.110 | -5.149 | -4.805 |
| IKUN | -4.320 | -4.109 | -5.072 | -4.622 |
| IKUN-C | -6.059 | -5.827 | -6.549 | -6.148 |
| MSLC | -15.468 | -12.289 | -15.651 | -12.265 |

| system | major-only none | major-only clipped | count none | count clipped |
|---|---|---|---|---|
| GPT-4 | -0.267 | -0.267 | -1.436 | -1.436 |
| Dubformer | -0.341 | -0.341 | -1.448 | -1.446 |
| ONLINE-B | -0.344 | -0.344 | -1.030 | -1.030 |
| TranssionMT | -0.365 | -0.365 | -1.049 | -1.049 |
| Unbabel-Tower70B | -0.335 | -0.335 | -1.158 | -1.158 |
| refB | -0.354 | -0.354 | -1.218 | -1.203 |
| Mistral-Large | -0.379 | -0.379 | -1.280 | -1.280 |
| CommandR-plus | -0.400 | -0.400 | -1.517 | -1.517 |
| ONLINE-W | -0.388 | -0.388 | -1.649 | -1.649 |
| Claude-3.5 | -0.364 | -0.364 | -1.476 | -1.476 |
| refA | -0.438 | -0.438 | -1.419 | -1.403 |
| IOL-Research | -0.420 | -0.420 | -1.637 | -1.637 |
| Gemini-1.5-Pro | -0.305 | -0.305 | -1.046 | -1.046 |
| Aya23 | -0.559 | -0.559 | -1.616 | -1.616 |
| ONLINE-A | -0.609 | -0.609 | -1.959 | -1.959 |
| Llama3-70B | -0.757 | -0.757 | -2.121 | -2.107 |
| IKUN | -0.772 | -0.772 | -1.984 | -1.984 |
| IKUN-C | -1.053 | -1.053 | -2.335 | -2.335 |
| MSLC | -2.863 | -2.863 | -4.199 | -4.189 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Claude-3.5 over Aya23 | dominates | no | 0 | 0 | yes |
| CommandR-plus over Aya23 | dominates | no | 0 | 0 | yes |
| Dubformer over Aya23 | dominates | no | 0 | 0 | yes |
| GPT-4 over Aya23 | dominates | no | 0 | 0 | yes |
| Gemini-1.5-Pro over Aya23 | dominates | no | 0 | 0 | no |
| Aya23 over IKUN | dominates | yes | 0 | 0 | yes |
| Aya23 over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over Aya23 | unresolved | no | 2 | 0 | yes |
| Aya23 over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Aya23 over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over Aya23 | dominates | yes | 0 | 0 | yes |
| Aya23 over ONLINE-A | dominates | no | 0 | 0 | no |
| ONLINE-B over Aya23 | dominates | yes | 0 | 0 | yes |
| ONLINE-W over Aya23 | unresolved | no | 2 | 0 | yes |
| TranssionMT over Aya23 | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over Aya23 | dominates | yes | 0 | 0 | yes |
| refA over Aya23 | dominates | no | 0 | 0 | yes |
| refB over Aya23 | dominates | yes | 0 | 0 | yes |
| CommandR-plus over Claude-3.5 | unresolved | no | 7 | 0 | no |
| Dubformer over Claude-3.5 | unresolved | no | 1 | 0 | no |
| GPT-4 over Claude-3.5 | dominates | no | 0 | 0 | yes |
| Claude-3.5 over Gemini-1.5-Pro | contested | no | 6 | 2 | no |
| Claude-3.5 over IKUN | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over IKUN-C | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over IOL-Research | dominates | no | 0 | 0 | no |
| Claude-3.5 over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over Claude-3.5 | unresolved | no | 4 | 0 | no |
| Claude-3.5 over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-B over Claude-3.5 | unresolved | no | 1 | 0 | no |
| ONLINE-W over Claude-3.5 | contested | no | 7 | 2 | no |
| TranssionMT over Claude-3.5 | unresolved | no | 3 | 0 | no |
| Unbabel-Tower70B over Claude-3.5 | unresolved | no | 1 | 0 | no |
| Claude-3.5 over refA | unresolved | no | 2 | 0 | no |
| refB over Claude-3.5 | unresolved | no | 1 | 0 | no |
| Dubformer over CommandR-plus | dominates | no | 0 | 0 | no |
| GPT-4 over CommandR-plus | dominates | no | 0 | 0 | yes |
| CommandR-plus over Gemini-1.5-Pro | contested | no | 7 | 4 | no |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| CommandR-plus over IKUN | dominates | yes | 0 | 0 | yes |
| CommandR-plus over IKUN-C | dominates | yes | 0 | 0 | yes |
| CommandR-plus over IOL-Research | dominates | no | 0 | 0 | no |
| CommandR-plus over Llama3-70B | dominates | yes | 0 | 0 | yes |
| CommandR-plus over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over CommandR-plus | dominates | no | 0 | 0 | no |
| CommandR-plus over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-B over CommandR-plus | dominates | no | 0 | 0 | no |
| CommandR-plus over ONLINE-W | unresolved | no | 2 | 0 | no |
| TranssionMT over CommandR-plus | dominates | no | 0 | 0 | no |
| Unbabel-Tower70B over CommandR-plus | dominates | no | 0 | 0 | no |
| CommandR-plus over refA | unresolved | no | 3 | 0 | no |
| refB over CommandR-plus | dominates | no | 0 | 0 | no |
| GPT-4 over Dubformer | dominates | no | 0 | 0 | no |
| Dubformer over Gemini-1.5-Pro | contested | no | 6 | 2 | no |
| Dubformer over IKUN | dominates | yes | 0 | 0 | yes |
| Dubformer over IKUN-C | dominates | yes | 0 | 0 | yes |
| Dubformer over IOL-Research | dominates | no | 0 | 0 | yes |
| Dubformer over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Dubformer over MSLC | dominates | yes | 0 | 0 | yes |
| Dubformer over Mistral-Large | unresolved | no | 3 | 0 | no |
| Dubformer over ONLINE-A | dominates | yes | 0 | 0 | yes |
| Dubformer over ONLINE-B | contested | no | 4 | 3 | no |
| Dubformer over ONLINE-W | dominates | no | 0 | 0 | no |
| Dubformer over TranssionMT | contested | no | 4 | 2 | no |
| Dubformer over Unbabel-Tower70B | contested | no | 6 | 2 | no |
| Dubformer over refA | unresolved | no | 2 | 0 | yes |
| Dubformer over refB | contested | no | 4 | 1 | no |
| GPT-4 over Gemini-1.5-Pro | contested | no | 4 | 2 | no |
| GPT-4 over IKUN | dominates | yes | 0 | 0 | yes |
| GPT-4 over IKUN-C | dominates | yes | 0 | 0 | yes |
| GPT-4 over IOL-Research | dominates | no | 0 | 0 | yes |
| GPT-4 over Llama3-70B | dominates | yes | 0 | 0 | yes |
| GPT-4 over MSLC | dominates | yes | 0 | 0 | yes |
| GPT-4 over Mistral-Large | unresolved | no | 2 | 0 | yes |
| GPT-4 over ONLINE-A | dominates | yes | 0 | 0 | yes |
| GPT-4 over ONLINE-B | contested | no | 4 | 2 | no |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| GPT-4 over ONLINE-W | dominates | yes | 0 | 0 | yes |
| GPT-4 over TranssionMT | contested | no | 3 | 2 | no |
| GPT-4 over Unbabel-Tower70B | contested | no | 4 | 2 | no |
| GPT-4 over refA | unresolved | no | 2 | 0 | yes |
| GPT-4 over refB | contested | no | 3 | 1 | no |
| Gemini-1.5-Pro over IKUN | dominates | no | 0 | 0 | no |
| Gemini-1.5-Pro over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over Gemini-1.5-Pro | contested | no | 7 | 4 | no |
| Gemini-1.5-Pro over Llama3-70B | dominates | no | 0 | 0 | no |
| Gemini-1.5-Pro over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over Gemini-1.5-Pro | contested | no | 7 | 2 | no |
| Gemini-1.5-Pro over ONLINE-A | dominates | no | 0 | 0 | no |
| ONLINE-B over Gemini-1.5-Pro | unresolved | no | 4 | 0 | no |
| ONLINE-W over Gemini-1.5-Pro | contested | no | 7 | 4 | no |
| TranssionMT over Gemini-1.5-Pro | unresolved | no | 6 | 0 | no |
| Unbabel-Tower70B over Gemini-1.5-Pro | unresolved | no | 6 | 0 | no |
| refA over Gemini-1.5-Pro | contested | no | 7 | 4 | no |
| refB over Gemini-1.5-Pro | unresolved | no | 6 | 0 | no |
| IKUN over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over IKUN | dominates | yes | 0 | 0 | yes |
| Llama3-70B over IKUN | unresolved | no | 5 | 0 | no |
| IKUN over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over IKUN | dominates | yes | 0 | 0 | yes |
| ONLINE-A over IKUN | dominates | no | 0 | 0 | yes |
| ONLINE-B over IKUN | dominates | yes | 0 | 0 | yes |
| ONLINE-W over IKUN | dominates | yes | 0 | 0 | yes |
| TranssionMT over IKUN | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over IKUN | dominates | yes | 0 | 0 | yes |
| refA over IKUN | dominates | yes | 0 | 0 | yes |
| refB over IKUN | dominates | yes | 0 | 0 | yes |
| IOL-Research over IKUN-C | dominates | yes | 0 | 0 | yes |
| Llama3-70B over IKUN-C | dominates | no | 0 | 0 | yes |
| IKUN-C over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over IKUN-C | dominates | yes | 0 | 0 | yes |
| ONLINE-A over IKUN-C | dominates | yes | 0 | 0 | yes |
| ONLINE-B over IKUN-C | dominates | yes | 0 | 0 | yes |
| ONLINE-W over IKUN-C | dominates | yes | 0 | 0 | yes |
| TranssionMT over IKUN-C | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over IKUN-C | dominates | yes | 0 | 0 | yes |
| refA over IKUN-C | dominates | yes | 0 | 0 | yes |
| refB over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over Llama3-70B | dominates | yes | 0 | 0 | yes |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| IOL-Research over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over IOL-Research | dominates | no | 0 | 0 | no |
| IOL-Research over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-B over IOL-Research | dominates | no | 0 | 0 | yes |
| ONLINE-W over IOL-Research | unresolved | no | 2 | 0 | no |
| TranssionMT over IOL-Research | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over IOL-Research | dominates | no | 0 | 0 | yes |
| refA over IOL-Research | unresolved | no | 2 | 0 | no |
| refB over IOL-Research | dominates | no | 0 | 0 | yes |
| Llama3-70B over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over Llama3-70B | dominates | yes | 0 | 0 | yes |
| ONLINE-A over Llama3-70B | dominates | no | 0 | 0 | yes |
| ONLINE-B over Llama3-70B | dominates | yes | 0 | 0 | yes |
| ONLINE-W over Llama3-70B | dominates | yes | 0 | 0 | yes |
| TranssionMT over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over Llama3-70B | dominates | yes | 0 | 0 | yes |
| refA over Llama3-70B | dominates | yes | 0 | 0 | yes |
| refB over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Mistral-Large over MSLC | dominates | yes | 0 | 0 | yes |
| ONLINE-A over MSLC | dominates | yes | 0 | 0 | yes |
| ONLINE-B over MSLC | dominates | yes | 0 | 0 | yes |
| ONLINE-W over MSLC | dominates | yes | 0 | 0 | yes |
| TranssionMT over MSLC | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over MSLC | dominates | yes | 0 | 0 | yes |
| refA over MSLC | dominates | yes | 0 | 0 | yes |
| refB over MSLC | dominates | yes | 0 | 0 | yes |
| Mistral-Large over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-B over Mistral-Large | dominates | no | 0 | 0 | no |
| Mistral-Large over ONLINE-W | dominates | no | 0 | 0 | no |
| TranssionMT over Mistral-Large | dominates | no | 0 | 0 | no |
| Unbabel-Tower70B over Mistral-Large | dominates | no | 0 | 0 | no |
| Mistral-Large over refA | dominates | no | 0 | 0 | no |
| refB over Mistral-Large | dominates | no | 0 | 0 | no |
| ONLINE-B over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-W over ONLINE-A | dominates | yes | 0 | 0 | yes |
| TranssionMT over ONLINE-A | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over ONLINE-A | dominates | yes | 0 | 0 | yes |
| refA over ONLINE-A | dominates | yes | 0 | 0 | yes |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| refB over ONLINE-A | dominates | yes | 0 | 0 | yes |
| ONLINE-B over ONLINE-W | dominates | no | 0 | 0 | no |
| ONLINE-B over TranssionMT | dominates | no | 0 | 0 | no |
| ONLINE-B over Unbabel-Tower70B | unresolved | no | 3 | 0 | no |
| ONLINE-B over refA | dominates | no | 0 | 0 | no |
| ONLINE-B over refB | unresolved | no | 1 | 0 | no |
| TranssionMT over ONLINE-W | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over ONLINE-W | dominates | no | 0 | 0 | no |
| ONLINE-W over refA | unresolved | no | 4 | 0 | no |
| refB over ONLINE-W | dominates | no | 0 | 0 | no |
| TranssionMT over Unbabel-Tower70B | unresolved | no | 5 | 0 | no |
| TranssionMT over refA | dominates | no | 0 | 0 | no |
| TranssionMT over refB | unresolved | no | 4 | 0 | no |
| Unbabel-Tower70B over refA | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over refB | unresolved | no | 1 | 0 | no |
| refB over refA | dominates | no | 0 | 0 | yes |

**translation, WMT24 ja-zh, 15 systems.** Published order at ('wmt', 'none'): Claude-3.5 > refA > GPT-4 > DLUT-GTCOM > Unbabel-Tower70B > Gemini-1.5-Pro > CommandR-plus > IOL-Research > Llama3-70B > Aya23 > Team-J > NTTSU > ONLINE-B > IKUN-C > MSLC. 559 items. Accuracy by system and cell, from which every quantity below follows:

| system | wmt none | wmt clipped | plain none | plain clipped |
|---|---|---|---|---|
| Claude-3.5 | -1.415 | -1.366 | -1.417 | -1.363 |
| refA | -1.537 | -1.512 | -1.493 | -1.479 |
| GPT-4 | -1.679 | -1.679 | -1.716 | -1.716 |
| DLUT-GTCOM | -1.734 | -1.667 | -1.779 | -1.745 |
| Unbabel-Tower70B | -1.903 | -1.874 | -1.940 | -1.913 |
| Gemini-1.5-Pro | -2.071 | -1.850 | -1.735 | -1.691 |
| CommandR-plus | -2.197 | -2.074 | -2.282 | -2.157 |
| IOL-Research | -2.448 | -2.308 | -2.287 | -2.166 |
| Llama3-70B | -3.426 | -3.266 | -3.121 | -3.050 |
| Aya23 | -3.459 | -3.263 | -3.179 | -3.091 |
| Team-J | -4.534 | -4.301 | -4.027 | -3.893 |
| NTTSU | -5.126 | -4.009 | -3.855 | -3.717 |
| ONLINE-B | -5.821 | -5.223 | -5.225 | -4.934 |
| IKUN-C | -7.652 | -6.824 | -5.469 | -5.290 |
| MSLC | -10.727 | -8.184 | -7.558 | -7.220 |

| system | major-only none | major-only clipped | count none | count clipped |
|---|---|---|---|---|
| Claude-3.5 | -0.190 | -0.190 | -0.657 | -0.657 |
| refA | -0.210 | -0.210 | -0.653 | -0.653 |
| GPT-4 | -0.226 | -0.226 | -0.812 | -0.812 |
| DLUT-GTCOM | -0.258 | -0.258 | -0.748 | -0.748 |
| Unbabel-Tower70B | -0.271 | -0.271 | -0.857 | -0.857 |
| Gemini-1.5-Pro | -0.243 | -0.243 | -0.762 | -0.762 |
| CommandR-plus | -0.337 | -0.337 | -0.933 | -0.933 |
| IOL-Research | -0.339 | -0.339 | -0.931 | -0.931 |
| Llama3-70B | -0.456 | -0.456 | -1.296 | -1.296 |
| Aya23 | -0.476 | -0.476 | -1.273 | -1.273 |
| Team-J | -0.673 | -0.673 | -1.336 | -1.336 |
| NTTSU | -0.612 | -0.612 | -1.409 | -1.409 |
| ONLINE-B | -0.877 | -0.877 | -1.716 | -1.716 |
| IKUN-C | -0.995 | -0.995 | -1.488 | -1.488 |
| MSLC | -1.331 | -1.331 | -2.233 | -2.233 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Claude-3.5 over Aya23 | dominates | yes | 0 | 0 | yes |
| CommandR-plus over Aya23 | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over Aya23 | dominates | yes | 0 | 0 | yes |
| GPT-4 over Aya23 | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over Aya23 | dominates | yes | 0 | 0 | yes |
| Aya23 over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over Aya23 | dominates | yes | 0 | 0 | yes |
| Llama3-70B over Aya23 | unresolved | no | 3 | 0 | no |
| Aya23 over MSLC | dominates | yes | 0 | 0 | yes |
| Aya23 over NTTSU | dominates | no | 0 | 0 | yes |
| Aya23 over ONLINE-B | dominates | yes | 0 | 0 | yes |
| Aya23 over Team-J | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over Aya23 | dominates | yes | 0 | 0 | yes |
| refA over Aya23 | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over CommandR-plus | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over DLUT-GTCOM | dominates | no | 0 | 0 | no |
| Claude-3.5 over GPT-4 | dominates | no | 0 | 0 | no |
| Claude-3.5 over Gemini-1.5-Pro | dominates | no | 0 | 0 | yes |
| Claude-3.5 over IKUN-C | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over IOL-Research | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over MSLC | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over NTTSU | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over ONLINE-B | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over Team-J | dominates | yes | 0 | 0 | yes |
| Claude-3.5 over Unbabel-Tower70B | dominates | yes | 0 | 0 | yes |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Claude-3.5 over refA | unresolved | no | 2 | 0 | no |
| DLUT-GTCOM over CommandR-plus | dominates | yes | 0 | 0 | yes |
| GPT-4 over CommandR-plus | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over CommandR-plus | dominates | no | 0 | 0 | no |
| CommandR-plus over IKUN-C | dominates | yes | 0 | 0 | yes |
| CommandR-plus over IOL-Research | unresolved | no | 2 | 0 | no |
| CommandR-plus over Llama3-70B | dominates | yes | 0 | 0 | yes |
| CommandR-plus over MSLC | dominates | yes | 0 | 0 | yes |
| CommandR-plus over NTTSU | dominates | yes | 0 | 0 | yes |
| CommandR-plus over ONLINE-B | dominates | yes | 0 | 0 | yes |
| CommandR-plus over Team-J | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over CommandR-plus | dominates | no | 0 | 0 | no |
| refA over CommandR-plus | dominates | yes | 0 | 0 | yes |
| GPT-4 over DLUT-GTCOM | unresolved | no | 3 | 0 | no |
| DLUT-GTCOM over Gemini-1.5-Pro | unresolved | no | 4 | 0 | no |
| DLUT-GTCOM over IKUN-C | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over IOL-Research | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over Llama3-70B | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over MSLC | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over NTTSU | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over ONLINE-B | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over Team-J | dominates | yes | 0 | 0 | yes |
| DLUT-GTCOM over Unbabel-Tower70B | dominates | no | 0 | 0 | no |
| refA over DLUT-GTCOM | dominates | no | 0 | 0 | no |
| GPT-4 over Gemini-1.5-Pro | unresolved | no | 3 | 0 | no |
| GPT-4 over IKUN-C | dominates | yes | 0 | 0 | yes |
| GPT-4 over IOL-Research | dominates | yes | 0 | 0 | yes |
| GPT-4 over Llama3-70B | dominates | yes | 0 | 0 | yes |
| GPT-4 over MSLC | dominates | yes | 0 | 0 | yes |
| GPT-4 over NTTSU | dominates | yes | 0 | 0 | yes |
| GPT-4 over ONLINE-B | dominates | yes | 0 | 0 | yes |
| GPT-4 over Team-J | dominates | yes | 0 | 0 | yes |
| GPT-4 over Unbabel-Tower70B | dominates | no | 0 | 0 | no |
| refA over GPT-4 | dominates | no | 0 | 0 | no |
| Gemini-1.5-Pro over IKUN-C | dominates | yes | 0 | 0 | yes |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Gemini-1.5-Pro over IOL-Research | dominates | no | 0 | 0 | no |
| Gemini-1.5-Pro over Llama3-70B | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over MSLC | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over NTTSU | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over ONLINE-B | dominates | yes | 0 | 0 | yes |
| Gemini-1.5-Pro over Team-J | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over Gemini-1.5-Pro | unresolved | no | 7 | 0 | no |
| refA over Gemini-1.5-Pro | dominates | no | 0 | 0 | yes |
| IOL-Research over IKUN-C | dominates | yes | 0 | 0 | yes |
| Llama3-70B over IKUN-C | dominates | no | 0 | 0 | yes |
| IKUN-C over MSLC | dominates | yes | 0 | 0 | yes |
| NTTSU over IKUN-C | dominates | no | 0 | 0 | yes |
| ONLINE-B over IKUN-C | contested | no | 2 | 2 | yes |
| Team-J over IKUN-C | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over IKUN-C | dominates | yes | 0 | 0 | yes |
| refA over IKUN-C | dominates | yes | 0 | 0 | yes |
| IOL-Research over Llama3-70B | dominates | yes | 0 | 0 | yes |
| IOL-Research over MSLC | dominates | yes | 0 | 0 | yes |
| IOL-Research over NTTSU | dominates | yes | 0 | 0 | yes |
| IOL-Research over ONLINE-B | dominates | yes | 0 | 0 | yes |
| IOL-Research over Team-J | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over IOL-Research | dominates | no | 0 | 0 | yes |
| refA over IOL-Research | dominates | yes | 0 | 0 | yes |
| Llama3-70B over MSLC | dominates | yes | 0 | 0 | yes |
| Llama3-70B over NTTSU | dominates | no | 0 | 0 | yes |
| Llama3-70B over ONLINE-B | dominates | yes | 0 | 0 | yes |
| Llama3-70B over Team-J | dominates | no | 0 | 0 | yes |
| Unbabel-Tower70B over Llama3-70B | dominates | yes | 0 | 0 | yes |
| refA over Llama3-70B | dominates | yes | 0 | 0 | yes |
| NTTSU over MSLC | dominates | yes | 0 | 0 | yes |
| ONLINE-B over MSLC | dominates | yes | 0 | 0 | yes |
| Team-J over MSLC | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over MSLC | dominates | yes | 0 | 0 | yes |
| refA over MSLC | dominates | yes | 0 | 0 | yes |
| NTTSU over ONLINE-B | dominates | no | 0 | 0 | no |
| Team-J over NTTSU | unresolved | no | 5 | 0 | no |
| Unbabel-Tower70B over NTTSU | dominates | yes | 0 | 0 | yes |
| refA over NTTSU | dominates | yes | 0 | 0 | yes |
| Team-J over ONLINE-B | dominates | yes | 0 | 0 | yes |
| Unbabel-Tower70B over ONLINE-B | dominates | yes | 0 | 0 | yes |
| refA over ONLINE-B | dominates | yes | 0 | 0 | yes |

*continued on the next page*

*continued from the previous page*

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| Unbabel-Tower70B over Team-J | dominates | yes | 0 | 0 | yes |
| refA over Team-J | dominates | yes | 0 | 0 | yes |
| refA over Unbabel-Tower70B | dominates | no | 0 | 0 | no |

**docking, 7 programs.** Published order at ('rmsd2+valid', 'none'): vina > gold > diffdock > deepdock > tankbind > unimol > equibind. 428 items. Accuracy by system and cell, from which every quantity below follows:

| system | rmsd2 none | rmsd2 energy minimization | rmsd2+valid none | rmsd2+valid energy minimization |
|---|---|---|---|---|
| vina | 0.523 | 0.484 | 0.512 | 0.421 |
| gold | 0.512 | 0.479 | 0.477 | 0.423 |
| diffdock | 0.379 | 0.381 | 0.136 | 0.334 |
| deepdock | 0.178 | 0.196 | 0.047 | 0.129 |
| tankbind | 0.149 | 0.199 | 0.026 | 0.119 |
| unimol | 0.229 | 0.280 | 0.019 | 0.175 |
| equibind | 0.026 | 0.054 | 0.002 | 0.047 |

| system | rmsd1 none | rmsd1 energy minimization | rmsd2+chem none | rmsd2+chem energy minimization |
|---|---|---|---|---|
| vina | 0.344 | 0.306 | 0.519 | 0.458 |
| gold | 0.344 | 0.292 | 0.509 | 0.470 |
| diffdock | 0.145 | 0.187 | 0.355 | 0.357 |
| deepdock | 0.047 | 0.072 | 0.133 | 0.154 |
| tankbind | 0.026 | 0.072 | 0.072 | 0.133 |
| unimol | 0.042 | 0.101 | 0.063 | 0.194 |
| equibind | 0.000 | 0.023 | 0.026 | 0.051 |

Per-pair classification. A pair is *contested* only when some cell puts it the other way round with an interval excluding zero; a pair that is neither dominating nor contested is one the benchmark does not resolve, which is a different failure and is not counted as a reversal.

| pair (as the published cell orders it) | verdict | sep. every cell | cells reversing | with an interval | own cell resolves |
|---|---|---|---|---|---|
| diffdock over deepdock | dominates | yes | 0 | 0 | yes |
| deepdock over equibind | dominates | yes | 0 | 0 | yes |
| gold over deepdock | dominates | yes | 0 | 0 | yes |
| deepdock over tankbind | unresolved | no | 2 | 0 | no |
| deepdock over unimol | contested | no | 5 | 2 | yes |
| vina over deepdock | dominates | yes | 0 | 0 | yes |
| diffdock over equibind | dominates | yes | 0 | 0 | yes |
| gold over diffdock | dominates | yes | 0 | 0 | yes |
| diffdock over tankbind | dominates | yes | 0 | 0 | yes |
| diffdock over unimol | dominates | yes | 0 | 0 | yes |
| vina over diffdock | dominates | yes | 0 | 0 | yes |
| gold over equibind | dominates | yes | 0 | 0 | yes |
| tankbind over equibind | dominates | yes | 0 | 0 | yes |
| unimol over equibind | dominates | yes | 0 | 0 | yes |
| vina over equibind | dominates | yes | 0 | 0 | yes |
| gold over tankbind | dominates | yes | 0 | 0 | yes |
| gold over unimol | dominates | yes | 0 | 0 | yes |
| vina over gold | unresolved | no | 3 | 0 | no |
| tankbind over unimol | contested | no | 6 | 4 | no |
| vina over tankbind | dominates | yes | 0 | 0 | yes |
| vina over unimol | dominates | yes | 0 | 0 | yes |

**The nine ESA translation boards.** Their per-cell accuracy matrices are not printed: nine boards of eleven to sixteen systems by eight cells is four hundred numbers that no reader checks, and they are in `results/robust_order_wmt24_esa.json` in full. What the text uses from them is here.

| pair | systems | segments | annotators | dominate | contested (cert.) | unresolved | tiers |
|---|---|---|---|---|---|---|---|
| cs-uk | 12 | 1954 | 14 | 55 of 66 | 1 (0) | 10 | 7 |
| en-cs | 16 | 634 | 19 | 103 of 120 | 4 (0) | 13 | 9 |
| en-es | 14 | 634 | 14 | 76 of 91 | 3 (0) | 12 | 8 |
| en-hi | 11 | 634 | 15 | 39 of 55 | 9 (2) | 7 | 5 |
| en-is | 11 | 634 | 4 | 50 of 55 | 2 (1) | 3 | 8 |
| en-ja | 13 | 634 | 14 | 63 of 78 | 5 (1) | 10 | 7 |
| en-ru | 14 | 634 | 7 | 74 of 91 | 1 (0) | 16 | 7 |
| en-uk | 11 | 634 | 8 | 45 of 55 | 0 (0) | 10 | 6 |
| en-zh | 13 | 634 | 14 | 48 of 78 | 8 (0) | 22 | 5 |

**The eight boards of the previous edition.** Same treatment, and the column a reader wants that the later edition cannot supply is the last but one: how many ratings a segment carries, which is what gives this protocol its criterion axis.

| pair | systems | segments | annotators | ratings | dominate | contested (cert.) | unresolved | tiers |
|---|---|---|---|---|---|---|---|---|
| ces-ukr | 14 | 1009 | 146 | 1.642 | 78 of 91 | 5 (0) | 8 | 8 |
| deu-eng | 14 | 509 | 82 | 1.905 | 85 of 91 | 1 (0) | 5 | 10 |
| eng-ces | 16 | 1098 | 162 | 1.453 | 98 of 120 | 4 (0) | 18 | 8 |
| eng-deu | 13 | 549 | 87 | 1.999 | 66 of 78 | 2 (0) | 10 | 8 |
| eng-jpn | 17 | 1098 | 164 | 1.399 | 124 of 136 | 2 (0) | 10 | 10 |
| eng-zho | 16 | 1098 | 154 | 1.397 | 108 of 120 | 0 (0) | 12 | 11 |
| jpn-eng | 18 | 1120 | 174 | 1.382 | 138 of 153 | 0 (0) | 15 | 13 |
| zho-eng | 16 | 884 | 128 | 1.452 | 107 of 120 | 6 (1) | 7 | 10 |

### D.1 THE THREE CONVENTIONS THE DOCKING TABLE FORCES

The released table forces three conventions, which are declared rather than chosen. Its two flatness checks are recorded only in the un-minimised arm, so a criterion using them would not exist in half the grid. They are dropped from every cell rather than from one, which keeps each criterion the same object in each cell. Including them changes the published cell by $0.0000$. A row with no scorable

pose counts as a miss and not as an absent observation, because dropping those rows would score each program on the subset it managed to produce output for. The crystal ligand ships as an eighth row. It is the reference re-scored through the same pipeline rather than a method, and it is excluded.

None of the three is a cell of the grid. Each is a property of the released table that any re-scoring of it inherits, which is why they are stated here rather than swept.

# E DIAGNOSIS OF THE RECALL SHORTFALL

Table 6 lists nine diagnostic measurements and the factor each one constrains, as a summary of what this appendix then reports in full; it also states three propositions that account for the measurements, and describes two interventions that test the propositions.

Table 6: Nine diagnostic measurements, each attached to the factor it constrains. Sample sizes differ between rows, and so does the evidence: *paired* carries an interval on the difference, *single* is one run against another with none, and *bound* is a best-of-several figure that limits instead of estimates.

| measurement | factor | evidence | result |
|---|---|---|---|
| learned selection against a frequency prior | selection | paired | $0.266$ against $0.410$, $\Delta = -0.144$ $[-0.196, -0.095]$ ($n$=245) |
| selection breadth, 30 to 300 rules | selection | single | recall $0.352 \rightarrow 0.413$; substrates with a hit in the pool $0.482 \rightarrow 0.608$ ($n$=245) |
| training set, 2,418 to 4,787 substrates | selection | single | recall $0.330 \rightarrow 0.334$, inside the $\pm 0.010$ seed spread |
| depth-2 rule application | coverage | single | ceiling $+0.012$, at $8.5\times$ the candidate cost ($n$=150) |
| composition of the uncovered set | coverage | single | $18.3\%$ of references outside the bank ($n$=1,170); the largest mass bin holds $6.0\%$ of it |
| site-of-metabolism prior | ranking | bound | $+0.011$ at the best of five weightings, itself chosen here ($n$=245) |
| abstention on filter score | ranking | single | precision never above $0.123$; recall $0.388 \rightarrow 0.041$ ($n$=1,170) |
| propensity-scored selection loss | selection | paired | $0.249$ against $0.266$; deficit to the prior widens to $-0.161$ $[-0.215, -0.107]$ ($n$=245) |
| similarity-to-training stratification | coverage | paired | bank coverage rises $0.712 \rightarrow 0.819$ with similarity, and an untrained rule engine falls on a slope the data cannot separate from the learned ones', $+0.022$ $[-0.041, +0.086]$ on F1 ($n$=1,170) |

## E.1 COVERAGE COMPOSITION, DATA SCALE, AND REGIOSELECTIVITY

Three of the nine measurements are reported here; the remaining six are expanded in the subsections that follow and in Propositions 2 and 3.

**Composition of the uncovered set.** On the full clean test split under the tautomer-aware default, $18.3\%$ of reference metabolites lie outside what the bank reaches in one step: $475$ of 2,597, the complement of 0.817. The misses are a diverse tail rather than one absent rule family. They divide into two kinds that call for different fixes. Typing each missed transformation by the reaction-centre change it requires and asking whether the bank holds any rule of that type: $337$ of the $475$ need a type the bank does not have at all, $98$ need a type it does have but whose rule did not fire on that substrate, and $40$ cannot be typed. So a fifth of the ceiling's gap is reachable by loosening rules that already exist and three quarters of it is missing chemistry, and a bank-builder choosing between writing new templates and generalising old ones has the split rather than a single number. Binning them by the substrate-to-metabolite mass difference, on a $500$-substrate subset under strict INCHIKEY matching where that binning was run, the largest bin holds $6.0\%$ of the uncovered set. We report the distribution, not named reaction classes, because a mass difference does not identify a transformation:

the same $+14$ is produced by an N-methylation and by an oxidation of a methylene to a carbonyl. Proposition 3 types the gap by reaction type and bounds how much of it a more general template could close.

**Training-set size.** Recall moves from 0.330 to 0.334 between 2,418 and 4,787 training substrates. Both are ablation runs rather than the deployed checkpoint, so neither is $0.339$ and neither should be read against it. They are also two single runs, not a paired comparison, so no interval is quoted; the change is smaller than the $\pm 0.010$ spread the same configuration shows across three seeds, which is the sense in which it is flat. Doubling the number of substrates does not move selection, which rules out under-training as the explanation for the shortfall. Proposition 2 accounts for that: with one to three rules firing per substrate over a label space of $7{,}581$, more substrates add little to the per-rule positive counts a learned selector would need.

**Site-of-metabolism prior.** Weighting candidates by a learned prior over which site of the substrate is metabolised lifts recall@15 from 0.147 to 0.158 on 245 test substrates. The weight was the best of five values tried on that same split, and the response to it is not monotone, so $+0.011$ is an upper bound on what this signal is worth rather than an estimate of it. Ordering candidates regioselectively within the bank remains hard, and we did not isolate why.

### E.2 Selection breadth and abstention

The number of rules the generator selects is the pipeline's most direct control over the recall–precision trade. Sweeping it on the deployed pipeline, with everything else held fixed, separates selection breadth from the two neighbouring factors.

| Rules applied | pool coverage | recall@15 | mean pool | mean output |
|---|---|---|---|---|
| 30 (deployed) | 0.482 | 0.352 | 12.5 | 7.99 |
| 100 | 0.547 | 0.388 | 35.3 | 9.47 |
| 300 (all applicable) | 0.608 | 0.413 | 107.6 | 9.53 |

Table 7: Selection breadth on the deployed pipeline, $n = 245$ test subset.

Widening the selector from 30 to 300 rules lifts recall@15 by 0.061 and pool coverage by 0.126, at 8.6 times the pool size. This is recall bought with precision, the same trade SyGMa makes by overproducing. The gain saturates: even when essentially every applicable rule is fired, the fraction of substrates with at least one reference somewhere in the pool plateaus at $0.608$, and recall@15 at $0.413$ remains far below SyGMa's $0.572$. Those two are not a ratio (the first counts substrates, the second references) so what the pair shows is that widening the selector stops helping well before either is exhausted. That is not the deployed ranking factor, which is measured on the narrower pool the deployed selector produces and is close to one: widening the selector fills the pool faster than the fifteen output slots can carry, so what selection recovers, the cut-off gives back.

Abstention moves in the opposite direction and gains little. Dropping candidates whose filter score falls below a threshold, then ranking the survivors as usual, traces a frontier along which precision barely moves. It is measured on the broad pool, not the deployed one (all 300 applicable rules, as in the last row of Table 7) which is why its no-threshold baseline sits above the deployed macro recall of $0.339$ at a larger output size. With no threshold, recall@15 is 0.388 at precision 0.098 and 9.96 outputs per substrate; at a threshold of 0.5, recall is 0.146 at precision 0.114 and 1.55 outputs; at 0.9, recall is 0.041 at precision 0.040 and 0.30 outputs. Precision never exceeds about 0.123. Raising the threshold removes true and apparently false candidates in near-equal proportion, because an unannotated candidate is scored a false positive whether or not it is a real metabolite. The best threshold selected on validation lifts F1 from 0.129 to 0.131 by trimming the output tail. Precision here is bounded by the annotation, not by the threshold, which is why the deployed system ranks rather than abstains.

On the same broad pools, the learned ranker can be compared directly against the frequency prior (Table 8). At narrow breadth the two are indistinguishable; as the pool broadens the learned filter pulls ahead while the prior plateaus. Substituting the prior for the learned score at the ranking stage would cost recall, not gain it.

| Rules applied | learned (filter × generator) | frequency prior |
|---|---|---|
| 30 | 0.352 | 0.359 |
| 100 | 0.388 | 0.374 |
| 300 | 0.413 | 0.374 |

Table 8: Learned ranker against the frequency prior on identical pools, by selection breadth.

Taken together, the deficit against SyGMa on this axis is a compound of over-narrow rule selection at the deployed setting, worth about 0.06 in recovered recall, and a residual coverage gap that no breadth setting closes. It is not a deficiency of the ranking model: the learned filter outranks the frequency prior once the pool is broad. These are $n = 245$ measurements, and the deployed baseline of 0.352 is inflated on this subset relative to the full-test macro recall of $0.339$.

### E.3 The decomposition's analogue for a sequence-to-sequence method

The factorisation runs only where the candidate set can be enumerated, which excludes the strongest systems in the comparison. The structural analogue is available anyway, and the three gates are the same three with the bank replaced by what a wide decode can reach: coverage becomes the references appearing anywhere in a wide beam, selection becomes what survives narrowing to the deployed beam, and ranking is the same truncation to $k$. The sets nest for the same reason, so the identity is exact rather than approximate.

MetaPredictor decodes in two stages, $8 \times 8$ then $2 \times 8$, giving $16$ ranked candidates. Re-running it at $16 \times 16$ then $4 \times 16$ on the same checkpoints, pipeline and seed changes nothing but the beam and yields $64$, of which $34.3$ are distinct under the matching criterion against the deployed $10.9$.

| | reach | selection | ranking | recall@15 | what the first factor is |
|---|---|---|---|---|---|
| MetaPredictor | 0.612 | 0.804 | 1.000 | 0.492 | a beam budget |
| GRAIL | 0.701 | 0.454 | | 0.318 | a rule bank |
| SyGMa | 0.520 | 0.947 | | 0.493 | a rule bank |

Read across the row the arithmetic holds exactly, $0.612 \times 0.804 \times 1.000 = 0.492$, which is the realised micro recall to four places. Read down the first column it does not mean the same thing twice, and that is the caveat the analogy needs, not a detail. A beam is a budget, so $0.612$ bounds *this decode* and rises with a wider one; a rule bank is a fixed object, but its reach is not, and §E.5 measures how much of it belongs to the engine rather than the rules. The two columns are structurally parallel and semantically different, and putting them in one column of one table without saying so would be the confusion this paper is about.

Two readings survive that caveat. MetaPredictor's narrowing costs $0.20$ where GRAIL's rule selection costs $0.55$, so a beam is a far less lossy budget than a top-$k$ over $7{,}581$ rules, which is what one would expect of a budget the model itself was trained against. And its ranking factor is $1.000$ exactly, for the reason GRAIL's is near one on this subset: the deployed decode leaves $10.9$ distinct candidates against a budget of $15$, so the field's $k$ binds neither of them. At $k = 15$ SyGMa is truncated on $92.7\%$ of substrates on the full split against MetaPredictor's $1.4\%$ and GRAIL's none, which is the population that fraction is measured on. The budget parameter of §A.5 is therefore acting on one method out of three even here.

All three rows are the $150$ shared substrates; the full-split figures for GRAIL and SyGMa in §A.3 are a different population and are not comparable with these.

### E.4 A third rule-grounded method

The decomposition runs on two methods and one of them is degenerate in the factor the paper is about, SyGMa's selection being $1$ by construction. BioTransformer supplies a third rule-grounded row, and a non-degenerate one: it selects, conditioning reactions on biosystem and enzyme, not firing everything applicable, and its templates ship enumerably: $994$ distinct SMIRKS across its human, environmental-microbial and standardisation reaction files, of which $16$ occur only inside records the

shipped database comments out. Every reach figure here is computed over all 994 and is therefore an upper bound on what the 978 active ones express, since a template can only add products.

| method | rules | bank reach | realised@15 | converts |
|---|---|---|---|---|
| GRAIL | 7,581 | 0.701 | 0.318 | 0.45 |
| SyGMa | 175 | 0.520 | 0.493 | 0.95 |
| BioTransformer | 994 | 0.479 | 0.444 | 0.93 |

The row is a bound on what this template set expresses through *our* engine, and it is not a third point on a trend. Its reach carries an interval, $0.479$ $[0.414, 0.550]$, and §E.5 measures the engine term at fixed rules at $+0.188$ $[+0.170, +0.205]$: a term the size of the $0.479$-to-$0.701$ spread among the three reach figures above, though measured on a different population, and larger than this row's own interval. The direction of that bias is not known for this row, and the reason is instructive. Our engine under-realises SyGMa's rules by about a factor of two, but that is a mismatch of hydrogen convention rather than a property of the engine: $69.4\%$ of BioTransformer's templates carry a hydrogen *atom* on the reactant side against not one of SyGMa's 175, and our loop supplies exactly the hydrogens theirs are written to consume (§E.5). The under-realisation therefore does not transfer here, $0.479$ is not established as a lower bound, and where this row sits in the reach ordering is fixed in neither direction. No monotone reading across the three rows is available, and we state the mechanism as the two-system contrast §A.3 draws until a third measurement of comparable provenance exists. The three reach figures here are of mixed provenance in any case: two are through their own engines and one through ours.

The two properties that follow are the row's actual claim, not caveats on it. Applying BioTransformer's templates uniformly with RDKit is not its own Java implementation, which layers enzyme predicates, cofactor logic and depth control on the same SMIRKS; this is therefore what the template set expresses under a stated application procedure, not what its pipeline enumerates. And its budgeted pool is not observable from frozen predictions, so selection and ranking stay merged in the conversion column rather than separating as they do for GRAIL, which is why the row cannot speak to the selection factor at all. The check that makes it credible is the inequality: realised recall cannot exceed what the bank reaches, which is also what catches a bank read from an incomplete set of the released reaction files, since a partial bank reaches less than the same system realises. The margin at 994 templates is $0.035$, and it holds at the point estimate only: the lower end of the reach interval, $0.414$, falls below the realised $0.444$. If further templates ship elsewhere in that distribution the reach is higher and the conversion lower still: in the direction the table already reports.

### E.5 How much of a bank's reach belongs to the engine

§A.3 compares $0.817$ against $0.539$ and the natural reading is bank breadth. Three measurements say that reading is wrong, and the third is the one that matters. Every reach in this subsection is measured with the substrate expanded, which is not the convention $0.817$ is quoted under; the comparison here is between two *applications* at fixed rules, and holding the expansion fixed across both arms is what makes it one.

**The two banks are not separate objects.** SyGMa's published rule set is one of the curated banks our own is a deduplicated union of, so containment is testable by verbatim string membership, not inference: 152 of its 175 rules are present in the bank, all in the curated portion and none among the mined templates. The comparison is a superset against a subset.

**Depth is not the confound.** The ceiling compared against here is one uncapped depth-1 pass of the whole bank, in the convention every deployed firing path uses. SyGMa's deployed pool is built by a scenario in which phase II fires on phase I products, which is composed rather than single-step, so the two were not measured the same way. Matching them costs little and moves the number the wrong way for us: SyGMa reaches $0.5325$ $[0.507, 0.558]$ at one step against $0.5391$ $[0.514, 0.565]$ deployed, composition worth $+0.0065$ $[+0.0038, +0.0096]$, and the gap against that same ceiling widens from $0.2780$ to $0.2846$ $[0.264, 0.305]$. The re-run reproduces the frozen pool on every substrate, which is the gate this was run behind.

**Reach is a property of the system, not of the rule base.** Hold the rules fixed at the 152 both systems have and swap only the engine. Arm A is the loop as it stood, without the contraction; the paragraph below separates what that costs from what the convention costs, and every figure in this subsection measured before that separation describes the loop as it stood:

| | rules, engine | reach |
|---|---|---|
| A | 152 shared, expanded, loop as it stood | 0.200 $[0.182, 0.218]$ |
| A′ | 152 shared, expanded, loop completed | 0.377 $[0.354, 0.400]$ |
| B | 152 shared, SyGMa's | 0.387 $[0.365, 0.410]$ |
| C | all 175, SyGMa's | 0.532 $[0.507, 0.558]$ |
| D | all 175, SyGMa's composed | 0.539 $[0.514, 0.565]$ |

Every row is the clean test split. The engine at fixed rules is worth $+0.188$ $[+0.170, +0.205]$, the 23 rules absent from our bank $+0.145$ $[+0.128, +0.162]$, and composition $+0.0065$ $[+0.0038, +0.0096]$; the three sum to $D - A$ because they are constructed as its consecutive differences. The engine term alone is the size of the entire reach gap the main text reports.

**Almost all of the term is an unfinished loop.** The arms above expand the substrate and never contract the product. Restoring the contraction lifts the expanded arm from $0.200$ $[0.182, 0.219]$ to $0.377$ $[0.354, 0.400]$ against the unexpanded arm's $0.390$ $[0.367, 0.413]$: the unfinished loop is worth $+0.177$ $[+0.159, +0.195]$ and the convention $+0.013$ $[+0.007, +0.020]$, both paired and both clear of zero. On the worked example below, benzoic acid under aromatic hydroxylation, the incomplete loop yields no product that sanitises and the completed one yields three, which are the three the unexpanded arm yields.

**Completing the loop is itself an undeclared choice, and its obvious implementation is wrong.** The contraction is one line and there are two of them. `AddHs` marks every heavy atom as taking no implicit hydrogens; `RemoveHs` then deletes the drawn hydrogens without lifting that mark, so a template that consumes a mapped hydrogen and puts nothing in its place leaves the atom one short, and the toolkit records an unpaired electron rather than refilling the valence. Clearing the mark first restores the capacity the expansion suspended. Applying the deployed bank to twelve substrates of the clean test split gives $69{,}468$ firings, of which the one-line version yields $30{,}434$ products that sanitise and the other $34{,}503$; $6{,}880$ of the first carry an unpaired electron and none of the second do; and of the $30{,}371$ both produce, $6{,}854$ differ, which is the same population. A metabolite corpus contains no radicals, so a product that acquires one cannot match any reference and is lost without an error. This is a third undeclared choice of the kind §A.2 is about, found inside the repair of the second, and unlike the first two it is not a fork between defensible options: an engine predicting metabolites should not emit radicals. Every arm reported here uses the contraction that does not.

**One call out of three.** The loop makes three discretionary choices before it reports a product: whether the substrate is expanded, whether products are normalised by tautomer canonicalisation or left as plain canonical SMILES, and what counts as a valid fragment. Toggled one at a time over the 152 shared rules on the clean test split, the first carries the entire term and the other two are inert.

| choice | from → to | reach | against the default |
|---|---|---|---|
| explicit hydrogens | yes → no | 0.390 | $+0.190$ $[+0.173, +0.208]$ |
| product normalisation | tautomer → canonical | 0.200 | 0.000 |
| validity floor | on → off | 0.200 | 0.000 |

**The term is a single decision.** The three choices are toggled one at a time over the same 152 rules, with the default configuration gated to reproduce arm A to four decimals and the comparator arm gated against arm B, which puts the whole term on the first (§A.2). The gate binds the rule set, the substrates, the matcher and the hydrogen decision, because moving any of those moves the number it checks; it cannot bind the other two, since every configuration with hydrogens explicit returns that number exactly. That is the second and third rows of that table rather than a gap in the check.

The engine term is $+0.188$, so that one call carries all of it, and what remains against SyGMa's own engine is $+0.002$ $[0.000, +0.005]$: at that setting the two programs are no longer separable on these rules. SyGMa's source calls neither `AddHs` nor `RemoveHs` anywhere, so the contrast is between what the two programs do and not an inference from what they emit. Every figure in this paragraph is micro, a difference of ratio-of-sums, which is the estimator the engine term itself uses. The macro counterparts sit beside them in the artifact because they are a different quantity and not a rounding of the same one: the switch is worth $+0.159$ under it, not $+0.190$, and the residual $+0.006$ $[+0.002, +0.010]$ rather than $+0.002$: an interval that excludes zero where the micro one does not. The separability claim is the micro one, and the two estimators should not be quoted against each other.

That the other two choices move nothing is a statement about these $152$ rules under one setting of the first choice, and it generalises in neither direction on its own. Across the whole $7{,}581$, under the convention the pipeline fires rules in, the normalisation knob is inert exactly: CANONICAL and STANDARDIZE both recover $454$ of $567$, the same references, and reproduce the committed ceiling to four decimals, which is also what gates this loop at the scale of the bank, not of the shared subset. Expand the hydrogens first and the same two modes separate by one recovered reference. Whether a discretionary choice matters is therefore conditional on the other discretionary choices, and "this one does not matter" is a claim about a configuration rather than about a knob.

**What the switch does.** Not matching, which is the natural guess and is wrong. Over the clean test split the $152$ templates fire $20{,}372$ times with hydrogens explicit against $19{,}629$ without, and every firing writes a SMILES either way. The loss is one step later and it is a structure the toolkit refuses: of the $21{,}905$ fragments those products separate into, $13{,}804$ are structures RDKit will not read back, against $87$ of $21{,}106$: $63\%$ against $0.4\%$, and $99.6\%$ of the excess is that, not the heavy-atom floor. A substitution template written against implicit hydrogens does not mention the hydrogen it replaces; it attaches the new group and lets RDKit decrement a count held as a property rather than an atom. Expand the substrate first and that hydrogen is an atom, nothing removes it, and the product carries an aromatic carbon holding both its hydrogen and the new substituent: benzoic acid under aromatic hydroxylation gives `[H]OC(=O)c1c([H])c([H])c([H])(O)c([H])c1[H]`, which does not parse. On an aromatic carbon the toolkit reports the failure as an unkekulisable ring rather than as a valence error, since the over-substituted atom leaves the ring without a kekulé structure; on an aliphatic one it is a valence error directly. The repair is the same and the diagnostic is not. The template is not wrong and the engine is not wrong. They disagree about how a hydrogen is spelled.

**The decision belongs to the bank as much as to the engine.** It does not follow that the expansion is a mistake. Which setting a bank wants is a property of how its templates were written, and no rule file records it. Running the same switch with our engine throughout, over three banks assembled by three groups, against the same substrates and their $567$ references:

| bank | rules | hydrogen atom on reactants | recovered, explicit → implicit | change |
|---|---|---|---|---|
| *explicit arm uncontracted, on the 245-substrate subsample* | | | | |
| SyGMa | 175 | none | $125 \to 299$ | $+0.307$ $[+0.260, +0.355]$ |
| ours | 7,581 | 675 | $414 \to 453$ | $+0.069$ $[+0.041, +0.098]$ |
| BioTransformer | 994 | 690 | $268 \to 74$ | $-0.342$ $[-0.388, -0.298]$ |

The comparison in that table is the expanded arm *without* the contraction against the implicit one, which is what a paper reporting a bank's coverage without an application procedure would produce; §A.2 compares the contracted arm instead, and the two answer different questions. Not one of SyGMa's $175$ templates carries a hydrogen *atom* on its reactant side, and its reach more than doubles once the hydrogens are gone under the uncontracted comparison. $690$ of BioTransformer's $994$ do, and its reach falls to a quarter of itself. Ours lies between them on both counts, and that row was written down before it was measured: at $675$ templates in $7{,}581$ our bank is in SyGMa's convention and should move the same way by much less, which is what it does. Three banks cannot establish the share as a predictor and it is not offered as one. What they establish is that the direction is not a property of our engine, so a reported coverage figure identifies a bank together with a convention, and the second half of that pair is recorded nowhere: including, until this measurement, by us. Our own bank is where it costs us: on these substrates it recovers $414$ references with hydrogens expanded and

453 without. 0.817 is therefore quoted without the expansion, matching the convention the deployed generator fires rules in; the figures this appendix reports for the shared 152 are the expanded ones, and each is labelled where it appears.

**The supervision is built in the other convention, and this is where it costs us most.** The label matrix the generator learns from asks, for each training substrate and each rule, whether applying that rule yields an annotated metabolite. It applies the rule to the substrate expanded with explicit hydrogens. The deployed pipeline fires rules on the substrate as written. Applying the whole bank both ways to 150 training substrates, of the 282 rule–substrate positives that exist in the convention the pipeline fires in, the supervision sees 140 and misses 142, and asserts a further 25 that do not exist at inference; the two matrices agree at a Jaccard of 0.456, and 29 of the 150 substrates carry a different row. The selector is therefore taught about rules other than the ones it will fire, roughly half the time.

What this does and does not touch. The decomposition of §A.3 measures the deployed system through its own application, so the selection factor is what that system achieves and the identity still holds. What it undermines is the reading: "selection is where the loss sits" and the learned selector's deficit to the frequency prior are both conditioned on supervision that disagrees with the firing convention on half its positives, and the prior inherits the same matrix, so neither arm of that comparison is pointed at the rules the pipeline runs. Rebuilding the labels and retraining is the measurement that would settle it and is not reported here. The cheap half of that measurement is reported: the frequency prior is the per-rule share of positives in this matrix and needs no gradient step to recompute. Over the same 150 substrates, 202 rules fire productively and 110 of them are never labelled positive, while 17 carry a label and never fire; the two priors rank the rules at a Spearman of 0.356 and share 39 of their top hundred. The prior arm is therefore not a weaker reading of the same ranking, it is a different one. The defect is the one this paper is named for, in our own pipeline, found by applying the paper's own question to it.

**What the decomposition survives, and what it does not license.** Each method's three factors are measured through its own application and multiply to its own realised recall, so the decomposition itself is unaffected. What it does not license is reading the first factor *across* methods as a property of their rule sets: it is a property of rule set and engine together, which is what §A.3 says. The direction of the effect is opposite to the one that would suit the present system rhetorically: our engine under-realises SyGMa's rules by a factor of two under the expansion, and the same expansion applied to our own bank costs it 0.082 of reach on the clean test split, which is why 0.817 is quoted under the convention the pipeline fires rules in, not under that one, and why the headline that the bank is not the binding constraint is strengthened rather than weakened. The 152 are curated, and the expansion is what the curated half gains from and the mined majority loses to (§E.9), so the factor of two is the size of an effect our own hand-written rules do not suffer: the asymmetry runs against us here, not for us. The mined majority was itself kept only if applying it through this engine regenerated its own annotated product (Appendix B), so it was validated against this engine, not a foreign one and supplies no offsetting overstatement.

**What is convention-dependent is a construct, not a domain.** Which SMARTS constructs an expansion touches is readable from a rule file: counts that already include the hydrogens about to be drawn are inert, counts of explicit connections or of implicit hydrogens lose every match, and the hydrogen *atom* primitive fires more often rather than less. Across six libraries and 51,368 templates the construct almost everyone uses is the inert one, and the two that move are used by disjoint sets of libraries, so a survey by domain would have found nothing: what predicts exposure is which construct a library was written with.

### E.6 THE SAME CHOICE OUTSIDE METABOLISM, FAILING THE OTHER WAY

Everything above is measured on metabolite rule banks, which alone would establish a fact about one small literature. The USPTO retrosynthesis templates are a library from another field, extracted by other people from other reactions for another task, and the instrument applies to them unchanged. What it finds is not a repetition, because the mechanism is the other one.

**The convention travels with the transcription.** One release attributes part of its bank to another tool, and how a rule was carried over decides which convention it carries. Of those rules, the 57 copied verbatim carry the atom primitive not once, while 88 of the 121 re-typed do, absent from all 175 rules of the tool they came from. A convention is therefore acquired in the act of rewriting a template, by people who are not choosing one.

| | metabolite banks | retrosynthesis templates |
|---|---|---|
| patterns | 175 / 7,581 / 994 | 42,554 |
| carrying a hydrogen atom | none / 675 / 690 | none |
| constraining an atom by degree | rare | all 42,554 |
| what the expansion breaks | the product | the match |

Applied to 500 products of the USPTO-50k test split this repository holds (the three-system group's, of the three that circulate under that name (§H)) the library makes 146,346 template–product matches with hydrogens implicit and 26,788 with them explicit, and the number of products with any applicable template falls from 500 to 489. The cause is one construct and the attribution is exact rather than approximate: of the 119,558 matches the expansion costs, 119,558 come back when the degree constraint alone is deleted from the very patterns that lost them. The check is a gate, the measurement is refused below 90%, and it returns 100%.

The three constructs of Table 9 behave differently under the expansion, and the last two rows of it are the controls. One release's copy of another tool's rules carries the atom primitive in half of them where the tool itself carries it in none, and a second extraction from the corpus that pins a connection count in every template pins it in none.

Table 9: Share of templates carrying each construct, by library. Rows below the rule are not counted as independent libraries: one is another release's copy of a library already listed, the other a second extraction from the corpus already listed. BioTransformer's row is measured on the release retrieved for this census, which holds 976 active templates against the 978 of the snapshot the reach figures above were computed on; the sixteen that occur only inside commented-out records reproduce exactly in both.

| library | templates | atom | hydrogen count | connection count |
|---|---|---|---|---|
| this work | 7,581 | 0.089 | 0.075 | 0.000 |
| curated partition | 1,715 | 0.394 | 0.327 | 0.001 |
| mined partition | 5,866 | 0.000 | 0.002 | 0.000 |
| SyGMa, phase 1 | 148 | 0.000 | 0.797 | 0.000 |
| SyGMa, phase 2 | 27 | 0.000 | 0.926 | 0.000 |
| BioTransformer | 976 | 0.696 | 0.439 | 0.007 |
| GLORY, CYP rules | 73 | 0.562 | 0.082 | 0.000 |
| GLORYx, own rules | 9 | 0.444 | 0.111 | 0.000 |
| USPTO extracted | 42,554 | 0.000 | 0.967 | 1.000 |
| *not counted as independent, and reported beside what they derive from* | | | | |
| GLORYx, rules attributed to SyGMa | 178 | 0.494 | 0.747 | 0.000 |
| RetroSim, same corpus as USPTO | 1,351 | 0.000 | 0.881 | 0.000 |

So a degree primitive counts explicit connections, expanding a molecule changes every degree in it, and a library built on them stops matching before it can build anything. That is a different failure from the metabolite one, where the patterns fire just as often and the product is what will not parse, and the two together give the claim its general form. What is convention-dependent is a class of SMARTS construct, not a domain: a primitive that counts explicit structure breaks matching, a template that rewrites without naming the hydrogen it consumes breaks the product. Both are readable from the rule file without running anything, and neither is declared by any library we have looked at.

| what breaks | the construct responsible | measured on |
|---|---|---|
| the match, before any product | a primitive counting explicit structure | 42,554 retro templates |
| the product, once built | an application loop left unfinished | this work's own engine |
| which answer the bank gives | a union of two written conventions | our own 7,581 rules |

**Three ways the choice bites, and one of them can be repaired.** A third mechanism turned up inside our own measurement rather than in someone else's library, and the three together are what makes this a class, not three defects, and each has one measured instance (§A.2).

The first two are read off a rule file and avoided by declaring the convention. The third is not avoidable by any declaration, because no single setting is right for an object assembled from two: one collection is mismeasured whichever is chosen, and the choice does not shift the answer so much as exchange it, which is what §E.9 shows it doing to the two halves of our bank. It is also the only one of the three that can be *recovered* rather than merely avoided, by dispatching the convention per template (Appendix G).

Two things this does not claim. It is applicability, not coverage against references, which is the first step of reach and not the whole of it. And no retrosynthesis system expands its target by convention, so this is a latent hazard in that field rather than an error in it: the point is that the dependence is a property of the patterns, and the patterns are not ours.

### E.7 The bank without a selector, at a matched budget

Everything above localises the loss at selection by observation. The counterfactual states it directly: apply the bank with no selector at all, rank the whole pool with the deployed scorer, and truncate at the number of candidates SyGMa emits. The breadth sweep does not answer this. It ranks the pool the deployed 30-rule selector produces, $12.5$ candidates on average of which $8.0$ are emitted, which has nothing to give at $k = 64$; and the widest row of Table 7 is not the whole bank either, since the generator's calibrated threshold is itself a selector and prunes rules however large $top\text{-}k$ is.

Dropping both, the bank produces $661.7$ raw products per substrate, $81.6$ distinct under the matching criterion; the $661.7$ is what enters the ranker, not what is emitted. The budget is set to SyGMa's emitted count, not to its pool: at $k = 82$ this arm emits $81.6$ distinct structures against the $80.7$ SyGMa emits, both counted after the harness's deduplication, which is applied identically to all five methods (Appendix H). The two counts are on different populations, this arm's $245$ substrates against SyGMa's full split, so the arm's budget of $82$ was chosen to match a published figure and is not a matching measured on one population. At that budget the cut is not even binding, the arm's recall there being equal to its whole distinct pool's coverage. Scored against SyGMa on the same $245$ substrates, macro recall, paired bootstrap over substrates:

| budget $k$ | bank, no selector | SyGMa | difference | |
|---|---|---|---|---|
| 8 | 0.415 | 0.521 | $-0.106$ $[-0.177, -0.034]$ | SyGMa ahead |
| 15 | 0.528 | 0.576 | $-0.048$ $[-0.115, +0.020]$ | indistinguishable |
| 32 | 0.652 | 0.589 | $+0.063$ $[+0.001, +0.125]$ | the bank ahead |
| 64 | 0.735 | 0.593 | $+0.141$ $[+0.085, +0.198]$ | the bank ahead |
| 82 | 0.760 | 0.593 | $+0.167$ $[+0.112, +0.223]$ | the bank ahead |

Table 10: The bank applied without a selector against SyGMa at matched output budgets, $n = 245$.

The comparison that isolates selection needs no second system, and it sits at the field's budget. At $k = 15$, macro on these same $245$ substrates, one bank and one engine give $0.352$ at the deployed selector breadth, $0.413$ with every applicable rule fired (Table 7) and $0.528$ with the selector and its threshold removed. That is the cost of the selection stage measured inside one system, against the macro ceiling reported for these substrates below, and it is not a cross-method quantity: no second engine and no second bank enter it.

The cross-method claim survives its own test, and the condition survives with it. The bank does convert into recall above SyGMa once the selector is gone, but only above the budget the field uses. At $k = 15$ the two are indistinguishable and below it SyGMa wins, so the advantage is bought with output, which is the parameter of §A.5 acting in favour of this system where §A.5 has it acting against us. Reporting only the $k = 82$ row would apply a standard to ourselves that this paper refuses to apply to SyGMa.

Two controls place it. At $300$ rules the same comparison is negative at every budget, $-0.118$ $[-0.188, -0.048]$ at $k = 82$, so it is removing the selector rather than widening it that does the work;

and that arm reproduces the deployed configuration exactly, pool $107.6$ and recall@15 $0.4134$ against the $107.6$ and $0.413$ of Table 7. This is a counterfactual and not an operating point: $82$ candidates per substrate drawn from $661.7$ is not a system anyone would deploy, and its precision would be worse than SyGMa's.

Read against the ceiling, not against SyGMa, the arm reaches $0.760$ against a macro ceiling of $0.854$ on these substrates, both per-substrate means, so $0.094$ of the ceiling is still unreached with the selector gone and the rest of what separates the arm from it is ranking and the cut-off. The micro ceiling for the same substrates, $0.801$, is a different aggregation and not comparable with either.

### E.8 The decomposition applied to the comparator, and what it bounds

Applied to a second method it stops being self-diagnosis. SyGMa is rule-grounded with an enumerable bank and decomposes on the same split in the same frame, with the inverse profile. Its reach is the lower, $0.539$ against $0.817$, comparator minus GRAIL throughout. The two are measured through their own engines and the paired interval on that difference is reported with the engine term it contains (Appendix E.5) rather than here. But it applies every applicable rule and emits the result, converting almost all of that reach where we convert a third of a wider one. Its selection factor is $1$ by construction, because it emits every product its applicable rules generate; the same computation returns $0.539$ for its coverage and $0.953$ for its ranking, and those two are measured. The gap between us is therefore not a worse rule base but the cost of *this* selection stage, which loses more than no selection stage does at the same budget ($0.352$ against $0.528$, macro at $k = 15$ on the same $245$ substrates (Appendix E.7)) and which recall charges only to methods that select.

The frame also bounds this architecture: a perfect selector and ranker would stop at 0.817 as our own engine realises it (§A.2), limited by the corpus, not by tuning, $71\%$ of uncovered transformations being types absent from the bank, so generalising the templates it holds is worth $0.038$. It is carried by the mined majority, $0.785$ on $245$ substrates against the curated fifth's $0.471$, four published collections adding under two per cent beyond what mining the training split reaches (Appendices E, G).

### E.9 Which half of the bank earns the ceiling

The bank is a deduplicated union of four published rule sets and a much larger body of templates mined from the training split, and nothing so far says which part reaches the chemistry. The mined templates are recorded verbatim, so the partition is exact rather than estimated: $5{,}866$ mined and $1{,}715$ curated, partitioning the $7{,}581$ with nothing left over. Applying each subset alone, through the same primitives and on the same substrates as the ceiling itself, the full clean test split:

| subset | rules | share of bank | coverage | share of ceiling |
|---|---|---|---|---|
| curated | 1,715 | 22.6% | 0.487 [0.463, 0.512] | 59.6% |
| mined | 5,866 | 77.4% | 0.800 [0.777, 0.824] | 98.0% |
| union | 7,581 | 100% | 0.817 [0.795, 0.839] | 100% |

The union reproduces the value already stored per substrate for exactly these substrates, to four decimals, which is the gate the measurement refuses to report without. Exclusively, mined templates reach $0.330$ of the references that the curated sets do not; the curated sets reach $0.017$ that the mined templates do not. Four published rule collections, restricted to this corpus, add under two per cent of the ceiling beyond what mining the training split already reaches, and $0.4705$ is reachable by both, carrying the fourth decimal because the third falls exactly on a half.

This does not contradict Appendix G, where discarding the $4{,}271$ rules that never carry a positive label costs recall. Contributing to a deployed pool and contributing to the ceiling are different things.

**The ordering of the two subsets depends on the hydrogen convention.** Applied through the helper the data-preparation path uses, which expands hydrogens onto the substrate before matching, the same rules on the same substrates give curated $0.668$ against mined $0.332$: the curated fifth carrying the ceiling and the mined majority re-deriving it. Both passes are correct arithmetic; they differ only in a convention neither subset is asked about, and it does not move a number, it exchanges

the two halves of the answer. That is the same defect this paper names on the leaderboard, met inside our own measurement, and it is why the primitives are now fixed at the ones §A.2 deploys.

Which of the helper's two steps does it can be read off by holding one fixed and moving the other. The gap between the subsets is $+0.313$ as deployed and $-0.336$ expanded; the validity floor carries $+0.000$ of that and expanding hydrogens carries the remaining $-0.650$. The floor is inert here in the strict sense, it removes about one product in four hundred from either subset, and the whole of the reversal is the hydrogen convention.

Why that convention should divide the bank along its provenance line is answered by a test this paper already fixed in advance. The dispatch criterion of Appendix G asks a purely syntactic question of each template: does its reactant pattern carry the hydrogen *atom* primitive, `[H]` or `[#1]`, which matches nothing at all unless the substrate's hydrogens are drawn. It never sees a provenance label, and it separates the bank as if it had: $675$ of the $7{,}581$ templates require the expansion, every one of them curated and none of the $5{,}866$ mined. The bank is a union of two collections written under different hydrogen conventions, and one of them is mismeasured under whichever single global setting is chosen.

That the notation and not the provenance is doing the work can be checked by holding provenance fixed. Splitting the curated subset in two by the same criterion, on the same substrates:

| curated templates | rules | as deployed | hydrogens expanded |
|---|---|---|---|
| carrying the hydrogen atom primitive | 675 | 0.032 | 0.558 |
| not carrying it | 1,040 | 0.486 | 0.265 |
| mined, for comparison | 5,866 | 0.800 | 0.332 |

The two halves of one provenance move in opposite directions, by a factor of seventeen up and a factor of two down, and the half that carries no hydrogen atom behaves like the mined majority it shares no provenance with. What decides which convention a template needs is how the template is written, and provenance predicts it only because hand-written collections use the primitive and extracted ones never do.

**Why the dispatch recovers anything, registered in advance and decided against us.** Two accounts fit a bank whose halves want opposite conventions. On the first, *mixing* creates the loss: one collection is mismeasured whichever global setting is chosen, and the repair returns what the losing half forfeits. On the second and duller, $675$ hand-written templates simply want the expansion, and the mined majority bears on neither the loss nor its repair. They differ in exactly one measurable way, and that difference was registered in advance of any measurement: if mixing creates the loss the residual is *superadditive* across the partition, $\mathrm{res}(\text{full}) > \mathrm{res}(\text{curated}) + \mathrm{res}(\text{mined})$.

**This adjudication waits on one re-measurement.** It is decided on our own bank's dispatch residuals, and the prediction is a comparison between three of them. Only the whole-bank arm has been measured under the corrected contraction, at $+0.0035$ $[+0.0015, +0.0058]$; the curated and mined arms stand as registered. The registered numbers are printed as what a re-measurement is compared against rather than as a result: the curated subset alone was $+0.0123$ $[+0.0079, +0.0170]$ and the mined subset exactly $+0.0000$ $[0.0000, 0.0000]$. The fourth decimal is carried because the third is what separates them. That last is zero on any population by construction, since none of its $5{,}866$ templates is dispatched anywhere, so it was measured on the $245$-substrate subsample, not the split: it checks that the subsetting does what it says rather than contributing a quantity. Of the three registered predictions the null and the curated one hold and the superadditivity does not, so the second account is the one that survives. Mixing does not create the loss, it *dilutes* it: the mined majority independently reaches part of what the expanded curated templates would have recovered, so their repair is partly redundant once both are present. What the dispatch recovers is therefore the expansion those $675$ templates require and the bank's global setting denies them: a smaller claim than a mixing account would have supported, and the one the measurement carries.

On the two published banks measured since the correction, per-template dispatch does not beat the better global setting at all, and on ours it does by $+0.0035$ $[+0.0015, +0.0058]$ (Appendix G); whether that residual is superadditive across the partition is what the two subset arms have still to say.

### E.10 THE SHORTFALL OF THE SYSTEM USED AS THE CASE STUDY

GRAIL trails both published comparators at the field's budget and leads at none: SyGMa reaches $0.572$ on the $1{,}168$ substrates they share, a paired $-0.242$ $[-0.271,\ -0.212]$. The second confirmatory comparison locates that shortfall in selection, where a learned selector reaches $0.266$ through the same enumeration and filter that rule-firing frequency alone carries to $0.410$, a paired $-0.144$ $[-0.196,\ -0.095]$. Choosing which of 7,581 rules apply is extreme multi-label classification with a near-empty positive set, a regime in which frequency baselines are hard to beat (Jain et al., 2016).

### E.11 SELECTION AS MEASURED AGAINST SELECTION AS DEPLOYED

The learned-versus-prior comparison in Table 6 is a controlled comparison of rule selection, in which each mode picks its own thirty rules and shares the downstream product loop and filter. It is not the deployed recall. The deployed pipeline applies its own top thirty rules, and the breadth sweep confirms the correspondence: the 30-rule baseline of 0.352 on this subset sits close to the deployed macro recall of $0.339$ on the full test, above it by the subset inflation noted above rather than reproducing it exactly.

At that operating point the deployed system is close to prior-independent. Two checkpoints differing only in whether a trained prior buffer was present gave 0.335 and 0.334 on a 291-substrate evaluation. Deployed recall is therefore limited by coverage, not by the prior term, even though the selection experiment shows the learned selector losing to the prior when selection is what is measured. The experiment exposes a weakness that deployment conceals by not depending on the selector's ordering.

Neither multi-step application nor any single rule family moves coverage much, and training-set size is saturated. The dominant addressable loss remains selection, at $0.338$: a learned rule selector that performs worse than counting. The three propositions below explain why.

### E.12 THREE PROPOSITIONS

**Proposition 1: the training objective is mismatched to the metric.** A filter trained under a strictly proper scoring rule learns a globally calibrated posterior, which is Bayes-optimal for AUC and for calibration. Ranking each substrate's pool by that posterior and taking the top $k$ is optimal for recall@$k$ only when pools are homogeneous. GRAIL's pools are not: they range from 17 to 150 candidates and from 1 to 18 positives. A pointwise-calibrated scorer can therefore be suboptimal for recall@$k$ while a listwise surrogate dominates it. A two-substrate counterexample makes this concrete: the reordering that improves recall is verifiably not globally calibrated, so a proper scoring rule rejects it.

A controlled swap of the objective, holding everything else fixed, confirms the prediction. The same two auxiliary heads, the same frozen generator and filter, and the same candidate pool are used twice, differing only in how the heads were trained. Trained pointwise and multiplied into the rank, they add nothing: $+0.0002$, 95% CI $[-0.0073, +0.0077]$. Fine-tuned with a listwise ranking loss against the frozen generator and filter context, the identical heads gain $+0.0089$ $[+0.0032, +0.0153]$ over their pointwise counterpart and $+0.0091$ $[+0.0027, +0.0160]$ over the baseline, on a paired bootstrap over the full test set of $1{,}170$ substrates. Capacity, features and pool are constant, so the margin is attributable to the objective. A second re-ranking arm reproduces the direction at the same sample size: adding type and site terms lifts recall@15 from $0.388$ to $0.404$ on the same broad pool, a paired $+0.0165$ $[+0.006, +0.027]$, while the factorised signal on its own is not significant ($+0.001$, $[-0.010, +0.013]$).

**Proposition 2: rule selection is extreme multi-label classification, where counting is a strong baseline.** Selecting which of $7{,}581$ rules apply to a substrate is an extreme multi-label classification problem with a nearly empty positive set: a median of five rules carries a positive label per substrate, $0.07$ per cent of the label space, and a mean of $11.3$, $0.15$ per cent. In this regime the classification literature has long established that frequency and propensity baselines are hard to beat and that naive losses are dominated by label sparsity; propensity-scored losses were introduced for precisely this failure mode (Jain et al., 2016). Metabolism rule selection inherits the pathology, which the domain literature does not name. Identifying it here, and testing it, is the contribution.

The mechanism is specific. Under positive-unlabelled annotation with an approximately constant labelling propensity, a learner with constant unlabelled weighting recovers a propensity-distorted score whose dominant component is the marginal rule-firing rate. That rate is the frequency prior, so the prior is competitive by construction at the selection stage, and the learned selector improves on it only when substrate-conditional variation in prevalence exceeds estimation noise. The saturation with training-set size indicates that it does not at present scale.

The supporting measurement is the learned-only recall of 0.266 against the prior's 0.410 ($\Delta = -0.144$, $[-0.196, -0.095]$, $n = 245$). The advantage is specific to selection and does not generalise to ranking: once rules have been applied broadly, the learned filter is the better scorer, at 0.413 against the prior's 0.374 on the same 245 substrates, a paired $+0.039$ $[+0.007, +0.072]$. That interval clears zero and not by much, and where it comes from is worth saying. The two arms order one identical pool, so they can differ only where the pool exceeds the fifteen-candidate budget by enough for order to matter. That is 32 substrates: the filter ahead on 22, behind on 10, tied with the prior on the remaining 213. The proposition makes a refutable prediction, that reweighting the labelled loss by the inverse estimated propensity should shrink the prior's edge; this is an open test. The propensity model itself is an unmeasured modelling assumption.

**Proposition 3: single-step rule application has a ceiling below one.** Single-step recall cannot exceed single-step coverage, and coverage is bounded away from one because a non-vanishing fraction of references are multi-generation products, for instance oxidation followed by conjugation, unreachable by any single rule application. No improvement in ranking recovers that residual.

Two measurements witness the bound. Depth-2 application raises the ceiling by 0.012 at 8.5 times the candidate cost. On the external GLORYx set the uncapped single-step ceiling is 0.709, and its references include multi-generation metabolites that no single step can reach.

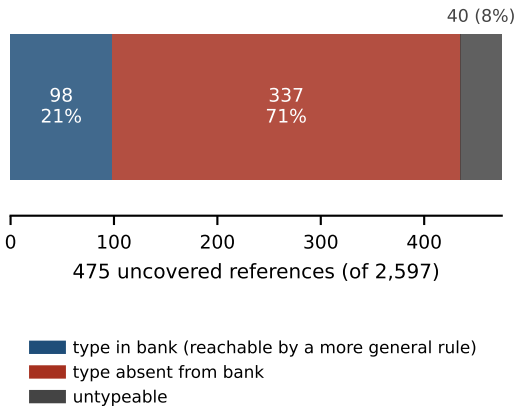



Figure 6: What the bank fails to reach in one step, typed by element-aware changed-bond multiset. Only the first group is reachable by generalising an existing template, which is what bounds the in-bank ceiling at 0.855.

The composition of the gap determines how much of it template generalisation could close. Typing each uncovered transformation by its element-aware changed-bond multiset, and comparing against the 4,417 types the bank already contains, splits the 475 uncovered test transformations into 98 (21%) whose reaction type the bank already has and 337 (71%) whose type is absent from it; the remaining 40 admit no radius-0 type and stay unclassified. The first group is reachable in principle by a more general template, which bounds an in-bank ceiling at $(2122 + 98)/2597 = 0.855$; closing both groups would bound it at 0.985. Template generalisation is therefore a real but tightly bounded gain of 0.038 in ceiling headroom, and seven-tenths of the residual is genuinely out-of-bank chemistry rather than an artefact of template specificity. The bound is conditional on single-step application: it constrains the paradigm, not the problem.

### E.13 TWO INTERVENTIONS

**A factorised generator tests Propositions 2 and 3.** Rebuilding the generator as a dense factorised model, $P(\text{type} \mid s) \cdot P(\text{site} \mid \text{type}, s)$ over a coarse reaction-type vocabulary, turns the observational decomposition into a causal test.

Selection proves fixable. The learned type head beats the frequency prior by roughly threefold at $k = 5$ (recall@5 0.086 against 0.030), and converges to the prior by $k = 10$ (0.298 against 0.309). Site localisation is close to, but below, its reference: 0.78 hit@3 against 0.81. The positive-unlabelled degeneracy of Proposition 2 is therefore an artefact of the target's support, not a fundamental limit, which confirms that proposition's prediction.

Coverage does not bind it. As a replacement generator the pipeline reaches recall@15 $0.489$ $[0.464, 0.514]$ on the full test against $0.339$ for the deployed pipeline, but it emits $14.0$ candidates per substrate against $8.4$, so that pair is not budget-matched and supports no claim. The conservative comparison is: truncated to five candidates it reaches $0.337$, matching the deployed pipeline's $0.339$ while emitting fewer than half as many, and at ten it reaches $0.432$. Both runs were scored separately, not paired, so no interval is quoted on a difference. The convention this is measured in is the one the deployed generator fires rules in; under the expanding one the same model reaches $0.256$ and emits $8.0$, with $46$ substrates receiving no prediction at all against 2 here, which is the size of the artefact rather than a property of the architecture. That the coverage term binds is consistent with Proposition 3 but does not test it: the proposition concerns what one application step can reach, and a type vocabulary is not application depth.

Kept at broad coverage and used only to re-rank, the same signal is worth a small significant margin: recall@15 rises from $0.388$ to $0.404$, a paired $+0.0165$ $[+0.006, +0.027]$ over the full test set. Both are broad-pool figures, not deployed ones. How the heads are trained decides whether that margin appears. Multiplying independently trained type and site probabilities into the rank adds nothing on a matched pool ($+0.0002$, not significant). Fine-tuning the same two heads with a listwise ranking loss against a frozen generator and filter beats that bolt-on by $+0.0089$ $[+0.003, +0.015]$ and the baseline by $+0.0091$ $[+0.003, +0.016]$. Rank-aware training, not the type and site heads themselves, recovers the margin.

Selection is addressable and ranking yields a small margin, but no *generator redesign* tested here matches SyGMa without a broader bank: for this redesign, coverage is what binds. That is not in tension with the summary above or with §A.3, which put the dominant addressable loss at the deployed operating point in selection. The two statements are about different quantities (where the deployed pipeline is losing recoverable recall, and what caps a generator rebuilt to fix selection) and §A.3 gives the reconciled verdict, that selection binds jointly with the cut-off, not alone.

**Set generation with a GFlowNet does not win here, and the reason is cardinality.** We formulated the generator as a set-terminal GFlowNet trained by Trajectory Balance against a positive-unlabelled set-coverage reward $\log R(S) = \beta(\text{TP} - \lambda|S|)$ with an analytic backward policy; the formulation is new for this domain and is released for reuse. Across three seeds on 40 validation substrates its single-set recall@15 is $0.292 \pm 0.032$, indistinguishable from beam search at $0.311 \pm 0.039$ and below a reranker at $0.411 \pm 0.044$. It is reported as a negative result, because the endpoint is one its objective works against: the reward pays a size penalty and recall@$k$ does not charge one. The diagnosis is cardinality rather than mis-specification: $\beta$ cancels in the argmax, so the reward is the *more* conservative of the two, and what happened is that the policy emitted $14.2$ against a configured maximum of $15$, leaving the size term almost no room to act. Set-level training without an explicit cardinality estimate is therefore untested here, not refuted, and predicting set size is the component this architecture was missing.

### E.14 THE OUTPUT BUDGET IS MIS-SET, AND ADAPTING IT PER SUBSTRATE DOES NOT HELP

Section A.5 reports that the truncation is the aggregate's second dial. This gives the measurement behind that, and the negative result that follows from it.

Every method here emits a number of candidates fixed by convention rather than by the substrate, and no budget is selected in what follows: the F1-versus-$k$ curve is swept for each method, the orderings are read off the curves, and the one pre-fixed comparison is made at a $k$ fixed in advance. A budget

chosen against the numbers it is reported on would be optimistically biased, and a reordering could be an artifact of the choice, so the choice is removed instead of corrected.

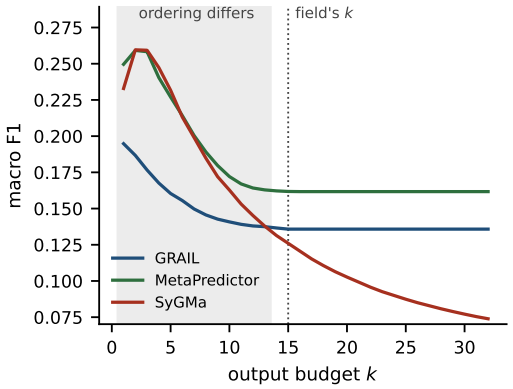



Figure 7: Macro F1 against the output budget. The shaded region is where the ordering differs from the one at the field's $k = 15$, marked by the dotted line.

Three orderings occur across $k \in [1, 32]$. At $k = 1$ it is MetaPredictor, SyGMa, GRAIL; from $k = 2$ to 5 SyGMa leads; from 6 to 13 it is MetaPredictor, SyGMa, GRAIL again; and from 14 on, where the field's 15 sits, it is MetaPredictor, GRAIL, SyGMa. GRAIL is last at every $k \leq 13$. At $k = 5$, fixed before these curves were computed, two of the three pairs separate: SyGMa above GRAIL by $0.071$ $[0.057, 0.086]$ and MetaPredictor above GRAIL by $0.067$ $[0.051, 0.083]$; SyGMa and MetaPredictor do not, SyGMa nominally above by $0.005$ $[-0.006, 0.015]$.

The mechanism is in the same sweep, and it is stronger than a statement about GRAIL. A budget only binds a method when the method emits more than the budget, so at $k = 15$ the comparison is not budget-matched at all: SyGMa is cut from $80.7$ candidates to 15 on $92.7\%$ of substrates, while re-truncating GRAIL's *emitted* list at $k = 15$ removes nothing from any of them, because its emission is already capped there. That last is a statement about re-truncation and not about the pool: the output reaches the cap upstream, on 308 of 1,170 substrates ($26.3\%$), and every reference the decomposition loses between the budgeted pool and the output falls on one of those substrates (§A.3). The convention silently advantages whichever methods emit less than the budget, and it is the budget rather than the ranking that decides. The magnitude is not a rounding effect: SyGMa's F1 at its best budget over $k \in [1, 32]$ is $5.4$ times its F1 scored at the full output it emits.

The binding is partial well below 15 too, which is why the comparison is reported at several budgets, not one. GRAIL is cut on $63.5\%$ of substrates at $k = 5$ and $45.6\%$ at $k = 8$, against SyGMa's $98.8\%$ and $96.5\%$. Against GRAIL, SyGMa is ahead by $0.083$ $[0.065, 0.100]$ at $k = 3$, $0.071$ $[0.057, 0.086]$ at $k = 5$, $0.039$ $[0.026, 0.052]$ at $k = 8$ and $0.022$ $[0.010, 0.034]$ at $k = 10$, every interval excluding zero, so the reversal is not an artifact of one budget inside the region.

**What a compact output is and is not worth.** GRAIL returns $8.4$ candidates against SyGMa's $80.7$, which we have described as an operating point recall@$k$ does not express. The sweep prices it, and against us throughout. Truncation is free: any user preferring a shorter list can take SyGMa's top-$k$, and at every budget up to 13 that shorter list beats GRAIL's, separated from $k = 1$ to $k = 10$. Emitting few candidates therefore confers no advantage that cutting a longer list would not also confer, at any budget below the field's, and GRAIL's lead at $k = 15$ is the cap declining to bind it.

The head of the list is no exception. At $k = 1$ both methods are cut, on $92.8\%$ and $99.6\%$ of substrates, and SyGMa leads there too, by $0.038$ $[0.011, 0.064]$. Every ranked list is scored with the parent removed, on tautomer-InChIKey equality with the substrate, so a method is never credited or charged for returning the compound it was given.

For GRAIL the deployed operating point can also be checked against a fixed cut, because its per-candidate ranking scores are released with the frozen predictions. Over the 1,166 scored substrates,

emitting as deployed scores $0.137$ against $0.196$ at $k = 1$: a paired difference of $-0.058$, 95% interval $[-0.075, -0.042]$.

The obvious next step is to make the cut depend on the substrate, and it does not work. Per substrate the F1-optimal cut $k^\star$ is computable exactly from the ranked list and the annotated set, so the ceiling is known: knowing $k^\star$ is worth $48$ to $76\%$ relative over the best constant across the five methods. Almost none of that is reachable. Cut each list instead at a per-substrate count predicted out of fold from substrate descriptors, and the part of the headroom that recovers is $-10\%$ for GRAIL, $-2\%$ for SyGMa and $+5\%$ for MetaPredictor: for two of the three, predicting the set size does worse than the best constant. The true count is a property of the substrate and the ranking is a property of the method, and out of fold the first carries almost none of the gap.

Cross-fitting confirms this out of sample. A gradient-boosted regressor on twelve substrate descriptors, predicting $k^\star$ and evaluated on held-out folds, recovers $3.2\%$ of the oracle headroom for GRAIL, $4.3\%$ for MetaPredictor and $1.1\%$ for SyGMa; predicting the set size directly falls below the constant baseline for two methods of three. Descriptors are not a graph encoder, so this bounds the direction rather than refuting it, but the mechanism it agrees with is structural: a head reading only the substrate cannot see which of its own predictions are correct, and that is where the headroom lives. The lever the measurement points to instead is a stopping rule on calibrated candidate scores, which is a different object from the abstention that failed in Table 6, that having been a global threshold judged on recall@$k$, where truncation can only lose.

### E.15 A SIMILARITY-TO-TRAINING SPLIT DOES NOT MEASURE TRANSFER HERE

Rule-encoded chemistry is general where learned patterns are not, so a rule-grounded method might be expected to lose less on substrates far from its training data. That would be a claim about the paradigm, not a leaderboard position, and it is testable on frozen predictions by stratifying the test split on maximum Morgan–Tanimoto similarity to any training substrate. It does not survive, and the way it fails is the part worth reporting.

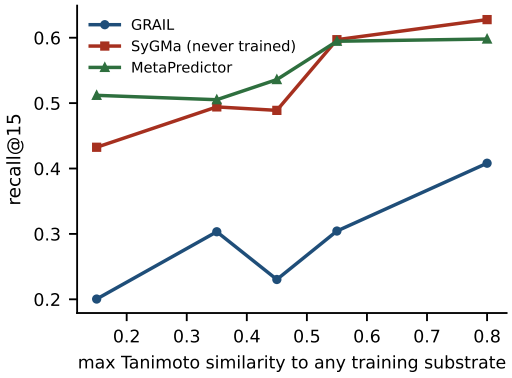



Figure 8: The control that decides the axis. If similarity to training measured transfer, SyGMa's curve would be flat: it is a rule engine that never saw the training split. It falls on a slope this measurement cannot separate from the learned methods', so the axis is measuring how well studied a substrate is.

The design carries a control that decides whether the axis means anything at all. SyGMa is a rule engine that never saw the training split, so if the axis measures transfer its curve must be flat. It is not: SyGMa degrades at $+0.180$ on F1 and $+0.291$ on recall against GRAIL's $+0.158$ and $+0.307$. Resampled on the same draws, so that the comparison is paired, the difference is $+0.022$ $[-0.041, +0.086]$ on F1 and $-0.016$ $[-0.144, +0.115]$ on recall. Those are failures to detect a difference, not demonstrations that none exists, and the width of the interval is the informative part: the F1 interval still admits a gap of $0.086$ and the recall interval one of $0.144$, against slopes of $0.158$ and $0.307$ themselves. What the data support is that a rule engine with nothing to transfer from falls on a slope this measurement cannot separate from a learned method's: not that the two slopes are equal.

A second stratification points the same way and carries its own confound rather than removing it. The fraction of a substrate's references the rule bank reaches at all rises from $0.712$ below similarity $0.3$ to $0.819$ above $0.6$. That is not a quantity untouched by training: $5{,}866$ of the $7{,}581$ rules are mined from the training split, and those are the rules that do the reaching, covering $0.785$ of the references against the curated remainder's $0.471$ (§E.9), and Appendix F says in terms that applying the bank to a substrate it was partly mined from is not independent of training, so a coverage gradient in similarity to training is exactly what a training-derived rule set would produce. It cannot serve as the leg of the argument that no learning touches, and is not used as one; the control above is what carries the section. The common cause is visible in the same stratification: reference sets are smaller at the far end too, $2.40$ annotated metabolites above $0.6$ against $1.90$ below $0.3$, though across all five strata the series is not monotone and the contrast is between its endpoints, not a trend. Training corpus, rule bank and annotation all derive from the literature of well-studied metabolism, so the axis measures how well studied a substrate is and moves all three together.

Read at face value the slopes would say the transformer transfers best, at $+0.106$ and $+0.118$, roughly half the other two. Read correctly they say the measurement cannot support a transfer claim in either direction, and we make none.

The detector is not ours. Buttenschön et al. (2024) stratify a docking benchmark by maximum sequence identity to its training set and run never-trained classical docking alongside the learned methods, which is the same design. Their control comes out *flat*: the classical methods score alike across strata while the learned ones fall on low-identity targets, so there the axis does measure transfer and the degradation is real. That is what makes the instrument worth using rather than a formality: it returns opposite verdicts in the two domains, and a control that always agreed would discriminate nothing.

What is ours is the negative result in this domain and the mechanism behind it. Stratifying by similarity to training is the standard way to argue generalisation in molecular machine learning, and wherever a knowledge base and a training corpus share provenance the axis is confounded with difficulty; Buttenschön et al. (2024) did not need that explanation because their control was flat. Adjacent evidence points the same way without closing it: retrosynthesis accuracy tracks training template frequency even for template-free models (Tanovic et al., 2026), and molecular datasets cover structure space by study intensity rather than uniformly (Kretschmer et al., 2025).

Had we run the learned methods alone we would have concluded from $+0.158$ against $+0.106$ that the transformer is the more robust of the two: a conclusion against our own system, and therefore the kind that attracts least scrutiny, resting on a measurement that cannot carry it. That conclusion would not even have been a marginal one: paired, the transformer's recall slope is shallower than GRAIL's by $-0.190$ $[-0.330, -0.044]$, an interval excluding zero. The axis would have delivered a confirmatory transfer claim from a quantity that does not measure transfer, which is why the control and not the interval width is what decides here.

### E.16 Partial identification of the metric ordering under incomplete annotation

Section A.5 ranks methods by F1, and every precision figure feeding it is a lower bound: a predicted structure that is a genuine metabolite nobody recorded is charged as a false positive. The threat is not symmetric. The method the metric demotes is the one emitting eighty candidates per substrate, so it is the one with the most room for unrecorded hits, and an ordering that only holds because the references are incomplete would be an artifact, not a result.

Model the annotation as a uniform propensity: for a substrate with $T$ true metabolites, each is recorded independently with probability $c$. Fix a prediction set of size $n$ covering $t$ of the true metabolites. Then $\mathbb{E}[\mathrm{TP}_{\mathrm{obs}}] = ct$ and $\mathbb{E}[|\text{references}|] = cT$, so

$$R_{\mathrm{obs}} = \frac{ct}{cT} = \frac{t}{T} = R_{\mathrm{true}}, \qquad P_{\mathrm{obs}} = \frac{ct}{n} = c\,P_{\mathrm{true}}, \qquad F1_{\mathrm{obs}} = \frac{2ct}{n + cT}.$$

Three consequences follow. Recall is unbiased, so the field's reliance on it is defensible on this ground. Precision is the truth scaled by a factor common to every method, so while its level is unidentified its *ordering* is identified for any $c$. F1 is neither, since $c$ enters numerator and denominator differently, and its ordering is therefore a function of $c$ that has to be checked rather than assumed.

We check it by correcting each per-substrate precision to $\min(P_{\text{obs}}/c,\, 1)$, leaving recall untouched, and recomputing macro F1 over the full split. The cap makes the correction conservative in the direction that matters: a method cannot be credited past certainty, and the method with the lowest observed precision is the one that benefits from the correction longest.

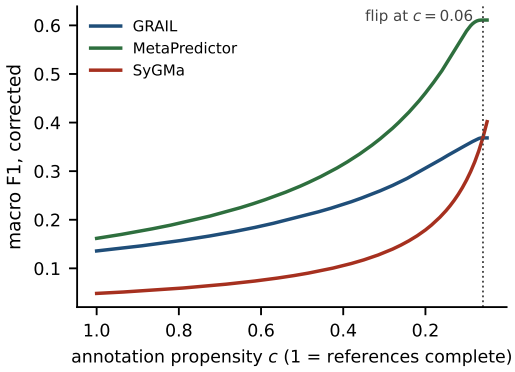



Figure 9: Macro F1 with each per-substrate precision corrected to $\min(P_{\text{obs}}/c, 1)$, swept over the annotation propensity. The reported ordering holds until $c = 0.06$, where SyGMa passes GRAIL; at that rate a substrate with three recorded metabolites would have fifty in truth.

| annotation propensity $c$ | GRAIL | MetaPredictor | SyGMa |
|---|---|---|---|
| 1.00 (as reported) | 0.136 | 0.162 | 0.048 |
| 0.75 | 0.163 | 0.202 | 0.062 |
| 0.50 | 0.208 | 0.271 | 0.088 |
| 0.30 | 0.263 | 0.374 | 0.132 |
| 0.20 | 0.306 | 0.461 | 0.179 |
| 0.10 | 0.354 | 0.589 | 0.282 |
| 0.05 | 0.368 | 0.611 | 0.402 |

The reported ordering survives to $c = 0.06$, where SyGMa overtakes GRAIL; MetaPredictor is never overtaken on a grid running down to $c = 0.05$. At $c = 0.055$ a substrate carrying three recorded metabolites would have about fifty-five in truth, which the references do not support.

The same model bounds what any set metric can measure. A predictor that emitted exactly the true metabolite set would still be scored against the annotated subset, so with completeness $c$ it measures Jaccard $c$ and F1 $2c/(1 + c)$ however good it is. At $c = 0.5$ that is a Jaccard ceiling of $0.5$ and an F1 ceiling of $0.667$; at $c = 0.3$, $0.30$ and $0.462$. Set-agreement endpoints are therefore bounded by annotation completeness in a way recall is not, which is a reason to report them beside recall, not in place of it, and a caution for any method trained directly on set agreement: optimising it on incomplete references trains partly on which metabolites were recorded.

The assumption doing the work is uniformity. A propensity that fell systematically with how implausible a product is would favour the wide-output method more than this correction allows, because its extra candidates are disproportionately the implausible ones. That is the direction in which the argument is weakest, and we state it rather than test it: estimating a structure-dependent propensity would need exactly the exhaustively annotated references whose absence creates the problem.

### E.17 The prescribed remedy for the selection pathology does not work

Section E.10 identifies rule selection as extreme multi-label classification and cites the literature in which propensity-scored losses were introduced for exactly that regime (Jain et al., 2016). Naming a pathology without trying its standard treatment would leave the diagnosis untested, so we ran the treatment.

Each positive rule label is weighted by the inverse of its estimated propensity, fit from the rule's positive count by the model of Jain et al. (2016) with $A = 0.55$, $B = 1.5$, and normalised over rules that carry a positive so the loss keeps its scale. Across the $7{,}581$ rules the weights run from $0.196$ to $1.669$ with a mean of $1.377$, and $81.5\%$ are up-weighted: most of them the $4{,}271$ that never carry a positive at all, which take the maximum weight, against $0.83$ for a rule with five positives. The training configuration is otherwise identical to the deployed one, differing in three fields, and the generator trained to completion over eight epochs with validation loss falling monotonically from $0.530$ to $0.472$.

It does not help. On the same 245 substrates, under the same protocol and with the same deployed filter, the learned-only selector reaches recall@15 of $0.249$ against $0.266$ under constant weighting. Its deficit to the frequency prior widens from $-0.144$ $[-0.196, -0.095]$ to $-0.161$ $[-0.215, -0.107]$; both are significant and the intervals overlap heavily, so the change itself is not distinguishable while the absence of improvement is clear. The prior-only arm returns $0.410$ in both runs, identical to four decimal places, which confirms the comparison isolates the training objective. Per-substrate scores from the two runs were not retained jointly, so no paired interval is quoted on the $-0.017$ difference itself.

The intervention was also run end to end, not only in the isolated probe, and there the discipline this paper argues for is what decides it. Retraining the deployed configuration with the propensity objective and evaluating it as deployed gives $0.348$ against $0.334$ on test, which read alone looks like a gain of $+0.014$. On validation, the split model selection is permitted to touch, the same comparison gives $0.343$ against $0.344$, a difference of $-0.001$. The test-split movement is a single seed against a seed-to-seed spread of $\pm 0.010$ and is not corroborated by the split available for selection, so it supports no claim. Both figures are on the 300-substrate evaluation that configuration runs and are not comparable with the full-split numbers elsewhere in this paper.

The conclusion concerns the domain and not the method. Inverse-propensity weighting exists to correct a label's rarity when that rarity is an artifact of incomplete annotation. In metabolism a rule's firing frequency is substantially not an artifact: common rules are common because the transformations they encode are common in the chemistry. Down-weighting them by a factor of five, which is what the fitted propensities do at the head of the distribution, therefore discards signal rather than correcting bias. This is a candidate explanation for the original finding, not a demonstration of it: separating the annotation and chemistry components of rule frequency would need an exhaustively annotated reference set, which is the same object whose absence creates the positive-unlabelled problem throughout this paper.

# F External validity of the coverage ceiling

## F.1 Internal and external coverage ceilings

The internal coverage ceiling is $0.817$, with a cluster-bootstrap 95% confidence interval of $[0.794, 0.839]$ over 10,000 resamples of substrates. It is a micro estimate, the pooled ratio of summed hits to summed references, not a mean over per-substrate ratios: pairs within a substrate are dependent, so a mean of ratios would misrepresent the uncertainty.

Recomputing the same quantity on an external set requires the same procedure throughout: the 37 GLORYx parent substrates, one uncapped depth-1 pass over the full bank in the same hydrogen convention, no cap on the candidate pool, and the same tautomer-aware match. Under those conditions the external ceiling is 0.709, with a 95% interval of $[0.616, 0.804]$, wide by design at this sample size.

A figure of 0.3715 appears elsewhere for the external ceiling. It is measured with a capped candidate pool and reflects the generator budget of that evaluation, not the reach of the rule bank, so it is not comparable to the internal ceiling and should not be read as an external ceiling.

## F.2 The external ceiling is not inflated by the drugs GRAIL trained on

Twenty-four of the 37 GLORYx parents are substrates in GRAIL's training or validation split, and most of the bank is mined automatically from annotated pairs in the training split. Applying the bank to a drug it was partly mined from is therefore not independent of training, whatever is true of the

absence of a neural network at inference. An adversarial reading of §A.4 raised exactly this, and it is right about the mechanism.

It is not right about the magnitude, and that is a measurement. Splitting the per-parent coverage counts stored for the external ceiling gives $0.719$ $[0.607, 0.842]$ on the twenty-four seen parents against $0.684$ $[0.544, 0.841]$ on the thirteen unseen, a difference of $+0.035$ $[-0.157, +0.222]$. The ceiling is therefore reported on all 37.

At these sample sizes that interval admits an inflation of nearly $0.2$ in either direction, so this is a failure to detect, not a demonstration of absence. What it rules out is the strong form of the objection, that the external ceiling is largely an artifact of mining from these drugs' own annotated pairs, and it does so on the only split available. It does not license the claim that no contamination exists.

### F.3 The criterion is implementation-dependent

Tautomer canonicalisation is not a mathematical operation on a molecule but a search with a budget. On the external references the enumerator reaches a canonical form for $191$ of $196$ unique metabolites and stops on its transform limit for the remaining $5$, $2.55\%$, which are left at whatever form the search reached. The canonical tautomer is also not stable across releases of the toolkit. Neither fact undermines the comparisons here, since every method is mapped through the same implementation at the same version, but both are reasons to report the implementation and its version alongside the criterion, which is the practice this protocol argues for.

### F.4 Composition and the internal–external gap

Both ceilings above are micro estimates. A separate question is whether the gap between them reflects the composition of the two populations: GLORYx parents are on average larger and more conjugated than the internal test substrates, and the spread of mass differences across the uncovered set reported in the main text suggests coverage should fall with molecular complexity.

The regression takes five molecular descriptors, namely molecular weight, ring count, aromatic-atom count, heteroatom count, and a count of sites available for phase-II conjugation, together with the number of reference metabolites, which is a property of the annotation rather than of the molecule. Fit to per-substrate coverage across both populations, it predicts macro coverages of about 0.858 internally and 0.808 externally in sample. These are macro quantities, per-substrate means, and are not directly comparable to the micro ceilings above. The in-sample external prediction of 0.808 sits above the measured external macro coverage of 0.779, so the in-sample fit overstates transferability.

Holding the external substrates out of the fit entirely, and predicting them from an internal-only model, gives 0.813 out of sample and recovers 58% of the gap in macro coverage between the two populations. That out-of-sample figure, not the in-sample one, is the quantity we report. It is a point estimate from a six-predictor model fit on the internal substrates and averaged over the 37 external ones, and it carries no interval; it indicates a direction, not a magnitude that would transfer.

### F.5 Interpretation

At $n = 37$ this is a suggestive partial composition effect. It is consistent with molecular composition accounting for part of the gap between internal and external coverage, and it is not a fitted law that transfers to other populations. It does not indicate a defect in the rule bank or in the matching protocol.

## G Released negative results

The main text locates the dominant recall shortfall at rule selection. A natural question is whether the granularity of the rule bank itself is responsible: the bank is assembled from several published sources and by automated mining, so it may contain functionally redundant rules that dilute the selection signal, and rules that never contribute and merely enlarge the label space. We tested three hypotheses of this kind. Each was specified together with the observation that would refute it, before the measurement was run, and none survived. They are reported here because the question

arises naturally from the decomposition, and because two of the three outcomes constrain what this architecture can be expected to achieve.

### G.1 Redundancy in the rule bank

Redundancy is present and measurable. Applying the full bank to a sample of 1,000 substrates, 2,747 rules fire at least once, and these fall into 2,056 distinct classes when rules are grouped by the exact set of substrate-product pairs they generate. A quarter of the firing rules are therefore functional duplicates of another rule on this chemistry. Splitting that sample in half and re-deriving the grouping on each half preserves 97.5% of the equivalences, so the grouping is not an artefact of the particular substrates sampled. A second, distribution-free procedure that canonicalises each reaction template and removes exact duplicates eliminates 7.5% of the whole bank. The two procedures can be compared where both apply, on the rules that fire: permutation variants account for 293 of the 691 functional collapses there, $42\%$, and equivalences no canonicalisation can see for the rest.

Whether removing this redundancy would improve selection is not established. Training marks every rule that produces an annotated metabolite as positive, and each rule carries its own embedding, so duplicated rules divide the positive signal for one transformation among several parameters. Against this, the noisy-or aggregation used at inference pools the duplicated scores again, which partially restores what training divided. The ablation reported below shows in addition that the per-rule embedding component is not load-bearing, which removes the mechanism by which the divided signal would matter. We therefore treat deduplication as a housekeeping measure affecting startup cost and memory, not as a route to better accuracy.

### G.2 Rules that never contribute

Across the 4,787 training substrates, 56.3% of the bank is never a positive label, and a further 20.1% is positive for exactly one substrate. Roughly three quarters of the 7,581-dimensional target is therefore either inert or supported by a single example, which is consistent with the extreme multi-label characterisation in the main text and suggests that the effective sparsity of the learning problem is reducible.

It does not follow that the corresponding rules can be removed. We tested removal directly, by discarding the 4,271 rules that are never a positive label in training, which reduces the bank from 7,581 to 3,310, and re-ranking on the validation split. Performance falls on both axes. Recall@15 declines from 0.388 to 0.352 (paired difference $-0.036$, 95% confidence interval $[-0.050, -0.024]$), and the fraction of references present anywhere in the generated pool declines from 0.439 to 0.378 ($-0.061$, $[-0.078, -0.046]$); both intervals exclude zero. The mean pool size falls from 46.9 to 40.0 candidates.

Rules carrying no positive label in training therefore still generate correct metabolites for substrates outside the training distribution. What appears as dead weight in the training target is in part generalisation capacity held in the tail of the bank. The distinction governs how the earlier statistic should be read: three quarters of the *target* is inert during training, but the bank is not correspondingly prunable, and treating the two as equivalent removes coverage on which the deployed system depends.

### G.3 The per-rule embedding component

Each rule's representation is the sum of a graph encoding of its reaction template, a learned per-rule embedding indexed by position in the bank, and a small vector of structural features. Decomposing the variance across rules attributes 82% to the per-rule embedding and 17 per cent to the graph encoding, which suggests that the model treats the bank largely as a lookup table. On that reading the rare rules, whose embeddings receive almost no gradient, would contribute noise to the selection score, and the extreme multi-label pathology would have an architectural explanation.

That reading does not hold. Setting the per-rule embedding to zero at inference, which removes the component carrying 82% of the variance, changes recall@15 by $-0.007$ on the test split. Repeating the ablation on the validation split with paired bootstrap intervals over 994 substrates gives no significant change at any cut-off: $+0.0011$ $[-0.0086, +0.0107]$ at $k = 1$, $+0.0055$ $[-0.0074, +0.0181]$ at $k = 5$, $-0.0116$ $[-0.0266, +0.0036]$ at $k = 15$, and $-0.0119$ $[-0.0282, +0.0044]$ at $k = 30$. Dom-

inance of a component in the variance of a representation does not imply that the component carries the decision. A single test-split run on 291 substrates had indicated that removing the embedding improved recall at small $k$ ($+0.037$ at $k = 1$, $+0.027$ at $k = 5$); that effect did not replicate on validation and is not reported as a result.

### G.4 Dispatching the hydrogen convention per template

§E.5 shows that what a bank reaches depends on a convention its engine applies globally, and that banks disagree about which convention they want. The obvious repair is to stop having a global one and apply each template under the convention it was written in. The policy was fixed before anything was measured (a template gets the expanded substrate if and only if its reactant side carries a bracket-hydrogen atom, blind to the product side, to the hydrogen-count primitive and to negations, with the 238 templates whose recursive SMARTS that test cannot see inside taking their own bank's majority convention) because a policy chosen after seeing which setting wins would beat any global setting by construction and would measure nothing.

The quantity that decides the claim is the residual rather than the win: reach under dispatch minus the better of the two global settings. A residual at zero says the bank is single-convention and choosing correctly is the whole story. Above zero says no global setting expresses the bank.

| bank | rules | dispatched | dispatch | best global | residual |
|---|---|---|---|---|---|
| SyGMa | 175 | 0 | 0.535 $[0.510, 0.561]$ | 0.535 | $+0.000$ $[0.000, 0.000]$ |
| ours | 7,581 | 675 | 0.8229 $[0.8009, 0.8445]$ | 0.8194 | $+0.0035$ $[0.0015, 0.0058]$ |
| BioTransformer | 994 | 726 | 0.4925 $[0.4669, 0.5182]$ | 0.4963 | $-0.0038$ $[-0.0067, -0.0015]$ |

One prediction was structural and two hold, and what the residual is measured against decides the second of those. SyGMa is the structural null: none of its templates is dispatched, so dispatch is the identity map and has to reproduce its implicit-hydrogen arm to four decimals, which it does, and that is the check that the classifier is not sending templates somewhere it should not. BioTransformer was predicted to clear its better global setting, because 726 of its 994 templates want the convention a single global arm would deny them: the 690 that name a hydrogen atom on their reactant side, and the 36 recursive patterns the classifier cannot see inside, which take the bank's majority convention by the rule registered before the run. It does not clear it. It loses by $-0.0038$ $[-0.0067, -0.0015]$, a paired difference measured in one pass over the same substrates rather than a gap between marginals, which is the wrong test for a within-bank contrast. Those templates do add products, and the products are real; almost none of them recovers a reference these substrates carry. Our own bank does clear it, by $+0.0035$ $[+0.0015, +0.0058]$. Every bank in the table records the revision that measured it and a released check refuses a citation of one whose revision does not match the others, because an entry produced by an earlier revision of the loop is exactly the mixture this appendix is about.

What "better global setting" means decides the second result, and getting it wrong is how this prediction was first reported as holding. An arm that expands the substrate and never contracts the product is not a setting anyone would choose; §A.2 calls it an unfinished loop. Two defects followed from it. Measured against that arm, the residual is a comparison with a program that has a bug in it. And the dispatch arm itself was built from the uncontracted half of its own enumeration, so a repaired control was being compared against an unrepaired treatment. With both sides completed the prediction fails, which is what the appendix reported before either defect was found, though not for the reason it gave. A comparison in which one side has been repaired and the other has not measures the repair.

The conclusion is close to the recommendation §A.3 already makes: declare the convention. Dispatching it per template is not worth having on either published bank tested here, being the identity on the single-convention one and a loss on the mixed one, and it pays only on ours. The failure is the useful half: a bank can be mixed in its syntax and single-convention in what it reaches, and the one bank where dispatch pays is the one whose 675 dispatched templates were written by hand for it. For a bank written in one convention there is nothing for a cleverer engine to recover, and SyGMa is the demonstration that the classifier knows the difference. Where our own residual comes from was registered as a further prediction and that one failed: it is not created by the bank being mixed but by

the 675 hand-written templates that want an expansion a global setting denies them, and it is *larger* on those templates alone than on the bank containing them (Appendix E.9).

### G.5 Evaluation caveat

The exploratory measurements in this appendix were scored on the test split before the discipline described in the main text was tightened; the pruning and ablation results were subsequently repeated on validation, and those are the values reported above. They support no claim in the main text.

**Estimating a leaderboard's fragility from the published table alone.** The instrument of §3 needs released prediction files, and most tables come with none. A pair reorders when a choice moves the two systems apart by more than the margin between them, and the margins are printed, so the share of pairs whose margin falls below a domain-level movement should predict the share that reorder. It does not. On the two retrosynthesis leaderboards the threshold rule predicts 21 fragile pairs against 16 observed, and replacing the single movement with its own distribution and summing the expected count gives 20.7: both over-predict by about a third, and by the same third on each, because a movement large enough to cross a margin need not be in the direction that flips it. A calibration factor fitted on two leaderboards would be an overfit, so none is offered. The instrument needs the predictions, which is one more reason for releasing them.

**Forecasting how many metabolites a substrate has, from the substrate.** If the emitted set should be sized per substrate, the obvious way is to predict the size. Least squares on heavy atoms, rings and rotatable bonds, fitted on the training split and applied to test, reaches 0.181 macro F1 against the deployed policy's 0.136 – and against 0.195 for simply emitting one candidate to every substrate. The forecast is worse than the constant it was meant to beat, so the count is not recoverable from the substrate's own descriptors. What works is relative: reading the candidate pool rather than the molecule (§A.5).

## H Limitations and scope

### H.1 Which boards the survey could have had, and why it has these

The composition of the survey invites a fair question: nineteen of the twenty-three boards come from one shared task and four from chemistry. What follows records the search for more, so that the composition is evidenced rather than asserted, and so that a reader can repeat it.

The binding constraint is not the criterion or the second axis. It is that a board must have released its *per-item* predictions, and most benchmarks never collect them. The Therapeutics Data Commons instructs a submitter to use its own evaluator to compute the performance of a model on the test set and to submit that performance, so the predictions exist only on the submitter's machine. The Open Graph Benchmark enumerates the fields a submission must carry, and every one is metadata or an aggregate: the test performance is a mean and a standard deviation over ten seeds. MoleculeNet has no submission process at all. In each case the leaderboard is a table of numbers whose inputs were never kept, which is a stronger statement than that the files could not be found.

Two further families are excluded by the rules of §D rather than by availability, and both exclusions are the one the molecular-generation board receives. In property and yield prediction the output is a number, so the only candidate criterion is a numeric tolerance, which is not a candidate answer to whether two objects are the same; the criterion axis exists only where the output is a structure. On the materials-discovery board the headline counts materials below a stability threshold, and that threshold is applied to the reference as well as to the prediction, so moving it rewrites which materials are stable rather than when a prediction matches one. A criterion changes when a prediction counts as hitting the reference; it never changes what the reference is.

The search found one further collection that qualifies: five systems with ranked predictions on an extrapolation split of the same corpus, sharing 4,948 targets, which is a further 10 pairs. Its items are not disjoint from the boards already surveyed, sharing 507, 484 and 467 targets with the three of them, so the families the union correction runs over are dependent. Holm is valid under arbitrary

dependence, so the correction stands; the overlap is reported because a reader is entitled to know the families are not disjoint rather than to assume it.

### H.2 Recall by criterion on the external set

Table 11: Recall@15 by matching criterion on the external GLORYx set ($n = 37$), the lower panel of Table 2. †GRAIL trained on 19 and validated on 5 of the 37 drugs, so its row is computed on the 13 it has not seen.

| **method** | InChIKey | no-stereo | tautomer | stereo step (95% CI) |
|---|---|---|---|---|
| SyGMa | 0.492 | 0.498 | 0.498 | $+0.007$ $[+0.000, +0.020]$ |
| BioTransformer | 0.346 | 0.374 | 0.374 | $+0.028$ $[+0.002, +0.064]$ |
| MetaPredictor | 0.362 | 0.504 | 0.504 | $+0.142$ $[+0.079, +0.215]$ |
| **GRAIL**† | 0.059 | 0.185 | 0.223 | $+0.126$ $[+0.028, +0.239]$ |

**The exchange in full.** Under canonical matching the leaders run GraphRetro $0.5377$, LocalRetro $0.5339$, Retroformer $0.5325$, Graph2SMILES $0.5293$; disregard stereochemistry and they run Graph2SMILES $0.5571$, Retroformer $0.5541$, GraphRetro $0.5501$, LocalRetro $0.5469$, so fourth becomes first. The family is twenty-one system pairs by six pairs drawn from the four conventions by four budgets, which is $504$ cells less the $56$ in which two conventions return identical results and no test exists. Stereo-blind matching is a candidate convention here and not a straw one: GLORYx scores on a stereo-blind skeleton, and neither GRAIL nor SyGMa emits an assigned stereocentre anywhere in the frozen predictions scored here, in $4{,}009$ and $4{,}001$ candidates scanned: the corpus standardises its substrates without stereochemistry, so there is none to carry through (Figure 11). That is a fact about this corpus and not a limitation of either method.

### H.3 The criterion at a matched output size

The criterion and the budget could be the same axis wearing two names, since a method that emits more has more chances to be rescued by a tolerant rule. Truncating each method to the same emitted count separates them. At about eight emitted candidates GRAIL reaches $k = 14$ and the two comparators $k = 8$, giving $8.13$, $7.91$ and $7.89$ emitted; the gains from canonical to tautomer matching there are $0.0161$ $[0.010, 0.023]$, $0.0356$ $[0.027, 0.045]$ and $0.0408$ $[0.031, 0.051]$, and the paired differences against GRAIL are $-0.0195$ $[-0.029, -0.010]$ against SyGMa and $-0.025$ $[-0.035, -0.015]$ against MetaPredictor. Both exclude zero, so the criterion still separates these methods once output size is held fixed and it is not the budget in disguise. This arm also removes a mismatch in how the main text reads the two axes together, which quotes each method's gain beside its uncapped emitted count: at $k = 15$ SyGMa emits $14.53$ of the $80.83$ it would emit uncapped.

### H.4 GRAIL does not win on recall

GRAIL loses on recall everywhere in this paper. Against SyGMa the margin is significant and separated: a paired difference of $-0.242$ in mean per-substrate recall, 95% confidence interval $[-0.271, -0.212]$. A McNemar test on a different endpoint, whether each substrate yields any hit at all within the top 15, rejects at $p \approx 1.7 \times 10^{-44}$; it corroborates the direction, not testing the margin. It also loses to the learned transformer baselines on the shared subset of $150$ substrates, where MetaPredictor reaches 0.585 and MetaTrans 0.561 against GRAIL's 0.365 under the same tautomer-aware protocol. Nothing here should be read as a claim to the state of the art. What the paper contributes is the coverage ceiling, the decomposition, and the matching protocol; GRAIL is one diagnosed row within them.

### H.5 Precision understates performance

Every precision number in this paper is a lower bound. Negatives are rule-applicable products that carry no annotation rather than confirmed non-metabolites, so an unannotated true metabolite counts as a false positive, and the size of that understatement is unknown. We lead with recall and with the

mean size of the output set for that reason. Precision and F1 nonetheless appear in §A.5, because under a uniform annotation propensity the understatement is a common factor that leaves the precision ordering identified, and the F1 ordering survives correction down to an annotation rate the references make implausible.

### H.6 THE RECOMMENDED CRITERION CANNOT PENALISE A LIMITATION OF OUR OWN SYSTEM

The default criterion disregards stereochemistry, and GRAIL's own standardisation writes SMILES without stereochemistry, so it never emits a stereocentre. Under a stereo-sensitive criterion GRAIL would be scored on a representation it cannot produce; under the default it is not. On the external set this is the whole of the difference, and it is large: on the thirteen drugs GRAIL has not trained on, it rises from $0.059$ to $0.185$ when stereochemistry is disregarded, against a further $0.038$ from tautomer canonicalisation. (The corresponding figures on all 37, $0.155$ to $0.310$, are inflated by training overlap and are superseded below.) GRAIL's external figure is therefore conditional on a tolerance that happens to suit its output convention. The comparisons in this paper remain valid, because every method is scored under the same criterion and all three rungs are reported, but a system that does emit stereochemistry is not rewarded for doing so anywhere in this evaluation.

### H.7 THE POPULATION IS UNDECLARED, AND DOES NOT SURVIVE BEING TESTED

Two of the four choices this paper names move an ordering and a third moves a reported quantity. The fourth is the one we could not show does either, and separating those statements is the whole point of testing it rather than asserting it from the others.

That it is undeclared is not in doubt. Eleven released files sit on three test sets sharing about a tenth of their reactions (§H.21), our own copy of a familiar name is one of the deviations, and fifteen products that appear in two of those files carry a *different* recorded reaction in each: a disagreement about the answer and not merely about the sample. A number quoted against one of those names cannot be read without knowing which set produced it.

Whether it *reorders* is a separate question with an estimand, and the estimand needs care. The smaller population is nested inside the larger, so a gap measured on the subset and a gap measured on the whole are not independent and their difference carries no honest interval. The complement, what the larger population has and the subset does not, is disjoint from the subset, so the contrast between subset and complement does. Reorderings below are reported against the whole population, which is the quantity a published table states; interactions are tested against the complement, which is the contrast that admits an interval.

Both instances are run this way. On our split the same three methods are scored under the same five criteria by the same estimator on the $150$ substrates all five methods share and on the full $1{,}170$: of $60$ comparisons $10$ reorder, and they are not scattered: MetaPredictor against SyGMa on recall reverses under every one of the five criteria, ahead by about $0.02$ on the shared subset and behind by about $0.02$ on the full split, with precision reordering four and F1 one. On the released retrosynthesis files the two clusters share $505$ products, $490$ of which agree on the recorded reaction; those $490$ are nested inside each cluster's own five thousand, giving $380$ comparisons of which $44$ reorder.

**Not one of the $440$ interactions survives Holm at $0.05$, and on our split not one has an interval excluding zero even before correction.**

**What the null could have detected.** A non-detection is a statement about the design as much as about the data, so the size it would have caught is computed rather than asserted. For each interaction the paired bootstrap already yields a sampling distribution; its standard deviation $\hat{\sigma}$ is the standard error of that interaction, estimated from the same resamples that give the interval and requiring no variance model beyond them. The two populations enter the contrast through disjoint sets of items, the shared subset and the complement, so the resampling is independent across them and no inter-arm correlation has to be assumed; where the two overlap, as the subset does with the whole, the contrast is not used. A two-sided test at the family's own Holm threshold rejects when the estimate exceeds $z_{1-\alpha/2m}\,\hat{\sigma}$ with $\alpha = 0.05$ and $m$ the family size, and reaches power $1-\beta$ against an effect of

$$\mathrm{MDE} \;=\; \left(z_{1-\alpha/2m} + z_{1-\beta}\right)\hat{\sigma},$$

evaluated at $\beta = 0.2$. With $m = 380$ on the released files the critical quantile is $z = 3.823$, and with $m = 60$ on our split $z = 3.341$. The median test in each family then detects $0.100$ and $0.066$, and the least sensitive $0.116$ and $0.202$. The median describes a typical comparison and the maximum is what the family-wide statement needs, since a claim that no interaction of a given size is present anywhere must hold at the blindest test in the family, so the bound this appendix asserts is the second pair. Against the criterion differentials that survive correction, $0.024$ internally and $0.120$ on the shared subset, the first bound is nearly five times the smaller and the second is of the same order as the larger. So these designs exclude an effect several times the size of the criterion's smaller interaction and are silent about one the size of its largest. That is the sense in which the null is bounded, and it is the only sense claimed. On the released files $18$ of $380$ have intervals excluding zero, which is about what $380$ tests produce by chance. The reorderings are what a subset of four hundred and ninety reactions, or a hundred and fifty substrates, does on its own.

So the axis is reported as a hazard to declaration and not as a lever on ranking. Two consequences follow. Reorderings obtained by re-scoring on a subset are not evidence that the population matters, although they resemble such evidence; and the criterion result is not explained away by the same mechanism, since it was measured with a certified interaction on the same estimator and $110$ of its $448$ tests survive the same correction.

The shared reactions do settle one thing no released table can. The three clusters have no system in common, so no published number compares a system in one against a system in another. On the $490$ they share, all ten are comparable at once: `gln` leads at $0.574$ top-1 under canonical matching, ahead of `chemformer` at $0.567$ and `gta` at $0.551$. On its own cluster's five thousand reactions `gln` is fifth of seven at $0.525$. We state this ordering and decline to defend it, for exactly the reason above: it is measured on four hundred and ninety reactions, and nothing here separates it from the same noise.

### H.8 Which number is which

Recall figures in this paper differ by population, aggregation, truncation and whether duplicate predictions are collapsed, and several coincide in value without being the same quantity. Table 12 states each one. Two readings follow from it directly. GRAIL's untruncated and $k$=15 figures are identical because it never emits more than fifteen, so the budget does not bind it. The SyGMa–MetaPredictor gap all but closes between the two rows, SyGMa ahead by $0.023$ untruncated and by $0.002$ at $k$=15, neither margin separated. That is the budget parameter of §A.5 acting on the leading pair rather than on ours.

Table 12: Every recall figure this paper reports for the three full-split methods on a distinct footing of population, aggregation and truncation, with what distinguishes it. A row whose second cell names a file rather than a section is a figure the released artifact carries and the paper does not print. I-TAUT is INCHIKEY-TAUTOMER throughout.

| figure | where | $n$ | aggregation | truncation | collapsed |
|---|---|---|---|---|---|
| GRAIL $0.265$ | Figure 5 | 1,170 | micro | $k$=15 | no |
| GRAIL $0.339$ | §H.10 | 1,170 | macro | $k$=15 | no |
| GRAIL $0.340$ | Table 2 | 1,170 | macro | $k$=15 | yes |
| GRAIL $0.344 \pm 0.010$ | §H.10 | 1,170 | macro, 3 seeds | $k$=15 | yes |
| GRAIL $0.365$ | Table 13 | 150 | macro | $k$=15 | yes |
| GRAIL $0.318$ | Appendix J | 150 | micro | $k$=15 | yes |
| SyGMa $0.592$ | `set_metrics_by_criterion` | 1,170 | macro | none | yes |
| SyGMa $0.571$ | Table 2 | 1,170 | macro | $k$=15 | yes |
| SyGMa $0.572$ | §E.10 | 1,168 common | macro | $k$=15 | yes |
| SyGMa $0.554$ | Table 13 | 150 | macro | $k$=15 | yes |
| MetaPredictor $0.569$ | Table 2 | 1,170 | macro | either | yes |
| MetaPredictor $0.585$ | Table 13 | 150 | macro | $k$=15 | yes |

### H.9 WHICH GRAIL NUMBER IS WHICH

Several figures appear and are not interchangeable. On the full split, 0.340 is the mean of per-substrate recalls after duplicate predictions are collapsed by the matching criterion, which is what the harness does for every method. 0.339 is the same mean over the output as emitted, without that collapse. 0.265 is the pooled ratio of hits to references, lower because substrates with many references weigh more. The decomposition uses the pooled uncollapsed form, the only one that multiplies out exactly.

### H.10 THE HEADLINE IS ONE CHECKPOINT

The deployed headline, recall@15 of 0.265 micro and 0.339 macro, comes from a single checkpoint, not an average over seeds. The correction bounds variance in the evaluation (resampling and discordance over a fixed model), not variance in training across independently seeded runs. Retraining on three seeds confirms the headline is stable, at $0.344 \pm 0.010$ macro and $0.269 \pm 0.006$ micro, but the headline itself remains a single-checkpoint number.

### H.11 THE FIVE-METHOD COMPARISON MIXES SAMPLE SIZES

Table 13: recall@15 by method and matching criterion on the shared internal subset ($n = 150$), with mean number of predictions returned per substrate.

| **method** | canon. | InChIKey (strict) | no-stereo (GLORYx) | Tanimoto=1 (fingerprint) | tautomer (ours) | mean outputs |
|---|---|---|---|---|---|---|
| **GRAIL** | 0.356 | 0.357 | 0.358 | 0.356 | 0.365 | 8.7 |
| BioTransformer | 0.315 | 0.435 | 0.439 | 0.315 | 0.444 | 10.8 |
| SyGMa | 0.514 | 0.547 | 0.548 | 0.514 | 0.554 | 74.2 |
| MetaTrans | 0.523 | 0.494 | 0.561 | 0.524 | 0.561 | 12.7 |
| MetaPredictor | 0.531 | 0.570 | 0.578 | 0.532 | 0.585 | 11.2 |

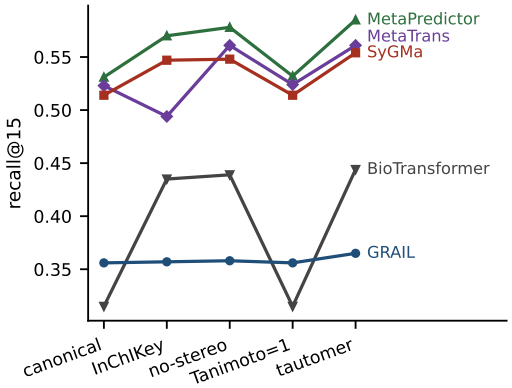



Figure 10: Recall@15 of five methods under the five matching criteria, on the shared subset. Predictions are identical across criteria.

The two reversals the main text reports on this subset rest on differences that are individually indistinguishable from zero at $n = 150$. GRAIL against BioTransformer is $[-0.044, +0.122]$ under canonical matching and $[-0.166, +0.007]$ under tautomer; MetaTrans against SyGMa is $[-0.064, +0.081]$ under canonical and $[-0.130, +0.023]$ under strict InChIKey. What the paper claims is the interaction that produces them; the reversals themselves illustrate it and are not evidence on their own. Of the two interactions, only GRAIL against BioTransformer survives the correction at the family the search spanned (Appendix H.16); MetaTrans against SyGMa is reported as an uncorrected interval.

The three additional comparators (BioTransformer, MetaPredictor and MetaTrans) are scored only on the shared subset of 150 substrates for which frozen third-party predictions exist, whereas GRAIL's and SyGMa's headline numbers elsewhere are reported on the full clean test split of 1,170. Within the comparison table itself all five methods are re-scored on the same 150 substrates, so the columns are paired on identical inputs.

MetaPredictor is also scored on 1,169 of the test substrates, at recall@15 of 0.568, close to SyGMa, which indicates that its subset value is representative. BioTransformer and MetaTrans remain on the subset. A five-method rerun at a single sample size is not available here, because BioTransformer's subset configuration was not scripted, and MetaTrans's inference pipeline no longer reproduces in-tree.

### H.12 The five-method comparison under four aggregates

Table 14 re-scores every method's deployed output set on the shared subset, with no top-$k$ truncation, since precision and Jaccard are set quantities. Its output sizes count structures as emitted, whereas Table 13 counts them after the harness collapses duplicates under the matching criterion, which is why the two differ slightly for every method that emits a duplicate. The orderings by recall and by F1 are different orderings.

Table 14: Macro precision, recall, F1 and Jaccard over each method's deployed output on the shared subset ($n = 150$), under INCHIKEY-TAUTOMER. Rank by recall and rank by F1 in the last two columns.

| **method** | prec. | recall | F1 | Jaccard | outputs | rk-R | rk-F1 |
|---|---|---|---|---|---|---|---|
| MetaTrans | 0.122 | 0.565 | **0.177** | 0.105 | 11.5 | 2 | 1 |
| MetaPredictor | 0.116 | **0.585** | 0.174 | 0.101 | 10.9 | 1 | 2 |
| BioTransformer | 0.127 | 0.453 | 0.162 | 0.104 | 10.7 | 4 | 3 |
| **GRAIL** | 0.110 | 0.365 | 0.143 | 0.092 | 8.7 | 5 | 4 |
| SyGMa | 0.029 | 0.562 | 0.051 | 0.028 | 74.2 | 3 | 5 |

Only one method moves by more than its own interval, and it is the one the decomposition names. On F1 the four selecting methods are mutually indistinguishable, every pairwise interval covering zero, the widest separation being GRAIL against MetaTrans at $-0.033$ $[-0.071, +0.004]$. SyGMa is below all four, $+0.092$ $[+0.063, +0.123]$ against GRAIL rising to $+0.125$ $[+0.103, +0.147]$ against MetaTrans. On recall it is above two of them, $-0.109$ $[-0.168, -0.052]$ against BioTransformer and $-0.197$ $[-0.274, -0.120]$ against GRAIL. Those two pairs are the separated reversals: each is significant in one direction on recall and in the other on F1.

The remaining exchanges of place, MetaTrans past MetaPredictor at the top and BioTransformer past GRAIL below it, are within noise and support no claim. The reading is not that the aggregate shuffles a leaderboard arbitrarily. It is that the aggregate is decisive for exactly the architecture the factorisation identifies, the one with no selection stage, and not detectably different among the architectures that select. That last is a failure to detect rather than a demonstration of absence: the widest of those intervals still admits a difference of 0.07 at $n = 150$, and separating the two readings needs a larger shared subset than the comparators currently supply.

### H.13 How much of the ordering is real

The reversals above are reported one at a time, which answers whether the phenomenon exists and not how much of the leaderboard it consumes. The aggregate is available from the same frozen predictions on the same 150 substrates: every pairing of a matching criterion with an aggregate is a leaderboard someone could have published, all five criteria are in use in this literature and all four aggregates are reported in it, and not one of the twenty cells is declared by the paper that stands on it.

| what is allowed to vary | cells | orderings | pairs changing sign | separated both ways |
|---|---|---|---|---|
| criterion and aggregate | 20 | 11 | 9/10 | 2/10 |
| criterion alone, at recall | 5 | 3 | 2/10 | 0/10 |
| criterion alone, at F1 | 5 | 3 | 5/10 | 0/10 |
| aggregate alone, at one criterion | 4 | 4 | 5/10 | 1/10 |

The two counts differ by more than a factor of four and only the smaller one is a finding. Five methods take eleven distinct orderings across the twenty cells and nine of the ten pairwise comparisons change sign somewhere in them, which is a description; two pairs reverse with disjoint intervals in both directions, and they are the two named above. Quoting the first without the second is the fallacy §A.4 rejects, and it is available here to anyone who wants it, ourselves included.

Which axis does the work is visible in the slices, and it is the one §A.5 says it is. Holding the aggregate at recall, the criterion alone moves two pairs and separates none: it erases gaps, not reordering. Holding the criterion fixed, the aggregate alone yields four distinct orderings from four cells, a different leaderboard for every aggregate, and separates one. The count of orderings is not a property of how finely a sweep was gridded, since both axes here are enumerable and small; what a finer budget sweep would add is more cells, not more information.

### H.14 Four methods, one population, and the budget that reorders all of them

The reorderings in §A.5 run through one method that emits far more than the others, which invites the reading that the finding is about SyGMa rather than about budgets. A second high-output method settles it. The SMIRKS variant of MetaTox supplies predictions for 291 test substrates, every one of which carries references and predictions from all four methods, so the comparison needs no new population beyond naming it. Scored through the same matcher, micro, with each method's own parent and its duplicates dropped before the budget is applied:

| method | emitted | r@5 | r@10 | r@15 | r@30 | r@50 |
|---|---|---|---|---|---|---|
| SyGMa | 72.98 | 0.3744 | 0.4556 | 0.4842 | 0.5068 | 0.5113 |
| MetaTox | 33.6 | 0.218 | 0.388 | 0.5143 | 0.6015 | 0.6271 |
| MetaPredictor | 10.75 | 0.3549 | 0.4707 | 0.4797 | 0.4797 | 0.4797 |
| GRAIL | 15.63 | 0.3203 | 0.4331 | 0.4797 | 0.4902 | 0.4902 |

The four methods take three distinct orderings across the sweep, and the method that moves furthest is the new one: MetaTox is last of the four at every budget up to 10 and first of them from 15 on, which is the budget this field reports at. Nothing about the chemistry changes between those columns. At $k = 5$ it has been truncated to a seventh of what it emits, and at $k = 30$ it has not. Its predictions are ranked by its own metabolite-likeness score, the quantity the delivery is named for; PASS writes that score only for the 5,177 of 10,601 predictions it declares above its own threshold, and the rest keep the order the file gives them, since PASS expresses no preference among predictions it declines to score.

Each method's own parent is dropped from its ranked list before the budget applies, and the choice does not run in favour of this system. GRAIL returns the substrate itself on 26 of the 291, MetaPredictor on 52 and MetaTox on 14; SyGMa's zero is a property of the released artifact rather than of SyGMa, whose tree is rooted at the substrate and whose parent was stripped when those predictions were frozen. A slot spent on the parent cannot score, so dropping it helps whichever method spends one, which is MetaPredictor twice over before it is ours. It does not carry the result: with the parent left in, the four methods take the same three orderings at the same budgets and GRAIL's recall@15 moves from 0.265 to 0.263.

### H.15 Every sentence that uses the word, and what entitles it

Section A.1 reserves *certified* for a difference that is separated and survives a correction inside a declared family, and lets *separated* stand alone anywhere. A definition of that kind is worth nothing unless the manuscript obeys it, so obedience is checked rather than asserted: a released script reads the manuscript, finds every sentence carrying the stronger word, and matches it against the declared confirmatory comparisons or against a correction the sentence names itself. It exits non-zero on a

sentence that does neither, which is what an earlier draft of this paper did in twenty-four places. The table gives each use the section and the run-in heading it appears under, so a reader can turn to the sentence and check it against the family beside it.

| where it is said | the comparison it refers to | family size | adjusted $p$ |
|---|---|---|---|
| §1, “Contributions” | the definition itself | – | – |
| | the correction over every grid at once | 15,752 | cutoff $6.7 \times 10^{-6}$ |
| §3.1 | names its own correction | stated in the sentence | – |
| §3.1, “The five terms” | names its own correction | stated in the sentence | – |
| §3.2, “Docking as a control with a published answer” | the docking control, after its own correction | 168 | smallest reversal $1.5 \times 10^{-3}$ against a $9.8 \times 10^{-4}$ threshold |
| §A.3 | the definition itself | – | – |
| §A.3, “Which choice the certified reversals belong to” | the definition itself | – | – |
| §A.1, “Metrics and inference” | names its own correction | stated in the sentence | – |
| | the definition itself | – | – |
| §D, “A leaderboard from outside chemistry, whose convention ships as a function” | names its own correction | stated in the sentence | – |
| | names its own correction | stated in the sentence | – |
| §D, “One correction per board, when the claim ranges over twenty-three” | the correction over every grid at once | 15,752 | cutoff $6.7 \times 10^{-6}$ |
| | names its own correction | stated in the sentence | – |
| | names its own correction | stated in the sentence | – |
| §D, “A test that assumes nothing about the shape” | the shape-free re-test | 42 | no verdict moves |
| | the definition itself | – | – |
| §D, “A null with a size” | the definition itself | – | – |
| | names its own correction | stated in the sentence | – |
| | the definition itself | – | – |
| §H.7, “What the null could have detected” | names its own correction | 448 | – |
| §H.15 | the definition itself | – | – |
| §H.21, “The run comes back positive, and at the budget this field headlines” | names its own correction | stated in the sentence | – |
| §H.21, “The second group does not reorder, and says why the first one does” | names its own correction | 66 | – |
| §H.21, “The inequality is arithmetic; how often it is met is not” | names its own correction | 112 | – |

### H.16 The multiplicity correction, and the family it is taken over

§A.4 corrects for multiplicity post hoc, and both inputs to that correction belong here, not in a sentence. The p-values are normal approximations to the stored paired-bootstrap intervals, $\mathrm{se} = (\mathrm{hi} - \mathrm{lo})/2z$,

and the external decisions do not depend on that approximation. Recomputed from the per-substrate vectors by flipping the sign of each substrate's difference-of-differences, which assumes nothing about the distribution, the same four interactions are rejected, and they are rejected in a family of twelve rather than six, since the sign-flip run tests both criterion steps rather than the one the table reports. The approximation does fail, and where it fails is instructive: three comparisons on the tautomer step move both methods by exactly the same amount on every substrate, so their bootstrap interval has zero width, and dividing by it turns the least informative comparison in the set into the most significant one. The sign-flip test returns $p = 1$ for all three, which is what a contrast carrying no evidence about its own sign deserves. None of the three enters any family this paper declares. The intervals themselves remain the primary evidence and the approximation is adequate only because most decisions sit far from their threshold: the one that does not is stated as such below.

We declare the family conservatively as the grid the two axes admit rather than the comparisons we report; no family was pre-specified. The two axes are method pairs and criterion steps, and two steps are reported on each set: canonical→tautomer and canonical→InChIKey internally, InChIKey→no-stereo and no-stereo→tautomer externally. Six pairs times two steps is a family of twelve externally; ten pairs times two steps is twenty internally. Taking the family as the pairs actually tested, or as the pairs one table admits, counts only one of the two axes.

Table 15: Every comparison entering a family, with its raw p-value, its Holm-adjusted p-value at the searched family size, and the step-down threshold its rank carries at $\alpha = 0.05$. Internal pairs are the 150-substrate subset, external pairs the 37 GLORYx drugs, and the last row is the second confirmatory comparison, on 245 test substrates. The third confirmatory comparison is not tabulated here because it is a family of its own: the 448 paired cross-domain interactions are corrected together in §H.21, where 110 survive.

| family | interaction | raw $p$ | Holm $p$ | threshold | decision |
|---|---|---|---|---|---|
| external ($m$=12) | MetaPredictor vs SyGMa | $1.13 \times 10^{-4}$ | $1.4 \times 10^{-3}$ | 0.0042 | rejected |
| | GRAIL vs SyGMa | $1.65 \times 10^{-4}$ | $1.8 \times 10^{-3}$ | 0.0045 | rejected |
| | BioTransformer vs GRAIL | $7.07 \times 10^{-4}$ | $7.1 \times 10^{-3}$ | 0.0050 | rejected |
| | BioTransformer vs MetaPredictor | $3.80 \times 10^{-3}$ | $3.4 \times 10^{-2}$ | 0.0056 | rejected |
| | BioTransformer vs SyGMa | 0.125 | 0.996 | 0.0063 | not rejected |
| | MetaPredictor vs GRAIL | 0.537 | 1.000 | 0.0071 | not rejected |
| internal ($m$=20) | GRAIL vs BioTransformer | $1.65 \times 10^{-6}$ | $3.3 \times 10^{-5}$ | 0.0025 | rejected |
| | MetaTrans vs SyGMa | $3.35 \times 10^{-3}$ | $6.4 \times 10^{-2}$ | 0.0026 | not rejected |
| selection ($m$=1) | learned selector vs frequency prior | $2.0 \times 10^{-8}$ | $2.0 \times 10^{-8}$ | 0.0500 | rejected |

The external family is unaffected by the widening: the same four interactions are rejected at $m = 6$ and at $m = 12$, and the two that fail are nowhere near their thresholds at either size. The internal one is not. At $m = 2$ or $m = 10$ both internal interactions survive; at the family the search actually spanned, the GRAIL–BioTransformer interaction survives with room to spare while MetaTrans–SyGMa misses its threshold of 0.0026 at $p = 0.00335$. We report that second interaction as an uncorrected interval from here on and do not count it as confirmatory. This is the one decision close enough to its threshold that the normal approximation matters, and it is decided against us, which is the direction in which an approximation is safe to use.

**The same endpoint on a scale-free footing.** The endpoint is a paired difference of raw gains on a $[0, 1]$-bounded metric, and the arms enter the relaxation step from different baselines. Internally that confound runs against the effect and §A.4 says so; externally it runs *with* it, MetaPredictor entering at 0.362 against SyGMa's 0.492 and gaining the more. Deploying the argument only where it exculpates would not do, so the obvious scale-free version is reported here: the gain as a fraction of the headroom left, $\Delta/(1-\text{recall}_{\text{strict}})$, resampled paired over substrates so numerator and denominator move together.

| method | recall (strict) | raw gain | normalised gain (95% CI) |
|---|---|---|---|
| *external,* $n = 37$, INCHIKEY→INCHIKEY-TAUTOMER | | | |
| SyGMa | 0.492 | $+0.007$ | $+0.013$ $[+0.000, +0.042]$ |
| BioTransformer | 0.346 | $+0.028$ | $+0.043$ $[+0.004, +0.099]$ |
| MetaPredictor | 0.362 | $+0.142$ | $+0.222$ $[+0.125, +0.333]$ |
| GRAIL | 0.155 | $+0.178$ | $+0.210$ $[+0.117, +0.313]$ |
| *internal,* $n = 150$, CANONICAL→INCHIKEY-TAUTOMER | | | |
| GRAIL | 0.356 | $+0.010$ | $+0.015$ $[+0.001, +0.040]$ |
| BioTransformer | 0.315 | $+0.130$ | $+0.189$ $[+0.121, +0.262]$ |

Every decision in Table 15 survives the change of scale, in both directions. Signing each interaction as the second method's normalised gain minus the first's, exactly as that table's pairs are ordered, the four externally rejected ones keep their sign and exclude zero: MetaPredictor–SyGMa $-0.209$ $[-0.322, -0.111]$, GRAIL–SyGMa $-0.197$ $[-0.305, -0.095]$, BioTransformer–GRAIL $+0.168$ $[+0.069, +0.273]$, BioTransformer–MetaPredictor $+0.179$ $[+0.075, +0.297]$, and internally GRAIL–BioTransformer $+0.174$ $[+0.108, +0.245]$. The two externally not rejected stay indistinguishable from zero, BioTransformer–SyGMa $-0.029$ $[-0.083, +0.003]$ and MetaPredictor–GRAIL $-0.012$ $[-0.163, +0.141]$. GRAIL's external row is on all 37 here, where the contaminated arm inflates it, and is not the 13-drug figure §A.4 reports; it is shown because the normalisation is what is at issue, not the level. The raw paired difference remains the endpoint, it is a differential re-scoring of frozen predictions and is well defined, and this is a robustness column beside it.

The last row of Table 15 is the second comparison §A.1 declares confirmatory, and it is not a protocol $\times$ method interaction, so it belongs to neither family above: the learned-versus-prior rule-selection gap of §E.10, $-0.144$ $[-0.196, -0.095]$ macro at $k = 15$ on 245 test substrates. No family was pre-specified for it either, and the committed artifact spans three selector modes and two stages, so a conservative family here is not of size one. Saying so is the answer, not an evasion: a comparison the paper declares corrected has to appear here with its threshold, and at $p = 2.0 \times 10^{-8}$ against $0.05$ the correction changes nothing whatever family it is put in.

### H.17 External validity rests on 37 substrates

The external ceiling is measured on 37 GLORYx parent substrates, so its 95% interval is wide by construction, at $[0.616, 0.804]$. The regression that partially explains the gap between internal and external coverage is fit on the same small external population. It is reported as suggestive and not as a law that transfers.

### H.18 One assumption behind Proposition 2 is not measured

Proposition 2's explanation assumes that the probability of a rule being labelled is proportional to its marginal firing rate. That propensity model is asserted, not estimated from data. The identifiability argument for why a learned selector underperforms a frequency prior is therefore consistent with the stated mechanism, not proof of it.

### H.19 Proposition 3 bounds single-step application only

Depth-2 application recovers 0.012 of ceiling at 8.5 times the candidate cost, so the bound constrains single-step rule application rather than metabolite prediction in general. Multi-step chemistry and chemistry outside the bank remain open sources of coverage.

### H.20 The factorised generator, measured in the convention it runs in

The factorised redesign of the generator reaches recall@15 of $0.489$ $[0.464, 0.514]$ on the full test split against the deployed $0.339$, at $14.0$ candidates per substrate against $8.4$. That pair is not budget-matched, so the conservative reading is the one stated: truncated to five candidates it matches $0.339$ while emitting fewer than half as many. It also contributes a re-ranking gain of $+0.0165$, 95% confidence interval $[0.006, 0.027]$, on the full clean test. Closing the larger gap to SyGMa still requires a broader rule bank, which is outside the scope of this paper.

### H.21 THE RANK-FLIP RESULT, AND THE RUN THAT DECIDES WHETHER IT LEAVES METABOLISM

The test was specified before it could be run, and specified so that it could fail: re-score the frozen predictions of at least two comparable published models in another domain under the same conventions, and if the ordering does not move, with an interaction interval covering zero, the broader claim fails.

That run needs at least two models of comparable strength whose match sensitivity differs and whose raw ranked predictions are public. Such predictions are public: the EvalRetro benchmark releases the ranked outputs of twelve single-step retrosynthesis systems on what it presents as the USPTO-50k test set (Hastedt et al., 2024), eleven of them as CSV, under a licence that permits re-scoring. Nothing needs training, and no checkpoint needs to run.

**The eleven are not on one test set, and one of the odd ones out is ours.** Reading the released files, not their table shows three populations where the name promises one. Clustering the eleven by which products they contain, the within-cluster overlap is $1.00$ and the between-cluster overlap is $0.10$: they are not disjoint and not the same set, and no threshold between those two numbers changes the grouping:

| group | systems | reactions agreed | share of its products in our split |
|---|---|---|---|
| seven-system | 7 | 4,999 | 0.10 |
| three-system | 3 | 4,982 | 1.00 |
| one-system | 1 | 4,998 | 0.10 |

Which group is the intended one is not in doubt: the benchmark's own instructions point at the typed Schneider USPTO-50k, whose test split is usually cited at $5{,}007$ reactions and whose whole is usually cited at $50{,}016$. No released file carries either number. The seven-system group's files carry between $5{,}001$ and $5{,}004$ distinct products and agree on $4{,}999$ of them; the three-system group's carry $4{,}983$ each and agree on $4{,}982$, and it is that group's products this repository's split contains. Our own copy of "USPTO-50k" totals $50{,}037$ reactions across its three splits, $5{,}004$ of them in the test split and as many again in validation. We are one of the deviations, which is the useful part of finding it: a shared name that does not fix a population is not a lapse by whoever assembled a table, it is a hazard of the ecosystem, and the deviation reported here is one of its instances.

A leaderboard is a leaderboard only within a group, so the eleven released files are not assembled into a single table. The comparison below runs on the seven-system group, reconstructing its test set from the reactions all seven carry and agree the answer to, rather than trusting any one file's ordering or its index column, which means a running prediction counter in one file, a reaction index in another and a float in a third. The single-model probe reported later in this appendix predates this and is on our own $5{,}004$-reaction copy; it stays, labelled, because the alternative is to quietly restate an old number on a new population.

**That the files are read as their authors wrote them is provable, not assumed.** Every number below depends on one structural reading, that the first row of a reaction's block is its recorded answer and the rest are ranked predictions, and seven files agreeing among themselves would agree just as well if all seven were read wrongly. The three-system group settles it independently, because for those reactions we hold the recorded answer ourselves: the first row of each block is our own recorded reactants on $5{,}004$ of its $5{,}004$ reactions, for all three of its systems. The recomputed canonical top-1 also lands on each system's published exact-match figure, which a misreading would not.

**The run comes back positive, and at the budget this field headlines.** Seven systems, $4{,}999$ reactions, frozen predictions, nothing trained. Top-1 accuracy under each convention:

| system | canonical | INCHIKEY | no-stereo | tautomer |
|---|---|---|---|---|
| Graph2SMILES | 0.5293 | 0.5293 | **0.5571** | **0.5595** |
| GraphRetro | **0.5377** | **0.5377** | 0.5501 | 0.5555 |
| Retroformer | 0.5325 | 0.5325 | 0.5541 | 0.5559 |
| LocalRetro | 0.5339 | 0.5339 | 0.5469 | 0.5495 |
| GLN | 0.5247 | 0.5247 | 0.5377 | 0.5397 |
| G2Retro | 0.5147 | 0.5147 | 0.5291 | 0.5309 |
| RetroXpert | 0.4467 | 0.4467 | 0.4619 | 0.4631 |

The leader changes. Under canonical matching (and identically under INCHIKEY, which on these predictions separates nothing that canonical SMILES does not) the order runs GraphRetro, LocalRetro, Retroformer, Graph2SMILES. Disregard stereochemistry and Graph2SMILES goes from fourth to first, with Retroformer second and GraphRetro third. Five of the twenty-one pairs exchange, on predictions that never change.

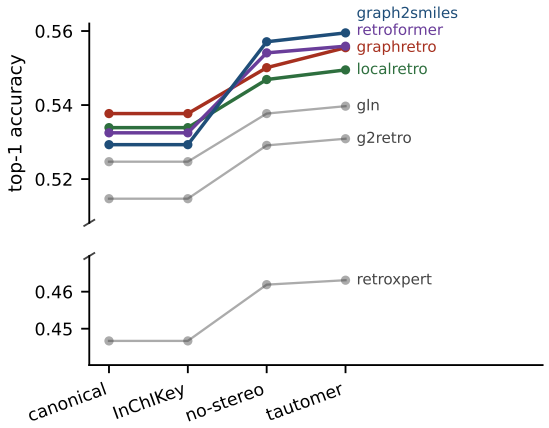



Figure 11: Top-1 accuracy of seven retrosynthesis systems under four matching criteria, on the 4,999 reactions their released prediction files share. Systems that exchange positions are drawn in colour, the remainder in grey.

The exchange is the visible part; the interaction is the claim, and it is certified under the strictest correction this paper applies anywhere. Of 448 paired interaction tests (every pair of systems, every pair of conventions, every $k$) 233 have intervals excluding zero and 110 **survive Holm at** 0.05 **across the whole family**. The $p$ each test is corrected on is not the bootstrap tail the intervals come from: a resampled tail cannot resolve past $1/B$, and on a family of this size every test such a correction keeps sits exactly on that floor, which counts empty tails rather than separated effects and moves with the resample count alone. These reactions are scored a hit or a miss, so the $p$ is the exact conditional test on the discordant reactions. The largest is Graph2SMILES against GraphRetro between canonical and stereo-blind matching, $-0.0154$ $[-0.020, -0.011]$.

Two qualifications travel with it. The effect is concentrated at top-1: at $k = 3$, 5 and 10 the seven take a single ordering under all four conventions and no pair exchanges. And what moves the ordering here is stereochemistry, with tautomer canonicalisation adding little on top: the same asymmetry the external metabolite set shows, and the opposite of the internal one. Top-1 is the number this field leads with, so a criterion that reorders only there reorders the reported leaderboard; but the reordering is a top-1 phenomenon on this benchmark, not a property of the whole accuracy curve.

**The second group does not reorder, and says why the first one does.** The same run on the three-system group is a replication we did not pre-register, and it comes back negative for reordering: one ordering under all four criteria, no pair exchanged. It is not negative about the criterion. Fifteen of its 66 paired interactions survive Holm, so the criterion moves those systems by certifiably different amounts too, and by almost exactly the same amounts as in the seven, a largest differential of 0.013 against 0.015.

What differs is how tightly the leaderboard is packed, and that, not the criterion, is what decides the outcome.

**The inequality is arithmetic; how often it is met is not.** A criterion reverses a pair exactly when it moves the two apart by more than the larger of the margins it connects. One direction is the familiar one: a difference changes sign only if the change exceeds it, so the movement must pass both margins and therefore their maximum. The other closes it: if the two margins share a sign, the movement between them is their difference, which is below the larger. So the condition is an equivalence and not a screen, and comparing the movement against one margin rather than the larger of the two is what produces false alarms. That needs no evidence. What needs evidence is whether real leaderboards are packed tightly enough for it to happen, which is a fact about how a field's methods converge. Both quantities are already computed for every leaderboard in this paper, so we compare them on all of them (11,803 method-pair by criterion-pair comparisons across five domains, with nothing re-scored:

| leaderboard | methods | comparisons | med. gap | med. move | exchanges |
|---|---|---|---|---|---|
| translation, en-de | 19 | 1026 | 0.3881 | 0.4464 | 131 |
| translation, ja-zh | 15 | 630 | 0.4469 | 0.8183 | 29 |
| translation, nine more | 11–16 | 4131 | — | — | 346 |
| translation, 2023, eight | 13–18 | 5454 | — | — | 165 |
| docking, PoseBusters | 7 | 123 | 0.1986 | 0.0724 | 7 |
| generation, MOSES | 8 | 277 | 0.0094 | 0.0001 | 4 |
| metabolites, GLORYx | 4 | 18 | 0.1303 | 0.0361 | 2 |
| retrosynthesis, seven | 7 | 126 | 0.0130 | 0.0016 | 20 |
| retrosynthesis, three | 3 | 18 | 0.0470 | 0.0011 | 0 |

The movement exceeds the larger margin in 704 of the 11,803, about one comparison in seventeen, and a reversal follows in every one of them and in none of the rest. That is the identity holding, which it must, and the run is a regression test on this code rather than evidence about leaderboards. Comparing the movement against one margin rather than both flags 3,632 instead, but that is not a weaker test so much as an arbitrary one: all 2,928 extra flags are movements that widen a gap, so the one-sided form plus a check on the sign recovers the same 704, and which margin counts as "first" is the alphabetical order of two criterion names — taking the other gives 2,623, a swing of a quarter. No number in this paper rests on that form; it is reported so that the choice is visible.

The distribution across leaderboards is the finding, and it is a ratio rather than a level. The seven-system retrosynthesis group is the most exposed at 15.9% because its systems sit a hundredth apart. The least exposed is the three-system group, at 0.0%: its members sit three and a half times further apart while the criterion moves them just as much. Generation is next at 1.4%, because its median movement is 0.0001 and the criterion has almost no purchase there. The two MQM translation tables have the widest margins here by a factor of three, a median of 0.3881 and 0.4469, and are exposed at 12.8% and 4.6%, because a change in what counts as an error moves a system further still. Neither the gap nor the movement decides it alone, which is why the exposed leaderboard is the one whose entries converge faster than the conventions separating them rather than the one whose entries are simply close.

Two earlier measurements outside metabolism come back negative, and they are negative about different things. The first carries a leaderboard: it scores eight generative models and can therefore exhibit an exchange or fail to. The second carries one model, and one model has no ordering to disturb, so it cannot test reordering at all: what it measures is how far the convention moves a score in that domain, which is an upper bound on what any pair separated by more than that could suffer, and a bound is not an observation. We keep it for the bound and label it as one.

The first targets molecular generation, where the matching criterion is not a way of checking an answer but the metric itself: MOSES reports Unique@$k$ as the count of distinct canonical SMILES among a model's generations (Polykovskiy et al., 2020). A model that emits a keto and an enol form of one compound is credited with two molecules. We re-scored the published generations of all eight models MOSES ships samples for, 240,208 molecules, under the same five criteria.

The criterion's effect here is certified across a family of 112 interactions and it is entirely stereochemical. Of 112 paired interactions, every pair of models against every criterion step, 22 have intervals excluding zero and 19 survive Holm at 0.05 across the family, and **all nineteen are the step from**

**canonical to stereo-blind**. Not one interaction survives under InChIKey, Tanimoto or tautomer matching. Tautomer canonicalisation is inert to four decimals on every one of the eight, the single non-zero difference being JT-VAE's $0.9992$ against $0.9991$.

What stereo-blindness does is visible per model and explains itself. The hidden Markov baseline falls $0.9226 \rightarrow 0.8884$ and the $n$-gram baseline $0.9856 \rightarrow 0.9565$; LatentGAN falls $0.9975 \rightarrow 0.9880$; and the combinatorial baseline moves by exactly $0.00000$ $[0.00000, 0.00000]$, because it enumerates without stereocentres and a stereo-blind rule has nothing of its to collapse. A criterion that is inert on one model by construction and costs another a point is a differential effect in the strict sense, and it produces the one rank exchange this set admits: LatentGAN past the combinatorial baseline, an interaction of $-0.0093$ $[-0.0109, -0.0076]$ that survives the same correction, swapping positions 5 and 6 under stereo-blind matching and under no other criterion.

One exchange from a corrected effect is the packing condition rather than a weak result. Seven of the eight sit at $0.986$ or above, so almost every pair is separated by less than the effect can close, and the pair that is not, exchanges. This is the same asymmetry the external metabolite set shows, stereochemistry doing the work and tautomerism none; what differs between the domains is not whether the criterion acts but how much of the leaderboard is close enough to feel it.

The second targets retrosynthesis. Scoring ReactionT5v2 on 200 substrates from the USPTO-50k test set this repository holds, which is the three-system group's, not the seven-system one scored above, the same predictions move only about 0.02 in top-1 accuracy across canonical, stereo-stripped, InChIKey and tautomer matching, between 0.695 and 0.715, roughly an order of magnitude less than the interaction we measure in metabolism. A plausible reason is that retrosynthesis references are canonical, clean small molecules, whereas a metabolism rule engine emits tautomer and stereochemical variants of its references. What this licenses is narrow, and the seven-system result above says how narrow. A leaderboard reorders when the convention moves two methods by *different* amounts, and one model cannot show a differential. One model's small movement might be read as a bound on what a pair could suffer, and that reading is arithmetically true and practically empty: the movement here is about $0.02$, while five of the six adjacent gaps in the seven-system published order are at most $0.01$ with a median of $0.0042$, and five pairs exchange. A bound larger than the gaps it is supposed to protect protects nothing.

**What the three measurements say together.** Matching conventions do not reorder every leaderboard, and the condition is a property of the references rather than of the task, but it is narrower than the bare statement suggests. The effect needs references carrying representational variety *of a kind the compared systems differ on*. Generative models trained on canonical ZINC SMILES reproduce that canonical form, so there is nothing to diverge from and the ordering barely moves. Retrosynthesis references are canonical, which is why the earlier single-model probe found so little, but they carry stereochemistry and the systems differ in reproducing it, and that alone reorders the top-1 table. Metabolite references carry tautomeric *and* stereochemical variety, written with whatever the reporting assay produced, and there the effect is largest. Canonical is not the same as variety-free, which is the distinction our own earlier reading missed.

### H.22 The external set is not external to GRAIL

The criterion ladder of §A.4 is reported on the 37 GLORYx drugs as an external replication. Checking two independent curations against each other turned up something that check was not looking for: $19$ of those $37$ parents are substrates in GRAIL's training split and $5$ more are in validation. Thirteen are unseen. The overlap costs a minute to compute and had never been computed.

The contamination is one-sided, which decides how much has to be repaired. SyGMa and BioTransformer do not train on this corpus, and MetaPredictor was trained by its authors on their own data, so their figures are unaffected *by this overlap* and stay at $n = 37$. Only GRAIL's are inflated, and inflated in our own favour. Restricting the other three to the thirteen would have been an error in the opposite direction, discarding two thirds of the sample to correct a bias they do not have.

The audit is scoped to GRAIL's splits because that is the only overlap that can inflate a number in this paper, and it remains open whether a learned comparator had seen these well-known drugs. The reported effect is the wrong shape for that: a method that had memorised stereo-bearing reference structures would match them under strict INCHIKEY and have *less* left for a stereo-blind criterion to

recover, so memorisation attenuates a stereo-step gain, not producing one. MetaPredictor enters that step at $0.362$, below the rule engine's $0.492$, which is the opposite signature. Nor is the comparator arm a system that trained on nothing: SyGMa's rule probabilities are themselves fitted to an observed-metabolite database (Ridder & Wagener, 2008).

Recomputing on the thirteen with the same loaders, matcher and frozen predictions, so that only the substrate list changes, GRAIL's external recall falls by up to a factor of 2.6: $0.155 \rightarrow 0.059$ under strict INCHIKEY, $0.310 \rightarrow 0.185$ once stereochemistry is disregarded, $0.333 \rightarrow 0.223$ under the tautomer-aware default. The other three methods move by less than $0.07$, which is the sample change rather than a correction.

Two claims survive and one does not. GRAIL's stereo step remains significant at $+0.126$ $[+0.028, +0.239]$ against $+0.155$ on the full set, so the mechanism is real, not memorised: GRAIL emits no assigned stereocentre on this corpus, and a stereo-blind criterion rewards that whether or not it has seen the drug. Its interaction with SyGMa survives too, but carries no information beyond that step and should not be counted as a second result: SyGMa's own stereo step on these thirteen is exactly zero, with an interval of $[0, 0]$, so the interaction is numerically the same measurement, $+0.126$ $[+0.028, +0.239]$. The interaction between BioTransformer and GRAIL does not: $+0.150$ $[+0.069, +0.242]$ on all 37, it is $-0.104$ $[-0.223, +0.000]$ on the unseen drugs, reversing sign and losing significance. It is among the two largest on the full set and is not claimed on either. Of six pairs, four were significant on the full set and two are on the clean one, both against SyGMa.

The headline external result is unaffected because it involves neither contaminated arm. Under strict INCHIKEY SyGMa leads MetaPredictor by $-0.130$ $[-0.196, -0.065]$ and one rung later the advantage is gone; both methods are external to this corpus in the ordinary sense.

We report this rather than quietly recomputing because the failure is instructive about the protocol the paper argues for. An audit of train-test overlap is cheap, is not implied by any of the other checks we run, and surfaced here only as a by-product of measuring something else.

### H.23 What two independent curations of the same drug agree on

Twenty-five of the external drugs are also curated by this paper's corpus, independently, which makes their two metabolite lists comparable to each other in exactly the way a prediction is compared to a reference. Nothing in this subsection involves a model.

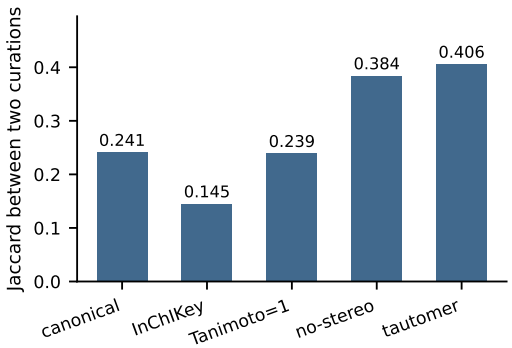



Figure 12: Two independent expert curations of the same twenty-five drugs, scored against each other under each criterion. No model is involved. The same two lists agree about a seventh of their content under the strictest rule and two-fifths under the most tolerant.

The criterion moves that agreement by a factor of nearly three. Under strict INCHIKEY the two curations of the same drug reach a Jaccard of $0.145$; under canonical SMILES, which this harness keys stereo-blind, $0.241$; under Tanimoto=1, $0.239$; stereo-blind INCHIKEY, $0.384$; and under the tautomer-aware default $0.406$. The five are not nested for this measurement (canonical sits above strict INCHIKEY, because dropping stereochemistry from a SMILES string merges more than dropping it from a hash) so what the figure shows is the span, not a ladder. One-sided, the corpus recovers

$0.647$ $[0.535, 0.756]$ of the GLORYx list and GLORYx $0.539$ $[0.439, 0.643]$ of the corpus list, both under the default. All figures are on the same twenty-five drugs, so none of this is a population effect. Two experts looking at the same chemistry appear to agree about a seventh of it under the strictest criterion and about two-fifths under the most tolerant; the difference is notation, not chemistry, and it is the cleanest argument in this paper for declaring the criterion.

Those two numbers also bound what the benchmark can register, and the bound needs no model. Neither curation is a predictor (a source's list is what its inclusion criteria admitted, not an attempt to reconstruct the other's) so what the pair fixes is the level at which two independent readings of the same chemistry coincide: $0.539$ one way, $0.647$ the other. References that no method retrieves cannot all be counted as model error.

Where models sit against that level is a statement about budget, and the budget is the parameter §A.5 declares. Uncapped, scored against the corpus annotation with the same matcher, MetaPredictor recovers $0.562$ $[0.443, 0.676]$, SyGMa $0.524$, BioTransformer $0.480$, GRAIL $0.338$; MetaPredictor's paired gap to the curation is $-0.023$ $[-0.153, +0.123]$, a tie inside noise, not a lead, and it is bought with $10.6$ metabolites per drug against the curation's $6.0$, SyGMa's near-tie with $95.9$. Capped instead at each drug's own curated set size (the budget the curation actually spent, with the curation itself not truncated) its $0.539$ stands against MetaPredictor's $0.301$, SyGMa's $0.271$, GRAIL's $0.284$ and BioTransformer's $0.193$, and every paired gap excludes zero (MetaPredictor $+0.238$ $[+0.123, +0.354]$, the smallest of the four). A fixed $k{=}6$, the curation's mean output, truncates the curation as well and points the same way without reaching significance: $0.472$ against $0.394$, gap $+0.078$ $[-0.050, +0.210]$. It is the more conservative of the two matches, since a curation asserts no ranking and the order its list is cut in is not one it claims.

Models therefore reach curator-level recall only by spending more output than a curator does. At the curator's own budget, matched at each drug's own curated set size, they are behind by a quarter to a third of the reference set; the symmetric $k{=}6$ match in the sentence before this one points the same way and does not reach significance, and this verdict does not cover it.

That per-substrate cap is not an arbitrary handicap, which is the objection it invites. The size supplied is the curator's own, so it is the per-substrate cardinality an oracle would choose, and Appendix E measures what that oracle is worth and how little of it is obtainable. Both figures are taken out of sample, from the same cross-fit, so they share a baseline: for MetaPredictor on the full split the oracle cutoff lifts macro-F1 from $0.254$ at the best constant $k$ to $0.435$, and a head trained to predict that cutoff recovers $0.008$ of the $0.181$ available, $4.3\%$ ($3.2\%$ for GRAIL, $1.1\%$ for SyGMa). No method here can set this budget for itself. What the cap does *not* do is deny the model outputs it could have justified, under recall a cap can only cost, so the comparison reduces to the plain statement: spending what the curator spent, the curator retrieves $0.539$ of the other source's list and the best model $0.301$.

One reading is not available, and it is the one the framing invites. The curators had the literature on each of these drugs in front of them; they were recording what had been observed, not predicting it from structure. So this measures the references, not the contestants: an independent competent reading of the same chemistry omits $46\%$ of what the corpus records, which is why recall past that level is increasingly a measurement of one source's inclusion criteria. It is not evidence about whether models have caught up with expert judgement, and should not be read as any.

Two limits bear on the interpretation. Twenty-five drugs give intervals around $\pm 0.1$, and these are drugs two sources both chose to curate, so they carry $5.7$ to $6.5$ references each against $2.2$ across the corpus. Better-studied compounds are exactly where two curations have the most opportunity to agree, so $0.539$ is more likely an upper bound on inter-source agreement than a typical value. The direction of that bias is not identified: on less-studied chemistry both the curations and the models have less to go on, and we have not measured which of the two degrades faster.

# I Worked example: one prediction set, two verdicts

A single pair of predictions makes the dependence on the matching convention concrete.

Suppose a method predicts D-alanine and the enol tautomer of acetone, and the references are L-alanine and the keto tautomer of acetone. Under strict INCHIKEY matching both predictions miss:

the alanine pair differs at a stereocentre, and the acetone pair differs by a keto–enol tautomerisation, which lies outside what plain InChI normalises, so it misses under stereo-blind matching as well.

Tautomer-aware matching canonicalises tautomers on both sides and generates its canonical SMILES without stereochemistry, so it closes both gaps at once. The acetone pair becomes a hit because the tautomers collapse, and the alanine pair becomes a hit because the stereocentre is no longer distinguished. The same predictions score full recall.

Nothing about the chemistry changed between the two readings. The same fixed predictions score as a near-total miss or as a perfect result depending only on which convention is applied. That is the effect the main text quantifies across five methods and an external set.

## J BUDGET-MATCHED COMPARISON

Recall at a fixed cut-off is not a comparison at matched cost. In their deployed configurations SyGMa reaches 0.554 by emitting about 74 predictions per substrate; GRAIL reaches 0.365 at about 8.7. Normalising by output size reorders the table: SyGMa's precision is 0.029, roughly four times below the low-output methods: GRAIL at 0.109, BioTransformer at 0.127, MetaTrans at 0.122 and MetaPredictor at 0.116. Precision understates all of these under incomplete annotation, but the 8.6-fold difference in output volume it reflects is real. At a smaller matched budget the recall spread compresses as well: at $k = 5$, SyGMa's 0.465 is close to MetaPredictor's 0.470, with GRAIL at 0.321. This matches the independent finding of Boyce et al. (2022) that SyGMa buys coverage with prediction volume.

Table 16: Recall at matched cut-offs, precision, and mean output size under tautomer-aware matching, for each method in its deployed configuration on the shared subset ($n = 150$). Recall here is *macro*, the mean of per-substrate recalls; the decomposition below reports the same methods on the same substrates in *micro*, the pooled ratio, which is why GRAIL reads $0.365$ in this table and $0.318$ there.

| Method | Recall@5 | Recall@10 | Recall@15 | Precision | Mean output |
|---|---|---|---|---|---|
| GRAIL | 0.321 | 0.356 | 0.365 | 0.109 | 8.7 |
| SyGMa | 0.465 | 0.533 | 0.554 | 0.029 | 74.2 |
| BioTransformer | 0.315 | 0.396 | 0.444 | 0.127 | 10.8 |
| MetaPredictor | 0.470 | 0.575 | 0.585 | 0.116 | 11.2 |
| MetaTrans | 0.424 | 0.533 | 0.561 | 0.122 | 12.7 |

### J.1 THE FRONTIER ACROSS BUDGETS

Table 16 compares methods as they are deployed, each with its own output policy. A second measurement removes that difference: rank each method's own candidate pool by its own score and truncate at a common $k$. This is a separate re-scoring run on the same $150$ substrates and the same matching criterion, and its numbers are not the deployed ones, because truncating a pool at $k$ is not what either system does in normal use.

Under that common budget, measured at seven cut-offs between 1 and 64, SyGMa leads GRAIL at every one with no crossover: 0.233 against 0.127 at $k = 1$, 0.493 against 0.350 at $k = 15$, and 0.520 against 0.410 at $k = 64$. GRAIL does not win at a matched output budget, and its selector and ranker are the binding weakness.

### J.2 DECOMPOSING BOTH METHODS

Applying the decomposition to each method, rather than to GRAIL alone, shows the mechanism. The values below are not the frontier's, and they are *micro* where the table above is macro, so GRAIL reads $0.318$ here against $0.365$ there on the same substrates. They describe each method's *deployed* output on the same $150$ substrates under the same criterion, whereas the frontier truncates each method's full candidate pool at a common $k$. The two differ for GRAIL, which reports $0.318$ as deployed against $0.350$ at $k = 15$ on the frontier, because the frontier draws fifteen candidates from a pool the deployed selector never emits.

SyGMa has no selection stage: it applies its whole applicable rule set and emits the result, at about 64 outputs per substrate on this subset. Almost nothing is lost between what its pool reaches and what it reports, 0.520 against a recall@15 of 0.493. Its recall is therefore close to the coverage its bank realises.

GRAIL's bank reaches more of these transformations than SyGMa emits: its ceiling on this subset is 0.701, against a full-test ceiling of $0.817$ and SyGMa's realised 0.520. GRAIL nevertheless reports 0.318, because selection and ranking together retain 0.454 of what the bank could reach. Nothing is lost at the cut-off itself *on these* $150$ *substrates*: emitting about eight candidates on average, GRAIL's budgeted pool is not truncated here, so its recall@15 equals that pool's coverage exactly. That is a property of this subset and not of the method: on the full split GRAIL's output policy, which is itself a cap at fifteen, is reached on $308$ of $1{,}170$ substrates, and that is where the third factor of §A.3 loses its references.

The gap to SyGMa is thus not a worse rule bank. It is the cost of *this* selection stage, which recall charges only to methods that select, and which GRAIL currently pays without recovering. The stronger reading, that any selection stage costs this, is not available and the counterfactual is what rules it out: removing the selector and ranking the raw pool at the same budget gives $0.528$ against the deployed selector's $0.352$ at $k = 15$ on $245$ substrates (§E.7), so what is measured here is a selector that retains less than no selector, not a tax the benchmark levies on selecting architectures. A selector that lost nothing would carry a selection factor of one and would not be charged at all.

## K THE POOL-RELATIVE EMISSION RULE

Every method compared here emits a number of candidates fixed by convention rather than by the substrate: a budget the field reports at, or whatever a threshold admits. The references are not shaped that way. Enzymes compete for one substrate, so its annotated products are alternatives drawn from a shared pool, and how many there are varies: the median substrate on the clean test split has one and the mean $2.22$, against a deployed policy that emits $8.4$ for every substrate alike.

This appendix reports what that costs, with the ranking held fixed in every arm so that any difference is attributable to the size of the emitted set and to nothing else. Three quantities bound the question. The deployed budget is what the system ships. The best global constant is the best any budget-based policy could have done, chosen by looking at this split. Truncating each substrate at its own annotated count is the oracle: it is what perfect knowledge of the set size would be worth at this ranking quality, and it is not achievable.

**The comparison that decides the claim.** A relative rule emits while a candidate scores within a factor $\alpha$ of the leader, which is what a model normalised over its siblings does at inference. Against the deployed budget it gains $+0.074$ $[+0.059, +0.089]$ in macro F1, but that is the weak arm and the paper does not rest on it: the deployed budget is a policy this appendix is arguing against, and beating it shows only that fifteen is too many. The arm that decides whether the *pool* is the right object is the best global constant, because a constant is precisely the policy that ignores the pool. Both arms are chosen with the same hindsight on the same split, which is stated rather than hidden, since neither is a forecast.

| policy | macro F1 | over the best constant | 95% CI | separated |
|---|---|---|---|---|
| deployed budget, $k = 15$ | 0.136 | $-0.059$ | — | — |
| best global constant, $k = 1$ | 0.195 | — | — | — |
| relative rule, $\alpha = 0.8$ | 0.201 | $+0.0061$ | $[+0.0012, +0.0113]$ | yes |
| relative rule, $\alpha = 0.75$ | 0.202 | $+0.0074$ | $[+0.0019, +0.0129]$ | yes |
| relative rule, $\alpha = 0.6$ | 0.206 | $+0.0114$ | $[+0.0042, +0.0188]$ | yes |
| **relative rule,** $\alpha = 0.5$ | **0.209** | $\mathbf{+0.0147}$ | $[+0.0059, +0.0237]$ | yes |
| relative rule, $\alpha = 0.4$ | 0.206 | $+0.0108$ | $[+0.0015, +0.0204]$ | yes |
| relative rule, $\alpha = 0.3$ | 0.206 | $+0.0113$ | $[+0.0005, +0.0221]$ | yes |
| oracle: each substrate's own count | 0.214 | $+0.0189$ | $[+0.0126, +0.0256]$ | yes |
| forecast of that count | 0.181 | $-0.0140$ | $[-0.0262, -0.0015]$ | yes, worse |

**The threshold is not fitted.** A rule that works at one value of a free parameter is a fit to the split it was tuned on. Sweeping $\alpha$ over ten values from $0.95$ to $0.1$ gives a plateau, not a spike: six consecutive values from $0.8$ down to $0.3$ beat the best constant with an interval excluding zero, the gain rises to $\alpha = 0.5$ and falls away symmetrically, and the two extremes fail in the two ways they should. At $\alpha = 0.95$ the rule is a constant of one and gains $+0.0013$ $[-0.0007, +0.0036]$; at $\alpha = 0.1$ it admits nearly everything and loses. What the sweep shows is that the pool carries information about how many to emit across a wide band of thresholds, which is the claim. But $0.5$ is an argmax over the sweep, which is a tuned value whatever the plateau around it looks like, so it is cross-fitted rather than defended. Choosing both $\alpha$ and the competing constant on four fifths of the substrates and comparing only on the remaining fifth, over $100$ random splits, reproduces the point estimate at $+0.0151$ and selects $\alpha = 0.5$ as the training-part argmax in all $100$. The spread across held-out fifths is wide, $[-0.0015, +0.0334]$, as a fifth of a thousand substrates must give; what this establishes is that the threshold is not a fit to this split, and it is not a second interval on the gain.

A companion sweep over *absolute* thresholds is released beside it (`results/stopping_rule.json`) and is a negative result: cutting the pool at a fixed score gains $+0.0035$ $[-0.0047, +0.0120]$ and calibrating it $+0.0039$ $[-0.0042, +0.0122]$, neither separated from zero. Only the two pool-relative rules separate — the one reported here and an expected-F1 rule at $+0.0126$ $[+0.0041, +0.0214]$. That is the comparison the claim needs: the information is in where a candidate sits relative to the pool, not in its score.

**Four fifths of the oracle, without the oracle.** The relative rule reaches $+0.0147$ of an available $+0.0189$. That is the useful form of the result: the headroom over the best constant is small in absolute terms, and almost all of it is reachable without knowing anything the evaluator knows. Forecasting the count directly from the substrate, in contrast, is worse than the constant it replaces by an interval that excludes zero, so the count is not predictable from the substrate in the way the pool ordering is (Appendix G).

**Where it does not hold.** The advantage is not uniform across the split's two populations (§3). On the $845$ parent drugs the rule beats the best constant by $+0.0210$ $[+0.0101, +0.0316]$; on the $325$ substrates that are themselves metabolites of another substrate here it does not, at $-0.0016$ $[-0.0160, +0.0126]$, so on the stratum where the model is weaker its scores say nothing about how many to emit. Under strict INCHIKEY matching rather than the tautomer-aware default the margin is $+0.0151$ $[+0.0063, +0.0241]$, so the result is not an artifact of the most permissive criterion. It remains a property of one method's frozen scores on one corpus: it says that this pool is informative about this set size, not that relative emission is generally preferable.

## L THE REFERENCE SET IS A GRAPH, AND HOW MUCH OF A PRECISION FIGURE THAT EXPLAINS

Every paper in this field defends a low precision figure the same way: the references are incomplete, so a chemically real product that no curator recorded is counted as an error. This paper has leaned on that defence too (§A.1). It is untestable in general, because what was never annotated is not observable. One component of it is testable, and this appendix measures it.

Metabolism composes. The corpus records conversions as edges, substrate to product, and those edges form a graph rather than a list of independent rows. If it records $A \to B$ and, elsewhere, $B \to C$, then it asserts a two-step route from $A$ to $C$ without any curator having written the edge $A \to C$. A method that emits $C$ for $A$ is scored wrong, and whether that is right depends on a fact about how the corpus was assembled rather than on the chemistry.

**The corpus is a graph and it is not transitively closed.** Node identity is the tautomer-aware InChIKey used throughout, so the structure is not an artifact of how a curator drew a molecule, and the graph is taken over all three clean splits, because the corpus is what asserts the second edge while a split only asks about the first. It carries $21{,}964$ edges and $7{,}296$ two-step compositions, which resolve to $7{,}234$ distinct composed pairs. The direct edge is annotated for $302$ of them. The remaining $6{,}932$ are routes the corpus asserts in two steps and denies in one.

**What that is worth to a method, measured.** If the defence carried weight, a method's scored-wrong output would be full of these. It is not. Of everything the metric counts wrong, the share the corpus itself reaches from that substrate in two steps is $0.0016$ $[0.0007, 0.0028]$ for GRAIL, $0.0035$ $[0.0021, 0.0052]$ for MetaPredictor and $0.0015$ $[0.0009, 0.0023]$ for SyGMa. The quantity is method-dependent in the way the argument predicts, and two of the three pairwise differences have intervals excluding zero, the largest being MetaPredictor over SyGMa at $+0.0020$ $[+0.0008, +0.0034]$. It is also two orders of magnitude too small to matter. Crediting a method for everything the annotation graph reaches in two steps raises precision from $0.1060$ to $0.1075$ for GRAIL, from $0.1061$ to $0.1090$ for MetaPredictor and from $0.0777$ to $0.0791$ for SyGMa; a third step adds a further $0.0002$; and the ordering by F1 is the same at every depth.

**Why it is worth reporting a null.** The point is not that annotation is complete; it is that the one part of the incompleteness the data can audit is small, so a precision figure on this corpus cannot be explained away by it. That removes an excuse rather than supplying one, including from this paper, which has used it. It also settles a design question cheaply: a benchmark rebuilt by closing its annotation graph under composition would turn $302$ annotated compositions into $7{,}234$ and would not change what any of these methods is measured to do, because none of them emits the difference.

**The model's own output is a graph too, and it composes.** The corpus is not the only thing here with edges. A method that predicts $A \to B$ and, separately, $B \to C$ has predicted a two-step route from $A$ to $C$ without emitting $C$ for $A$, and a one-step evaluation scores that as a miss. Composition is restricted to intermediates that are themselves substrates of this split, the only molecules every method has a prediction to compose with, which leaves $389$ substrates for GRAIL, $428$ for MetaPredictor and $367$ for SyGMa.

Composing enlarges the emitted set, and a larger set matches more for no good reason, so the recovery is reported against a control that enlarges it by the same amount: composing through the same number of substrates drawn at random. That control recovers $0.0000$ for all three methods, to four decimals, which is what makes the rest a statement about the method rather than about set size. Of the references a one-step evaluation records as missed, composing the method's own predictions recovers $0.0086$ $[0.0041, 0.0138]$ for GRAIL, $0.0243$ $[0.0169, 0.0328]$ for MetaPredictor and $0.0136$ $[0.0080, 0.0202]$ for SyGMa, every one of them separated from its control. The quantity differs by method: MetaPredictor recovers $0.0158$ $[0.0069, 0.0249]$ more than GRAIL and $0.0107$ $[0.0028, 0.0188]$ more than SyGMa, both separated.

It is not a policy. Recovering one reference costs $86$ additional candidates for MetaPredictor and $187$ for GRAIL, so emitting the composition would destroy precision many times over what it buys. What it is is a property of a method that recall does not report: how nearly its one-step output is closed under its own dynamics. Two methods with the same recall can differ in it, and here they do.

**One further reading.** The corpus is closed at $4.2\%$ and the test split alone at $9.7\%$, on $390$ composed pairs against the corpus's $7{,}234$. The gap between the two is not noise about depth: within one split, chains tend to be annotated together because they come from one record, while a composition that crosses splits joins records from different sources that no curator ever linked. Which of the two a substrate falls under is another undeclared property of the population it sits in (§3).

## M WHAT THE MEASUREMENTS POINT AT, AND WHAT THEY DO NOT

Two measurements in this paper constrain how a rule-grounded metabolite predictor should be built, and neither of them is a statement about the model that produced them. This appendix states the design they point at and marks the boundary between what is measured here and what is proposed.

**What is measured.** Emitting a set sized to the candidate pool beats the best global constant on frozen scores by $+0.0147$ $[+0.0059, +0.0237]$ macro F1, which is four fifths of what an oracle count would take over the same constant (Appendix K). Against the deployed budget the margin is $+0.074$, but that is the arm §A.5 discredits and it is not what this design rests on. Forecasting the count from the substrate's own descriptors is worse than the constant it replaces (Appendix G). So the number of metabolites a substrate has is not recoverable from the substrate; it is recoverable from the distribution of scores over the candidates the substrate generates.

**What that rules out.** A scorer that maps a (substrate, product) pair to an independent probability cannot express it. Its output for one candidate does not change when a better sibling appears, so the size of the emitted set has to come from a threshold or a budget, and both are global objects for a quantity that is not global. The deployed filter here is such a scorer, and the $+0.0147$ over the best constant it cannot express is the price.

**What it points at.** Score the candidates of one substrate against each other. Encode the substrate once, score each candidate from the atom it fires at and the rule that fires, and take a softmax over the pool rather than a sigmoid over each pair, with one extra logit for emitting nothing further. Training is then a listwise loss over the pool: the negative log of the normalised score, averaged over the annotated products, which asks the model to rank a substrate's true products above its own alternatives rather than to calibrate an absolute probability for each. That matches what this corpus can supply. Labels here are positive-unlabelled, so an absolute probability has no ground truth to be calibrated against, while a relative order does.

**Why the flow machinery is not needed.** Metabolism is a directed acyclic graph and the obvious tool for generating one is a flow network trained by trajectory balance, which exists to solve credit assignment when only the terminal state is rewarded. That is not this problem. The annotations here are edges: a triple names a substrate and one of its direct products, and a metabolite that is itself metabolised appears again as a substrate, which $1{,}435$ of the $17{,}081$ annotated products in the training split do. Supervision is therefore dense at every node of the graph, and the objective can be a per-node quantity. The graph follows by recursion: apply the pool softmax at the root, expand the candidates that carry mass, and stop where the extra logit does, which makes depth a learned quantity rather than the fixed one every system here uses.

**What is not claimed.** None of this is trained or evaluated in this paper. The emission rule is measured, the listwise objective is not; a model trained that way could fail for reasons the frozen-score experiment cannot see, and the honest status of the paragraph above is a design that two measurements make plausible and no measurement here confirms.