FREA: A MULTI-SOURCE EXPERT BENCHMARK FOR REACTION FEASIBILITY VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reaction feasibility verification is essential for synthesis planning, yet progress is constrained by limited evaluation resources and scarce negative data. We introduce FREA, an expert-annotated benchmark of 751 reactions spanning retrosynthesis proposals, zero-yield experimental records, perturbations by large language models (LLMs), and candidates from five generation methods. This multi-source design enables a broad assessment of verifier strengths and failure modes. Each reaction is manually annotated by expert chemists under carefully specified criteria. We evaluate frontier LLMs, forward models, and learned classifiers on FREA. Frontier LLMs show strong performance, with the best attaining the highest mean balanced accuracy across sources and closely matching expert rankings of retrosynthesis models. Forward models achieve the best performance on retrosynthesis proposals, with likelihood scoring providing an efficient option for large-scale screening. However, no model performs best across all four sources, and each exhibits distinct failure patterns. We also release a corpus of over 14 million recorded reactions and generated negative candidates. Exploratory training experiments suggest some transfer from generated negatives to other reaction sources, and generating negative data that reliably improve generalization remains an open question. Together, the benchmark and corpus provide a shared platform for diagnosing verifiers and studying how to improve them.

1 INTRODUCTION

Reaction feasibility verification is essential for turning retrosynthesis proposals into useful synthesis plans. A synthesis planner may connect a target to available starting materials yet include a chemically infeasible step (Long et al., 2025; Maziarz et al., 2025; Baker et al., 2026). Verifiers determine which proposals remain available for planning and are also used to compare retrosynthesis models. Their errors therefore affect both route selection and the measurement of progress.

Two common approaches are forward prediction and classification. Forward models accept proposals when they recover the target product (Schwaller et al., 2019b; Maziarz et al., 2025). Classifiers learn to distinguish recorded reactions from constructed negatives (Segler et al., 2018; Genheden et al., 2020a; Gaiński et al., 2025). Assessing these approaches requires an independent test of agreement with chemical feasibility. Otherwise, a model can succeed on its training objective without reliably identifying which proposed reactions can proceed.

Building this test is difficult because negative reaction data are scarce (Voinarovska et al., 2024). Moreover, failure under one set of conditions does not establish infeasibility under all reasonable conditions. The field therefore needs a public testbed that evaluates model proposals, experimental records, and controlled perturbations under a common feasibility criterion. Such coverage is essential for distinguishing progress on generated data from improved verification of actual model proposals, and for comparing methods on a shared basis.

We introduce FREA, a public benchmark of 751 reactions labeled by two chemists under an explicit criterion. A proposal must list its substantive reactants and admit reasonable conditions that produce the target in isolable form. Its four sources are retrosynthesis proposals, zero-yield experimental records, perturbations by large language models (LLMs), and candidates from five generation methods. This multi-source design connects evaluation on practical proposals with diagnosis of failures

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.

on controlled perturbations. Expert labels are assigned independently of a candidate's intended label, preserving feasible alternatives among generated reactions. The same generation methods also produce a large corpus for training verifiers and studying negative-data design.

Our contributions are:

1. **A public expert benchmark with an explicit task definition.** We provide 341 feasible and 410 infeasible reactions across four complementary sources, with an annotation protocol and recorded reasons and notes (Sections 3 and 4).
2. **A systematic evaluation of verification methods.** We compare forward models, classifiers, LLMs and rule-based filters, diagnose differences across sources and generation methods, and evaluate verifier-based rankings of retrosynthesis models (Section 6).
3. **Exploratory insights into negative-data design.** Matched training comparisons examine how negative supervision and generator choice affect generalization. Observed gains vary across sources and do not consistently extend to retrosynthesis proposals. We develop hypotheses for improving transfer through label auditing, descriptor matching, and generation guided by errors in model proposals (Section 7).

Our evaluation reveals distinct strengths and limitations across verification approaches, with no model performing best across all four reaction sources. By evaluating these sources under a common expert criterion, FREA enables systematic comparison of verifiers and diagnosis of their failure modes. It also provides a shared testbed for determining whether improvements in models or training with generated negatives transfer across reaction sources.

The benchmark[1] and generated corpus[2] are available on Hugging Face. The code accompanies the submission as supplementary material.

2 Related Work

Feasibility filters. Classifiers trained on recorded reactions and constructed negatives have been used in synthesis planning since the in-scope filter of Segler et al. (2018). AiZynthFinder uses template-generated negatives (Genheden et al., 2020b;a). RetroGFN combines forward-template negatives with replacements of the recorded product (Gaiński et al., 2025). RetroTrim combines several scorers and evaluates complete routes by expert review (Sadowski et al., 2025). These studies establish the value of filtering in planning, but evaluation on generated negatives does not test whether the assigned negative labels agree with experts. Expert review of complete routes addresses a different unit of evaluation from the individual reaction. FREA evaluates individual reactions across four sources under a common expert criterion.

Expert and precedent-based evaluation. ChemCensor scores reactions by checking their reaction centers and functional-group context against recorded reactions (Zagribelnyy et al., 2026b). Concurrent work also evaluates reaction plausibility against expert judgments (Zagribelnyy et al., 2026a). FREA combines model proposals with experimental records and multiple forms of controlled perturbation under an explicit feasibility criterion. This broader coverage supports both diagnosis of source-dependent failures and evaluation of transfer from generated negative training data.

Negative reaction data. Published failures are scarce (Voinarovska et al., 2024). Existing resources include private laboratory records (Jiang et al., 2021), high-throughput experiments for acid–amine couplings (Zhong et al., 2025), and data for enzymatic reactions (Chainani et al., 2025). Experimental failures describe particular conditions. FREA uses zero-yield records as candidates for expert annotation and combines them with model proposals and constructed reactions. Appendix A gives a detailed comparison of these resources and adjacent approaches.

[1] https://huggingface.co/datasets/FREA-benchmark/FREA-benchmark
[2] https://huggingface.co/datasets/FREA-benchmark/FREA-corpus

3 REACTION FEASIBILITY TASK DEFINITION

Given reactants $\mathbf{R} = \{r_1, \ldots, r_k\}$ and products $\mathbf{P} = \{p_1, \ldots, p_m\}$, a verifier predicts

$$f(\mathbf{R}, \mathbf{P}) \in \{0, 1\}, \tag{1}$$

where 1 denotes a feasible reaction and 0 an infeasible one. The input lists all species on the left-hand side and the target products, without reaction conditions. By-products need not be listed.

We define feasibility for a *retrosynthesis proposal*. A proposal is feasible when it satisfies both of the following conditions.

Completeness. Every substantive reactant must be listed. A substantive reactant supplies atoms beyond hydrogen that the product retains, or acts in no routine supporting role. For example, a Suzuki coupling proposal is incomplete if its organoboron partner is missing. Auxiliary components such as catalysts, solvents, bases, reducing agents and water may be assumed, including when they contribute hydrogen to the product. Their status depends on their role and on what they supply: water may be implicit in a hydrolysis, but the alcohol supplying an ester group must be listed. Appendix B.1 specifies the permitted roles and gives boundary examples. Completeness is checked structurally: each product fragment must have a credible source among the listed reactants or permitted auxiliaries. Matching element counts alone cannot establish this correspondence.

Reactivity. There must be reasonable conditions under which the listed reactants, with any permitted auxiliary components, produce the target products in isolable form. The target need not be the major product. The reaction must be a single step or a credible one-pot operation. Sequential addition, formation of reactive species in situ, and workup are allowed; a sequence requiring isolation of an intermediate is not. Assumed conditions must be compatible with the listed reactants and products.

This definition matches the information supplied by retrosynthesis models: the proposed reactants are given, while conditions are usually left to the chemist. Allowing routine auxiliary components also accommodates common omissions in reaction databases (Schneider et al., 2016; Wigh et al., 2024). Requiring substantive partners prevents an incomplete proposal from being accepted merely because additional starting materials could make it work.

The labels record expert judgments under this criterion. They do not predict yield or certify experimental success. A zero-yield record can therefore be labeled feasible if reasonable alternative conditions could produce the target. Appendix B gives the full decision procedure and annotation protocol.

4 THE FREA BENCHMARK

FREA provides a common test of whether verifiers agree with experts across different sources of reaction candidates. It contains 751 expert-labeled reactions, with 341 feasible and 410 infeasible under Section 3. **The four sources connect practical screening, targeted diagnosis, and evaluation of transfer** (Table 1). Combining reactions encountered in practice with deliberate perturbations allows us to test whether a verifier's strengths extend beyond one source.

4.1 FOUR REACTION SOURCES

Retrosynthesis proposals (RP). We collect proposals from eight single-step retrosynthesis models for targets in ORDerly (Wigh et al., 2024), a curated dataset derived from the Open Reaction Database (Kearnes et al., 2021). The models span template-based (Chen & Jung, 2021), semi-template (Chen et al., 2023), and template-free approaches (Zeng et al., 2024; Han et al., 2024; Irwin et al., 2022; Yu et al., 2024; Baker et al., 2026). Proposals are retained whether or not they reproduce the recorded reactants. RP therefore tests the proposals that verifiers must screen during planning, including valid alternatives to the recorded reaction.

Zero-yield records (ZR). These are ORDerly reactions with a reported yield of zero. They were attempted experimentally, but failure under the recorded conditions does not determine feasibility

Table 1: Composition of the FREA benchmark. *In practice*: reactions produced by retrosynthesis models or attempted in the laboratory. *Controlled*: candidates obtained from a recorded reaction by a known negative candidate generation method. Table C.1 gives the per-model and per-method breakdown, including the fraction judged infeasible for each GR generation method.

Source	Produced by	In practice	Controlled	Feas.	Infeas.	Total
Retrosynthesis proposals (RP)	Eight retrosynthesis models	✓		197	85	282
Zero-yield records (ZR)	Experimental records	✓		79	14	93
LLM perturbations (LP)	Two LLMs editing recorded reactions			32	55	87
Generated reactions (GR)	Five generation methods		✓	33	256	289
Total				341	410	751

under our criterion. ZR tests whether a verifier distinguishes chemical feasibility from the outcome of one experimental attempt.

LLM perturbations (LP). GPT-5.4 (OpenAI, 2026a) and Claude Opus 4.6 (Anthropic, 2026a) edit recorded reactions to make them infeasible while keeping them superficially coherent. LP tests whether a verifier detects infeasible edits while preserving variants that remain feasible under expert review.

Generated reactions (GR). We propose five methods for constructing negative candidates, building on earlier uses of template application and replacement of reaction components (Segler et al., 2018; Genheden et al., 2020a; Gaiński et al., 2025). The methods differ in what they change (Figure 1):

- **Random Addition (RA)** adds atoms at reactant reaction centers.
- **Random Replacement (RR)** replaces the reactant set with a similar set from another recorded reaction.
- **New Center (NC)** chooses a different product disconnection and completes the resulting fragments into reactants (Chen et al., 2023; Baker et al., 2024).
- **Forward Template (FT)** applies an alternative forward template to the recorded reactants.
- **Elementary Template (ET)** applies a sequence of elementary reaction templates (Joung et al., 2024) to obtain a different product.

RA, RR and NC change the reactants while keeping the product. FT and ET change the product while keeping the reactants. We sample the same number of candidates from each method for annotation, retaining both feasible and infeasible examples. The known generation methods allow failures to be traced to specific constructions. Comparing GR with the other sources then tests whether gains on these constructions extend to other reactions. Applying the methods to the curated ORDerly corpus also produces the training resource studied in Section 7. Appendices C and D give collection and generation details.

4.2 Expert Annotation

Two chemists, a pharmacy faculty member and a chemistry postdoctoral researcher with wet-laboratory experience, labeled reactions without seeing their sources or model predictions. The workflow combined guideline calibration, independent annotation and joint review of disagreements, yielding 751 expert-labeled benchmark reactions. The release includes recorded failure reasons and notes where available. Appendix B gives the calibration procedure, inter-annotator agreement and adjudication details.

The expert labels illustrate why source alone cannot determine feasibility: 79 of 93 zero-yield records are judged feasible, as are 33 of 289 generated candidates intended to be negative. Expert labels therefore provide a reference that neither recorded outcomes nor generation procedures supply alone. Source proportions are not estimates of their frequency in practice.

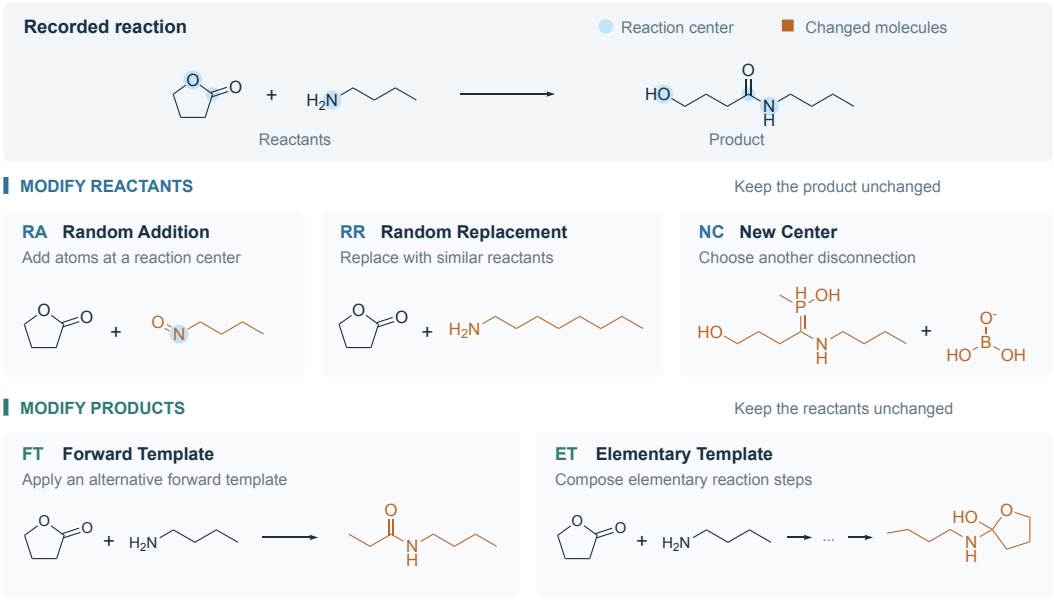

Figure 1: The five methods used to construct GR candidates, illustrated from a common parent reaction (top). RA, RR and NC change the reactants while keeping the product; FT and ET change the product while keeping the reactants. The methods generate intended negatives, whose feasibility is determined by expert annotation.

5 Evaluation Protocol

Evaluated systems. We compare three groups of systems in the main results:

- **LLMs.** Claude Opus 5 (Anthropic, 2026b) and GPT-6 Astra (OpenAI, 2026b) receive the reactants, products and feasibility criterion, and return a decision with an explanation.
- **Forward models.** Chemformer (Irwin et al., 2022) and BARTSmiles (Chilingaryan et al., 2024) are fine-tuned on recorded reactions. Top-10 inclusion accepts a proposal if its target matches one of the ten highest-ranked product predictions after canonicalization. Log-likelihood (LL) scores the supplied target by its mean per-token log-probability given the reactants.
- **Models trained on generated negatives.** BARTSmiles (LL + neg.) continues the BARTSmiles forward model with a loss that favors recorded reactions over generated candidates, and is scored by LL. BARTSmiles (disc.) is trained to output a feasibility score directly. Descriptor GBDT is a gradient-boosted tree using 58 chemical descriptors. All three use our generated corpus.

We also evaluate a rule-based baseline and the released filter from the AiZynthFinder planning tool (Genheden et al., 2020b), with results in the appendix.

Metrics. The primary metric is AUROC. We also report balanced accuracy (BAcc), the mean of the acceptance rate for feasible reactions and the rejection rate for infeasible reactions. BAcc covers all systems, including the LLM and top-10 decisions for which we extract no continuous score. Continuous-score thresholds are fixed on held-out generated pairs before application to FREA. We report results for each source and their unweighted mean across the four sources.

Appendices E and F provide implementation details and the full evaluation protocol, including inference and training settings, model and threshold selection, and statistical tests.

Table 2: Results on FREA by source (feasible/infeasible counts in parentheses). Avg. denotes the four-source mean. Dashes indicate no continuous score. Best values per column are bold.

	RP (197/85)		ZR (79/14)		LP (32/55)		GR (33/256)		Avg.	
System	AUROC	BAcc	AUROC	BAcc	AUROC	BAcc	AUROC	BAcc	AUROC	BAcc
Large language models, criterion in the prompt										
Claude Opus 5	–	69.9	–	64.3	–	**76.1**	–	61.8	–	**68.0**
GPT-6 Astra	–	66.8	–	63.7	–	69.6	–	**62.7**	–	65.7
Forward-model proxies										
Chemformer (top-10)	–	70.6	–	67.2	–	63.3	–	50.1	–	62.8
Chemformer (LL)	76.7	71.6	**82.4**	70.2	64.5	63.8	67.0	49.8	72.6	63.9
BARTSmiles (top-10)	–	**73.1**	–	**74.4**	–	59.0	–	57.1	–	65.9
BARTSmiles (LL)	**78.9**	71.5	79.5	72.5	66.1	60.4	66.9	53.4	72.8	64.4
Trained on our generated negative candidates										
BARTSmiles (LL + neg.)	77.0	67.1	80.8	63.7	69.4	66.1	**73.5**	54.9	**75.2**	62.9
BARTSmiles (disc.)	67.8	62.5	79.0	60.1	66.0	54.8	66.0	56.6	69.7	58.5
Descriptor GBDT	70.3	64.0	72.2	56.3	**75.4**	70.2	64.7	62.4	70.7	63.2

6 RESULTS

6.1 PERFORMANCE ACROSS SOURCES

We first compare the systems in Section 5 across the four sources to identify their strengths and limitations under the common expert criterion. Table 2 reports AUROC and BAcc within each source, together with their unweighted means.

Forward models accept only 28.1–37.5% of feasible LLM perturbations, despite strong performance on retrosynthesis proposals. BARTSmiles achieves the highest RP AUROC with likelihood scoring (78.9) and the highest RP BAcc with top-10 inclusion (73.1). Across the four forward-model readouts, acceptance rates for feasible reactions range from 94.9–98.7% on ZR but only 28.1–37.5% on LP (Table G.3). Their strong performance on feasible recorded reactions therefore does not extend to feasible edits of those reactions.

Within each backbone, we also compare top-10 inclusion with likelihood scoring. Their BAcc differs by at most 3.7 points on any source, with neither readout consistently better. Likelihood scores the supplied target in a single teacher-forced pass, avoiding autoregressive beam search. This makes it much faster and well suited to large-scale evaluation. Its continuous score also supports an adjustable acceptance threshold.

Claude Opus 5 achieves the highest four-source mean BAcc (68.0). It also achieves the highest LP BAcc (76.1). BARTSmiles top-10 and GPT-6 Astra follow in mean BAcc, at 65.9 and 65.7. These comparisons use the LLMs' binary decisions, which do not provide a continuous score for adjusting thresholds.

BARTSmiles (LL + neg.) achieves the highest mean AUROC, whereas the discriminative verifier underperforms forward prediction on RP. BARTSmiles (LL + neg.) reaches 75.2 mean AUROC and 73.5 on GR, while remaining within 1.9 points of the best AUROC on RP and ZR. By comparison, BARTSmiles (disc.) falls 11.1 AUROC points below BARTSmiles LL on RP (95% CI $[+4.3, +17.9]$ for the forward model's advantage). The feature-based Descriptor GBDT provides a simple reference at 70.7 mean AUROC.

The AUROC advantage of BARTSmiles (LL + neg.) does not translate into higher BAcc at the thresholds selected on generated pairs. Its mean BAcc is 62.9, compared with 64.4 for BARTSmiles LL. Better score ordering therefore does not by itself yield better binary decisions under the threshold rule used here. Because these systems differ in training subsets and objectives, their comparison cannot isolate the effect of either factor. Section 7 explores these training choices through matched runs from the same forward-model initialization.

Table 3: AUROC on GR by the method that produced each candidate, with DeLong 95% confidence intervals. Each column ranks the infeasible candidates of one method (count in parentheses) against the same 33 feasible GR candidates, so 50 is chance and a value below 50 ranks infeasible candidates above feasible ones. The measure is threshold-free, so it covers only systems that return a score; rejection rates at each operating point, including those of the language models and the top-10 readouts, are in Table G.4.

	Reactant molecules changed			Transformation changed	
System	RA(50)	RR(53)	NC(54)	FT(44)	ET(55)
Forward-model proxies					
Chemformer (LL)	44.7 [32, 57]	83.5 [75, 93]	26.2 [15, 37]	87.5 [79, 96]	94.9 [90, 100]
BARTSmiles (LL)	46.1 [33, 60]	79.8 [70, 90]	32.5 [20, 45]	83.5 [74, 93]	93.7 [88, 99]
Trained on our generated negative candidates					
BARTSmiles (LL + neg.)	72.7 [62, 84]	53.4 [40, 66]	94.2 [90, 99]	67.7 [55, 81]	78.1 [67, 89]
BARTSmiles (disc.)	63.9 [51, 77]	46.5 [33, 60]	78.0 [66, 90]	68.5 [56, 81]	72.9 [61, 85]
Descriptor GBDT	50.4 [36, 65]	45.6 [32, 59]	87.0 [79, 95]	67.8 [56, 80]	72.0 [60, 84]

6.2 Performance by Generation Method

The overall GR scores do not reveal whether a verifier performs consistently across the five generation methods. To examine this variation, Table 3 compares each method's infeasible candidates with the same 33 feasible GR reactions drawn from all five methods.

Both forward models score below chance on NC negatives, ranking infeasible candidates above feasible reference reactions. NC generates alternative product disconnections. Chemformer achieves 26.2 AUROC (95% CI $[15, 37]$) and BARTSmiles achieves 32.5 ($[20, 45]$). Both intervals lie entirely below 50, indicating an inversion of the desired ranking on this comparison. The same models achieve 83.5–94.9 AUROC on FT and ET, which change the recorded product, and 79.8–83.5 on RR, which replaces the reactant set. Both are near chance on RA, which makes local reactant edits. Evaluation on FT, ET or RR alone would miss this failure on alternative disconnections.

BARTSmiles (LL + neg.) improves NC and RA scores but reduces RR performance relative to BARTSmiles LL. NC increases from 32.5 to 94.2 and RA from 46.1 to 72.7, while RR decreases from 79.8 to 53.4. The higher overall GR AUROC thus combines substantial gains and losses across methods. The discriminative verifier exhibits a similar profile and scores below BARTSmiles (LL + neg.) on four of the five methods.

To include systems that return only decisions, we also examine rejection rates (Table G.4). Claude rejects at least 83% of infeasible candidates from each method, but also rejects two-thirds of the feasible GR reactions. Its high negative rejection rates must therefore be considered together with its rejection of feasible alternatives.

6.3 Ranking Retrosynthesis Models

The preceding analyses evaluate individual reactions. We next examine whether a verifier can also compare retrosynthesis models by the quality of their proposals. For each of the eight models represented in the 282-reaction RP subset, we calculate verifier and expert acceptance rates, then compare the resulting model rankings using Spearman's correlation (Table G.1).

The LLM-based rankings agree most closely with the expert ordering among the evaluated rankings. Experts accept 48.3–92.6% of each model's sampled proposals. Spearman's ρ is 0.93 for Claude (95% CI $[0.31, 0.96]$) and 0.76 for GPT ($[0.26, 0.93]$). The thresholded continuous-score systems have correlations between -0.19 and $+0.17$, with intervals spanning zero. Top-10 rank correlations are not reported. Thus, strong reaction-level RP scores do not necessarily imply agreement with expert model rankings at the likelihood thresholds used here.

The same analysis also quantifies the tradeoff when these decisions are used to filter proposals. BARTSmiles LL retains 50.0% of RP proposals, of which 87.9% are feasible. Claude retains 70.2%,

Table 4: One further epoch of training without and with generated negative candidates, with matched positive reactions and training schedules. AUROC by source, log-likelihood readout. *w/ neg.* is BARTSmiles (LL + neg.) of Tables 2 and 3.

	RP	ZR	LP	GR	Avg.
BARTSmiles	78.9	79.5	66.1	66.9	72.8
+ w/o neg.	77.8 −1.1	76.5 −3.0	66.8 +0.7	64.5 −2.4	71.4 −1.4
+ w/ neg.	77.0 −1.8	80.8 +1.4	69.4 +3.4	73.5 +6.7	75.2 +2.4
Δ (w/ − w/o)	−0.7	+4.3	+2.6	+9.0	+3.8

Table 5: The same recipe with the negatives restricted to the candidates of one generation method, and with an equal share of all five. AUROC by source, log-likelihood readout. Grey: difference from *w/o neg.*, which is an independent run of the setting of Table 4 under the shared positive stream these groups use (Appendix E.2.4).

	RP	ZR	LP	GR	Avg.
w/o neg.	77.6	**79.0**	65.5	64.8	71.7
w/ only RA neg.	78.6 +1.0	76.8 −2.3	67.3 +1.8	68.0 +3.2	72.7 +0.9
w/ only RR neg.	**78.9** +1.3	**79.0** +0.0	**74.4** +8.9	67.4 +2.6	**74.9** +3.2
w/ only NC neg.	74.8 −2.8	76.4 −2.6	67.4 +1.9	**76.9** +12.1	73.9 +2.2
w/ only FT neg.	70.7 −6.9	76.2 −2.8	62.2 −3.4	52.4 −12.4	65.4 −6.4
w/ only ET neg.	77.3 −0.3	77.3 −1.7	68.5 +3.0	64.5 −0.3	71.9 +0.2
w/ all-method neg.	75.3 −2.3	77.3 −1.7	71.0 +5.5	72.5 +7.7	74.0 +2.3

of which 81.8% are feasible. Precision must therefore be considered together with the fraction retained when evaluating a verifier for planning.

7 Exploratory Training with Generated Negatives

The benchmark diagnoses verification failures across sources. We next use the accompanying corpus to investigate how negative training data might address them. This training study is exploratory and hypothesis-generating. Matched runs examine negative supervision and generator choice. The per-method comparison follows a pre-registered design (Appendix E.2.4). We use these comparisons to develop hypotheses for negative-data design and transfer across sources.

7.1 Adding Negative Supervision

We first test whether negative supervision improves verification beyond continuing to train on recorded reactions alone. We compare BARTSmiles (LL + neg.) with a control that starts from the same BARTSmiles forward model and uses the same positive reactions and training schedule. Both runs continue for one epoch over 800,000 pairs (Appendix E.2.3). The control optimizes only the forward-prediction loss. BARTSmiles (LL + neg.) adds a preference loss that ranks each recorded reaction above its generated candidate, using the same mean per-token log-likelihood as at evaluation. This preserves forward-prediction training while introducing supervision from the negative candidates.

The observed mean gain is concentrated outside retrosynthesis proposals. The four-source mean AUROC increases from 71.4 for the continued-training control to 75.2 with negatives (Table 4). The starting model scores 72.8, so the gain is not reproduced by the additional epoch of positive-only training. Scores increase on ZR, LP and GR, but decrease slightly on RP.

Repeating the comparison with 2,400,000 pairs gives the same directions of change (Table G.2). None of the within-source differences survives Holm correction (Appendix F.5). These observations motivate examining which negative candidates support transfer to model proposals.

7.2 Choosing the Generation Method

We next examine whether the generation method changes the pattern of gains across sources. We compare training on each method alone, an equal mixture of all five, and a control without negatives. The seven runs share the initialization, positive stream for the forward-prediction loss, and number of training steps. Each run with negatives uses $800{,}000$ pairs. Only the pairs supplied to the preference loss change, so the comparison holds the positive forward-prediction data fixed. All runs are evaluated at their final training step (Appendix E.2.4).

The observed generator rankings differ between GR and RP. In these runs, NC raises GR AUROC by 12.1 points but lowers RP by 2.8 points (Table 5). RR raises mean AUROC by 3.2 points, followed by the mixture at 2.3 points, while ET leaves the mean nearly unchanged. Both RR and RA slightly improve RP over the control. The equal mixture has a lower mean than RR and also lowers RP by 2.3 points.

Only the FT decline of 6.4 points in mean AUROC is significant after Holm correction ($p_{\mathrm{adj}} = 0.001$). The RR gain has $p_{\mathrm{adj}} = 0.08$, leaving its benefit over the control inconclusive. FT also has the lowest estimated negative purity at 77.2%, compared with about 93% for RR, NC and ET (Table C.1). Its performance decline therefore cannot be attributed to the generation method independently of label noise.

7.3 Hypotheses for Improving Generalization

To formulate hypotheses for improving transfer, we interpret the preceding comparisons alongside expert label audits and the descriptor-classifier analysis. The discriminative verifier's 98.8 AUROC on held-out generated pairs versus 69.7 mean AUROC on FREA further motivates evaluating generalization against independent expert labels.

Validating negative labels may improve transfer by preserving feasible alternatives. Generated candidates can remain feasible, so assigning them negative labels can penalize valid alternatives to the recorded reaction. FT combines the lowest estimated negative purity with the largest performance decline. This motivates expert audits of recurring labeling errors, followed by comparisons that filter or down-weight uncertain candidates while holding the generation method fixed.

Reducing simple descriptor differences may encourage more transferable learning. RR, NC and ET have comparable estimated purities but different observed transfer patterns. Among these three, RR candidates are the hardest for the descriptor classifier to separate from recorded reactions (Table E.1), and RR gives the largest observed mean gain. This motivates testing generators that match recorded reactions and negative candidates on simple molecular descriptors. Whether reducing these differences improves transfer remains to be tested.

Targeting errors in model proposals may improve transfer more than simply mixing generators. In the observed runs, NC improves GR while lowering RP, and the equal mixture also lowers RP relative to the control. Expert-reviewed proposals from a separate development set could identify recurring errors and guide perturbations and method weights. Held-out proposals would then test whether these targeted choices improve transfer.

8 Conclusion

We introduce FREA, a public multi-source expert benchmark for reaction feasibility verification, together with a large corpus of recorded reactions and generated negative candidates. Under an explicit feasibility criterion, we compare verification approaches across reaction sources, analyze their failure modes, and assess their use in ranking retrosynthesis models. Exploratory training comparisons examine how generated negatives affect verification and transfer, motivating further study of negative-data design. Together, the benchmark and corpus support the development and evaluation of verifiers against expert judgments.

The study is limited by benchmark coverage, inherent judgment uncertainty, and the exploratory nature of the training comparisons. See Appendix I for details.

AI USE STATEMENT

We used generative AI in three ways, all disclosed here and reflected in the submission form.

As an object of study. Two large language models produced the LP source of the benchmark by editing recorded reactions (Section 4.1), and two further models are evaluated as feasibility verifiers (Section 5). The generation setup and LLM evaluation settings are described in Appendix C.3 and Appendix E.1.1. All outputs used as benchmark data were labeled by human chemists before entering the benchmark; no model output was accepted as a label.

For data processing. The negative candidate generation methods of Appendix D are algorithmic and do not use language models, except where noted for LP. Scripts that assemble and check the data were written with AI assistance and reviewed by the authors.

For writing and analysis code. We used AI assistance to draft and revise prose, to write analysis and table-generation code, and to check the manuscript against the primary sources it cites. Every number reported in the paper is produced by the scripts included in the supplementary material and was checked against its source by an author. Every claim about prior work was verified against the original publication. The authors take responsibility for the content of this paper, including any text, claims, or artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

The task definition is in Section 3 and the full annotation protocol, decision procedure, and failure-reason taxonomy in Appendix B, including the inter-annotator agreement analysis. Appendix C documents how each source was collected and how the splits were made, and reports the overlap check between the benchmark and the training data. Appendix D specifies the five negative candidate generation methods and the quality-control rules. Appendix E gives model architectures, losses, training configurations and baseline implementations. Appendix F states the checkpoint rule, the threshold rule, and the statistical tests, and reports the sensitivity of our conclusions to the checkpoint. The benchmark and generated corpus are available on Hugging Face at the links in Appendix H. Code is provided in the submission's supplementary material, and the annotation protocol is documented in Appendix B.

REFERENCES

Anthropic. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, Feb 2026a. Accessed: 2026-05-01.

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, July 2026b. Accessed: 2026-09-26.

Frazier N Baker, Ziqi Chen, Daniel Adu-Ampratwum, and Xia Ning. Rlsync: Offline–online reinforcement learning for synthon completion. *Journal of Chemical Information and Modeling*, 64 (17):6723–6735, 2024.

Frazier N Baker, Daniel Adu-Ampratwum, Reza Averly, Botao Yu, Huan Sun, and Xia Ning. Larc: Towards human-level constrained retrosynthesis planning through an agentic framework. In *Proceedings of AI for Accelerated Research Symposium (AIAS)*, volume 3, pp. 153–176, 2026. doi: 10.29007/z3hb.

Yash Chainani, Zhuofu Ni, Kevin M Shebek, Linda J Broadbelt, and Keith EJ Tyo. Dora-xgb: an improved enzymatic reaction feasibility classifier trained using a novel synthetic data approach. *Molecular Systems Design & Engineering*, 10(2):129–142, 2025.

Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

Ziqi Chen, Oluwatosin R Ayinde, James R Fuchs, Huan Sun, and Xia Ning. G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6(1): 102, 2023.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Karen Hambardzumyan, Zaven Navoyan, Armen Aghajanyan, Hrant Khachatrian, and Lusine Khondkaryan. Bartsmiles: Generative masked language models for molecular representations. *Journal of Chemical Information and Modeling*, 64(15):5832–5843, 2024.

Connor W Coley, William H Green, and Klavs F Jensen. Rdchiral: An rdkit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *Journal of chemical information and modeling*, 59(6):2529–2537, 2019.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Piotr Gaiński, Michał Koziarski, Krzysztof Maziarz, Marwin Segler, Jacek Tabor, and Marek Śmieja. Diverse and feasible retrosynthesis using gflownets. *Information Sciences*, 714:122194, 2025.

Samuel Genheden, Ola Engkvist, and Esben Bjerrum. A quick policy to filter reactions based on feasibility in AI-guided retrosynthetic planning. *ChemRxiv*, 2020a. doi: 10.26434/chemrxiv.13280495.v1.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020b.

Yuqiang Han, Xiaoyang Xu, Chang-Yu Hsieh, Keyan Ding, Hongxia Xu, Renjun Xu, Tingjun Hou, Qiang Zhang, and Huajun Chen. Retrosynthesis prediction with an iterative string editing model. *Nature Communications*, 15(1):6404, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

Shu Jiang, Zhuosheng Zhang, Hai Zhao, Jiangtong Li, Yang Yang, Bao-Liang Lu, and Ning Xia. When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*, 9:85071–85083, 2021.

Joonyoung F Joung, Mun Hong Fong, Jihye Roh, Zhengkai Tu, John Bradshaw, and Connor W Coley. Reproducing reaction mechanisms with machine-learning models trained on a large-scale mechanistic dataset. *Angewandte Chemie International Edition*, 63(43):e202411296, 2024.

Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

Lanxin Long, Rui Li, and Jian Zhang. Artificial intelligence in retrosynthesis prediction and its applications in medicinal chemistry. *Journal of Medicinal Chemistry*, 68(3):2333–2355, 2025.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin HS Segler. Re-evaluating retrosynthesis algorithms with syntheseus. *Faraday Discussions*, 256:568–586, 2025.

Mistral AI. Mistral NeMo. `https://mistral.ai/news/mistral-nemo`, July 2024.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026a. Accessed: 2026-05-01.

OpenAI. GPT-6 Astra. https://openai.com/index/gpt-6-astra/, September 2026b. Accessed: 2026-09-26.

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

Dmitrij Rappoport and Alán Aspuru-Guzik. Predicting feasible organic reaction pathways using heuristically aided quantum chemistry. *Journal of Chemical Theory and Computation*, 15(7): 4099–4112, 2019.

RDKit. Rdkit: Open-source cheminformatics, 2024. URL https://doi.org/10.5281/zenodo.14779836. Accessed on 20 April 2026.

Michal Sadowski, Tadija Radusinović, Maria Wyrzykowska, Lukasz Sztukiewicz, Jan Rzymkowski, Paweł Włodarczyk-Pruszyński, Mikołaj Sacha, Piotr Kozakowski, Ruard van Workum, and Stanislaw Kamil Jastrzebski. Trustworthy retrosynthesis: Eliminating hallucinations with a diverse ensemble of reaction scorers. *arXiv preprint arXiv:2510.10645*, 2025.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56(12):2336–2346, 2016.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019a.

Philippe Schwaller, Vishnu H Nair, Riccardo Petraglia, and Teodoro Laino. Evaluation metrics for single-step retrosynthetic models. In *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)*, 2019b.

Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

Lukas M Sigmund, Tina Seifert, Riya Halder, Giulia Bergonzini, Magnus J Johansson, Per-Ola Norrby, Kjell Jorner, and Mikhail Kabeshov. Predicting reaction feasibility and selectivity of aromatic C–H thianthrenation with a qm–ml hybrid approach. *Angewandte Chemie International Edition*, 64(44):e202510533, 2025.

Varvara Voinarovska, Mikhail Kabeshov, Dmytro Dudenko, Samuel Genheden, and Igor V Tetko. When yield prediction does not yield prediction: an overview of the current challenges. *Journal of Chemical Information and Modeling*, 64(1):42–56, 2024.

Daniel S Wigh, Joe Arrowsmith, Alexander Pomberger, Kobi C Felton, and Alexei A Lapkin. Orderly: data sets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling*, 64(9):3790–3798, 2024.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Bogdan Zagribelnyy, Ivan Ilin, Nikita Bondarev, Anton Morgunov, Arkadii Lin, Maksim Kuznetsov, Rim Shayakhmetov, Vladimir Aladinskiy, Alex Aliper, and Alex Zhavoronkov. URSA: Chemistry-aware benchmark for utilitarian retrosynthesis assessment. *arXiv preprint arXiv:2607.04688*, 2026a. URL https://arxiv.org/abs/2607.04688.

Bogdan Zagribelnyy, Ivan Ilin, Maksim Kuznetsov, Nikita Bondarev, Mathieu Reymond, Roman Schutski, Thomas MacDougall, Rim Shayakhmetov, Zulfat Miftakhutdinov, Mikolaj Mizera, Vladimir Aladinskiy, Alex Aliper, and Alex Zhavoronkov. When single answer is not enough: Rethinking single-step retrosynthesis benchmarks for LLMs. *arXiv preprint arXiv:2602.03554*, 2026b.

Kaipeng Zeng, Bo Yang, Xin Zhao, Yu Zhang, Fan Nie, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. Ualign: pushing the limit of template-free retrosynthesis prediction with unsupervised smiles alignment. *Journal of Cheminformatics*, 16:80, 2024.

Haowen Zhong, Yilan Liu, Haibin Sun, Yuru Liu, Rentao Zhang, Baochen Li, Yi Yang, Yuqing Huang, Fei Yang, Frankie S Mak, et al. Towards global reaction feasibility and robustness prediction with high throughput data and bayesian deep learning. *Nature Communications*, 16(1):4522, 2025.

A Related Work, Extended

This appendix expands Section 2 and compares the resources it mentions (Table A.1).

Filters trained on constructed negatives. Segler et al. (2018) introduced an in-scope filter: a neural classifier trained to distinguish recorded reactions from constructed ones. That paper reports only that its 100 million negatives came from an established protocol together with negative sampling; Genheden et al. (2020a), who reimplemented the filter, describe the two strategies as taking the further products that the recorded template yields on the recorded reactants (30 million) and applying a randomly chosen template to a random reactant set (70 million). AiZynthFinder ships a filter built on the same principle, and Genheden et al. (2020a) compare three ways of drawing the negatives, all template-based and all keeping the recorded reactants: applying the recorded template to give a further product, applying a random template, and applying a template suggested by a recommender. RetroGFN's reaction feasibility model is trained on negatives from two procedures, applying forward templates to the recorded reactants to obtain a product that was not recorded, and swapping the recorded product for another product of high Tanimoto similarity (Gaiński et al., 2025). The same paper separately builds a set of hard negatives in the way our RR method does, pairing a template-generated product with reactant sets of high Tanimoto similarity to the ones it was generated from, and uses that set to estimate how often a forward model accepts an infeasible reaction, which is the measurement Section 6 makes against expert labels instead. RetroTrim assembles an ensemble of scorers, some learned and some derived from reaction databases, and shows that different scorers catch different classes of hallucinated reaction; its evaluation is a structured expert review of complete routes for 32 targets rather than of individual reactions (Sadowski et al., 2025). Across this line of work the negatives are constructed by a procedure and the resulting filter is assessed either by its effect on downstream search or on held-out data built by the same procedure.

Evaluation based on precedent. Zagribelnyy et al. (2026b) argue that scoring single-step retrosynthesis by exact match against one recorded answer penalizes valid alternatives, and propose instead to extract a proposed reaction's reaction center and its surrounding functional-group context and check both against a relational database of reported reactions, assigning a score of zero when no precedent is found. They release CREED, a corpus of 22,712,526 reactions for 1,493,715 products generated from roughly 3,000 expert-written bidirectional templates. Precedent-based scoring is complementary to our expert judgments, which can accept a sound proposal without a matching recorded precedent. CREED supplies plausible generated reactions for training. FREA combines feasible and infeasible expert-labeled candidates across four sources to evaluate verifiers and the transfer of generated-negative training.

Negative reactions from experiment. Jiang et al. (2021) trained on failed reactions recorded in a private industrial laboratory notebook, a resource that is not public. Zhong et al. (2025) released high-throughput experimentation data for acid–amine couplings with Bayesian deep learning, which gives genuine negatives with conditions attached but for a single reaction class. DORA-XGB builds an enzymatic reaction feasibility classifier, taking unreported reactions as infeasible (Chainani et al., 2025); the assumption is strong and the domain is retrobiosynthesis. Zero-yield records are plentiful in public databases (Kearnes et al., 2021; Wigh et al., 2024) but report an outcome under particular conditions, so they cannot by themselves establish that a transformation is impossible. This is why our ZR source enters the benchmark as candidates for annotation rather than as negatives.

Feasibility from first principles. Rappoport & Aspuru-Guzik (2019) explore reaction pathways with heuristically aided quantum chemistry, and Sigmund et al. (2025) combine quantum-chemical descriptors with machine learning to predict feasibility and selectivity for one reaction type. Mechanistic datasets support models that reproduce elementary steps (Joung et al., 2024). These methods answer a more fundamental question than ours and at a cost per reaction that rules out screening thousands of proposals, but they are a plausible source of labels for difficult cases in future extensions.

Adjacent tasks. Yield prediction asks how much product forms under stated conditions rather than whether a proposal is well formed, and its own difficulties are documented (Voinarovska et al., 2024). Forward reaction prediction asks what products form (Schwaller et al., 2019a; Ir-

win et al., 2022; Chilingaryan et al., 2024), and is repurposed for verification by the round-trip convention (Schwaller et al., 2019b; Maziarz et al., 2025) that Section 6 evaluates.

Table A.1: Selected resources related to reaction feasibility. *Expert labels*: whether individual reactions carry feasibility judgments by chemists. *Negatives*: whether the resource contains reactions asserted not to proceed.

Resource	Scope	Expert labels	Negatives	Source of negatives
Jiang et al. (2021)	One laboratory	—	✓	failed experiments (private)
Zhong et al. (2025)	Acid–amine couplings	—	✓	high-throughput experiments
DORA-XGB (Chainani et al., 2025)	Enzymatic	—	✓	unreported reactions
RetroGFN (Gaiński et al., 2025)	Small molecule	—	✓	templates, product swap
CREED (Zagribelnyy et al., 2026b)	Small molecule	—	—	—
FREA (ours)	Small molecule	✓	✓	four sources (Section 4)

B Feasibility Criterion and Annotation Protocol

This appendix states the criterion of Section 3 in full and gives its rationale (Appendices B.1 and B.2), specifies what annotators record (Appendices B.3 to B.5), and describes how the evaluation set was annotated (Appendices B.6 to B.9).

B.1 Criterion

For each reaction, the annotator decides whether, under some reasonable conditions, the listed reactants can give the listed products in isolable form, in a single step or a credible one-pot operation. Only the reactants and the products are given; no solvent, catalyst, temperature or yield is supplied. The judgment combines the two conditions of Section 3, which are applied as follows.

Completeness. Every species on the left-hand side counts as a reactant, whether or not a reaction database would record it as a reagent; what matters is whether an omitted species is essential. An omitted species may be assumed only if both conditions hold: it acts in one of the roles of Table B.1, and it supplies nothing beyond hydrogen that is retained in the products. Such a species is *auxiliary*. Anything else is *substantive* and may not be assumed, whether it falls outside those roles or would supply a retained heavy atom while acting in one of them. A transformation evidently intended as a Suzuki coupling is not a complete proposal without its organoboron partner, and an esterification written without its alcohol is not one either, because the alcohol supplies carbon that the product retains. Status follows from the role a species plays here, not from its name, size or customary classification (Table B.2), so methanol is auxiliary as a solvent and substantive as the source of a methyl ester. A missing substantive reactant makes a reaction infeasible even when the intended transformation is recognizable.

A proposal that is complete but for a permitted auxiliary carries a `missing_auxiliary_reagent` note naming the species and its role (Appendix B.5).

The test is structural, not stoichiometric. The test cannot be implemented by comparing element counts, which fail in both directions. Counts can balance although a substantive partner is missing, because leaving groups and other listed species routinely supply the same elements. In one proposal from our pool, a substrate and the coupling reagent HATU are offered as giving a product that bears a 2-picolylamine group; the element counts balance, since HATU supplies carbon, nitrogen and oxygen, yet no listed species contains a picolylamine fragment, so the amine partner is missing. Conversely, counts can fail to balance although nothing is missing: the test concerns only atoms retained in the products, so omitted by-products, unspecified stoichiometry and differences

Table B.1: Permitted auxiliary roles. An omitted species may be assumed only when it acts in one of these roles and supplies nothing beyond hydrogen that is retained in the products. Any species that would supply a retained heavy atom, oxygen included, is substantive and may not be assumed.

Role	Scope
Catalysis, ligation, solvation	Species that supply no retained heavy atom.
Acid or base assistance	Proton transfer or pH control.
Reduction, hydrogen donation	Conventional reducing systems and hydrogen sources that transfer no retained heavy atom.
Oxidation	Conventional oxidants that transfer no retained heavy atom. An oxidant supplying oxygen that the products retain is not admitted.
Water	Water in conventional hydrolysis or workup, including the hydrogen it contributes, but not where the products retain its oxygen.
Activation, dehydration, coupling assistance	Species that enable the listed reactants to react without supplying retained heavy atoms, where the activation can occur within the permitted operation.

Table B.2: The substantive–auxiliary boundary applied to common transformations. Within a row, the species in the middle column may be assumed when omitted; the species in the right column may not, because it supplies atoms beyond hydrogen that the products retain.

Transformation	Auxiliary (may be assumed)	Substantive (may not)
Cross-coupling	Pd catalyst, ligand, base, solvent	organoboron or halide partner
Esterification	acid catalyst, activating reagent	alcohol
Amide coupling	coupling reagent, base	amine
Reductive amination	$NaBH_4$, H_2/Pd	amine
Wittig olefination	base	ylide or phosphonium salt
Mitsunobu reaction	DEAD, PPh_3	pronucleophile
Methylation	base	methyl source
Oxidation	conventional oxidant	oxygen-containing organic fragment
Hydrolysis, hydration	water, acid or base	—
O-Demethylation	BBr_3	—
Phthalimide cleavage	hydrazine	—
Transfer hydrogenation	isopropanol as hydrogen donor	—

Role dependence. The same species can fall on either side: methanol is auxiliary as a solvent and substantive as the source of a methyl ester; isopropanol is auxiliary as a hydrogen donor and substantive as the source of an isopropyl ester.

in hydrogen count are not defects, provided a credible hydrogen source exists among the permitted additions. Annotators are therefore asked where each fragment of the products comes from, structurally, rather than whether the formula balances.

Reactivity. The annotator may assume any reasonable conditions that are not stated, including solvent, temperature, time, atmosphere and order of addition, but may not replace, modify or disregard a listed species. A reaction satisfies reactivity if, under some such conditions, the listed reactants, completed only by auxiliary species, give the listed products in a form that can be isolated. They may form alongside other products, and by-products need not be listed; what is required is that the listed products can be obtained and separated. Assumed conditions must remain compatible with the reactants and products; hypothetical extreme conditions cannot rescue an otherwise unsupported transformation. The reaction must take place in a single step or a credible one-pot operation, which may include sequential addition and the *in situ* formation of a reactive species; workup does not count as a separate step, whereas the isolation of an intermediate does. A reaction fails this condition only if no reasonable conditions would yield the listed products.

B.2 RATIONALE AND SCOPE

Why an explicit definition is needed. Feasibility is used in the synthesis-planning literature in at least four senses: a reaction may be called feasible because it is recorded in a database, recoverable by a forward model from its reactants, precedented in its reaction center and functional-group context, or executable in a laboratory. These senses agree on routine transformations but diverge on the cases that matter most for evaluation: an unprecedented but sound disconnection may be executable, whereas a precedented transformation applied to an incompatible substrate is not. A benchmark must therefore state which sense it measures. Ours is the validity of a retrosynthesis proposal, judged by experts under the rules of this appendix.

Why the two conditions are defined this way. The criterion reflects chemical practice and the requirements of synthesis planning. *Chemical practice.* A chemist reviewing a proposed step reads everything on the left-hand side as reactants and judges an omission by whether the omitted species is essential: a solvent, water or a routine additive is supplied without comment, a partner that builds part of the product is not. Because a proposal specifies no conditions, the chemist likewise asks whether reasonable conditions exist, not whether particular conditions are stated. *Application.* In retrosynthesis planning, a proposal that omits a routine component still describes a usable step and should be expanded, whereas one that omits the coupling partner does not describe a step at all and should be pruned. A step is useful in a route if its product can be isolated and carried forward, and how the material divides among competing products depends on conditions that a proposal does not fix, so reactivity asks only that the listed products can be isolated. *Informativeness.* If any unlisted species could be assumed, or any conditions however extreme, almost any product would be reachable from almost any reactants, and a negative label would carry no information.

What the label asserts. A positive label requires more than a recognizable transformation, as the completeness condition makes explicit. A negative label denotes a defect in the proposal as stated; it does not assert that the transformation is impossible under every conceivable condition, a claim that expert annotation cannot establish. Proposal feasibility is also distinct from *experimental* feasibility, which additionally requires that suitable conditions be identified and that the reaction proceed in acceptable yield in practice. Retrosynthesis proposals carry no conditions, and holding a model that reads them responsible for experimental outcomes would penalize it for information it is not given. The same reasoning applies to experimental records: a reported success establishes that suitable conditions exist, whereas a reported failure establishes only that the conditions tried were unsuitable. Records of failed experiments therefore enter the evaluation set as candidates and are labeled under the same criterion as every other source.

Independence from the source of a reaction. The criterion refers only to the listed reactants and products, not to how the reaction was produced. It therefore applies unchanged to proposals from retrosynthesis models, reactions edited by language models, experimental records and generated candidates, which is what allows performance to be compared across the sources of the evaluation set.

B.3 DECISION PROCEDURE AND RECORDED FIELDS

For each proposal, the annotator records a label (*feasible*, *infeasible* or *uncertain*); one or more failure reasons if the label is infeasible (Appendix B.4); one or more notes if it is feasible (Appendix B.5); a difficulty rating; and a free-text note, which is required for *uncertain* and recommended otherwise. *Uncertain* is not a softened negative. It is reserved for proposals on which the annotator genuinely cannot decide, with the reason stated; an annotator who leans towards one label records that label.

Every proposal is taken through the ordered checks of Table B.3, and the annotator stops at the first that fails; a proposal that passes all six is feasible. The order is part of the protocol. A defective proposal can often be described in more than one way—a missing amine partner, for instance, as an absent reactant or as a transformation without mechanistic basis—and without a fixed precedence, two annotators who agree that a proposal is infeasible may still record different reasons. Fixing the order makes the recorded reason reproducible. Two further rules apply. The reasons at check 3 are alternatives: `missing_substantive_reactant` applies when the in-

Table B.3: Ordered decision procedure. Annotators apply the checks in order and stop at the first that fails, recording the corresponding reason (defined in Table B.4). The two reasons at check 3 are alternatives (Appendix B.3); reasons from checks 4–6 may be recorded together.

#	Check	Reason recorded if the check fails
1	Can every listed species exist and be present?	`unstable_or_nonexistent_reactant`
2	Can the listed products exist as drawn?	`unstable_or_nonexistent_product`
3	Can the listed reactants give the products at all?	`missing_substantive_reactant` or `reactants_cannot_give_product`
4	Does another process intercept the intended pathway?	`reactivity_conflict`
5	Is the specified regio- and stereochemistry consistent with the credible pathways?	`regio_stereo_violation`
6	Can the products be obtained in one operation, without isolating an intermediate?	`requires_isolated_steps`
—	All checks pass	*Feasible*; one or more notes are recorded (Appendix B.5).

tended transformation is recognizable and its only defect is an unlisted substantive partner, and `reactants_cannot_give_product` applies otherwise; if both are selected, only the latter is kept. Reasons from checks 4–6 may be recorded together when more than one applies.

B.4 Failure reasons

Table B.4 defines the seven reasons. They map onto the two conditions of Section 3: `missing_substantive_reactant` records a failure of completeness, and each of the other six records a failure of reactivity. Two distinctions guide the choice of reason. First, `reactivity_conflict` applies when the intended transformation is sound but a competing process prevails, whereas `reactants_cannot_give_product` applies when the transformation has no valid basis to begin with. Second, competing outcomes do not by themselves make a proposal infeasible. The proposal is feasible, with a `selectivity_concern` note, if reasonable conditions could still give the products in isolable form; it is infeasible, with `reactivity_conflict` or `regio_stereo_violation`, if a concrete conflict or competing pathway rules this out.

B.5 Notes on Feasible Proposals

Notes document the components or conditions assumed when assigning a feasible label. The four note categories are:

`no_identified_caveat` Every species that must be physically added is already listed, and no material concern is identified. It does not claim that a complete procedure is given or that the experiment would succeed, and it excludes all other notes.

`missing_auxiliary_reagent` A species permitted by Table B.1 is required but not listed, such as the base of an alkylation or the reductant of a reductive amination. The annotator names it and its role.

`selectivity_concern` Competing chemo-, regio- or stereochemical outcomes are possible, but credible conditions could still give the products in isolable form.

`harsh_conditions` Credible but demanding conditions, such as elevated temperature or pressure, are required. The conditions must remain compatible with the reactants and the products; hypothetical extreme conditions cannot rescue an otherwise unsupported transformation.

These notes document the assumptions accompanying a feasible label; the evaluation target is the binary feasibility label. Difficulty ratings record annotation effort and are not used in the reported evaluations.

Table B.4: Failure reasons recorded for infeasible proposals, in the order of the checks in Table B.3.

Reason	Definition
`unstable_or_nonexistent_reactant`	A listed species, substrate or reagent, is drawn incorrectly, cannot exist, or is too unstable to be present. Lack of bottle stability alone does not qualify a species that can credibly form *in situ*.
`unstable_or_nonexistent_product`	A listed product has a structural or stability defect that prevents it from reasonably existing or being obtained as drawn. Unfamiliarity, or failure of software to parse it, does not qualify.
`missing_substantive_reactant`	The intended transformation is recognizable, but a partner that supplies atoms beyond hydrogen to the products, and is not a permitted auxiliary, is not listed.
`reactants_cannot_give_product`	The listed species can exist, but cannot give the products in one step: the skeletons do not correspond, a listed reagent cannot perform its required role, or there is a concrete mechanistic obstacle.
`reactivity_conflict`	A functional group or listed component causes a competing reaction, quenching or decomposition that intercepts the intended pathway, so that the products cannot be obtained in isolable form. Includes conflicts that only prior protection could resolve.
`regio_stereo_violation`	The specified regioisomer or stereochemical outcome is inconsistent with the credible pathways, so that the specified product cannot be obtained.
`requires_isolated_steps`	Obtaining the products requires an additional, independent step with isolation of an intermediate.

B.6 Annotators and interface

Proposals were labeled by two annotators, a pharmacy faculty member and a postdoctoral researcher in chemistry, both with extensive knowledge of organic chemistry and wet-laboratory experience. Each proposal was presented in a web interface that renders the reactants and the products as structures from their SMILES, with the SMILES strings available on request. The annotators had no access to proposal sources or model predictions. The 41 calibration reactions in the pool were settled during calibration. The 759 reactions outside the calibration set were labeled independently under the final protocol, without access to each other's labels.

B.7 Protocol calibration

Before the main annotation, both annotators labeled a calibration set of 50 proposals under a draft of the guideline, 41 of which belong to the annotation pool. The other nine calibration proposals are outside the pool and are not included in the released benchmark. Disagreements were discussed by cause, and the discussion was used to amend the guideline and align the annotators' understanding. The labels of the calibration proposals were settled in the same discussion, so the 41 that belong to the pool are part of the benchmark but are excluded from the agreement analysis of Appendix B.8. The amendments made the substantive–auxiliary test and its structural reading explicit (Appendix B.1), introduced the ordered decision procedure (Appendix B.3), and consolidated overlapping failure reasons into the seven of Table B.4. Both were in force for the whole of the main annotation.

B.8 Agreement and adjudication

Table B.5 gives the independent labels of the two annotators, before any adjudication, on the 759 reactions of the pool that are not in the calibration set. Agreement was 77.9%, with Cohen's $\kappa = 0.55$ (95% CI [0.49, 0.61]). These statistics measure independent judgments before joint review.

Table B.5: Independent labels of the two annotators before adjudication, on the 759 reactions of the annotation pool outside the calibration set. Rows give the first annotator's label and columns the second's.

	Annotator 2			
Annotator 1	**Feasible**	**Infeasible**	**Uncertain**	**Total**
Feasible	255	77	0	332
Infeasible	90	336	0	426
Uncertain	0	1	0	1
Total	345	414	0	759

Of the 168 disagreements, 119 were adjudicated by joint review against the written guideline. The remaining 49 await adjudication and are excluded from the current benchmark and its reported evaluations. The benchmark therefore contains 591 reactions with matching independent labels, 119 with adjudicated labels, and the 41 calibration reactions, totaling 751 reactions from the 800-reaction pool. Recorded reasons and notes are provided as supporting metadata where available.

B.9 Screening

Before annotation, every reaction from every source was screened with the general validity rules applied to the corpus (Table D.1). A reaction was discarded if its SMILES failed to parse or canonicalize, if it contained an unstable motif such as an O–O–O or B–O–O chain, if a product appeared unchanged among its reactants, or if a product contained more heavy atoms of some element than the reactants. The screen removes parsing failures and several predefined structural defects. The remaining reactions still require expert judgment. No additional structural screening was applied after annotation.

These filters define the structural scope of the benchmark. Reactions that pass them are evaluated by the experts for completeness and reactivity under the stated criterion.

C Construction of the Benchmark

Section 4 presents the benchmark by source. This appendix describes how it was built, in the order of construction: curating positive reactions from ORDerly (Appendix C.1), generating negative candidates (Appendix C.2), collecting the reactions of each source (Appendix C.3), screening and annotation (Appendix C.4), and checking overlap with the training data (Appendix C.5).

C.1 Positive reactions

Positive reactions are taken from ORDerly (Wigh et al., 2024), a curated subset of the Open Reaction Database (Kearnes et al., 2021) with standardized atom mapping and yield annotations, starting from its retrosynthesis train and test splits. All SMILES are canonicalized with RDKit (RDKit, 2024). Reactions with a recorded yield of zero are removed from the positives and archived as candidates for the benchmark (Appendix C.3). We then discard exact duplicates and every reaction that violates a rule marked *All* or *Positive* in Table D.1: reactions that fail parsing, reactions in which a product component appears verbatim among the reactants, reactions containing an O–O–O or B–O–O chain, reactions in which a product contains more atoms of some element than the reactants, and reactions with three or more highly similar reactant components. The curated corpus contains 801,276 training and 76,630 test positives.

C.2 Negative candidates

The five generation methods of Section 4.1 are applied to every positive reaction in both splits (Appendix D). Each candidate inherits the split of the positive reaction from which it was derived and passes the quality-control pipeline of Appendix D.7. Candidates derived from test-split reactions supply the generated source of the benchmark; the full output forms the generated corpus analyzed in Section 7 and Appendix D.8.

Table C.1: Breakdown of the benchmark by the model, language model or negative candidate generation method that produced each reaction. *Infeas. (%)* is the share of a row's reactions the experts judged infeasible, which for a generation method is the share of its candidates that are genuinely negative.

Source	Origin	Feasible	Infeasible	Total	Infeas. (%)
RP	LARC (Claude 3.5 Sonnet)	43	12	55	21.8
	LARC (Mistral NeMo)	42	12	54	22.2
	G2Retro	17	13	30	43.3
	EditRetro	18	12	30	40.0
	LlaSMol	16	13	29	44.8
	UAlign	14	15	29	51.7
	Chemformer	22	6	28	21.4
	LocalRetro	25	2	27	7.4
ZR	ORDerly zero-yield records	79	14	93	15.1
LP	GPT-5.4	15	30	45	66.7
	Claude Opus 4.6	17	25	42	59.5
GR	RA	8	50	58	86.2
	RR	4	53	57	93.0
	NC	4	54	58	93.1
	FT	13	44	57	77.2
	ET	4	55	59	93.2
Total		341	410	751	54.6

C.3 Collecting the four sources

Table C.1 breaks the benchmark down by the model, language model or generation method that produced each reaction.

Retrosynthesis proposals. For targets from the test split, we collected reactant proposals from eight single-step retrosynthesis models: LocalRetro (Chen & Jung, 2021), G2Retro (Chen et al., 2023), UAlign (Zeng et al., 2024), EditRetro (Han et al., 2024), Chemformer (Irwin et al., 2022), LlaSMol (Yu et al., 2024), and LARC (Baker et al., 2026) with two backbone language models, Claude 3.5 Sonnet (Anthropic, 2024) and Mistral NeMo (Mistral AI, 2024). Proposals were pooled across models and deduplicated by canonical SMILES. They were retained whether or not they reproduce the recorded reactants, since both kinds belong to the distribution a model faces in deployment.

Zero-yield records. From ORDerly we retrieve the reactions with a recorded yield of exactly zero that were removed during positive curation, apply the same canonicalization and validity rules, and remove any reaction whose canonical reactant–product pair also occurs among the positives. A zero-yield record shows only that the conditions tried did not give the product (Appendix B.2); these reactions are labeled under the same criterion as every other source.

LLM perturbations. Recorded reactions from the test split were given to GPT-5.4 (OpenAI, 2026a) and Claude Opus 4.6 (Anthropic, 2026a), which were instructed to modify the reactants or the product so that the reaction becomes infeasible while remaining superficially coherent. The instruction states the models' intent only; whether a modified reaction is infeasible is decided by the annotators.

Generated reactions. Generated candidates were sampled in equal numbers from the test-split output of each of the five generation methods (Appendix D). All 289 are outputs of these methods, including those the annotators judged feasible.

C.4 SCREENING AND ANNOTATION

Before annotation, every reaction from every source passed the screen described in Appendix B.9. The screened pool of 800 reactions was then labeled under the protocol of Appendix B. The calibration, independent annotation and adjudication procedure yields the 751 benchmark reactions (Appendix B.8).

C.5 SPLITS AND OVERLAP

The corpus follows the ORDerly retrosynthesis train/test split, and all benchmark reactions except the zero-yield records derive from the test split. The benchmark was deduplicated against the data used to train our models; no benchmark reactant–product pair occurs in those training data. Product overlap remains possible even when the complete reaction differs. 115 benchmark reactions have a main product that is also the product of a training reaction, 97 of them retrosynthesis proposals. BARTSmiles log-likelihood scores 76.6 AUROC on those 97 proposals against 80.8 on the other 185; this comparison is descriptive because the two subsets may differ in difficulty.

D NEGATIVE CANDIDATE GENERATION

The generated reactions in the benchmark and the training corpus are produced by five generation methods. Each produces candidates intended to be infeasible, and whether a candidate is in fact infeasible is established only by expert annotation (Section 4.2). This appendix summarizes the procedures, their main settings, and the quality-control pipeline.

D.1 DESIGN

Each method perturbs a positive reaction $(\mathbf{R}^*, \mathbf{P}^*)$ along one of two axes. Reactant perturbations (RA, RR, NC) replace $\mathbf{R}^*$ with $\tilde{\mathbf{R}}$ while holding the products fixed, whereas outcome perturbations (FT, ET) replace $\mathbf{P}^*$ with $\tilde{\mathbf{P}}$ while holding the reactants fixed. All five methods are applied to every positive reaction in both splits, and each candidate retains a pointer to its parent positive, which supports pairwise training and method-level analysis.

D.2 RANDOM ADDITION (RA)

RA modifies one or more reactants at the reaction center. Each reactant is reduced to its synthon by removing leaving groups, atoms drawn from a frequency-weighted distribution of common leaving-group atoms (e.g., C, O, N, halogens) are appended iteratively at the reaction center under valence constraints, for at most six addition steps per synthon, and the modified synthons are reassembled into reactants $\tilde{\mathbf{R}}$, giving the candidate $(\tilde{\mathbf{R}}, \mathbf{P}^*)$. Up to three candidates are generated per positive reaction. The modification is intended to make a reactant incompatible with the required bond-forming event. It does not guarantee a difficult candidate, and it can produce a reactant that still undergoes a related transformation.

D.3 RANDOM REPLACEMENT (RR)

RR replaces the reactants with structurally similar reactants drawn from the positive training set. For each positive reaction, we retrieve the 200 most similar reactions by Tanimoto similarity on a joint reactant–product fingerprint and take as $\tilde{\mathbf{R}}$ the reactants of the highest-ranked ones that are similar to, but different from, $\mathbf{R}^*$. Up to nine candidates are formed per positive reaction, eight in this way and one from a reactant set drawn at random from the training set. The procedure does not verify that the replacement fails to give $\mathbf{P}^*$, and it discards replacements whose Tanimoto similarity to the original reactant falls below 0.1 (Table D.1).

D.4 NEW CENTER (NC)

NC proposes an alternative disconnection of the product. The reaction-center model of G2Retro (Chen et al., 2023) predicts the top-10 candidate centers for $\mathbf{P}^*$, and up to three centers that

differ from the recorded one are retained, taken from the lowest-ranked end of the list. The resulting synthons are completed by RLSynC (Baker et al., 2024), which returns ten completions ranked by its value estimate, and the low-ranked completions are selected. A low rank indicates only that a completion is unlikely under RLSynC, not that the resulting proposal is infeasible.

D.5 Forward Template (FT)

FT applies forward reaction templates extracted from the positive training set to the recorded reactants, using RDChiral (Coley et al., 2019). We apply every template that occurs at least ten times in the training set together with up to ten randomly drawn rarer templates per reaction, and every product $\tilde{\mathbf{P}}$ that differs from $\mathbf{P}^*$ under canonical SMILES comparison gives a candidate $(\mathbf{R}^*, \tilde{\mathbf{P}})$. A product predicted by a template but not recorded for these reactants may nevertheless be feasible.

D.6 Elementary Template (ET)

ET enumerates elementary steps with 739 templates from the library of elementary templates of Joung et al. (2024). Starting from $\mathbf{R}^*$, the templates are applied breadth-first for up to four steps, and every product reached in at least two steps that differs from $\mathbf{P}^*$ and is not recorded as a product of these reactants gives a candidate, with at most 50 candidates per reaction. As for FT, an unrecorded product is not necessarily infeasible.

D.7 Quality control

All candidates pass a unified pipeline of 16 rules in three tiers: general validity, negative–positive consistency, and removal of method-specific artifacts (Table D.1). In particular, no candidate may encode a product set that is a recorded outcome of the same reactants in the positive corpus.

Table D.1: Summary of the 16 filtering rules applied to the FREA dataset. "All" denotes filters applied to both positive and negative samples. Methods are described in Appendix D.

Rule	Scope	Condition for invalidation
General validity		
`all_invalid_smiles`	All	Canonical reactant or product SMILES is null
`product_in_reactant`	All	Any product component appears verbatim among the reactant components
`unstable_substructure`	All	Reactant or product contains an O–O–O or B–O–O chain
`atom_conservation`	All	Product has more atoms of any element than the reactants
`reactant_internal_similarity`	Positive	Median pairwise reactant Tanimoto > 0.5 with ≥ 3 components
Negative–positive consistency		
`negative_parent_not_none`	Negative	`parent` field is null
`negative_split_match_parent`	Negative	Negative split $\neq$ parent split, or parent not found
`negative_parent_not_invalid`	Negative	Parent positive has been marked invalid
`ra_nc_reaction_centers`	RA, NC	Parent has null `reaction_centers` annotation
`neg_is_not_pos`	Negative	Product set $\subseteq$ some positive's product set for the same reactant SMILES
Method-specific artifacts		
`ft_small_product`	FT	Product has < 6 heavy atoms
`ft_stereo_only`	FT	Product equals parent product after stripping stereochemical annotations
`ft_et_tautomer_equiv`	FT, ET	Product is a strict tautomer of the parent product
`rr_reactant_too_dissimilar`	RR	Tanimoto(replacement reactant, original reactant) < 0.1
`nc_small_fragment`	NC	Reactant set contains a blacklisted, chemically unreasonable fragment
`nc_spurious_ci`	NC	Methyl iodide present in reactants but contributes no net methylation to the product

Table D.2: Size of the generated corpus by split. Per-method rows count candidates before cross-method deduplication; the *Negative* row counts unique canonical reactions.

	Train	**Test**
Positive	801,276	76,630
Negative (unique)	12,047,157	1,189,994
RA	1,705,924	163,253
RR	3,024,969	295,737
NC	1,433,340	149,175
FT	1,781,362	181,411
ET	4,122,334	401,826
Total (unique)	12,848,433	1,266,624

D.8 Size of the generated corpus

Table D.2 gives the size of the generated corpus by split and method. Per-method rows count candidates before cross-method deduplication, whereas the *Negative* row counts unique canonical reactions. The per-method rows therefore sum to 20,772 (train) and 1408 (test) more than the *Negative* row, the difference being the reactions produced by more than one method.

E System Implementations

This appendix describes the evaluated systems and their training settings. Model selection, decision thresholds and statistical tests are given in Appendix F.

E.1 LLMs and Baselines

E.1.1 Large Language Models

We query Claude Opus 5 and GPT-6 Astra through their APIs with extended reasoning at the `high` setting, one call per reaction, with no examples from FREA and no parameter updates. Both models receive the same system prompt, verified identical byte for byte. It states the criterion of Section 3 operationally: what may be assumed without being listed, what may not, and six conditions under which the reaction is not feasible, which correspond to the failure reasons of Appendix B.4 with the third condition covering two of them. The user message carries the reactant and product SMILES.

The decision is read from the final answer tag. One Claude Opus 5 response contained no parsable tag and is excluded from its metrics. Claude Opus 5 used on average 1073 prompt and 532 completion tokens per reaction, at a total cost of US$30.31 for the 751 reactions. GPT-6 Astra used on average 709 prompt and 1032 completion tokens, 956 of them for reasoning, at a total cost of US$44.07 and 32.6 seconds per call.

E.1.2 Forward-Model Proxies

Chemformer (Irwin et al., 2022) and BARTSmiles (Chilingaryan et al., 2024) are fine-tuned for forward prediction on the recorded reactions of the training split, so neither sees a negative candidate. Chemformer starts from its released checkpoint and is trained for 20 epochs. BARTSmiles starts from the Stage-1 checkpoint of Appendix E.2 and is trained for 15,490 steps. Each is taken at the checkpoint with the lowest validation loss: epoch 17 for Chemformer (validation loss 0.04685, against 0.04687 at the last epoch) and step 12,000 for BARTSmiles (0.04678, against 0.04811 at step 15,000). The validation reactions of BARTSmiles largely overlap with those used in Stage 1, so its validation loss is optimistic in absolute terms. It serves only to pick a checkpoint.

For the top-10 readout, each model generates 50 beams, and a reaction is accepted if the canonical SMILES of its products matches one of the ten highest-ranked predictions. The log-likelihood readout is the mean per-token log-probability of the product string given the reactant string under teacher forcing, computed with a maximum length of 1700 tokens for BARTSmiles and the default length of Chemformer.

E.1.3 THE AIZYNTHFINDER IN-SCOPE FILTER

Evaluation setup. We evaluate the released AiZynthFinder filter as an existing planning baseline. The model embeds ECFP4 fingerprints of the product and the difference between product and reactant fingerprints, and scores them by cosine similarity. It has about 4.2M parameters and was trained on 911,869 USPTO reactions and 1,621,484 template-generated negatives (Genheden et al., 2020a). We use `uspto_filter_model.onnx` from version 4.4.1 without retraining, through AiZynthFinder's own reaction objects and filter implementation. No reactant failed to parse. We report results using both the threshold rule of Appendix F.4 and the filter's default cutoff of 0.05; AUROC is independent of the cutoff.

Potential training-data overlap. The training set is not released. We check against Lowe's raw USPTO extraction (Lowe, 2012), a superset of the training data, to bound possible overlap. Matching up to reagents, 132 benchmark reactions appear there: 92 of 93 ZR, 33 of 87 LP, 7 of 282 RP and none of the 289 GR reactions. These matches indicate possible exposure, particularly for ZR and LP, rather than verified membership in the filter's training set. On the 619 reactions with no match, its pooled AUROC is 63.3.

Performance across generation methods. On the held-out generated pairs of Appendix F.4, AUROC is 89.0 for FT, 83.6 for ET and 58.0 for NC. Three additional released filters trained on individual negative types show the same ordering.

Scope of the comparison. These results characterize how the released filter transfers to FREA. Its training data and input conventions differ from ours: its inputs omit reagents, its fingerprints ignore stereochemistry, and its negative labels come from template-generated candidates. Removing carbon-free fragments from the benchmark's reactant lists changes pooled AUROC from 65.9 to 65.3. The benchmark comparison therefore evaluates the existing tool under our criterion. Section 7 explores changes to negative training data using matched training runs.

E.1.4 FEATURE-BASED CLASSIFIER

Features. Twenty-two additive descriptors are computed for the products and as the difference between products and reactants: molecular weight, heavy atoms, rings, aromatic rings, rotatable bonds, hydrogen-bond donors and acceptors, topological polar surface area, logP, formal charge, bonds, stereocenters, and the counts of C, N, O, S, P, F, Cl, Br, I, and other heavy elements. To these 44 features we add the numbers of reactants and products, the product-to-reactant ratios of molecular weight and heavy atoms, the lengths of both SMILES strings and their ratio, the excess of heavy atoms in the products and the number of elements whose count increases, the maximum, mean, and minimum Tanimoto similarity between the Morgan fingerprints (radius 2, 2048 bits) of the largest product and each reactant, and two indicators of whether the product is a substructure of a reactant or the reverse, for 58 features in total, all computed with RDKit (RDKit, 2024).

Model and data. We train a histogram gradient-boosted tree classifier (Pedregosa et al., 2011) with 400 iterations, learning rate 0.06, at most 31 leaves, at least 20 samples per leaf, L2 regularization 1.0, and early stopping on an internal 10% split. The training data are 100,000 pairs from the training split, 20,000 per method, in which every recorded reaction appears at most once. None of the 200,000 rows coincides with a benchmark reaction, even after sorting reactant fragments. The predicted probability of feasibility is the score, thresholded at 0.510, the value that maximizes BAcc on the held-out generated pairs of Appendix F.4.

Benchmark performance. Descriptor GBDT reaches 70.7 mean AUROC on FREA. On the per-method GR comparison, it reaches 87.0 on NC but is near chance on RA and RR (Table 3). The descriptors therefore provide a useful reference, with performance varying substantially by candidate type.

Separability of each method's candidates. On the held-out 10% of the training pairs, about 2000 candidates of each method are scored against 10,000 recorded reactions (Table E.1). Before running this analysis we predicted that RR candidates, the only ones assembled from real molecules, would

be the hardest to separate. The prediction holds among the three methods with matched purity. RA is as hard to separate as RR, which may reflect its lower purity or the small surface trace of its edits.

Table E.1: AUROC of Descriptor GBDT in separating each method's negative candidates from recorded reactions on held-out training pairs, with DeLong 95% confidence intervals.

Method	Purity (%)	Candidates	AUROC
NC	93.1	1981	98.5 [98.2, 98.8]
FT	77.2	1970	92.5 [91.7, 93.2]
ET	93.2	1985	91.5 [90.8, 92.2]
RR	93.0	2002	86.4 [85.6, 87.2]
RA	86.2	2062	85.9 [85.1, 86.7]

What the model uses. Permutation importance on the benchmark ranks the change in bond count, the change in carbon count, and the two substructure indicators highest. The heavy-atom excess and the number of elements whose count increases have exactly zero importance, so the model does not detect violations of atom balance, which the pre-annotation screen removes in any case (Appendix B.9). Fitting the same model directly on the benchmark by five-fold cross-validation gives a pooled AUROC of 75.0, below the 79.8 of the model trained on generated pairs, so these descriptors do not make the benchmark easy to learn.

E.1.5 Rule-Based Baseline

Rule Check is a non-learning cascade implemented with RDKit. It rejects a reaction if (1) a product cannot be parsed, (2) a molecule contains an unstable motif such as an O–O–O or B–O–O chain, (3) the products contain more atoms of some heavy element than the reactants, (4) a product appears unchanged among the reactants, (5) no reactant has a Morgan-fingerprint similarity to the product above a small threshold, or (6) a functional group formed in the product, such as an amide, ester, ether, imine, urea, carbamate, sulfonamide, nitrile, or nitro group, has no compatible precursor among the reactants, unless the product is a substructure of a reactant or is highly similar to one. Rules (1) to (4) coincide with the pre-annotation screen, so on FREA only rules (5) and (6) can reject a reaction. The baseline averages 60.2 BAcc over the four sources (Table G.2), and we omit it from the main text.

E.2 Models Trained with Generated Negative Candidates

This appendix describes BARTSmiles (disc.), the six other discriminative verifiers reported in the supplementary results, and the controlled comparisons of Section 7. All variants share one configuration, which was not tuned for any of them, and each model is trained with random seed 42.

E.2.1 The Discriminative Verifier

Backbone and input. BARTSmiles (disc.) is built on BARTSmiles (Chilingaryan et al., 2024), a BART-style encoder–decoder model pre-trained on SMILES strings. A reaction is written as {reactants}[[>>]]{products}[[Judge]], where molecules are canonicalized with RDKit (RDKit, 2024) and joined by periods, and no reagents or conditions are included. The tokens [[>>]], [[Judge]], and [[Yes]] are added to the vocabulary with randomly initialized embeddings. The whole string is the encoder input. The decoder receives its start token, and the logit of [[Yes]] at the first decoding position is the feasibility score $s(\mathbf{R}, \mathbf{P}) \in \mathbb{R}$. Scoring a reaction therefore costs one encoder pass and one decoder step.

Stage 1: reaction denoising. Reaction strings are longer than the single molecules seen in pre-training, so the model is first adapted to them by denoising. On the recorded reactions of the training split, contiguous token spans with lengths drawn from a Poisson distribution ($\lambda = 3.5$) are corrupted, up to 20% of the non-special tokens. Corrupted positions are replaced by a <mask> token or by random tokens, and the decoder reconstructs the original string under the usual cross-entropy loss. The checkpoint with the lowest validation reconstruction loss initializes Stage 2.

Stage 2: pairwise training. Each negative candidate in the training split is paired with the recorded reaction from which it was generated. Candidates that coincide after canonicalization are merged, and a candidate produced by several methods is assigned to one of them so that all five methods contribute equally. The resulting 7,095,035 pairs contain 15,763 recorded reactions that are absent from the positive corpus of Appendix C.1, none of which occurs in the test split. For a batch of B pairs with scores s_i^+ (recorded reaction) and s_i^- (negative candidate), the loss is

$$\mathcal{L} = \alpha_{\mathrm{BT}}\,\mathcal{L}_{\mathrm{BT}} + \alpha_{\mathrm{BCE}}\,\mathcal{L}_{\mathrm{BCE}} + \alpha_{\mathrm{C}}\,\mathcal{L}_{\mathrm{C}}, \qquad (\alpha_{\mathrm{BT}}, \alpha_{\mathrm{BCE}}, \alpha_{\mathrm{C}}) = (0.8, 0.1, 0.1), \tag{2}$$

with

$$\mathcal{L}_{\mathrm{BT}} = -\frac{1}{B}\sum_{i=1}^{B} \log \sigma\big(s_i^+ - s_i^-\big)\,, \tag{3}$$

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{2B}\sum_{i=1}^{B} \Big[\log \sigma\big(s_i^+\big) + \log\big(1 - \sigma\big(s_i^-\big)\big)\Big]\,, \tag{4}$$

$$\mathcal{L}_{\mathrm{C}} = \left(\frac{1}{2B}\sum_{i=1}^{B} \big(s_i^+ + s_i^-\big)\right)^2, \tag{5}$$

where σ is the logistic function. The Bradley–Terry term $\mathcal{L}_{\mathrm{BT}}$ ranks each recorded reaction above its candidate and is unchanged when a constant is added to all scores. The cross-entropy term removes this invariance by tying the scores to a decision boundary at zero, and the centering term keeps the mean score of a batch near zero. The cross-entropy term alone would not bound the scores, since it keeps decreasing as they grow apart. The centering term controls the batch mean, while weight decay regularizes the model parameters.

Dynamic re-weighting. Methods differ in how quickly the model learns their candidates. Every T steps, the pairwise accuracy acc_m of the current model on held-out pairs of method m is measured, and pairs of method m are then sampled with probability

$$w_m \propto \exp\left(\frac{1 - \mathrm{acc}_m}{\tau}\right), \tag{6}$$

where τ is a temperature, and no method falls below a sampling probability of 0.05. BARTSmiles (disc.) uses $T = 5000$ steps.

Decision rule. A reaction is predicted feasible if $s(\mathbf{R}, \mathbf{P}) > t$, where t is the raw logit threshold frozen on held-out generated pairs (Table F.4).

E.2.2 The Discriminative Verifier Family

Table E.2 lists the seven discriminative verifiers. All start from the same Stage-1 checkpoint and use the same Stage-2 loss, optimizer settings, and effective batch of 512 pairs. They are compared at step 20,000, where each has seen 10.24 million pairs. The number of distinct recorded reactions in the training pairs differs across methods. FT has the fewest, each appearing 8.6 times on average.

Table E.2: Training configuration of the seven discriminative verifiers. All are evaluated at step 20,000.

Negative candidates	Pairs	Re-weighting	Steps trained
One method (RA, RR, NC, FT, or ET)	1,419,007	none	22,168
All five methods, equal shares	1,419,007	every 1000 steps	22,168
All five methods (BARTSmiles (disc.))	7,095,035	every 5000 steps	41,571

E.2.3 The Controlled Pairs

The models of Section 7 start from the BARTSmiles forward model of Appendix E.1.2 and are trained for one epoch, with identical order, seed, learning-rate schedule, warmup, and numerical precision within a pair. The smaller pair uses 800,000 pairs (6250 steps) and the larger

2,400,000 (18,750 steps), at an effective batch of 128. The first optimizes the forward-prediction loss $-\log p_\theta(\mathbf{P} \mid \mathbf{R})$ on the recorded reaction of each pair, and its negative candidates are loaded but unused. The second adds a weighted Bradley–Terry term like $\mathcal{L}_{\mathrm{BT}}$, computed on the score used at inference, the mean per-token log-likelihood $\bar{\ell}_\theta(\mathbf{R}, \mathbf{P})$, of the recorded reaction and of its candidate. Every run is evaluated at the end of its epoch, so no checkpoint is selected. Selecting by validation loss would break the pairing, because the loss of the run with negatives contains the preference term and the loss of the run without it does not, so the two are not the same quantity. Gold AUROC moves by at most 0.010 over the second half of each run, within the evaluation's own noise. All four are thresholded like every other continuous score.

E.2.4 The Per-Method Comparison, and Its Pre-Registration

The seven groups of Table 5 share the recipe of Appendix E.2.3 at the 800,000-pair scale and differ only in the pairs the preference term sees: none (the reference), 800,000 from one method, or 160,000 from each of the five. The forward-prediction loss is computed on one stream of recorded reactions, in one order, shared by all seven, so the positives are held fixed and the comparison isolates the negatives. This removes a confound present in the discriminative family, whose groups draw their positives from their own method's pairs and therefore differ in how many distinct recorded reactions they see. The mixture uses plain concatenation with no re-weighting. Every group is trained for one epoch of 6250 steps from the same initialization with identical optimizer settings and seed, and is scored at the final step.

The design, the endpoints and the analysis were written down before any training data was sampled and before any run was launched; the file and its hash are included in the supplementary material. The primary endpoint is each group's four-source mean AUROC minus the reference group's, tested by a paired bootstrap stratified by source with Holm correction across the six groups. The registration also fixed the rule that conclusions about methods are drawn from the three whose candidates have matched purity, RR, NC and ET, with FT and RA reported for reference. It recorded three predictions: that RR would not be lower than any other group receiving negatives, that FT would be the lowest of the five single-method groups, and that the mixture would not exceed the reference on RP. The observed point estimates follow all three predictions.

One deviation is recorded. The registration required the sampled training pairs to contain no benchmark reaction and the experiment to stop otherwise. The first sample contained four such pairs, the runs were stopped, and each was replaced by the next unused clean pair of the same permutation before the runs were restarted. No stream shares a reaction with the benchmark.

The reference recipe, run once here and once as the smaller pair of Table 4, differs by 0.3 on the four-source average and by 2.5 on ZR with nothing changed, which bounds how much any margin of a few points can carry.

F Evaluation Protocol Details

F.1 Metrics

AUROC is the probability that a randomly drawn feasible reaction receives a higher score than a randomly drawn infeasible one, with ties counted as one half. We compute it within a source. TPR is the share of feasible reactions a system accepts, TNR the share of infeasible reactions it rejects, and BAcc their mean. Within a fixed source, BAcc weights the two classes equally. The average of a system is the unweighted mean over the four sources and was fixed before the evaluation. It includes GR, whose generation methods also supply the negative training data for several systems. The AUROC of a negative candidate method on GR compares that method's infeasible candidates with all 33 feasible GR candidates, which come from all five methods (RA 8, RR 4, NC 4, FT 13, ET 4).

Why we evaluate within sources. Feasible rates range from 11.4% in GR to 84.9% in ZR. To illustrate the effect on a pooled score, each reaction receives the fraction of feasible labels in its source in Table 1. We then compute AUROC over all reactions. Every within-source comparison is tied and contributes one half, but the between-source comparisons raise pooled AUROC to 82.1. These source rates are estimated from the benchmark labels; the calculation illustrates what pooled

AUROC can reward without assessing individual reactions. Pooling also reverses a model comparison: BARTSmiles (disc.) exceeds Chemformer LL in pooled AUROC (82.5 versus 81.9), but is lower in four-source mean AUROC (69.7 versus 72.6). Evaluating within each source prevents between-source comparisons from dominating the reported performance.

F.2 Sensitivity to Auxiliary-Reagent Annotations

We assess the effect of auxiliary-reagent annotations by excluding the 104 feasible reactions carrying `missing_auxiliary_reagent` and evaluating all systems at the same frozen thresholds (Table F.1). This subset is defined from annotation metadata without reference to model predictions. The note records a permitted auxiliary omission under Appendix B.1.

Claude Opus 5 retains the highest mean BAcc, while some intermediate rankings change. The four leading mean AUROCs range from 74.6 to 75.0 on the restricted subset: BARTSmiles (LL + neg.) is comparable to the forward baselines here, rather than retaining its full-benchmark lead. Its advantage over BARTSmiles (disc.) remains 4.2 points.

Table F.1: Source-mean scores after excluding the 104 feasible reactions annotated with `missing_auxiliary_reagent`, using the same frozen thresholds. Full AUROC is reproduced from Table 2; restricted AUROC and BAcc use the remaining reactions.

System	Full AUROC	Restricted AUROC	Restricted BAcc
Claude Opus 5	–	–	70.0
GPT-6 Astra	–	–	67.2
Chemformer (top-10)	–	–	64.5
Chemformer (LL)	72.6	74.6	65.7
BARTSmiles (top-10)	–	–	67.9
BARTSmiles (LL)	72.8	75.0	66.7
BARTSmiles (LL + neg.)	75.2	74.9	63.6
BARTSmiles (disc.)	69.7	70.7	59.6
Descriptor GBDT	70.7	74.7	66.6

F.3 Checkpoints

Forward models are taken at the checkpoint with the lowest validation loss for forward prediction (Appendix E.1.2). For the seven discriminative verifiers, we scored all 129 saved checkpoints on the 10,000 held-out generated pairs of Appendix F.4. For every model, between 38.9% and 75.0% of the checkpoints after step 5000 lie within the DeLong 95% confidence interval of the best checkpoint, and for BARTSmiles (disc.) the best and the final checkpoints differ by 2×10^{-5} AUROC. In-domain validation provides little separation among these checkpoints. We report a common budget of 20,000 steps, the latest step at which all seven runs saved a checkpoint. At this step every model has seen 10.24 million pairs, whereas the final checkpoints of the runs differ by a factor of 1.9 in compute. For BARTSmiles (disc.), step 20,000 also gives the highest pooled AUROC on RP, ZR, and LP among its nine checkpoints, which could favor it.

Table F.3 compares the reported checkpoints with the final ones and with an oracle chosen on the benchmark labels. BARTSmiles (disc.) stays at or below 67.8 AUROC on RP at every checkpoint, and even its oracle average (71.2) falls short of forward-model likelihood (72.6 and 72.8). The verifier trained on RR candidates remains the best of the seven under all three rules, the one trained on FT candidates remains near chance, and the small mixture stays close to the full one.

F.4 Thresholds

Every continuous score is thresholded at the value that maximizes BAcc on 10,000 held-out pairs of recorded reactions and negative candidates from the generated test split. The set is balanced between recorded reactions and candidates and across the five methods, and shares no reactant or product string with the benchmark. BAcc gives equal weight to the two classes, but the selected threshold can still depend on the mixture of generation methods. The threshold is then frozen (Table F.4).

A frozen threshold loses some BAcc relative to one tuned on each source of the benchmark. For BARTSmiles LL the loss is 1.8 points on RP, 1.9 on ZR, 7.3 on LP, and 15.0 on GR.

F.5 Statistical Testing

Confidence intervals on a single metric are percentile intervals from 10,000 bootstrap resamples of the reactions in the evaluated subset. AUROC differences between two systems on the same reactions are tested with the paired DeLong test (DeLong et al., 1988). Table F.2 lists twelve paired DeLong tests, with p-values adjusted by the Holm procedure (Holm, 1979) over the twelve of them together. One comparison survives correction: on RP, BARTSmiles LL exceeds BARTSmiles (disc.) by 11.1 AUROC points. The difference between Descriptor GBDT and BARTSmiles (disc.) on RP is not significant; this does not establish equivalence. Neither controlled pair of Section 7 reaches significance on any source, including the 9.0-point difference on GR, where only 33 feasible candidates are available. The per-method training comparison uses the separate paired-bootstrap analysis and six-comparison correction specified in Appendix E.2.4.

Table F.2: Paired DeLong tests, within source. Δ is AUROC(A) $-$ AUROC(B) in points. p_{adj} is Holm-corrected over all twelve tests.

Source	A	B	Δ	95% CI	p	p_{adj}
RP	BARTSmiles (LL)	BARTSmiles (disc.)	$+11.1$	$[+4.3, +17.9]$	0.0015	**0.018**
RP	Descriptor GBDT	BARTSmiles (disc.)	$+2.5$	$[-4.3, +9.4]$	0.470	1.00
RP	BARTSmiles (LL)	Descriptor GBDT	$+8.6$	$[+0.8, +16.3]$	0.031	0.343
RP	RR only	BARTSmiles (disc.)	$+6.7$	$[+0.3, +13.2]$	0.042	0.418
RP	BARTSmiles (LL + neg.)	no negatives	-0.7	$[-4.2, +2.7]$	0.671	1.00
ZR	BARTSmiles (LL + neg.)	no negatives	$+4.3$	$[-2.5, +11.2]$	0.213	1.00
LP	BARTSmiles (LL + neg.)	no negatives	$+2.6$	$[-3.2, +8.5]$	0.381	1.00
GR	BARTSmiles (LL + neg.)	no negatives	$+9.0$	$[-3.0, +21.1]$	0.142	1.00
RP	+ negatives, 3×	no negatives, 3×	-4.0	$[-8.0, -0.0]$	0.050	0.445
ZR	+ negatives, 3×	no negatives, 3×	$+1.1$	$[-3.3, +5.4]$	0.624	1.00
LP	+ negatives, 3×	no negatives, 3×	$+4.5$	$[-2.0, +11.0]$	0.178	1.00
GR	+ negatives, 3×	no negatives, 3×	$+9.0$	$[-3.4, +21.4]$	0.154	1.00

For the ranking of the eight retrosynthesis models in Table G.1, the interval on Spearman's ρ comes from 5000 bootstrap resamples of the proposals within each retrosynthesis model, which propagates the uncertainty in both the expert accept rates and the system's. Under the same resampling the expert ranking reproduces itself at $\rho = 0.86$, illustrating the sampling uncertainty in the reference ordering.

G Additional Results

Full results by source. Table G.2 gives AUROC with bootstrap confidence intervals and BAcc for every system, including the rule-based baseline and both controlled pairs, and Table G.3 gives the TPR and TNR behind each BAcc. The intervals on ZR and LP are wide because these sources contain only 14 and 32 examples of their minority label.

Generated reactions by method, at each operating point. Table G.4 gives the share of each method's candidates that every system rejects, and the share of the 33 feasible GR candidates it accepts, complementing the threshold-free comparison of Table 3. Systems calibrated on generated pairs accept few feasible GR candidates. On the calibration data, generated candidates are assigned negative labels by construction.

LLM perturbations by producing model. LP was constructed with GPT-5.4 and Claude Opus 4.6, and we evaluate newer models of the same two families. Table G.5 splits LP by the model that produced each perturbation. Neither evaluated LLM has higher BAcc on the perturbations of its own family in this sample.

The verifier family by method. Table G.6 gives the AUROC of the discriminative verifiers of Appendix E.2.2 and the controlled pairs on the candidates of each method. A verifier is often not the best on its own method. For example, the verifier trained on RR candidates scores higher on RA candidates than the one trained on them.

Table F.3: Sensitivity of the verifier models to the checkpoint, in AUROC by source. *Reported*: the fixed budget of 20,000 steps used throughout. *Final*: the last checkpoint of each run. *Oracle*: the checkpoint with the highest average AUROC on FREA, chosen with the benchmark labels and therefore an upper bound that no legitimate rule could attain.

Model	Rule	Step	RP	ZR	LP	GR	Avg.
RA only	Reported	20,000	51.0	72.8	57.9	54.3	59.0
	Final	22,168	50.8	73.6	57.8	54.1	59.1
	Oracle	11,000	53.4	74.2	63.0	54.9	61.4
RR only	Reported	20,000	74.5	87.3	68.2	68.6	74.6
	Final	22,168	74.7	87.6	68.3	68.4	74.7
	Oracle	5000	72.5	88.1	73.9	73.6	77.0
NC only	Reported	20,000	54.1	68.9	60.5	58.3	60.4
	Final	22,168	53.8	69.3	60.7	58.4	60.6
	Oracle	9000	55.4	65.0	62.7	60.7	61.0
FT only	Reported	20,000	52.4	43.6	49.5	43.5	47.2
	Final	22,168	52.3	42.9	49.8	43.4	47.1
	Oracle	6000	53.2	49.9	53.5	48.5	51.3
ET only	Reported	20,000	52.0	50.3	58.1	58.0	54.6
	Final	22,168	52.0	49.1	58.1	57.7	54.2
	Oracle	4000	59.8	63.6	61.2	60.4	61.2
Mixture, 1.42M pairs	Reported	20,000	68.0	75.7	66.8	61.2	67.9
	Final	22,168	67.8	76.0	67.1	60.3	67.8
	Oracle	7000	65.6	90.3	64.5	67.1	71.9
Mixture, 7.10M pairs	Reported	20,000	67.8	79.0	66.0	66.0	69.7
	Final	41,571	66.9	74.0	64.6	61.0	66.6
	Oracle	10,000	66.8	79.4	65.5	73.2	71.2

Table F.4: Frozen decision thresholds. Each maximizes balanced accuracy on the 10,000 held-out generated pairs and is applied unchanged to FREA. Scores are log-likelihoods for the forward models, raw logits for the verifier models, and probabilities for the shallow baseline.

System	Threshold
Descriptor GBDT	0.5101
Chemformer (LL)	−0.0787
BARTSmiles (LL)	−0.0829
BARTSmiles (disc.)	−0.0913
BARTSmiles (LL), extra epoch	−0.0942
BARTSmiles (LL + neg.)	−0.4863
BARTSmiles (LL), extra epoch, 3×	−0.1354
BARTSmiles (LL + neg.), 3×	−0.5464
Mixture, 1.42M pairs	−0.0684
RA only	3.3424
RR only	2.1345
NC only	5.4880
FT only	3.8126
ET only	3.7918

Ranking the retrosynthesis models. Table G.1 gives acceptance rates on RP and the reported Spearman correlations with the expert ranking of the eight retrosynthesis models. Rank correlations cover the LLM decisions and thresholded continuous scores; the top-10 rows report acceptance rate and precision only.

H Data Card

Contents. The release has three parts. The **benchmark** is the 751 expert-labeled reactions of Section 4. Each record holds the canonical reactant and product SMILES, the binary feasibility label, the source and the model or method that produced the reaction, a failure reason for infeasible reactions where one was recorded (Appendix B.4), and free-text notes where the annotators left them. The **generated corpus** contains negative candidates paired with their parent reactions, each carrying the method that produced it, split into training and test portions that share no reaction (Appendices C.5 and D.8). The **protocol** is the annotation guideline given to the chemists, the decision procedure of Appendix B.3, and the failure-reason taxonomy.

Table G.1: Ranking the eight retrosynthesis models by the share of their RP proposals a system accepts. *Accept* is that share over all 282 proposals, against the 69.9% the experts accept, and *Precision* the share of accepted proposals that are feasible. ρ is the Spearman correlation between the system's ranking of the eight models and the experts' ranking; brackets give 95% intervals from 5000 bootstrap resamples of the proposals within each model. Dashes denote unreported top-10 rank correlations.

System	Accept (%)	Precision (%)	ρ vs. experts
Expert labels	*69.9*	*100.0*	—
Claude Opus 5	70.2	81.8	$+0.93$ [$+0.31$, $+0.96$]
GPT-6 Astra	78.7	78.8	$+0.76$ [$+0.26$, $+0.93$]
Chemformer (top-10)	65.2	83.2	—
Chemformer (LL)	51.4	87.6	-0.04 [-0.68, $+0.69$]
BARTSmiles (top-10)	65.2	84.8	—
BARTSmiles (LL)	50.0	87.9	$+0.01$ [-0.67, $+0.67$]
BARTSmiles (LL + neg.)	62.8	81.4	-0.19 [-0.67, $+0.52$]
BARTSmiles (disc.)	68.1	77.6	-0.12 [-0.47, $+0.66$]
Descriptor GBDT	59.6	79.8	-0.07 [-0.54, $+0.69$]
AiZynthFinder filter	53.9	75.0	$+0.17$ [-0.61, $+0.64$]

Table G.2: AUROC with 95% bootstrap confidence intervals (10,000 resamples) and BAcc by source for all systems, including the two controlled pairs of Section 7.

	RP		ZR		LP		GR	
System	AUROC	BAcc	AUROC	BAcc	AUROC	BAcc	AUROC	BAcc
Claude Opus 5	–	69.9	–	64.3	–	76.1	–	61.8
GPT-6 Astra	–	66.8	–	63.7	–	69.6	–	62.7
Chemformer (top-10)	–	70.6	–	67.2	–	63.3	–	50.1
Chemformer (LL)	76.7 [70.6, 82.2]	71.6	82.4 [69.9, 93.0]	70.2	64.5 [50.4, 77.6]	63.8	67.0 [58.8, 74.6]	49.8
BARTSmiles (top-10)	–	73.1	–	74.4	–	59.0	–	57.1
BARTSmiles (LL)	78.9 [73.1, 84.2]	71.5	79.5 [64.8, 92.1]	72.5	66.1 [52.8, 78.6]	60.4	66.9 [57.5, 75.8]	53.4
AiZynthFinder filter	57.0 [49.6, 64.3]	56.6	74.0 [55.7, 89.5]	69.9	62.3 [49.6, 74.3]	57.1	43.1 [33.9, 52.8]	42.6
BARTSmiles (LL + neg.)	77.0 [71.1, 82.4]	67.1	80.8 [67.9, 92.0]	63.7	69.4 [56.8, 80.9]	66.1	73.5 [64.2, 82.4]	54.9
BARTSmiles (disc.)	67.8 [61.0, 74.4]	62.5	79.0 [63.1, 92.1]	60.1	66.0 [53.2, 78.4]	54.8	66.0 [55.1, 76.3]	56.6
Descriptor GBDT	70.3 [63.8, 76.4]	64.0	72.2 [59.4, 83.6]	56.3	75.4 [64.1, 85.3]	70.2	64.7 [53.9, 75.4]	62.4
Rule Check	–	60.0	–	61.5	–	58.0	–	61.3
BARTSmiles (LL), extra epoch	77.8 [71.9, 83.3]	71.5	76.5 [59.9, 90.7]	70.2	66.8 [53.6, 79.3]	58.9	64.5 [55.3, 73.2]	49.1
BARTSmiles (LL), extra epoch, 3×	78.2 [72.5, 83.5]	71.4	78.5 [63.1, 91.6]	74.4	65.3 [52.1, 77.8]	62.6	64.5 [55.3, 73.1]	50.7
BARTSmiles (LL + neg.), 3×	74.2 [68.2, 80.0]	65.5	79.6 [65.6, 91.4]	64.3	69.8 [57.4, 80.9]	65.8	73.5 [64.3, 82.2]	56.6

Collection and labeling. Reactions derive from ORDerly (Wigh et al., 2024), a curated subset of the Open Reaction Database (Kearnes et al., 2021), which draws on patents and published literature. No new experiments were run and no human subjects are involved. Two chemists among the authors labeled reactions without seeing their sources, following the calibration, independent annotation and adjudication procedure in Appendix B.

Intended use. The benchmark supports evaluation of reaction feasibility verifiers under Section 3. The generated corpus supports training and comparisons of negative generation methods, as studied in Section 7. Annotation quality is documented in Appendix B.8; coverage and application scope are described in Appendix I.

Distribution and maintenance. The benchmark is available at `https://huggingface.co/datasets/FREA-benchmark/FREA-benchmark`, and the generated corpus at `https://huggingface.co/datasets/FREA-benchmark/FREA-corpus`. Code is provided in the submission's supplementary material. The benchmark and the generated corpus are released under CC BY 4.0. A changelog will record label corrections, and we expect the benchmark to grow as annotation continues.

Table G.3: TPR and TNR by source at each system's operating point. TPR is the share of feasible reactions accepted, TNR the share of infeasible reactions rejected.

	RP (197/85)		ZR (79/14)		LP (32/55)		GR (33/256)	
System	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Claude Opus 5	82.2	57.6	100.0	28.6	61.3	90.9	33.3	90.2
GPT-6 Astra	88.8	44.7	98.7	28.6	71.9	67.3	51.5	73.8
Chemformer (top-10)	77.7	63.5	98.7	35.7	37.5	89.1	18.2	82.0
Chemformer (LL)	64.5	78.8	97.5	42.9	31.2	96.4	12.1	87.5
BARTSmiles (top-10)	79.2	67.1	98.7	50.0	34.4	83.6	33.3	80.9
BARTSmiles (LL)	62.9	80.0	94.9	50.0	28.1	92.7	18.2	88.7
AiZynthFinder filter	57.9	55.3	68.4	71.4	46.9	67.3	12.1	73.0
BARTSmiles (LL + neg.)	73.1	61.2	98.7	28.6	59.4	72.7	12.1	97.7
BARTSmiles (disc.)	75.6	49.4	98.7	21.4	75.0	34.5	18.2	94.9
Descriptor GBDT	68.0	60.0	91.1	21.4	75.0	65.5	42.4	82.4
Rule Check	88.3	31.8	87.3	35.7	68.8	47.3	90.9	31.6
BARTSmiles (LL), extra epoch	62.9	80.0	97.5	42.9	25.0	92.7	9.1	89.1
BARTSmiles (LL), extra epoch, 3×	67.5	75.3	98.7	50.0	34.4	90.9	15.2	86.3
BARTSmiles (LL + neg.), 3×	72.1	58.8	100.0	28.6	62.5	69.1	15.2	98.0

Table G.4: Behavior on GR at each system's operating point. The first five columns give the share of each methods infeasible candidates that a system rejects (counts in parentheses), the last the share of the 33 feasible GR candidates it accepts. Both depend on where the threshold falls, and Table 3 gives the threshold-free comparison for the systems that return a score.

	Reactant molecules changed			Transformation changed		Feasible
System	RA(50)	RR(53)	NC(54)	FT(44)	ET(55)	(33)
Large language models						
Claude Opus 5	84.0	98.1	83.3	90.9	94.5	33.3
GPT-6 Astra	54.0	96.2	40.7	84.1	94.5	51.5
Forward-model proxies						
Chemformer (top-10)	70.0	98.1	44.4	100.0	100.0	18.2
Chemformer (LL)	82.0	100.0	57.4	100.0	100.0	12.1
BARTSmiles (top-10)	66.0	98.1	42.6	100.0	100.0	33.3
BARTSmiles (LL)	84.0	100.0	61.1	100.0	100.0	18.2
Deployed filter						
AiZynthFinder filter	74.0	64.2	46.3	95.5	89.1	12.1
Trained on our generated negative candidates						
BARTSmiles (LL + neg.)	96.0	94.3	100.0	97.7	100.0	12.1
BARTSmiles (disc.)	94.0	90.6	100.0	93.2	96.4	18.2
Descriptor GBDT	76.0	56.6	100.0	84.1	94.5	42.4

I LIMITATIONS

Evaluation scope. FREA evaluates whether individual proposals admit reasonable conditions under the stated completeness and reactivity criterion, without specifying those conditions or testing the reactions experimentally. A feasible label therefore does not establish a practical procedure, an acceptable yield, or the quality of a complete synthesis route, where compatibility between steps and purification also matter.

Expert judgment and annotation boundaries. Even with a written criterion, chemists may differ on whether conditions are reasonable, a target is isolable, or an omitted component serves a permitted auxiliary role in a particular reaction. Our labels rely on two chemists, whose independent agreement was 77.9% with Cohen's $\kappa = 0.55$ before adjudication. Joint review produces a shared reference, but does not eliminate possible annotation errors or establish agreement with other experts; the final binary labels also do not express the uncertainty of borderline judgments.

Coverage and source composition. All four sources derive from ORDerly's patent and literature records, and screening removes parsing failures and selected structural defects before annotation, so the benchmark does not represent the full range of chemistry or unfiltered planner outputs. The source proportions are chosen for evaluation, rather than estimated from deployment, and the four-source mean need not reflect performance in a particular planning workflow. GR also uses the same

Table G.5: Results on LP split by the LLM that produced the perturbation (feasible/infeasible counts in parentheses). The evaluated LLMs are newer models of the same two families. Neither shows an advantage on perturbations produced by its own family. One Claude Opus 5 response on a GPT-5.4 perturbation could not be parsed and is excluded.

	Produced by GPT-5.4 (15/30)			Produced by Claude Opus 4.6 (17/25)		
System	BAcc	TPR	TNR	BAcc	TPR	TNR
Claude Opus 5	76.9	57.1	96.7	74.4	64.7	84.0
GPT-6 Astra	63.3	46.7	80.0	73.1	94.1	52.0
Chemformer (top-10)	65.0	46.7	83.3	62.7	29.4	96.0
Chemformer (LL)	66.7	40.0	93.3	61.8	23.5	100.0
BARTSmiles (top-10)	53.3	33.3	73.3	65.6	35.3	96.0
BARTSmiles (LL)	61.7	33.3	90.0	59.8	23.5	96.0
AiZynthFinder filter	55.0	46.7	63.3	59.5	47.1	72.0
BARTSmiles (LL + neg.)	66.7	60.0	73.3	65.4	58.8	72.0
BARTSmiles (disc.)	58.3	80.0	36.7	51.3	70.6	32.0
Descriptor GBDT	68.3	80.0	56.7	73.3	70.6	76.0

Table G.6: AUROC on GR by negative candidate generation method for the verifier models of Table 5, and of the two controlled pairs. Each cell compares one method's infeasible candidates with the feasible candidates of the same method (4–13 of them), so single cells are imprecise. Italic cells are a single-method model evaluated on its own method.

Trained on	RA	RR	NC	FT	ET
RA only	*65.5*	55.7	81.5	37.1	71.4
RR only	72.5	*72.2*	30.6	73.8	80.9
NC only	52.5	51.9	*81.0*	61.7	42.7
FT only	61.0	49.5	33.3	*45.8*	72.7
ET only	43.0	74.1	40.7	64.9	*55.0*
Mixture, 1.42M pairs	52.0	93.4	87.5	54.4	83.6
Mixture, 7.10M pairs	45.2	84.0	90.3	65.0	80.9
BARTSmiles (LL), extra epoch	66.5	57.1	25.5	84.3	82.3
BARTSmiles (LL + neg.)	51.7	75.0	92.1	83.0	77.3
BARTSmiles (LL), extra epoch, 3×	68.0	56.6	29.6	83.7	80.9
BARTSmiles (LL + neg.), 3×	54.2	78.8	94.4	82.0	77.7

generation methods as our negative training corpus, so gains on this source do not by themselves demonstrate transfer to model proposals or experimental records.

Sample size and comparison design. Several comparisons depend on small minority classes: ZR has only 14 infeasible reactions, and the per-method GR analyses share 33 feasible reference reactions. Those references come from all five generators, so each per-method AUROC measures discrimination against a mixed reference set and can reflect differences between generators as well as feasibility. The reported confidence intervals condition on the fitted models and final labels; they do not account for annotation errors or variation across training runs.

Generated labels and exploratory training. Generated candidates can remain feasible, and treating them as negatives during training can penalize valid alternatives; differences in candidate purity also prevent us from attributing performance differences solely to the generation method. Our controlled training comparisons use one backbone, one corpus and one seed per setting, and most observed gains are not statistically significant after correction for multiple comparisons. The resulting hypotheses about label reliability and transfer therefore require tests across seeds, architectures and corpora.

Decision rules and model comparisons. Thresholds calibrated on generated pairs may transfer poorly to other sources, so a low acceptance rate at these thresholds does not establish that a model cannot achieve a different precision–recall tradeoff. The LLM results likewise describe two models under one shared prompt, and may change with the prompt or inference settings. The retrosynthesis ranking analysis covers eight models on their sampled proposals at the evaluated operating points, rather than establishing a general ordering of planners or verifiers.