Recovering Method Parameters from Protocol Prose: Task Definition, Audited Extraction, and the Comparability It Licenses

Fabio Rovai
The Tesseract Academy, London, United Kingdom
fabio@thetesseractacademy.com

Abstract

Repositories usually report metadata quality as completeness, the share of fields filled. We measure instead whether two deposited analyses record enough to be compared. Over all 167,331 pairs of gas chromatography analyses in Metabolomics Workbench, 120 pairs meet the necessary conditions for comparing retention indices and two meet those for comparing retention values; one field, the retention index reference series, recorded for 19 of 579 analyses, removes 99.9% of the pairs that share a stationary phase. The oven temperature programme is usually written in the protocol prose instead, and the series sometimes is, so we define their recovery as an extraction task with ordered output and explicit abstention. Audited by two blind LLM raters whose error detection we measure on planted errors, the programme extractor is 65% precise on 40 random MetaboLights recoveries with no fabricated value, recovers in full 49% of the programmes stated in a separate random sample of 40 protocols, and the series extractor is correct on 57 of 66. On the census corpus itself, recovered series raise the pairs that can compare indices from 120 to 171, or from 78 to 92 when the two series must be identical. Of 239 populated Workbench programme fields, 130 hold only a range, a single temperature or a pointer to a paper, yet a completeness audit counts all of them as filled. Every released value carries its source sentence and is attributed to the extracting pipeline, not to the depositor.

1 Introduction

The standard quality report for a scientific data repository is a completeness audit. Fields are enumerated, populated values are counted, a percentage is published, and improvement programmes target the percentage. Minimum information checklists are written in the same currency (Sumner et al., 2007), and compliance with them has been measured across public metabolomics repositories in exactly these terms (Spicer et al., 2017).

A populated field, however, is not a usable field. In a public reference spectral archive of 137,030 spectra (Horai et al., 2010), the collision energy of a mass spectrum is recorded on 84.2% of records and written in 1,023 distinct ways, so joining on it exactly retains 9.3% of the comparisons that a normalised join retains (Rovai, 2026a). Completeness reports 84.2%. The figure a consumer can act on is 9.3%.

What a consumer needs is not whether a field is filled but whether two records can be used together, and for what. That property is relational rather than intrinsic, graded rather than binary, and computable from the record alone. Section 3 makes it precise as four cumulative conditions over pairs of analyses, first reported with a census in Rovai (2026a): of the 167,331 pairs of gas chromatography analyses in the Metabolomics Workbench corpus (Sud et al., 2016), two meet the condition for comparing measured values. Among pairs that share a stationary phase, the collapse is concentrated in a single field, the retention index reference series, which 19 of 579 analyses record.

The contribution of this paper begins where that diagnosis ends. The field is missing from the structured record and present in the text. In the MetaboLights corpus (Haug et al., 2013), none of 355 gas chromatography studies declares a protocol parameter for the oven temperature programme or for the reference series, while their protocol prose usually describes the programme and sometimes the series. The repair for a metadata deficit that no curation campaign has closed is therefore an information extraction problem over documents the repository already stores, and it is measurable in the units that matter, which are comparisons licensed rather than fields filled.

We make five contributions. We define recovery of two chromatographic method parameters from protocol prose as an extraction task with an ordered

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 247–261
November 10, 2026 ©2026 Association for Computational Linguistics

output and a stated correctness criterion (Section 4). We build an extractor whose governing difficulty is exclusion rather than recognition, because the decoy values sit in the same clauses as the targets at similar magnitudes (Section 5). We audit both extractors by reading, separating precision, structural agreement and fabrication, with two blind LLM raters whose error detection we measure on planted errors (Section 6). We close the loop on the census corpus itself, reproducing the census as a control, adjudicating every recovery that changes it, and measuring the result under three readings of the reference-series condition, together with the cost of transferring an extractor between repositories (Section 7). And we release the recovered values with the sentence each came from, attributed in PROV-O to the extracting pipeline, which makes the release an adjudicable evaluation resource (Section 9).

2 Related work

Extraction as archive infrastructure. This workshop has repeatedly addressed the layer this paper works on. Research-data metadata has been generated from citation contexts as a defined task with a constructed dataset (Watanabe et al., 2025), articles have been linked to the data publications that describe them (McGillivray et al., 2025), URL citations have been classified to promote research artefacts (Tsunokake and Matsubara, 2022), property values have been extracted from tables (Phi et al., 2025), leaderboards assembled by linking table entities (Yang et al., 2022), and concepts extracted from four hundred thousand papers (Ting et al., 2025). Extraction has been placed in an archive's ingest path (Susemiehl and Mazzarella, 2025), reproducibility signals read off articles (Akella et al., 2022), and quantities (Bölücü et al., 2023) and bibliographic structure (Joshi et al., 2023) extracted from scientific text. Two features of that work carry over to ours: the deliverable is typically a corpus or a service component rather than a model, and evaluation often has to be built alongside the extractor, because for many of these targets no reference annotation exists in advance.

Defining a target and building its resource. Our framing follows the same practice. The PICO corpus was constructed specifically to support automatic data extraction from biomedical literature (Mutinda et al., 2022), and MuLMS annotates materials-science text in layers for information extraction (Schrader et al., 2023). Both make the bet we make, that the bottleneck is a precisely defined target with adjudicated examples rather than model capacity.

Method and procedure extraction. Reading method detail out of scientific prose is established practice in chemistry and materials science. ChemDataExtractor recovers chemical identifiers, properties and spectra from articles and tables (Swain and Cole, 2016), synthesis procedures have been converted into ordered action sequences (Vaucher et al., 2020), and a text-mined corpus of inorganic synthesis recipes was built from published methods sections (Kononova et al., 2019). Our task differs in two ways that matter for a repository: the output is an ordered numeric programme, so the right numbers in the wrong order are wrong, and the target is a parameter of the instrument behind a deposited measurement, so the extracted value has a determinate consumer, the comparison the repository's schema cannot license.

Why completeness is the wrong target. The checklists and compliance studies of Section 1 (Sumner et al., 2007; Spicer et al., 2017) count fields filled. The community-level frame is the FAIR principles (Wilkinson et al., 2016), whose interoperability criterion is where our measurement bites: a field that is present, populated and written in 1,023 ways is findable and accessible without being interoperable, and a completeness percentage cannot tell the difference.

Graded comparability. The notion we use is imported rather than invented. Psychometrics tests configural, metric and scalar invariance as separate cumulative claims, each licensing a specific comparison (Meredith, 1993). The same problem is solved under other names elsewhere, including calibration transfer in chemometrics (Wang and Kowalski, 1992) and retrospective harmonisation in epidemiology (Fortier et al., 2017). We adopt the psychometric structure because it alone treats comparability as graded and separately testable; its factor-analytic machinery is not needed, since for an instrument record the predicates are decidable from the metadata.

3 Where comparability is lost

A ladder over pairs. We instantiate the psychometric levels for gas chromatography as four cumulative predicates over an unordered pair of analyses

(a, b), following Rovai (2026a). Writing m for the recorded modality, ϕ for the stationary phase family, σ for the retention index reference series, π for the thermal programme, $\perp$ for a value the record does not carry, and a subscript for which of the two analyses a value belongs to:

$$
\begin{aligned}
L_1 &\iff m_a = m_b \\
L_2 &\iff L_1 \wedge \phi_a = \phi_b \neq \perp \\
L_3 &\iff L_2 \wedge \sigma_a, \sigma_b \neq \perp \\
L_4 &\iff L_3 \wedge \pi_a = \pi_b \neq \perp
\end{aligned}
$$

The rung of a pair is $\max\{i : L_i(a, b)\}$. The first rung says the records are about the same kind of measurement, the second is what comparing elution orders needs, the third what comparing retention indices needs, and the fourth what comparing retention values directly needs. Assignment uses only what the record states. There is no imputation, and at this stage no inference from publication text, which is exactly what separates this section from the rest of the paper.

What the predicates assert. Each predicate is a necessary condition for the comparison it names, read off the record, and none is sufficient. L_2 asks for the same phase family because selectivity, and with it elution order, is a property of the phase (Rohrschneider, 1966; McReynolds, 1970); it does not test column length, film thickness or carrier flow. L_3 asks for a reference series on both sides because an index is defined only relative to its series (Kováts, 1958; van den Dool and Kratz, 1963), and as written it asks for presence, not identity. That is permissive in a way that matters. The FiehnLib library indexes retention against fatty acid methyl ester (FAME) markers from C8 to C30 rather than against n-alkanes, and its authors interconvert the two scales with a fitted fifth-degree polynomial rather than treating them as one (Kind et al., 2009); an index computed against FAMEs and one computed against alkanes are comparable only through such a transformation. L_4 adds programme equality, which puts both runs in the same interpolation regime, logarithmic for isothermal operation and linear for temperature-programmed operation, but which regime was applied is recorded for one analysis in the corpus. The ladder is therefore an operational approximation: a pair that fails a rung does not, on its record, support its comparison, and a pair that passes may still not, and Section 7 reports how the result changes when L_3 also requires the two series to be of the same kind, and then to be the same series.

Rung	Pairs at or above	Share
L_1 configural	167,331	100%
L_2 metric, scale	98,479	58.85%
L_3 metric, calibrated	120	0.072%
L_4 scalar	2	0.001%

Table 1: Comparability census over all pairs of gas chromatography analyses in the Metabolomics Workbench corpus (Rovai, 2026a). The step from L_2 to L_3 removes 99.9% of what survived the first step, and its cause is one unrecorded field.

The census. We classified every unordered pair of the 579 gas chromatography analyses in the Metabolomics Workbench corpus, giving 167,331 pairs. Because every pair is classified rather than sampled, the counts in Table 1 are a census and carry no sampling error; the uncertainty in them is definitional, in ϕ, σ and π.

Two pairs in one hundred and sixty-seven thousand meet the necessary conditions for comparing values. Programme equality is tested on the ordered numbers in a free-text field, which can equate programmes whose numbers play different roles and separate identical programmes written in different units; both pairs it finds carry character-identical programme text.

One field does the damage. The losses are unequal. Differing stationary phases remove 68,852 pairs, differing thermal programmes remove 118, and the absence of a retention index reference series on at least one side of the pair removes 98,359. That single field is stated for 19 of 579 analyses, and the convention by which the index was computed is stated for one. The distinction is not pedantry. Kováts interpolates logarithmically and applies to isothermal operation (Kováts, 1958); van den Dool and Kratz interpolate linearly and apply to temperature-programmed operation (van den Dool and Kratz, 1963). Which one is correct depends on how the oven ran, and a bare index does not say, so the reference series and the programme are jointly the calibration.

4 Task definition

The two fields that govern the census are missing from the structured record in different ways in the two repositories we use (Table 2). MetaboLights offers no slot for either: none of its 355 gas chro-

Workbench, 579 analyses	Programme	Series
Key absent, empty or nullish	340	560
Key populated	239	19
yields a structured value	62	15

Table 2: The state of the two fields in Metabolomics Workbench. A programme is structured when the transferred extractor of Section 7 recovers one from the field; a series when the value names both its kind and its carbon span. Of the other four populated series values, two name only a kind and two hold an index range (1000–3100) rather than a series. MetaboLights has no slot for either field in any of its 355 gas chromatography studies.

matography studies declares a protocol parameter for the reference series or for an oven programme, and two declare a parameter named only *Temperature*. Metabolomics Workbench has a key for each, and both keys are mostly absent or insufficient. The reference-series key is absent from 560 of 579 analyses, and of the 239 analyses whose programme field is populated, 130 hold only a range, a single temperature or a pointer to a paper (Section 7). In MetaboLights the programme, and sometimes the series, is written in the protocol prose instead. We therefore define their recovery as an extraction task. A system receives the free-text protocol of one deposited study and returns, for each of two targets, either a structured value or an explicit abstention.

Target 1: the thermal programme. The output is an ordered sequence of steps. Each step carries an oven temperature in degrees Celsius, and optionally a ramp rate in degrees per minute leading to it and a hold duration in minutes at it. Order is part of the output, because the same temperatures in a different sequence describe a different experiment. A returned programme is correct when its steps, their order, and the rates and holds attached to them agree with the protocol as it is written.

Target 2: the retention index reference series. The output is a normalised carbon series, that is, the homologous series against which a reported retention index was computed, with its kind (n-alkane or FAME) and its span, so that two studies naming it differently become comparable. Correctness is agreement in kind and span with the series the protocol states.

Abstention is a first-class output. A protocol that does not state a programme must yield no programme. This matters more than it usually does, because the consumer of the value is a comparability predicate: a fabricated programme does not merely add noise, it promotes a pair to a rung it has not earned and licenses a comparison that is invalid. An omission leaves a visible gap; an invention is indistinguishable from a recorded measurement. We therefore treat the two error directions as unequal throughout, and prefer an extractor that drops a hold time to one that invents one.

Why this is a task and not a cleanup. The two targets are instances of a pattern we expect to generalise: the field that governs whether two records can be combined is frequently the field the depositing schema has no slot for, or a slot nobody fills, while the prose that accompanies the deposit carries it. The task is therefore defined by the consumer rather than by the text. Given a comparability predicate, the extraction targets are whichever of its arguments the record leaves as $\perp$, and success is measured as comparisons licensed rather than as fields filled.

The task is small enough to run as a shared task of the shape this workshop has carried before (Grezes et al., 2022, 2025): the inputs are public, the outputs short and structured, the adjudication protocol is that of Section 6, and the adjudicated items we release are a starting gold standard.

5 Corpus and method

Corpus. We use the 355 gas chromatography studies of the MetaboLights corpus, a public, general-purpose metabolomics repository whose deposits pair machine-readable assay tables with free-text protocol descriptions. The census of Section 3 is over Metabolomics Workbench, and the extractors were developed on MetaboLights. Section 7 then applies them to the Workbench records themselves, so that the effect of recovery on the census is measured on the census corpus and not inferred across repositories.

Extractor. The programme is read as an ordered token stream rather than by independent pattern matches. The walk proceeds left to right through the protocol, emitting temperatures, ramp rates and durations as they are encountered and binding each rate and duration to the temperature it governs by position in the stream. Reading in order is what makes the binding sound. Independent pattern matches carry no relative order between them, so on a multi-ramp programme a rate can bind

Protocol text (study MTBLS3, verbatim, as stored with the recovered value)

column temperature was 110 °C; this was held for 2 min and then increased by 30 °C/min to 310 °C. The column effluent was introduced into a Trace DSQ quadrupole mass spectrometer (Thermo Electron; **transfer line temperature = 320 °C**, **ion source temperature = 240 °C**, *electron beam = 70 eV)*

Step	Ramp	Target	Hold
0		110 °C	2 min
1	30 °C/min	310 °C	

Figure 1: A recovered programme and the sentence it came from, as released. The two temperatures in bold are decoys inside the same evidence window, one above the oven's final temperature and one inside its range, and both must be rejected. Storing the evidence is what makes a single extraction adjudicable by reading rather than by rerunning.

to the following clause's target, and two named methods can be spliced into a single programme that no instrument ran. The procedure is given as pseudocode in Appendix A, so that it can be reimplemented from this paper alone.

Exclusion is the hard part. Recognition of a temperature is trivial and almost all of the engineering is in refusing the wrong ones. Three failure classes were diagnosed and each has a specific mitigation, which we state because the mitigations are the method.

Decoy capture. Oven temperatures share their clauses, at similar magnitudes, with inlet, injector, transfer line, interface, ion source, detector, quadrupole, septum, desorption and incubation temperatures, among others (Appendix A lists all 22 governing terms). An extractor that takes the first temperature after an anchor term returns a programme beginning at the injector. We reject any candidate governed by a decoy term within a bounded window, and verify the rule on cases where a decoy occurs in the same sentence as the target.

Figure 1 shows the difficulty concretely. In that study the transfer line sits at 320 °C, above the oven's final temperature, and the ion source at 240 °C, inside its range, in the clause immediately following the programme. Nothing about magnitude, position or units separates the wanted values from the unwanted ones; only the governing noun does.

Fabrication from adjacency. A study whose final temperature is followed by the clause *"the total run time was 50 min"* yielded a fifty-minute hold that never happened. We exclude durations governed by run-time and ramp-duration language, a rule that caught a second instance immediately on introduction.

Substring collapse. The third class arose in the stationary-phase normaliser of Section 3 rather than in the extractor: a pattern containing `5ms` and bounded only on the right matches inside `DB-35ms`, a chemically different phase, and yields a plausible value rather than an obvious error. Every pattern that reads a chemical identifier is now bounded at both ends.

The reference-series extractor buys precision with recall. It parses a carbon range or list only when a mention of an alkane, a fatty acid methyl ester, a retention index or Kováts appears in a window running 120 characters before the match and 60 after it (Appendix B). That construction is why its precision is high, and it is also its limit: a series described without carbon numbers is not recovered at all. We report this as a design choice rather than a defect, because for a value that will license a comparison we prefer a narrow extractor with a visible miss rate to a broad one with an unknown error rate. The same reasoning governs abstention on two studies whose protocols name two distinct methods: rather than choose between them, the extractor emits nothing and the studies are withheld.

6 Evaluation

Yield. The programme extractor recovers 206 ordered programmes from the 355 studies, 58.0% of the corpus, comprising 529 temperature steps of which 370 carry a hold time, giving 135 distinct programmes. Sixty-six of the 206 are multi-ramp, which matters because every failure of the extractor's first version (Table 4) fell in that class. The series extractor recovers a normalised carbon series for 66 studies, 18.6%. Neither target has a structured field in this corpus, so both figures are gains over nothing that a machine could previously join on.

How often the prose states each target is a separate question, and we have answered it only heuristically. The programme's anchor pattern (oven, column temperature, temperature programme and their variants; Appendix A) matches the protocol text of 279 studies, 78.6%, and a series term (retention index, alkane, Kováts, FAME, fatty acid

Thermal programme	
Studies in corpus (gas chromatography)	355
Anchor phrase in protocol text (regex)	279
Programmes recovered	206
of which multi-ramp	66
Temperature steps recovered	529
of which carrying a hold time	370
Distinct programmes	135
Median initial, final temperature (°C)	60, 315
Median ramp rate (°C/min)	10
Retention index reference series	
Series term in protocol text (regex)	181
Studies with a series recovered	66
By kind: n-alkane / FAME	51 / 21
By form: range / list	63 / 9
Commonest span	C10–C40

Table 3: What the two extractors recovered from 355 MetaboLights gas chromatography studies, none of which has a structured field for either target. Kind and form counts exceed the study count because some protocols state more than one series.

Audit	**Verdict**	**Rate**	**95% CI**
Programmes, MetaboLights, 40 draws			
fully correct	26 of 40	65%	50–78
structural agreement	38 of 40	95%	83–99
free of fabrication	40 of 40	100%	91–100
Programmes, Workbench transfer, 25 draws			
fully correct	22 of 25	88%	70–96
Reference series, every recovery			
MetaboLights	57 of 66	86%	76–93
Workbench	5 of 5	100%	57–100
Development checks, single pass			
programme, final	10 of 14	71%	45–88
programme, tightened	11 of 12	92%	65–99
programme, first	7 of 12	58%	32–81
reference series	12 of 12	100%	76–100
internal std., withdrawn	4 of 12	33%	14–61

Table 4: Every precision audit in this work, with Wilson intervals in per cent (Wilson, 1927). The upper blocks are the adjudicated verdicts. Only the "fully correct" and series rows are precisions; structural agreement and freedom from fabrication are separate properties, defined in Section 6. The series blocks are censuses of every recovery, not samples.

methyl ester) occurs in 181. These are regular-expression counts, not annotations.[1] To measure recall instead, both raters read the full protocols of 40 studies drawn at random from all 355, and agree on all 160 judgements. Of these, 35 state an oven programme; the extractor returns a programme that agrees with the text in every temperature, rate and step order for 18 of them, a recall of 51% (95% interval 36–67; 49% if an omitted hold counts as a miss), and returns a programme for no study that states none. Of the 40, 7 state a reference series and the series extractor finds 5. On the same sample the anchor pattern matches 30 of the 35 studies that state a programme and 1 that does not, so the regex count of 279 undercounts the studies that state a programme. The extractor returns nothing for 17 stated programmes, 5 of them because the anchor never matches, and none of the 18 it returns invents a value.

Table 3 summarises what was recovered. The distributional properties are a weak but useful check that the extractor is reading oven programmes and not something else, since the median recovered programme runs from 60 °C to 315 °C at 10 °C per minute, which is an ordinary temperature-programmed separation.

Audit protocol. Precision cannot be read off the output distribution, which looks reasonable whether or not the values are right. We draw a random sample (seed 20260923) and show the raters described next each extracted value beside the sentence it was taken from, with up to 2,800 characters of the surrounding protocol for MetaboLights items. An extractor that changes is re-audited from a fresh draw, because a verdict on the previous version measures the previous version, and an extraction whose audited precision is poor is withdrawn rather than qualified.

Who adjudicated. Every audited item was judged by two blind LLM raters, Claude Opus 5.5 and Claude Fable 5.1, which agree on 133 of 136 verdicts; quoted protocol text settles the rest. The raters were tested rather than trusted: on 40 items, 24 of them deliberately corrupted, each catches 24 of the 24 with no false alarm, and a mechanical check with no model in it agrees with both on 129 items. The author's own reading of 48 cards agrees on 40 cards with the verdicts as they stood before it, 7 of them planted corruptions it caught, and one of its differences exposed an error both raters shared, now corrected. Appendix E gives the protocol in full, and the development checks at the foot of Table 4 carry no claim.

Three verdicts, kept apart. The programme audits use three separate verdicts. *Correct* means every step, its order, and the rates and holds at-

[1] Rovai (2026a) reports 284 studies for the series. That count also matched *internal standard*, which names a different object; excluding it leaves the 181 here.

MTBLS748, a dropped step

column temperature was set to 70 °C. A temperature ramp of 1 °C/min for 5 min followed by a temperature ramp of 5.63 °C/min to a final temperature of 330 °C with a hold time of 10 min

Output: 70 °C, then 5.63 °C/min to 330 °C, hold 10 min. The first ramp is lost; what survives is correct.

MTBLS12353, a wrong initial temperature

column temperature was maintained at 40 °C, and the injection volume was 2 µL. [. . .] *The oven temperature ramp was held at 90 °C for 1 min, raised to 100 °C at a rate of 25 °C/min*

Output opens at 40 °C, taken from a sentence that is not the programme, and the programme's own opening step at 90 °C for 1 min is dropped.

Figure 2: Two structurally wrong outputs from the 40-draw audit, quoted from their stored evidence. Neither invents a number: one loses a step and one binds the wrong sentence.

tached to it agree with the protocol as it is written, and this is the only programme figure we call precision. *Structural agreement* is weaker and reported separately: every temperature, every ramp rate and the order of every step agree, while a hold time stated in the text may be absent from the output. *Fabrication* means a value appears in the output that is in no reading of the text, and it is reported as its own rate because it is the failure the consumer of these values cannot detect.

On the 40-draw sample the extractor's precision is 26 of 40, 65%, with an interval of 50–78. Structural agreement is 38 of 40, and 0 of 40 outputs contain a fabricated value. Of the outputs that are not correct, 12 differ from the protocol only by an omitted hold time and 2 are structurally wrong. Because a reader cannot check a failure mode that is only described, the structural failures are reproduced in Figure 2. The second is decoy capture in a form the decoy list cannot catch, since the offending phrase is the anchor itself occurring in a sentence about something other than the oven.

Reference series. Every recovery was adjudicated rather than a sample. Of the 66 MetaboLights studies, 57 are correct and 8 get the span wrong; for the remaining 1, the protocol states no reference series. A list interrupted by a concentration or continued with *and*, as in *C22, C28 (each 0.1 mg/ml final concentration; Sigma Aldrich) and C32*, is cut at the interruption; and a mixture the text gives as two ranges, such as *C8-C16: 1 mg/mL, C18-C24: 0.5 mg/mL*, is output as two series. The five Workbench recoveries are taken up in Section 7.

A withdrawn target. A third extractor aimed at internal standard identities reported candidates in 138 studies. A random twelve returned four compounds and eight fragments of English, including *"and vortexed"* and *"see below"*. A precision near one third is not a measurement, so no yield is reported for it and nothing it produced appears in the release. A compound name needs a chemical entity recogniser and a registry lookup (Swain and Cole, 2016), not a pattern over tokens: the targets that shipped are numeric and positional, and the one that failed is nominal.

7 The same corpus, before and after

The question a consumer actually asks is how many more comparisons recovery licenses. We answer it on the census corpus itself.

Control. Before changing anything we recomputed the census of Section 3 from the same cache with the same field derivation. It reproduces exactly: 579 analyses, 167,331 pairs, 98,479 at L_2, 120 at L_3 and 2 at L_4. Had it not, nothing downstream would be reportable.

Transfer is not free, and the reasons are nameable. Applied to Workbench unchanged, the extractor of Section 5 structures 22 of 579 analyses, 3.8%, against 58.0% on MetaboLights. Two causes account for nearly all of the gap and neither is about the science. The extractor first locates the programme in a long protocol using an anchor phrase, and that phrase occurs in only 32 of the 239 populated programme fields, 13.4%, because a repository that gives the programme its own field has already done the locating. And the degree sign is written differently: Workbench records use `U+2103`, and some are mojibake, a UTF-8 degree sign decoded as Latin-1 arriving as `ˆA°`. Bypassing the anchor when the field is itself the window, and normalising the degree sign (Appendix C), raises the yield to 62 of 579 analyses, which is 25.9% of the 239 populated fields. An audit of 25 of those extractions finds 22 correct, 88% with an interval of 70–96 (Table 4); every failure omits a hold written as *minute* or *minutes* in words, which the hold pattern does not match.

Most of the remainder has no programme to extract. Table 5 classifies the 177 populated fields that yield nothing. A temperature range with no

What the populated field holds	n	%
A temperature range, no rate or hold	73	41.2
A single temperature	56	31.6
Other, including inlet-only values	34	19.2
A rate is present, extractor declined	6	3.4
Ramps given as durations, not rates	5	2.8
Nullish text in a populated slot	2	1.1
Defers to a cited publication	1	0.6
Total declined	**177**	

Table 5: The 177 populated Workbench programme fields from which no programme is recovered. Only the fourth and fifth rows, 11 fields, are extractor misses; the first, second and last rows, 130 fields, hold no programme at all.

rate and no hold accounts for 73, a single temperature for 56, and one field defers to a cited paper, reading *"According to temperature gradient. DOI: 10.1385/1-59745-003-0:439"*. Only 11 of the 177 are extractor misses with a recoverable programme present. A completeness audit scores all 239 populated fields as filled, and 130 of them hold no programme at all.

The five recoveries, one by one. The series extractor finds a reference series in the prose of five analyses whose series key is empty, raising the count from 19 to 24. Because the downstream result rests on these records, each is listed in Table 6 with the phrase it came from and its verdict; the full evidence windows are in Appendix D. Every L_3 pair in the census lies within one phase family, 5% phenyl, where 16 analyses state a series. Three of the five recoveries are on that family and add 18 pairs each, three of which are pairs among themselves, for 51 new pairs. The other two move no pair: one is on a family where no other analysis states a series, and the phase of the other is not resolved by the normaliser.

What recovery licenses, under three readings. Under the census's own predicate the pairs at L_3 rise from 120 to 171, an increase of 42.5%. Because L_3 asks only that a series be present on both sides (Section 3), we recomputed it twice more (Table 7). The 16 analyses on 5% phenyl that state a series are 13 stating FAME C8–C30, 2 stating FAME without a span and 1 stating n-alkanes C10–C40; the three recoveries add 1 FAME C8–C30 and 2 n-alkane C8–C40. Requiring the same kind of series separates the alkane analyses from the FAME ones, and requiring the same series leaves the FAME C8–C30 block and the single pair of C8–C40 alkanes. Under the strictest reading the gain is 14 pairs, 17.9%. The direction holds under every reading; its size depends on the predicate as much as on the extraction.

The scalar rung does not genuinely move. Both census L_4 pairs carry character-identical programme text; the transferred extractor recovers both programmes of one pair and neither of the other, whose fields write the degree sign as a superscript zero (`U+2070`), which the normalisation of Appendix C does not map. The recomputed count of 1 is therefore an extractor miss, not a stricter comparison, and on the evidence of this corpus two pairs meet the necessary conditions for a direct comparison of retention values.

8 Why the field is empty

Fixing the schema is the obvious remedy, but across four disciplines with machine-readable formats (Rovai, 2026a) the field that makes a record interpretable is present on 100% of seismic channels, 53.8% of metrology records, 16.6% of meteorological deployments and none of the 355 MetaboLights gas chromatography studies, although in none of the four is the field required (International Federation of Digital Seismograph Networks, 2019; World Meteorological Organization, 2019) and none is validated. We offer a hypothesis these domains suggest rather than establish: a field is complete when its primary consumer cannot complete the primary task without it. A seismometer's counts mean nothing without the transfer function; a chromatogram without its programme still yields peaks, and the programme is needed only to *compare*, which is somebody else's task, later. Recovery from prose is then a realistic remedy, because a methods section is written for a reader who must judge the work.

9 Release

The recovered values are released in version 0.2.0 of the Zenodo deposit (Rovai, 2026b) that accompanies Rovai (2026a), as two graphs in a vocabulary of OWL 2 modules with SHACL (World Wide Web Consortium, 2017) validation layers: 4,256 triples for the 206 thermal programmes and their 529 ordered steps, and 581 triples for the 72 reference series of 66 studies. Every verdict from every rater and check on every audited item, the five Workbench recoveries and the scripts that regenerate every figure in this paper are deposited with them.

Analysis	Recovered series	Evidence (verbatim)	Phase family	New L_3	Verdict
ST001369	FAME C8–C30	*a standard mix of fatty acid methyl esters (FAME C8-C30)*	5% phenyl	18	correct
ST000496	n-alkane C8–C40	*A standard alkane mixture (C8-C40) was injected at the beginning of the analysis*	5% phenyl	18	correct
ST001134	n-alkane C8–C40	*Retention indices (RIs) were calculated relative to those of a C8-C40 alkane series*	5% phenyl	18	correct
ST000980	FAME C8–C30	*a standard mix of fatty acid methyl esters (FAME C8-C30)*	not resolved	0	correct
ST001432	n-alkane C7–C30	*a standard alkane series mixture (C7-C30) was injected before every batch*	100% dimethyl	0	correct

Table 6: Every reference series recovered from prose on the census corpus. New L_3 counts pairs involving the analysis; the three 5% phenyl recoveries share three pairs, so the distinct total is 51.

L_3 requires	Before	After	Gain
a series on both sides	120	171	42.5%
the same kind of series	105	123	17.1%
the same series	78	92	17.9%

Table 7: Pairs at L_3 on the census corpus before and after recovery, under the census predicate and two stricter readings. L_4 is 2 before and 1 after under all three; the lost pair is an extractor miss (see text).

Every recovered value carries the sentence it was extracted from, so any extraction can be adjudicated by reading rather than rerunning, and the adjudicated value and sentence pairs seed a gold standard for the task.

Every recovered value is attributed, in PROV-O (Lebo et al., 2013), to the software agent that extracted it, and carries the time and the extraction run that generated it, rather than being written into a metadata field as though the depositor had stated it. An extraction is a claim about what a text says, and recording it as the depositor's own value launders a pipeline's inference into a scientist's assertion. Attribution keeps the two separable and lets a later extractor supersede an earlier one without rewriting the record it read. A SHACL shape rejects any recovered value that lacks the attribution, the time or the run. Version 0.1.0 of the deposit named the pipeline only in a file comment, and the shape reports 834 violations against it; against the version cited here it reports none, and no recovered value differs between the two.

The extraction graphs are validated on their own. Against all five SHACL layers the programme graph conforms, and two deliberate corruptions, a step with no temperature and a temperature written as text, are each rejected, so the step shapes can fail. The series graph is checked by the provenance shape alone, since no other shape targets its class. The deposit's wider bridge graph, which contains none of the extracted values, has no violations and 531 open warnings, all recording that a percept stored without a stimulus concentration cannot be compared with another panel's; the rule is left firing so that the count can serve as a progress metric, in the manner of crystallography's structured validation alerts (International Union of Crystallography, 2026).

10 Conclusion

A repository's quality report should say what its records can be used for, not how many of its fields are filled. Asked of a complete corpus of gas chromatography analyses, the answer is that two pairs in 167,331 meet the conditions for a direct comparison of values, and that one field, missing from the record and sometimes recoverable from the prose, removes almost every pair that shares a stationary phase. On the same corpus, three recovered series raise the pairs that can compare indices by 42.5%, or 17.9% when the series must be identical, while 130 populated programme fields that a completeness audit counts as filled hold no programme at all.

Limitations

The gain at the calibrated rung rests on three recovered records out of five. Because the rung is a property of a pair, three records move 51 pairs, which is the mechanism working as intended, but the figure is sensitive to those records and to the predicate: the 42.5% under the census predicate is 17.9% when the two series must be identical.

The comparability ladder is an operational approximation. Each predicate is a necessary condition read off the record, and none is sufficient: L_2 and L_4 do not test column dimensions or car-

rier flow, and the census form of L_3 does not test whether the two series are on the same scale. The logic of graded, separately tested levels transfers from psychometrics; the factor-analytic apparatus that establishes those levels there is neither carried across nor needed here.

Recall is measured on a sample of 40 studies, so its interval is wide, and it rests on the same two raters as precision. The 58.0% and 25.9% yields are recovery rates against their denominators, not recall.

The verdicts rest on two LLM raters and a mechanical check, with quoted text settling disagreements. One person's reading checked them and exposed an error both raters shared; on 4 cards the reading and the text disagree. Neither is sufficient alone, and no item was annotated by two people. Planted errors measure detection only for the kinds of error we anticipated.

Four failure modes are diagnosed and deliberately not fixed: holds written as *minute* or *minutes* in words, a degree sign written as a superscript zero, a series list interrupted by a concentration or continued with *and*, and a mixture written as two ranges. Each repair is a small change to a pattern. We leave them because changing the extractor would invalidate the audit just reported, and the paper's own rule is that a changed extractor is re-audited from a fresh draw.

In MetaboLights the stored programme window is 300 characters, which truncates the longest programmes; the raters saw the surrounding protocol text in those cases, but a reader of the release who has only the window may not be able to settle them.

An LLM used as the extractor, scored by the same raters, checks and planted errors, is the obvious comparison and our next experiment.

The extractor is a bounded token walk with lexical exclusion rules, developed on English protocol prose and carrying a decoy list drawn from one instrument configuration. It has been measured on two repositories, and the transfer cost was two concrete assumptions rather than a redesign, but a third repository, another language or another analytical technique would call for its own audit. The one target we explored that is nominal rather than numeric, internal standard identities, needs a chemical entity recogniser with a registry lookup, which marks where this family of methods ends and another begins.

Ethics statement

The work uses only material already published in public repositories, and no human subjects data was collected. One data source used elsewhere in the wider deposit declined further automated requests partway through a crawl its robots policy permitted; we stopped at that point, and we withhold the graph derived from it because the source declares no licence. Recovered values are published as attributed claims of an extracting pipeline and not as statements of the original depositors, so no scientist is recorded as having asserted something they did not write. LLM assistants were used to write and run the analysis code and to draft text, and made the adjudication passes described in Section 6: Claude Opus 5 the development pass, Claude Opus 5.5 and Claude Fable 5.1 the two blind passes. The author read the audit cards described there and is responsible for the content.

References

Akhil Pandey Akella, Hamed Alhoori, and David Koop. 2022. Reproducibility signals in science: A preliminary analysis. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 140–144, Online. Association for Computational Linguistics.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. 2023. Investigating the impact of syntax-enriched transformers on quantity extraction in scientific texts. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 1–13, Bali, Indonesia. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Isabel Fortier, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, et al. 2017. Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, 46:103–105.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, Anjali Pandiri, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset & shared task. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 5–15, Mumbai, India and virtual. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first shared task on detecting entities in the astrophysics

literature (DEAL). In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, et al. 2013. MetaboLights: an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41:D781–D786.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45:703–714.

International Federation of Digital Seismograph Networks. 2019. FDSN StationXML schema, version 1.2. https://www.fdsn.org/xml/station/.

International Union of Crystallography. 2026. checkCIF and the validation reply form. https://checkcif.iucr.org/.

Bikash Joshi, Anthi Symeonidou, Syed Mazin Danish, and Floris Hermsen. 2023. An end-to-end pipeline for bibliography extraction from scientific articles. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 101–106, Bali, Indonesia. Association for Computational Linguistics.

Tobias Kind, Gert Wohlgemuth, Do Yup Lee, Yun Lu, Mine Palazoglu, Sevini Shahbaz, and Oliver Fiehn. 2009. FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry*, 81(24):10038–10048.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6:203.

Ervin Kováts. 1958. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, 41:1915–1932.

Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. PROV-O: The PROV ontology. W3C Recommendation, 30 April 2013, https://www.w3.org/TR/prov-o/.

Barbara McGillivray, Kaveh Aryan, Viola Harperath, Marton Ribary, and Mandy Wigdorowitz. 2025. Finding the paper behind the data: Automatic identification of research articles related to data publications. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 34–43, Mumbai, India and virtual. Association for Computational Linguistics.

W. O. McReynolds. 1970. Characterization of some liquid phases. *Journal of Chromatographic Science*, 8:685–691.

William Meredith. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58:525–543.

Faith Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. PICO corpus: A publicly available corpus to support automatic data extraction from biomedical literature. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 26–31, Online. Association for Computational Linguistics.

Van-Thuy Phi, Dinh-Truong Do, Hoang-An Trieu, and Yuji Matsumoto. 2025. A hybrid LLM and supervised model pipeline for polymer property extraction from tables in scientific literature. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 94–102, Mumbai, India and virtual. Association for Computational Linguistics.

L. Rohrschneider. 1966. Eine Methode zur Charakterisierung von gaschromatographischen Trennflüssigkeiten. *Journal of Chromatography A*, 22:6–22.

Fabio Rovai. 2026a. Comparability in the public olfactory record: volatile measurement, human perception, and the join between them. *Preprint*, arXiv:2609.29702.

Fabio Rovai. 2026b. Olfactory measurement ontology: a vocabulary, conformance suite and graded audit of the public olfactory record. Zenodo deposit, version 0.2.0, doi:10.5281/zenodo.23270951.

Timo Pierre Schrader, Matteo Finco, Stefan Grünewald, Felix Hildebrand, and Annemarie Friedrich. 2023. MuLMS: A multi-layer annotated text corpus for information extraction in the materials science domain. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 84–100, Bali, Indonesia. Association for Computational Linguistics.

Rachel A. Spicer, Reza Salek, and Christoph Steinbeck. 2017. Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data*, 4:170137.

Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, et al. 2016. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44:D463–D470.

Lloyd W. Sumner, Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, et al. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3:211–221.

Nicholas Susemiehl and Joseph Mazzarella. 2025. AI for data ingestion into IPAC archives. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 87–93, Mumbai, India and virtual. Association for Computational Linguistics.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56:1894–1904.

Yuan-Sen Ting, Alberto Accomazzi, Tirthankar Ghosal, Tuan Dung Nguyen, Rui Pan, Zechang Sun, and Tijmen de Haan. 2025. AstroMLab 5: Structured summaries and concept extraction for 400,000 astrophysics papers. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 170–185, Mumbai, India and virtual. Association for Computational Linguistics.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 8–19, Online. Association for Computational Linguistics.

H. van den Dool and P. Dec. Kratz. 1963. A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *Journal of Chromatography A*, 11:463–471.

Alain C. Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H. Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications*, 11:3601.

Yongdong Wang and Bruce R. Kowalski. 1992. Calibration transfer and measurement stability of near-infrared spectrometers. *Applied Spectroscopy*, 46:764–771.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2025. Metadata generation for research data from URL citation contexts in scholarly papers: Task definition and dataset construction. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 72–79, Mumbai, India and virtual. Association for Computational Linguistics.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

World Meteorological Organization. 2019. WIGOS metadata representation, version 1.0. `https://schemas.wmo.int/wmdr/`.

World Wide Web Consortium. 2017. Shapes constraint language (SHACL). W3C Recommendation, `https://www.w3.org/TR/shacl/`.

Sean Yang, Chris Tensmeyer, and Curtis Wigington. 2022. TELIN: Table entity LINker for extracting leaderboards from machine learning publications. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 20–25, Online. Association for Computational Linguistics.

A Extractor pseudocode

The ordered token walk of Section 5, stated so that the extractor can be reimplemented from this paper alone. DECOY and RUNTIME are closed lexical sets. DECOY holds *inlet, injector, injection port, transfer line, interface, ion source, source temp, detector, quadrupole, manifold, septum, MS source, auxiliary, aux, trap temp, desorption, incubat(ion), agitat(ion), storage, freezer, water bath* and *oven of the mass*. RUNTIME holds *run time, analysis time, acquisition time, cycle time, chromatography time, total time, solvent delay, equilibrat(ion), post-run, run lasted* and *during*. The anchor that locates the programme matches *oven* (optionally followed by *temperature* and *program* or *programme*), *column temperature*, *temperature program(me)*, *chromatograph… was programmed* and *thermal program(me)*.

```
input : protocol text T
output: ordered programme P, or abstain

# 1. scan once, left to right
S <- []
for each match m in scan(T):
    if m is a temperature   : S += TEMP(m)
    if m is a rate per min  : S += RATE(m)
    if m is a duration      : S += DUR(m)

# 2. reject governed candidates
for each token t in S:
    if governed(t, DECOY, window)  : drop t
    if t is DUR and
       governed(t, RUNTIME, window): drop t

# 3. bind by position, not by pattern
P <- []
for each TEMP(v) in S, in stream order:
    step <- Step(target = v)
    r <- nearest preceding unbound RATE
    if r exists and no TEMP lies between:
        step.ramp <- r ; mark r bound
    d <- nearest following unbound DUR
    if d exists and no TEMP lies between:
        step.hold <- d ; mark d bound
    P += step

# 4. abstain rather than guess
if T names two distinct methods: abstain
if P is empty                  : abstain
return P
```

B Reference-series pseudocode

```
input : chromatography text T
output: normalised carbon series, or abstain
```

```
out <- []
for each match m of RANGE in T:        # C7-C40
    win <- T[m.start-120 : m.end+60]
   if no ANCHOR in win: continue    # alkane | FAME
                                # | retention index
                                        # | Kovats
    lo, hi <- carbon numbers of m
    if lo >= hi or hi > 60 or lo < 4: continue
    kind <- FAME if FAME in win else n-alkane
    out += series(kind, lo, hi, range, win)

for each match m of LISTED in T:     # C8, C10, C12
    win <- T[m.start-120 : m.end+60]
    if no ANCHOR in win: continue
    nums <- carbon numbers of m in 4..60
    if |nums| < 3: continue
    kind <- FAME if FAME in win else n-alkane
    out += series(kind, min, max, list, win)

return dedup(out) on (kind, lo, hi, form)
```

The sanity bounds are what make abstention the default. A carbon range that runs backwards, exceeds C60 or begins below C4 is not a reference series, and two carbon numbers that merely happen to sit near each other are rejected by the anchor requirement rather than reported with low confidence.

C Transfer pseudocode

The adaptation of Section 7 in full. Only the two marked changes differ from Appendix A: decoy rejection, the run-time exclusion, binding by position and abstention are reused unmodified.

```
input : F, value of a dedicated programme field
output: ordered programme P, or abstain

# changed 1: normalise degree-sign conventions
F <- replace(F, U+2103, "degC")
       # the single-character degree-Celsius
F <- replace(F, "A-circumflex degree", "degree")
       # mojibake: UTF-8 read as Latin-1
F <- replace(F, U+00BA, "degree")
       # masculine ordinal used as a degree sign

# changed 2: the field is the window
win <- F[:460]
       # was: locate with ANCHOR, then window

return walk(tokenise(win), win)   # unchanged
```

D Evidence for the five Workbench recoveries

The evidence window stored with each recovery in Table 6, verbatim and including its truncation at both ends. ST001369 and ST000980 carry the same protocol text.

ST001369 FAME C8–C30
ss range of m/z 50-600 at a rate of 2.7 scan/s. Several IS injections, a standard mix of fatty acid methyl esters (FAME C8-C30), extraction blanks and 3 QCs samples were injected at the

ST000496 n-alkane C8–C40
by a 70 kV electron impact (EI), and 20 scans/sec were recorded over the mass range 85-500. A standard alkane mixture (C8-C40) was injected at the beginning of the analysis for subseque

ST001134 n-alkane C8–C40
aTOF® software package (Leco Corp., St Joseph, MI, USA). Retention indices (RIs) were calculated relative to those of a C8-C40 alkane series. Deconvolution, feature extraction and peak i

ST000980 FAME C8–C30
ss range of m/z 50-600 at a rate of 2.7 scan/s. Several IS injections, a standard mix of fatty acid methyl esters (FAME C8-C30), extraction blanks and 3 QCs samples were injected at the

ST001432 n-alkane C7–C30
detect retention time shifts and facilitate Kovats retention index (RI) calculation, a standard alkane series mixture (C7-C30) was injected before every batch during sample analysis. Th

E How the verdicts were checked

Each of the 136 audited items, the 40- and 25-draw programme samples and every reference-series recovery in both repositories, was judged by two LLM raters, Claude Opus 5.5 and Claude Fable 5.1, each blind to every other verdict. They agree on 65 of 65 programme verdicts ($\kappa = 1.00$) and 68 of 71 series verdicts ($\kappa = 0.71$; Cohen, 1960). Two checks test the raters rather than trust them. A mechanical reconciliation with no model in it lists every temperature, rate and duration in the window the extractor read and compares them with the output; it agrees with both raters on 129 items, and each of the other 7 is settled by words of its protocol (Table 8). A planted-error set of 40 items, 24 of them corrupted in known ways (a dropped hold or step, swapped targets, a fabricated rate or hold, a shifted span, a flipped kind) and shown under opaque identifiers, measures each checker: Opus catches 24 of 24 and Fable 24, both flag all 6 planted fabrications with no false alarm, and the mechanical check catches 22, missing only dropped holds. Where the raters agree, their verdict stands, except the 2 items corrected after the author's check (Appendix F); the 3 items on which they disagree are settled by the quoted words of the protocol (Table 8). A development pass made with an earlier model (Claude Opus 5) agrees with the final verdicts on 65 of 65 programme items. The author then read 48 cards, each asking whether the output agrees with the text on one named point: every non-correct verdict, 16 items found correct

and 8 outputs we had deliberately corrupted. The reading agrees with the verdicts as they stood before it on 40 cards, 7 of them planted corruptions it caught. It also found an error both raters shared, a mixture the text gives as two ranges and the output gives as two series, and the 2 items of that form are counted wrong. Of the other differences, the reading and the quoted text disagree on 4, and 2 turn on whether a marker added at another concentration belongs to the series (Appendix F).

F Every verdict

The author's reading led to two corrections, both items that both raters had judged correct: MTBLS3627 (*"C8-C16: 1 mg/mL, C18-C24: 0.5 mg/mL"*); MTBLS13384 (*"(FAME, C8-C16, C18-C30) for internal retention marker index"*). Where the reading still differs from the final verdicts: MTBLS866 (arguable: *"and C32 (0.4 mg/ml final concentration)"*); MTBLS748 (the reading and the text disagree: *"A temperature ramp of 1 °C/min for 5 min"*); MTBLS12353 (the reading and the text disagree: *"ACQUITY UPLC BEH Amide"*); MTBLS687 (arguable: *"and C32 (0.4 mg/ml final concentration)"*); MTBLS8006 (a planted corruption the reading missed: *"maintained for 4 min"*); MTBLS1228 (the reading and the text disagree: *"standard was added to the solution as internal standard"*); MTBLS7299 (the reading and the text disagree: *"to 305 °C at a rate of 20 °C/min and finally held at 305 °C for 5 min"*).

Every reference-series item that is not correct, or that either rater did not judge correct: MTBLS171 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS208 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS687 (final wrong kind or span; O wrong kind or span, F correct); MTBLS748 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS866 (final wrong kind or span; O wrong kind or span, F correct); MTBLS1228 (final not a series; O correct, F not a series); MTBLS3627 (final wrong kind or span; O correct, F correct); MTBLS9884 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS13384 (final wrong kind or span; O correct, F correct). All other series items are correct under all three.

Item	Mechanical	Opus 5.5	Fable 5.1	Settled	Words of the protocol
MTBLS13035	hold omitted	correct	correct	correct	*"between 5-32 min"*
MTBLS2323	correct	hold omitted	hold omitted	hold omitted	*"at which it was initially held for 4 min"*
MTBLS2324	correct	hold omitted	hold omitted	hold omitted	*"at which it was initially held for 4 min"*
MTBLS1228	correct	correct	not a series	not a series	*"standard was added to the solution as internal standard"*
MTBLS13081	wrong span	correct	correct	correct	*"calculated from an in-house C8–C40 alkane mix"*
MTBLS687	correct	wrong span	correct	wrong span	*"and C32 (0.4 mg/ml"*
MTBLS866	correct	wrong span	correct	wrong span	*"and C32 (0.4 mg/ml"*

Table 8: The items on which the mechanical check and the two raters are not unanimous, each settled by the quoted words of its protocol.

Set	Item	Final	O	F	D	Set	Item	Final	O	F	D
A	MTBLS1778	C	C	C	C	A	MTBLS2021	C	C	C	C
A	MTBLS849	H	H	H	H	A	MTBLS7299	C	C	C	C
A	MTBLS1206	C	C	C	C	A	MTBLS1610	H	H	H	H
A	MTBLS1980	C	C	C	C	A	MTBLS8105	C	C	C	C
A	MTBLS14861	C	C	C	C	A	MTBLS2323	H	H	H	H
A	MTBLS848	H	H	H	H	A	MTBLS10841	C	C	C	C
A	MTBLS2324	H	H	H	H	A	MTBLS532	C	C	C	C
A	MTBLS4894	C	C	C	C	B	ST002871	C	C	C	C
A	MTBLS779	C	C	C	C	B	ST003544	C	C	C	C
A	MTBLS13	H	H	H	H	B	ST003415	C	C	C	C
A	MTBLS12365	C	C	C	C	B	ST002540	C	C	C	C
A	MTBLS15	H	H	H	H	B	ST004201	C	C	C	C
A	MTBLS7006	C	C	C	C	B	ST003863	H	H	H	H
A	MTBLS780	C	C	C	C	B	ST001310	C	C	C	C
A	MTBLS13991	C	C	C	C	B	ST000586	C	C	C	C
A	MTBLS8958	C	C	C	C	B	ST003561	C	C	C	C
A	MTBLS95	C	C	C	C	B	ST004845	C	C	C	C
A	MTBLS4332	H	H	H	H	B	ST004136	C	C	C	C
A	MTBLS12437	C	C	C	C	B	ST004739	C	C	C	C
A	MTBLS13035	C	C	C	C	B	ST004845#2	C	C	C	C
A	MTBLS226	C	C	C	C	B	ST000092	C	C	C	C
A	MTBLS96	C	C	C	C	B	ST004089	C	C	C	C
A	MTBLS1579	H	H	H	H	B	ST004376	H	H	H	H
A	MTBLS808	C	C	C	C	B	ST000054	C	C	C	C
A	MTBLS14	H	H	H	H	B	ST004599	C	C	C	C
A	MTBLS8056	H	H	H	H	B	ST004056	C	C	C	C
A	MTBLS14913	C	C	C	C	B	ST004738	C	C	C	C
A	MTBLS1737	C	C	C	C	B	ST003998	C	C	C	C
A	MTBLS5	H	H	H	H	B	ST004643	C	C	C	C
A	MTBLS748	S	S	S	S	B	ST002993	H	H	H	H
A	MTBLS8006	C	C	C	C	B	ST003544#2	C	C	C	C
A	MTBLS12291	C	C	C	C	B	ST004123	C	C	C	C
A	MTBLS12353	S	S	S	S						

Table 9: Every programme verdict. Set A is the 40-draw MetaboLights audit and set B the 25-draw Workbench transfer audit. Final is the reported verdict; the raters agree on every programme item, so none needed settling. O and F are the blind passes by `claude-opus-5-5` and `claude-fable-5-1`, and D is the development pass. C is correct, H hold omitted, S structural error.