Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing Spectral Evidence Cited by Large Language Models in Pyrazolone Glycoside Assignment

Heba Ahmed Aboukhousa

heba.aboukhousa@mbzuh.ac.ae

Mohamed Bin Zayed University for Humanities

Research Article

Keywords:

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11419388/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Evaluations of large language models (LLMs) in structure elucidation score the answers returned; none of those examined describes checking cited spectral values signal by signal. Every value cited in the justification fields of 72 responses from two LLMs (gpt-5, gemini-3.1-pro-preview), which chose between drawn N- and O-glycoside candidates for two pairs of pyrazolone glycosides from infrared (IR) spectra alone or with NMR spectra, was checked against the labelled values of the supplied spectra and, for the anomeric carbon or proton, against the signal named. Thirteen fields, all with NMR spectra, gave another signal's value for the anomeric signal or a value without a recorded counterpart, and an aryl carbon near 104 ppm, also printed for the sugar-free aglycone, was given as the anomeric carbon in 9 of 12 committed (non-abstaining) responses with the spectra of 117a, whose panels labelled and marked the anomeric carbon (8 of 12 in an independent second coding; Cohen's $\kappa = 0.72$ over the 36 NMR-added responses, 0.32 over the 22 that contain its value). On the 2-deoxy pair, whose NMR items proved to combine spectra of two compounds, the count was 10 of 18 (19 of 35 overall); a post hoc IR feature, coded once like the field check, was named in 15 of 18 IR-only and 1 of 18 NMR-added responses on the ribose pair, a difference this design cannot attribute to the NMR spectra; and agreement with the thesis assignments, which leave one riboside's attachment site and the 2-deoxy assignment open, is secondary.

1 Introduction

Large language models (LLMs) are now evaluated on structure elucidation from spectra. The benchmarks and evaluations examined for this work score the answer returned, a structure, a name or SMILES string, or the answer to a question, in some cases together with scored spectrum-interpretation sub-tasks or an analysis of error types.[1–7] One quotes the explanations of a model in two case studies, and another notes that reasoning traces need not describe how an answer was produced; none of them describes a signal-by-signal comparison of the spectral values that a model cites with the values present in the spectra it was given (Table S37; sections, tables and figures whose numbers begin with S are in the Supporting Information, Supporting File 1). For a regioisomer question this matters: a correct label justified by the discriminating observation and one justified by a misassigned signal receive the same score. The aim of this paper is therefore narrow. For two pairs of isomeric glycosides, assigned in a doctoral thesis as N- and O-glycosides, it takes the one field of each response that the prompt reserves for the decisive feature and records which observation two LLMs cite there, whether each value cited is among the labelled values of the supplied spectra, and whether a value cited as the anomeric carbon is that carbon. The audit concerns the values cited in that field; it is not an assessment of whether the reasoning built on them is valid. For one question, the identity of the signal given as the anomeric carbon, the complete text of the responses that had nuclear magnetic resonance (NMR) spectra available was read as well.

The test case is a regiochemical assignment from nucleoside chemistry. 4-Arylhydrazono-5-(trifluoromethyl)pyrazol-3-ones can be glycosylated at either ring nitrogen or at the oxygen of the lactam group, and the products are constitutional isomers of identical molecular formula (Fig. 1). The glycosides examined here were prepared by silyl-Hilbert–Johnson (Vorbrüggen) glycosylation,[8,9] and the closest precedents for this scaffold report N-linked products: ribofuranosyl nucleosides,[10] a 2-(tri-O-acetyl-β-D-

ribofuranosyl) derivative[11] with the anomeric carbon assigned at 84.4 ppm, and N-glucosides of arylhydrazono analogues that include the 3-fluorophenyl aglycone studied here.[12] O-Glycosylation has been reported on the same scaffold as part of N,O-bis-glucosides.[13] As far as their titles, abstracts and the data consulted show (Table S23), none of these reports characterises mono-glycosides of one aglycone and one sugar attached at different sites, so none shows directly which observations separate the connectivities.

The isomers cannot be told apart by molecular formula or by the mass of the molecular ion, and structure-similarity measures do not encode which isomer is correct; their values for the two candidate structures as drawn for the models range from 0.54 to 0.94, and from 0.35 to 0.87 when the O-isomer is encoded in its arylazo form (Table 1). What distinguishes the isomers has to be read from the spectra. Only the N2-glycoside can retain the hydrazone–lactam system of the parent pyrazolone; the O-glycoside has no ring carbonyl group, and an N1-glycoside can exist only in arylazo forms (Fig. 1).

Such an audit needs assignments whose support is known. The compounds and spectra used here were reported in the author's doctoral thesis, deposited in an institutional repository in 2017 and not published in journal form[9] (hereafter, the source). Establishing how far the recorded data support the assignments made there was a precondition of the audit, and it altered the design: the support differs between the compounds, for one pair the recorded data do not establish the assignment, and for the other they leave one of the two assignments open (Section 3.1). Agreement of a response with the source assignments is therefore a secondary result here. The status of the source assignments is summarised first (Section 3.1; the full re-examination is in Section S17) and the evaluation follows (Sections 3.2–3.4).

The source is cited as a primary record, as an external dataset would be; its syntheses, biological assays and structural conclusions are not reproduced or claimed. The contribution is an audit of the evidence that the models cite. It checks every value in the justification field of 72 responses against the values recorded for the supplied spectra, and whether a value given for the anomeric carbon or proton is that signal; it inspects what each stimulus image displays; and it reads the complete responses for one question, which a chemist who had not seen the first coding coded a second time (Sections 2.3, 2.4 and 3.2–3.4). Re-examining the source record was a necessary by-product. It leaves open the attachment site of one riboside and the assignment and ring size of the 2-deoxy pair, and it shows from original records that neither ^{13}C spectrum of that pair is printed with the compound whose ^{1}H spectrum is printed under the same number, the attribution of the spectrum printed for 123a being shown directly and that of the spectrum printed for 122a following by elimination (Section 3.1; Sections S16 and S17). The records are deposited with checksums,[14] apart from those listed as not available in the Data and Software Availability section. A re-analysis of the reaction records of the source, which does not bear on the evaluation, is given in Section S17.1.

2 Methods

2.1 Source data, compounds and numbering

All synthetic and spectroscopic data originate in the source.[9] Two pairs of glycosides of 4-[2-(3-fluorophenyl)hydrazinylidene]-5-(trifluoromethyl)-2,4-dihydro-3*H*-pyrazol-3-one (named 4-(3'-fluorophenylhydrazono)-5-trifluoromethyl-2,4-dihydropyrazol-3-one in the source and, in shortened form, in the abstract) were used: the ribose pair 116a/117a (2,3,5-tri-O-acetyl-D-ribosides) and the 2-deoxyribose pair 122a/123a (di-O-acetyl-2-deoxy-D-ribosides). The source assigns 116a and 122a as O-glycosides (3-glycosyloxy derivatives) and 117a and 123a as N-glycosides attached at N2 (2-glycosyl derivatives). For the evaluation the compounds were coded PX-A02 (116a), PX-A01 (117a), PX-B01 (122a) and PX-B02 (123a). Two compounds assigned in the source as O- and N-benzoyl derivatives of the 3-nitrophenyl analogue, 124f (PX-C02) and 127c (PX-C01), were used for one supplementary task (Section 3.4). Infrared wavenumbers and ^{13}C shifts are the peak labels printed on the source figures; ^{1}H shifts are those listed in the experimental section of the source. According to the source, the infrared spectra were recorded as KBr discs and the NMR spectra on a 400 MHz spectrometer (100 MHz for ^{13}C), in $CDCl_3$ for the four glycosides; the spacing of the labelled solvent lines agrees with that frequency (Section S8). The characterisation reported for the compounds, including their mass spectra, is given in the source (its Section 2.2 and Figs. 18–30 and 39–47). After the evaluation the chemical part of the source (its Sections 2.2 and 2.3, with their figures) was read again in full for this article, including the spectra that had not been supplied to the models; what that reading adds is stated in Sections 3.1 and 3.2 and in the Supporting Information (Sections S15 and S17). The original spectrometer data that the author (H.A.A.) holds, 56 one-dimensional data sets of January and February 2016 (47 of them different), were made available after the text had been written; they include a ^{13}C data set that bears the file name printed on the ^{13}C figure of 117a and reproduces its labelled lines, but no two-dimensional spectrum and no ^{13}C data of 116a, 122a or 123a. The data sets bearing on the four compounds were reprocessed, and what they add is stated in Sections 3.1 and 3.2 and in the Supporting Information (Section S16). Scans of spectrometer printouts of 2016 and processed files of 2014, found afterwards among her files, hold ^{1}H records of both 2-deoxy compounds and ^{13}C records of one of them; they show that the ^{13}C spectra of this pair are printed in the source with the wrong compound (Section 3.1; Section S16.3). Atom numbering follows Fig. 1. Digitisation and the later re-examination raised fourteen points that limit what may be claimed; they are listed with their treatment in the Supporting Information (Section S6.1) as audit items A1–A14.

Table 1 gives standard structure-level similarity measures for the ribose pair, computed with RDKit[15] (release 2026.03.6; Morgan fingerprints[16] of radius 2 and 3, atom-pair and topological-torsion fingerprints, all as 2048-bit hashed bit vectors without chirality; MACCS keys;[17] Tanimoto coefficient). The values depend on the tautomer encoded, so both encodings are given; the SMILES strings are listed in the Supporting Information (Section S8).

Table 1
Tanimoto similarity of the two candidate structures (N2- and O-riboside) under two tautomer encodings of the O-isomer. Both candidates, and the N1-riboside, have the same formula, $C_{21}H_{20}F_4N_4O_8$, the same neutral monoisotopic mass (532.1217 Da) and the same calculated *m/z* for $[M + H]^+$ (533.1290); their InChIKeys[18] differ.

Encoding of the O-isomer	MACCS keys	Atom pair	Morgan, radius 2	Morgan, radius 3	Topological torsion
Hydrazone (4*H*-pyrazole) form	0.937	0.785	0.654	0.536	0.590
Arylazo (1*H*-pyrazole) form	0.866	0.725	0.433	0.349	0.464

2.2 Source assignments, their status, and the outcomes recorded

The assignments against which answers are compared are those of the source (Table 2), and are called source labels below. They were fixed before any model was run and were held separately from every model input. The source reports no crystal structure, annotates or describes no heteronuclear correlation from the anomeric proton to the aglycone, and contains no computed-shift comparison, so the assignments are inferential. Two kinds of outcome are therefore kept apart. The primary outcomes are what a response cites as decisive, whether each cited value is among the values recorded for the supplied spectra and, for a value cited as the anomeric carbon, which signal it is; they do not depend on the source labels, and all three are reported for both pairs. The count over the complete response extends the third of them beyond the field and was introduced afterwards (step (v) below); within the field the aryl carbon is given as the anomeric carbon in 10 of 35 committed NMR-added responses (6 of 12 with the spectra of 117a). Agreement of the answer with the source label is a secondary outcome, conditional on a label that for one compound of the ribose pair is not established (Section 3.1). Thirteen decisions were taken after responses had been seen. (i) When re-examination showed how unevenly the assignments are supported, the 2-deoxy pair was treated as open: agreement with its labels is tabulated as agreement with the source labels and is not interpreted as chemical accuracy. Infrared naming, which does not depend on the assignment, is reported for each pair (Section 3.3). (ii) The source label of 116a was restated as open between attachment at O and at N1 (Section 3.1). (iii) Half of the conditions were added after the first batch of responses (Section 2.3). (iv) The coding of Section 2.4 replaced an earlier script-based coding, and the infrared criterion of Table 2, a labelled band at 1670–1690 cm^{-1}, was formulated from the four spectra themselves; it is an operational description of these spectra and differs from the criterion stated in the source, the absence of absorption at 1640–1690 cm^{-1} (audit item A10; Section S6.1). (v) The reading of the complete response for the anomeric-carbon question, with its decision categories, was introduced after the field coding (Section 2.4). (vi) The stimulus images were inspected after the coding (Section 2.3). (vii) The second coding was carried out after the first coding had been completed, and the decision not to use the return of one coder for the anomeric-carbon question was taken after that return had been seen (Section 2.4). (viii) The chemical part of the source and those of its figures that had not been supplied to the models were re-examined after the first coding and the inspection of the images; the observations on its

heteronuclear single-quantum coherence (HSQC) figures (Sections 3.1 and 3.2) were made at that stage and played no part in the design of the items or in the coding. (ix) Original records of the 2-deoxy pair, found in the author's files after the text had been written, showed that the ^{13}C spectra of this pair are printed in the source with the wrong compound (Section 3.1); the stimuli, the value tables, the coding and the counts are those of the panels as supplied and were not changed. (x) The stricter reading of infrared naming (Section 3.3), the division of the anomeric-carbon count by what the panels displayed, its denominator of 30 responses that argue from the ^{13}C shift of the anomeric carbon, and the count of 16 flags obtained when the reading of the *Reasoning* field is applied to the field check (Section 3.2) were defined after the counts had been made. (xi) The variation of the tolerances and the check that no cited value is in a value table without being displayed (Section 2.4) were made after the coding. (xii) After the first return of coder A, his task on the infrared feature was restated to him in a message that also told him how his answers to it had appeared (Sections 2.4 and S14). (xiii) The pilot, seen before the main evaluation, had already produced two responses that took the aromatic CH near 104 ppm for the anomeric carbon (Section S3.3), so the anomeric-carbon question was not formulated without knowledge of that pattern. No protocol was registered before the evaluation; the study is exploratory, and its counts describe what was found without testing a hypothesis fixed in advance.

Table 2

Source assignments, the operational infrared feature (formulated after the responses had been collected, from these four spectra; Section 2.2), and the status of each assignment after re-examination. Candidate letters are those of the drawings supplied to the models. COSY, correlation spectroscopy.

Compound	Source assignment (candidate)	Operational infrared feature	Status
PX-A01 (117a)	N-riboside (A)	band at 1680.47 cm^{-1}	consistent with the N2-glycoside in its hydrazone–lactam form (Section 3.1)
PX-A02 (116a)	O-riboside (B)	no labelled band at 1670–1690 cm^{-1}	arylazo form favoured; attachment at O or at N1 open, N2 disfavoured but not excluded (Section 3.1)
PX-B01 (122a)	O-deoxyriboside (A)	no labelled band at 1670–1690 cm^{-1}	open; the ^{13}C spectrum printed for it is not that of the compound whose ^{1}H, COSY and HSQC spectra are printed for it, and is attributed, by elimination, to the compound whose ^{1}H spectrum is printed for 123a; the corrected record favours an N-glycoside; ring size and anomeric configuration not established (Section 3.1)
PX-B02 (123a)	N-deoxyriboside (B)	band at 1679.91 cm^{-1}	open; aglycone signals of the hydrazone–lactam type (^{13}C on the attribution by elimination); the ^{13}C spectrum printed for it is, by the original records, that of the compound whose ^{1}H, COSY and HSQC spectra are printed for 122a; ring size and anomeric configuration not established (Section 3.1)

2.3 Evaluation design: a forced choice between two drawn candidates

Stimuli. Spectra were supplied as image crops of the source figures: the infrared spectrum over about 1950–1030 cm^{-1} (1500 is the only complete axis label inside the crop), the ^{1}H and ^{13}C spectra and one two-dimensional spectrum where the source contains any (gradient heteronuclear multiple-bond correlation, gHMBC, for 116a and 117a; HSQC for 122a; none for 123a). The source also contains a correlation spectroscopy (COSY) and an HSQC spectrum of 117a and a COSY spectrum of 122a, which were not supplied. Structure drawings, panel letters and embedded data-file names were masked. An identifier band was added above every NMR panel, showing the file stem in the single-case items (for example PX-A01_13C) and the compound number and kind of spectrum in the paired items; the infrared images of the paired infrared-only items are headed Spectrum 1 or Spectrum 2, those of the paired NMR-added items carry the band “Compound 1 · IR” or “Compound 2 · IR”, and the single-case infrared images and the candidate drawings carry no identifier band.

The crops do not all display the same things. The deposited images were inspected for this article after the responses had been collected and coded (Section S13), and six properties bear on the results. (i) The gHMBC panels keep the source’s annotations C-1" to C-5" beside a carbon axis that has no printed scale inside the crop, so they mark which row is the anomeric carbon without giving its shift (Fig. 3, broken frames). (ii) The HSQC panel of 122a contains the ^{1}H projection, on which the proton labels 3" and 4" of the source are legible, and an empty contour region: the crop covers the upper left of the source figure, and every cross-peak of that figure lies outside it. No correlation is visible in it, although the prompts of the items concerned announce an HSQC spectrum. (iii) The ^{13}C panel of 123a shows the range 96–172 ppm and an inset of the solvent region, which contains the solvent triplet and two further labelled lines (77.21 and 76.50 ppm); most of the peak labels of the main trace are cut off at the upper edge (those of 112.673 and 103.831 ppm are complete), and apart from the two further lines of the inset no sugar carbon of that spectrum is displayed. The ^{13}C panels of the other three compounds show the oxygenated sugar carbons with their labels; that of 122a ends near 35 ppm, so its C-2" is not displayed. (iv) The ^{1}H panels carry peak labels and integrals. (v) The candidate drawings show the N2-glycoside in the hydrazone–lactam form and the O-glycoside as the 4-hydrazono-4*H*-pyrazole tautomer, so both candidates are drawn with a hydrazone N–H; in the drawing of the N2-riboside the 2"-O-acetyl group overlaps the ring carbonyl group. Both candidate drawings of the 2-deoxy pair show a furanosyl ring (Section 3.1). (vi) The ^{13}C panels of 122a and 123a were cut from the figures printed for these compounds in the source. Original records found after the evaluation show that the figure printed for 123a is the spectrum of the compound whose ^{1}H and HSQC spectra are printed for 122a, and that the figure printed for 122a is not a spectrum of that compound (it is attributed, by elimination, to the other compound; Section 3.1); in the NMR-added items of this pair the ^{1}H panel and the ^{13}C panel presented for a compound therefore came from different compounds. Supplying images rather than peak lists means that performance combines the reading of a spectrum with chemical reasoning, and the two cannot be separated in this design. The 41 images, with pixel dimensions and checksums, are listed in the Supporting Information (Section S2).

Systems and requests. A pilot through the chat interfaces of two services (22–26 September 2026; 75 responses) was used to correct the stimuli; it is described in the Supporting Information (Section S3.3) and contributes to no count in this article. The main evaluation (27–28 September 2026) used the providers' application programming interfaces (APIs) with the model identifiers gpt-5-2025-08-07 (OpenAI; referred to as gpt-5) and gemini-3.1-pro-preview (Google; referred to as Gemini); a model addressed through its hosted interface is called a system below. Each response came from a single request with no prior turns. The first identifier is a dated snapshot; the second is a preview identifier, and the identifier alone does not show that the model behind it was unchanged between requests. The identifier returned by each service was logged for every request that returned a response, together with a timestamp, a hash of the prompt file and the ordered list of attached images. Each request consisted of one user message that contained the images, in the order recorded in the run log, followed by the text of the prompt; no system message was sent, and the images were sent as image data without file names. No temperature, reasoning, output-length or tool parameter was sent (the run log records the sampling setting of all 90 requests as default), and the image-detail setting of gpt-5 was left at its default, so these were at the providers' defaults; neither finish reasons nor reasoning-token counts were logged, and the internal reasoning of the systems was neither requested nor recorded; the default values in force and the versions of the client libraries were not recorded. Repeating the requests cannot be expected to reproduce the responses, which are deposited verbatim. The request code is deposited and is the authoritative record of what was sent.[14]

Conditions. On each glycoside pair, two presentations were crossed with two levels of evidence, with three replicates per system: a single compound with both candidate drawings, or both compounds with both candidate drawings; and the infrared spectrum alone, or the infrared spectrum together with the NMR spectra. This gives 36 responses per pair and 72 in all, with 18 further infrared-only responses on the benzoyl pair. A paired response assigns two compounds jointly and is counted as agreeing only if both assignments agree; single-case and paired responses are therefore different units and are reported separately, and any sum over them is labelled as a descriptive total. The conditions were collected in two batches. The paired infrared-only and the single-case NMR-added conditions were run first; the single-case infrared-only, the paired NMR-added and the benzoyl conditions were designed after the first batch had been collected and were run later the same day for gpt-5 and, for Gemini, begun later the same day and completed on the following day after an interruption (Section S3.4). The requests made after the interruption include all six Gemini responses to the paired NMR-added items; any error rows written at the interruption were removed before the run was resumed, and whether a request was repeated in an interval of 26 minutes within that group is not recorded. Within each presentation the level of evidence therefore coincides with the batch, and replicates of a condition were requested consecutively.

Labels and order. The mapping of candidate letters and of spectrum order to the isomers was fixed for each pair and was not counterbalanced (Table S2). In every single-case item with identifier suffix 01 the source label is candidate A and with suffix 02 candidate B; in both glycoside pairs the first spectrum or compound is the one assigned as the O-isomer. A constant answer would thus agree with the source label on half of the single-case items, and for the ribose pair a default that assigns the first compound to candidate A and the second to candidate B gives the inverted assignment. In the single-case NMR-added items the suffix was visible in the identifier band of the NMR panels; the prompts forbid inferring identity

from file names, and the answers do not follow the suffix (8 of the 24 responses to these items agree with the source label; Section 3.3). Effects of label and order cannot be separated from effects of content with the present data. The letters returned differ between conditions: in the single-case infrared-only items candidate B was chosen in 16 of the 21 committed responses, in the single-case NMR-added items Gemini chose A in 10 of 12, and in the paired items "A / B" was returned in 19 of 24, among them all 12 NMR-added responses.

Prompts. Each of the nine prompt files was fixed before the first request of its condition and was not changed afterwards; the four files of the first batch preceded all API responses, and the five of the second batch were derived from them. They are reproduced in the Supporting Information (Section S1). All of them tell the model to use only the evidence supplied, not to infer identity from file names or external lookup, to identify "the decisive spectral feature", and to report a confidence between 0 and 100; all offer UNDETERMINED as an answer. They state that the candidates "differ only in the atom through which the sugar is attached" ("the benzoyl group" in the two benzoyl files); chemically the two drawings differ also in the arrangement of double bonds (a ring carbonyl group in one, a C-3 = N bond in the other), which is the difference that the infrared feature reflects. The paired prompts state that one compound is candidate A and the other candidate B. Each item is therefore a forced choice between two drawn candidates, with UNDETERMINED allowed, and not open structure elucidation; neither an N1-glycoside nor a pyranosyl form was among the candidates (Section 3.1). The single case prompts additionally ask the model to explain any ambiguity and contain a *Limitations* field, and they forbid inventing "peaks, correlations, or experimental details"; the paired prompts have neither the ambiguity instruction nor the *Limitations* field, and their prohibition names peaks and experimental details only. A probe through the chat interfaces, made before the API runs, of whether the services could reproduce thesis-specific facts without the spectra elicited none in the nine responses retained; three further responses were not retained, and a note made at the time, which cannot be verified and is not used, records that one of them gave a table of yields and isomer ratios without a source (Section S5). This does not show that the source was absent from training data.

2.4 Coding of the responses

The unit coded is the *Decisive evidence* field (the justification field) of a response, which is reproduced verbatim for all 90 responses in the Supporting Information (Table S4). Four things were recorded for each of the 72 glycoside responses, and a fifth for the 36 that had NMR spectra available.

Naming of the infrared feature. The field names the operational infrared feature if it cites a band within 5 cm^{-1} of the 1680 cm^{-1} band as present in the compound that shows it, or states that carbonyl absorption is absent from a range that includes 1670–1690 cm^{-1} or that the compound has no ring or lactam carbonyl band, or states that the spectrum contains a single, ester-only carbonyl absorption. A field that asserts a second, non-ester carbonyl elsewhere, for example at 1639 or 1646 cm^{-1}, does not qualify under the second or third of these criteria. For 122a a statement that absorption is absent from such a range is coded as naming the feature, although the trace has an unlabelled shoulder there; the same holds for 116a, in whose trace both second coders reported a shoulder (Section 3.1). Naming is recorded whether or not the answer agrees with the source label and whether or not the feature is offered as the basis of the answer.

Anomeric-carbon citation. Where the field cites a ^{13}C value as the anomeric carbon, the value is compared with the labelled values of the panels (tolerance 0.6 ppm) and recorded as the carbon assigned C-1" in the source, as the aryl carbon near 104 ppm, or as a value not among the labelled values.

Field check. Every value cited in the field was compared with the value table of the compound concerned: the peak labels of its infrared and ^{13}C panels and the ^{1}H list of the source (tolerances 5 cm^{-1}, 0.6 ppm and 0.1 ppm); the two-dimensional panels have no value table. The value table is that of the source figure: cited values are checked against the values recorded for the supplied panels, not against what the cropped image displays, which is assessed separately (Section 2.3). Where a crop does not display part of the figure (the sugar carbons of 123a), a value could have a counterpart in the table without being visible in the image; no field cites such a value. A field is flagged if a cited value has no counterpart in the table, or if the value it gives for the anomeric carbon or the anomeric proton belongs to another signal. The anomeric carbon is the one assigned C-1" in the source (for the 2-deoxy pair, in the list of the compound under whose number the panel was supplied; Section 3.1); the signal near 104 ppm is assigned to the aryl ring here (Section 3.2). The identity of other signals is not checked, because for the infrared bands and the low-field proton the recorded data do not establish one; a field can therefore be flagged for a misattributed signal only if it cites NMR data, and an infrared-only field only for a value with no counterpart. A field that cites no value is recorded as such. Separately, a consistency note is made where a field contains a statement that contradicts another statement of the field or the answer given, or places the absent carbonyl band in a range that does not contain 1670–1690 cm^{-1}; such a note concerns the coherence of the field and not its values. The check was applied to all 72 responses, whether or not they agree with the source label. The tolerances allow for values that the responses round or give as approximate (for example “~104 ppm”); the 13 flags are the same at tolerances between 3.5 and 25 cm^{-1} and between 0.25 and 1.0 ppm, and below 3.5 cm^{-1} one further field would be flagged (Section S4.7). For the 2-deoxy pair the C-1" of the list of the source ceased to be a valid reference with the findings of Section 3.1; the one field affected is noted in Section 3.2.

Agreement with the source label. A response agrees, disagrees or abstains (UNDETERMINED). Abstentions are counted separately and remain in the denominators; a response that does not abstain is called committed.

Reading of the complete response. One question was put to the complete text of all 36 NMR-added responses: whether the response gives the aryl carbon near 104 ppm as the anomeric carbon. Each committed response was recorded as doing so in the *Decisive evidence* field, as doing so in the *Reasoning* field, as not doing so, or as discussing no anomeric carbon, and each decision is recorded with the sentence of the response on which it rests (Table S24; Supporting File 2). A first reading was prepared with the artificial intelligence (AI) assistant (Section 2.6). The assistant then repeated the reading twice, in separate sessions that received the responses under random identifiers and not the first decisions; both repetitions returned the category of the first reading for all 36 responses (Section S12). These repetitions are a check of consistency by the same assistant and not a human check. H.A.A. checked every decision against the response; she declares that she made this check without having looked at the codings returned by the second coders. She agreed with 35 of the 36 decisions; for the one abstention she entered “no

anomeric carbon discussed", the response remains recorded as an abstention, and no count changes (Section S12). The first reading was not blind to system or condition; the sentences are reproduced so that a reader can verify all 36 decisions.

The classification of every response is listed in the Supporting Information (Table S9), and the counts of infrared naming and the classification of the anomeric-carbon citations do not change when the tolerances are varied between 1 and 25 cm^{-1} and between 0.25 and 1.0 ppm (Section S4.7). The rule for the infrared feature is also given as a short script (Supporting File 3; Section S4.6). Applied to the 72 fields it returns the same named or not-named outcome as the first coding, on which it was tuned. It implements that one rule and checks nothing else: not the flags of the field check, not the identity of the aryl carbon, not the anomeric-carbon counts of Section 3.2 and not the chemistry. A second script (Supporting File 4) recomputes the counts of Sections 3.2 and 3.3 from the coding file (Supporting File 2); it tabulates the codes and does not derive them. The first coding was done by H.A.A. from the verbatim fields, starting from a first-pass classification prepared with an AI assistant (Section 2.6).

Second coding. Two chemists experienced in the interpretation of NMR and infrared spectra joined the study for this purpose. Both are members of the department in which the source was prepared; neither supervised nor examined it, and neither is an author of this article (both had been invited as co-authors before they coded and declined after reading the manuscript; Section S14). Neither had taken part in its design, in the collection of the responses or in the first coding, and both declared that they had not seen the spectra or the compounds before. They are called coder A and coder B below; the letters were settled by lot (Section S14), and at the request of one of them the returns are not attributed by name. Both received the same workbook. It contained the infrared and ^{13}C panels of the four compounds, the two-dimensional panel of three of them and the candidate drawings, as supplied to the systems (the ^{1}H panels were not included), the 72 *Decisive evidence* fields and the 36 complete NMR-added responses under random identifiers without the names of the systems, and the two coding rules; the first coding, the counts and the manuscript were withheld. The workbook asked each coder first to answer six questions on what the spectra of each compound display (Sections 3.1 and 3.2), and then to code (i) whether each of the 72 fields names the infrared feature and (ii) whether each of the 36 NMR-added responses gives the signal near 104 ppm as the anomeric carbon, and where; the order in which the coders worked was not recorded. Coder A returned the workbook twice and coder B once; the circumstances are given in the Supporting Information (Section S14). After his first return, whose pattern of answers on the infrared feature the AI assistant had reported to H.A.A., coder A was sent a restatement of that task; his second return completed one unanswered cell and changed two answers on the infrared feature and his answer on the anomeric carbon of 123a, and it changed no answer on the reading of the responses. Every return is reported in the Supporting Information (Section S14), and the answers are given in full in Supporting File 5.

On (ii), the coding of coder A agrees with the first coding, that is, with the reading of the complete responses prepared with the assistant and confirmed decision by decision by H.A.A., in 31 of the 36 responses on yes or no (Cohen's kappa,[19] $\kappa = 0.72$) and in 28 of 36 when the place of the statement is included ($\kappa = 0.64$); on the 35 committed responses the two codings agree on 30 ($\kappa = 0.72$). The codings answer yes together for 15 responses and no together for 16; the first coding alone answers yes for 4 and

coder A alone for 1, so the proportions of positive and of negative agreement are both 0.86. Fourteen of the responses contain no value of the signal and are no in both codings; on the 22 in which a value of the signal occurs the two codings agree on 17 (κ = 0.32), so the overall κ owes much to the responses in which no value of the signal occurs. Resampling the twelve system–item cells (10 000 bootstrap samples) give 95% intervals of 0.50–0.89 for the κ of 0.72 on all 36 responses and of − 0.13 to 0.73 for the κ of 0.32 (Section S14). The 5 responses on which the two codings differ on yes or no were not discussed with the coder, and no entry of the first coding was changed: the counts of this article are those of the first coding, and the count of coder A is given beside the main count (Section 3.2; Table S30). The return of coder B gives “yes” for 34 of the 36 responses, among them 13 of the 14 responses in whose text no value of the signal occurs. It does not discriminate between the responses and is not used as a coding. This was decided after the return had been seen. A plan written after the returns of coder A and before that of coder B had provided for the majority of the three codings, which gives 20 of 35 and is reported in the Supporting Information (Section S14).

On (i), none of the three returns applied the written rule. Twenty-six of the 72 fields contain no reference to the infrared spectrum and are “no” under the rule; the three returns answer “yes” for 21, 23 and 25 of them. The comments of coder A show that he judged whether the statements of the fields are right; coder B answered “yes” for all but one of the fields he answered. The sheet of this task did not convey the question to either coder, and the reason was not established (Section S14). The naming of the infrared feature therefore has no second coding: its counts (Section 3.3) rest on the first coding and on the rule published as a script. The check of the cited values (the 13 flagged fields of Section 3.2) and the consistency notes were not part of the second coding; ten of the 13 flags are the misidentification covered by (ii), and coder A coded nine of these ten responses as giving the signal as the anomeric carbon.

An earlier, script-based coding (codes J1–J4; Sections S4.2–S4.5) combined the infrared feature and the anomeric-carbon value in one category. Because that category does not distinguish the two, it is not used in this article.

The records have three limits. They concern one field, in which the prompt asks for a single decisive feature; an observation not named there may still appear in the *Reasoning* field, which was coded for the one question stated above only and is deposited. Other statements quoted in this article from outside the field (Sections 3.2 and 3.4) were found by reading or by text search and are identified as such. Naming an observation does not show that the inference drawn from it is valid or that it produced the answer, since a model’s stated reasoning need not be the cause of its output;[20,21] in the response format of every prompt, moreover, the answer comes first and the *Decisive evidence* and *Reasoning* fields follow it. The field check establishes whether a cited value is present and whether a value given for the anomeric signal is that signal; it does not establish that the argument built on the value is sound.

2.5 Statistical treatment

The evaluation covers two compound pairs with one aglycone, two systems and three replicates per condition. Replicates are repeated requests on the same stimulus and measure the variability of the responses, not chemical diversity. Counts are reported as counts, and no inferential statistic is computed

on them. With three replicates in each of the 24 system–item cells (twelve items, two systems), no difference between the two systems is tested or claimed. Agreement between two codings is given as the number of agreements, as the cross-tabulation of the two codings and as Cohen's kappa with an interval from resampling the system–item cells; because kappa depends on how often the category occurs, it is also given for the responses in which a value of the signal concerned occurs (Section 2.4). The statistical treatment of the reaction records of the source is described with them (Sections S6.2, S6.3 and S17.1).

2.6 Use of generative AI tools

The two systems evaluated are the objects of study. Separately, H.A.A. used Claude (Anthropic) as a research assistant; the model version was not recorded for every working session and is therefore not stated. The assistant was used for literature searching and reference checking, for drafting and editing text, and for the Python code of the statistical re-analysis of the reaction records (Section S17.1), the similarity calculations of Table 1, the API runner, the coding and table scripts, the structure scheme (Fig. 1), the data plots (Fig. 4 and Figures S5, S6 and S8), the graphic for the table of contents, and the scripts that assemble the composite spectral figures of the article and of the Supporting Information from the deposited stimulus images, the screenshots and the two-dimensional NMR figures of the source without redrawing them. It was also used for internal critical reviews of earlier drafts. These reviews identified the inconsistencies reported in Section 3.1, and the assistant prepared the first-pass classification of Table S9, the first reading of the *Reasoning* fields (Table S24), the inspection of the deposited stimulus images (Section S13), the reading of the text and figures of the source and the measurements on its HSQC figures (Section S15), the literature comparisons of Section 3.1, of Sections S17.2 and S17.3 and of Tables S15–S19, the quotations of Table S23, the reprocessing of the original spectrometer data and the reading of the original printouts and files of the 2-deoxy pair (Section S16). It also prepared the workbook sent to the second coders, read the first return of coder A and reported its completeness and the pattern of its answers to the task on the infrared feature to H.A.A. before she had opened it, drafted the plan of 9 October 2026, the comparison of Table S37 and the verification table S38, compared the returns with the first coding and computed the agreement figures of Section 2.4 (Section S14). H.A.A. checked the following: the classification of the responses against the verbatim fields, and the values of the source used in the comparisons against the labelled values of its figures; the 36 decisions of the reading of the complete responses against the responses; eight statements of the description of the stimulus images against the images; four statements on the HSQC figure printed for 122a against that figure in the source, and three places of the deposited file of the source; the two lines of the doublet near 104 ppm, of nearly equal height, and the absence of a splitting of about 27 Hz in the line of C-1", in data set A of the original ^{13}C data of 117a (Section S16.1); the row of cross-peaks below the row annotated C-1" in the gHMBC figure of 117a, against that figure in the source (Section 3.2); the values of Table S23 that the article compares for the sugar-free aglycone (its ^{13}C list, carbonyl band and N–H signal), for its N-glucoside (the two lines near 104 ppm and the carbonyl band), for the 2-ribofuranosyl derivative (C-1' at 84.4 ppm and the pyrazolone carbonyl band) and for the two ribosyltriazoles (the ^{13}C values used), against the papers; the wavenumbers read from the three original infrared printouts of the 2-deoxy pair (Section S16.3); and the sample names, dates, labels and integrals read from the two printouts of 29 February 2016 against the printouts (Section S16.3). She did not check any of the following: the other quotations of Table S23 against the cited papers, among them the other

sugar carbons and the H-1' coupling of the 2-ribofuranosyl derivative and the other ^{13}C values of the N-glucoside; the other entries of the description of the images (Section S13); the other statements about the source and the other positions read from its HSQC figures (Section S15); the other results of the reprocessing of the original spectrometer data (Sections S16.1 and S16.2); the comparison with published evaluations (Table S37); the statements of the responses on two-dimensional spectra (Section 3.4) and the other statements found by text search outside the justification field (Sections 3.2–3.4); the comparison of the returns of the second coders with the first coding and the agreement figures computed from it (Section S14); the other readings of the original records of the 2-deoxy pair (Section S16.3). These parts rest on the work of the assistant. No figure contains AI-generated imagery. H.A.A. designed the study and ran every session; the compounds, spectra and assignments examined are those of her thesis. She audited the spectroscopic support for the source assignments with the assistance described above and checked the reported numbers against the source data and the raw responses, with the exceptions named above. The two second coders declared that they did the second coding without an AI tool. The author is accountable for the final text.

3 Results and Discussion

3.1 Status of the source assignments

The source assignments were re-examined after the responses had been collected. The full account, with the literature comparisons, is in Section S17, and Table 2 summarises its outcome; only what bears on the evaluation is stated here. Table S38 lists, for the main claims of this article, the record on which each rests and whether H.A.A. checked the reading.

Infrared feature. All four glycosides show the ester carbonyl absorption between 1741 and 1751 cm^{-1} (Fig. 2). 117a and 123a show a further labelled band, at 1680.47 and 1679.91 cm^{-1}; 116a and 122a show none between 1670 and 1690 cm^{-1} but one at 1639.05 and 1646.32 cm^{-1}. The origin of that band is not established. A conjugated, hydrogen-bonded lactam carbonyl absorbing below 1650 cm^{-1} cannot be excluded from the infrared data alone, although the carbonyl band of the related hydrazone-type pyrazolones compared in Section S17.2 is reported at 1658–1682 cm^{-1}; in a KBr disc a band near 1640 cm^{-1} can also contain the bending band of adsorbed water; and for the sugar-free aglycone the source lists no carbonyl band (its list gives 3459 cm^{-1}, NH/OH, and 1633 cm^{-1}, C = C), whereas a carbonyl band at 1670 cm^{-1} is reported for it elsewhere,[12] so the two reports give no fixed reference position (Section S17.2). The source states its own criterion as the absence of absorption at 1640–1690 cm^{-1} in the O-isomers; the band of 122a lies inside that range and that of 116a 1 cm^{-1} below it (audit item A10, Section S6.1). The criterion used here, a labelled band at 1670–1690 cm^{-1}, was formulated after the responses had been collected, from these four spectra (Section 2.2); it is an operational description of them and not a validated diagnostic range. The trace of 122a has an unlabelled shoulder that cannot be placed outside this window, and both second coders reported one in the trace of 116a (Table S27; Section S13), so for these two compounds the operational feature is the absence of a labelled band, not the absence of absorption.

Ribose pair. For 117a the recorded data are consistent with the source assignment, the N2-glycoside in its hydrazone–lactam form (Fig. 1a): the band at 1680 cm^{-1}, a ^{13}C signal at 157.8 ppm, a low-field proton at 13.48 ppm and the anomeric carbon at 84.39 ppm have counterparts in reports on related compounds,[11,12,22] two of which share an author and one of which comes from the research group in which the source was prepared, so the comparisons do not establish the connectivity independently (Section S17.2; Table S23). For 116a, the O-riboside of the source, the data favour an arylazo form and disfavour attachment at N2 without excluding it, but they do not distinguish attachment at O from attachment at N1, an alternative that the source does not consider (Fig. 1c). Its sugar carbons, with the anomeric carbon at 90.22 ppm, lie within 0.5 ppm of values read from a published list for an N-ribosyl-1,2,3-triazole[23] (compared by H.A.A. with the paper), and no anomeric-carbon value for an O-ribofuranoside of a hydroxypyrazole was found. The comparison with the sugar donor that bears on 122a (below) was not made for 116a: the source gives no ^{13}C data for its ribose donor, a commercial tetra-O-acetylribofuranose, and no literature value was compared. If C-1 of such a donor lies well above 90 ppm, the same reasoning would disfavour an O-glycoside for 116a as well (Section S17.3). The label of 116a is therefore open, an abstention on it is defensible, and if the compound is the N1-riboside the candidates offered to the models are incomplete. The gHMBC spectra (Fig. 3) show no annotated or described correlation from H-1" to a pyrazole carbon; no inference is drawn from the absence.

2-Deoxyribose pair. The aglycone carbons of the two printed ^{13}C spectra coincide within 0.35 ppm at eight positions, the aromatic proton shifts are the same, and the low-field proton is at 13.56 and 13.54 ppm, so both compounds show the aglycone signals of 117a, including 122a, which the source assigns as the O-glycoside from its infrared spectrum (for the second compound the ^{13}C part rests on the attribution below). Their sugar signals differ.

Neither ^{13}C spectrum of this pair is printed with the compound whose ^{1}H spectrum is printed under the same number: original records show directly that the spectrum printed for 123a is that of the compound whose ^{1}H spectrum is printed for 122a and that the spectrum printed for 122a is not, and that the latter belongs to the other compound follows by elimination (Sections S15, S16.3 and S17.3). The sugar cross-peaks of the HSQC figure printed for 122a, of which only an empty corner reached the models (Section 2.3), lie at about 76.3, 66.5, 66.7, 63.3 and 31.1 ppm (read to about 0.5 ppm) and agree within 0.25 ppm with lines labelled in the spectrum printed for 123a, not with the sugar lines of the spectrum printed for 122a (78.54, 68.56, 67.09, 66.13 and 29.69 ppm). The positions were read with the AI assistant from the figure in the author's file of the thesis; H.A.A. then compared the positions of the sugar cross-peaks and of the cross-peak at 103.9 ppm with the figure of the source and found them as given. Original records found afterwards in the author's files explain this: a ^{1}H and a ^{13}C printout recorded on 29 February 2016 under one sample designation have the signals of the ^{1}H spectrum printed for 122a and the lines labelled in the spectrum printed for 123a (20 of its 21 labels within 0.008 ppm), and a ^{13}C measurement of August 2014 of a sample with the same ^{1}H spectrum has the same sugar lines (the line at 76.50 ppm is not resolved from the solvent signal in it) and none of those labelled at 78.54, 68.56, 67.09 and 66.13 ppm (read from the files with the assistant and not checked by H.A.A.). H.A.A. compared the values read from the two printouts with the printouts and found them as given; on her statement both are measurements that she made herself on

one sample. The ^{13}C spectrum printed for 123a is therefore that of the compound whose ^{1}H, COSY and HSQC spectra are printed for 122a, whose anomeric carbon is at 76.50 ppm and not at 78.54 ppm, the C-1" listed in the source; the spectrum printed for 122a is attributed, by elimination, to the other compound, for which no original ^{13}C record was found. The records do not show at which stage the ^{13}C figures came to be printed with the wrong compound. H.A.A. states that she recorded all the spectra of 122a herself, on one sample tube, and that the handwritten corrections of the sample letter on two of the three original infrared printouts are right; the infrared spectra are used as printed.

For the compound whose ^{1}H spectrum is printed for 122a, the corrected record favours an N-glycoside without establishing it, and without excluding an O-glycoside that isomerises in solution (Table S17). Its anomeric carbon lies 15 ppm below C-1 of the 2-deoxy donor listed in the source (91.5 ppm), a carbon that bears two oxygen atoms, as C-1" of an O-glycoside would, whereas C-1" of an N-glycoside bears one oxygen and one nitrogen and is expected at higher field; no reference value for an O-glycoside of a 2-deoxypentose with a heteroaryl aglycone, and none for an N-glycoside of a 2-deoxypentopyranose, was compared. The N-ribofuranosides compared (84.4–90.5 ppm; Tables S17 and S23) do not argue against an N-glycoside, because a 2-deoxy sugar and a pyranosyl ring both move C-1 to higher field. Every original infrared printout of the two samples labels a band at 1677–1680 cm^{-1} (compared by H.A.A. with the printouts; Section S16.3); only the infrared figure printed for 122a points to an O-glycoside. If both compounds are N2-glycosides, they differ in the sugar. The sugar signals, the record of the donor and the ^{1}H couplings of this compound in two original data sets (Section S16.2; processed with the assistant and not checked by H.A.A.) are more easily reconciled with a pyranosyl ring in a chair form than with the furanosyl ring that the source names and both candidate drawings show, without excluding the latter (Section S17.3). The assignment, the ring size and the anomeric configuration of this pair remain open.

Consequences for the evaluation. Agreement with the source labels is therefore a secondary result (Section 2.2); for the 2-deoxy pair it is tabulated and not interpreted as chemical accuracy. What a response cites, and whether a value given for the anomeric carbon or proton is that signal, do not depend on the assignments and are reported for both pairs. For the 2-deoxy pair two premises of the items may fail, that one compound is candidate A and the other candidate B (if both are N2-glycosides) and that one drawing depicts the compound (if the ring is pyranosyl), and a third does fail: in the NMR-added items the ^{13}C panel presented for a compound is not the spectrum of the compound whose ^{1}H panel it accompanies. These items are defective stimuli, and their results are shown separately from those of the ribose pair throughout (Tables 3–6). None of the 18 responses to them remarks on the mismatch; the empty HSQC panel and the undisplayed sugar region of one ^{13}C panel gave little means of detecting it.

3.2 Identity of the signals cited

The field check of Section 2.4 was applied to the *Decisive evidence* field of all 72 glycoside responses (Table S9). Seven fields cite no value. In 52 of the other 65, every value cited has a counterpart in the value table and the anomeric signal, where one is given with a value, is the signal so assigned; 13 are flagged. Eleven of the 13 give for the anomeric carbon or proton a value that belongs to another signal, and two cite a value with no counterpart. Six are on the ribose pair and seven on the 2-deoxy pair, in whose NMR-added

items the ^{1}H and ^{13}C panels came from different compounds (Section 3.1). Two other kinds of discrepancy are kept apart from the flags. No field cites a value that is in the value table but not displayed in the image supplied (Section 2.4). Four responses report a correlation that the HSQC panel supplied does not show (Section 3.4); this concerns what a panel shows rather than whether a value is in the value table, and three of these responses have a flagged field, one of them among the two with no counterpart. Ten of the flagged fields belong to responses that disagree with the source label and three to responses that agree with it, all three on the 2-deoxy pair. A consistency note was made on five of the fields that pass: four contain a statement that contradicts another statement of the field or the answer given, and one places the absent carbonyl band at 1690–1715 cm^{-1} (Sections 3.3 and 3.4). All 13 flagged fields belong to NMR-added items; in the 36 infrared-only fields every value cited has a counterpart, and only a missing counterpart could have been flagged there (Section 2.4). Ten of the eleven misattributions are one and the same error: an aryl carbon given as the anomeric carbon. Twelve of the 13 flags are on gpt-5 fields; the nine identifications of the aryl carbon made by Gemini are all in the *Reasoning* field, which the check does not cover, and Gemini fields cite fewer values (four of its 18 NMR-added fields cite none), so the flags do not compare the two systems (Section 2.5). If the reading of the *Reasoning* field is applied to the values cited in the field, the three single-case Gemini fields on 122a, which cite 103.80 ppm alone, are flagged as well, giving 16, six of them in responses that agree with the source label. One field that passes, which cites 78.5 ppm as the anomeric carbon of 122a, passes on the list of the source only (Section 3.1).

All 17 committed NMR-added responses on the ribose pair (eleven single-case, six paired) named the ^{13}C shift of the anomeric carbon in the *Decisive evidence* field, and 14 gave a value there. Five gpt-5 fields state a numerical range for at least one connectivity (100–105 ppm for an O-glycoside; 85–92, 86–92 or 85–90 ppm for an N-glycoside), three describe the values as typical of one or the other, and for the nine Gemini responses the ranges appear only in the *Reasoning* field (Table S5, entries marked †). The signal was not always the anomeric carbon (Table 3).

Table 3

The signal cited as the anomeric carbon in the NMR-added items (three responses per system per item). C-1" is the carbon so assigned in the source; C-2' is the aryl CH between C-1' and C-3', as assigned in the source and here (Tables S15 and S16). In the fifth column a response is counted if it gives C-2' as the anomeric carbon in its *Decisive evidence* field or in its *Reasoning* field, all 36 responses having been read for this question (Section 2.4; Table S24); the last column gives the count in the independent coding of coder A (Section 2.4). The two rows before the last give the pairs separately.

Item	**C-1"; aryl C-2' (ppm)**	**gpt-5**	**Gemini**	**Aryl C-2' given as anomeric / committed**	**Coder A**
Single case, 117a	84.39; 104.09	104.09 (three)	84.39 (three)	3 / 6	4 / 6
Single case, 116a	90.22; 107.67	90.2 (two); one abstention	90.2 (three)	0 / 5	0 / 5
Paired, 116a and 117a	as above	90.2 for 116a and 104.1 for 117a (three)	the same (three)[a]	6 / 6	4 / 6
Single case, 122a	78.54[b]; 103.80	78.5 (one); "88–89", no labelled value (one); no anomeric carbon discussed (one)	103.80 (three), value in the field, read as the anomeric carbon in the *Reasoning* field[a]	3 / 6	2 / 6
Single case, 123a	77.2 in the list of the source, without a correlation spectrum; sugar region not displayed; 103.83	103.8–103.9 (three)	103.8 (one)[a]; none, every signal above 100 ppm assigned to the aglycone (one); none, the spectrum described as incomplete (one)	4 / 6	3 / 6
Paired, 122a and 123a	as above	103.8 for 122a (one); no anomeric carbon discussed (two)	103.8 for 122a (two)[a], one of them with 76.5 for 123a[a]; no anomeric carbon discussed (one)	3 / 6	3 / 6
Ribose pair, three items		6 / 8, all in the field	3 / 9, all in the *Reasoning* field[a]	9 / 17 (9 / 12 with the spectra of 117a)	8 / 17 (8 / 12)
2-Deoxy pair, three items[c]		4 / 9, all in the field	6 / 9, all in the *Reasoning* field[a]	10 / 18	8 / 18
All items		10 / 17, all in the field	9 / 18, all in the *Reasoning* field[a]	19 / 35	16 / 35

[a] Reading taken from the *Reasoning* field. [b] The value listed in the source and labelled in the panel supplied for 122a; by the original records that panel is not a spectrum of the compound whose ^{1}H spectrum is printed for 122a, and the anomeric carbon of that compound lies at 76.50 ppm, in the spectrum printed for 123a (Section 3.1). [c] In the NMR-added items of this pair the ^{13}C panel presented for a compound is not the spectrum of the compound whose ^{1}H panel it accompanies (Section 3.1).

On 117a, the three single-case gpt-5 responses gave the signal at 104.09 ppm as the anomeric carbon and assigned the compound to the O-glycoside; one of them adds that no signal is observed in the range it expects for an N-glycoside (86–92 ppm). None mentions in the field the signal at 84.39 ppm, which is labelled in the ^{13}C panel and is the carbon assigned C-1" in the source; outside the field, one response lists it among the sugar carbons and one names it as a signal that “could in principle be misassigned”. The six paired responses likewise gave 104.09 ppm as the anomeric carbon of 117a (the three Gemini responses in the *Reasoning* field). On the 2-deoxy pair the corresponding signal (103.80 ppm in 122a, 103.83 ppm in 123a) was given as the anomeric carbon in ten responses (Table 3). In all, 19 of the 35 committed NMR-added responses gave this signal as the anomeric carbon: ten in the *Decisive evidence* field, all from gpt-5 (10 of its 17 committed responses), and nine in the *Reasoning* field, all from Gemini (9 of 18); in three of the nine the value itself is in the field and its reading as the anomeric carbon in the *Reasoning* field. Of the other 16 committed responses, twelve do not give the aryl signal as the anomeric carbon and four discuss no anomeric carbon (Table S24; column aryl_C2_given_as_anomeric of Supporting File 2). In the independent coding of coder A the count is 16 of 35 (Section 2.4). The two codings differ on 5 responses (Table S30). In three of them the response states the matter in so many words: two give the signal as the anomeric carbon and were not counted by coder A, and one assigns it to the aryl ring and was counted by him. In the other two, both counted in the first coding only, the identification is made by implication: the sentence on which the first coding rests does not contain the word anomeric, and an earlier sentence of the same passage does. Without these two the first coding gives 17 of 35. The 19 come from eight of the twelve NMR-added system–item cells, in five of which all three replicates made the identification. Of the nine on 117a, six are from the paired item, whose answers coincide with the positional default and whose prompt has neither the ambiguity instruction nor the *Limitations* field; in the single-case item all three gpt-5 responses and none of the three Gemini responses made the identification. The identification was thus made in three of the six single-case and in all six paired responses on 117a; the design does not establish whether the paired presentation contributed to this difference (Section 2.3). Over the 30 committed responses that argue from the ^{13}C shift of the anomeric carbon (29 from a signal described as that carbon and one from the absence of such a signal above 100 ppm; Section S3.2), a denominator defined after the count (Section 2.2), the count is 19 of 30. The first coding and the coding of coder A give 19 and 16 of 35 (9 and 8 of 17 on the ribose pair, 10 and 8 of 18 on the 2-deoxy pair; Table 3); the majority of the three codings, which includes the return of coder B, gives 20 (Section S14).

What the images displayed differs between the compounds and qualifies this count (Section 2.3). For 117a the ^{13}C panel labels the signal at 84.39 ppm and the gHMBC panel marks the row of C-1"; two of the three single-case Gemini responses state that they located C-1" from that panel, and the three gpt-5 responses mention the panel only in their *Limitations* field, as insufficiently annotated or as lacking a correlation from

H-1" to the heterocycle. For 122a the ^{13}C panel labels the signal at 78.54 ppm, but nothing supplied identifies it as the anomeric carbon, because the HSQC panel contains no contours; by the original records the panel does not show the spectrum of the compound whose ^{1}H spectrum was supplied with it, and in the complete HSQC figure the cross-peak of the proton at 5.83 ppm does not lie at this value (Section 3.1). For 123a the sugar region of the ^{13}C spectrum was not displayed, so no signal identifiable as the anomeric carbon was shown, and the source lists 77.2 ppm as C-1" without a correlation spectrum (Section S17.3); the two inset lines at 77.21 and 76.50 ppm were visible, and one paired response gave 76.499 ppm as the anomeric carbon of 123a; this line coincides, within 0.3 ppm, with the position at which the HSQC figure has the cross-peak of the proton at 5.83 ppm, the H-1" of the ^{1}H spectrum printed for 122a. By the original records the panel supplied for 123a shows the ^{13}C spectrum of the compound with that ^{1}H spectrum, so this line is the anomeric carbon of that compound, although nothing supplied showed it and the panel stood under the other number. Four of the 19 responses concern this compound and six concern 122a, so for ten of the 19 the images displayed neither a correlation nor an identified anomeric carbon against which the identification could be checked directly. A count of the labelled carbons argued against the identification in two ways. Between 100 and 165 ppm the labelled lines, with the two lines near 163 ppm counted once as the C–F doublet of C-3', number nine, as many as the aglycone carbons other than the unlabelled CF_3 carbon, only if the signal near 104 ppm is among them. In the panel of 123a these two lines are 2.47 ppm (about 248 Hz) apart, as expected for a one-bond C–F coupling; in the panel of 122a they are 1.98 ppm (about 199 Hz) apart (audit item A7), a spacing that does not identify them as one carbon, and read as two carbons, as one gpt-5 response on 122a did in its *Reasoning* field (“an additional carbonyl-type signal at 164.3 ppm”), they leave one line more than the aglycone requires. The ^{13}C panel of 122a, however, also shows four oxygenated sugar carbons, the full set of a di-O-acetyl-2-deoxypentose, which leaves no place for a fifth. Two of the six single-case responses on 123a drew the conclusion that the image allows: one assigned every signal above 100 ppm to the aglycone, and one stated that the ^{13}C spectrum supplied is incomplete. Giving the signal near 104 ppm as the anomeric carbon is a misidentification on all three compounds, but only for 117a was the carbon assigned C-1" both displayed and marked. The 19 thus fall into two groups: 9 of the 12 committed responses that had the spectra of 117a (three of the six single-case and all six paired), made against panels that labelled and marked the anomeric carbon, and 10 of the 18 on the 2-deoxy pair, made on panels that gave no direct evidence and whose ^{13}C panel was, by the original records, not a spectrum of the compound whose ^{1}H panel it accompanied (Section 3.1); in the coding of coder A the figures are 8 of 12 and 8 of 18 (Table 3).

Four observations place the signal near 104 ppm in the aryl ring and not in the sugar; Section S18.1 gives them in full. First, signals at 103.8 and 104.1 ppm are printed for the parent aglycone, which carries no sugar, and at 103.7 and 104.2 ppm for its N-glucoside;[12] the signal recurs in all three hydrazone-type compounds here (Table S16). Second, the gHMBC panel of 117a supplied to the models has, directly below the row annotated C-1", a row at about 104 ppm whose two cross-peaks lie in the columns of the aromatic multiplets that the source assigns to H-4' and H-6', the protons three bonds from C-2' (Fig. 3b; Section S13). Third, the source itself assigns this signal to C-2', and in its HSQC figure of 122a the cross-peak at 103.9 ppm lies in the column of an aromatic proton (Section S15). Fourth, in the original ^{13}C printout of 29

February 2016 of the 2-deoxy compound whose ^{1}H spectrum is printed for 122a the signal is a doublet of 26.6 Hz (104.098 and 103.833 ppm), and the six aryl signals show the one- to four-bond C–F couplings of a 3-fluorophenyl ring (Section S16.3); the reprocessed original data of 117a and a measurement of 2014 give the same doublet, with two lines of nearly equal height (Sections S16.1 and S16.3). A doublet requires coupling to a single fluorine nucleus, that is, to the aryl fluorine on C-3' (coupling to the CF_3 group would give a quartet), and in each of the three attachment isomers of Fig. 1 every sugar carbon is at least nine bonds from that fluorine. Of the four observations, the third rests on the source itself, including a cross-peak position that H.A.A. checked, and the fourth on a printout whose labels she checked, on the original data of 117a, in which she found the doublet near 104 ppm, and no splitting of that size in the line of C-1", herself, and on the processing, by the assistant, of the files of 2014, which she did not check; the first rests on published lists, whose values she compared with the papers, and the second on a reading of the gHMBC image made with the assistant, which she compared with the figure of the source (Section 2.6). For the ^{13}C panel supplied for 122a, which by the original records is not a spectrum of the compound whose HSQC figure is printed for 122a, the third and fourth observations do not apply directly; there the assignment rests on the first observation and on the coincidence of 103.80 ppm with the aryl signal of the other spectra (Table S16; Section S18.1). The signal is accordingly assigned, in the source and here, to C-2', the aryl CH between C-1' and C-3'. Whatever its exact position within the aryl ring, it belongs to the aglycone and is not an anomeric carbon, so this finding does not depend on whether the compounds are N- or O-glycosides or on which sugar carbon is C-1"; besides the reading of the responses (Section 2.4), it depends on the assignment of the 104 ppm signal to the aryl ring. The answer that accompanied it differs between the compounds: on 117a and 123a it disagrees with the source label, and on 122a it agrees. The source labels one line of the doublet in each spectrum, and the splitting (0.27 ppm) is not resolved in the supplied image, so no claim is made that the multiplicity was available to the models.

The answers of the two second coders show what the panels allowed a chemist to read (Section S18.2; Table S27). Before coding, and without the ^{1}H panels, the source or this article, they were asked which labelled value is the anomeric carbon and, in a second question that named the signal near 104 ppm and offered the answers “aromatic (aglycone) carbon”, “anomeric (sugar) carbon” and “cannot tell”, what that signal is. Both gave the signal at 103.80 ppm as the anomeric carbon of 122a. For 123a, coder B and, in his second return, coder A marked the signal near 104 ppm as the anomeric carbon. On the ribose pair coder A named 104.09 and 107.67 ppm as the anomeric carbons of 117a and 116a, whereas coder B named 84.39 ppm for 117a. The same workbook contained the 36 responses, and the order in which the coders worked was not recorded. These answers are not a human baseline: they are the answers of two chemists to a few questions under these conditions and are not compared statistically with the responses. For 122a they show that the supplied carbon panel, without reference data, allowed the reading that the systems made; the four observations above rest on data that the panels of the 2-deoxy pair do not display. The ten responses on 122a and 123a made an identification for which the supplied panels gave no direct evidence, only counts of the labelled carbons: for 123a the count between 100 and 165 ppm, and for 122a, besides that count with its anomalously spaced pair near 163 ppm, its four oxygenated sugar carbons, which leave no place for a fifth. The nine on 117a made it against a ^{13}C panel that labels the carbon assigned C-1" and a gHMBC panel that marks its row, and coder A made it there as well.

On 116a, the five committed single-case responses and the six paired responses (three in the field, three in the *Reasoning* field) cited the signal at 90.2 ppm, which is the carbon assigned C-1", and read it as an N-glycosidic anomeric carbon. These responses disagree with the source label, and none of them cited in the field the infrared or aryl data in which 116a differs from 117a. Outside the field, a text search finds two single-case Gemini responses that read the band at 1639 cm^{-1}, in the *Reasoning* field, as the carbonyl group of the pyrazolone and two single-case gpt-5 responses that state in the *Limitations* field that the infrared spectrum does not distinguish the linkages; this search is not part of the coding. The value they cited is the one on which Section 3.1 raises attachment at N1. Their inference that the sugar is on nitrogen cannot be scored as chemically wrong; what the record shows is that in the field they justified it by one value judged against a general range, without reference to the other evidence supplied.

In the paired NMR-added item on the ribose pair, all six responses returned the inverted assignment at stated confidences of 80–95: 116a as the N-glycoside, citing 90.2 ppm, and 117a as the O-glycoside, citing 104.09 ppm (for the three Gemini responses both values are in the *Reasoning* field). The two carbons assigned C-1" in the source, at 84.39 and 90.22 ppm, were not compared with each other in any *Decisive evidence* field. The same answer is also what a positional default would give (Section 2.3), and the two cannot be distinguished here.

The correspondence between the ranges stated and the answers returned was established after the fact, and for the Gemini responses from the *Reasoning* field; testing whether the stated rule produced the answers would require items in which the anomeric label is removed, edited or exchanged while everything else is held fixed (Section S7.2).

3.3 Naming of the infrared feature, and agreement with the source labels

Table 4 gives the results for each pair by presentation and Table 5 by item and system; Fig. 4 shows the 36 responses on the ribose pair. On the ribose pair, with the infrared spectrum alone, the *Decisive evidence* field named the infrared feature in 10 of 12 single-case responses and in 5 of 6 paired responses; with the NMR spectra added to the same infrared spectrum it did so in 1 of 12 and in 0 of 6. On the 2-deoxy pair the corresponding figures are 7 of 12 and 6 of 6 with the infrared spectrum alone, and 2 of 12 and 6 of 6 with NMR spectra added. The difference between the levels of evidence is thus seen in the single-case presentation of both pairs and, in the paired presentation, on the ribose pair only (on the 2-deoxy pair as well, from 6 to 4 of 6, under the stricter reading that follows). Two of the six paired NMR-added fields on the 2-deoxy pair that cite the band near 1680 cm^{-1} for 123a also describe the band of 122a at 1646 cm^{-1} as a ring carbonyl; if the exclusion that Section 2.4 applies under its second and third criteria, to a field that asserts a second, non-ester carbonyl, were applied under the first as well, they would not count, and the figure would be 4 of 6 (6 of 18 for the 2-deoxy pair with NMR spectra added). This is the stricter reading referred to below. In the paired presentation 19 of the 24 responses assigned candidate A to the first compound and B to the second, including all twelve on the 2-deoxy pair, where this is also the source assignment, and all six NMR-added responses on the ribose pair; those results cannot be separated from a positional default (Section 2.3). For 122a, moreover, the operational feature is a property of the printed

figure supplied: the original infrared printout of this sample whose printed name is not altered, and, on the statement of H.A.A. about the handwritten corrections, a second one, label a band near 1680 cm^{-1} (read from the scans with the assistant and compared by H.A.A. with the printouts; Section 3.1). Naming does not depend on the assignments, which is why it is reported for both pairs; the low figure for 122a with the infrared spectrum alone (1 of 6) coincides with four of the six responses citing the band at 1646 cm^{-1} and choosing the candidate with a ring carbonyl (Section 3.4). As a secondary result, agreement with the source labels on the ribose pair was 9 of 12 and 3 of 12 in the single-case presentation and 5 of 6 and 0 of 6 in the paired presentation; for the 2-deoxy pair agreement is tabulated and not interpreted as chemical accuracy (Section 3.1). The differences in infrared naming describe these items and are not attributed to the NMR spectra (Section 3.5). The naming was coded once: the second coding of this judgement did not succeed (Section 2.4), so the counts of this section rest on the first coding and on the rule published as a script (Supporting File 3). A text search of the complete responses, which is not part of the coding, finds the band near 1680 cm^{-1} cited outside the field in one further infrared-only response on the ribose pair (the second abstention, in its *Reasoning* field) and in two further NMR-added ones (two of the agreeing Gemini responses on 117a, in the *Reasoning* field), so that on the ribose pair 16 and 3 of 18 responses cite it somewhere; on the 2-deoxy pair it finds only ranges that include 1680 cm^{-1}. The difference between the levels of evidence is therefore not confined to the field. By system–item cells (two of three replicates or more), the feature was named in 6 of the 6 infrared-only and none of the 6 NMR-added cells on the ribose pair, and in 4 and 3 of 6 on the 2-deoxy pair.

Table 4

Naming of the infrared feature and agreement with the source label for each pair (both systems, three replicates per system per item). A paired response assigns both compounds jointly and agrees only if both assignments agree with the source labels. Total rows are descriptive totals over the two presentations, which are different units. For the 2-deoxy pair, agreement is agreement with the source labels only, and in its NMR-added items the ^{1}H and ^{13}C panels came from different compounds (Section 3.1). One of the ten single-case infrared-only responses on the ribose pair counted as naming the feature is an abstention that cites the band as expected for both candidates.

Pair	Presentation	Evidence	*n*	Operational infrared feature named	Agrees with source label	Abstains	Disagrees	Named and agrees
Ribose	Single case	Infrared only	12	10	9	2	1	8
Ribose	Single case	NMR added	12	1	3	1	8	1
Ribose	Paired	Infrared only	6	5	5	0	1	5
Ribose	Paired	NMR added	6	0	0	0	6	0
Ribose	Total	Infrared only	18	15	14	2	2	13
Ribose	Total	NMR added	18	1	3	1	14	1
2-Deoxy	Single case	Infrared only	12	7	6	1	5	6
2-Deoxy	Single case	NMR added[b]	12	2	5	0	7	2
2-Deoxy	Paired	Infrared only	6	6	6	0	0	6
2-Deoxy	Paired	NMR added[b]	6	6[a]	6	0	0	6[a]
2-Deoxy	Total	Infrared only	18	13	12	1	5	12
2-Deoxy	Total	NMR added[b]	18	8[a]	11	0	7	8[a]

[a] 4 (paired) and 6 (total) under the stricter reading of Section 3.3, which does not count two paired fields that also describe the band of 122a at 1646 cm^{-1} as a ring carbonyl. [b] Items whose ^{1}H and ^{13}C panels came from different compounds (Section 3.1).

Table 5

Responses by item and system (three replicates per cell): number naming the infrared feature, and number agreeing with the source label, with the infrared spectrum alone / with NMR spectra added. On the 2-deoxy pair the figure after the solidus refers to items whose ^{1}H and ^{13}C panels came from different compounds (Section 3.1).

Item	System	Operational infrared feature named	Agrees with source label
Single case, 117a	gpt-5	2 / 0	0 / 0
Single case, 117a	Gemini	3 / 1	3 / 3
Single case, 116a	gpt-5	2 / 0	3 / 0
Single case, 116a	Gemini	3 / 0	3 / 0
Paired, 116a and 117a	gpt-5	2 / 0	2 / 0
Paired, 116a and 117a	Gemini	3 / 0	3 / 0
Single case, 122a	gpt-5	0 / 0	0 / 0
Single case, 122a	Gemini	1 / 0	1 / 3
Single case, 123a	gpt-5	3 / 0	2 / 0
Single case, 123a	Gemini	3 / 2	3 / 2
Paired, 122a and 123a	gpt-5	3 / 3[a]	3 / 3
Paired, 122a and 123a	Gemini	3 / 3	3 / 3

[a] 3 / 1 under the stricter reading of Section 3.3.

Three things in the item-level record of the ribose pair qualify these totals. First, naming and agreeing are not the same set. Of the 15 infrared-only responses that named the feature, 13 agreed with the source label; one gpt-5 response on 117a named the band at 1680 cm^{-1} and chose the candidate without a ring carbonyl, and one abstained in a field that begins “None” and cites the band as present but expected for both candidates. Without the abstention, the infrared-only figure is 14 of 18 (9 of 12 single-case, 5 of 6 paired). Of the 14 infrared-only responses that agreed, one did not name the feature: a gpt-5 response on 116a described the band at 1639 cm^{-1} as an amide carbonyl, which would argue for the lactam, and chose candidate B. Another agreeing gpt-5 response on 116a stated that the spectrum has a single carbonyl band but placed the absent lactam band at 1690–1715 cm^{-1}. Second, the difference in agreement in the single-case presentation lies entirely on 116a (6 of 6 to 0 of 6); on 117a agreement was 3 of 6 at both levels of evidence, gpt-5 never agreeing and Gemini always. Third, the three NMR-added responses that agreed, all from Gemini on 117a, cited the anomeric carbon at 84.39 ppm, and all three also cite the band at 1680 cm^{-1}, one in the field and two in the *Reasoning* field; two state that the gHMBC panel assigns this carbon as C-1".

3.4 Further observations

The band at 1646 cm $^{-1}$. In the single-case infrared-only items on 122a, four of six responses cited the band at 1646.32 cm^{-1} and chose the candidate with a ring carbonyl (two gpt-5 fields call it an amide carbonyl; two Gemini fields cite the band alone), and a fifth abstained; none of the six paired infrared-only responses, in which the partner spectrum was also supplied, did so; all six gave the positional answer, candidate A for the first compound and B for the second, which is also the source assignment. In the paired NMR-added item two gpt-5 responses again described the 1646 cm^{-1} band as a ring carbonyl. Given Section 3.1, that reading is consistent with the NMR fingerprint of 122a and cannot be scored as an error. Both responses nevertheless assign 122a to the candidate without a ring carbonyl; one carries a consistency note for this, and the other is already flagged for the anomeric signal.

A named band assigned to the other candidate. Two infrared-only gpt-5 responses, one on 117a and one on 123a, named the band near 1680 cm^{-1} as the only non-ester carbonyl absorption and then chose the candidate that has no ring carbonyl; they are two of the four contradictions noted in Section 3.2.

Abstention. Four of the 90 responses abstained, all from gpt-5 and all under the single-case prompts, which alone ask for ambiguity to be explained; four abstentions do not allow this to be attributed to the prompt. Three were on the ribose pair. Abstentions are not counted as disagreement, and for an infrared-only item they are defensible, since the infrared data alone do not exclude a low-frequency carbonyl (Section 3.1).

Values not among the labelled values. In two of the 72 glycoside *Decisive evidence* fields, both from gpt-5 on the 2-deoxy pair, a cited value has no labelled value within tolerance: a ^{13}C resonance “at ~ 88–89 ppm” said to correlate with the anomeric proton in the HSQC panel, where the nearest labelled value is 78.54 ppm, and a carbonyl signal “at ~ 168.5 ppm”, where the nearest labelled value is the acetate carbonyl at 169.89 ppm. The second corresponds to a truncated label: in the supplied ^{13}C panel of 123a the label of this signal is cut off after “16” and part of a third digit (Section 2.3), and the value cited is compatible with an estimate from the axis. The first is different in kind, because the correlation it reports cannot have been read from the image (next paragraph). The check cannot in general distinguish a value read from an unlabelled feature of a trace from one with no basis in the image. The eleventh misattribution of Section 3.2 is a Gemini response on the 2-deoxy pair that gives the anomeric proton of 123a as 6.96 ppm, which is an aromatic proton (6.98 ppm).

Correlations reported that the panels do not show. The HSQC panel of 122a, as supplied, shows the ^{1}H projection and an empty contour region (Section 2.3). Four of the twelve responses that received it nevertheless report what the HSQC spectrum shows. One single-case gpt-5 field gives a correlation of the anomeric proton with a carbon at 88–89 ppm; one paired gpt-5 field states that the HSQC spectrum shows the carbon at 103.8 ppm to be a CH; and two paired Gemini responses state in the *Reasoning* field that the HSQC spectrum confirms an anomeric carbon at 103.8 ppm. None of these statements can have been read from the image, and the two Gemini statements are at variance with the spectrum itself: in the complete figure of the source the cross-peak of this carbon lies in the column of an aromatic proton, and the cross-peak of the anomeric proton lies near 76.3 ppm. The gpt-5 statement that the carbon is a CH agrees with the complete figure, in which, however, its cross-peak lies in the column of an aromatic proton. The single-case prompt forbids inventing correlations, and the paired prompts name peaks and experimental details

only (Section 2.3). The other eight responses do not mention the HSQC spectrum, and none of the twelve remarks that the panel is empty. The empty panel is a defect of the stimulus, found only when the deposited images were inspected for this article, and the four statements are reported as what the responses did with it. A statement of the same kind concerns the gHMBC panel of 116a: the one abstaining response reports “correlations from the anomeric region to several heteroaryl carbons”, whereas the panel shows cross-peaks on the rows of C-5", C-2" and C-1" only.

Stated confidence. Gemini reported 85 or more in every response and 95 in 37 of 45. For gpt-5 the committed responses ranged from 60 to 90; on the ribose pair, single-case and paired responses together, its five agreeing responses averaged 74 and its ten disagreeing responses 78.5. The ranges for agreeing and disagreeing responses overlap for both systems. The prompt did not define what the confidence refers to, and no calibration analysis is attempted.

Benzoyl pair. The supplementary task on the O- and N-benzoyl derivatives asked only whether the number of carbonyl bands visible in an infrared image (one against two) is mapped to the candidate drawing with the corresponding number of carbonyl groups. Seventeen of 18 responses agreed with the source labels. Because the source characterisation of this pair is internally inconsistent (audit item A5), the task says nothing about the chemistry of the compounds and is reported in the Supporting Information (Section S3.5).

3.5 Implications and limits

The results describe what two systems cited on twelve items of a forced choice between two drawn candidates, under default settings and with three replicates per item; like every count here, they are not error rates. Table 6 lists each result with what it depends on and with its share on the 2-deoxy NMR-added items, whose ^{1}H and ^{13}C panels came from different compounds.

Table 6

What each result depends on. Open source assignments: the site of attachment of 116a, and the assignment, ring size and anomeric configuration of the 2-deoxy pair (Section 3.1).

Result	Depends on the open source assignments	Of which on the 2-deoxy NMR-added items[a]	Rests on
2 fields cite a value with no counterpart in the value table	no	2	the value tables of the source figures
11 fields give another signal's value for the anomeric carbon or proton (10 of them the aryl carbon near 104 ppm)	no; for the 2-deoxy pair the C-1" of the list of the source is the reference, and one field that cites it passes on that list only	5 (4 of them the aryl carbon)	the value tables; the assignment of the signal near 104 ppm to the aryl ring (Section 3.2)
19 of 35 committed NMR-added responses give the aryl carbon as the anomeric carbon (9 of 12 with the spectra of 117a; 10 of 18 on the 2-deoxy pair)	no	10	the assignment of the signal near 104 ppm to the aryl ring; the reading of the *Reasoning* field in the first coding (16 of 35 in the coding of coder A)
4 responses report HSQC correlations that the panel supplied cannot show	no	4	the inspection of the stimulus image (Section 2.3)
Naming of the operational infrared feature (ribose pair 15 and 1 of 18; 2-deoxy pair 13 and 8 of 18, 6 under the stricter reading of Section 3.3)	no	the 8 (6) NMR-added of the 2-deoxy pair	the operational criterion of Table 2; coded once
Agreement with the source labels, ribose pair (14 and 3 of 18)	yes: 117a is consistent with its label, the label of 116a is open	none	the source labels; in the paired items a positional default cannot be excluded
Agreement with the source labels, 2-deoxy pair (12 and 11 of 18)	yes: the assignment of the pair is open	the 11 NMR-added	not interpreted as chemical accuracy; the NMR-added items mixed the spectra of two compounds

[a] Items whose ^{13}C panel is not the spectrum of the compound whose ^{1}H panel it accompanies (Section 3.1).

The flags of the field check are the most direct result of the audit: each compares a value cited in the justification field with the value table and, for the anomeric carbon or proton, with the signal for which the value is given. The count of 19 extends one question, the identity of the signal given as the anomeric carbon, to the *Reasoning* field. Five observations follow, each confined to this case.

The identity of a cited signal. The aryl carbon near 104 ppm was given as the anomeric carbon by both systems, in both presentations and on three of the four compounds. The case that bears most directly on what the responses did with the evidence displayed is 117a: its ^{13}C panel labelled the carbon assigned C-1", its gHMBC panel marked the row of that carbon, and 9 of the 12 committed responses that had these spectra still gave the aryl carbon (Section 3.2). The 10 of 18 on the 2-deoxy pair were made on panels that gave no direct evidence and, by the original records, combined the spectra of two compounds; they show what the responses did with a defective stimulus. The result does not depend on the open assignments, because the signal is also printed for the sugar-free aglycone. A tabulated shift range is only as good as the signal it is applied to: the ranges that the responses stated for N- and O-glycosidic anomeric carbons are not shown here to be wrong for ribofuranosides, but they were applied to a signal that is not the anomeric carbon, and no field compared the two carbons assigned C-1" (84.39 and 90.22 ppm); for the 2-deoxy pair, whose ring is more easily reconciled with a pyranose (Section 3.1), they may not apply, since a 2-deoxy sugar and a six-membered ring both move C-1 to higher field.

Agreement and evidence. On the 2-deoxy pair, six responses that agree with the source label give the aryl carbon as the anomeric carbon, in the field or in the *Reasoning* field, and two of them are among the flagged fields; on the ribose pair the three agreeing NMR-added responses cite the carbon assigned C-1" (Section 3.3). A check confined to the justification field would have missed 9 of the 19 identifications, which are made in the *Reasoning* field only.

The observation named. The infrared-only and the NMR-added items differ in the observation named in the field: on the ribose pair the operational infrared feature was named in most infrared-only responses and in almost none of the NMR-added ones, and on the 2-deoxy pair it was named less often with NMR spectra added in the single-case presentation (2 against 7 of 12) and, in the paired one, only under the stricter reading (4 against 6 of 6; Section 3.3). The design does not allow this difference to be attributed to the NMR spectra.

The stimulus. Inspection of the deposited images, made after the coding, showed an HSQC panel without contours, a ^{13}C panel without its sugar region and with truncated labels, and a candidate drawing with overlapping groups (Section 2.3); four responses reported HSQC correlations that the panel cannot show, one flagged value corresponds to a truncated label, and two responses noted that the ^{13}C panel of 123a is incomplete (a single-case Gemini response, Section 3.2, and a paired gpt-5 response that describes the window as “cropped”). Only the original records showed that the ^{1}H and ^{13}C panels of every NMR-added item of the 2-deoxy pair came from different compounds (Section 3.1). An evaluation built on image crops should verify, before any request is sent, that each crop displays what the item presupposes and that the record from which the crops are taken is itself consistent.

Transfer. The audit can in principle be applied to other evaluations under four conditions: a value table for every spectrum supplied, against which each cited value can be looked up; a structured field in which the answer is justified, so that the values to be checked can be located; the identity of the signals that an answer may cite, established independently of the responses (here available for the anomeric carbon and proton, and for the signal near 104 ppm from data that the panels did not contain; Sections 2.4 and 3.2);

and labelled peaks, so that the presence of a cited value can be decided within a stated tolerance. It has not been tested on other compounds or compound classes, with a third model, on free-form output without a designated justification field, or on unlabelled spectra, for which the values would have to be read from the trace and the tolerance would have to cover that reading.

The patterns recorded here (a tabulated shift range judged against single values, a signal attributed to the wrong carbon, and a band near 1640 cm^{-1} read as a second carbonyl) resemble difficulties described in studies of how students interpret spectra;[24,25] the third cannot be scored as an error on the present data (Sections 3.1 and 3.4). A classroom exercise built on the present materials is outlined in Section S9; it has not been trialled.

Limits. The analysis rests on two pairs of compounds with one aglycone, twelve items, two systems and three replicates of each request, so the 72 responses are repeated tests of four related compounds and not 72 cases. Agreement can be interpreted for the ribose pair only, and there it is conditional, because the label of 116a is open (Section 3.1). The task offered neither the N1-glycoside nor a pyranosyl form and drew the O-glycoside as a hydrazone tautomer with an N–H; the label and order mapping was fixed; the level of evidence coincided with the batch of collection; the NMR-added items carried more images and text as well as more chemical information; the prompt asked for one decisive feature; and the stimuli were images that retained the source's sugar annotations and were uneven in what they displayed (Section 2.3). One field of each response was coded, the rest being read for the anomeric-carbon question only; only that judgement has a usable second coding, and the infrared naming and the field check were coded once (Section 2.4). The first-pass classification, the reading of the complete responses and the inspection of the images were prepared with an AI assistant; which parts H.A.A. checked is stated in Section 2.6 and Table S38. A control run with a transcribed peak-list arm, exchanged candidate letters, reversed order and all conditions in one randomised sequence is specified in Section S7.2 and has not been run.

4 Conclusions

Two large language models chose between drawn candidates for isomeric pyrazolone glycosides, and every value cited in the justification fields of their 72 responses was checked against the values recorded for the spectra supplied. Thirteen fields, all of responses with NMR spectra, cite a value with no counterpart or give for the anomeric signal the value of another signal. An aryl carbon near 104 ppm, also printed for the sugar-free aglycone, was given as the anomeric carbon in 9 of the 12 committed responses that had the spectra of 117a, whose panels labelled and marked the anomeric carbon, and in 10 of the 18 on the 2-deoxy pair, whose panels gave no direct evidence and combined the spectra of two compounds (19 of 35 in all, nine of them in the *Reasoning* field only; 16 of 35 in the second coding). These results do not depend on whether the compounds are N- or O-glycosides. They depend on the assignment of the 104 ppm signal to the aryl ring, which rests on the signal printed for the sugar-free aglycone, on its carbon–fluorine coupling in the original data, on the HSQC figures of the source and on the gHMBC panel of 117a, and for the panel supplied for 122a on its coincidence with the aryl signal of the other spectra. A post hoc operational infrared feature, coded once, was named in 15 and 1 of 18 responses on the ribose pair without and with NMR spectra, and in 13 and 8 of 18 on the 2-deoxy pair; the design does not allow these differences to be

attributed to the NMR spectra. Agreement with the source labels on the ribose pair, 14 and 3 of 18, is a secondary result.

Re-examination of the source record qualifies those labels (Section 3.1; Sections S16 and S17). The data are consistent with 117a as the N2-glycoside in its hydrazone–lactam form; for 116a they favour an arylazo form and do not distinguish attachment at O from attachment at N1. Original records show that the ^{13}C spectrum printed for 123a of the 2-deoxy pair belongs to the compound whose ^{1}H spectrum is printed for 122a and that the one printed for 122a does not (its attribution to the other compound follows by elimination); for the compound whose ^{1}H spectrum is printed for 122a, assigned in the source as the O-glycoside, the corrected record favours an N-glycoside and a pyranosyl ring without establishing either. A $^{1}H–^{15}N$ correlation from H-1" could establish attachment at nitrogen, although its absence would not establish an O-glycoside, and a crystal structure could settle the site of attachment in the solid; neither is available.

The audit separates what an answer cites from whether the answer agrees with a label. It needs a value table for every spectrum, a structured justification field, independently identified signals and labelled peaks, and it has been applied here to one compound class, two models and image stimuli only. What the article demonstrates is this auditing procedure on one case; it does not measure the reliability of the two models.

Declarations

Data and Software Availability

Released. The prompts, the stimulus images exactly as supplied to the models, the 90 verbatim model responses (the 72 analysed here and 18 of the supplementary task on the benzoyl pair), the run log with model identifiers, timestamps and prompt hashes, the request code, the value tables of the source spectra, the original script-based coding with its scripts, the pilot workbook and the session screenshots are openly available in Zenodo, with a manifest of SHA-256 checksums: version 1.0.1, version DOI https://doi.org/10.5281/zenodo.23015879. Version 1.1.0, version DOI https://doi.org/10.5281/zenodo.23440612, changes none of these files and adds those of Supporting File 6 (below); the concept DOI for all versions is https://doi.org/10.5281/zenodo.23014095.[14]

Supplied with this article. The coding used in this article is not part of version 1.0.1. It is given in full in the Supporting Information (Table S9, with the verbatim fields in Table S4 and the reading of the *Reasoning* fields in Table S24) and in machine-readable form in Supporting File 2 (ir_feature_coding.csv). Supporting File 3 (code_ir_feature.py) states the rule for the infrared feature as a script, and Supporting File 4 (summarise_coding.py) recomputes the counts of Sections 3.2 and 3.3 from the coding file. Supporting File 5 holds the workbook sent to the second coders, their answers, comments and declarations in all three returns, the comparisons with the first coding, the plan of 9 October 2026 as written on that date (with one name removed and one wrong description replaced, both marked; its blank date field and its section numbers are those of the draft of that day) and the scripts used, among them the one that measures the infrared traces of 116a and 122a. Supporting File 6 holds, with its own README and SHA-256 checksums,

the further files on which the article rests that version 1.0.1 does not contain: Supporting Files 2–5; the ten original data sets that were reprocessed or are described (Section S16), with the processing scripts, the second computation and a catalogue of the 56 data sets that H.A.A. holds; the two printouts of 2016 reproduced in Figure S7, the values read from the processed files of 2014 and the scripts used to read them (Section S16.3); the HSQC images of the thesis, the script used to read the positions of Tables S31 and S32 and the scripts that draw Figures S3 and S4 (Section S15); the two repetitions by the assistant of the reading of the complete responses, with their blind packet (Section S12); and the script of the measurement of the infrared traces of 116a and 122a (Section S13). The archive of Supporting File 6 is the file added to the deposit in version 1.1.0 (MD5 2641a7f46858b075250230e4c1b6237c).

Not available. Three probe responses were not retained (Section S5). The original two-dimensional NMR data and the data files of the ^{13}C spectra printed for 116a, 122a and 123a are not among the original data that H.A.A. holds. The returned workbooks of the second coding carry personal details, and the files of 2014 carry folder paths of the computer on which they were saved; both are kept by the corresponding author, and the numbers read from the files of 2014 are in Supporting File 6. The other scanned printouts read in Section S16.3, among them the three infrared printouts of the 2-deoxy pair and the ^{1}H printouts of the second sample, are kept by the corresponding author as well.

The two models evaluated are proprietary and were accessed through the providers' APIs under the identifiers given in Section 2.3; the similarity calculations used RDKit[15] (release 2026.03.6), and all scripts are written in Python 3 and were run with Python 3.11, NumPy 2.4, SciPy 1.17, Matplotlib 3.10, pandas 3.0, openpyxl 3.1, Pillow 12.2 and nmrglue 0.12. The uploaded supporting files carry the prefix Supporting_File_N_ before the file names used here. The infrared spectra of Figure 2 are the deposited stimulus images, which are crops of figures of the author's doctoral thesis;[9] the gHMBC spectra of Figure 3 are reproduced complete from the thesis. The copyright of the thesis is held by the author.

Supporting Information

Supporting File 1 (PDF): prompts as sent; stimulus images, label and order mapping, source figures; the 90 API responses with their coding; value tables and coding rules; probe for retrievable thesis-specific facts; reaction records of the source; specifications of controls and measurements not performed; spectroscopic record of the four glycosides; outline of a classroom exercise; manifest of the deposited archive; literature values as printed in their sources; reading of the *Reasoning* fields; properties of the stimulus images; second coding; the text and figures of the source re-examined; original spectrometer data reprocessed, with original records of the 2-deoxy pair; re-examination of the source record; the assignment of the signal near 104 ppm and the answers of the second coders in full; comparison with published evaluations, and status of verification of the main claims (Sections S1–S19). Supporting File 2 (CSV): coding of the 72 glycoside responses, one row per response. Supporting File 3 (PY): the rule for naming of the infrared feature. Supporting File 4 (PY): recomputation of the counts of Sections 3.2 and 3.3 from Supporting File 2. Supporting File 5 (ZIP): files of the second coding. Supporting File 6 (ZIP): the files of version 1.1.0 of the data deposit (original spectrometer data reprocessed, records of the 2-deoxy pair, HSQC reading,

repetitions of the reading of the *Reasoning* fields, scripts, and Supporting Files 2–5), with README and SHA-256 checksums.

Author Information

Corresponding Author

Heba A. Aboukhousa – Mohamed Bin Zayed University for Humanities, Abu Dhabi, United Arab Emirates; orcid.org/0000-0001-7667-7243; Email: heba.aboukhousa@mbzuh.ac.ae

Author Contributions

H.A.A. conceived and designed the study and ran all model sessions. With the AI assistance described in Section 2.6 she digitised and re-analysed the source data, compiled the source assignments and audited their spectroscopic support, carried out the first coding and wrote the manuscript; the analysis code and the figures were prepared with that assistance (Section 2.6). The second coding was done by two chemists who are not authors (Section 2.4). H.A.A. is accountable for the work.

Funding

The author received no funding for this work.

Notes

The author declares no competing financial interest. H.A.A. is the author of the doctoral thesis that this article re-examines. The supervisors of the thesis took no part in the present study. The two second coders, who are not authors, are members of the department in which that thesis was prepared; neither supervised nor examined it. No ethics approval was sought: the study involved no animals, and the only persons whose answers are reported are the two second coders, whose answers are reported without attribution and with their consent.

Acknowledgments

The author thanks Dr. Soleiman Hisaindee and Dr. Mohammad Khasawneh (Department of Chemistry, College of Science, United Arab Emirates University) for the independent second coding (Section 2.4), and the supervisory committee and examiners of the doctoral thesis on which the experimental data rest. The author used Claude (Anthropic) as a research and writing assistant; what it prepared, and what the author checked, are stated in Section 2.6 and Table S38.

References

1. Guo K, Nan B, Zhou Y, Guo T, Guo Z, Surve M, Liang Z, Chawla NV, Wiest O, Zhang X (2024) Can LLMs solve molecule puzzles? A multimodal benchmark for molecular structure elucidation. Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Datasets and Benchmarks Track, pp 134721–134746

2. OdanChem, NMRArena (2026) NMR benchmark for general LLMs and ML specialized models. GitHub repository, https://github.com/odanchem/NMRArena (accessed 2026-09-28).
3. Sharlin S, Agbere F, Ishimwe K, Osifová Z, Socha O, Dračínský M, Josephson TR (2026) NMR-Challenge for LLMs: evaluating chemical reasoning in humans and AI. J Chem Inf Model 66:6299–6312. https://doi.org/10.1021/acs.jcim.6c00481
4. Runcie NT, Deane CM, Imrie F (2026) Assessing the chemical intelligence of large language models. J Chem Inf Model 66:216–227. https://doi.org/10.1021/acs.jcim.5c02145
5. Alampara N, Schilling-Wilhelmi M, Ríos-García M, Mandal I, Khetarpal P, Grover HS, Krishnan NMA, Jablonka KM (2025) Probing the limitations of multimodal language models for chemistry and materials research. Nat Comput Sci 5:952–961. https://doi.org/10.1038/s43588-025-00836-3
6. Priessner M, Lewis RJ, Johansson MJ, Goodman JM, Janet JP, Tomberg A (2026) Enhancing molecular structure elucidation with reasoning-capable LLMs. Digit Discovery 5:1237–1251. https://doi.org/10.1039/d5dd00359h
7. Huang L, Zhang Z, Wang S (2026) CheMM-R1: enhancing chemical structure recognition and elucidation with reasoning multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*; ; pp 26902–26921. https://doi.org/10.18653/v1/2026.findings-acl.1341
8. Vorbrüggen H, Ruh-Pohlenz C (2001) Handbook of Nucleoside Synthesis. Wiley, New York
9. Aboukhousa HAG Regioselective synthesis of biologically active pyrazolone nucleosides and their benzoyl analogues. Ph.D. Dissertation, United Arab Emirates University, Al Ain, 2017. https://scholarworks.uaeu.ac.ae/philosophy_dissertations/5
10. Ahmed AMS, Abou-Elkhair RAI, El-Torky AM, Hassan AE (2019) A. 3-Trifluoromethylpyrazolones derived nucleosides: synthesis and antiviral evaluation. Nucleosides Nucleotides Nucleic Acids 38:590–603. https://doi.org/10.1080/15257770.2019.1591445
11. Haikal A, Zohdi HF, Badi Z (2003) 2-(2',3',5'-Tri-*O*-acetyl-β-D-ribofuranosyl)-4-(*p*-tolylazo)-5-trifluoromethyl-2,4-dihydropyrazol-3-one. *Molbank 2003*, M306. https://doi.org/10.3390/M306
12. Abdou IM, Saleh AM, Zohdi HF (2004) Synthesis and antitumor activity of 5-trifluoromethyl-2,4-dihydropyrazol-3-one nucleosides. Molecules 9:109–116. https://doi.org/10.3390/90300109
13. Khalil NSA (2009) M. A facile synthesis, structure, and antimicrobial evaluation of novel 4-arylhydrazono-5-trifluoromethyl-2,4-dihydropyrazol-3-ones, their *N*- and *N*,*O*-. Carbohydr Res 344:1654–1659. bis-β-D-glucosides https://doi.org/10.1016/j.carres.2009.06.003
14. Aboukhousa H (2026) Data and code for the evaluation of two large language models on the N- versus O-glycoside assignment of a pyrazolone [Data set]. Zenodo, Concept DOI for all versions https://doi.org/10.5281/zenodo.23014095; version 1.0.1: https://doi.org/10.5281/zenodo.23015879; version 1.1.0: https://doi.org/10.5281/zenodo.23440612
15. RDKit Open-Source Cheminformatics Software, release 2026.03.6. https://www.rdkit.org (accessed 2026-09-30)
16. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t

17. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280. https://doi.org/10.1021/ci010132r
18. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC international chemical identifier. J Cheminf 7:23. https://doi.org/10.1186/s13321-015-0068-4
19. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46. https://doi.org/10.1177/001316446002000104
20. Turpin M, Michael J, Perez E, Bowman SR Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS* (2023)); 2023; pp 74952–74965); 2023; pp 74952–74965
21. Chen Y, Benton J, Radhakrishnan A, Uesato J, Denison C, Schulman J, Somani A, Hase P, Wagner M, Roger F, Mikulik V, Bowman SR, Leike J, Kaplan J, Perez E (2025) Reasoning models don't always say what they think. *arXiv* arXiv:2505.05410. https://doi.org/10.48550/arXiv.2505.05410
22. Burgart YV, Elkina NA, Shchegolkov EV, Krasnykh OP, Makhaeva GF, Triandafilova GA, Solodnikov SY, Boltneva NP, Rudakova EV (2023) Saloutin, V. I. Powerful potential of polyfluoroalkyl-containing 4-arylhydrazinylidenepyrazol-3-ones for pharmaceuticals. Molecules 28:59. https://doi.org/10.3390/molecules28010059
23. Beghennou A, Rondot O, Corcé V, Botuha C (2024) -1,2,3-Triazolyl-1,6-naphthyridin-7(6*H*)-ones as potential fluorescent nucleoside analogues: synthesis and optical properties. Molecules 29(1):687. https://doi.org/10.3390/molecules29030687
24. Connor MC, Finkenstaedt-Quinn SA, Shultz GV (2019) Constraints on organic chemistry students' reasoning during IR and ^{1}H NMR spectral interpretation. Chem Educ Res Pract 20:522–541. https://doi.org/10.1039/C9RP00033J
25. Osifová Z, Socha O, Dračínský M (2024) NMR-Challenge.com: exploring the most common mistakes in NMR assignments. J Chem Educ 101:2561–2569. https://doi.org/10.1021/acs.jchemed.4c00092

Figures

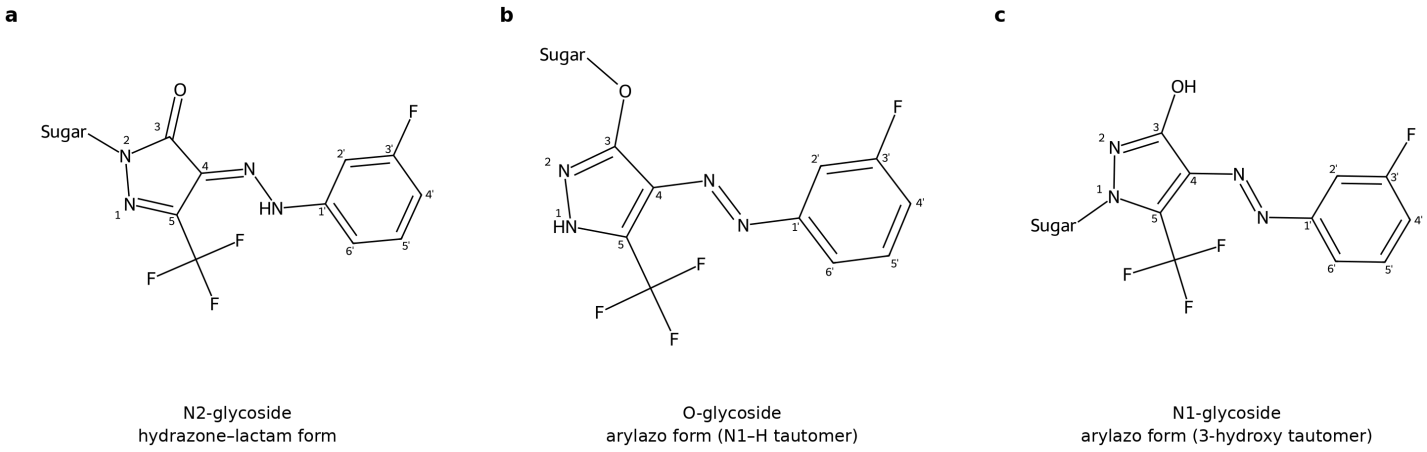


Figure 1

The three attachment isomers of the glycosides of 4-(3-fluorophenylhydrazono)-5-(trifluoromethyl)pyrazol-3-one, with the atom numbering used in this paper: pyrazole atoms unprimed (N2 adjoins C-3), aryl carbons single-primed (fluorine on C-3'); sugar atoms are double-primed (C-1" to C-5"). (a) N2-glycoside in the hydrazone–lactam form. (b) O-glycoside in the arylazo form, drawn as the N1–H tautomer. (c) N1-glycoside, which the source does not consider, drawn as the 3-hydroxy–arylazo tautomer; its other uncharged tautomer is the N2–H 3-oxo–arylazo form, and no uncharged hydrazone form can be written for it, because N1 cannot form the C-5=N1 double bond and C-5 bears no hydrogen. The two candidate drawings supplied to the models, an N2- and an O-glycoside, are deposited as candidates_ribose.png;[14] in them the O-glycoside is drawn as the 4-hydrazono-4*H*-pyrazole tautomer and not in the arylazo form of panel b (Section 2.3).

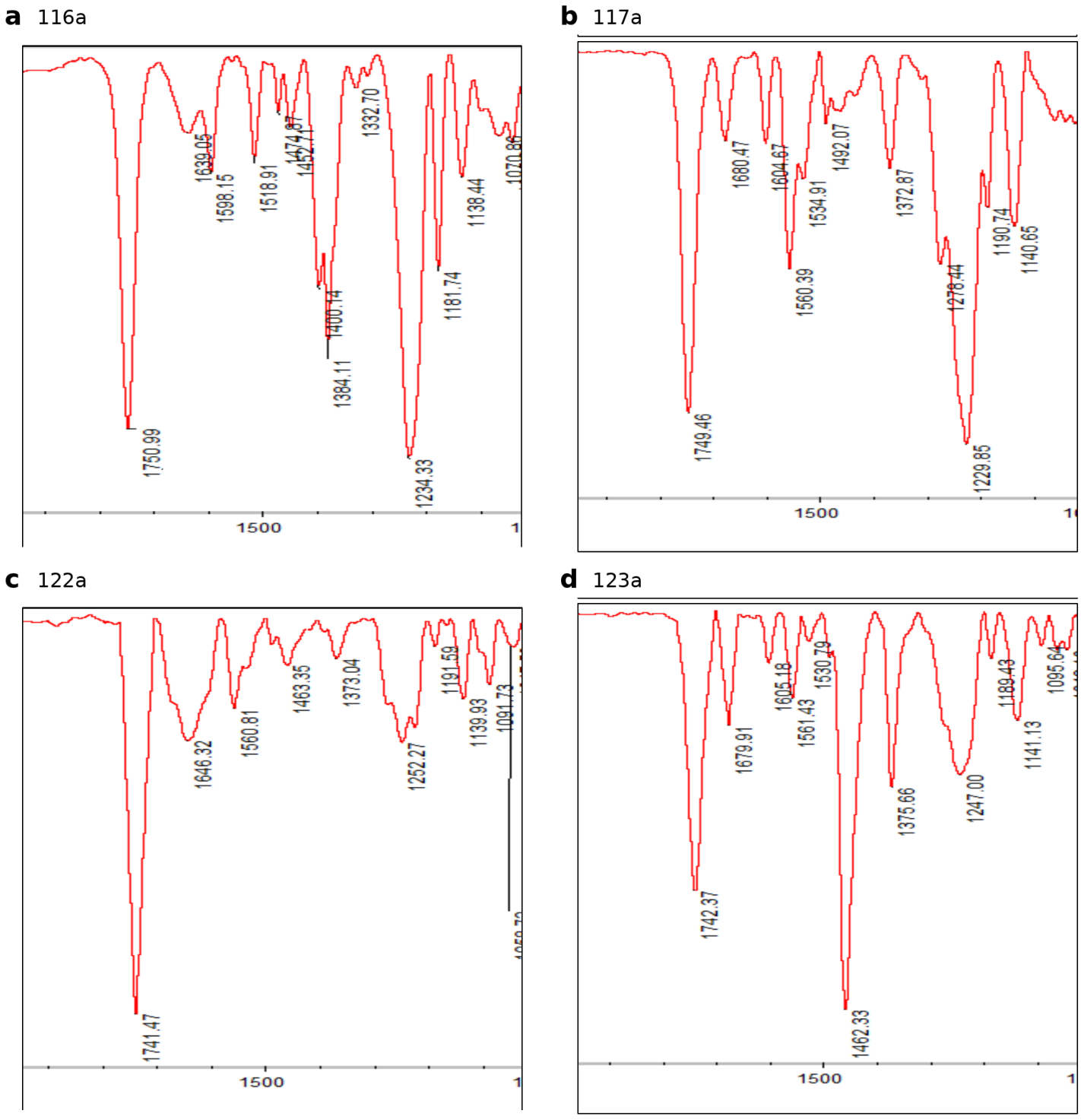
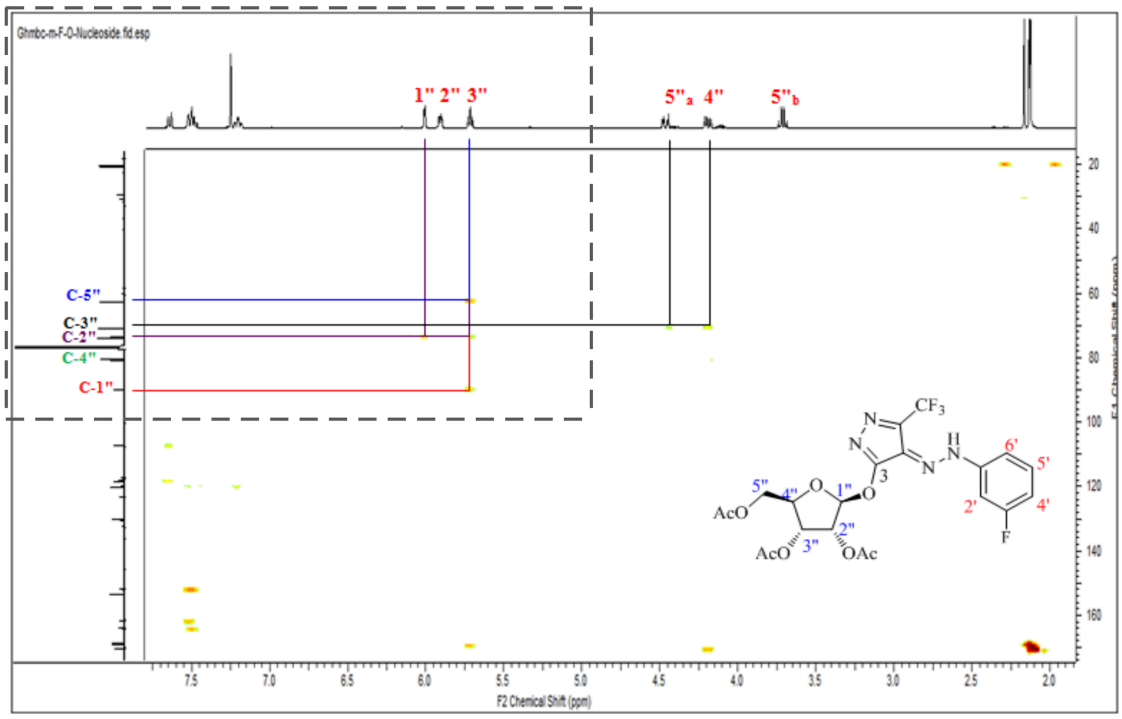


Figure 2

Infrared spectra of the four glycosides as supplied to the models: crops of Figures 23 and 44 of the source[9] covering about 1950–1030 cm^{-1}, in which the axis ticks are 100 cm^{-1} apart and 1500 is the only complete axis label. (a) 116a and (b) 117a (ribose pair); (c) 122a and (d) 123a (2-deoxyribose pair). 117a and 123a show a labelled band at 1680.47 and 1679.91 cm^{-1}; 116a and 122a show no labelled band at 1670–1690 cm^{-1} and instead one at 1639.05 and 1646.32 cm^{-1}.

a 116a



b 117a

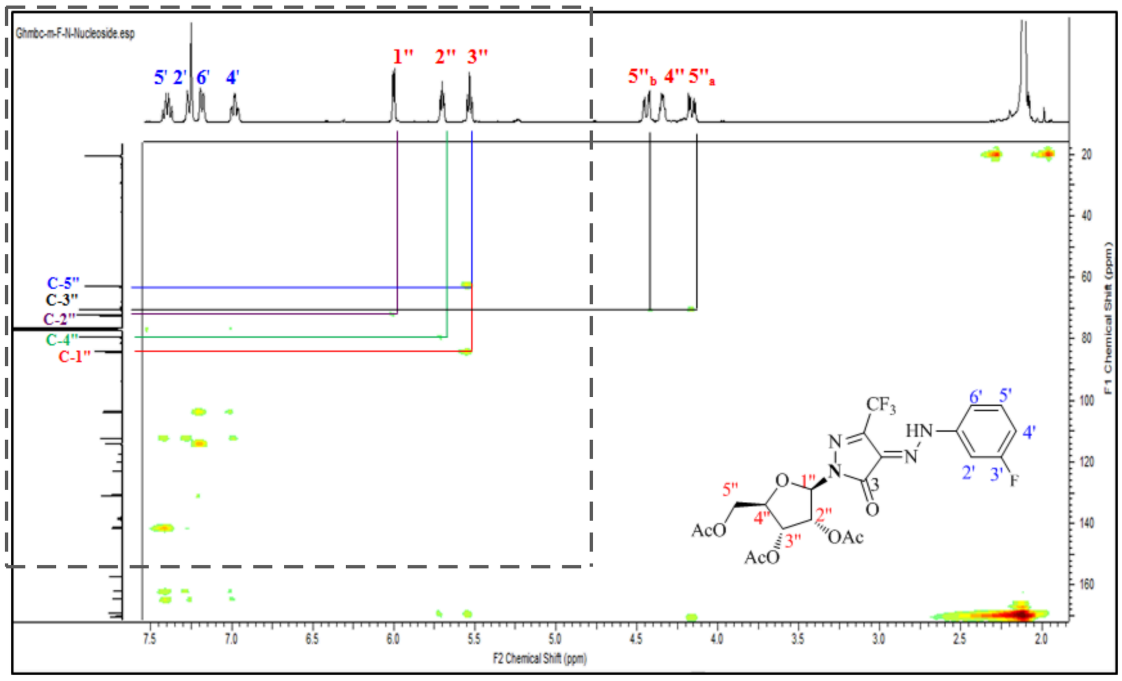


Figure 3

gHMBC spectra of (a) 116a and (b) 117a, reproduced complete from the source (its Figures 20 and 29[9]) with the atom labels, guide lines and structure drawings of the source. The broken grey frame, added here, marks the part of each figure that was supplied to the models; in the images supplied the file name was masked and, in panel b, the labels of the aromatic protons and of H-1" were not visible (the images as supplied are in the data deposit[14]). The annotated correlations connect sugar protons with sugar carbons

(C-1" to C-5"); neither axis has a printed scale inside the frame. In panel b the cross-peaks below the row annotated C-1" lie in the columns of the aromatic protons (Section 3.2).

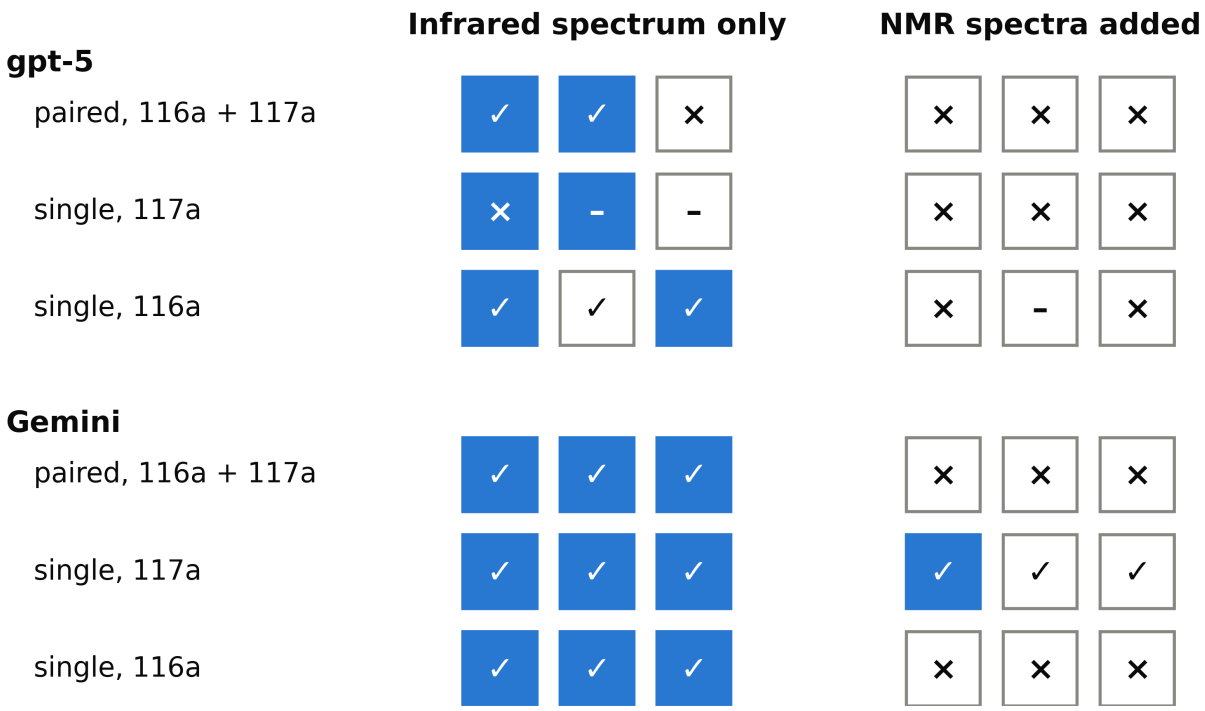


Figure 4

The 36 responses on the ribose pair, one square per response, by system, item and evidence supplied. Filled squares, the infrared feature is named in the *Decisive evidence* field; open squares, it is not. The symbol in each square shows whether the response agrees with the source label (✓), disagrees with it (×) or abstains (–).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CPSupportingInformation20261018t.pdf
- TOCgraphic.png