



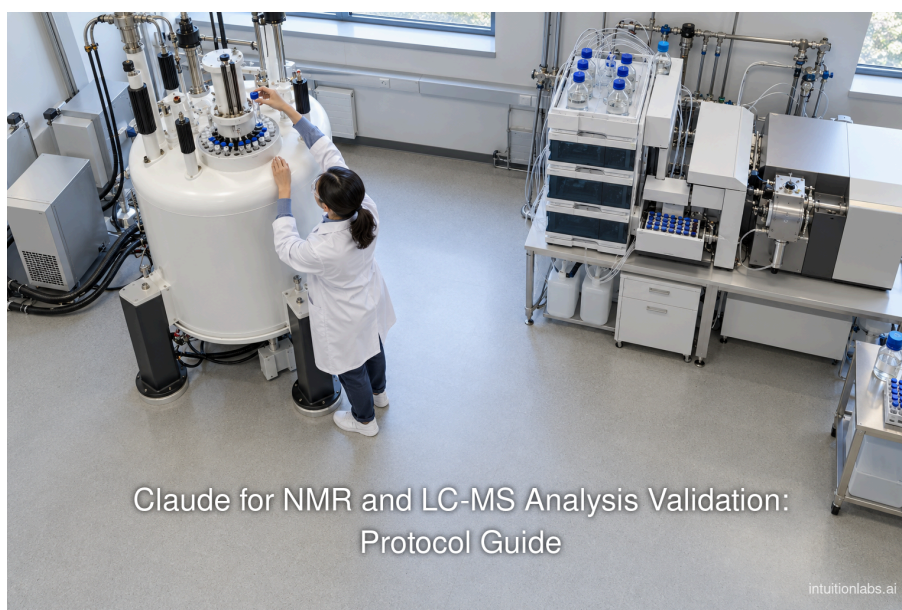
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude for NMR and LC-MS Analysis Validation: Protocol Guide

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-31 · [claude nmr](#) · [ai nmr analysis](#) · [ai lc-ms analysis](#) · [ai structure elucidation](#) · [claude chemistry](#) · [ich q2r2 ai validation](#) · [gxp generative ai validation](#) · [21 cfr part 11 ai](#) · [nmr lc-ms validation protocol](#)

Original article: <https://intuitionlabs.ai/articles/claude-nmr-lc-ms-validation-protocol>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Key Changes: What the 2026 Evidence Does and Does Not Establish
 4. Implementation Considerations: A Four-Task Validation Protocol
 5. Reference Tools and Competitive Context
 6. Data Analysis and Evidence
 7. Implications and Future Directions
 8. Frequently Asked Questions (FAQs)
 9. Conclusion
-

Executive Summary

Anthropic published two sets of results in 2026 that are often discussed together as "Claude for NMR" but support four different claims: forward nuclear magnetic resonance (NMR) prediction, inverse structure elucidation, raw NMR file processing, and liquid chromatography-mass spectrometry (LC-MS) processing with a purity calculation. The **June 5, 2026** white paper assessed **Claude Opus 4.7** against **ChemDraw 25.0.2** and **MestReNova 17.0.0** (www-cdn.anthropic.com), using 20 forward-prediction compounds and 15 elucidation problems (www.anthropic.com). The **August 18, 2026** report followed one routine quality-control sample from a contract research organization through **Claude Opus 5** (www-cdn.anthropic.com).

The June numbers are favorable but small. Opus 4.7 posted a mean ¹H error of **0.079 ppm**, the lowest of any tool tested (www-cdn.anthropic.com), and on ¹³C it was comparable with MestReNova at 1.37 and 1.48 ppm (www-cdn.anthropic.com). It recovered all eight simpler inverse problems on every attempt (www.anthropic.com). Anthropic itself calls the results indicative rather than precise and excludes 2D experiments and stereochemistry (www.anthropic.com). The August workflow returned processed NMR and LC-MS results in **23 and 19 minutes** (www.anthropic.com), with a purity of 96.4% against the laboratory's 96.33% (www.anthropic.com). The technical report shows those two figures rest on different peak sets: three peaks for Claude and seven for the laboratory (www-cdn.anthropic.com).

No regulator has published a validation standard for this use. The usable anchors are the International Council for Harmonisation (ICH) Q2(R2) principle that a procedure must be fit for its intended purpose (database.ich.org), the validation requirement in 21 CFR Part 11 (www.ecfr.gov), and the draft EU GMP Annex 22, which as drafted does not apply to generative AI and large language models. EMA said in mid-2026 that it was still weighing consultation results (www.ema.europa.eu).

This report proposes a four-task pilot protocol: a written intended-use statement per task, a blinded post-cutoff reference set, replicate runs with pinned model identifiers, peak-level error metrics, top-k tables, raw-file checksums with reconciliation, and named human review. It sets no universal acceptance thresholds, because those belong to each laboratory's own risk assessment. Nothing here suggests an AI result establishes identity, purity, release, or GMP suitability. All performance figures are vendor research on small samples.

Introduction and Background

Searches for "Claude NMR" bundle analytical jobs that differ in input, reference method, and consequence of error. A chemist who pastes a SMILES string (a line-text structure notation) and asks for a predicted ^1H spectrum is doing a different task from an analyst who uploads a raw free induction decay (FID, the time-domain signal an NMR spectrometer records) and asks for an integrated peak table. Both differ again from asking a model to read a vendor LC-MS binary and report area-percent purity.

This report examines how a laboratory can convert the 2026 results into a pilot. It is written for medicinal and process chemists, analytical development leaders, chemistry informatics teams, lab automation owners, research IT, and quality partners. The perspective is that of an adjacent advisor: IntuitionLabs is a life-sciences and AI consultancy, not a vendor of NMR, LC-MS, or AI software, and has no product in the comparisons below. The report does not claim that Claude replaces ChemDraw, MestReNova, instrument software, or a qualified analyst. Anthropic's June post is similarly modest, describing Claude as starting to assist with translation, recall, and integration work that complements chemists' judgment (www.anthropic.com).

Three points frame the analysis. The instrument run is short (two to three minutes for a routine proton spectrum, about 10 for LC-MS) and the tedious part is analyzing the output (www.anthropic.com). The evidence base is two vendor-authored studies, with no independent replication located in the sources reviewed. And the regulatory frame was unsettled as of October 2026: the main GMP-specific artificial intelligence (AI) text is a consultation draft, as is the main FDA AI document.

Terms used below: **forward prediction** is structure in, spectrum out; **inverse elucidation** is spectrum and formula in, candidate structures out; **raw processing** is vendor file in, calibrated spectrum or chromatogram out; and **purity** means UV area percent, which differs from content by weight. Dates and model availability are stated as of October 2026.

Key Changes: What the 2026 Evidence Does and Does Not Establish

Change 1: A June benchmark of forward and inverse NMR tasks

The white paper, authored by David Kamber (www-cdn.anthropic.com), drew compounds from ChemRxiv preprints posted after the models' training cutoff, taking the first fully characterized novel compounds per preprint (www-cdn.anthropic.com). Compounds were chosen and locked before predictions were generated. Each Claude model was queried three times per compound, while the two deterministic classical tools ran once (www-cdn.anthropic.com). Scoring used fixed denominators of 401 hydrogen and 225 carbon atoms (www-cdn.anthropic.com), and a peak counted as correct within ± 0.20 ppm for hydrogen or ± 1.0 ppm for carbon (www.anthropic.com). About 80% of Claude's sub-peak spacings fell within half a hertz (www.anthropic.com).

The inverse task asked Opus 4.7 for up to three ranked SMILES candidates per attempt, with stereochemistry excluded (www-cdn.anthropic.com). Eight simpler targets received only the high-resolution mass spectrometry (HRMS) formula and 1D spectra and were all recovered on every attempt. Seven denser targets also received

the starting-material structure; four were solved on all three runs and the rest on two of three (www.anthropic.com). Without that hint the model could loop without committing to a structure (www.anthropic.com).

Several caveats belong in any pilot design. MestReNova's 1H score excluded experimental multiplets wider than 0.3 ppm, so tools were not scored on identical atoms (www-cdn.anthropic.com). Every tool predicted carbonyl carbons slightly low, so agreement between two tools on that shift is not independent evidence (www-cdn.anthropic.com). Anthropic's own wish list is several hundred compounds across 20 to 30 scaffold classes, with at least 15 compounds per class (www.anthropic.com).

Change 2: An August raw-file workflow on one sample

The August 18, 2026 technical report, titled "Automated processing of raw NMR and LC-MS data with Claude Opus 5 on Claude Science", follows one routine QC sample through the Claude Science platform running Claude Opus 5, including its built-in tooling and auditor. A pilot on a different surface, such as the plain API or chat, would therefore be testing a different system. Both measurements begin as raw proprietary instrument files (www.anthropic.com): the NMR session received the first FID with its header (www-cdn.anthropic.com), and the LC-MS session received a single instrument file (www-cdn.anthropic.com) acquired on a single-quadrupole system (www-cdn.anthropic.com). Where Claude's peak table and the operator's integral regions could be matched they agreed within 0.08 H on a common scale, while four weak operator regions of 0.1 to 0.2 H had no counterpart in the table (www-cdn.anthropic.com).

For LC-MS, the report states that decoded intensities matched instrument-stored totals for all 2,664 scans and that all 3,751 photodiode-array blocks parsed with exact byte consumption (www-cdn.anthropic.com). The headline purity was **96.4%** against the laboratory's **96.33%**. The technical report states that the 96.4% sums only three peaks detected from 0 min, void peak included, whereas the laboratory's figure is computed over seven peaks from 3.46 min. On the laboratory's seven-peak basis, Claude's own areas give **98.8%** (www-cdn.anthropic.com). The match of 96.4% to 96.33% is therefore a match of two differently defined numbers, so a laboratory should fix the integration basis in advance.

The self-correction example is instructive. In the heavy-water exchange comparison, Claude's first reading overstated the result and was corrected (www.anthropic.com); the technical report attributes the catch to a built-in auditor (www-cdn.anthropic.com). The comparison is against the laboratory's own processing records, not an independent ground truth (www-cdn.anthropic.com).

Change 3: A moving model and platform baseline

The June study used Opus 4.7, the August report used Opus 5, and Anthropic's October 2026 documentation recommends **Opus 5.5** as the starting point for most workloads (docs.claude.com). Opus 4.7 is listed as active with retirement not sooner than **April 16, 2027** (docs.claude.com). Each model identifier is a pinned version whose underlying model remains constant (docs.claude.com), although serving infrastructure around it can change (docs.claude.com).

Repeatability controls have narrowed. Anthropic's Opus 5.5 migration guide states that temperature zero never guaranteed identical outputs on prior models (docs.claude.com), and the API version dated 2026-10-21 returns a 400 error when temperature, top_p, or top_k is sent, even at the default value (docs.claude.com). Replicate

runs must therefore supply the evidence of consistency. Academic work agrees: one study found accuracy variations up to 15% across naturally occurring runs (arxiv.org), and another found nondeterminism persisting at temperature 0.0 (arxiv.org).

Access is a further variable. The August post says life science research tasks were blocked in Anthropic's most capable model (www.anthropic.com), and the June launch notes describe biology and chemistry requests on Fable 5 falling back to Opus 4.8 (www.anthropic.com). A Life Sciences Verification Program offers vetted organizations more permissive access (www.anthropic.com). A pilot should record which model actually served each run.

Change 4: A regulatory frame without a finished AI rulebook

Method validation comes first. **ICH Q2(R2)** states that validation demonstrates a procedure is fit for its intended purpose, and EMA lists a legal effective date of June 14, 2024 (www.ema.europa.eu). For electronic records, **21 CFR 11.10** requires system validation and secure, time-stamped audit trails (www.ecfr.gov), and FDA's 2003 scope guidance states it interprets Part 11's scope narrowly and recommends a documented risk assessment (www.fda.gov).

The AI-specific layer remains in draft. The EU **draft Annex 22** applies to AI models in critical applications with direct effect on patient safety, product quality, or data integrity (health.ec.europa.eu). It covers only static (health.ec.europa.eu) and deterministic (health.ec.europa.eu) models and says it does not apply to generative AI and large language models (health.ec.europa.eu). For non-critical uses it expects qualified personnel to remain responsible for output suitability (health.ec.europa.eu). EMA's mid-2026 workshop report records consultation comments supporting a wider approach and states that the drafting group has discussed allowing generative AI and LLMs in GMP applications under a documented, risk-based control strategy, so the draft exclusion may change (www.ema.europa.eu), but no adopted text was located. FDA's AI credibility draft is dated January 2025 and still described as a draft (www.fda.gov); it excludes drug discovery and operational-efficiency uses that do not affect patient safety, drug quality, or study-result reliability (www.fda.gov).

Implementation Considerations: A Four-Task Validation Protocol

Intended-use statements and risk classification

Write one intended-use statement per task and do not let one task's evidence support another. ICH Q14 frames this through an analytical target profile (ATP) that drives technology choice (database.ich.org). The FDA draft starts from the question of interest and the context of use, the specific role and scope of the model (www.fda.gov), and treats model risk as a combination of model influence and decision consequence (www.fda.gov). A laboratory can borrow that logic without claiming the draft applies to its work.

Table 1 below summarizes the four claims side by side, using the vendor's own design descriptions. The risk column is this report's analytical judgment, not a regulatory classification.

Task	Input	Reference and ground truth	Replicates and n	Documented limits	Risk if wrong, and minimum control
1. Forward NMR prediction	SMILES string plus solvent	ChemDraw and MestReNova; published experimental peak lists transcribed by a chemist	Three runs per compound; 20 compounds, four scaffold families	No 2D or stereochemistry ; three solvents (www-cdn.anthropic.com)	Low to moderate for planning; chemist compares with the experimental spectrum
2. Inverse elucidation	HRMS formula, 1H and 13C peak lists; starting material for hard cases	Published target structures	15 problems, three attempts, up to three candidates	Hard cases needed a hint; no stereochemistry	Moderate to high; candidates are hypotheses until confirmed by orthogonal data
3. Raw NMR processing	Raw FID with header	Contract laboratory operator processing	One sample, one run	No replicate statistics reported; weak peaks absent	Moderate; analyst overlays output on instrument-software spectrum
4. LC-MS processing and purity	Proprietary single-quadrupole run file	Laboratory seven-peak PDA table	One sample, one run	Nominal mass only (www-cdn.anthropic.com); different peak basis	Moderate to high if area percent informs a go or no-go decision; written integration basis and reconciliation to instrument software

Table 1 shows that only tasks 1 and 2 have any replicate design and only tasks 3 and 4 involve raw instrument files, so evidence cannot be pooled across them. The risk column is what each organization must replace with its own assessment.

Within a GxP (good practice quality) context, two gates apply before any task touches regulated work. First, whether the output enters a regulated record at all: FDA's data integrity guidance lists the ALCOA attributes (attributable, legible, contemporaneous, original or true copy, accurate) (www.fda.gov), and MHRA adds complete, consistent, enduring, and available (assets.publishing.service.gov.uk). Second, the Annex 22 scope question: for critical GMP uses the draft excludes language models, so a pilot there needs a documented regulatory position first. Teams should consult qualified quality and regulatory professionals for that determination.

Reference dataset design

The June design is a template and shows what to improve. A laboratory dataset should include:

- **Post-cutoff or in-house compounds**, with the model's stated knowledge cutoff recorded from vendor documentation.
- **Scaffold diversity** matched to the lab's chemotypes. Anthropic's four families each probe one NMR challenge, such as a slow-exchange NH proton, and the authors name untested scaffolds (www-cdn.anthropic.com).
- **Solvents** used locally, tested separately, since the June set covered only DMSO-d6, CDCl3, and D2O.

- **Each instrument model** feeding the pilot, since file headers differ. The August sample was a 400 MHz DMSO-d6 spectrum.
- **Impurities, residual solvents, and water**, because purity claims depend on them.
- **Blinded ground truth.** The Annex 22 draft says test data must not be used in development, training, or validation (health.ec.europa.eu), and that generating test labels with generative AI is not recommended (health.ec.europa.eu).
- **Adequate sample size.** The draft requires adequate statistical confidence but gives no numeric minimum (health.ec.europa.eu).

Table 2 converts the June coverage and limits into a checklist a laboratory can extend.

Dimension	June benchmark coverage	Gap to close in a lab pilot
Compounds	20 forward and 15 inverse, from four ChemRxiv preprints	Own chemotypes; several hundred if feasible
Scaffolds	Chloropyridazines, Boc-N-aryl maleimides and N-Boc ynamides, spiroketones, alpha-silyl methanesulfonamides	Other NH-active heteroaromatics; natural products
Solvents	DMSO-d6, CDCl3, D2O	Methanol-d4, benzene-d6, acetone-d6, others
Experiments	1D 1H and 13C only	2D COSY, HSQC, HMBC if the use case needs them
Stereochemistry	Excluded by design	Separate task with separate evidence
Scoring	401 hydrogens and 225 carbons	Report coverage and omitted atoms per tool
Replicates	Three per Claude model	Enough runs to estimate spread; pin the identifier

Table 2 makes the external-validity problem concrete: a laboratory working with heteroaromatic amines in methanol-d4 has essentially no direct coverage in the June paper.

Repeatability and model-version controls

ICH Q2(R2) distinguishes repeatability from intermediate precision. Its repeatability design calls for a minimum of 9 determinations across the reportable range or 6 at 100% of test concentration (database.ich.org), and it encourages design-of-experiments studies for intermediate precision (database.ich.org). Those numbers describe chemical procedures and do not transfer to language-model runs, but the structure does: repeat identical inputs, then vary the factors that could matter. Suggested controls:

- **Pin the identifier** and record it for every run. Anthropic states it gives at least 60 days of notice before retiring a publicly released model (docs.claude.com).
- **Record platform state**, including timestamps and tool-call logs, since infrastructure can change under an unchanged identifier.

- **Run replicates** and report the spread, not only the mean. A 2026 agentic NMR paper notes that some tasks converge on one answer while others do not (arxiv.org).
- **Freeze the prompt and attachments.** The August prompts were short plain-language requests, which are easy to version and equally easy to vary by accident.
- **Treat changes as events.** The Annex 22 draft expects change control before deployment (health.ec.europa.eu).

Data handling needs review. Claude Science keeps raw datasets on the user's own systems and sends only the context needed for each step (www.anthropic.com), while prompts and responses are processed under standard retention (claude.com). Anthropic's code-execution documentation states that container data, including uploaded files and outputs, is retained for up to 30 days (docs.claude.com), and uploaded files cannot be modified after upload (docs.claude.com). Compliance documentation is routed through Anthropic's trust portal (support.claude.com); the pages reviewed contained no statement on 21 CFR Part 11 conformance for Claude products.

Forward-prediction acceptance metrics

Compute, per nucleus and per scaffold: mean absolute error (MAE), root-mean-square error (RMSE), the fraction of atoms within a stated tolerance, and coverage (the fraction of experimental atoms for which a prediction was returned). Tolerances are a laboratory decision derived from how the prediction will be used, such as distinguishing regioisomers.

(Hypothetical Example) A chemist scores six ¹H peaks with absolute errors of 0.03, 0.05, 0.12, 0.08, 0.31, and 0.02 ppm. MAE is 0.61 divided by 6, or 0.102 ppm. RMSE is the square root of 0.1207 divided by 6, or 0.142 ppm. Five of six peaks, 83.3%, fall within ± 0.20 ppm. The single 0.31 ppm outlier lifts RMSE about 40% above MAE, so both are reported and the outlier peak is named so a reviewer can see whether it is an exchangeable proton.

Report in the scorecard: MAE and RMSE for each tool on identical atoms, the within-tolerance percentage with a confidence interval, atoms excluded, per-scaffold results, the replicate range, and a systematic-bias table. Always keep the experimental spectrum in the comparison, because two predictors can share an error.

Inverse-elucidation ranking and confirmation

Define success before running. Anthropic reports success counts out of three attempts with up to three ranked candidates, and the public text reviewed does not state whether success required the top-ranked candidate, so a laboratory should specify top-1 and top-3 separately. A candidate that matches constitution is not a confirmed structure, because stereochemistry is outside the design.

From Anthropic's reported counts, the derived arithmetic is 24 of 24 attempts on the eight simple problems and 12 plus 6 correct attempts out of 21 on the seven harder problems (www-cdn.anthropic.com), or **42 of 45** attempts (93.3%). Attempts cluster within problems, so the effective sample is 15 problems, not 45. As a worked bound, a tool that solves 8 of 8 independent problems has an exact one-sided 95% lower confidence bound on its success rate of about **69%**, so small perfect scores do not substitute for a larger set.

Metric choice changes the picture elsewhere in the literature. A 2025 model (ChefNMR) reported 40% top-1 accuracy on simulated natural-product spectra ([arxiv.org](#)), and 56% top-1 on a teaching set but 21% on experimental NMRShiftDB2 spectra ([arxiv.org](#)). Another reported 60.4% within the first 15 predictions ([arxiv.org](#)), a hit counted if the target appears among 15 candidates ([arxiv.org](#)). A 2026 LLM-agent system reported 71% top-1 on a 15-molecule set ([arxiv.org](#)). These are not interchangeable with Anthropic's results.

Confirmation workflow:

- **Log all ranked candidates**, not only the one selected.
- **Forward check** each candidate with an independent predictor against the experimental peak list.
- **Require orthogonal data**, such as 2D NMR. Established computer-assisted structure elucidation (CASE) software documents 2D inputs; one open-source system builds structures from 2D connectivity without a shift database ([nuzillard.github.io](#)).
- **Insert controls**: problems with known answers and problems with no valid answer, recording whether the model commits to a wrong structure.

Raw FID and LC-MS integrity, parsing, calibration, and reconciliation

The August report documents reconciliation steps any laboratory can require.

1. **Hash at ingest.** Compute a **SHA-256** digest of each raw file before upload. FIPS 180-4 notes that a changed message yields a different digest with very high probability ([csrc.nist.gov](#)).
2. **Preserve originals** in the controlled system. 21 CFR 211.194 requires a record of all graphs, charts, and spectra from laboratory instrumentation ([www.ecfr.gov](#)), and FDA describes reprocessible chromatographic records as dynamic ([www.fda.gov](#)).
3. **Reconcile the parse** against quantities the instrument stored, as the LC-MS run did with all 2,664 scans. For NMR, echo acquisition parameters from the header and compare with the instrument record.
4. **Demand processing parameters** (phase, apodization, baseline, referencing, fitting) in the report. The August report, for example, reported a coupling constant only where a reconstructed first-order pattern reproduced every observed line.
5. **Check calibration.** Single-quadrupole data give nominal mass only; accurate-mass instruments are described as capable of precision of 10 ppm or better ([www.chromatographyonline.com](#)), and unit mass resolution has a fundamental limitation when candidates give near-identical spectra ([www.thermofisher.com](#)).
6. **Define the purity basis**: wavelength, start time, peak list, and void-peak treatment. The August report shows why: the same sample gives different purities on a three-peak and a seven-peak basis.
7. **Cross-check in instrument software** and an independent converter. ProteoWizard's msConvert converts between file formats ([proteowizard.sourceforge.io](#)), though vendor formats need Windows and vendor libraries ([proteowizard.sourceforge.io](#)). The open nmrglue library reads Bruker, Varian, JCAMP-DX, nmrML and other NMR formats ([nmrglue.readthedocs.io](#)).

Purity needs separate care. Diode-array peak-purity checks can miss impurities at much lower absorbance than the main compound ([www.chromatographyonline.com](#)) or with matching spectra ([www.chromatographyonline.com](#)). The Journal of Medicinal Chemistry editors note that HPLC typically will not

detect solvent residues, water, or inorganic impurities ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)), and, in a 2014 editorial, the journal named quantitative ¹H NMR following its protocol an accepted purity method ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). A literature qNMR example found a commercial sample sold as 97% pure to contain 78.4% w/w of the labeled compound ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). An AI-computed area percent should never be presented as content or release purity.

Human review, exception handling, and audit artifacts

Assume the model can be wrong in a fluent way. NIST defines confabulation as confidently stated but erroneous or false content (nvlpubs.nist.gov), and a Nature Chemistry benchmark found many models cannot reliably estimate their own limitations (www.ncbi.nlm.nih.gov). Model self-reports are not a control. Review controls that map to existing GxP expectations:

- **Second-person review.** 21 CFR 211.194 expects a second person to show that original records were reviewed for accuracy, completeness, and compliance (www.ecfr.gov).
- **Output checking.** 21 CFR 211.68 requires checks of computer input and output for accuracy (www.ecfr.gov).
- **Audit trail review**, which FDA likens to assessing cross-outs on paper (www.fda.gov).
- **Authority checks and signatures** (www.ecfr.gov), with record changes that do not obscure previously recorded information (www.ecfr.gov).
- **Reconstruction.** MHRA states raw data must permit full reconstruction of the activities (assets.publishing.service.gov.uk), so store prompts, files, tool logs, model identifier, and outputs together.
- **Exception triggers.** The Annex 22 draft asks that models log confidence scores (health.ec.europa.eu). Language models do not provide calibrated confidence by default, so define explicit triggers: unmatched peaks, integrals outside tolerance, replicate disagreement, or any unparsed data stream.

Claude Science records an auditable history of how each output was made (www.anthropic.com), a useful artifact that does not by itself make a system validated for Part 11 purposes. Annex 11 requires validated applications and qualified infrastructure (health.ec.europa.eu). ISPE's GAMP AI guide is described as suitable for all types of computerized systems using AI (ispe.org), and FDA's computer software assurance guidance applies a risk-based approach to AI/ML tools (www.fda.gov), though it addresses device manufacturing rather than drug GMP.

Pilot scorecard and stop conditions

End the pilot with a decision per task, not one decision for "Claude". Define pass criteria before data are unblinded, derived from the intended use. Record:

- **Accuracy:** peak-level error, within-tolerance fraction, top-k success, integral agreement, purity agreement on a stated basis.
- **Repeatability:** replicate spread and the share of inputs where replicates disagree on the final answer.
- **Coverage:** atoms, peaks, or scans for which a result was returned.

- **Review burden:** measured analyst minutes and correction cycles. The laboratory operator's 1 min 49 s of hands-on NMR processing (www-cdn.anthropic.com) excludes review and interpretation time, so each lab should measure its own.
- **Exceptions:** count, category, and time to resolve.

(Hypothetical Example) If processing takes 4 minutes of analyst time today, and the pilot measures 3 minutes of review plus 0.4 correction cycles at 2 minutes each, the pilot cost is 3.8 minutes, which is not meaningfully better. Savings appear only if review time is lower than manual time and the exception rate stays low. The numbers are invented to show the arithmetic.

Stop conditions should be explicit: a confident wrong structure on a control problem with no review flag; unexplained replicate disagreement on a final answer; a raw-file parse that cannot be reconciled; a purity basis that cannot be reproduced; a model change that alters results on the locked reference set; or AI output entering a regulated record without the agreed review step. The outcome of a stop is a documented decision to narrow scope, retrain staff, or end that task's pilot.

Reference Tools and Competitive Context

The pilot should run against current reference software. *Table 3* summarizes documented capabilities from vendor and project pages, which are producers' statements about their own tools. It does not rank them, and no prices are listed because the pages reviewed published none.

Tool	Documented scope	Inputs and formats	Documented limits or notes
Mestrelab Mnova	Verify checks a proposed structure against analytical data (mestrelab.com) with a quality score from -1.0 to +1.0 (mestrelab.com); Structure Elucidation ranks candidates using ¹³ C shift predictions (mestrelab.com); qNMR purity uses a library method (mestrelab.com)	Many vendor NMR formats (support.mestrelab.com); MSChrom opens LC and GC-MS data from several vendors (mestrelab.com)	Pages reviewed state no accuracy figures for Verify or Structure Elucidation
NMRium	Open-source browser NMR processing; online app available for non-commercial use (nmrium-docs.pages.dev)	Auto-processes raw FIDs; prefers a post-FT spectrum if both exist (nmrium-docs.pages.dev)	No validation statements in the documentation reviewed
ChemDraw NMR prediction (Revvity Signals)	1D ¹ H and ¹³ C; equimolar mixtures only (revvitysignals.com)	Drawn or imported structures	Revvity lists separate NMR Predictors with incremental, HOSE code, and neural network algorithms (revvitysignals.com)
ACD/Labs Spectrus Structure Elucidator Suite (now under Revvity Signals)	Generates and ranks candidate structures from NMR and MS data (revvitysignals.com); numerical match factor for verification (revvitysignals.com)	Structure and spectral data	Offers 3D configuration selection using NOESY or ROESY data (revvitysignals.com)

Tool	Documented scope	Inputs and formats	Documented limits or notes
Bruker software	CMC-se supports elucidation and verification workflows (bruker.com); CMC-assist is listed as end-of-life (www.bruker.com); deep learning added to TopSpin for 1H signal detection (www.bruker.com)	Bruker instrument data	NMRtist is a cloud AI platform for protein NMR, not small-molecule verification (www.bruker.com)
Open-source CASE and tools	SENECA ranks candidates by evolutionary search (pmc.ncbi.nlm.nih.gov); DP4-AI automates raw NMR processing and assignment (pubs.rsc.org); nmrshiftdb2 is an open database (nmrshiftdb.nmr.uni-koeln.de)	Varies by project	Academic software with small validation sets, for example 41 molecules for SENECA

The table indicates that established tools differ from Claude mainly in what they document and how they are controlled. Commercial packages specify supported formats, and the two classical predictors in the June study are deterministic, while Claude's strengths in the June paper were on peak shape and coupling prediction and on 1D elucidation without 2D data. A pilot should compare task by task, not tool against tool.

Data Analysis and Evidence

The public quantitative record is small.

Forward prediction. Opus 4.7's mean 1H error of 0.079 ppm is under half the 0.20 ppm window. On 13C the gap to MestReNova is 0.11 ppm (1.37 versus 1.48), small against a 1.0 ppm tolerance. Claude values are averages of three runs while classical tools are single points, and the authors report that Opus 4.7's run-to-run error varied by less than its margin over the next-best tool (www-cdn.anthropic.com), with ranges shown only in figures. For context, a graph-network 13C predictor reported accuracy comparable to the best density functional methods at 1.5 ppm (pubmed.ncbi.nlm.nih.gov). Test sets differ, so no ranking is implied.

Inverse elucidation. The derived counts are 24 of 24 attempts on simple problems and 18 of 21 on harder ones. Trained models report different metrics. MSNovelist ranked the correct structure first for 25% of 3,863 spectra and retrieved it for 45% (www.nature.com). DiffMS reported more than doubling accuracy on a harder benchmark (arxiv.org), while its authors caution that another test set contains molecules nearly identical to training molecules (arxiv.org). A 2026 benchmark found in-domain retraining far outperformed general models (arxiv.org), and models trained on computed spectra drop in performance on experimental NMR measurements (arxiv.org). Test-set provenance can move accuracy more than architecture.

Raw processing and LC-MS. The quantitative statements are the timings, the matched-region agreement, and the purity pair. The arithmetic matters: 96.4 minus 96.33 is 0.07 percentage points, but 98.8 minus 96.33 is 2.47 percentage points on the laboratory's own basis. Timing definitions also need care: the 23 and 19 minutes run from prompt to delivered answer, while the laboratory's finished report is reported to have arrived four days after the first spectrum (www.anthropic.com). Calendar lag and analyst effort are different quantities, and neither measures review burden.

Variability and capability. One study of five language models configured to be deterministic found none gave repeatable accuracy on every task (arxiv.org). A 2,788-question chemistry benchmark found the best models outperformed the best human chemists on average (www.ncbi.nlm.nih.gov), yet predicting the number of NMR signals from SMILES proved hard even for the best models, for example 22% correct for o1-preview (www.ncbi.nlm.nih.gov).

Regulatory numbers. ICH Q2(R2) gives design minima for chemical procedures, such as at least five concentrations for linearity (database.ich.org) and a signal-to-noise ratio of at least 10:1 for the quantitation limit (database.ich.org). Integration quality depends on a flat baseline (nmrlab.technion.ac.il), and quantitative NMR work uses long relaxation delays, with shorter ones justified only when reduced accuracy is acceptable (pmc.ncbi.nlm.nih.gov).

Implications and Future Directions

Intended use. The evidence most directly supports low-consequence, expert-reviewed use: forward prediction for planning, first-pass peak tables from raw files that an analyst overlays on instrument-software output, and candidate generation for 1D elucidation with a documented starting material. It does not support unreviewed identity, purity, or release statements.

Regulated work. The draft Annex 22 treats language models as outside its scope, while EMA's workshop record shows the question is under discussion (www.ema.europa.eu). The joint FDA and EMA principles call for risk-based approaches with proportionate validation (www.ema.europa.eu) and performance assessment of the complete system including human-AI interactions (www.ema.europa.eu). The human review step is therefore part of the system under test, and documentation should say so. MHRA's 2026 commentary on AI-assisted inspection responses likewise expects technical review by experienced people (mhrainspectorate.blog.gov.uk).

Evidence gaps. Replicate statistics for raw-file processing do not exist in the public record, independent replication of the June benchmark is absent from the sources reviewed, and the benchmark has not been extended to the scaffolds, solvents, and 2D experiments Anthropic lists as next steps. Laboratories that publish de-identified pilot results on shared reference sets would close these gaps faster than additional vendor demonstrations. Anthropic expects Claude's scientific judgment to become more acute (www.anthropic.com), so any pilot should be repeated when the model identifier changes.

Tool owners. A hybrid arrangement is plausible: a language model for flexible parsing and hypothesis generation, then confirmation with deterministic software. Quality partners belong in the first protocol draft.

Frequently Asked Questions (FAQs)

Can Claude be used for NMR spectra interpretation?

In Anthropic's June study, Opus 4.7 matched or beat ChemDraw and MestReNova on average for forward 1D shift prediction and solved the simple inverse problems. The study covered 20 and 15 compounds and excluded 2D NMR and stereochemistry. Interpretation of a real spectrum should remain an analyst's task, with Claude as a first-pass assistant.

Can an LLM analyze LC-MS data?

In one documented case Claude decoded a proprietary single-quadrupole file and reproduced instrument totals for all scans, and the resulting purity depends on how peaks are counted. This is a single-sample demonstration. Trained spectrum-to-structure models report modest accuracy on broad MS/MS benchmarks, for example 26% top-1 on the CASMI 2016 challenge (www.ncbi.nlm.nih.gov), so structure assignment from MS/MS is a different task from file processing.

How should AI tools be validated in an analytical chemistry laboratory?

Start from an intended-use statement and ATP, build a blinded reference set, run replicates, report error and top-k metrics, reconcile raw-file parsing, and document human review. Acceptance criteria are laboratory decisions; the Annex 22 draft sets no numeric limits and requires only that criteria be at least as high as the process replaced (health.ec.europa.eu).

Does ICH Q2(R2) apply to AI-assisted analysis?

Q2(R2) covers analytical procedures and states the fit-for-purpose principle. It was not written for AI tools, but its characteristics (specificity, accuracy, repeatability, intermediate precision, range) supply a vocabulary for pilot design, and a quality unit should decide how far to extend it (www.fda.gov).

What does GxP validation of generative AI in analytical labs require?

No finalized AI-specific GMP rule was located in the sources reviewed. The practical anchors are risk-based computerized system validation (health.ec.europa.eu), data integrity principles (www.fda.gov), change control, and documented human oversight.

How does 21 CFR Part 11 relate to AI review of analytical data?

Part 11 calls for system validation, audit trails, and signed records that show signer, time, and meaning (www.ecfr.gov). FDA interprets the scope narrowly and endorses risk-based validation. A model session producing records used in regulated decisions falls under the same expectations as other software. No Part 11 conformance statement for Claude products was found.

What other AI approaches exist for structure elucidation from NMR and mass spectrometry?

Simulation-trained transformers, LLM agents with tool environments, and established CASE systems all exist (see Table 3 and the inverse-elucidation section). Results are not directly comparable because datasets and metrics differ, and one agent paper lists repeated-run robustness as an open limitation (arxiv.org).

Is Claude a replacement for ChemDraw, MestReNova, or an analyst?

No. Anthropic states its goal as ensuring chemists know where Claude can save them time (www.anthropic.com), and the analyst remains the reviewer of record.

Conclusion

"Claude for NMR and LC-MS" is not one claim. Forward prediction has a 20-compound, three-run benchmark with favorable mean errors. Inverse elucidation has a 15-problem benchmark that works best on simpler targets or with a starting-material hint. Raw NMR processing and LC-MS processing each have one documented sample, with timings that compare favorably to calendar lag but not to measured analyst effort, and the purity pair compares two differently defined numbers.

For a laboratory, the implication is a staged pilot with four independent decisions. Each task needs its own intended-use statement, blinded reference set, replicate design, metrics, and stop conditions. Model identifiers should be pinned and changes treated as change events, because sampling parameters (temperature, top_p, top_k) are rejected on API version 2026-10-21 and, on earlier versions, when set to non-default values on Claude 4.7 and later models, so the API version in use should be verified. Raw files should be hashed, parses reconciled, purity bases defined, and results reviewed by a qualified person against instrument-software output. In GMP contexts, the draft Annex 22 currently excludes language models, although the drafting group is considering changing that position, so critical uses should wait for a documented regulatory position until guidance is finalized.

Nothing in the public record shows that an AI result establishes identity, purity, release, or GMP suitability. What the 2026 results show is that a general-purpose model can be a credible assistant on routine NMR and LC-MS data, and that the open question is how much of that holds on a laboratory's own compounds, solvents, instruments, and review process. The protocol above answers it with measurements. Questions about regulatory status should go to qualified quality and regulatory professionals.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://www-cdn.anthropic.com/07441e654ad3dfef0cd090e9361511562825d012.pdf>
2. <https://www.anthropic.com/research/making-claude-a-chemist>
3. <https://www-cdn.anthropic.com/9f08da5189ac269b3242ca760de9823805c3f5f6.pdf>
4. <https://www.anthropic.com/research/Claude-accelerates-protein-design>
5. https://database.ich.org/sites/default/files/ICH_Q2%28R2%29_Guideline_2023_1130.pdf
6. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11>
7. <https://www.ema.europa.eu/en/events/good-manufacturing-practice-multistakeholder-workshop-expert-contributions-artificial-intelligence-guidance-development-annex-22>
8. <https://docs.claude.com/en/docs/about-claude/models/overview>
9. <https://docs.claude.com/en/docs/about-claude/model-deprecations>
10. <https://docs.claude.com/en/docs/about-claude/models/model-ids-and-versions>
11. <https://docs.claude.com/en/docs/models/opus-5-5/migration-guide>
12. <https://docs.claude.com/docs/en/api/migration-2026-10-21>
13. <https://arxiv.org/abs/2408.04667>
14. <https://arxiv.org/abs/2601.19934v1>

15. <https://www.anthropic.com/news/claude-fable-5-mythos-5>
16. <https://www.anthropic.com/news/life-sciences-verification-program>
17. <https://www.ema.europa.eu/en/ich-q2r2-validation-analytical-procedures-scientific-guideline>
18. <https://www.fda.gov/media/75414/download>
19. https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf
20. https://www.ema.europa.eu/en/documents/report/report-multistakeholder-workshop-expert-contributions-artificial-intelligence-guidance-development-annex-22_en.pdf
21. <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development>
22. <https://www.fda.gov/media/184830/download>
23. https://database.ich.org/sites/default/files/ICH_Q14_Guideline_2023_1116.pdf
24. <https://www.fda.gov/media/119267/download>
25. https://assets.publishing.service.gov.uk/media/5aa2b9ede5274a3e391e37f3/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf
26. <https://arxiv.org/html/2607.19406>
27. <https://www.anthropic.com/news/claude-science-ai-workbench>
28. <https://claude.com/product/claude-science>
29. <https://docs.claude.com/en/docs/agents-and-tools/tool-use/code-execution-tool>
30. <https://docs.claude.com/en/docs/build-with-claude/files>
31. <https://support.claude.com/en/articles/10015870-what-certifications-has-anthropic-obtained>
32. <https://arxiv.org/html/2512.03127v1>
33. <https://arxiv.org/abs/2512.18531>
34. <https://arxiv.org/html/2512.18531v2>
35. <https://arxiv.org/abs/2607.19406>
36. https://nuzillard.github.io/LSD/index_ENG.html
37. <https://csrc.nist.gov/files/pubs/fips/180-4/final/docs/fips180-4.pdf>
38. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-C/part-211/subpart-J/section-211.194>
39. <https://www.chromatographyonline.com/view/mass-spectrometry-primer-part-iii>
40. <https://www.thermofisher.com/blog/analyteguru/when-resolution-changes-the-story-single-quadrupole-vs-hram-gc-ms>
41. <https://proteowizard.sourceforge.io/tools/msconvert.html>
42. https://proteowizard.sourceforge.io/doc_users.html
43. <https://nmrglue.readthedocs.io/en/latest/tutorial.html>
44. <https://www.chromatographyonline.com/view/peak-purity-liquid-chromatography-part-i-basic-concepts-commercial-software-and-limitations>
45. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4255678>
46. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4255677>
47. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
48. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12226332>

49. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-C/part-211/subpart-D/section-211.68>
50. https://health.ec.europa.eu/system/files/2016-11/annex11_01-2011_en_0.pdf
51. <https://ispe.org/publications/guidance-documents/gamp-guide-artificial-intelligence>
52. <https://www.fda.gov/media/188844/download>
53. <https://mestrelab.com/software/mnova/verify>
54. <https://mestrelab.com/main-product/structure-elucidation>
55. <https://mestrelab.com/main-product/qnmr>
56. <https://support.mestrelab.com/kb/article/419-nmr-supported-file-formats%E2%80%8B>
57. <https://mestrelab.com/main-product/mschrom>
58. <https://nmrium-docs.pages.dev>
59. https://nmrium-docs.pages.dev/20_1d-spectra/loading_a_spectrum
60. <https://revvitysignals.com/products/spectrus-analytical-chemistry/nmr-predictors>
61. <https://revvitysignals.com/products/spectrus-analytical-chemistry/structure-elucidator-suite>
62. <https://bruker.com/de/products-and-solutions/mr/nmr-software/cmc-se.html>
63. <https://www.bruker.com/en/products-and-solutions/mr/nmr-software/cmc-assist.html>
64. <https://www.bruker.com/en/landingpages/bbio/artificial-intelligence-in-nmr.html>
65. <https://www.bruker.com/en/products-and-solutions/mr/nmr-software/nmrtist.html>
66. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4089557>
67. <https://pubs.rsc.org/en/content/articlehtml/2020/sc/d0sc00442a>
68. <https://nmrshiftdb.nmr.uni-koeln.de>
69. <https://pubmed.ncbi.nlm.nih.gov/34667567>
70. <https://www.nature.com/articles/s41592-022-01486-3>
71. <https://arxiv.org/html/2502.09571v2>
72. <https://arxiv.org/html/2601.00941v1>
73. <https://arxiv.org/abs/2602.10158>
74. <https://nmrlab.technion.ac.il/files/2017/10/proc.pdf>
75. https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf
76. <https://mhrainspectorate.blog.gov.uk/2026/06/29/use-of-ai-for-gxp-inspection-responses-setting-standards-without-stifling-innovation>
77. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9262714>
78. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q2r2-validation-analytical-procedures>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.