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, which infer molecular substructures or full structures directly from spectra by linking peak patterns to chemical environments [17, 3], are becoming increasingly capable in routine settings [8, 7]. 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 [6, 5, 4].

Most popular interpretability tools (feature importances, saliency, SHAP) summarize correlations rather than testing how evidence *causes* model decisions to change. They are not linked to an *intervention* a spectroscopist could perform—such as adding or recognizing a peak in a known ppm window (e.g., carbonyl ^{13}C 190–220,ppm)—limiting trust and experimental guidance [12, 2]. We instead build on counterfactual and recourse methods [20, 18, 10] to perform *explicit, chemistry-valid edits* that probe the model's internal causal reasoning rather than physical molecular causality, revealing which spectral measurements *cause the model* to change its prediction, not which peaks physically arise from structure [15, 14].

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*, in the sense used by causal-recourse methods [20, 18, 10] and domain-constrained scientific counterfactuals [9, 15]. In spectroscopy, restricting edits to expert ppm windows grounded in empirical NMR databases and textbook chemical shift theory [16, 13, 6, 11] preserves physical feasibility and ensures counterfactuals correspond to experimentally realizable spectral variations. We pair these operators with a fast search for one-edit counterfactuals and report practice-oriented causal estimands (per-sample effect, cohort ATE with Confidence Intervals (CI) 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 [11], 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.

Intuition and practical use. Consider a near-boundary ketone prediction where the model assigns $p_{\text{ketone}} = 0.41$. Counterfactual NMR identifies that adding a single ^{13}C peak near 200 ppm (the canonical carbonyl region) would raise the probability to 0.80 ($\Delta p = +0.39$). For a spectroscopist, this is a concrete next step: check whether a weak carbonyl resonance exists in that range or acquire a rapid ^{13}C scan focused on 190–220 ppm. If the signal appears, the model's reasoning is chemically sound—carbonyl presence supports the ketone hypothesis; if not, the missing evidence explains its uncertainty. Each counterfactual edit thus maps directly to a measurable or acquirable piece of spectral evidence, turning interpretability into targeted experimental guidance.

Contributions. We introduce (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 [19], supporting trustworthy deployment, targeted curation, and principled *next-experiment* choices.

2 Method

Setup and interventional semantics. Let X denote 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 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 in Appendix A). 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 $\widehat{\text{ATE}}^t = \frac{1}{|D|}\sum_{i\in D}\hat{\tau}_i^t$. 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* $= \widehat{\text{ATE}}_{\text{targeted}} - \widehat{\text{ATE}}_{\text{random}}$. 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, ^{1}H 2.0–2.8 ppm; *aromatic* ^{13}C 110–150 ppm, ^{1}H 6.0–8.5 ppm; *methyl* ^{13}C 10–25 ppm, ^{1}H 0.6–1.3 ppm; *carbonyl* ^{13}C 160–220 ppm; *ester* ^{13}C 160–180 ppm, ^{1}H 4.0–4.5 ppm; *amide* ^{13}C 165–180 ppm, ^{1}H 2.5–3.5 ppm; *methoxy* ^{13}C 55–60 ppm, ^{1}H 3.3–3.8 ppm; *alcohol* ^{13}C 60–75 ppm, ^{1}H 1.0–5.0 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.

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): aromatic 0.76; methoxy 0.75; halogen 0.51; all others 0.80.
Rule-consistency, intensity realism, mutual exclusivity: 1.00 for all labels.

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. The candidate set $\mathcal{C}_t$ and operator semantics are defined formally in Appendix A. 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 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 ($0.35 < p_t(x) < 0.65$) of n spectra. 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

Table 2: Channel ablations (Δp mean per case). Carbonyl is driven entirely by ^{13}C windowing (as no viable ^{1}H candidates were present; "Both" equals the ^{13}C effect (union fallback)); 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

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. To interpret which spectral channels drive these effects, we refer to the channel ablations in Table 2, which decompose Δp by ^{1}H and ^{13}C contributions. *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.

Counterfactuals vs. coefficient analysis for linear models. Since h is a logistic regression model, one might ask what counterfactual edits offer beyond inspecting coefficients. Coefficients reflect the *global, correlational importance* of spectral bins averaged across the dataset, whereas counterfactuals probe *local, causal sensitivity* for a specific spectrum under domain-valid perturbations. This distinction yields several advantages. First, *instance-specificity*: coefficients summarize overall trends, while counterfactuals answer "what minimal change would flip this decision?"—crucial for near-boundary cases. Second, *interventional semantics*: coefficients describe associations, but counterfactuals test explicit interventions (e.g., "add a ^{13}C peak at 200 ppm") that correspond to realizable chemist actions. Third, *chemical plausibility*: edits are restricted to expert windows ($\mathcal{W}_t$), ensuring physically valid perturbations rather than arbitrary high-weight bins. Fourth, *minimality and effect size*: the edit search jointly minimizes perturbation magnitude and reports concrete probability shifts (Δp), yielding interpretable effect estimates rather than abstract weights. Finally, *specificity and planning*: comparing targeted to randomized edits isolates true causal signal from generic sensitivity and suggests actionable next experiments (e.g., acquire ^{13}C data in 190–220 ppm for carbonyl

hypotheses). Thus, even for linear models, counterfactual edits convert static global coefficients into *instance-level, experimentally actionable reasoning*.

Mechanistic interpretation (Table 2). Following the aggregate effects in Table 1, Table 2 decomposes these responses by spectral channel, quantifying which modality carries the 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{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. This occurs because the search is constrained to a single edit: when both channels are available, candidate ^{1}H peaks compete with stronger ^{13}C features for the same edit slot, occasionally displacing the optimal ^{13}C intervention and thereby reducing the average Δp.

Evidence for chemically correct reasoning. The claim that models rely on *chemically correct regions* rather than *spurious shortcuts* is supported by multiple lines of evidence. First, *mechanistic alignment*: edits in ^{13}C windows drive model changes for carbonyl-bearing groups (Table 2), exactly as predicted by physical chemistry—carbonyl carbons are strongly deshielded ($\sim$190–220 ppm) due to π-electron withdrawal [1, 16]. Conversely, for labels without strong ^{13}C signatures (e.g., *halogen*), both ATE and lift are small, suggesting the model correctly does *not* rely on weak or absent signals. Second, *specificity*: targeted edits within expert windows produce effects 2–4$\times$ stronger than random edits in the *same* windows (Table 1, Lift column), indicating the model recognizes *which locations* within a window are most informative, not just that the window is broadly relevant. Third, *actionability*: edits correspond to realizable chemist actions (acquiring ^{13}C focused on 190–220 ppm for carbonyl hypotheses), and practitioners can verify whether signals appear where counterfactuals suggest. Fourth, *consistency with empirical databases*: edit locations cluster within expert windows (e.g., ketone ^{13}C edits concentrate in 190–220 ppm) as would be expected from measured spectra in NMRShiftDB2 [11].

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.

Appendix: Additional Details

A Edit Operators and Objective

Edit operator (definition). Let $x \in \mathbb{R}^{820}$ be the concatenated binned spectrum (^{1}H then ^{13}C) after max-normalization; let b_H, b_C map a ppm value to its bin index for ^{1}H and ^{13}C, respectively. For channel $c \in \{H, C\}$, ppm location δ, and amplitude $a \in [0, a_{\max}]$, define the single-peak edit vector

$$\Delta(\delta, a, c)_j \;=\; \begin{cases} a & \text{if } j = b_c(\delta), \\ 0 & \text{otherwise,} \end{cases}$$

(optionally convolved with a 1–3 bin triangular kernel; ablations unchanged). The edit operator is elementwise addition with clipping to the valid range:

$$x' \;=\; x \oplus \Delta(\delta, a, c) \;\; \min\{\,1,\; x + \Delta(\delta, a, c)\,\}.$$

In causal terms, this corresponds to an intervention $\mathrm{do}(X \leftarrow X \oplus \Delta)$ on the model input, and we estimate the induced change in predicted outcome $p_t(x)$. The candidate set for label t is

$$\mathcal{C}_t \;=\; \big\{\Delta(\delta, a, c) \;:\; \delta \in \mathcal{W}_t^{(c)},\; a \in [0, a_{\max}],\; c \in \{H, C\}\big\},$$

where $\mathcal{W}_t^{(c)}$ are expert ppm windows for t (specified below and used in Table 1).

Minimal intervention and objective (capsule). We use a single-bin addition at ppm $\delta \in \mathcal{W}_t$ with amplitude a=0.8 and `max_edits`= 1. For each candidate $\Delta(\delta, a)$ we maximize the score

$$S(\Delta) = \big[p_t(x \oplus \Delta) - p_t(x)\big] - \lambda \, \|x' - x\|_1,$$

and select the best location on a ppm grid; λ=0.1 unless noted. For each spectrum we also draw K=5 randomized edits with the *same* windows and amplitude to form a *paired* baseline.

Edit semantics. Let $b(\delta)$ be the bin index for ppm location δ. The intervention applies

$$x'_{b(\delta)} \leftarrow \min\{1,\; x_{b(\delta)} + a\}, \qquad x'_j \leftarrow x_j \text{ for } j \neq b(\delta),$$

with $a \in \{0.4, 0.6, 0.8\}$ in sensitivity analyses (default a=0.8). We never overwrite existing signal; additions saturate at 1.0. If a candidate bin already exceeds $1-a$, we skip it to preserve intensity realism. An optional 3-bin triangular *line-shape* kernel can mimic mild broadening:

$$x'_{b(\delta)+k} \leftarrow \min\Big\{1,\; x_{b(\delta)+k} + a\, w_k\Big\}, \quad k \in \{-1, 0, 1\}, \quad \mathbf{w} = [0.25,\, 0.5,\, 0.25].$$

Search objective. The search objective balances two competing goals: maximizing the probability increase

$$p_t(x \oplus \Delta) - p_t(x)$$

while minimizing the spectral perturbation

$$\lambda \, \|x - (x \oplus \Delta)\|_1.$$

The effectiveness term measures how much the edit flips the prediction, while the L_1 proximity penalty quantifies the total absolute change across all spectral bins. We use L_1 distance because it directly counts "how many bins changed and by how much," making it interpretable for chemists who prefer targeted, minimal interventions. The parameter $\lambda = 0.1$ creates a linear trade-off: we accept a 0.1 increase in proximity for every 1.0 increase in effectiveness. This formulation favors edits that achieve large probability shifts with minimal spectral changes, ensuring counterfactuals are both actionable and chemically plausible.

Proximity. We report $\mathrm{prox}(x, x') = \|x' - x\|_1$ on the max-normalized, concatenated vector; *smaller* values indicate more minimal edits. For completeness we also compute a normalized variant $\mathrm{nprox} = \|x' - x\|_1 / (\|x\|_1 + \epsilon)$ (not used in the main tables).

B Randomized Baseline and Inference

Paired randomized baseline. For each spectrum, we sample $K{=}5$ randomized edits using the *same* window set $\mathcal{W}_t$, amplitude a, grid, and proximity penalty as the targeted search. Effects are paired within-spectrum:

$$\text{Lift} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\tau}_i^{\text{tgt}} - \hat{\tau}_i^{\text{rnd}}\big), \quad \hat{\tau}_i = p_t(x_i') - p_t(x_i).$$

Uncertainty and tests. We compute 95% CIs via nonparametric bootstrap (1,000 resamples) unless otherwise stated and report two-sided p-values. ATE lift uses a *paired* t-test (paired permutation test recommended for small n); flip-rate lift (at calibrated threshold 0.5) uses McNemar's test with continuity correction. Multi-label results are exploratory; we do not correct for multiplicity.

C Sensitivity Analyses

Boundary band robustness. ATE and Lift are stable across near-boundary definitions for *ketone*:

Table 3: ATE and Lift robustness across boundary bands for *ketone*.

Band	ATE	Lift
$[0.35, 0.65]$	0.336	+0.318
$[0.40, 0.60]$	0.329	+0.309
$[0.30, 0.70]$	0.341	+0.321

Amplitude sweep. Varying the edit amplitude $a \in \{0.4, 0.6, 0.8\}$ yields monotone ATE increases with similar Lift, indicating that specificity is not an artifact of large edits. ($a{=}0.8$ used by default.)

Grid granularity. Using ^{1}H steps of 0.1–0.2 ppm and ^{13}C steps of 0.5–1.0 ppm changes ATE by < 0.01 on average; default steps balance fidelity and speed.

D Windows, Grids, and Realism

Chemical windows. 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 (^{1}H) and 0.5–1.0 ppm (^{13}C).

Realism checks. Mutual-exclusivity rules prevent contradictory constellations; intensity realism is enforced by clipping/saturation and skip-logic on high bins; all feasibility checks scored 1.0 in Table 1.

E Modeling and Implementation Notes

All experiments use an 80/20 train–test split with a fixed random seed and calibrated probabilities. Proximity is computed using the ℓ_1 distance on the concatenated, max-normalized spectrum vector, and all spectra are normalized prior to applying edits.

Search efficiency. Beam-k (typically $k \in \{3, 5\}$) reduces myopia with linear overhead in k; candidate scoring is vectorized across ppm bins per case. Optional multi-edit mode (up to three edits) adds an edit-count penalty and consistency checks.

F Data and Licensing

Data. Spectra and SMILES are sourced from `NMRShiftDB2` (license: CC-BY-SA) with standard quality filters. Substructure labels are derived via SMARTS on SMILES only; we do not use NMR annotations to create labels, avoiding leakage.

G Extensions and Limitations

Extension to non-boundary cases. Our evaluation focuses on *near-boundary* cases ($0.35 < p_t(x) < 0.65$), where counterfactual guidance is most decision-relevant—these are the ambiguous spectra where minimal interventions can meaningfully change model outputs. For *high-confidence* cases ($p_t(x) > 0.65$), counterfactuals remain informative but may require larger or multiple edits to alter predictions; they instead answer "what evidence would weaken this confident call?" For *low-confidence* cases ($p_t(x) < 0.35$), one-peak edits may be insufficient, and coordinated or attenuating interventions could be required. This near-boundary focus is intentional: these are precisely the cases where practitioners face real diagnostic uncertainty and need actionable recourse.

Extensions to complex models. For *non-linear models* beyond logistic regression (e.g., neural networks, transformers, or graph neural networks), counterfactual interventions become even more valuable because coefficients no longer exist. Gradient- or perturbation-based explanations (e.g., saliency, integrated gradients, LIME) remain correlational and lack physical plausibility or minimality guarantees. By contrast, our framework enforces chemistry-valid windows and quantifies effect size, extending naturally to any differentiable model by replacing the linear predictor h with a neural network. In complex models, such plausibility constraints are even more critical—they prevent the identification of spurious, nonphysical shortcuts. For graph-based ASE models predicting molecular structures, counterfactuals can isolate which spectral features drive particular structural decisions (e.g., "adding a carbonyl ^{13}C signal makes the model predict a ketone rather than an aldehyde"). In all cases, counterfactual edits provide *actionable, chemically valid, and model-agnostic* interpretability, bridging model reasoning and experimental design.

Limitations. We acknowledge limitations: (i) We test *interventions* (adding peaks), not *removals* (attenuating existing signals); a model relying on spurious shortcuts might respond differently to removals. (ii) Some functional-group cohorts (e.g., *amide*, *aromatic*) are small ($n<25$), which limits statistical power and generalization beyond the present dataset; reported p-values should thus be interpreted as within-model significance rather than population-level inference. (iii) The binned representation omits fine structure (coupling, multiplicities) that could reveal spurious pattern-matching. However, the combination of mechanistic alignment (channels match textbook chemistry), specificity (targeted $>$ random within windows), and reproducibility across labels provides strong evidence that the models we audit have learned chemically meaningful patterns. Future work could validate against ground-truth assignments, test interventions on out-of-distribution data, and examine model attention weights to directly visualize which spectral regions drive predictions.

References

[1] Raymond J. Abraham and Mehdi Mobli. *Analysis of High Resolution NMR spectra*. Elsevier, Amsterdam, 2011.

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

[3] Yoshihiro Asai, Atsushi Fujita, and Yutaka Yamamoto. Learning to retrieve with atomic structure-aware neural networks for chemical reaction prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 628–635, 2018.

[4] 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.

[5] 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.

[6] Timothy D. W. Claridge. *High-Resolution NMR Techniques in Organic Chemistry*. Elsevier, Amsterdam, 3rd edition, 2016.

[7] Susanna Di Vita, Florian Grötschla, Luca A. Lanzendörfer, and Roger Wattenhofer. Leveraging pre-trained language models for rapid and accurate structure elucidation from 2d nmr data. In *NeurIPS 2024 Workshop on AI for Scientific Discovery (AI4Mat)*, 2024. URL `https://openreview.net/forum?id=t8qSHNCjuB`.

[8] Wenhao Gao and Connor W. Coley. Chemllm: Foundation models for chemical and spectroscopic reasoning. *arXiv preprint arXiv:2402.09392*, 2024.

[9] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *AISTATS*, 2020.

[10] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, Isabel Valera, and Bernhard Schölkopf. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. doi: 10.1145/3527848.

[11] Stefan Kuhn and Nils E. Schlorer. Nmrshiftdb2—an enhanced open-source nmr database. *Journal of Cheminformatics*, 16(1):19, 2024. doi: 10.1186/s13321-024-00818-1.

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

[13] Ernő Pretsch, Philippe Bühlmann, and Martin Badertscher. *Structure Determination of Organic Compounds: Tables of Spectral Data*. Springer, Berlin, 4th edition, 2009.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[15] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *PNAS*, 2021.

[16] Robert M. Silverstein, Francis X. Webster, and David J. Kiemle. *Spectrometric Identification of Organic Compounds*. John Wiley & Sons, Hoboken, NJ, 7th edition, 2005.

[17] Christoph Steinbeck, Young J. Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon L. Willighagen. The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2): 493–500, 2003. doi: 10.1021/ci025584y.

[18] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 10–19. ACM, 2019. doi: 10.1145/3287560.3287566.

[19] Susanna Di Vita. Counterfactual nmr: Chemistry-constrained spectral interventions for model auditing. `https://github.com/SusannaDiV/counterfactualNMR`, 2025. Accessed: 2025-11-02.

[20] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2018. Earlier version: arXiv:1711.00399 (2017).