Peak Risk Score: A Peak-Space Verification Layer for AI Scientists in NMR-Guided Molecular Discovery

Anonymous Authors[1]

Abstract

AI systems for molecular discovery increasingly generate candidate structures together with simulated evidence, but existing evaluation methods struggle when observations contain missing peaks, spurious peaks, and uncertainty in peak cardinality. In AI-driven discovery pipelines, models generate hypotheses, simulate observations, and must decide which candidates are consistent with experimental measurements; errors at this stage can lead to overconfident but incorrect selections. In NMR-guided structure elucidation, standard metrics such as shift MAE or RMSE can reward low average error while missing chemically meaningful peak-pattern mismatches. We introduce the Peak Risk Score (PRS), a model-agnostic probabilistic scoring framework for spectroscopy-grounded molecular hypotheses. PRS treats simulated spectra as approximate forecasts, constructs predictive ensembles that capture realistic variability, and evaluates them in peak space using proper scoring-rule components for peak existence, position, and amplitude, yielding interpretable peak-wise contributions. On a synthetic point-process benchmark, PRS detects over-predictive models that add spurious events while preserving low matched-event error, identifying the correct forecaster in 99.6% of cases compared with 28–93% for standard baselines. On $1{,}849$ paired simulated/experimental $^{13}\mathrm{C}$ NMR spectra, PRS achieves 95.5% top-1 accuracy in a structure-identification task with one correct candidate and four decoys, outperforming the best classical shift-based metric (62.5%) and Wasserstein-1 distance (93.3%), while matching DP4 and DP5 probabilities (95.0% and 95.5%). Combining PRS with DP5 further improves accuracy to 95.8%. PRS provides a domain-grounded verification layer for AI-scientist workflows by explicitly modeling uncertainty, penalizing structural mismatches, and handling variable-cardinality experimental observations.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

AI systems for molecular discovery increasingly operate as hypothesis-generating scientific agents: they propose candidate structures, simulate the measurements those structures should produce, and rank the hypotheses that best explain experimental data. This generate–simulate–evaluate loop appears in automated structure elucidation, closed-loop molecular design, simulation-based screening, self-driving laboratories, and emerging AI-scientist workflows (King et al., 2009; Burger et al., 2020; Roch et al., 2020; Gómez-Bombarelli et al., 2018; Segler et al., 2018; Stokes et al., 2020; Schwalbe-Koda & Gómez-Bombarelli, 2020; Boiko et al., 2023; Bran et al., 2024; Lu et al., 2024). Its reliability depends not only on the generative model or simulator, but also on the evaluation layer that decides whether a simulated hypothesis is consistent with experimental evidence.

This paper focuses on a specific but central bottleneck for AI-scientist systems: verification. An autonomous molecular-discovery agent may generate candidate structures and simulate their expected spectra, but it still needs to determine whether those simulations are genuinely supported by experimental observations. We study this problem in NMR-guided molecular discovery and introduce PRS as a domain-grounded, model-agnostic evidence-consistency layer.

Nuclear magnetic resonance (NMR) spectroscopy is a representative setting where this evaluation problem is difficult. Although NMR spectra can be viewed as time-series signals, chemists interpret them through structured peak patterns: each peak has a chemical shift, an integral or intensity, and a chemical meaning tied to an atomic environment (Keeler, 2010b; Claridge, 2009b; Pavia et al., 2014b). Simulated spectra produced by DFT, empirical predictors, or learned models rarely match experiments exactly. Differences arise from referencing errors, conformational effects, solvent and temperature conditions, line broadening, imperfect peak

picking, missing or overlapping peaks, impurities, mixtures, and processing artefacts. Comparing simulated and experimental spectra is therefore not fixed-dimensional regression, but a variable-cardinality matching problem in which missing and spurious peaks may be as important as shift error.

Despite this, many NMR prediction and structure-elucidation pipelines still rely on deterministic shift errors such as MAE, RMSE, or maximum absolute deviation (Klukowski et al., 2025; Smith & Goodman, 2010; Howarth et al., 2020; Howarth & Goodman, 2022; Klukowski et al., 2022; Schützke et al., 2022; Jonas & Kuhn, 2019). These metrics are useful for measuring average shift accuracy, but can fail in cases that matter for automated hypothesis selection: a candidate may achieve low matched-peak error while omitting diagnostic experimental peaks or hallucinating plausible-looking spurious peaks. Distributional distances and kernel similarities partially address alignment issues, including recent spectral benchmarks based on shift-marginalized MMD (Di Vita, 2025), but they remain deterministic comparisons between spectra and typically do not provide probabilistic, peak-wise explanations. Classical Bayesian methods such as DP4 and DP5 remain strong structure-ranking baselines, but they are designed around calibrated chemical-shift error models and do not directly expose local peak-existence failures as an evaluation signal (Smith & Goodman, 2010; Howarth et al., 2020; Howarth & Goodman, 2022).

We argue that AI-driven scientific workflows require evaluation metrics that are probabilistic, decomposable, and aligned with downstream decisions. This view is consistent with probabilistic forecasting, where proper scoring rules such as CRPS and related criteria evaluate predictive distributions rather than only point predictions (Gneiting & Raftery, 2007; Matheson & Winkler, 1976; Gneiting & Katzfuss, 2014). For NMR-guided molecular structure elucidation, this means evaluating whether a candidate's full predicted peak pattern provides a plausible explanation of the observed spectrum, including uncertainty in peak existence, position, and integral.

We introduce the *Peak Risk Score* (PRS), a model-agnostic probabilistic scoring framework for evaluating spectroscopy-grounded molecular hypotheses. PRS treats a simulated peak list as an approximate forecast of the experimental peak list, constructs an ensemble of perturbed peak lists capturing realistic experimental variability, aligns simulated and experimental peaks into local groups, and scores each group using peak-existence uncertainty together with chemical-shift and integral errors. The resulting risk decomposes into interpretable peak-wise contributions and can optionally be chemically weighted, allowing an AI-scientist workflow to identify whether a candidate fails because of missing carbonyl peaks, spurious aromatic signals, poor integral agreement, or systematic shift mismatch.

Figure 1 summarizes PRS as a verification layer for AI-scientist workflows. We evaluate PRS in two settings. On a synthetic point-process benchmark, PRS reliably detects over-predictive forecasters that add spurious events while preserving low matched-event error, identifying the correct forecaster in 99.6% of sequences. On 1,849 paired simulated/experimental ^{13}C NMR peak lists, PRS achieves 95.5% top-1 accuracy in a DP4-style structure-ranking task with one correct candidate and four decoys, substantially outperforming classical shift-based metrics and Wasserstein-1 distance while matching DP4/DP5 probabilities. Combining PRS with DP5 further improves top-1 accuracy to 95.8%, suggesting that peak-space risk captures information complementary to classical Bayesian shift-error models.

Our contributions are as follows:

1. We introduce PRS as a probabilistic verification layer for AI-scientist molecular-discovery workflows, enabling generated molecular hypotheses to be checked against variable-cardinality experimental evidence.

2. We show on a synthetic point-process benchmark that PRS detects over-predictive forecasters that standard matched-error metrics fail to penalize.

3. We demonstrate on 1,849 paired simulated/experimental ^{13}C NMR spectra that PRS reaches DP4/DP5-level structure-ranking accuracy while providing local interpretability and complementarity to existing Bayesian methods.

2. Related Work

NMR structure elucidation and Bayesian ranking. NMR spectroscopy is central to organic structure elucidation because chemical shifts, integrals, and correlations encode molecular connectivity and local chemical environments (Keeler, 2010a; Claridge, 2009a; Pavia et al., 2014a; Silverstein et al., 2014). Computer-assisted structure elucidation has long used NMR constraints to propose or rank structures, including logic-based, rule-based, and expert-system approaches (Jaspars, 1999; Nuzillard & Massiot, 1991; Steinbeck, 2001; Lindel et al., 1999; Köck et al., 2021; Elyashberg et al., 2004; Elyashberg & Argyropoulos, 2021; Kuhn & Schlörer, 2015). DP4, DP4-AI, and DP5 established probabilistic candidate ranking from computed and experimental shifts using calibrated DFT error models (Smith & Goodman, 2010; Howarth et al., 2020; Howarth & Goodman, 2022). PRS is complementary: instead of only modeling residual shift errors, it scores whether the candidate reproduces the observed peak pattern, including missing and spurious peaks.

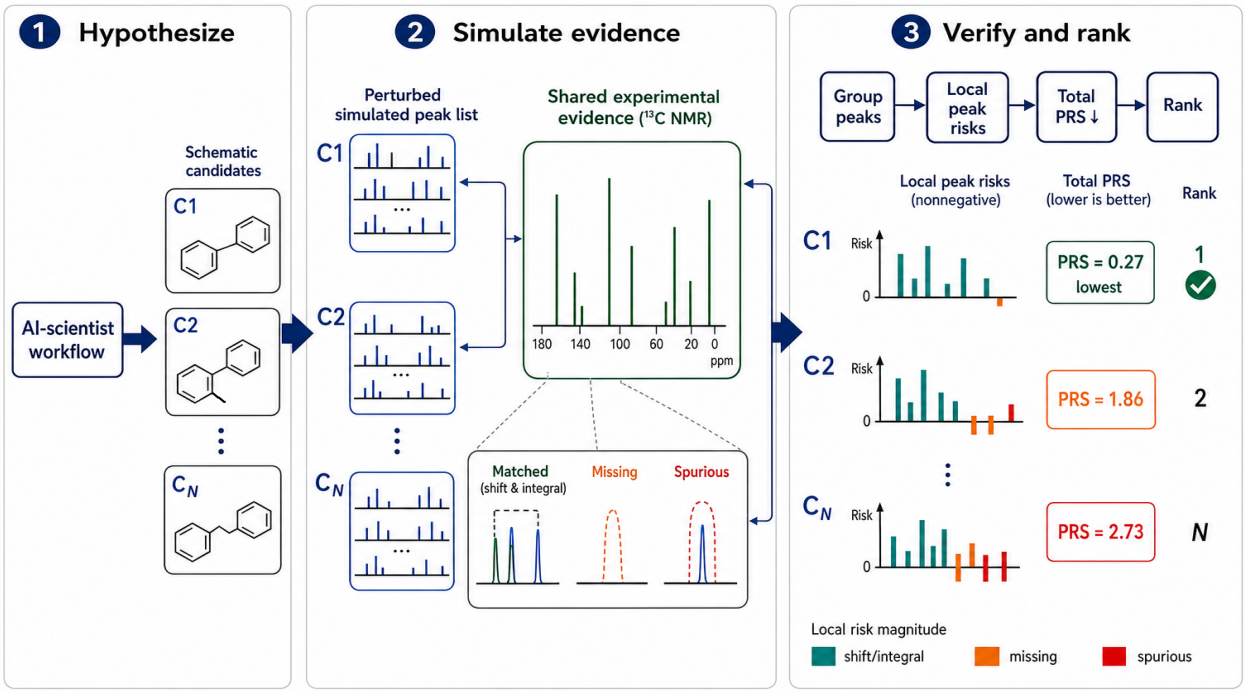

Figure 1. Overview of PRS as a peak-space verification layer for AI-scientist workflows. An AI-scientist workflow proposes candidate molecular hypotheses, each candidate is converted into a predictive ^{13}C NMR peak-list ensemble, and the ensemble is compared against shared experimental evidence. PRS groups peaks, computes decomposable local risks for matched, missing, and spurious peaks, aggregates them into a total Peak Risk Score, and ranks candidate structures, with lower PRS indicating better agreement with the experimental spectrum.

Machine learning and metrics for spectroscopy. Machine learning is increasingly used for NMR prediction, spectral interpretation, and automated structure elucidation, including nmrshiftdb2, uncertainty-aware NMR prediction, and ARTINA (Kuhn & Schlörer, 2015; Jonas & Kuhn, 2019; Klukowski et al., 2022; Kuhn et al., 2008). Surveys of deep learning in NMR and broader spectroscopy emphasize that evaluation often relies on generic grid-level, image-style, or regression losses rather than metrics aligned with physicochemical spectral structure (Liu et al., 2020; Bejar-Grimalt et al., 2025; Qu, 2020; Luo et al., 2022; Cai et al., 2025; Passos et al., 2022). Recent benchmarks introduce synthetic and semi-synthetic spectral datasets (Schützke et al., 2022; 2023; Chen et al., 2024; da Silva et al., 2023); Sim-to-Exp-MMNMR further pairs simulated and experimental ^{1}H and ^{13}C NMR spectra and introduces shift-marginalized MMD (Di Vita, 2025) to measure their distributional shift. These distances improve robustness but remain deterministic spectrum-level comparisons, whereas PRS provides probabilistic and local peak-wise diagnostics.

Probabilistic verification and variable-cardinality observations. Proper scoring rules evaluate predictive distributions and reward calibration and sharpness (Gneiting & Raftery, 2007; Matheson & Winkler, 1976; Gneiting & Katzfuss, 2014; Jordan et al., 2019; Raftery et al., 2005; Winkler, 1996). The Brier score handles binary event occurrence (Brier, 1950), while CRPS and related scores evaluate continuous ensemble forecasts (Hersbach, 2000; Bröcker & Smith, 2008; Gneiting & Raftery, 2005). Related work on point-process and kernel/energy scoring emphasizes that event forecasts must account for locations, counts, and distributional structure (Brehmer et al., 2021; Gneiting & Ranjan, 2011). PRS adapts this perspective to NMR peak lists, where forecasts have uncertain cardinality, chemically heterogeneous regions, and explicit missing/spurious peak failures.

3. Methodology

3.1. Problem setup

We consider molecules indexed by n, each with a simulated ^{13}C NMR spectrum and a corresponding experimental spectrum. Since chemists interpret NMR spectra through peaks, we represent each experimental spectrum as a peak list

$$\mathcal{Y}^{(n)} = \{y_j^{(n)}\}_{j=1}^{J_n}, \qquad y_j^{(n)} = (\omega_j, a_j, c_j),$$

where ω_j is the chemical shift in ppm, a_j is a normalized integral or intensity, and $c_j \in \mathcal{C}$ is a coarse chemical-region label such as aliphatic, aromatic, heteroatom-attached, or carbonyl. The corresponding simulated spectrum is represented as

$$\mathcal{X}^{0,(n)} = \{x_i^{0,(n)}\}_{i=1}^{I_n}, \qquad x_i^{0,(n)} = (\tilde{\omega}_i, \tilde{a}_i, c_i).$$

Unlike fixed-dimensional regression metrics, this representation allows simulated and experimental spectra to have different numbers of peaks. We treat $\mathcal{X}^{0,(n)}$ as an approximate mean forecast of $\mathcal{Y}^{(n)}$ and evaluate whether the simulated peak pattern provides a plausible probabilistic explanation of the experimental one.

3.2. Predictive ensembles in peak space

For each candidate structure, PRS constructs an empirical predictive distribution over peak lists by perturbing the simulated peak list. Formally, we write the ensemble as samples from an empirical perturbation distribution around the simulated spectrum,

$$\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(M)} \sim \mathcal{P}_\theta(\mathcal{X} \mid \mathcal{X}^0),$$

where $\theta = (\sigma_\omega, \sigma_a, p_{\text{drop}}, p_{\text{add}})$ denotes fixed perturbation hyperparameters. For ensemble member $m \in \{1, \ldots, M\}$, each simulated peak is perturbed as

$$\omega_i^{(m)} = \tilde{\omega}_i + \varepsilon_{\omega,i}^{(m)}, \qquad \varepsilon_{\omega,i}^{(m)} \sim \mathcal{N}(0, \sigma_\omega^2),$$

$$a_i^{(m)} = \tilde{a}_i \exp(\varepsilon_{a,i}^{(m)}), \qquad \varepsilon_{a,i}^{(m)} \sim \mathcal{N}(0, \sigma_a^2).$$

Each peak is independently dropped with probability p_{drop}, modeling missing or overlapping signals, and spurious peaks are optionally added with probability p_{add}, modeling impurities, processing artefacts, or simulator mismatch. This yields an ensemble

$$\hat{\mathcal{X}}^{(1:M)} = \{\mathcal{X}^{(m)}\}_{m=1}^{M},$$

which acts as a forecast distribution over variable-cardinality peak lists. In all main experiments, we use

$$M = 50, \quad \sigma_\omega = 1.0 \text{ ppm}, \quad \sigma_a = 0.1,$$
$$p_{\text{drop}} = 0.05, \quad p_{\text{add}} = 0.03,$$

and do not train any additional model.

3.3. Peak grouping and existence probabilities

To compare the forecast ensemble with the experimental peak list, we align peaks into local groups. Given a maximum matching distance τ, experimental and simulated peaks may be matched only if their chemical shifts differ by at most τ ppm. Matched peaks across ensemble members are merged into groups $g \in \mathcal{G}$, each corresponding either to an observed experimental peak or to a false-positive simulated peak group. Each group has an experimental target

$$y_g = \begin{cases} (\omega_g^\star, a_g^\star, c_g), & \text{if an exp. peak is present,} \\ \varnothing, & \text{otherwise,} \end{cases}$$

and at most one assigned simulated peak from each ensemble member:

$$\mathcal{X}_g^{(m)} = \begin{cases} (\omega_g^{(m)}, a_g^{(m)}, c_g), & \text{if a sim. peak is assigned to } g, \\ \varnothing, & \text{otherwise.} \end{cases}$$

The empirical peak-existence probability is

$$\hat{p}_g = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left[\mathcal{X}_g^{(m)} \neq \varnothing\right].$$

Conditioned on existence, the assigned ensemble peaks define empirical distributions over shifts and integrals:

$$\Omega_g = \{\omega_g^{(m)} : \mathcal{X}_g^{(m)} \neq \varnothing\}, \qquad A_g = \{a_g^{(m)} : \mathcal{X}_g^{(m)} \neq \varnothing\}.$$

3.4. Peak-wise scoring

For each group, PRS combines peak-existence uncertainty with continuous uncertainty over chemical shift and integral. Let $Z_g = 1$ if an experimental peak exists in group g and $Z_g = 0$ otherwise. We score existence with the Brier score,

$$S_{\text{exist}}(g) = (\hat{p}_g - Z_g)^2.$$

For groups with an experimental peak, we score the forecast shift distribution with an empirical CRPS term:

$$\widehat{\text{CRPS}}_\omega(g) = \frac{1}{M_g} \sum_{m=1}^{M_g} \left|\omega_g^{(m)} - \omega_g^\star\right| - \frac{1}{2M_g^2} \sum_{m,m'} \left|\omega_g^{(m)} - \omega_g^{(m')}\right|,$$

where $M_g = |\Omega_g|$. The integral term is defined analogously:

$$\widehat{\text{CRPS}}_a(g) = \frac{1}{M_g} \sum_{m=1}^{M_g} \left|a_g^{(m)} - a_g^\star\right| - \frac{1}{2M_g^2} \sum_{m,m'} \left|a_g^{(m)} - a_g^{(m')}\right|.$$

The peak-wise risk is

$$S_{\text{peak}}(g) = \lambda_{\text{exist}} S_{\text{exist}}(g) + \lambda_\omega \widehat{\text{CRPS}}_\omega(g) + \lambda_a \widehat{\text{CRPS}}_a(g).$$

For false-positive groups, where $y_g = \varnothing$, only the existence term is active. This makes PRS explicitly sensitive to hallucinated peaks, unlike matched-error metrics that only evaluate

Table 1. Synthetic point-process benchmark. We report the fraction of 500 sequences where each metric identifies GOOD as best, recovers the full ordering GOOD<MODERATE<POOR, and scores GOOD better than the OVER-PREDICTOR.

Metric	Top-1	Full order	Good<OverP
PRS (ours)	**0.996**	**1.000**	**0.996**
Mean $\lvert\Delta t\rvert$	0.286	0.890	0.286
RMSE(Δt)	0.276	0.896	0.278
Signal RMSE	0.926	0.902	0.930
W_1	0.804	0.644	0.910

assigned peaks. The existence term is a Brier score and the shift/integral terms are empirical CRPS components, so PRS inherits the calibration-oriented behavior of proper scoring rules while remaining decomposable over peak groups; the precise propriety assumptions are discussed in Appendix A. Because peak grouping is data-dependent, PRS should be interpreted as a proper-scoring-rule-based evaluation metric under the induced grouping, rather than as a globally calibrated likelihood over raw spectra.

3.5. Aggregating peak risk and ranking structures

The full Peak Risk Score is a chemically weighted sum of local risks:

$$\mathrm{PRS}(\hat{\mathcal{X}}^{(1:M)}, \mathcal{Y}) = \sum_{g \in \mathcal{G}} w(c_g) S_{\mathrm{peak}}(g),$$

where $w(c_g) > 0$ can upweight chemically diagnostic regions such as carbonyls or heteroatom-attached carbons. Lower PRS indicates that the candidate's simulated spectrum better explains the experimental peak list under the forecast ensemble.

For structure-ranking experiments, each molecule is associated with one correct candidate and $K-1$ decoys. We compute PRS_k for each candidate k and convert risks into compatibility scores

$$\ell_k = -\mathrm{PRS}_k, \qquad p_k = \frac{\exp(\ell_k)}{\sum_{k'} \exp(\ell_{k'})}.$$

We then report top-1 accuracy, i.e., the fraction of molecules for which the correct structure receives the highest compatibility. This protocol allows direct comparison with classical shift-error metrics, Wasserstein distances, SM-MMD, and DP4/DP5 probabilities.

4. Results

4.1. Synthetic Point-Process Benchmark

To demonstrate that PRS benefits arise from the *structure* of the scoring rule rather than NMR-specific tuning, we first evaluate it on a simple 1D point-process toy benchmark (Poisson event sequences with perturbed forecasters; full setup in Appendix B). In this setting, PRS is the only metric that reliably detects an OVER-PREDICTOR that adds ∼50% spurious events while maintaining low position error on matched events: it correctly identifies GOOD as the best forecaster in 99.6% of sequences and ranks GOOD better than OVER-PREDICTOR in 99.6% of cases, compared to 28–93% for standard match-based and signal-level baselines (Table 1). This supports PRS as a robust evaluator for variable-cardinality event forecasts, even outside the NMR domain.

4.2. NMR Evaluation

Evaluation setup. We evaluate PRS and baselines on a benchmark of 1,849 molecules with paired simulated and experimental ^{13}C peak lists, each spectrum containing at least two peaks (Di Vita, 2025). For each candidate structure, we generate an ensemble of $M = 50$ peak-list forecasts using fixed perturbation hyperparameters ($\sigma_\omega = 1.0$ ppm, $\sigma_a = 0.1, p_{\mathrm{drop}} = 0.05, p_{\mathrm{add}} = 0.03$) and group peaks with a maximum matching distance of $\tau = 5.0$ ppm. Unless otherwise stated, all results refer to this configuration.

We compute PRS for each molecule and compare it to classical peak-space metrics: average absolute shift error (Avg $\lvert\Delta\mathrm{ppm}\rvert$), RMSE in ppm, maximum absolute shift error (Max $\lvert\Delta\mathrm{ppm}\rvert$), a region-weighted absolute prediction error (APE), and the W_1 Wasserstein distance between simulated and experimental shift distributions. We also compare to an SM-MMD similarity score and to DP4 and DP5 probabilities for the same candidate set.

DP4-style structure ranking. To test whether PRS captures the aspects of forecast quality that matter for structure elucidation, we use a DP4-style ranking protocol. For each molecule, we construct one correct structure and $K = 4$ random molecule decoys (total 5 candidates), compute PRS and all baselines, and report top-1 accuracy.

Table 2 summarizes the results. PRS alone achieves a top-1 accuracy of 0.9551, substantially higher than all purely shift-based metrics. Among distance-based baselines, W_1 reaches 0.9329 top-1 accuracy, while SM-MMD similarity attains 0.9567, very slightly higher than PRS. DP4 and DP5 reach 0.9502 and 0.9551 respectively on the same candidate sets. PRS and DP methods are complementary: simply combining PRS with DP4 or DP5 via a logistic fusion of their scores improves top-1 accuracy further to 0.9567 (DP4+PRS) and 0.9578 (DP5+PRS). Although the absolute gain over DP5 is modest, it is meaningful because PRS adds a different verification signal: it penalizes missing and spurious local peak patterns that can be invisible to calibrated shift-error likelihoods. Thus, the improvement indicates

Table 2. DP4-style structure ranking accuracy (top-1) with one correct structure and four decoys, evaluated across 1,849 molecules. Metrics are grouped by type. Best result per category is **bolded**; overall best is underlined.

Category	Metric	Top-1 Acc.
Point-estimate	Avg $\lvert\Delta\delta\rvert$	**0.6247**
	RMSE(δ)	0.5592
	Max $\lvert\Delta\delta\rvert$	0.3716
	Region-weighted APE	0.6160
Distributional	W_1	0.9329
	SM-MMD similarity	**0.9567**
Probabilistic	DP5 probability	**0.9551**
	DP4 raw	0.1769
	DP4 probability	0.9502
Proposed	PRS	0.9551
Combo	$\text{DP4}_{\text{proper}}\times\text{PRS}$	0.9567
	DP5 $\times$ PRS	**<u>0.9578</u>**

complementarity rather than metric redundancy. Additional ablation studies and local peak analyses are reported in Appendix C.

4.3. Robustness to PRS Design Choices

Because PRS contains several user-facing design choices, we test whether its behaviour is stable under changes to ensemble size, perturbation scale, matching threshold, and chemical-region weighting. All ablations are performed on a 200-molecule subsample unless otherwise stated. Increasing the number of ensemble members stabilizes the score but gives diminishing returns: the median PRS decreases from 0.86 at $M = 5$ to 0.71 at $M = 20$ and 0.67 at $M = 200$, while the standard deviation decreases from 11.0 to 6.4. This suggests that moderate ensembles already provide stable uncertainty estimates, and the default choice $M = 50$ is sufficient for ranking. PRS is also robust to the perturbation hyperparameters used to construct the predictive ensemble. Sweeping the shift noise $\sigma_\omega \in \{0.3, 0.5, 1.0, 2.0, 3.0\}$ and peak-drop probability $p_{\text{drop}} \in \{0.0, 0.05, 0.10, 0.20\}$ keeps median PRS values within a narrow range of approximately 0.69–0.90, with best behaviour around $\sigma_\omega \in [0.5, 2.0]$. Very small perturbations underestimate experimental variability, while very large perturbations over-smooth the forecast distribution, but neither regime qualitatively changes the ranking behaviour. The peak-matching threshold τ shows similar stability. An overly strict threshold of $\tau = 1$ ppm produces a lower median PRS but a much higher mean due to many unmatched peaks being treated as large errors. Beyond $\tau = 5$ ppm, the median PRS saturates around 0.68–0.69, suggesting that PRS is not sensitive to the exact threshold once it lies in a chemically reasonable range. Finally, different chemical-region weighting schemes—uniform, chemically motivated, inverse-frequency, and carbonyl-heavy weights—all achieve similar DP4-style top-1 accuracies in the range 0.95–0.96. This indicates that chemical weights can be used for interpretability or task-specific emphasis without destabilizing the metric.

4.4. Generalization Beyond the Main ^{13}C Setting

We also test whether PRS remains informative outside the main ^{13}C ranking setting. Applying the same pipeline to ^{1}H NMR yields a median PRS of 2.96 and a mean Avg $\lvert\Delta\text{ppm}\rvert$ of 0.074. In a DP4-style ranking task, PRS reaches a top-1 accuracy of 0.60, compared with 0.47 for Avg $\lvert\Delta\text{ppm}\rvert$. Although ^{1}H NMR is more affected by overlap, multiplicity, and integration ambiguity than ^{13}C NMR, PRS still provides a stronger ranking signal than a matched-shift baseline. We further evaluate how PRS behaves as the number of decoy structures increases. On ^{13}C NMR, top-1 accuracy degrades gracefully from 0.97 for $K = 2$ candidates to 0.94 for $K = 4$, 0.91 for $K = 8$, and 0.89 for $K = 16$. This trend is expected because the ranking problem becomes harder as the candidate set grows, but the degradation is smooth rather than catastrophic. These results suggest that PRS captures a robust peak-space compatibility signal rather than overfitting to the default one-correct-plus-four-decoys setting.

4.5. Local Interpretability of Peak Risk

A key advantage of PRS over spectrum-level distances is that the total score decomposes into local peak-wise contributions. This makes it possible to identify why a candidate structure fails: whether because of missing experimental peaks, spurious predicted peaks, poor chemical-shift agreement, or integral mismatch. Across the full ^{13}C dataset, PRS is strongly correlated with classical shift errors, including Spearman $\rho = 0.73$ with Avg $\lvert\Delta\text{ppm}\rvert$ and $\rho = 0.73$ with RMSE, but it also captures peak-existence failures that averaged shift metrics ignore. PRS is negatively correlated with peak-matching F1 ($\rho = -0.55$) and SM-MMD similarity ($\rho = -0.67$), consistent with its interpretation as a loss-like measure of spectral mismatch. To examine chemically meaningful local errors, we analyze 987 carbonyl peaks. The peak-wise shift CRPS is highly correlated with Avg $\lvert\Delta\text{ppm}\rvert$ on these peaks (Spearman $\rho = 0.91$), but unlike absolute shift error it also accounts for predictive spread and missingness. In qualitative examples, the largest PRS contributions often arise from missing or duplicated heteroatom and carbonyl peaks, precisely the local failures that can drive incorrect structure assignments. This decomposability is useful for AI-scientist workflows: rather than only returning a scalar confidence score, PRS can indicate which parts of the experimental evidence are not explained by a generated molecular hypothesis.

4.6. Complementarity with DP4/DP5

PRS and DP4/DP5 encode different notions of agreement between simulated and experimental NMR data. DP4 and DP5 model per-atom chemical-shift errors under calibrated DFT error distributions, making them strongest when the main uncertainty is a systematic or method-level shift error. PRS instead evaluates whether the predictive ensemble reproduces the observed peak pattern, including peak existence, local shift uncertainty, and integral agreement. It is therefore strongest when the relevant failure is local peak-pattern inconsistency, such as missing diagnostic peaks, spurious peaks, or poor local agreement.

These different inductive biases explain why score fusion improves ranking. When a candidate has plausible matched shifts but fails to reproduce important experimental peaks, PRS assigns a large local risk and can down-rank an overconfident DP5 candidate. Conversely, when all candidates suffer from a shared systematic DFT shift bias, DP5 can remain reliable while PRS is high for all candidates. The improvement from DP5 alone (0.9551) to DP5+PRS (0.9578) is numerically small but meaningful: it indicates that peak-space risk resolves a subset of errors not captured by Bayesian shift-error modeling.

From the perspective of AI-driven scientific discovery, this complementarity is desirable. A generated molecular hypothesis should not be accepted only because its average shifts are plausible; it should also explain the observed experimental evidence at the level of local, chemically meaningful peaks. The fused score therefore acts as a simple combined verifier: DP4/DP5 contribute global shift-error likelihoods, while PRS contributes local peak-space consistency. PRS can therefore serve as a model-agnostic diagnostic layer on top of existing NMR predictors or Bayesian ranking systems.

5. Conclusion

We introduced the Peak Risk Score (PRS), a model-agnostic peak-space evaluation framework for spectroscopy-grounded molecular hypothesis selection. Motivated by AI-driven discovery workflows, where systems generate candidate structures, simulate their observable consequences, and must decide which hypotheses are consistent with experimental evidence, PRS provides a probabilistic verification layer for NMR-guided structure elucidation. Rather than reducing spectra to average shift errors, PRS evaluates whether a candidate reproduces the observed peak pattern under uncertainty, explicitly accounting for peak existence, chemical-shift position, and integral agreement.

On 1,849 paired simulated/experimental peak-list examples, PRS is strongly correlated with classical shift errors while providing a better proxy for downstream structure identification. In a DP4-style ranking task with one correct structure and four decoys, PRS achieves top-1 accuracy comparable to DP4 and DP5, and simple score fusion with DP5 performs best overall. Ablation studies show that PRS is robust to ensemble size, perturbation scale, matching threshold, and chemical-region weighting. The synthetic point-process benchmark further shows that PRS can detect over-predictive forecasters that add spurious events while maintaining low matched-event error, supporting its relevance beyond NMR as an evaluation tool for variable-cardinality scientific observations.

The main value of PRS is not only higher ranking accuracy, but also interpretability and reliability. Because the score decomposes into local peak-wise contributions, it can identify whether a candidate structure fails due to missing diagnostic peaks, spurious signals, poor shift agreement, or integral mismatch. This makes PRS useful as a diagnostic module inside AI-scientist pipelines: an autonomous system could use it to reject overconfident but poorly supported candidates, request additional measurements, rerun simulations, or escalate uncertain cases to human experts.

PRS also highlights an important limitation of purely model-based scientific evaluation. In realistic spectra, discrepancies may arise not only from an incorrect molecular structure, but also from impurities, mixtures, solvent effects, incomplete simulations, or processing artefacts. In such cases, PRS will correctly flag a mismatch between the simulated and experimental evidence, even if the underlying proposed structure is chemically correct. This behavior is desirable for verification, but it also suggests that future AI-scientist systems should distinguish between structural error and experimental or preprocessing error.

Future work includes learning data-driven perturbation models that better approximate experimental error distributions, extending PRS to multidimensional and multimodal NMR, integrating PRS into autonomous agents for NMR-guided structure elucidation and synthesis planning, and combining peak-space risk with representation-based spectral distances or time-series foundation models. More broadly, PRS illustrates a design pattern for AI scientists: generated hypotheses should not be accepted solely because they are plausible under a simulator or generative model, but should be verified against experimental evidence using decomposable, uncertainty-aware, domain-grounded scoring rules.

References

Bejar-Grimalt, J. et al. SpectraML: Challenges and opportunities for machine learning in spectroscopic analysis. *Journal of Magnetic Resonance*, 2025. doi: 10.1016/j.jmr.2025.108123. in press.

Boiko, D. A., MacKnight, R., and Gomes, G. Emergent

autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6: 525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Brehmer, J., Cranmer, K., Louppe, G., and Pavez, J. Scoring rules for point processes. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Bröcker, J. and Smith, L. A. From ensemble forecasts to predictive distribution functions. *Tellus A: Dynamic Meteorology and Oceanography*, 60(4):663–678, 2008. doi: 10.1111/j.1600-0870.2008.00333.x.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., et al. A mobile robotic chemist. *Nature*, 583(7815): 237–241, 2020.

Cai, Y., Lin, Y., Cai, H., and Ni, H. Deep learning in vibrational spectroscopy: Benefits, limitations, and recent progress. *Journal of the Chinese Chemical Society*, 72 (6):611–626, 2025.

Chen, Z., Badman, R. P., Foley, B. L., Woods, R. J., and Hong, P. Glyconmr: Dataset and benchmark of carbohydrate-specific nmr chemical shift for machine learning research. *Journal of Data-centric Machine Learning Research*, 2024.

Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry*. Tetrahedron Organic Chemistry Series. Elsevier, Oxford, 2 edition, 2009a.

Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry*. Tetrahedron Organic Chemistry Series. Elsevier, Oxford, 2 edition, 2009b.

da Silva, T., Wilson, M., Heckemann, R., et al. Agnostic: A large-scale benchmark dataset for machine learning-assisted spectral analysis in magnetic resonance spectroscopy. *Magnetic Resonance in Medicine*, 89(6):3123–3138, 2023.

Di Vita, S. Simexp-mmnmr: A benchmark for simulation-to-experiment generalization in multimodal nmr with chemistry-aware metrics. In *Special Issue Wiley Advanced Intelligent Systems, Proceedings of the NeurIPS 2025 AI4Mat Workshop*, 2025. arXiv:2501.01234.

Elyashberg, M. E. and Argyropoulos, D. Computer-assisted structure elucidation: Present and future perspectives. *Magnetic Resonance in Chemistry*, 2021.

Elyashberg, M. E., Williams, A. J., and Blinov, K. A. Structure elucidator: A versatile expert system for molecular structure elucidation from 1D and 2D NMR spectra. *Journal of Chemical Information and Computer Sciences*, 2004.

Gneiting, T. and Katzfuss, M. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Gneiting, T. and Raftery, A. E. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005. doi: 10.1126/science.1115255.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Gneiting, T. and Ranjan, R. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

Howarth, A. and Goodman, J. M. The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. *Chemical Science*, 13(12):3507–3518, 2022. doi: 10.1039/D1SC05951A.

Howarth, A., Ermanis, K., and Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chemical Science*, 11(17):4351–4359, 2020. doi: 10.1039/D0SC00442A.

Jaspars, M. Computer-assisted structure elucidation of natural products using two-dimensional NMR spectroscopy. *Natural Product Reports*, 1999.

Jonas, E. and Kuhn, S. Rapid prediction of nmr spectral properties with quantified uncertainty. *arXiv preprint arXiv:1910.00659*, 2019.

Jordan, A., Krüger, F., and Lerch, S. Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, 90(12):1–37, 2019.

Keeler, J. *Understanding NMR Spectroscopy*. John Wiley & Sons, Chichester, 2 edition, 2010a.

Keeler, J. *Understanding NMR Spectroscopy*. John Wiley & Sons, Chichester, 2 edition, 2010b.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., et al. The automation of science. *Science*, 324(5923):85–89, 2009.

Klukowski, P., Riek, R., and Güntert, P. Machine learning in nmr spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 148–149:101575, 2025. doi: 10.1016/j.pnmrs.2025.101575.

Klukowski, P. et al. ARTINA: AI-aided structure elucidation from routine NMR spectra. *Nature Communications*, 13 (1):7974, 2022. doi: 10.1038/s41467-022-35557-7.

Köck, M., Junker, J., Gräf, C., and Lindel, T. WEBCOCON: A new HETCOR-based computer-assisted structure elucidation protocol. *Molecules*, 2021.

Kuhn, S. and Schlörer, N. E. nmrshiftdb2—restructuring a free and open-source NMR database. *Journal of Cheminformatics*, 2015.

Kuhn, S., Colbourne, A., Wilkins, J., and Marriott, P. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 2008.

Lindel, T., Junker, J., Köck, M., and Paul, G. M. 2d-nmr-guided constitutional analysis of organic compounds employing the computer program cocon. *European Journal of Organic Chemistry*, 1999.

Liu, F. et al. Deep learning in nuclear magnetic resonance spectroscopy. *Chemical Communications*, 56(68):9866–9882, 2020. doi: 10.1039/D0CC03398A.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Luo, Y., Zhang, J., Li, Z., and Wang, X. Deep learning in raman spectroscopy: Fundamentals, progress, and prospects. *TrAC Trends in Analytical Chemistry*, 157: 116795, 2022. doi: 10.1016/j.trac.2022.116795.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976.

Nuzillard, J.-M. and Massiot, D. Logic for structure determination. *Tetrahedron*, 1991.

Passos, D., Rodrigues, A., and Lopes, J. Perspectives on deep learning for near-infrared spectral data. *Journal of Near Infrared Spectroscopy*, 30(3–4):169–192, 2022. doi: 10.1177/09670335221109562.

Pavia, D. L., Lampman, G. M., Kriz, G. S., and Vyvyan, J. R. *Introduction to Spectroscopy*. Cengage Learning, Boston, 5 edition, 2014a.

Pavia, D. L., Lampman, G. M., Kriz, G. S., and Vyvyan, J. R. *Introduction to Spectroscopy*. Cengage Learning, Boston, 5 edition, 2014b.

Qu, X. Review and prospect: Deep learning in nmr spectroscopy. *Chemistry – A European Journal*, 26(46):10391–10401, 2020. doi: 10.1002/chem.202001039.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005. doi: 10.1175/MWR2906.1.

Roch, L. M., Häse, F., Kreisbeck, C., Tamayo-Mendoza, T., Yunker, L. P. E., Hein, J. E., and Aspuru-Guzik, A. Chemos: An orchestration software to democratize autonomous discovery. *PLOS ONE*, 15(4):e0229862, 2020.

Schützke, F., Blümel, M., de Sá, F., et al. Validating neural networks as models for spectroscopic classification. *npj Computational Materials*, 9(1):47, 2023. doi: 10.1038/s41524-023-009XX-x.

Schützke, F. H. et al. A universal synthetic dataset for machine learning-assisted analysis of spectroscopic data. *Digital Discovery*, 1(5):504–514, 2022. doi: 10.1039/D2DD00074D.

Schwalbe-Koda, D. and Gómez-Bombarelli, R. Generative models for automatic chemical design. In Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., and Müller, K.-R. (eds.), *Machine Learning Meets Quantum Physics*, pp. 445–467. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-40245-7_21.

Segler, M. H., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

Silverstein, R. M., Webster, F. X., Kiemle, D. J., and Bryce, D. L. *Spectrometric Identification of Organic Compounds*. John Wiley & Sons, Hoboken, NJ, 8 edition, 2014.

Smith, S. G. and Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR shift calculation: the DP4 probability. *Journal of the American Chemical Society*, 132(37):12946–12959, 2010. doi: 10.1021/ja105035r.

Steinbeck, C. Seneca: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation. *Journal of Chemical Information and Computer Sciences*, 2001.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, J. R., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4): 688–702, 2020.

Winkler, R. L. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996. doi: 10.1007/BF02562681.

A. Additional Methodological Details

A.1. Peak-list perturbation details

The perturbation model used in PRS is intentionally simple and model-agnostic. The same perturbation configuration is used for all molecules and all ranking experiments; no candidate-specific error model is learned or fitted to improve top-1 accuracy. Ablations in Appendix C show that PRS is robust to moderate changes in ensemble size, perturbation scale, dropout probability, matching threshold, and chemical-region weights.

Spurious peaks are sampled from broad empirical priors over shift and integral ranges estimated from the experimental dataset. They are included to model impurities, peak-picking artefacts, weak overlapping signals, or simulator incompleteness, and to stress-test whether a metric penalizes over-prediction. PRS can also be computed with $p_{\text{add}} = 0$ when the goal is to evaluate only perturbations around the simulator output.

A.2. Peak grouping implementation

Peak grouping is performed independently for each molecule. Unless otherwise stated, we use $\tau = 5.0$ ppm for ^{13}C NMR. When multiple simulated peaks are eligible for the same experimental peak, we select a disjoint assignment by minimizing total shift distance using a greedy or Hungarian-style matching procedure. Matched simulated peaks across ensemble members are merged into a group associated with the same experimental peak. Simulated peaks that cannot be matched to any experimental peak form false-positive groups, and unmatched experimental peaks define groups with low or zero forecast existence probability.

The grouping step is not intended to infer atom-level assignments. Instead, it defines local peak neighborhoods in which probabilistic forecast quality can be evaluated. This is important because NMR peak comparison is inherently variable-cardinality: the simulated and experimental spectra may contain different numbers of peaks, and unmatched peaks are often chemically meaningful. The local Brier and CRPS components remain proper for their corresponding grouped variables once the grouping is fixed.

A.3. General CRPS form

For a univariate predictive distribution F and observation y, the CRPS is

$$\begin{aligned}\mathrm{CRPS}(F, y) = \mathbb{E}_{X\sim F}|X - y| \\ - \frac{1}{2}\mathbb{E}_{X,X'\sim F}|X - X'|.\end{aligned}$$

Given ensemble samples $\{x_m\}_{m=1}^{M_g}$, we use

$$\begin{aligned}\widehat{\mathrm{CRPS}}(F, y) = \frac{1}{M_g}\sum_{m=1}^{M_g}|x_m - y| \\ - \frac{1}{2M_g^2}\sum_{m,m'}|x_m - x_{m'}|.\end{aligned}$$

This estimator is used separately for chemical shifts and integrals in the main text. If no simulated peak is assigned to a group, the CRPS terms are omitted and the missing peak is penalized through the existence term.

A.4. Chemical-region weights

The main experiments use chemically motivated region weights $w(c) > 0$ to allow diagnostic regions, such as carbonyls and heteroatom-attached carbons, to contribute more strongly to the final risk. A representative weighting scheme is

$$w(c) = \begin{cases} w_{\text{carb}} & \text{if } c = \text{carbonyl}, \\ w_{\text{hetero}} & \text{if } c = \text{heteroatom-attached}, \\ w_{\text{arom}} & \text{if } c = \text{aromatic/olefinic}, \\ w_{\text{aliphatic}} & \text{if } c = \text{aliphatic}. \end{cases}$$

In the simplest setting, we use uniform weights $w(c) \equiv 1$. We also evaluate inverse-frequency weights and carbonyl-heavy weights in the ablations. All weighting schemes yield similar DP4-style top-1 accuracies, suggesting that PRS is not overly sensitive to this choice.

A.5. Proper-scoring interpretation

The Brier score is a strictly proper scoring rule for Bernoulli events, and CRPS is a strictly proper scoring rule for univariate predictive distributions. For a fixed peak grouping and fixed non-negative component weights, PRS combines these scoring-rule components into a decomposable group-wise risk. Under the additional assumption that grouped variables are the forecast targets and the forecast factorizes across groups, the resulting weighted score is proper for the induced grouped representation. In practice, however, peak grouping is data-dependent and the perturbation ensemble is a hand-crafted approximation to experimental uncertainty. We therefore interpret PRS as a proper-scoring-rule-based evaluation metric, rather than as a globally strictly proper score over raw spectra.

In practice, peak grouping is data-dependent and the perturbation ensemble is a hand-crafted approximation to experimental uncertainty. We therefore interpret PRS primarily as a model-agnostic probabilistic evaluation metric, rather than as a fully calibrated likelihood model of NMR experiments. This distinction is important for AI-scientist workflows:

PRS is designed to flag whether simulated evidence supports a generated hypothesis, not to replace full Bayesian modeling of all experimental artefacts.

A.6. Structure-ranking protocol

For all ranking experiments, the same candidate sets are used across PRS, classical shift-error metrics, Wasserstein distance, SM-MMD, and DP4/DP5 probabilities whenever applicable. This ensures that differences in top-1 accuracy reflect differences between scoring metrics rather than differences in the candidate pools.

B. Synthetic Point-Process Benchmark Details

We generate 500 ground-truth event sequences on $[0, T]$ with $T = 200$. Event counts are Poisson-distributed with rate $\lambda = 8$, positions are sampled uniformly, and amplitudes are sampled uniformly from $[0.5, 2.0]$. Four forecasters produce predictions by perturbing the ground truth: GOOD ($\sigma_{\text{pos}} = 0.3$, $p_{\text{drop}} = 0.02$, $p_{\text{add}} = 0.02$), MODERATE ($\sigma_{\text{pos}} = 1.5$, $p_{\text{drop}} = 0.10$, $p_{\text{add}} = 0.05$), POOR ($\sigma_{\text{pos}} = 3.0$, $p_{\text{drop}} = 0.25$, $p_{\text{add}} = 0.15$), and OVER-PREDICTOR ($\sigma_{\text{pos}} = 0.2$, $p_{\text{drop}} = 0.03$, $p_{\text{add}} = 0.50$).

The OVER-PREDICTOR is designed to have low matched-event position error while adding many false-positive events. We compare PRS with $M = 50$ ensemble members against mean $|\Delta t|$, RMSE on matched events, pixelwise signal RMSE, and Wasserstein-1 distance W_1.

C. Additional Results and Ablations

C.1. Metric behaviour and correlations

Across the full dataset, PRS values are concentrated at small risk but with a heavy tail:

$$\text{PRS mean} = 57.39 \pm 189.41, \qquad \text{median} = 9.83.$$

Standard shift-based metrics show typical ^{13}C prediction errors: Avg $|\Delta\text{ppm}| = 0.94 \pm 0.56$, RMSE$= 1.22 \pm 0.66$, and Max $|\Delta\text{ppm}| = 2.46 \pm 1.26$. The region-weighted APE is 0.95 ± 0.57, and the W_1 Wasserstein distance between simulated and experimental shift distributions is 4.50 ± 7.13. The F1 score for peak matching is 0.90 ± 0.16, indicating that most spectra are globally aligned but with non-trivial local mismatches. The SM-MMD similarity is high on average (0.96 ± 0.05), and DP5 probabilities are broad (0.78 ± 0.40).

PRS is strongly correlated with standard shift errors: Spearman $\rho(\text{PRS}, \text{Avg}|\Delta\text{ppm}|) = 0.73$, $\rho(\text{PRS}, \text{RMSE}) = 0.73$, and $\rho(\text{PRS}, \text{Region-wt APE}) = 0.73$. It also shows moderate correlation with W_1 ($\rho = 0.59$) and negative correlation with peak-match F1 ($\rho = -0.55$), as expected for a loss-like quantity. PRS is strongly anti-correlated with SM-MMD similarity ($\rho = -0.67$), meaning that spectra with high similarity under SM-MMD tend to have low PRS. Importantly, PRS is negatively correlated with DP5 and DP4 log-probability ($\rho(\text{PRS}, \text{DP5 prob}) = -0.35$, $\rho(\text{PRS}, \text{DP4 log-prob}) \approx -0.74$), consistent with all three metrics rewarding well-matched simulated/experimental peak patterns.

C.2. Additional ablation details

The main text summarizes the ablation results. Here we only report the additional implementation details and interpretation not already discussed there.

Ensemble size. The ensemble-size sweep uses $M \in \{5, 20, 200\}$ on the same 200-molecule subsample. The main effect of increasing M is reduced variance in the empirical CRPS and existence estimates; beyond moderate ensemble sizes, rankings change little. This supports the use of $M = 50$ in the main experiments as a practical compromise between stability and computational cost.

Perturbation scale and dropout. For the perturbation sweep, we vary

$$\sigma_\omega \in \{0.3, 0.5, 1.0, 2.0, 3.0\},$$
$$p_{\text{drop}} \in \{0.0, 0.05, 0.10, 0.20\}.$$

Small σ_ω values make the forecast distribution overly sharp, while very large values make it overly diffuse. The observed stability across this grid suggests that PRS does not require precise calibration of the hand-crafted perturbation model to remain useful as a ranking metric.

Matching threshold. The threshold sweep uses $\tau \in \{1, 3, 5, 10, 15\}$ ppm. Very small thresholds create many unmatched experimental and simulated peaks, increasing the contribution of missing-peak and false-positive penalties. Once τ is large enough to define chemically reasonable local neighborhoods, the PRS distribution changes only modestly.

Region weights. We evaluate four weighting schemes: chemically motivated weights, uniform weights, inverse-frequency weights, and carbonyl-heavy weights. Since all produce similar ranking accuracy, the choice of $w(c)$ should mainly be viewed as a way to emphasize task-relevant or chemically diagnostic regions rather than as a sensitive tuning parameter.

^{1}H NMR and decoy scaling. The ^{1}H experiment uses the same PRS pipeline as the main ^{13}C setting, without changing the scoring rule. The lower absolute ranking accuracy is expected because ^{1}H spectra contain more overlap, multiplicity effects, and integration ambiguity. The decoy-scaling experiment varies the candidate-set size while keeping the same scoring protocol, providing a harder test of whether

PRS remains informative as the number of plausible alternatives increases.

C.3. Carbonyl quantification details

Carbonyl peaks are analyzed separately because they are sparse and chemically diagnostic. We compare the peak-wise shift term CRPS_ω against Avg $|\Delta\mathrm{ppm}|$ on 987 carbonyl peaks. The two quantities are highly correlated, but CRPS_ω additionally reflects predictive spread and the peak-wise uncertainty model.

This analysis supports the local-diagnostic interpretation of PRS: high-risk carbonyl or heteroatom groups can identify candidate structures that look acceptable under averaged shift errors but fail to explain chemically important experimental evidence.