Counterfactual NMR: Benchmarking Minimal Spectral Interventions for Interpretable Structure Elucidation

Anonymous Author(s)
Affiliation
Address
email

Abstract

NMR-based structure elucidation models are often accurate yet opaque: they rarely indicate *which specific measurements* drove a call or *what to acquire next*. We introduce *Counterfactual NMR*, a causal audit that asks: *What minimal, chemically plausible edit to the spectrum would flip this prediction?* Edits are single-peak interventions constrained to expert ppm windows for ^{1}H/^{13}C (e.g., carbonyl ^{13}C 190–220 ppm); a fast search selects the smallest change that maximally increases a target probability. Effects are quantified by per-sample treatment $\hat{\tau}$, cohort Average Treatment Effect (ATE) with 95% CIs, and *lift* versus a window-matched randomized baseline to separate specificity from generic sensitivity; mechanistic ablations (^{1}H-only/^{13}C-only/both) test alignment with textbook chemistry. On near-boundary cohorts ($0.35<p<0.65$), minimal ^{13}C interventions produce large, precise shifts—e.g., *ketone* ATE 0.336 (95% CI $[0.315, 0.357]$, flip 0.684) and *alcohol* ATE 0.272 (95% CI $[0.261, 0.283]$, flip 0.800)—with targeted effects 2–4$\times$ stronger than random edits under the same constraints. Ablations confirm chemistry (carbonyl/ketone are ^{13}C-driven; alcohol shows balanced ^{1}H/^{13}C+synergy); edits remain sparse and realistic. Counterfactual NMR turns interpretability into *actionable recourse*, enabling trustworthy auditing, targeted data curation, and principled next-experiment selection in functional group prediction workflows.

1 Introduction

Automated and semi-automated *structure elucidation* (ASE/CASE) systems map NMR spectra to molecular structures or substructures and are increasingly capable in routine settings. Yet practitioners face three persistent pain points: (i) models are *opaque* about *which specific measurements* drive a call; (ii) failures under *distribution shift* (solvent/field changes, impurities, unusual motifs) are hard to diagnose; and (iii) experiment planning remains *heuristic*—chemists still ask, *"Which additional measurement would most change or confirm this prediction?"* Even with widely adopted 1D/2D experiments (^{1}H, ^{13}C, HSQC, HMBC) central to connectivity inference, current ML explanations largely remain correlational and non-actionable [4, 3, 2].

Most popular interpretability tools (feature importances, saliency, SHAP) summarize correlations but do not answer the experimentalist's question: *What minimal, chemically plausible change to the spectrum would flip this decision?* They are not linked to an *intervention* a spectroscopist could realize by acquiring a different experiment or by recognizing/adding a peak in a known ppm window (e.g., carbonyl ^{13}C $\sim$190–220 ppm; aromatic ^{13}C $\sim$110–150 ppm; aromatic ^{1}H $\sim$6–8.5 ppm), limiting trust, slowing root-cause analysis under shift, and offering little guidance for principled *next-experiment* selection [6, 1].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

We introduce *Counterfactual NMR*, a causal audit and benchmark that makes NMR interpretability *actionable* for ASE. We define chemistry-constrained spectral edit operators—add/shift/attenuate peaks strictly within expert ppm windows for ^{1}H/^{13}C—so counterfactuals are both *minimal* and *plausible*. We pair these operators with a fast search for one-edit counterfactuals and report practice-oriented causal estimands (per-sample effect, cohort ATE with CIs, and *lift* vs. randomized edits) that separate *specificity* from generic sensitivity. We add *mechanistic alignment* checks via channel ablations (^{1}H-only vs. ^{13}C-only vs. both) anchored to textbook shift regions (e.g., carbonyl ^{13}C $\sim$190–220 ppm), making the inferred "causal evidence" falsifiable against chemistry.

On open NMR data in the style of NMRShiftDB2 [5], single-peak, chemically valid ^{13}C interventions induce large, reliable probability shifts on near-boundary cases for carbonyl-bearing classes and meaningful shifts for aromatics, indicating that models move with chemically correct evidence rather than spurious cues. Because edits are minimal and windows enforced, effects are interpretable as *decision-relevant evidence*, not artifacts of rescaling.

Contributions. (1) *Counterfactual, chemistry-constrained spectral interventions* for NMR with efficient minimal-edit search. (2) *Causal estimands* (ATE with CIs; *lift* vs. randomized edits) that quantify decision-relevant effect sizes and isolate specificity. (3) *Mechanistic alignment* via ^{1}H/^{13}C ablations tied to established regions. (4) A *benchmark and code* for drop-in auditing within ASE, supporting trustworthy deployment, targeted curation, and principled *next-experiment* choices.

2 Method

Setup and interventional semantics. Let Z denote the latent molecular structure, X the observed spectrum (peak list or binned vector over ^{1}H/^{13}C), and $\hat{Y} = h(X)$ a learned predictor (substructure logits or ASE outputs). We assume $X = f(Z) + \epsilon$ for NMR physics f and noise ϵ, and probe h by applying explicit, chemistry-constrained interventions to X. An edited spectrum is $x' = x \oplus \Delta$, where Δ is a domain-valid spectral edit (defined below). For a target label t, let $p_t(x) = P(\hat{Y}_t = 1 \mid X = x)$. The per-sample counterfactual effect is $\hat{\tau}_i^t = p_t(x_i') - p_t(x_i)$. Over a cohort D, the (C)ATE is $\widehat{\mathrm{ATE}}^t = \frac{1}{|D|}\sum_{i\in D}\hat{\tau}_i^t$. To assess specificity (i.e., causal signal beyond generic sensitivity), we compute a *paired* randomized baseline using the same windows and edit budget but random locations and report *Lift* $= \widehat{\mathrm{ATE}}_{\text{targeted}} - \widehat{\mathrm{ATE}}_{\text{random}}$. When D contains near-boundary cases ($0.35 < p_t(x) < 0.65$), the estimand is a CATE.

Chemistry-constrained edit operator. We use a single actionable operator that adds one peak at ppm δ with bounded normalized amplitude a inside expert windows $\mathcal{W}_t$ for the target: $x' = x \oplus \Delta(\delta, a)$ with $\delta \in \mathcal{W}_t$ and $a \in [0, a_{\max}]$. Examples include *ketone* ^{13}C 190–220 ppm; *aromatic* ^{13}C 110–150 ppm and ^{1}H 6.0–8.5 ppm; *aldehyde* ^{1}H 9.0–10.5 ppm. We set $a_{\max} = 0.8$ on the max-normalized scale. This "add-one-peak" operator maximizes interpretability and identifiability; shift and attenuate variants are retained as extensions.

Minimal-edit search. For target t and spectrum x, candidates are evaluated on a fixed ppm grid within $\mathcal{W}_t$ and we select

$$\Delta^\star = \arg\max_{\Delta\in\mathcal{C}_t} \Big(p_t(x\oplus\Delta) - p_t(x)\Big) - \lambda\|x - (x\oplus\Delta)\|_1 - \gamma\,\mathbf{1}\{\#\text{edits} > 1\},$$

with $\lambda = 0.1$ and a large γ enforcing `max_edits=1`. We use a greedy selection (beam-k optional), and log proximity (ℓ_1 change), sparsity (edit count), and rule-consistency diagnostics.

Cohorts, randomized baseline, and inference. We evaluate on near-boundary cases ($0.35 < p_t(x) < 0.65$), where recourse is most decision-relevant. For each case, we generate K random edits (same windows, amplitude, and edit budget) to obtain a paired randomized effect on the exact same spectrum. We report ATE with 95% CIs (nonparametric bootstrap or t-interval when appropriate) and Lift. Significance for ATE lift uses a paired t-test or a paired permutation test (recommended for small n). Flip-rate lift (at threshold 0.5) uses McNemar's test on paired decisions (targeted vs. random).

Data, representation, and predictor. We use an open NMR set with SMILES and ^{1}H/^{13}C peak lists (1849 molecules in our split). Substructure labels are derived

Table 1: Counterfactual effects on near-boundary cohorts. Edit budget: one peak, domain windows; proximity penalized. Lift compares targeted vs randomized edits within the same windows/budget.

Label	n	ATE ↑	95% CI	Flip ↑	ATE_{rand}	$\text{Flip}_{\text{rand}}$	ATE Lift (p)	Flip Lift (p)
Ketone	19	0.336	[0.315, 0.357]	0.684	0.018	0.105	+0.318 ($< 10^{-6}$)	+0.579 (3×10^{-4}
Aromatic	11	0.382	[0.341, 0.423]	0.273	0.085	0.091	+0.298 (3.5×10^{-9})	+0.182 (0.269
Methyl	45	0.351	[0.337, 0.365]	0.467	0.117	0.244	+0.234 (1.5×10^{-20})	+0.222 (0.0277
Carbonyl	50	0.414	[0.398, 0.431]	0.500	0.057	0.140	+0.357 (1.7×10^{-37})	+0.360 (1.1×10^{-4}
Ester	33	0.315	[0.300, 0.330]	0.576	0.097	0.273	+0.219 (1.4×10^{-17})	+0.303 (0.0128
Amide	17	0.309	[0.289, 0.330]	0.647	0.101	0.294	+0.208 (1.0×10^{-8})	+0.353 (0.0393
Methoxy	32	0.442	[0.413, 0.472]	0.375	0.138	0.250	+0.304 (7.8×10^{-14})	+0.125 (0.281
Halogen	50	0.041	[0.032, 0.050]	0.240	0.001	0.080	+0.040 (5.7×10^{-9})	+0.160 (0.0291
Alcohol	20	0.272	[0.261, 0.283]	0.800	0.069	0.300	+0.203 (8.7×10^{-13})	+0.500 (0.0015

Proximity (avg): ketone 0.80; aromatic 0.76; methyl 0.80; carbonyl 0.80; ester 0.80; amide 0.80; methoxy 0.75; halogen 0.51; alcohol 0.80
Rule-consistency, intensity realism, mutual exclusivity: 1.00 for all labels.

Table 2: Channel ablations (Δp mean per case). Carbonyl is driven entirely by ^{13}C windowing; ester/amide/methoxy also show strong ^{13}C contributions with mild ^{1}H synergy.

Label	^{1}H-only Δp	^{13}C-only Δp	Both Δp
Ketone	0.157	0.345	0.421
Aromatic	0.131	0.351	0.307
Methyl	0.098	0.337	0.372
Carbonyl	0.000	0.448	0.448
Ester	0.089	0.295	0.343
Amide	0.072	0.305	0.346
Methoxy	0.171	0.481	0.499
Halogen	0.064	0.064	0.126
Alcohol	0.162	0.274	0.390

For carbonyl, no viable ^{1}H candidates were present; "Both" equals the ^{13}C effect (union fallback).

from SMARTS for `aromatic`, `carbonyl`, `aldehyde`, `ketone`, `ester`, `amide`, `alkene`, `alkyne`, `methoxy`, `halogen`, `nitro`, `alcohol`, `amine`, `methyl`. Spectra are binned as ^{1}H 0–12 ppm at 0.02 ppm (600 bins) and ^{13}C 0–220 ppm at 1.0 ppm (220 bins), concatenated and max-normalized per sample (dim=820), with peaks assigned to nearest bins. Unless noted, h is one-vs-rest logistic regression with Platt calibration (80/20 split; fixed seed).

Mechanistic ablation. To test channel roles, we repeat the search under ^{1}H-only, ^{13}C-only, and *both*. If a channel has no candidates, "both" falls back to the other (union fallback). We summarize mean Δp per channel, a dominance ratio (^{13}C:^{1}H), and a synergy score defined as (both$-\max\{$H,C$\}$).

Defaults and reproducibility. Unless specified: $a = 0.8$, $\lambda = 0.1$, `max_edits=1`, $K = 5$ random baselines per case, ^{1}H grid 0.1–0.2 ppm, ^{13}C grid 0.5–1.0 ppm, calibrated probabilities enabled. All reported metrics are averaged over the near-boundary cohort with CIs and paired tests as above.

3 Discussion

We report in Table 1 the *model-level causal effects* of *minimal, chemistry-constrained* spectral edits on *near-boundary* cases ($0.35 < p_t(x) < 0.65$). The ATE is the average change in calibrated probability p_t after a one-peak intervention within expert ppm windows; larger ATE means the model's decision is more sensitive to the targeted (chemically valid) evidence. *Lift* subtracts the effect of a matched *randomized* edit (same windows/budget) on the *same* spectrum, isolating *specificity* from generic sensitivity; p-values test whether this targeted-vs-random difference is nonzero. *Flip* is the fraction of near-boundary cases that cross the 0.5 decision threshold under the targeted edit; the corresponding

lift compares to random edits. The footnotes summarize *proximity* (smaller perturbations indicate more minimal edits) and feasibility checks (rule-consistency, intensity realism, mutual exclusivity). Table 2 ablates channels by re-running the search under ^{1}H-only, ^{13}C-only, and their combination. Here Δp is the mean per-case probability gain; "Both" uses a union fallback when one channel has no viable candidates.

Main findings (Table 1). Targeted one-peak edits yield large, precise probability shifts on ambiguous spectra for chemically anchored labels. *Carbonyl* shows the strongest effect (ATE 0.414, lift $+0.357$, $p \ll 10^{-10}$) and high flip lift ($+0.360$, $p = 1.1\times10^{-4}$), indicating that adding a carbonyl-region ^{13}C signal is both highly *effective* and *specific* in moving the model's decision. *Ketone* behaves similarly (ATE 0.336, lift $+0.318$, $p \approx 0$; flip lift $+0.579$, $p = 3\times10^{-4}$), consistent with carbonyl-driven evidence. *Ester*, *amide*, and *alcohol* also exhibit substantial targeted effects and significant flip lifts, reflecting clear, actionable model sensitivity to their characteristic windows. Notably, *aromatic* achieves a large ATE (0.382) despite near-ceiling baseline AP; however, its flip lift is not significant (small n and high baseline confidence), which is consistent with strong probability movements that do not always cross the 0.5 threshold. At the other extreme, *halogen* shows a small but statistically specific effect (ATE 0.041, lift $+0.040$), and a modest flip lift, suggesting that substantially larger or additional evidence is needed to decisively alter halogen calls.

Minimality and plausibility. Across labels the edits remain sparse and small (one-peak budget; proximity around ~ 0.8 for most labels), and all feasibility checks pass (1.0 for rule-consistency, intensity realism, and mutual exclusivity). The lower proximity reported for *halogen* indicates comparatively larger perturbations were needed to achieve any movement, matching its small ATE; by contrast, *carbonyl*/*ketone*/*amide* achieve large, specific effects with small edits, evidencing tight alignment between chemical windows and model reasoning.

Mechanistic interpretation (Table 2). Channel ablations quantify *which* modality carries decision-relevant evidence. For *carbonyl* and *ketone*, ^{13}C alone accounts for essentially all of the effect (e.g., *carbonyl* ^{13}C-only $\Delta p = 0.448$; ^{1}H-only has no viable candidates), matching textbook chemical shifts for carbonyl carbons. *Amide* and *ester* show the same ^{13}C predominance with mild ^{1}H synergy (Both $> \max\{\text{H,C}\}$), while *alcohol* exhibits a more balanced contribution: ^{1}H contributes meaningfully (broad 1–5 ppm signatures) and combining channels increases the effect further (Both $0.390 > \max\{0.162, 0.274\}$). For *aromatic*, ^{13}C-only exceeds ^{1}H-only (specificity of sp^2 carbons), and "Both" is slightly smaller than ^{13}C-only, consistent with partial redundancy between channels under a one-peak budget. These patterns strengthen the claim that observed effects arise from *chemically correct* regions, not spurious shortcuts.

Implications for ASE and experiment planning. Because the reported effects are computed on near-boundary spectra, they function as a *triage score* for what evidence is most likely to change an uncertain call. The large ATE and flip lifts for carbonyl-bearing hypotheses recommend prioritizing a quick ^{13}C acquisition in the 190–220 ppm window for ketone/aldehyde differentials; the balanced but synergistic *alcohol* result suggests value in acquiring *both* channels when feasible. More broadly, the targeted-vs-random lift provides a principled check before investing time in additional experiments: if lift is small or non-significant (e.g., halogen), further edits or 2D connectivity data (HSQC/HMBC) may be required to affect the decision.

4 Conclusion

Our findings are *model-level* (algorithmic recourse) rather than claims about the physical $Z \to X$ mechanism: they reveal which measurements *cause the model* to change its output. Estimates are conditioned on the edit family (one-peak additions within expert ppm windows) and on near-boundary cohorts, i.e., a CATE rather than a global ATE. The binned representation omits fine multiplet structure and J-coupling; extending the operators and adding 2D spectra (e.g., HSQC/HMBC) is a natural next step. Small-n cohorts can yield wider flip-rate intervals even when ATEs are precise, so we emphasize paired targeted-vs-random *lift* and confidence intervals to guide interpretation. Overall, minimal, chemistry-constrained counterfactual edits move model probabilities in the *right* spectral regions, yielding large, statistically specific effects for carbonyl-derived labels and interpretable ^{1}H/^{13}C roles across the board.

Defaults and reproducibility. Unless specified: a=0.8, λ=0.1, `max_edits`= 1, K=5 randomized edits per case, ^{1}H grid step 0.1–0.2 ppm, ^{13}C step 0.5–1.0 ppm, 80/20 split with fixed seed, calibrated probabilities. All reported metrics are averaged over the near-boundary cohort with CIs and paired tests as above.

Appendix: Additional Details

Operator variants and realism. Shift and attenuate operators are available but disabled by default to preserve identifiability of causes; when enabled, shifts are capped to small $\Delta\delta$ within $\mathcal{W}_t$, and attenuations are bounded so as not to create chemically impossible constellations. Mutual-exclusivity checks prevent contradictory edits, and intensity draws can be sampled from empirical distributions learned from experimental libraries.

Window tables and grids. We maintain per-label windows for ^{1}H and ^{13}C (e.g., carbonyl C: 190–220 ppm; amide C: 165–180 ppm; aromatic C: 110–150 ppm; aldehyde H: 9.0–10.5 ppm; methyl C: 10–40 ppm; aromatic H: 6.0–8.5 ppm). Candidate grids default to 0.1–0.2 ppm for ^{1}H and 0.5–1.0 ppm for ^{13}C.

Beam search and complexity. Beam-k search (typically $k \in \{3, 5\}$) lowers myopic failures and is linear in k times the number of candidate ppm bins. Vectorized scoring batches all candidates per case for efficient inference.

Statistics and testing. We compute CIs via nonparametric bootstrap (1,000 resamples) unless n is large, and use paired t-tests or paired permutation tests for ATE lift. Flip-rate lift uses McNemar's test with continuity correction. Reported p-values are two-sided.

Implementation notes. All spectra are max-normalized per sample before edits; proximity uses ℓ_1 on the concatenated vector. We fix random seeds for splits and baselines. If one channel lacks candidates, the "both" condition defaults to the available channel to avoid spurious zero effects. Optional multi-edit mode (up to three edits) is supported with an additional penalty term and consistency checks.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Ad Bax and Michael F. Summers. Proton and carbon-13 assignments by two-dimensional heteronuclear multiple-bond correlation. *Journal of the American Chemical Society*, 108(8): 2093–2094, 1986. doi: 10.1021/ja00268a060.

[3] Geoffrey Bodenhausen and David J. Ruben. Natural abundance nitrogen-15 nmr by enhanced heteronuclear spectroscopy. *Journal of Chemical Physics*, 72(10):4472–4473, 1980. doi: 10.1063/1.439679.

[4] Timothy D. W. Claridge. *High-Resolution NMR Techniques in Organic Chemistry*. Tetrahedron Organic Chemistry Series. Elsevier, 3 edition, 2016. ISBN 978-0-08-099986-9.

[5] Stefan Kuhn. Twenty years of nmrshiftdb2: A case study of an open nmr database. *Magnetic Resonance in Chemistry*, 2024. doi: 10.1002/mrc.5378. In press / early view at time of writing.

[6] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of NeurIPS*, 2017.