MOLP2SBENCH: DO LLMS REALLY REASON FROM MOLECULAR PROPERTIES TO STRUCTURE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used in molecular discovery and structure elucidation, where decisions depend on physical observations. Existing evaluations primarily focus on end-to-end molecular reconstruction, where correct predictions do not necessarily indicate that these results are grounded in the supplied properties. To address this, we introduce MolP2SBench, a benchmark for evaluating molecular property-to-structure (P2S) reasoning in LLMs. Each task presents a masked molecular scaffold with given properties, requiring the model to select the missing structure from plausible candidates. To assess whether predictions genuinely rely on the provided properties, we test how models respond when properties are shuffled or replaced, thereby evaluating their reliance on correct property-structure pairing and their responsiveness to counterfactual property changes. We also establish a channel-wise leaderboard using Property-Grounded Recovery (PGR), which gives equal weight to source and counterfactual recovery. In our evaluation of seven models, we find that LLMs can leverage physical properties for molecular structure prediction, but their performance varies substantially across property types. More importantly, higher accuracy with a given property does not necessarily imply greater responsiveness to that property. For example, a model may benefit more from ^{1}H NMR than from ^{13}C NMR, yet be less likely to change its prediction when the ^{1}H data is replaced. Interestingly, LLM performance does not necessarily reflect the relative difficulty of these properties for human chemists. For instance, although UV–Vis spectra provide less direct structural information than NMR spectra, Astra achieves higher accuracy with UV–Vis data than with NMR peaks. In summary, MolP2SBench provides a framework for studying how LLMs extract and use information from different physical properties, revealing performance patterns that may differ from expectations based on conventional chemical intuition and offering a foundation for developing more reliably property-grounded models for molecular and quantum-chemical inference. The code and dataset are available at https://anonymous.4open.science/r/MolP2SBench-7591.

1 INTRODUCTION

LLMs are increasingly used in molecular science, including molecular generation (Cai et al., 2026), molecular discovery (Ye et al., 2025), and structure elucidation (Guo et al., 2024; Su et al., 2025; Noh et al., 2026). In many of these settings, decisions depend on more than recognizing molecular motifs; they require relating physical observations, such as spectra, electronic properties, or molecular geometry, to competing structural hypotheses. Different property channels encode different structural information, and models may not exploit them equally well. Thus, a key question is not simply whether an LLM can produce the correct molecular structure, but whether its decision actually responds to the physical evidence provided.

Existing evaluations (Liu et al., 2026; Bushuiev et al., 2024; Han et al., 2026) do not directly answer this question. End-to-end molecular reconstruction entangles evidence interpretation with structural search and output generation: a model may reach the correct answer because of structural priors, even if the supplied property evidence plays little role in its decision (Geirhos et al., 2020). Aggregate accuracy does not reveal which property channels are actually useful to the model, nor whether

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

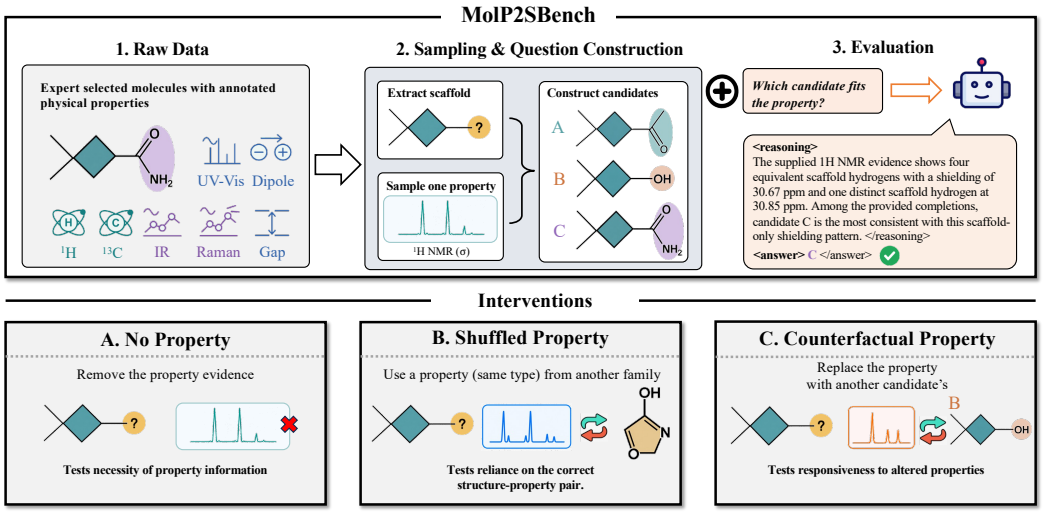


Figure 1: **MolP2SBench: a fixed molecular choice under changing property inputs. Top:** a masked scaffold, candidate SMILES, and physical observations define a local-completion task; the model selects a candidate and provides a short justification. The illustrative ^{1}H input contains calculated shieldings and counts for retained scaffold nuclei. Molecular drawings visualize the text inputs, and the response illustrates the output format. **Bottom:** structure-only input, cross-task shuffling, and same-scaffold counterfactual (CF) property substitution test the benefit of property inputs, pairing dependence, and whether the choice follows replacement property information. The scaffold and candidates remain fixed throughout.

the model would change its prediction if the evidence changed; gains under property-conditioned input therefore do not by themselves establish evidence use (Feng et al., 2019).

To systematically evaluate how well models use the supplied physical properties, we formulate property-to-structure prediction as a multiple-choice masked completion task (Figure 1). Each task presents a masked molecular scaffold with given properties, requiring the model to select the missing structure from plausible candidates. To evaluate how well models use the supplied properties, we apply three controlled intervention schemes. Removing them tests whether the supplied properties confer any benefit beyond the scaffold alone; shuffling them across tasks measures the model's reliance on correct property-structure correspondence; and replacing them with counterfactual (CF) properties from an alternative candidate within the same family assesses the model's responsiveness: whether its selection dynamically adapts to the newly supplied evidence (Gardner et al., 2020). Collectively, these comparisons address not merely whether a model is correct, but whether its decision tracks the physical observations it was given.

Our results show that while LLMs can leverage physical properties for molecular structure prediction, their reliance on this data varies substantially across models and property types. Reliance and responsiveness also show different observed patterns across property channels. For example, Sol has a larger accuracy gain from correctly paired atomwise ^{1}H NMR than from ^{13}C peaks, yet a lower conditional flip rate toward the corresponding CF candidate when the atomwise ^{1}H properties are substituted. Interestingly, model performance does not consistently reflect the relative difficulty of these properties for human chemists. For instance, although UV–Vis spectra provide less direct structural mapping than NMR, models such as Astra achieve higher accuracy with UV–Vis data than with familiar NMR peaks. Beyond these interventions, combined-input analyses reveal model-specific gains and losses: extending the strongest property bundle to all properties improves recovery for Sol and Astra but reduces it for Luna and Opus. Furthermore, in paired NMR comparisons, net accuracy gains often hide significant shifts in model behaviors, as higher-recovery models still fail on sizeable subsets of tasks successfully solved by weaker models. Case studies show that while Astra demonstrates a superior ability to utilize local chemical structures for NMR interpretation, all evaluated models systematically err when global molecular dipoles contradict local polarity heuristics.

In summary, whether a model selects the correct structure, whether its decision is genuinely grounded in the supplied evidence, and whether its stated justification is consistent with the evidence are distinct questions. MolP2SBench disentangles these dimensions through controlled intervention schemes and targeted case studies, exposing variations in property utilization that traditional accuracy metrics fail to reflect. Our contributions are fourfold:

1. A channel-wise recovery leaderboard. We establish a leaderboard using Property-Grounded Recovery (PGR), which balances source recovery with recovery under candidate-targeted counterfactual property substitutions. By averaging counterfactual recovery within each family prior to cross-family aggregation, PGR ensures equal weighting regardless of varying candidate counts. Furthermore, reporting scores independently across property channels reveals model-specific strengths without collapsing heterogeneous properties into a single aggregate ranking.

2. Disentangling reliance and responsiveness. We introduce two complementary metrics to characterize distinct capabilities in how models utilize supplied molecular properties. To measure static reliance, Reliance Margin (RM) quantifies the performance change when the correct property–task pairing is disrupted by shuffling. To evaluate dynamic responsiveness, Counterfactual Flip Rate (CFR) measures whether predictions follow targeted substitutions of the supplied property among cases that are initially solved correctly. The observed channel rankings differ between the two metrics, motivating separate evaluation of reliance and responsiveness.

3. Substitution-ready benchmark construction. We construct MolP2SBench with 136 distinct task families spanning 11 molecular-property channels, yielding 1,496 property-conditioned source tasks paired with matched shuffled controls and candidate-targeted counterfactual variants. Additionally, we analyze within-family property separability and verify that our benchmark produces robust metric estimates and stable model rankings under substantial subsampling.

4. Beyond aggregate accuracy. We show that recovery depends strongly on the supplied properties, and that a higher-recovery model does not simply solve a superset of a weaker model's tasks: across paired comparisons, net gains span 11.0 to 32.4 pp while 40.4% to 45.6% of tasks change outcome in one direction or the other. Model explanations can also conflict with candidate-level physical values, with all four closed-weight models failing more often when molecular dipoles contradict a local carbonyl-based polarity heuristic.

2 Related Work

Table 1 compares MolP2SBench with related molecular benchmarks. We provide a more detailed discussion of the related work in Appendix A.

Table 1: **Related work comparison.** We compare benchmark support for resolving structure from input properties. ✓ marks an evaluated feature; × means it was not reported or not part of the task.

Benchmark	Local editing/ completion	Masked-part reconstruction	Structural distractors	Property-driven answer switch	Explanation checks
FGBench (Liu et al., 2026)	×	×	×	×	✓
ChemCoTBench (Hao et al., 2026)	✓	×	×	×	×
MolLangBench (Cai et al., 2026)	✓	×	×	×	×
ChemIQ (Runcie et al., 2026)	×	×	×	×	✓
SpectrumBench (Yang et al., 2026)	×	×	✓	×	×
MolecularIQ (Bartmann et al., 2026)	×	×	×	×	✓
ChemDIRT (Inae et al., 2026)	×	×	✓	×	×
ChemCoTBench V2 (Guo et al., 2026)	✓	×	×	×	✓
MolP2SBench (ours)	✓	✓	✓	✓	✓

3 Benchmark Design

MolP2SBench formulates property-to-structure prediction as a multiple-choice masked completion task to systematically evaluate how well models use the supplied physical properties. Each task's input is carefully constructed around three core elements: (i) a **molecular scaffold** where property-

critical fragments are masked; (ii) **physical properties** chosen for their high sensitivity to structural changes; (iii) a **candidate family** list containing a unique ground-truth completion. Through controlled intervention on the provided properties, this design explicitly isolates and evaluates two key dimensions of model capability beyond pure reconstruction: **reliance** on correctly paired properties; **responsiveness** to counterfactual properties. Figure 1 summarizes the model-visible inputs and the intervention scheme.

3.1 Problem Formulation

Masking isolates local structural inference from open-ended molecular generation. Reconstructing an entire molecule usually combines evidence interpretation with structural search and output generation. Evaluating such outputs is also ambiguous when a generated structure differs from the reference but exhibits similar properties. We therefore mask property-critical fragments while preserving the remaining scaffold and boundary bond orders. This focuses the task on a localized structural difference within a fixed molecular context.

Property Selection. We select physical properties from three groups that characterize different aspects of molecular structure: electronic properties comprising dipole magnitude, the HOMO–LUMO gap, and UV–Vis observations; vibrational properties comprising IR and Raman spectra; and NMR properties comprising ^{1}H and ^{13}C observations. These groups characterize electronic structure, bond vibrations, and local nuclear environments, respectively. We obtain **11 property channels** by evaluating each electronic and vibrational property separately and using three representations for each NMR type: peaks-only, integration-augmented, and atomwise for ^{1}H; peaks-only, degeneracy-augmented, and atomwise for ^{13}C. NMR inputs provide calculated shielding values for nuclei in the visible scaffold. Representation details appear in Appendix D.

Candidate families make property interventions directly comparable. Each task provides two to four chemically valid options compatible with the scaffold and its attachment constraints. Construction combines three stages: (i) scaffold matching groups molecules with an identical canonical scaffold after masking; (ii) family assembly selects structurally distinct completions, maintains molecule-disjoint tasks, and aligns their electronic, vibrational, and NMR properties; (iii) property-specific filtering excludes a family if any verifiable candidate pair falls below its channel-specific separation threshold.

The final benchmark contains 136 task families and 394 distinct candidate structures. Its 11 property representations provide 1,496 source-property instances, matched shuffled-property controls, and 2,766 eligible within-family CF substitutions. Across 1,000 subsamples of 100 families, aggregate rankings for the diagnostic metrics RM and CFR are retained in 97.7–99.3% of draws (Appendix G.5). Channel-wise PGR yields mean rank correlations of 0.940–0.999 and mean absolute score deviations of 0.91–1.34 pp (Appendix G.6). This organization supports controlled comparisons of property reliance and responsiveness with fixed structural alternatives. Construction and coverage details appear in Appendix B.

3.2 Task Construction

Let q_i, for $i = 1, \ldots, N$, denote a task with visible scaffold G_i^{obs} and ordered candidate family $\mathcal{C}_i = \{G_{i0}, \ldots, G_{i,n_i-1}\}$, where n_i ranges from two to four. The source candidate G_{is_i} is the molecule from which the task is constructed, with source index s_i. For property configuration p, let $\mathbf{P}_{ij}^{(p)}$ denote candidate j's aligned observations. We keep the scaffold and candidate order fixed while varying the supplied properties to construct two types of evaluation tasks. The model receives one property input, not a lookup table of candidate properties, and returns a zero-based candidate index with a concise justification.

Reliance tasks evaluate whether successful selection depends on correctly paired properties. We compare two conditions: **source properties** $\mathbf{P}_{is_i}^{(p)}$, the observations associated with the source candidate; and **shuffled properties** $\mathbf{P}_{\pi(i),s_{\pi(i)}}^{(p)}$, which break the task–property correspondence while preserving the property channel and input format. Here, π assigns each evaluated task a different source from a fixed reference property pool; donor assignments are held constant across channels (Appendix B). Both conditions are evaluated against s_i, and their accuracy difference measures the

benefit of correct property pairing. We denote the corresponding predictions by $\widehat{s}^{(p)}_{i,\mathrm{src}}$ and $\widehat{s}^{(p)}_{i,\mathrm{shuf}}$. We additionally include a **no properties** baseline $P_0 = \varnothing$, shared across configurations, to measure the benefit of supplying properties beyond the structural input alone. Its prediction, $\widehat{s}_{i,0}$, is also evaluated against s_i.

Counterfactual tasks evaluate whether the model's selection is sensitive to counterfactual changes in the provided properties. We replace $\mathbf{P}^{(p)}_{is_i}$ with **counterfactual (CF) properties** $\mathbf{P}^{(p)}_{ij}$ from another candidate G_{ij}, $j \neq s_i$, in the same family, termed the **CF candidate**. Each eligible alternative defines a separate substitution, with prediction $\widehat{s}^{(p)}_{ij,\mathrm{CF}}$ evaluated against j. We assess whether a model switches to the corresponding CF candidate after property substitution.

3.3 Metrics

We use Property-Grounded Recovery as the primary metric, with Reliance Margin and Counterfactual Flip Rate as diagnostic metrics.

3.3.1 Primary metric

Property-Grounded Recovery (PGR) jointly evaluates recovery under the original source property and candidate-targeted counterfactual substitutions, assigning equal weight to source and counterfactual recovery within each task family.

Let $\widehat{s}^{(p)}_{ij}$ denote the prediction when candidate j supplies the property: the source prediction when $j = s_i$, and the CF prediction otherwise. With n_i and s_i defined in Section 3.2,

$$\mathrm{PGR}_p = \frac{50}{|\mathcal{E}_p|} \sum_{i \in \mathcal{E}_p} \left[\mathbf{1}\left(\widehat{s}^{(p)}_{is_i} = s_i\right) + \frac{1}{n_i - 1} \sum_{j \neq s_i} \mathbf{1}\left(\widehat{s}^{(p)}_{ij} = j\right) \right]. \tag{1}$$

Here, $\mathcal{E}_p$ is the input-validated family set shared across all evaluated models for channel p, with source properties and CF properties available for every alternative candidate. Counterfactual correctness is first averaged within each family, ensuring that families with more candidates do not receive greater weight. Source recovery and family-level counterfactual recovery each contribute half of the score. Terminal-invalid outputs count as failures without reducing the denominator. PGR is reported separately for each channel as a percentage, without conditioning on source correctness or averaging across channels.

3.3.2 Diagnostic metrics

Reliance Margin (RM) measures the model's **reliance**: performance gain from providing the correctly paired property relative to a shuffled-property control. A positive RM indicates that the correct condition-target alignment explicitly improves selection accuracy (reported in percentage points in Table 3(a)).

Counterfactual Flip Rate (CFR) measures the model's **responsiveness**: how consistently a model flips to the corresponding CF candidate after property substitution, conditioned on a correct initial prediction. We compute the proportion of successful flips across eligible CF substitutions for these qualifying tasks, then average the results. CFR specifically quantifies responsiveness within each evaluated cohort. Table 3(b) reports this metric as a percentage.

Formal definitions, eligibility criteria, and invalid-response handling for RM and CFR are provided in Appendix G.4.

3.4 Implementation

We align molecular records from QM9 (Ramakrishnan et al., 2014), QM9-NMR (Gupta et al., 2021), and QM9S (Zou et al., 2023). QM9 supplies the HOMO–LUMO gap, QM9-NMR supplies gas-phase NMR shieldings, and QM9S supplies dipole, IR, Raman, and UV–Vis observations.

We evaluate the 11 single-channel inputs defined in Section 3.1. IR and Raman use extracted peak lists, while UV–Vis supplies the first ten excitation states. Final cohort selection uses the property-

specific checks in Appendix B.2. NMR filtering uses integration-augmented ^{1}H and degeneracy-augmented ^{13}C shieldings; the other NMR representations use the same selected families. Appendices B and D document data alignment, observation scopes, units, and preprocessing. Appendix C explains the theoretical background of the properties.

Evaluation follows a closed protocol without tools, retrieval, task-specific training, or solved few-shot examples. Appendix E provides inference settings, scoring procedures, and uncertainty details; Appendix G documents CF coverage and eligibility.

4 Results and Analysis

4.1 Evaluation Setup

We evaluate GPT-5.6-Luna and GPT-5.6-Sol (OpenAI, 2026b), GPT-6-Astra (OpenAI, 2026a), Claude Opus 5 (Anthropic, 2026), Qwen3.8-27B (Qwen Team, 2026), Gemma-4-31B-it (Google DeepMind, 2026), and ChemDFM-v2.0-14B (OpenDFM, 2025), abbreviated Luna, Sol, Astra, Opus, Qwen, Gemma, and ChemDFM. The closed protocol uses no tools, retrieval, task-specific training, or solved few-shot examples. Only input-validated records enter scoring. Appendix E provides model identifiers, inference settings, coverage, and output-handling details. Table 2 reports PGR, while Tables 3(a) and 3(b) report the RM and CFR diagnostics. All three metrics are reported separately by property channel and NMR representation.

4.2 Main Results

Table 2 reports the primary PGR leaderboard for all seven models and eleven property channels. Each channel is evaluated on a common model-independent family set with complete source and CF coverage, without conditioning on initial source correctness.

PGR provides a channel-wise leaderboard of source and counterfactual recovery. Astra has the highest point estimate in ten of the eleven channels, while Opus leads on atomwise ^{13}C NMR (69.3% versus Astra's 68.1%). IR yields the highest PGR for Luna, Sol, Opus, and Astra, at 82.4%, 90.7%, 88.5%, and 93.1%, respectively. Qwen and Gemma also achieve their highest scores with IR (58.2% and 68.6%), whereas their NMR scores remain close to the uniform-random baseline. ChemDFM scores range from 35.6% to 40.1% across channels. These results summarize source and CF recovery jointly; the following diagnostics examine dependence on correct property pairing and conditional responsiveness separately.

We distinguish two aspects of property use: whether correct structure–property pairing improves source recovery (Table 3(a)), and whether replacing those properties redirects an initially correct selection toward the counterfactual candidate (Table 3(b)). Both interventions hold each task's scaffold and structural alternatives fixed.

Table 2: **Property-Grounded Recovery (PGR, %) leaderboard.** Scores give equal weight to source and counterfactual recovery within each family (Eq. 1). Each channel uses the same complete-coverage, input-validated families across all seven models. Gold 👑 and silver 👑 mark the highest (bold) and second-highest model scores in each column. The random baseline is the expected score under uniform candidate selection.

Weights	Model	Electronic			Vibrational		Magnetic resonance					
							^{1}H NMR			^{13}C NMR		
		Dipole	Gap	UV–Vis	IR	Raman	Peaks	+Int.	Atomwise	Peaks	+Deg.	Atomwise
Uniform random		37.9	37.7	37.9	37.9	37.9	37.7	37.7	37.7	37.7	37.7	37.7
Open	Qwen3.8-27B	44.2	35.9	38.6	58.2	49.5	35.5	36.0	38.0	37.7	37.0	39.5
	Gemma-4-31B-it	45.5	37.9	46.0	68.6	55.0	37.8	36.1	37.8	36.9	36.9	38.6
	ChemDFM-v2.0-14B	37.5	37.4	35.6	38.9	38.1	36.9	36.9	40.1	35.6	36.0	36.0
Closed	GPT-5.6-Luna	62.2	47.1	58.4	82.4	66.5	38.4	43.8	40.2	44.9	41.9	47.2
	GPT-5.6-Sol	64.6 👑	69.0 👑	73.2 👑	90.7 👑	72.1 👑	55.2	57.5 👑	55.6	58.1	58.1	56.3
	Claude Opus 5	59.7	55.0	66.1	88.5	71.5	55.4 👑	55.0	58.8 👑	68.0 👑	62.5 👑	**69.3** 👑
	GPT-6-Astra	**69.4** 👑	**72.3** 👑	**80.0** 👑	**93.1** 👑	**74.7** 👑	**70.6** 👑	**71.2** 👑	**64.6** 👑	**68.7** 👑	**69.3** 👑	68.1 👑

Table 3: **Reliance and responsiveness across seven models. (a) RM (pp)** is the accuracy difference between source and shuffled properties; **(b) CFR (%)** is the task-averaged rate of flips to the CF candidate among initially correct sources. Gold 👑 and silver 👑 mark the highest (bold) and second-highest scores in each column, including ties. Metric definitions and evaluated cohorts appear in Appendix G.

Metric	Weights	Model	Electronic			Vibrational		Magnetic resonance					
								^{1}H NMR			^{13}C NMR		
			Dipole	Gap	UV–Vis	IR	Raman	Peaks	+Int.	Atomwise	Peaks	+Deg.	Atomwise
(a) RM (pp)	Open	Qwen3.8-27B	6.7	−6.7	3.7	15.4	15.4	−4.4	4.4	−3.0	4.4	0.0	8.1
		Gemma-4-31B-it	11.9	2.2	4.4	27.9	19.9	0.0	0.0	6.7	−0.7	−2.2	−1.5
		ChemDFM-v2.0-14B	3.7	−3.0	−2.2	2.2	0.7	−2.2	5.2	0.7	0.0	−3.0	−2.2
	Closed	GPT-5.6-Luna	30.4👑	8.1	14.1	39.7	30.1	−0.7	9.6	17.0	7.4	13.3	15.6
		GPT-5.6-Sol	28.1	**33.3**👑	32.6👑	49.3👑	37.5👑	17.8	25.2👑	25.2	12.6	17.8	10.4
		Claude Opus 5	23.7	16.3	28.9	47.1	36.8	25.2👑	13.3	28.9👑	**28.9**👑	**27.4**👑	**37.0**👑
		GPT-6-Astra	**36.3**👑	31.1👑	**41.5**👑	**52.2**👑	**44.9**👑	**36.3**👑	**35.6**👑	**34.8**👑	22.2👑	23.0👑	18.5👑
(b) CFR (%)	Open	Qwen3.8-27B	26.4	11.0	16.7	46.3	33.3	11.6	7.1	14.6	10.3	7.3	12.3
		Gemma-4-31B-it	31.7	6.0	28.7	61.4	34.2	6.2	5.4	6.4	5.8	11.2	5.0
		ChemDFM-v2.0-14B	13.3	7.8	10.9	10.5	10.9	12.6	9.4	14.8	6.0	5.1	4.8
	Closed	GPT-5.6-Luna	66.5	48.2	60.0	80.3	56.2	36.3	32.8	31.9	39.1	26.9	37.2
		GPT-5.6-Sol	73.2👑	63.6👑	72.2👑	90.1👑	63.8👑	54.7	55.4👑	53.6👑	67.3👑	61.1	56.8
		Claude Opus 5	64.8	49.3	61.6	89.8	58.7	56.2👑	50.9	52.9	66.5	65.2👑	68.8👑
		GPT-6-Astra	**74.5**👑	**76.6**👑	**81.1**👑	**91.2**👑	**65.1**👑	**76.7**👑	**77.5**👑	**67.4**👑	**70.7**👑	**70.1**👑	**69.0**👑

Models use physical properties, but performance varies across property channels. All seven models show positive IR RM values, with IR yielding the largest or joint-largest gain for six of them. Within the GPT family, Luna, Sol, and Astra achieve IR gains of 39.7, 49.3, and 52.2 pp, respectively; Opus reaches 47.1 pp, close to Sol, and has the largest gains across the three ^{13}C representations. Under IR substitution, CFR values are 80.3% for Luna and 89.8–91.2% for Sol, Opus, and Astra, evaluated on each model's initially correct sources with valid CF responses. Among open-weight models, Qwen and Gemma also benefit most from vibrational properties, with IR RM values of 15.4 and 27.9 pp, while several NMR views yield near-zero or negative gains. The chemistry-specialized ChemDFM shows more limited effects: its IR RM is 2.2 pp and its CFR is 10.5%. Thus, the results indicate property-dependent selection across model groups, but its strength varies substantially by model and channel.

Reliance and responsiveness show different observed property rankings. Recall that RM quantifies a model's static reliance on correct property-structure pairing, whereas CFR measures its dynamic responsiveness to counterfactual evidence. Luna, Sol, and Astra all show a higher RM for atomwise ^{1}H than for ^{13}C peaks, but a lower CFR. For Sol, the RM values are 25.2 versus 12.6 pp, whereas the CFR values are 53.6% versus 67.3%. CFR is evaluated on each view's initially correct source cohort. This observed reversal motivates reporting reliance and responsiveness separately: RM measures the benefit of correct property pairing, whereas CFR measures responses to targeted property substitutions. Disentangling these dimensions may inform downstream applications, such as human-in-the-loop molecular design, where models must adapt to updated experimental feedback.

LLMs can perform better with properties that are less intuitive to chemists. NMR offers familiar links between spectral features and local chemical environments, whereas UV–Vis provides less direct structural information. Nevertheless, Astra achieves 80.1% source recovery with UV–Vis, compared with 69.9% and 66.2% with ^{1}H and ^{13}C NMR peaks, respectively. Sol shows the same ordering. These observed results highlight UV–Vis as a useful input for molecular selection alongside the more familiar NMR representations.

In summary, these results establish property-dependent selection, while leaving the validity of the supporting chemical argument open. After examining combined-property inputs and the interpretation of model comparisons, we use case studies to examine how models connect the supplied observations to candidate-specific environments and geometry.

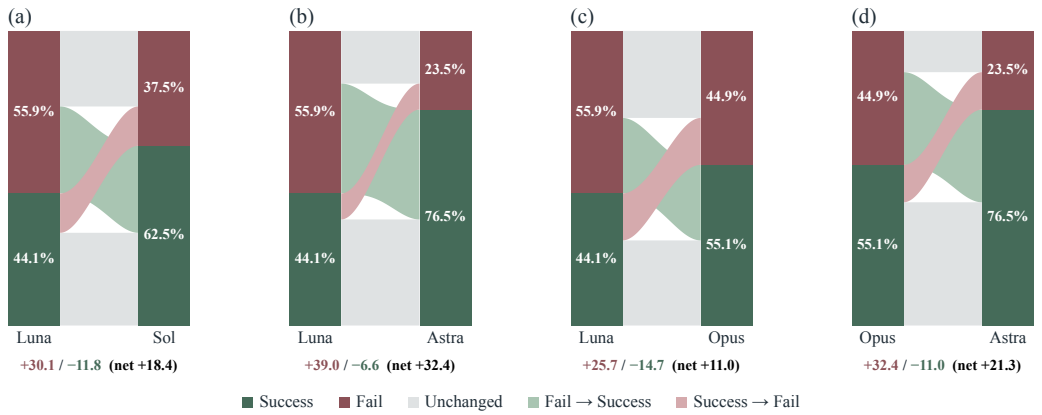


Figure 2: Success/failure transitions across four pairs of closed-weight models under NMR-bundle input: (a) Luna–Sol, (b) Luna–Astra, (c) Luna–Opus, and (d) Opus–Astra. Percentages refer to all tasks; $+/-$ indicate gains/losses from left to right, and net reports the difference in source-property accuracy in pp.

4.3 Combined Property Analysis

We compare single properties, related-property bundles, and all-properties input while keeping constituent values and candidate order matched (Appendix H).

Combining properties does not always improve recovery. Within the GPT family, Sol and Astra improve over the strongest constituent in all three bundles, with larger gains for Astra; for example, electronic recovery rises from 72.8% to 73.5% for Sol and from 80.1% to 88.2% for Astra. Luna instead improves only with IR and Raman (81.6% to 83.8%), while electronic and NMR recovery fall from 65.4% to 61.8% and from 49.3% to 44.1%, respectively. Opus also benefits from electronic and vibrational combinations, but its NMR recovery drops from 66.2% to 55.1%. Thus, combining related channels can help, but does not consistently outperform the strongest individual input.

A successful within-bundle combination does not guarantee a successful cross-bundle combination. Among closed-weight models, adding the remaining properties to the strongest bundle (vibrational for all four) improves Sol from 89.7% to 93.4% and Astra from 96.3% to 98.5%, but reduces Luna from 83.8% to 78.7% and Opus from 90.4% to 89.0%. Thus, Luna and Opus benefit from combining IR and Raman but not from extending that bundle to all properties. Among open-weight models, Qwen and Gemma show the same reversal: both benefit from IR and Raman together, yet recovery falls from 64.0% to 48.5% and from 74.3% to 54.4%, respectively, when the remaining families are added. The single-property–bundle–all-properties progression is therefore not generally monotonic; more comprehensive input need not produce better recovery.

4.4 Do Recovery-Rate Differences Reflect Capability Differences across Models?

Higher recovery does not subsume another model's successes. The paired NMR-bundle results in Figure 2 show that improvements in overall recovery accompany changes in both directions. Sol achieves 62.5% recovery compared with Luna's 44.1%, but 30.1% of all tasks are solved by Sol and not Luna, while 11.8% are solved by Luna and not Sol. The contrast is sharper for Opus: it improves recovery to 55.1%, yet 25.7% of tasks are solved by Opus and not Luna, and 14.7% by Luna and not Opus. Across the four paired NMR comparisons, net gains range from 11.0 to 32.4 pp, whereas the churn rate, defined as the share of all tasks whose success/failure outcome differs between the two models, ranges from 40.4% to 45.6%, so even a large recovery advantage does not amount to a simple expansion of the comparator's solved task set.

Bundle-level advantages do not necessarily extend to single properties. Opus exceeds Astra on ^{13}C peaks-only source recovery (69.1% versus 66.2%) and atomwise source recovery (70.6% versus 66.9%), despite its lower NMR-bundle recovery (55.1% versus 76.5%). This source-recovery

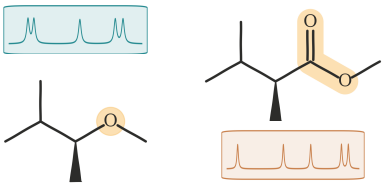


Figure 3: Ether (left) and ester (right). Orange highlights the candidate fragments.

Table 4: Predictions under each ^{13}C input. Checkmarks indicate selection of the property-providing candidate.

Model	Given ester ^{13}C	Given ether ^{13}C
Astra	Ester ✓	Ether ✓
Luna	Ether ×	Ester ×
Sol	Ether ×	Ether ✓
Opus	Ether ×	Ether ✓

advantage does not carry over to unconditional CF accuracy: Astra exceeds Opus with both peaks-only input (65.9% versus 63.6%) and atomwise input (65.1% versus 64.0%; Appendix G, Table 9). Model rankings therefore depend on both the supplied properties and the diagnostic being evaluated.

4.5 Case Study

What Astra's Advantage Reveals about Property Interpretation and Molecular Structure Understanding. Task 97CCF0B132ED0CC7 (Appendix J) contrasts the ether `CO[C@@H](C)C(C)C` with the ester `COC(=O)[C@@H](C)C(C)C`, whose property values are supplied. The retained methine carbon is directly oxygen-bound in the ether but adjacent to the carbonyl carbon in the ester. Given ^{13}C shieldings of 134.3052–168.7333 ppm, Astra selects the ester, arguing that the ether's oxygen-bound methine should be more deshielded and explicitly noting that the masked carbonyl carbon is not observed. Luna and Opus select the ether, invoking the absence or expected contribution of a carbonyl signal; Sol also selects the ether, assigning the two lowest shieldings to two oxygen-bound retained carbons. The distinction therefore requires connecting candidate-specific connectivity to the actual observation scope, not merely recognizing that an ester contains a carbonyl.

When the property input is replaced by the ether's scaffold shieldings, including 101.2638 and 130.2122 ppm, Astra switches to the ether and attributes the stronger deshielding to its oxygen-bound methine and methoxy carbons. Only Astra selects the property-providing candidate in both conditions: Luna reverses the two assignments, while Sol and Opus select the ether throughout.

Models make more errors when property values conflict with local heuristics. For task 49E99A99315E5E96 (Appendix I.3), the carbonyl-containing target has a dipole of 1.45724 D, below the reduced candidate's 5.37590 D. All four models select the reduced candidate under dipole-magnitude input; Astra argues that the additional carbonyl is less consistent with the low magnitude. The reference values contradict this local polarity heuristic. Across 35 within-task matched carbonyl/reduced pairs, ten similarly reverse the expectation that adding a carbonyl increases molecular dipole. Astra makes errors on 7/10 reversal tasks versus 5/25 nonreversal tasks; Luna, Sol, and Opus also have higher error rates in the reversal subset (Appendix I.3). These post-hoc contrasts identify a limitation of the heuristic, not geometry as its isolated cause: the substitutions change electronic structure and conformation together.

5 Conclusion and Limitation

MolP2SBench establishes a controlled framework for evaluating property-to-structure reasoning in LLMs beyond answer accuracy. Across seven models, reliance and responsiveness rank property channels differently, additional properties can reduce recovery, and case studies reveal physical inconsistencies in model explanations. By combining curated completion tasks, a channel-wise recovery leaderboard, and matched property interventions, MolP2SBench provides the community with a reusable testbed for diagnosing evidence use and evaluating future training strategies toward more reliable property-grounded molecular inference.

Limitation. MolP2SBench covers small-molecule completion with computed properties and curated candidates; experimental spectra and open-ended structure elucidation remain outside its scope. The current benchmark's coverage is constrained by the effort required for careful dataset construction and manual validation. Future work could expand molecular families and property channels through LLM-assisted dataset construction while retaining expert review and validation.

AI USE STATEMENT

Large language models were evaluated in the benchmark experiments on molecular property-to-structure reasoning. Models were prompted to select a candidate completion for a masked molecular scaffold and provide a concise justification. Their experimental use, including the task inputs, evaluation conditions, and scoring procedures, is described in the paper.

Generative AI tools also assisted with language editing and manuscript revision, and with discussions of the evaluation protocol and interpretation of results. The authors reviewed the AI-assisted revisions and take full responsibility for the methodology, reported results, and final manuscript.

ETHICS STATEMENT

This work uses public computational molecular datasets for academic research on molecular structure prediction. It does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

The benchmark data and code are available at `https://anonymous.4open.science/r/MolP2SBench-7591`. Appendix B details dataset construction, filtering, property alignment, and the evaluation manifest. Appendix K lists the source molecules, masked scaffolds, and candidate structures for all tasks. Appendix D describes the property representations and preprocessing, while Appendix E records the model identifiers, inference settings, coverage, output-selection rules, and statistical procedures.

REFERENCES

Anthropic. Introducing Claude Opus 5, July 2026. URL `https://www.anthropic.com/news/claude-opus-5`. Official release announcement.

EJ Baerends, OV Gritsenko, and R Van Meer. The kohn–sham gap, the fundamental gap and the optical gap: the physical meaning of occupied and virtual kohn–sham orbital energies. *Physical Chemistry Chemical Physics*, 15(39):16408–16425, 2013.

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. Molecularіq: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. 2026.

Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4):432–444, 2023.

Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. volume 37, pp. 110010–110027, 2024.

Feiyang Cai, Jiahui Bai, Tao Tang, Guijuan He, Joshua Luo, Tianyu Zhu, Srikanth Pilla, Gang Li, Ling Liu, and Feng Luo. Mollangbench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation. In *International Conference on Learning Representations*, volume 2026, pp. 83272–83309, 2026.

Shi Feng, Eric Wallace, and Jordan Lee Boyd-Graber. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5533–5538, 2019.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Google DeepMind. Gemma-4-31B-it. Hugging Face model card, 2026. URL https://huggingface.co/google/gemma-4-31B-it.

Hongyu Guo, Hao Li, He Cao, Gongbo Zhang, and Li Yuan. From answers to states: Verifiable process-level evaluation of chemical reasoning in large language models. *arXiv preprint arXiv:2606.03660*, 2026.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. volume 37, pp. 134721–134746, 2024.

Amit Gupta, Sabyasachi Chakraborty, and Raghunathan Ramakrishnan. Revving up 13c nmr shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Machine Learning: Science and Technology*, 2(3): 035010, 2021.

Taolin Han, Shuang Wu, Jinghang Wang, Yuhao Zhou, Renquan Lv, Bing Zhao, and Wei Hu. Molquest: A benchmark for agentic evaluation of abductive reasoning in chemical structure elucidation. *arXiv preprint arXiv:2603.25253*, 2026.

Li Hao, He Cao, Bin Feng, Daniel Shao, Robert Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. volume 38, 2026.

Eric Inae, Tim Gunn, Chris Bond, and Meng Jiang. Chemdirt: A diversified instruction, representation, and task benchmark for robust chemistry-llm evaluation. *arXiv preprint arXiv:2608.21504*, 2026.

Rahul Kumar Jinger, Heike Fliegl, Radovan Bast, Maria Dimitrova, Susi Lehtola, and Dage Sundholm. Spatial contributions to nuclear magnetic shieldings. *The Journal of Physical chemistry. A*, 125(8):1778, 2021.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. volume 38, 2026.

Omar Mahmood, Elman Mansimov, Richard Bonneau, and Kyunghyun Cho. Masked graph modeling for molecule generation. *Nature communications*, 12(1):3156, 2021.

Heewoong Noh, Namkyeong Lee, Gyoung S Na, Kibum Kim, and Chanyoung Park. Ir-agent: Expert-inspired llm agents for structure elucidation from infrared spectra. In *International Conference on Learning Representations*, volume 2026, pp. 14555–14583, 2026.

OpenAI. GPT-6 Astra: A new generation of intelligence, September 2026a. URL https://openai.com/index/gpt-6-astra/. Official release announcement.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026b. URL https://openai.com/index/gpt-5-6/. Official release announcement.

OpenDFM. ChemDFM-v2.0-14B. Hugging Face model card, 2025. URL https://huggingface.co/OpenDFM/ChemDFM-v2.0-14B.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Daniel Rothchild, Alex Tamkin, Julie Yu, Ujval Misra, and Joseph Gonzalez. C5t5: Controllable generation of organic molecules with transformers. *arXiv preprint arXiv:2108.10307*, 2021.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *Journal of chemical information and modeling*, 66(1):216–227, 2026.

Yunyue Su, Jiahui Chen, Zao Jiang, Zhenyi Zhong, Liang Wang, Qiang Liu, and Zhaoxiang Zhang. Spectrallm: Uncovering the ability of llms for molecular structure elucidation from multi-spectral data. 2025.

Edgar Bright Wilson, John Courtney Decius, Paul C Cross, et al. *Molecular vibrations: the theory of infrared and Raman vibrational spectra*, volume 102. McGraw-Hill New York, 1955.

Chengrui Xiang, Tengfei Ma, Yujie Chen, Tong Wang, Haowen Chen, and Xiangxiang Zeng. Specx: A large-scale benchmark for multi-modal spectroscopy and cross-paradigm evaluation. *arXiv preprint arXiv:2605.18791*, 2026.

Bing Yan, Angelica Chen, and Kyunghyun Cho. Inconsistency of llms in molecular representations. *Digital Discovery*, 4(10):2876–2892, 2025.

Zhuo Yang, Jiaqing Xie, Shuaike Shen, Daolang Wang, Yeyun Chen, Ben Gao, Shuzhou Sun, Biqing Qi, Dongzhan Zhou, Lei Bai, et al. Spectrumworld: Artificial intelligence foundation for spectroscopy. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12655–12666, 2026.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11):957–964, 2023.

SUPPLEMENTARY MATERIAL

Appendix A provides the extended related-work discussion. Appendices B–E document construction, physical background, representations, and the current execution snapshot. Appendices F–H give source-recovery, conditional-transition, and paired-combination results. Appendix I documents verified case records and exploratory geometry diagnostics. Appendix J gives NMR cohort definitions, paired statistics, and selected cases, and Appendix K lists all final-cohort source tasks, candidate structures, and scaffold drawings.

APPENDIX CONTENTS

A Related Work **14**

B Dataset Construction **14**
B.1 Candidate Families and Selection . . . 14
B.2 Property-Specific Filtering . . . 14
B.3 Molecular Identity and Spectral Alignment . . . 15
B.4 Coverage and Intervention Identity . . . 15
B.5 Input Boundary and Reproducible Artifacts . . . 15

C Physical Meaning and Theoretical Background of Molecular Properties **15**
C.1 Electronic structures under the DFT formulation . . . 16
C.2 Electric Dipole Moment . . . 16
C.3 HOMO–LUMO Gap . . . 17
C.4 Proton and Carbon-13 NMR . . . 17
C.5 Infrared and Raman Vibrational Properties . . . 17
C.6 UV–Visible Electronic Excitations . . . 18

D Property Representations **19**
D.1 Within-Family Property Separability . . . 20

E Inference Configuration and Coverage **21**
E.1 Model Identifiers and Protocol . . . 21
E.2 Coverage and Output Selection . . . 21
E.3 Bootstrap Intervals . . . 21

F Additional Source-Recovery and Representation Results **21**

G Property Evaluation: Coverage, Counterfactual Flips, and Metrics **22**
G.1 Conditional Cohorts and Source Weighting . . . 22
G.2 Unconditional CF Accuracy . . . 22
G.3 Comparison of Property Rankings . . . 22
G.4 Metric Definitions . . . 23
G.5 Model-Ranking Stability under Family Subsampling . . . 24
G.6 Channel-Wise PGR Stability under Family Subsampling . . . 24

H Combined Properties **25**
H.1 Matched Input Conditions . . . 26

I Case Records and Property Interpretation **26**
I.1 Verified Residual-Error Records . . . 26
I.2 Reference-Geometry Checks for Electronic and NMR Explanations . . . 26
I.3 Geometry and Heuristic Diagnostics . . . 27

J Detailed NMR Properties and Case Analysis **27**

K Complete Task Families **30**

A Related Work

Table 1 compares the benchmarks discussed below.

Molecular reasoning and structure elucidation. FGBench evaluates functional-group property reasoning (Liu et al., 2026); ChemCoTBench and MolLangBench test molecular operations (Hao et al., 2026; Cai et al., 2026); ChemIQ examines structural reasoning and NMR elucidation (Runcie et al., 2026). Spectrum-to-structure studies address generation, retrieval, and interactive property acquisition (Guo et al., 2024; Su et al., 2025; Bushuiev et al., 2024; Han et al., 2026; Xiang et al., 2026; Yang et al., 2026). MolP2SBench fixes a masked scaffold and complete candidate molecules, then substitutes aligned physical observations to test whether decisions follow the evidence.

Property-conditioned reconstruction. Masked graph modeling reconstructs molecular graphs under property conditions (Mahmood et al., 2021); C5T5 reconstructs corrupted molecular names conditioned on properties (Rothchild et al., 2021); and the Regression Transformer combines property prediction with conditioned generation (Born & Manica, 2023). These approaches develop models through conditional reconstruction. We instead use local completion to evaluate general-purpose language models, constructing candidate-aligned observations that enable controlled property substitutions. Our contribution lies in the benchmark formulation and matched evaluation design, rather than masking itself.

Diagnostic evaluation and verification. MolecularIQ uses graph-derived symbolic verification (Bartmann et al., 2026); consistency studies and ChemDIRT vary molecular representations or equivalent instructions (Yan et al., 2025; Inae et al., 2026); and ChemCoTBench V2 evaluates intermediate chemical states (Guo et al., 2026). MolP2SBench changes the supplied property values while preserving the structural hypotheses: an appropriate response may therefore require changing the answer. We connect these behavioral comparisons to candidate-level checks of physical compatibility and stated explanations, separating correct selection from the validity of the argument supporting it.

B Dataset Construction

B.1 Candidate Families and Selection

The final cohort contains 136 source tasks and 394 distinct candidate SMILES, with 58 two-candidate, 34 three-candidate, and 44 four-candidate tasks. There are 62 single-attachment and 74 two-attachment masks. Hidden connected groups contain one to three heavy atoms. Candidates are validated local completions, not structures generated by the evaluated models.

We group molecules by their canonical masked scaffold and attachment constraints, assemble structurally distinct completions, and select molecule-disjoint families spanning different fragment sizes and structural contrasts. We then align candidate properties and apply the property-specific checks below to determine the final cohort.

B.2 Property-Specific Filtering

For property channel p, let $\mathcal{V}_{ip}$ contain the unordered candidate pairs in family i for which both aligned property records are available. A family is retained when its source-property inputs are complete and every verifiable pair passes the corresponding channel threshold:

$$d_p\left(\mathbf{P}_{ij}^{(p)}, \mathbf{P}_{ik}^{(p)}\right) \geq \tau_p \quad \text{for every } p \text{ and } (j,k) \in \mathcal{V}_{ip}. \tag{2}$$

Thus, a verified below-threshold pair excludes the family.

NMR filtering uses scaffold-only computed shieldings, with integration for ^{1}H and degeneracy for ^{13}C. Wasserstein distances compare normalized empirical distributions. For IR and Raman, peak heights are normalized to unit total mass and matched one-to-one within a 20 cm^{-1} position tolerance; the distance is one minus the maximum matched mass, with each match contributing the smaller of the two peak masses. UV–Vis filtering uses the unweighted energy distribution of the first

Table 5: Property-specific criteria for final cohort selection. Each threshold applies to every verifiable within-family candidate pair.

Property	Distance	Minimum
QM9S dipole norm	Absolute magnitude difference	0.30 D
HOMO–LUMO gap	Absolute difference	0.0005 Hartree
^{1}H NMR	Shielding-distribution Wasserstein distance	0.02 ppm
^{13}C NMR	Shielding-distribution Wasserstein distance	0.10 ppm
IR	Peak-overlap distance	0.05
Raman	Peak-overlap distance	0.05
UV–Vis	Excitation-energy Wasserstein distance	0.08 eV

ten excitation states. The alternative NMR representations are evaluated on the same selected families rather than filtered independently. These channel-specific numerical criteria define the curated cohort.

B.3 Molecular Identity and Spectral Alignment

The mapping uses molecular derivative data to reconstruct IR and Raman profiles, searches their respective full tables independently, and requires agreement on the spectral row. Mapping acceptance requires IR cosine similarity at least 0.99, Raman similarity at least 0.995, and a best-versus-second-best margin of at least 0.001 in each search. File identity, atom counts, geometry checks, and source-file hashes provide additional checks. These thresholds validate identity alignment; they are distinct from the candidate-separation thresholds in Table 5. The observations come from the verified spectral rows, not the reconstructed profiles.

B.4 Coverage and Intervention Identity

All 136 source tasks have the required target-property inputs. Across the 11 evaluated representations, the cohort contains 1,496 source-property instances and 2,766 eligible CF substitutions. Bundle CF records require all constituent channels to correspond to the same source and CF candidate.

All-properties CF input combines the constituent properties of the same CF candidate, with candidate order and CF labels checked across channels.

B.5 Input Boundary and Reproducible Artifacts

Only the public task projection is supplied: record ID, masked scaffold, allowed attachment fields, candidate SMILES, and the selected property information. Source labels, CF-candidate identities, source-data alignment metadata, and internal validation records remain evaluator-only. Models are not given the candidates' full property table. Evaluation identities include dataset version, cohort, view, and record ID.

A fixed cohort manifest records the selected source IDs, candidate families, property-specific criteria, eligible CF IDs, shuffle assignments, and input SHA256 hashes. Validation covers spectrum mappings, molecular identities, task records, candidate order, CF-candidate labels, and model-visible property payloads. All reported statistics must use the selected source IDs and their eligible CF records.

C Physical Meaning and Theoretical Background of Molecular Properties

Our 11 property channels cover seven property types and describe molecular charge distribution, frontier electronic structure, local nuclear environments, and vibrational and electronic excitations. All values are quantum-chemical calculations from QM9 and its extensions, QM9NMR and QM9S (Ramakrishnan et al., 2014; Gupta et al., 2021; Zou et al., 2023). We summarize the underly-

ing density functional theory (DFT) quantities and their structural information below. Appendix D lists source-specific calculation levels and input representations.

C.1 Electronic structures under the DFT formulation

The starting inputs are nuclear charge numbers Z_A, isotope masses M_A, nuclear positions $\mathbf{R}_A$, and the electron number and spin state; A indexes atoms. For the neutral molecules considered here, the electron number is $N_e = \sum_A Z_A$. A density functional and an atomic-orbital basis set specify the computational approximation. At fixed nuclear positions and zero applied field, the nonrelativistic, spin-restricted closed-shell generalized Kohn–Sham equations give

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 + v_{\mathrm{nuc}}(\mathbf{r}) + v_{\mathrm{H}}[n](\mathbf{r}) + \hat{v}_{\mathrm{xc}}\right]\phi_p(\mathbf{r}) = \varepsilon_p\phi_p(\mathbf{r}),$$
$$n(\mathbf{r}) = \sum_p f_p|\phi_p(\mathbf{r})|^2. \tag{3}$$

Here p indexes normalized spatial orbitals ϕ_p, ε_p is an orbital energy, and f_p is its occupation (two for an occupied orbital and zero for an unoccupied orbital), with $\sum_p f_p = N_e$. The variable $\mathbf{r}$ is an electronic position, n is electron density, ∇^2 is the spatial Laplacian, m_e is the electron mass, and $\hbar = h/(2\pi)$ with Planck's constant h. The nuclear-attraction and electron-repulsion (Hartree) potentials are

$$v_{\mathrm{nuc}}(\mathbf{r}) = -\frac{e^2}{4\pi\epsilon_0}\sum_A \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|}, \qquad v_{\mathrm{H}}[n](\mathbf{r}) = \frac{e^2}{4\pi\epsilon_0}\int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}'. \tag{4}$$

Here e is the positive elementary charge, ϵ_0 is vacuum permittivity, $\mathbf{R}_A$ is the position of nucleus A, and $\mathbf{r}'$ is an integration coordinate. The exchange–correlation operator $\hat{v}_{\mathrm{xc}}$ is obtained from the chosen functional; for hybrid DFT it includes nonlocal exchange. Thus, the orbitals and density are solved together, rather than supplied as independent properties.

The DFT electronic energy E_{el} sums orbital kinetic energy, electron–nuclear attraction, Hartree repulsion, and the chosen exchange–correlation energy. Adding nuclear repulsion gives the Born–Oppenheimer potential energy,

$$U(\mathbf{R}) = E_{\mathrm{el}}(\mathbf{R}) + \frac{e^2}{4\pi\epsilon_0}\sum_{A<B}\frac{Z_A Z_B}{|\mathbf{R}_A-\mathbf{R}_B|}, \qquad \nabla_{\mathbf{R}}U(\mathbf{R}^0) = 0. \tag{5}$$

Here $\mathbf{R}$ collects all nuclear coordinates, $A < B$ counts each nuclear pair once, and $\mathbf{R}^0$ is a stationary geometry found by energy minimization; identifying it as a stable minimum additionally requires positive curvature in the vibrational subspace. $\nabla_{\mathbf{R}}$ differentiates with respect to those coordinates. Electronic properties and their derivatives below are evaluated using the appropriate source's optimized geometry and calculation level. The functional and basis are required model choices, so these approximate properties are not determined by masses and universal constants alone.

C.2 Electric Dipole Moment

The electric dipole operator is

$$\hat{\boldsymbol{\mu}} = e\sum_A Z_A\mathbf{R}_A - e\sum_{\ell=1}^{N_e}\hat{\mathbf{r}}_\ell. \tag{6}$$

Here ℓ indexes electrons and $\hat{\mathbf{r}}_\ell$ is their position operator; nuclear charges and positions are defined above. The permanent dipole is the ground-state expectation of this operator and measures the separation of positive and negative charge. Using the electron density from Eq. (3), it is

$$\boldsymbol{\mu} = e\sum_A Z_A\mathbf{R}_A - e\int \mathbf{r}\,n(\mathbf{r})\,d\mathbf{r}, \qquad \mu = \|\boldsymbol{\mu}\|. \tag{7}$$

Here, the vector $\boldsymbol{\mu}$ is the permanent dipole moment, and μ is its magnitude; $\|\cdot\|$ denotes the Euclidean norm. The dipole reflects bond polarity and the geometry-dependent cancellation of bond contributions; even polar bonds can give zero net dipole in a symmetric molecule. It therefore constrains overall polarity, substituent arrangement, and conformation.

C.3 HOMO–LUMO GAP

The frontier-orbital gap is the energy separation between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO):

$$\Delta\varepsilon = \varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}. \tag{8}$$

Here $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$ are the respective orbital energies, and $\Delta\varepsilon$ is their difference. These are the eigenvalues defined in Eq. (3). Their separation characterizes frontier electronic structure and is sensitive to conjugation and electron-donating or electron-withdrawing substituents. Within related molecules, extended conjugation often reduces this gap. It is a method-dependent electronic descriptor, generally distinct from both the fundamental gap (ionization energy minus electron affinity) and a neutral excitation energy (Baerends et al., 2013).

C.4 PROTON AND CARBON-13 NMR

The nuclear magnetic shielding tensor describes how the electronic response to an external magnetic field modifies the field at nucleus i:

$$\mathbf{B}_{\mathrm{loc},i} = (\mathbf{I} - \boldsymbol{\sigma}_i)\mathbf{B}_{\mathrm{ext}}, \qquad \sigma_i^{\mathrm{iso}} = \frac{1}{3}\,\mathrm{Tr}(\boldsymbol{\sigma}_i). \tag{9}$$

Here i indexes nuclei, $\mathbf{B}_{\mathrm{ext}}$ is the external magnetic field, and $\mathbf{B}_{\mathrm{loc},i}$ is the local field at nucleus i. The tensor $\mathbf{I}$ is the identity, $\boldsymbol{\sigma}_i$ is the dimensionless nuclear shielding tensor, and σ_i^{iso} is its isotropic average. The trace Tr sums the three diagonal tensor components. DFT magnetic-response calculations yield $\boldsymbol{\sigma}_i$; multiplying its dimensionless isotropic average by 10^6 gives the numerical value in ppm.

The tensor itself is obtained as a mixed derivative of the electronic energy with respect to magnetic perturbations (Jinger et al., 2021):

$$\sigma_{i,ab} = \left.\frac{\partial^2 E_{\mathrm{el}}(\mathbf{B}_{\mathrm{ext}}, \mathbf{m}_i; \mathbf{R}^0)}{\partial m_{i,a}\,\partial B_{\mathrm{ext},b}}\right|_{\mathbf{B}_{\mathrm{ext}}=\mathbf{m}_i=0}. \tag{10}$$

Here $a, b \in \{x, y, z\}$ are Cartesian components and $\mathbf{m}_i$ is the magnetic moment at nucleus i. The fields perturb the electronic Hamiltonian; differentiating the self-consistent electronic solution gives the magnetic response at fixed geometry. In this formula E_{el} excludes the direct nuclear Zeeman term $-\mathbf{m}_i \cdot \mathbf{B}_{\mathrm{ext}}$, which accounts for the unshielded field. The denominator has units of energy, explaining why $\sigma_{i,ab}$ is dimensionless.

^{1}H NMR. Proton shielding is sensitive to local bonding and electronic effects, including nearby electronegative atoms, aromatic ring currents, and intramolecular hydrogen bonding. It helps distinguish hydrogen environments and substitution patterns, including changes induced at retained hydrogens by the masked group.

^{13}C NMR. Carbon shielding helps distinguish saturated, unsaturated, aromatic, and carbonyl environments, thereby constraining the carbon framework and functional groups. For both nuclei, we expose only retained-scaffold values. Optional equivalence-class counts encode how many retained nuclei share a computed class; they do not encode spin–spin splitting.

C.5 INFRARED AND RAMAN VIBRATIONAL PROPERTIES

Both channels describe normal modes about an optimized geometry. In the harmonic approximation (Wilson et al., 1955), differentiating the potential energy in Eq. (5) gives the Cartesian Hessian, and the isotope masses give the mass matrix:

$$H_{Aa,Bb} = \left.\frac{\partial^2 U}{\partial R_{Aa}\,\partial R_{Bb}}\right|_{\mathbf{R}^0}, \qquad M_{Aa,Bb} = M_A \delta_{AB}\delta_{ab}. \tag{11}$$

Here A, B index atoms, a, b index Cartesian directions, R_{Aa} is a component of $\mathbf{R}_A$, and each Kronecker delta is one when its indices match and zero otherwise. The normal modes follow by diagonalization:

$$\mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}\mathbf{L}_j = \omega_j^2 \mathbf{L}_j, \qquad \widetilde{\nu}_j = \frac{\omega_j}{2\pi c_0}, \tag{12}$$

Here j indexes vibrational modes, $\mathbf{L}_j$ is the dimensionless eigenvector of the mass-weighted Hessian for mode j, ω_j is its angular frequency, $\widetilde{\nu}_j$ is its wavenumber, and c_0 is the speed of light. Each atomic mass appears three times on the diagonal of $\mathbf{M}$, once per Cartesian coordinate; $\mathbf{M}^{-1/2}$ contains their inverse square roots. Frequencies depend on atomic masses, bond stiffness, and coupling between atomic motions.

In a fixed molecular frame, remove overall translation and rotation and choose orthonormal eigenvectors at $\mathbf{R}^0$, $\mathbf{L}_j^{\mathsf{T}}\mathbf{L}_k = \delta_{jk}$, where k is another mode index and T denotes transpose. The mass-weighted normal coordinate is then explicitly

$$\begin{aligned} Q_j &= \mathbf{L}_j^{\mathsf{T}}\mathbf{M}^{1/2}(\mathbf{R}-\mathbf{R}^0) = \sum_{A,a} L_{Aa,j}\sqrt{M_A}\,(R_{Aa}-R^0_{Aa}), \\ \frac{\partial}{\partial Q_j} &= \sum_{A,a}\frac{L_{Aa,j}}{\sqrt{M_A}}\frac{\partial}{\partial R_{Aa}}. \end{aligned} \tag{13}$$

Here $L_{Aa,j}$ is the component of $\mathbf{L}_j$ for atom A and direction a. Thus Q_j is a projection of mass-weighted atomic displacements, with units of square root of mass times length, rather than a new independent input. The second relation uses these fixed reference eigenvectors and converts Cartesian property derivatives to mode derivatives with the other mode coordinates held fixed. These derivatives may be evaluated by analytic response equations or by finite differences of self-consistent calculations at displaced geometries.

Infrared (IR). In the double-harmonic approximation (harmonic nuclear motion and a dipole linear in displacement), the IR intensity of mode j is proportional to the squared dipole derivative along its normal coordinate Q_j:

$$I_j^{\mathrm{IR}} \propto \left\| \left(\frac{\partial \boldsymbol{\mu}}{\partial Q_j}\right)_0 \right\|^2. \tag{14}$$

Here I_j^{IR} is the integrated harmonic IR intensity of mode j. The derivative $\partial\boldsymbol{\mu}/\partial Q_j$ is the change in the permanent dipole per unit normal-coordinate displacement; the subscript 0 denotes evaluation at the equilibrium geometry. IR activity therefore requires a vibration to change the dipole. Band positions constrain functional groups such as carbonyl, O–H, and N–H groups, while the fingerprint region reflects coupled skeletal vibrations.

Raman. Raman activity instead arises from changes in the polarizability tensor $\boldsymbol{\alpha}$, the response of the dipole to an electric field. Its mode derivative must be nonzero, $(\partial\boldsymbol{\alpha}/\partial Q_j)_0 \neq 0$. Again, the subscript 0 denotes equilibrium geometry, and the derivative describes the change in polarizability per unit normal-coordinate displacement. Its isotropic and anisotropic parts determine the Raman activity. Raman often emphasizes symmetric skeletal and conjugated-bond vibrations. Thus, IR and Raman share normal-mode frequencies but provide complementary intensity patterns.

For the usual static, nonresonant approximation, first define the field-dependent dipole and then differentiate it at zero field:

$$\begin{aligned} \mu_a(\mathbf{R},\mathbf{E}) &= -\frac{\partial U(\mathbf{R},\mathbf{E})}{\partial E_a}, \\ \alpha_{ab}(\mathbf{R}) &= \left.\frac{\partial \mu_a(\mathbf{R},\mathbf{E})}{\partial E_b}\right|_{\mathbf{E}=0} = -\left.\frac{\partial^2 U(\mathbf{R},\mathbf{E})}{\partial E_a\,\partial E_b}\right|_{\mathbf{E}=0}. \end{aligned} \tag{15}$$

Here $\mathbf{E}$ is a uniform electric field, E_a and μ_a are Cartesian components, and $U(\mathbf{R},\mathbf{E})$ includes the field interaction of both electrons and nuclei at fixed $\mathbf{R}$. The permanent dipole in Eq. (7) is $\boldsymbol{\mu}(\mathbf{R},\mathbf{0})$; the polarizability derivative is taken before setting the field to zero. These are clamped-nuclei electronic derivatives: the electronic state responds to the field while the nuclei remain fixed. Combining them with Eq. (13) gives the IR dipole derivative and the Raman polarizability derivative; both ultimately depend on the same electronic calculation, nuclear geometry, and masses.

C.6 UV–Visible Electronic Excitations

Time-dependent DFT (TDDFT) describes the electronic response to a weak time-dependent perturbation. Its linear-response equations are built from occupied and unoccupied ground-state orbitals,

their energy differences, and the Hartree and exchange–correlation response kernel (the change of the corresponding potential/operator as the electronic state changes). Solving this response problem yields excitation energies and amplitudes; contraction of the amplitudes with orbital dipole integrals yields transition moments. In electronic-state notation, these quantities are

$$\begin{aligned} \Delta E_s &= E_s - E_0, \qquad \boldsymbol{\mu}_{0s} = \langle \Psi_0 | \hat{\boldsymbol{\mu}} | \Psi_s \rangle, \\ f_{0s} &= \frac{2 m_e \Delta E_s}{3 \hbar^2 e^2} \|\boldsymbol{\mu}_{0s}\|^2. \end{aligned} \tag{16}$$

Here 0 denotes the electronic ground state and s indexes excited states. The quantities E_0 and E_s are their energies at the same nuclear geometry, and ΔE_s is the excitation energy. The states Ψ_0 and Ψ_s are normalized electronic wavefunctions, $\hat{\boldsymbol{\mu}}$ is defined in Eq. (6), and the bracket denotes its transition matrix element $\boldsymbol{\mu}_{0s}$, distinct from the permanent dipole. In Eq. (16), f_{0s} is the dimensionless orientation-averaged oscillator strength of the $0 \to s$ transition and $\|\boldsymbol{\mu}_{0s}\|$ is the transition-dipole magnitude. The transition energies and strengths reflect chromophores, conjugation, and the character of excitations. Their joint pattern provides richer electronic information than a single frontier-orbital gap.

D Property Representations

The property information supplied to a model depends on both the computed source and its conversion into the prompt. The following descriptions specify the reported representations and their preprocessing (Appendix B).

Electronic gap. We use the QM9 HOMO–LUMO gap in Hartree. QM9 reports properties calculated at the B3LYP/6-31G(2df,p) level (Ramakrishnan et al., 2014). This scalar is used as property information for choosing a completion, not as a regression target for the evaluated language model.

Dipole. QM9S provides dipole information in Debye. The norm representation contains its scalar magnitude; the vector representation contains its three Cartesian components. A vector depends on the coordinate frame in which it is represented. We treat norm and vector as different property representations.

NMR. The NMR channel uses gas-phase isotropic shielding from QM9NMR, calculated at the mPW1PW91/6-311+G(2d,p) level (Gupta et al., 2021). Values are shielding constants in ppm, not experimentally referenced chemical shifts. Only nuclei belonging to the retained scaffold are exposed. Symmetry classes are determined in the completed molecule, so a hidden-group change can alter both a retained nucleus's shielding and its equivalence class. The peaks-only view reports class-averaged shieldings. The views labeled integration (^{1}H) or degeneracy (^{13}C) also report retained-nucleus counts for these equivalence classes. These are computed counts, not measured peak areas or spin–spin splitting multiplicities. The raw view contains one value per retained nucleus without internal atom identifiers. These views probe representation sensitivity without changing the molecular candidate set.

IR and Raman. The spectra are derived from QM9S calculations at the B3LYP/def-TZVP level (Zou et al., 2023). The supplied property information is a compact peak list from the broadened spectrum rather than a raster plot or its full 3,501-point grid. Wavenumbers are in cm^{-1} and intensities are relative. Peak extraction uses relative-prominence thresholds of 0.02 for IR and 0.03 for Raman.

UV–Vis. The UV–Vis view uses the first ten QM9S TDDFT transitions, exposing excitation energy in eV, wavelength in nm, and transition-dipole norm in Debye.

Combined properties. The all-properties condition concatenates the seven canonical channels: dipole norm, gap, ^{1}H NMR peaks with retained-nucleus counts, ^{13}C NMR peaks with degeneracy, IR, Raman, and UV–Vis.

D.1 WITHIN-FAMILY PROPERTY SEPARABILITY

Table 6 summarizes the 25 September 2026 analysis of the 136 final-cohort source tasks drawn from release `0.1.0-candidate.15-spectra-r1`, using the released source-property and CF-property inputs. A task meets a channel's separation criterion when the source observations fall outside the similarity cutoff of every distractor with an available released input. The percentage is calculated over tasks with at least one available distractor, not over candidate pairs.

Electronic properties. Dipole magnitude and gap use absolute differences, with separation cutoffs of 0.5 D and 0.1 eV, respectively. Gap values are converted from Hartree using 27.211386245988 eV per Hartree. UV–Vis compares the first ten excitation states in their supplied order. For a source–distractor pair, let d_E and d_μ be the largest absolute differences across these states in excitation energy and transition-dipole norm. The normalized distance is $\max\{d_E/(0.2\,\mathrm{eV}), d_\mu/(0.25\,\mathrm{D})\}$, with a pair inside the similarity cutoff when this distance is at most one. The smallest nearest-distractor distance across the evaluated tasks is 1.94, and the median is 6.49.

Vibrational properties. IR and Raman use the released peak lists. Nonnegative peak intensities are normalized to sum to one within each spectrum. A one-to-one peak assignment maximizes matched mass, with each pair contributing the smaller normalized intensity only when the wavenumber difference is at most $10\,\mathrm{cm}^{-1}$. Let M denote the total matched mass. A distractor is inside the similarity cutoff if $M \geq 0.9$; the distance reported in the table is $1 - M$. Larger distances indicate less overlap of the supplied peak lists.

NMR representations. Shielding observations are treated as unordered multisets. Peaks-only inputs contribute one value per peak; integration- and degeneracy-augmented inputs repeat each value according to its supplied count; atomwise inputs use each listed value once. For equal-sized multisets, a one-to-one assignment minimizes the sum of absolute shielding differences. A pair is inside the similarity cutoff when all assigned differences are at most 0.1 ppm for ^{1}H or 2 ppm for ^{13}C. The pairwise distance is the largest difference under that assignment. Unequal multiset sizes are treated as a mismatch. Distance summaries use tasks with finite nearest-distractor distances.

Table 6: **Channel-wise property separability.** Percentages refer to source tasks separated from every available distractor under the channel-specific reporting cutoffs described above. Distance summaries use the nearest available distractor for each task; Q1–Q3 is the interquartile range. Dipole distances are in D, gap in eV, and NMR in ppm; UV–Vis distances are cutoff-normalized, while IR and Raman use $1 - M$.

Channel	%	Median	Q1–Q3
Dipole magnitude	95.4	1.184	0.806–1.566
HOMO–LUMO gap	88.2	0.912	0.403–1.510
UV–Vis	100.0	6.486	5.309–8.980
IR	100.0	0.713	0.618–0.796
Raman	99.2	0.435	0.312–0.567
^{1}H peaks	94.1	0.808	0.291–1.342
^{1}H peaks + integration	94.1	0.857	0.324–1.377
^{1}H atomwise	99.3	0.922	0.416–1.427
^{13}C peaks	96.3	11.438	6.782–25.128
^{13}C peaks + degeneracy	96.3	11.897	6.829–24.983
^{13}C atomwise	98.5	11.718	6.829–24.983

The final-cohort analysis is implemented in `audits/final136_appendix_sync/recompute.py`, reusing the reporting rules in `audits/current/scripts/analyze_channel_separability.py`. Its archived summary records the channel counts, distance quantiles, dataset version, and SHA256 fingerprints of the source and CF inputs used for the calculation.

E Inference Configuration and Coverage

E.1 Model Identifiers and Protocol

The seven evaluated models are recorded as `gpt-5.6-luna`, `gpt-5.6-sol`, `gpt-6-astra`, `claude-opus-5`, `Qwen/Qwen3.8-27B`, `google/gemma-4-31B-it`, and the local deployment alias `chemdfm-v2:14b-f16`.

No tools or literature retrieval are enabled.

E.2 Coverage and Output Selection

Models return a zero-based candidate index and a concise justification. Source-property accuracy uses all 136 tasks. Paired source–shuffle comparisons use input-validated matched records. Appendix G defines CF eligibility and source weighting.

Scoring selects the first terminal occurrence for the applicable versioned input. API failures can be retried; a terminal invalid answer is retained and scored as incorrect, not replaced by a later convenient response. Ordinary accuracy and unconditional CF accuracy include invalid outputs in their full denominators. CFR instead requires a correct source prediction and valid responses to eligible CF substitutions.

E.3 Bootstrap Intervals

We retain task-level predictions for paired model comparisons. Where reported, 95% percentile intervals use paired source-task bootstrap resamples.

The intervals in Table 12 use 10,000 paired bootstrap resamples of the 136 source-task correctness differences, with seed 20260916 and percentile endpoints. Invalid answers are incorrect. Source tasks are the resampling units.

F Additional Source-Recovery and Representation Results

Table 7 supplies the broader model and representation context. Every reported value uses the same 136 source tasks. NMR values are computed scaffold shieldings; changing their representation changes retained information as well as presentation. The canonical-channel and bundle rows use the same corrected snapshot as the main-text tables; the additional representation rows are recomputed on the same final cohort after exact input replay.

Table 7: Source-property accuracy (%) across seven models. L: Luna; S: Sol; A: Astra; O: Opus; Q: Qwen; G: Gemma; C: ChemDFM. All denominators are 136, including invalid responses.

Property	L	S	A	O	Q	G	C
No properties (P_0)	39.0	42.6	41.2	33.1	33.8	38.2	39.0
Dipole norm	65.4	63.2	69.1	60.3	44.9	44.9	39.0
Dipole vector	62.5	69.1	71.3	61.8	37.5	40.4	38.2
Gap	41.9	71.3	69.1	53.7	36.8	41.2	36.0
^{1}H peaks	37.5	54.4	69.9	55.1	32.4	39.7	39.0
^{1}H +Int.	44.9	61.0	69.9	55.1	35.3	38.2	41.9
^{1}H atomwise	44.1	58.1	65.4	63.2	36.0	38.2	40.4
^{13}C peaks	48.5	54.4	66.2	69.1	34.6	33.8	34.6
^{13}C +Deg.	49.3	55.1	67.6	66.2	36.0	36.0	36.0
^{13}C atomwise	54.4	52.2	66.9	70.6	39.0	36.8	36.0
IR	81.6	89.0	93.4	88.2	55.1	69.9	40.4
Raman	69.1	77.9	80.9	77.2	51.5	61.8	40.4
UV–Vis	53.7	72.8	80.1	67.6	36.0	44.9	36.0
Electronic	61.8	73.5	88.2	68.4	36.0	44.9	40.4
NMR	44.1	62.5	76.5	55.1	37.5	39.0	41.2
Vibrational	83.8	89.7	96.3	90.4	64.0	74.3	40.4
All properties	78.7	93.4	98.5	89.0	48.5	54.4	39.7

G Property Evaluation: Coverage, Counterfactual Flips, and Metrics

G.1 Conditional Cohorts and Source Weighting

Table 3(b) reports Counterfactual Flip Rates (CFR) for all seven models. A source enters only when the model selects it correctly under source properties and has at least one valid response to an eligible CF substitution. After the same input-replay exclusions as in the main evaluation, valid CF responses are averaged within source and then across sources (Eq. 20). Table 8 additionally separates successful CF flips, source persistence, and other choices for selected spectral and combined inputs.

Table 8: Conditional source-macro transitions (%). The CFR column uses Eq. 20; source-persistence and other-choice rates use the same cohort and source weighting with the corresponding outcome indicators. Minor sum differences reflect rounding.

Model	Property	CFR	Persist	Other
Luna	IR	80.3	11.1	8.6
	Raman	56.2	27.9	15.9
	Vibrational	85.0	8.5	6.5
	All	75.8	16.2	8.0
Sol	IR	90.1	6.2	3.6
	Raman	63.8	25.0	11.2
	Vibrational	90.2	5.6	4.2
	All	88.1	7.9	4.0
Astra	IR	91.2	5.1	3.7
	Raman	65.1	23.8	11.1
	Vibrational	92.1	5.7	2.1
	All	97.9	1.7	0.4
Opus	IR	89.8	5.5	4.7
	Raman	58.7	27.7	13.7
	Vibrational	89.3	6.3	4.4
	All	83.6	10.3	6.1

G.2 Unconditional CF Accuracy

Table 9 reports unconditional CF accuracy on the final cohort. Unlike Table 3(b), this pair-micro metric includes initially incorrect sources and counts terminal-invalid CF responses as failures.

G.3 Comparison of Property Rankings

We examine the atomwise ^{1}H versus ^{13}C peaks contrast highlighted in Section 4. Luna, Sol, and Astra each have a larger Reliance Margin (RM) for atomwise ^{1}H, but a lower CFR (Tables 3(a) and 3(b)).

For Sol, RM values are 25.19 and 12.59 pp, respectively, on matched input-validated sources. CFR values are 53.59% and 67.34%, respectively. We draw 100,000 paired bootstrap samples of 136 sources with replacement from the full final-cohort source universe (seed 20260922; NumPy 1.26.4, PCG64; lexicographically sorted source IDs). The two views and both metrics share the same sampled source IDs and multiplicities. Each source retains all its CF records; metric-specific eligibility and denominators are recomputed in every replicate, preserving within-source averaging and source-macro weighting. No replicate has an empty denominator.

We define each difference as atomwise ^{1}H minus ^{13}C peaks. The RM difference is $+12.59$ pp (95% percentile interval: $[-1.50, 27.07]$ pp), and the CFR difference is -13.76 pp ($[-26.88, -0.56]$ pp). Both directions are preserved in 93,319 of 100,000 replicates (93.3%). As a sensitivity check, restricting both conditional rates to the sources initially correct in both views and with valid CF responses yields 60.33% versus 73.33%, a difference of -13.00 pp ($[-28.00, 2.00]$ pp). This common-cohort comparison uses a separate paired source bootstrap with the same seed.

Table 9: **Unconditional CF Accuracy.** CF-candidate selection accuracy (%) under same-scaffold property substitution, without conditioning on source correctness. Eligible, input-validated CF pairs receive equal weight; terminal-invalid responses count as failures. Model identifiers and property definitions follow Table 3(a). Green shading and boldface mark the highest value in each column.

	Electronic			Vibrational		Magnetic Resonance					
						^{1}H NMR			^{13}C NMR		
Model	Dipole	Gap	UV–Vis	IR	Raman	Peaks	+Int.	Atomwise	Peaks	+Deg.	Atomwise
Closed-weight LLMs											
GPT-5.6-Luna	54.2	45.7	60.0	80.4	61.7	33.7	37.6	32.2	38.4	31.0	32.2
GPT-5.6-Sol	60.0	65.9	70.0	90.0	66.7	50.0	46.9	48.4	56.2	53.9	54.3
GPT-6-Astra	**66.7**	**70.2**	**77.9**	**90.4**	**70.8**	**64.7**	**66.3**	**57.8**	**65.9**	**67.4**	**65.1**
Claude Opus 5	52.5	52.7	62.1	86.7	66.2	50.8	50.0	48.4	63.6	58.5	64.0
Open-weight LLMs											
Qwen3.8-27B	37.1	32.2	31.2	53.3	43.8	33.1	31.9	33.5	35.0	33.1	34.5
Gemma-4-31B-it	40.4	30.2	41.7	63.8	45.8	31.8	31.4	32.2	34.9	32.2	34.5
ChemDFM-v2.0-14B	30.4	34.1	32.1	33.3	30.8	31.8	29.1	35.3	34.1	31.8	32.9

G.4 Metric Definitions

We use the task and prediction notation defined in Section 3.2. Let $\mathbf{1}\{\cdot\}$ denote the indicator function. For source-property and shuffled-property inputs, define

$$c_{i,e}^{(p)} = \mathbf{1}\{\widehat{s}_{i,e}^{(p)} = s_i\}, \qquad e \in \{\mathrm{src}, \mathrm{shuf}\}, \tag{17}$$

and let $c_{i,0} = \mathbf{1}\{\widehat{s}_{i,0} = s_i\}$. Terminal-invalid responses count as incorrect in these accuracy calculations.

Reliance Margin (RM). For each comparison, we retain tasks whose source and shuffled inputs both pass input validation and index these matched source tasks by $i = 1, \ldots, N$. Thus, N denotes the number of evaluated source tasks in that comparison, after any input-validation exclusions. RM is

$$\mathrm{RM}^{(p)} = \frac{100}{N} \sum_{i=1}^{N} \left(c_{i,\mathrm{src}}^{(p)} - c_{i,\mathrm{shuf}}^{(p)} \right). \tag{18}$$

Both conditions are scored against the source candidate. Table 3(a) reports RM in pp. Unlike the comparison with P_0, RM retains property inputs in both conditions and tests their correct pairing with the structural task.

Counterfactual Flip Rate (CFR). Let $J_i^{(p)}$ contain the eligible CF candidates whose CF inputs pass validation and whose model responses yield valid candidate indices. Define the initially correct cohort

$$S_p = \{i : \widehat{s}_{i,\mathrm{src}}^{(p)} = s_i,\ |J_i^{(p)}| > 0\}. \tag{19}$$

CFR, expressed as a percentage, is

$$\mathrm{CFR}^{(p)} = \frac{100}{|S_p|} \sum_{i \in S_p} \frac{1}{|J_i^{(p)}|} \sum_{j \in J_i^{(p)}} \mathbf{1}\{\widehat{s}_{ij,\mathrm{CF}}^{(p)} = j\}. \tag{20}$$

We first calculate each task's fraction of valid CF responses selecting the corresponding CF candidate, then average across tasks with equal weight. The source-property response determines cohort inclusion but is not counted as a CF test. Selecting the source or another candidate counts as an unsuccessful flip; an invalid CF response is excluded from this conditional metric. Table 3(b) reports CFR in %. Empty conditional cohorts yield N/A. The initially correct cohort is defined separately for each model and configuration. In contrast, unconditional CF accuracy averages correctness over all input-validated source–CF candidate pairs and includes terminal-invalid responses as failures.

Auxiliary accuracy and baseline gain. Source-property accuracy, also termed source recovery, uses the input-validated source cohort of size N_{src}:

$$A_{\mathrm{src}}^{(p)} = \frac{1}{N_{\mathrm{src}}} \sum_{i=1}^{N_{\mathrm{src}}} c_{i,\mathrm{src}}^{(p)}. \tag{21}$$

The source-recovery tables use all 136 final-cohort tasks, whereas RM uses the source–shuffle input-valid intersection. On the common source/no-properties cohort, which contains the same 136 tasks here, improvement over the baseline P_0 is

$$\Delta_0^{(p)} = \frac{100}{N_{\mathrm{src}}} \sum_{i=1}^{N_{\mathrm{src}}} \left(c_{i,\mathrm{src}}^{(p)} - c_{i,0} \right). \tag{22}$$

Reported in pp, this source recovery gain measures whether source properties improve recovery beyond the scaffold and candidate structures alone. Uniform random candidate selection has expected accuracy $N_{\mathrm{src}}^{-1} \sum_i n_i^{-1} = 37.75\%$ on the final cohort.

G.5 Model-Ranking Stability under Family Subsampling

We assess how aggregate model rankings change when fewer task families are evaluated. For each subset size $n \in \{25, 50, 75, 100, 125\}$, we sample 1,000 subsets without replacement from the final cohort (seed 20260925). Each draw uses the same families across all seven models and retains their source, shuffled, and CF records together. Predictions and property assignments remain fixed; only the evaluated family subset changes.

We aggregate over eleven input configurations: the seven single-property channels (using peaks-only NMR), the electronic, magnetic-resonance, and vibrational bundles, and all properties. For each configuration, RM and CFR follow Appendix G.4; eligible families receive equal weight, with CF responses first averaged within family. Configuration scores are then averaged equally. RM excludes input-mismatched shuffle records, and CFR includes only initially recovered sources with valid, input-matched CF responses. If a sampled configuration has no eligible family for any model, it is omitted for all models in that draw.

Table 10: Model-ranking stability under family subsampling. For each score, ρ is the mean Spearman correlation with its full-cohort model ranking; Exact is the percentage of draws preserving the complete ranking; MAD is the mean absolute score deviation from the full-cohort value, averaged across models and draws, in percentage points.

	RM			CFR		
n	ρ	Exact	MAD	ρ	Exact	MAD
25	0.9845	59.9	3.16	0.9859	63.3	2.94
50	0.9938	82.1	2.01	0.9936	82.2	1.80
75	0.9981	94.5	1.40	0.9969	91.2	1.26
100	0.9998	99.3	0.91	0.9992	97.7	0.80
125	1.0000	100.0	0.45	1.0000	100.0	0.41

The full-cohort ordering is Astra, Sol, Opus, Luna, Gemma, Qwen, and ChemDFM for both aggregate metrics. Table 10 shows that mean rank correlations exceed 0.98 even at the smallest subset size. At $n = 100$, the complete RM and CFR rankings are retained in 99.3% and 97.7% of draws, respectively, with mean absolute deviations below one percentage point. At $n = 125$, both rankings are retained in every draw. These results show that the aggregate model ordering is stable to family subsampling within the evaluated cohort.

G.6 Channel-Wise PGR Stability under Family Subsampling

We evaluate PGR stability separately for the eleven single-property channels in Table 2, using the same complete-coverage, common-input-valid family sets and the score in Eq. 1. We draw 1,000

subsets of 100 families without replacement from the final 136-family cohort, using lexicographically sorted family IDs and NumPy 1.26.4's PCG64 generator with seed 20260925. Each subset is shared across channels and models; for channel p, scoring uses its intersection with $\mathcal{E}_p$. All source and CF outcomes remain paired within family. No predictions are regenerated, and channels are not averaged.

For each channel, Spearman correlation compares the seven-model ranking in each subset with that channel's full-set ranking. Ties receive average ranks. Exact agreement requires the same ranks, including ties. We calculate family scores in integer twelfths before ranking to avoid floating-point tie artifacts. Mean absolute deviation (MAD) averages absolute PGR differences over models and draws. These channel-wise PGR results are separate from the aggregate RM/CFR analysis above.

Table 11: Channel-wise PGR stability over 1,000 subsamples of 100 families. ρ is mean Spearman rank correlation; Exact is complete-ranking agreement (%); MAD is mean absolute score deviation (pp).

Property	ρ	Exact	MAD
Dipole	0.9865	62.4	1.34
Gap	0.9793	52.3	1.04
UV–Vis	0.9989	96.2	1.14
IR	0.9991	97.0	0.91
Raman	0.9887	66.1	1.08
^{1}H peaks	0.9398	20.0	1.10
^{1}H +Int.	0.9562	29.9	1.08
^{1}H atomwise	0.9477	23.5	1.12
^{13}C peaks	0.9711	38.2	1.08
^{13}C +Deg.	0.9627	34.7	1.10
^{13}C atomwise	0.9843	61.1	1.05

Mean rank correlations range from 0.940 to 0.999, with MAD from 0.91 to 1.34 pp. Complete-ranking agreement varies by channel, from 20.0% to 97.0%.

H Combined Properties

The three bundles group related channels: electronic uses dipole norm, gap, and UV–Vis; NMR uses integration-augmented ^{1}H and degeneracy-augmented ^{13}C observations from the visible scaffold; vibrational uses IR and Raman peak lists. All-properties input orders the channels as dipole, gap, ^{1}H NMR, ^{13}C NMR, IR, Raman, and UV–Vis. Table 7 reports source accuracy across input combinations; Table 12 reports the four paired correctness cells for all-properties versus vibrational input, including invalid outputs as incorrect.

A different reference gives a different contrast. Relative to IR alone, all-properties input gains eight tasks and loses one for Astra (126 correct in both and one wrong in both). Luna gains 12 and loses 16, Sol gains 12 and loses six, and Opus gains 11 and loses ten. This distinguishes marginal score changes from turnover in the error set.

Table 12: Paired source correctness when moving from the vibrational bundle to all-properties input. Each row totals 136. Gain/loss refer to all-properties input relative to vibrational input; loss denotes correct-to-incorrect corruption. These are within-model input comparisons, not cross-model transitions. Intervals are descriptive paired-bootstrap 95% intervals in pp.

Model	Both correct	Gain	Loss	Both wrong	Δ [95% interval]
Luna	98	9	16	13	-5.1 [-12.5, 2.2]
Sol	119	8	3	6	3.7 [-0.7, 8.1]
Astra	130	4	1	1	2.2 [-0.7, 5.9]
Opus	115	6	8	7	-1.5 [-6.6, 3.7]

H.1 Matched Input Conditions

Table 7 in Appendix F reports single-property, bundle, and all-properties source accuracy on the same 136 tasks. Constituent property values, masked scaffolds, and candidate order are matched across conditions. The NMR bundle uses integration-augmented ^{1}H and degeneracy-augmented ^{13}C inputs; All-properties input combines electronic, vibrational, and NMR properties using these two NMR representations.

I Case Records and Property Interpretation

Task references link to the corresponding structures and candidate lists in Appendix K.

I.1 Verified Residual-Error Records

The two Astra all-properties source errors are exhaustively documented here. For task DF6B46BD1891D2D1, the masked scaffold is `[*:1]C(C)(C)OC=O`; the supplied candidates, in order, are `CC(C)(OC=O)C(N)=O`, `CC(C)(C)OC=O`, `CC(C)(C=O)OC=O`, and `CC(=O)C(C)(C)OC=O`. The target index is 1; the joint answer is 3 while all three vibrational configurations answer 1. For task A7C3C4816799E0C2, the candidates are `N#CC(CO)CO`, `CC(=O)C(CO)CO`, `COC(CO)CO`, and `OCC(O)CO`. The target is 3; the four spectral/joint configurations all answer 2. These indices and transitions are saved-output observations.

I.2 Reference-Geometry Checks for Electronic and NMR Explanations

These supplementary cases complement Section 4.5 and were selected after inspecting errors. We verified each featured Astra user request against the current public input and each submitted answer against its saved raw response. Candidate geometry was read from the native property source, converted to a molecular graph, and checked against the corresponding candidate SMILES. No new conformer search or quantum calculation was performed. Atom indices below are zero-based in the source geometry file.

Dipole case. For task AE2202D59DBD5246, the masked scaffold is `[*:1]N(C=O)C=O`. Target index 3 is `CN(C=O)C=O` (QM9 `gdb_341`); Astra selects index 0, `O=CN(C=O)C=O` (`gdb_2501`). Reading the native QM9S files confirms dipole magnitudes 0.73125 and 3.46903 D, respectively. The latter conformer's directed C→O axes have pairwise angles $118.5547°$, $14.5898°$, and $133.1442°$, inconsistent with the balanced threefold arrangement invoked in the explanation. These axes diagnose geometry; they are not assumed to be additive electronic dipole contributions. The target–selected-candidate mismatch is directly checked. Luna and Opus choose the same alternative with similar symmetry arguments.

NMR case. For task 252307DC6FE392BF, the scaffold is `[*:1]C1(C(N)=O)CC1`. Target index 2 is `NC(=O)C1(C(N)=O)CC1` (`gdb_54299`); Astra selects index 0, `CC(=O)C1(C(N)=O)CC1` (`gdb_54290`). The target's retained NH shieldings are 22.2110 and 26.9237 ppm, whereas the selected candidate's corresponding values are 27.3063 and 27.3415 ppm. In the target's QM9NMR reference geometry, retained N7–H15 lies 1.8553 Å from O2 of the other carboxamide, at an N–H$\cdots$O angle of $134.93°$. In the selected candidate, the closest relevant amide H16 lies 3.6631 Å from acetyl O2, at $104.83°$. Astra interprets the unequal retained NH shieldings as evidence for hydrogen bonding with the acetyl carbonyl. Thus the proposed short NH$\cdots$acetyl-O contact is unsupported by the selected reference conformer. Luna, Sol, and Opus select the target in this case, so the limitation is not shared by every model.

Representation and scope. The two target NH values above are averaged to 24.56735 ppm with integration two in the symmetry-grouped view. This is graph-class averaging, not a measured conformer-ensemble average. The atomwise view preserves their difference but does not expose atom assignments. Both views omit nuclei in the missing group and contain neither coupling constants nor NOE observations. Dipole and NMR geometry checks use their respective native sources.

Interpretation. These analyses therefore highlight a gap between recognizing local chemical motifs and grounding their consequences in candidate-specific geometry. Graph symmetry need not imply dipole cancellation, and a plausible hydrogen-bond explanation must be supported by the candidate's spatial arrangement. Structural changes also entail electronic changes, and the reviewed explanations suggest that models tend to rely on local electronic heuristics without adequately accounting for the accompanying spatial rearrangements.

I.3 Geometry and Heuristic Diagnostics

These analyses use the full 136-task final cohort. Geometry descriptors were computed without loading model predictions. The geometry analyses use source-matched structures that pass candidate-property checks.

Shape and contact contrasts. Shape contrast is the maximum candidate-pair retained-heavy-scaffold RMSD after proper-rotation Kabsch alignment, minimized over chirality-preserving scaffold automorphisms. Median splits are 0.52281 Å for dipole/UV–Vis and 0.50689 Å for NMR. Astra's lower/upper-half error counts are 18/59 versus 14/58 for dipole, 13/59 versus 11/58 for UV–Vis, 26/66 versus 21/66 for atomwise ^{1}H, and 23/66 versus 21/66 for atomwise ^{13}C. Larger scaffold-shape contrast therefore does not consistently imply more errors.

A contact screen uses retained N/O-bound H, RDKit-defined N/O acceptors, donor–acceptor graph distance of at least three bonds, H$\cdots$A distance at most 2.5 Å, and D–H$\cdots$A angle at least 120°. Contact presence differs across candidates in nine NMR tasks. Atomwise-^{1}H error counts on contrast/noncontrast tasks are 5/9 versus 67/123 (Luna), 4/9 versus 53/123 (Sol), 6/9 versus 41/123 (Astra), and 3/9 versus 47/123 (Opus). Astra atomwise-^{13}C gives 6/9 versus 38/123 on the same tasks. With a 110° cutoff Astra atomwise-^{1}H gives 9/21 versus 38/111; with 130°, 3/4 versus 44/128. Distance cutoffs 2.3–2.7 Å do not change membership here.

Per-model/view linear-probability analyses adjust for log minimum target–distractor property distance, candidate count, retained heavy-atom count, and correctness under P_0; 2,000 task-bootstrap draws use seed 20260916. The atomwise-^{1}H contact coefficient is 0.1016 per descriptor standard deviation, with a nominal 95% interval of [0.0212, 0.1827] from 1,999 full-rank bootstrap fits.

Matched carbonyl/reduced pairs. Deleting one terminal carbonyl oxygen and sanitizing the reduced graph must reproduce the within-task comparison candidate's non-stereochemical graph. This rule yields 35 pairs in 35 tasks, ten with lower dipole in the carbonyl-rich candidate; every reversal exceeds 0.25 D. Under dipole-magnitude source input, reversal/nonreversal error counts are 7/10 versus 8/25 (Luna), 7/10 versus 7/25 (Sol), 7/10 versus 5/25 (Astra), and 8/10 versus 7/25 (Opus). Restricting to targets inside the pair leaves nine reversal and seventeen nonreversal tasks: errors are 7/9 versus 3/17, 7/9 versus 4/17, 7/9 versus 2/17, and 8/9 versus 2/17, respectively. These selected substitution contrasts challenge a monotonic polarity heuristic.

Task 49E99A99315E5E96. Candidate 0, `N#CCn1cnnc1`, has dipole 5.37590 D; target candidate 1, `N#CC(=O)n1cnnc1`, has 1.45724 D. All four models select candidate 0 under norm input. Astra argues that the additional carbonyl makes the low magnitude less consistent, contradicting the actual ordering. Only Astra switches to the target under vector input, while computing its magnitude and acknowledging frame dependence.

Norm/vector diagnostic. On 136 matched tasks, norm/vector correct counts are 89/85 (Luna), 86/94 (Sol), 94/97 (Astra), and 82/84 (Opus). Wrong-to-correct/correct-to-wrong counts are 11/15, 12/4, 7/4, and 13/11; invalid-inclusive exact McNemar p-values are 0.557, 0.077, 0.549, and 0.839. Opus has one invalid vector response and no invalid norm response in this cohort; its valid matched pairs give 82/84 correct and 13/11 gains/losses. Vector norms match scalar inputs within 10^{-10} D. In the ten reviewed reversal cases Astra reduces vectors to magnitude and notes frame dependence.

J Detailed NMR Properties and Case Analysis

NMR: restricted observations and errors in interpreting available property information. The task references below link to their source molecules, candidate lists, and scaffold drawings in Ap-

pendix K. Across several models, the supplied NMR representations yield lower source recovery than other informative property channels. For example, Astra obtains 69.9% with ^{1}H peaks and 66.2% with ^{13}C peaks, compared with 80.1% with UV–Vis; Sol obtains 54.4% for both NMR views, compared with 72.8% with UV–Vis. This ordering is not universal: Opus performs better with ^{13}C peaks than with UV–Vis, and Astra's ^{1}H result is similar to its HOMO–LUMO-gap result. The relevant question is therefore why the particular NMR property information supplied by this benchmark leaves substantial room for error, rather than whether NMR is intrinsically less informative than other molecular properties.

An important feature of the task is that the NMR observation is restricted to nuclei retained on the visible scaffold (Section 4.5). Nuclei within the masked group, including its attached hydrogens, are excluded. The model consequently has to infer a completion from its effect on the shielding of the remaining nuclei, without direct access to that group's own NMR signals. This restriction differs from the whole-molecule scope of the other property channels. For example, if a carbonyl carbon belongs to the masked group, its own ^{13}C shielding is unavailable even when the candidate structure contains a carbonyl. The observation also consists of calculated gas-phase isotropic shieldings (Gupta et al., 2021), with no experimental multiplet patterns, J-couplings, or atom assignments. Applying a familiar whole-spectrum recognition strategy therefore requires care: the model must determine which expected signals are actually within the supplied scope.

Removing hidden-group nuclei eliminates potential discriminative cues, but it does not necessarily eliminate all information about the completion. A concrete example of a difficult residual distinction is Task 79CD04C3DFC1C559.[1] The retained scaffold is `[*:1]C(=O)NC1CCC1`, and the candidates are a methyl carbamate, `COC(=O)NC1CCC1`, and a urea, `NC(=O)NC1CCC1`. The methoxy carbon of the carbamate lies outside the observed scaffold. The remaining ^{13}C peak lists are numerically close: the retained-carbonyl shieldings are 28.2095 and 28.0280 ppm, and the Wasserstein-1 (W_1) distance between the two scaffold distributions is only 0.3562 ppm. All four closed models select the urea when the property information comes from the carbamate. Astra invokes a more deshielded urea carbonyl, while other explanations also rely on general carbonyl-environment arguments. These responses illustrate how broad chemical trends can fail to resolve a fine quantitative distinction after direct substituent signals have been removed. The two computed reference distributions are different.

At the dataset level, candidate separation is associated with model performance. When tasks are grouped using the final cohort's one-third and two-thirds quantiles of the nearest-alternative W_1 distance (lower includes the first cutoff; middle includes the second), Astra's ^{1}H accuracy increases from 47.8% to 75.6% to 86.7%; its ^{13}C accuracy increases from 34.8% to 73.3% to 91.1%. These patterns are consistent with numerical proximity contributing to difficulty. Candidate count and chemical composition also differ between groups, however, so the relationship is descriptive. Moreover, all 136 targets have distinct reference distributions at zero tolerance in each NMR representation.

The submitted explanations identify additional problems that cannot be reduced to missing signals. Models sometimes misinterpret the supplied quantity or reason about nuclei that the observation explicitly excludes. Task E1413048E6C74A5A is particularly instructive. All four models correctly select the ester `CC(C)COC(=O)CO`, but Luna cites a "strong downfield 13C shielding peak at 167.54 ppm" as support for an ester-carbonyl assignment. The carbonyl is inside the masked group and is absent from the observed list. The explanation also applies a carbonyl-region chemical-shift interpretation to an absolute shielding value. For scale, using the matching calculated gas-phase TMS shielding of 186.9704 ppm gives an approximate referenced shift of $186.9704 - 167.54025 \approx 19.43$ ppm, rather than 167.54 ppm.[2] In contrast, Astra, Sol, and Opus discuss the two relatively low-shielding environments of retained oxygen-bound carbons. Because the final choice is correct, this example does not itself contribute to the accuracy deficit; it shows

[1] The five NMR tasks discussed here belong to an outcome-stratified packet of 24 task–view cases: three cases from each of four strata (Astra alone correct, all four correct, Astra incorrect but another model correct, and all four incorrect), separately for ^{1}H and ^{13}C peaks-only inputs. Selection within strata used SHA256 ordering with a fixed seed.

[2] The matching calculated gas-phase ^{13}C TMS reference is tabulated with the QM9NMR dataset at `https://moldis-group.github.io/qm9nmr/`. This conversion illustrates the shielding/shift distinction; it is not an experimental peak assignment.

that source recovery can conceal an incorrect account of the property information. The ester/ether comparison below provides a corresponding case in which a scope error accompanies an incorrect choice.

The case records document errors in interpreting the physical meaning of supplied shieldings, their association with local environments, and their observation boundary.

Matched property information and local structural comparisons. Astra's advantage is clearest for ^{1}H peak lists. It reaches 69.9% source recovery, compared with 54.4% for Sol, 55.1% for Opus, and 37.5% for Luna. Its advantage over Sol comprises 31 tasks solved only by Astra and ten solved only by Sol, a net difference of 15.4 pp with a descriptive paired 95% bootstrap interval of [6.6, 24.3].[3] Property substitutions provide additional support for a difference in property use. Among the source tasks initially solved correctly by both Astra and Sol and with valid CF substitutions, Astra selects the corresponding CF candidate in 76.6% of source-averaged outcomes and Sol in 56.0%. CF outcomes are averaged within each source before averaging across the common source cohort. Furthermore, all candidates have the same peak count in 94.9% of tasks; within this subset, Astra still obtains 68.2% source recovery, compared with 52.7% for Sol and 53.5% for Opus. The advantage therefore persists when simple peak-count discrimination is unavailable.

Task 97CCF0B132ED0CC7 illustrates how an accurate account of candidate connectivity and observation scope can support Astra's decision. The candidates are the ether `CO[C@@H](C)C(C)C` and the ester `COC(=O)[C@@H](C)C(C)C`, with the ester as the source. The retained methine is directly bonded to oxygen in the ether, but adjacent to a carbonyl carbon in the ester. The supplied ^{13}C shieldings range from 134.3052 to 168.7333 ppm. Astra selects the ester, explaining that the ether would more strongly deshield the directly oxygen-bound methine and explicitly noting that the masked carbonyl carbon is excluded from the observation. Luna and Opus instead select the ether and invoke the absent carbonyl signal or its expected contribution to the peak pattern. Sol also selects the ether, interpreting the two lowest supplied shieldings as two directly oxygen-bound retained carbons. The decisive distinction is therefore not merely recognizing that an ester contains a carbonyl; it is relating the candidate-specific bonding of the retained atoms to the property information that can actually be observed.

The same case also supplies a behavioral check on Astra's explanation. When the input is replaced with the ether candidate's scaffold property information, which includes shieldings at 101.2638 and 130.2122 ppm, Astra switches to the ether and relates the more strongly deshielded pattern to the oxygen-bound methine and methoxy carbon. Among the four models, Astra alone selects the property-providing candidate in both conditions: Luna reverses the two assignments, while Sol and Opus select the ether under both inputs. These are separate calls with substituted property information, not a conversational correction. This paired outcome strengthens the interpretation that Astra is responding to a structurally relevant change in the supplied property information.

Task 9FCBD2D389E1B5FA provides a complementary ^{1}H example. The source is `CO[C@@H](C=O)C1CC1`, and the alternative is `CC(=O)[C@@H](C=O)C1CC1`. Both retain the aldehyde and cyclopropyl portions of the scaffold, while the retained methyl is oxygen-bound in the source and carbonyl-adjacent in the alternative. Astra associates the observed peak at 28.2110 ppm with the oxygen-bound methyl and selects the source. Opus assigns this peak to a methine and selects the carbonyl-linked alternative; Luna associates the much lower 21.5655 ppm peak with deshielding of the chiral proton, whereas Astra identifies an aldehydic contribution. An independent check using the existing counted representation shows that the 28.2110 ppm peak represents three retained hydrogens, contradicting Opus's methine assignment. These counts were not supplied in the peaks-only condition and are used here only to evaluate the explanation. When the alternative candidate's property information is substituted, Astra switches to that candidate and relates the methyl/methine region near 29–30 ppm to the carbonyl-adjacent environments. Again, Astra alone selects the property-providing candidate in both conditions among the four models. Candidate reference values, W_1 distances, and the integration check used in this discussion are evaluator-side information, not a candidate-property lookup supplied to the models.

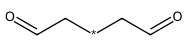
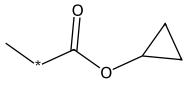
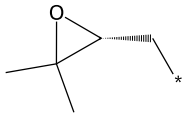
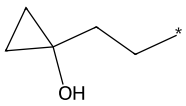
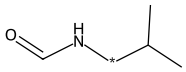
[3] This additional NMR interval uses 10,000 task-level bootstrap resamples with seed 20260917. It is post-hoc and pointwise, conditional on the saved runs, without multiple-comparison correction or repeated-generation uncertainty.

These cases suggest that Astra's stronger ^{1}H performance involves more accurate coordination of NMR interpretation with local structural context. Its successful explanations identify which atoms remain observable, distinguish direct oxygen attachment from carbonyl adjacency, and associate numerical patterns with candidate-specific environments. These outputs provide checkable correspondences between the supplied property information, candidate structures, and stated explanations. Sol and Opus correctly use related distinctions in other cases. For example, in Task 894BBF79C573C29A, they correctly select an isopropyl carbamate, while Astra selects the amide alternative and gives a different NH/methine assignment. The overall model ranking is also nucleus-dependent: Opus exceeds Astra in the ^{13}C peaks-only and atomwise conditions. The aggregate results establish higher ^{1}H peaks-only source recovery for Astra, while the selected cases document both property-consistent local comparisons and explanation errors.

K Complete Task Families

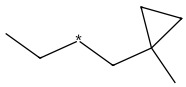
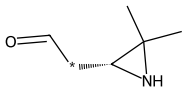
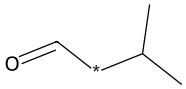
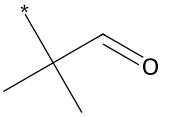
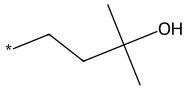
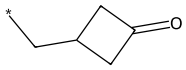
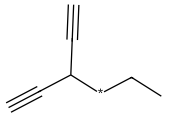
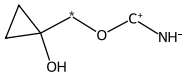
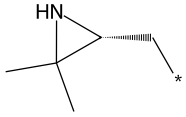
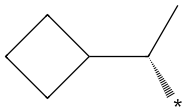
This section lists all 136 source tasks and their 394 candidate entries. For each task, we provide the source-molecule SMILES, the masked scaffold as SMILES and a structural drawing, and the complete candidate list in its original order. Tasks are identified by the same 16-character IDs used throughout the paper; task references link directly to the corresponding entries below. Candidate indices are zero-based, as in the task inputs; a superscript asterisk marks the source candidate. In the scaffold drawings, $*$ denotes the masked fragment. SMILES strings, including stereochemical and charge annotations, are reproduced unchanged from the task records.

Table 13: Complete task families: source molecules, masked scaffolds, and candidate SMILES.

Task and SMILES	Candidate list	Scaffold
Task 005A3B23AB24FFB0 Source: `O=CCCCC=O` Scaffold: `[*:1](CC=O)CC=O`	0*: `O=CCCCC=O` 1: `O=CCOC(=O)CC=O`	
Task 0293945FD5430B3D Source: `CNC(=O)OC1CC1` Scaffold: `[*:1](C)C(=O)OC1CC1`	0: `CC(=O)C(=O)OC1CC1` 1*: `CNC(=O)OC1CC1`	
Task 043694CC3A98D45D Source: `CC[C@H]1OC1(C)C` Scaffold: `[*:1]C[C@H]1OC1(C)C`	0: `COC[C@H]1OC1(C)C` 1: `CC(=O)C[C@H]1OC1(C)C` 2*: `CC[C@H]1OC1(C)C`	
Task 0556D16AF07422B2 Source: `OCCC1(O)CC1` Scaffold: `[*:1]CCC1(O)CC1`	0: `NC(=O)CCC1(O)CC1` 1: `N#CCCC1(O)CC1` 2*: `OCCC1(O)CC1` 3: `CCCC1(O)CC1`	
Task 0D39B8E892B8D428 Source: `CC(C)OC(=O)NC=O` Scaffold: `[*:1](NC=O)C(C)C`	0*: `CC(C)OC(=O)NC=O` 1: `CC(C)CNC=O` 2: `CC(C)C(=O)NC=O`	

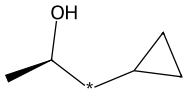
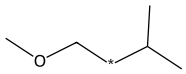
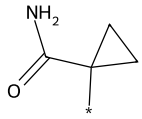
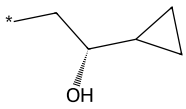
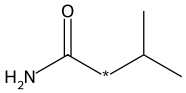
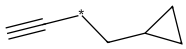
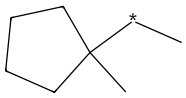
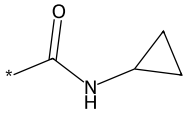
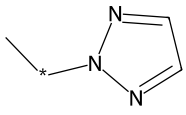
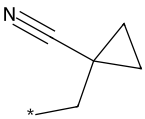
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 0D5F32ECEAFB257A Source: CCOCC1(C)CC1 Scaffold: [*:1](CC)CC1(C)CC1	0: CCC(=O)CC1(C)CC1 1*: CCOCC1(C)CC1	
Task 0F468D9186495EFA Source: CC1(C)N[C@H]1C(=O)C=O Scaffold: [*:1](C=O)[C@H]1NC1(C)C	0: CC1(C)N[C@H]1CC=O 1*: CC1(C)N[C@H]1C(=O)C=O	
Task 105A590726D5A52A Source: CC(C)OC(=O)C=O Scaffold: [*:1](C=O)C(C)C	0: CC(C)C(=O)C=O 1: CC(C)CC=O 2: CC(C)NC=O 3*: CC(C)OC(=O)C=O	
Task 107BFFB67D464E36 Source: CC(=O)C(C)(C)C=O Scaffold: [*:1]C(C)(C)C=O	0: COC(C)(C)C=O 1: CC(C)(C=O)C(N)=O 2*: CC(=O)C(C)(C)C=O	
Task 10C5E7E398F5B90E Source: CCCC(C)(C)O Scaffold: [*:1]CCC(C)(C)O	0: CC(C)(O)CCC=O 1*: CCCC(C)(C)O 2: CC(C)(O)CCC(N)=O 3: CC(C)(O)CCO	
Task 10D912DA9FC1655C Source: N#CCC1CC(=O)C1 Scaffold: [*:1]CC1CC(=O)C1	0: O=CCC1CC(=O)C1 1: O=C1CC(CO)C1 2: CC(=O)CC1CC(=O)C1 3*: N#CCC1CC(=O)C1	
Task 139B40967BE0655C Source: C#CC(C#C)OCC Scaffold: [*:1](CC)C(C#C)C#C	0*: C#CC(C#C)OCC 1: C#CC(C#C)C(=O)CC	
Task 1A28585E9E7612A7 Source: [NH-][CH+]OC(=O)C1(O)CC1 Scaffold: [*:1](O[CH+][NH-])C1(O)CC1	0*: [NH-][CH+]OC(=O)C1(O)CC1 1: [NH-][CH+]OCC1(O)CC1	
Task 1AE0AF63F033E4A9 Source: COC[C@H]1NC1(C)C Scaffold: [*:1]C[C@H]1NC1(C)C	0*: COC[C@H]1NC1(C)C 1: CC(=O)C[C@H]1NC1(C)C	
Task 209A0A1292E6922F Source: CC(=O)[C@@H](C)C1CCC1 Scaffold: [*:1][C@@H](C)C1CCC1	0: C[C@@H](O)C1CCC1 1*: CC(=O)[C@@H](C)C1CCC1 2: C[C@@H](C#N)C1CCC1	

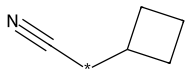
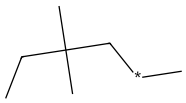
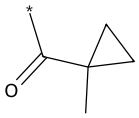
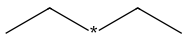
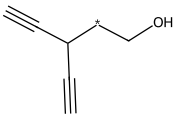
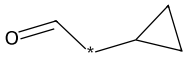
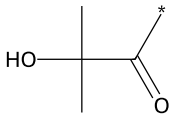
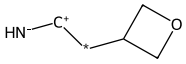
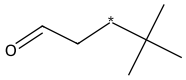
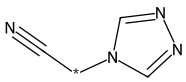
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 230A5DC9D459F8DE Source: C[C@@H](O)CC1CC1 Scaffold: [*:1](C1CC1)[C@@H](C)O	0: C[C@@H](O)C(=O)OC1CC1 1*: C[C@@H](O)CC1CC1	
Task 246F5B26DFAA12E2 Source: COCC(=O)C(C)C Scaffold: [*:1](COC)C(C)C	0: COCCC(C)C 1: COCC(=O)OC(C)C 2*: COCC(=O)C(C)C	
Task 252307DC6FE392BF Source: NC(=O)C1(C(N)=O)CC1 Scaffold: [*:1]C1(C(N)=O)CC1	0: CC(=O)C1(C(N)=O)CC1 1: NC(=O)C1(O)CC1 2*: NC(=O)C1(C(N)=O)CC1 3: N#CC1(C(N)=O)CC1	
Task 2773120914D0BE17 Source: OC[C@@H](O)C1CC1 Scaffold: [*:1]C[C@@H](O)C1CC1	0*: OC[C@@H](O)C1CC1 1: CC(=O)C[C@@H](O)C1CC1 2: COC[C@@H](O)C1CC1	
Task 2B36F250BC153F3E Source: CC(C)OC(=O)C(N)=O Scaffold: [*:1](C(N)=O)C(C)C	0: CC(C)NC(N)=O 1*: CC(C)OC(=O)C(N)=O 2: CC(C)C(=O)C(N)=O	
Task 2CD421DF3B64F435 Source: C#CC(=O)OCC1CC1 Scaffold: [*:1](C#C)CC1CC1	0*: C#CC(=O)OCC1CC1 1: C#CC(=O)CC1CC1	
Task 2F119352D7894AE6 Source: COC1(C)CCCC1 Scaffold: [*:1](C)C1(C)CCCC1	0: CC(=O)C1(C)CCCC1 1*: COC1(C)CCCC1	
Task 39C4D5D6A9DB1877 Source: CC(=O)NC1CC1 Scaffold: [*:1]C(=O)NC1CC1	0: COC(=O)NC1CC1 1*: CC(=O)NC1CC1 2: NC(=O)C(=O)NC1CC1 3: CC(=O)C(=O)NC1CC1	
Task 3A7FC857E2E5511F Source: CCn1nccn1 Scaffold: [*:1](C)n1nccn1	0: COC(=O)n1nccn1 1: CC(=O)n1nccn1 2*: CCn1nccn1	
Task 3AEB8B1B49C28495 Source: CC(=O)CC1(C#N)CC1 Scaffold: [*:1]CC1(C#N)CC1	0*: CC(=O)CC1(C#N)CC1 1: CCC1(C#N)CC1 2: COCC1(C#N)CC1 3: N#CC1(CC(N)=O)CC1	

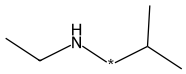
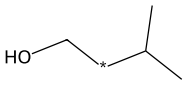
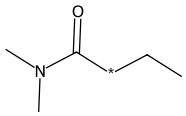
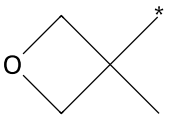
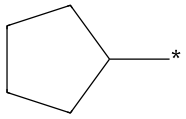
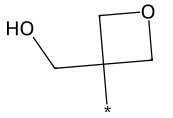
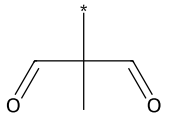
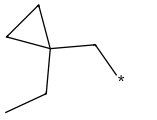
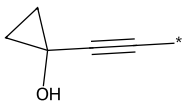
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 3B32A784772837ED Source: `N#CC(=O)C1CCC1` Scaffold: `[*:1](C#N)C1CCC1`	0*: `N#CC(=O)C1CCC1` 1: `N#CC(=O)OC1CCC1`	
Task 3C48CDD9CF8488CE Source: `CCC(C)(C)COC` Scaffold: `[*:1]C(C)(C)CC`	0: `CCC(C)(C)CC(C)=O` 1*: `CCC(C)(C)COC`	
Task 407DADE8EF2DD3A2 Source: `COC(=O)C1(C)CC1` Scaffold: `[*:1]C(=O)C1(C)CC1`	0*: `COC(=O)C1(C)CC1` 1: `CC1(C(=O)C(N)=O)CC1` 2: `CC1(C(=O)C#N)CC1` 3: `CC1(C(N)=O)CC1`	
Task 43E2D244483F435E Source: `CCOC(=O)CC` Scaffold: `[*:1](CC)CC`	0*: `CCOC(=O)CC` 1: `CCC(=O)CC` 2: `CCOCC`	
Task 446C8D96D87CEC39 Source: `C#CC(C#C)CCO` Scaffold: `[*:1](CO)C(C#C)C#C`	0*: `C#CC(C#C)CCO` 1: `C#CC(C#C)C(=O)CO`	
Task 45356781186D81E9 Source: `O=CNC1CC1` Scaffold: `[*:1](C=O)C1CC1`	0: `O=CC(=O)C1CC1` 1*: `O=CNC1CC1` 2: `O=CC(=O)OC1CC1` 3: `O=CCC1CC1`	
Task 45545E30575A679F Source: `CC(C)(O)C(=O)C=O` Scaffold: `[*:1]C(=O)C(C)(C)O`	0: `CC(C)(O)C(=O)C(N)=O` 1: `CC(C)(O)C(=O)C#N` 2: `CC(C)(O)C(N)=O` 3*: `CC(C)(O)C(=O)C=O`	
Task 465D23908FEDE6E8 Source: `[NH-][CH+]OC(=O)C1COC1` Scaffold: `[*:1]([CH+][NH-])C1COC1`	0*: `[NH-][CH+]OC(=O)C1COC1` 1: `[NH-][CH+]OC1COC1`	
Task 47B9859ADE985B3D Source: `CC(C)(C)OCC=O` Scaffold: `[*:1](CC=O)C(C)(C)C`	0*: `CC(C)(C)OCC=O` 1: `CC(C)(C)C(=O)CC=O`	
Task 49E99A99315E5E96 Source: `N#CC(=O)n1cnnc1` Scaffold: `[*:1](C#N)n1cnnc1`	0: `N#CCn1cnnc1` 1*: `N#CC(=O)n1cnnc1`	

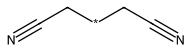
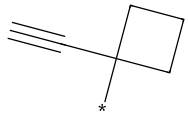
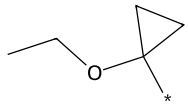
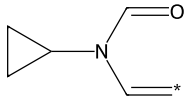
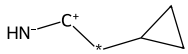
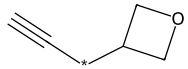
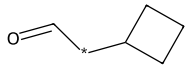
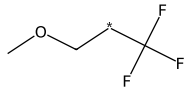
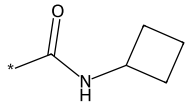
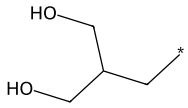
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 4A12D3737AA3A4CA Source: `CC1(O)CN(C=O)C1` Scaffold: `[*:1]N1CC(C)(O)C1`	0: `CC1(O)CN(C(N)=O)C1` 1*: `CC1(O)CN(C=O)C1`	
Task 4C393FCA9E44AE20 Source: `CCNC(=O)C(C)C` Scaffold: `[*:1](NCC)C(C)C`	0*: `CCNC(=O)C(C)C` 1: `CCNC(=O)OC(C)C`	
Task 4D88083DAD64937D Source: `CC(C)CCO` Scaffold: `[*:1](CO)C(C)C`	0: `CC(C)OC(=O)CO` 1*: `CC(C)CCO`	
Task 50B0F1F3F405BC45 Source: `CCNC(=O)N(C)C` Scaffold: `[*:1](CC)C(=O)N(C)C`	0*: `CCNC(=O)N(C)C` 1: `CCC(=O)C(=O)N(C)C`	
Task 52491A01317B6F72 Source: `CC1(C(N)=O)COC1` Scaffold: `[*:1]C1(C)COC1`	0: `CC1(C)COC1` 1: `CC1(O)COC1` 2*: `CC1(C(N)=O)COC1` 3: `CC1(C=O)COC1`	
Task 53E5A90033598201 Source: `NC(=O)C1CCCC1` Scaffold: `[*:1]C1CCCC1`	0: `OC1CCCC1` 1*: `NC(=O)C1CCCC1` 2: `N#CC1CCCC1` 3: `CC(=O)C1CCCC1`	
Task 55E46FB6790CFFDB Source: `NC(=O)C1(CO)COC1` Scaffold: `[*:1]C1(CO)COC1`	0: `CC1(CO)COC1` 1*: `NC(=O)C1(CO)COC1` 2: `OCC1(O)COC1` 3: `O=CC1(CO)COC1`	
Task 574B651325D3BD11 Source: `CC(O)(C=O)C=O` Scaffold: `[*:1]C(C)(C=O)C=O`	0: `CC(C=O)(C=O)C(N)=O` 1*: `CC(O)(C=O)C=O` 2: `CC(C#N)(C=O)C=O`	
Task 5C0A2E92C47860A1 Source: `CCC1(CC(N)=O)CC1` Scaffold: `[*:1]CC1(CC)CC1`	0*: `CCC1(CC(N)=O)CC1` 1: `CCC1(CC=O)CC1` 2: `CCC1(CO)CC1`	
Task 5CA52712D7BA537C Source: `NC(=O)C#CC1(O)CC1` Scaffold: `[*:1]C#CC1(O)CC1`	0*: `NC(=O)C#CC1(O)CC1` 1: `CC#CC1(O)CC1` 2: `N#CC#CC1(O)CC1`	

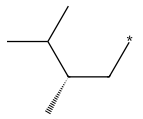
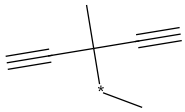
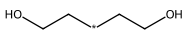
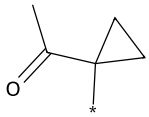
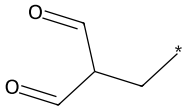
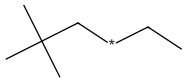
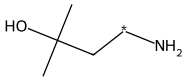
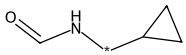
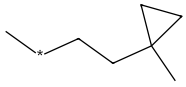
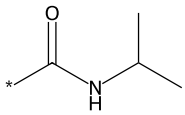
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 6156F4D256F86771 Source: N#CCC(=O)CC#N Scaffold: [*:1](CC#N)CC#N	0: N#CCCCC#N 1*: N#CCC(=O)CC#N 2: N#CCOC(=O)CC#N 3: N#CCOCC#N	
Task 667FB15D4ABC2DAA Source: C#CC1(C#N)CCC1 Scaffold: [*:1]C1(C#C)CCC1	0: C#CC1(C(C)=O)CCC1 1*: C#CC1(C#N)CCC1 2: C#CC1(C(N)=O)CCC1 3: C#CC1(O)CCC1	
Task 6CCE6F1CFF686EC3 Source: CCOC1(C#N)CC1 Scaffold: [*:1]C1(OCC)CC1	0*: CCOC1(C#N)CC1 1: CCOC1(C(C)=O)CC1 2: CCOC1(C(N)=O)CC1 3: CCOC1(C)CC1	
Task 6DFF62767CA262DC Source: N=CN(C=O)C1CC1 Scaffold: [*:1]=CN(C=O)C1CC1	0*: N=CN(C=O)C1CC1 1: O=CN(C=O)C1CC1	
Task 710125E7E697B6FC Source: [NH-][CH+]OC(=O)C1CC1 Scaffold: [*:1]([CH+][NH-])C1CC1	0: [NH-][CH+]OC1CC1 1*: [NH-][CH+]OC(=O)C1CC1	
Task 72EEA655E1BB79C9 Source: C#CCC1COC1 Scaffold: [*:1](C#C)C1COC1	0: C#CC(=O)C1COC1 1*: C#CCC1COC1 2: C#CC(=O)OC1COC1	
Task 770563210BBD880C Source: O=CC(=O)C1CCC1 Scaffold: [*:1](C=O)C1CCC1	0: O=CC(=O)OC1CCC1 1: O=COC1CCC1 2*: O=CC(=O)C1CCC1 3: O=CCC1CCC1	
Task 770746CF42201E53 Source: COCCC(F)(F)F Scaffold: [*:1](COC)C(F)(F)F	0: COCC(=O)C(F)(F)F 1*: COCCC(F)(F)F	
Task 79CD04C3DFC1C559 Source: COC(=O)NC1CCC1 Scaffold: [*:1]C(=O)NC1CCC1	0*: COC(=O)NC1CCC1 1: NC(=O)NC1CCC1	
Task 7AFAE7450D2BA2F9 Source: O=CCC(CO)CO Scaffold: [*:1]CC(CO)CO	0: CC(=O)CC(CO)CO 1: CCC(CO)CO 2*: O=CCC(CO)CO 3: NC(=O)CC(CO)CO	

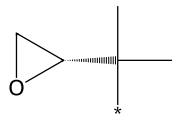
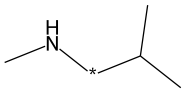
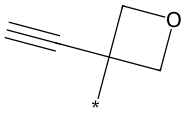
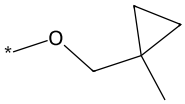
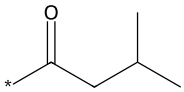
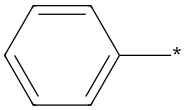
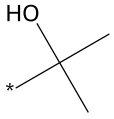
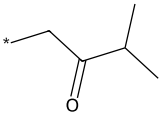
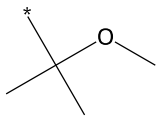
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 7B035A714C8D2D48 Source: CC(C)[C@@H](C)CC(N)=O Scaffold: [*:1]C[C@H](C)C(C)C	0: CC(C)[C@@H](C)CO 1*: CC(C)[C@@H](C)CC(N)=O 2: CC(=O)C[C@H](C)C(C)C 3: CC(C)[C@@H](C)CC#N	
Task 7B1F580DF30BC377 Source: C#CC(C)(C#C)C(C)=O Scaffold: [*:1](C)C(C)(C#C)C#C	0*: C#CC(C)(C#C)C(C)=O 1: C#CC(C)(C#C)OC	
Task 7BF9D99B3C712989 Source: O=C(CCO)OCCO Scaffold: [*:1](CCO)CCO	0: O=C(CCO)CCO 1: OCCCCCO 2*: O=C(CCO)OCCO	
Task 7C4091D1B273028B Source: CC(=O)C1(C)CC1 Scaffold: [*:1]C1(C(C)=O)CC1	0: CC(=O)C1(C(C)=O)CC1 1: CC(=O)C1(C#N)CC1 2*: CC(=O)C1(C)CC1	
Task 7CD2F406E74AFA61 Source: CC(=O)CC(C=O)C=O Scaffold: [*:1]CC(C=O)C=O	0: O=CC(C=O)CO 1: NC(=O)CC(C=O)C=O 2: O=CCC(C=O)C=O 3*: CC(=O)CC(C=O)C=O	
Task 804232E39F552D6A Source: CCOCC(C)(C)C Scaffold: [*:1](CC)CC(C)(C)C	0: CCC(=O)CC(C)(C)C 1*: CCOCC(C)(C)C	
Task 805AE1C24C13E756 Source: CC(C)(O)COC(N)=O Scaffold: [*:1](N)CC(C)(C)O	0: CC(C)(O)CC(N)=O 1*: CC(C)(O)COC(N)=O	
Task 83038E6D6357A2A6 Source: O=CNCC1CC1 Scaffold: [*:1](NC=O)C1CC1	0*: O=CNCC1CC1 1: O=CNC(=O)OC1CC1 2: O=CNC(=O)C1CC1	
Task 84B51E25F5AD2498 Source: COCCC1(C)CC1 Scaffold: [*:1](C)CCC1(C)CC1	0: CC(=O)CCC1(C)CC1 1*: COCCC1(C)CC1	
Task 85586DA089B69E42 Source: COC(=O)NC(C)C Scaffold: [*:1]C(=O)NC(C)C	0*: COC(=O)NC(C)C 1: CC(=O)C(=O)NC(C)C 2: CC(=O)NC(C)C 3: CC(C)NC(=O)C(N)=O	

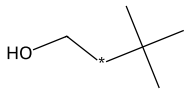
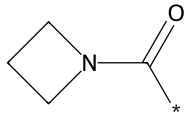
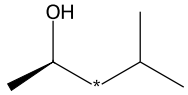
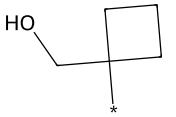
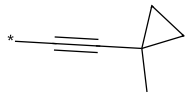
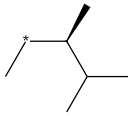
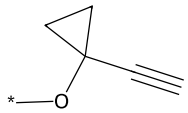
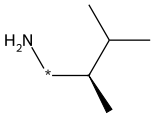
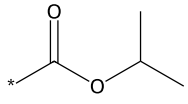
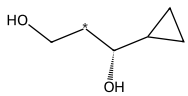
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 882245EA97520381 Source: `CC(C)(O)[C@H]1CO1` Scaffold: `[*:1]C(C)(C)[C@H]1CO1`	0: `CC(C)(C(N)=O)[C@H]1CO1` 1: `CC(C)(C#N)[C@H]1CO1` 2: `CC(=O)C(C)(C)[C@H]1CO1` 3*: `CC(C)(O)[C@H]1CO1`	
Task 88C97FE31AD8F75C Source: `CCCOCCC` Scaffold: `[*:1](CCC)CCC`	0: `CCCC(=O)CCC` 1*: `CCCOCCC` 2: `CCCOC(=O)CCC`	
Task 894BBF79C573C29A Source: `CNC(=O)OC(C)C` Scaffold: `[*:1](NC)C(C)C`	0: `CNC(=O)C(C)C` 1*: `CNC(=O)OC(C)C`	
Task 8E3C2B15F7EA9DCF Source: `C#CC1(O)COC1` Scaffold: `[*:1]C1(C#C)COC1`	0*: `C#CC1(O)COC1` 1: `C#CC1(C)COC1` 2: `C#CC1(C#N)COC1` 3: `C#CC1(C(N)=O)COC1`	
Task 8ECAAE94795D4478 Source: `CC1(COC=O)CC1` Scaffold: `[*:1]OCC1(C)CC1`	0*: `CC1(COC=O)CC1` 1: `COCC1(C)CC1` 2: `CC(=O)OCC1(C)CC1`	
Task 8F3EBB5FA1D15FEA Source: `CC(C)CC(=O)C(N)=O` Scaffold: `[*:1]C(=O)CC(C)C`	0: `CC(C)CC(=O)C#N` 1*: `CC(C)CC(=O)C(N)=O` 2: `CC(=O)CC(C)C` 3: `CC(=O)C(=O)CC(C)C`	
Task 9126F1FAF023C0C2 Source: `O=Cc1ccccc1` Scaffold: `[*:1]c1ccccc1`	0: `O=[N+]([O-])c1ccccc1` 1: `Oc1ccccc1` 2*: `O=Cc1ccccc1`	
Task 91935246108AF8AC Source: `CC(C)(O)C#N` Scaffold: `[*:1]C(C)(C)O`	0: `CC(=O)C(C)(C)O` 1: `CC(C)(C)O` 2*: `CC(C)(O)C#N`	
Task 92176A1FD6CFADCB Source: `CC(C)C(=O)CO` Scaffold: `[*:1]CC(=O)C(C)C`	0: `CC(=O)CC(=O)C(C)C` 1: `CC(C)C(=O)CC(N)=O` 2: `CC(C)C(=O)CC#N` 3*: `CC(C)C(=O)CO`	
Task 9232A0B17FA5AF7B Source: `COC(C)(C)C(N)=O` Scaffold: `[*:1]C(C)(C)OC`	0*: `COC(C)(C)C(N)=O` 1: `COC(C)(C)C` 2: `COC(C)(C)C#N` 3: `COC(C)(C)C(C)=O`	

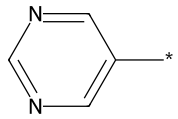
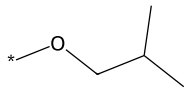
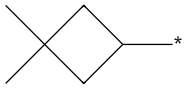
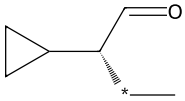
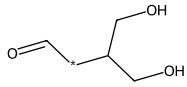
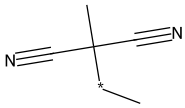
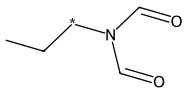
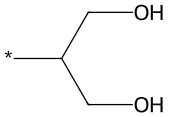
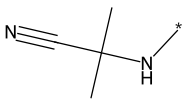
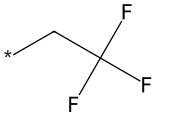
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 92EF44C16081DC55 Source: `CC(C)(C)OC(=O)CO` Scaffold: `[*:1](CO)C(C)(C)C`	0: `CC(C)(C)CCO` 1*: `CC(C)(C)OC(=O)CO` 2: `CC(C)(C)C(=O)CO`	
Task 93D3F171694A478A Source: `NC(=O)N1CCC1` Scaffold: `[*:1]C(=O)N1CCC1`	0: `CC(=O)C(=O)N1CCC1` 1*: `NC(=O)N1CCC1` 2: `COC(=O)N1CCC1` 3: `N#CC(=O)N1CCC1`	
Task 941D831C38D8914B Source: `CC(C)OC(=O)[C@@H](C)O` Scaffold: `[*:1](C(C)C)[C@@H](C)O`	0*: `CC(C)OC(=O)[C@@H](C)O` 1: `CC(C)C[C@@H](C)O`	
Task 948FFDE8CC5D8FA7 Source: `N#CC1(CO)CCC1` Scaffold: `[*:1]C1(CO)CCC1`	0*: `N#CC1(CO)CCC1` 1: `NC(=O)C1(CO)CCC1` 2: `OCC1(O)CCC1` 3: `CC1(CO)CCC1`	
Task 9683002FA47E674E Source: `CC(=O)C#CC1(C)CC1` Scaffold: `[*:1]C#CC1(C)CC1`	0: `CC1(C#CC#N)CC1` 1*: `CC(=O)C#CC1(C)CC1`	
Task 97CCF0B132ED0CC7 Source: `COC(=O)[C@@H](C)C(C)C` Scaffold: `[*:1](C)[C@@H](C)C(C)C`	0: `CO[C@@H](C)C(C)C` 1*: `COC(=O)[C@@H](C)C(C)C`	
Task 989AE8521BA7272A Source: `C#CC1(OC=O)CC1` Scaffold: `[*:1]OC1(C#C)CC1`	0: `C#CC1(OC)CC1` 1: `C#CC1(OC(C)=O)CC1` 2*: `C#CC1(OC=O)CC1`	
Task 9B92313F1C8E02FE Source: `CC(C)[C@@H](C)C(N)=O` Scaffold: `[*:1](N)[C@H](C)C(C)C`	0*: `CC(C)[C@@H](C)C(N)=O` 1: `CC(C)[C@@H](C)OC(N)=O`	
Task 9BB875535FD8F9F9 Source: `CC(C)OC(N)=O` Scaffold: `[*:1]C(=O)OC(C)C`	0*: `CC(C)OC(N)=O` 1: `CC(=O)C(=O)OC(C)C` 2: `CC(C)OC(=O)C#N`	
Task 9BCD7134AA9886D1 Source: `OCC[C@@H](O)C1CC1` Scaffold: `[*:1](CO)[C@@H](O)C1CC1`	0: `O=C(CO)[C@@H](O)C1CC1` 1*: `OCC[C@@H](O)C1CC1`	

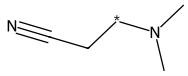
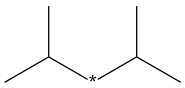
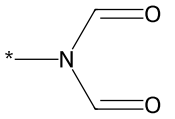
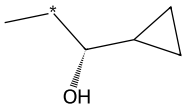
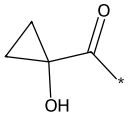
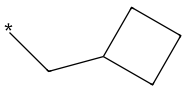
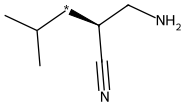
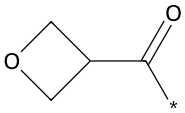
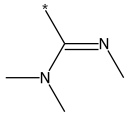
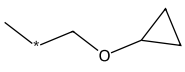
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task 9BF5F7BCD985EC74 Source: `COc1cncnc1` Scaffold: `[*:1]c1cncnc1`	0: `Nc1cncnc1` 1: `O=[N+]([O-])c1cncnc1` 2*: `COc1cncnc1` 3: `Cc1cncnc1`	
Task 9C74B641B7F823FB Source: `CC(C)COC(N)=O` Scaffold: `[*:1]OCC(C)C`	0: `CC(C)COC=O` 1: `COCC(C)C` 2*: `CC(C)COC(N)=O`	
Task 9D9E13DA2B72CDDA Source: `CC1(C)CC(C#N)C1` Scaffold: `[*:1]C1CC(C)(C)C1`	0: `CC1(C)CC(C(N)=O)C1` 1*: `CC1(C)CC(C#N)C1` 2: `CC1(C)CC(O)C1` 3: `CC(=O)C1CC(C)(C)C1`	
Task 9FCBD2D389E1B5FA Source: `CO[C@@H](C=O)C1CC1` Scaffold: `[*:1](C)[C@@H](C=O)C1CC1`	0*: `CO[C@@H](C=O)C1CC1` 1: `CC(=O)[C@@H](C=O)C1CC1`	
Task A2B5123D6703CB27 Source: `O=COC(CO)CO` Scaffold: `[*:1](C=O)C(CO)CO`	0*: `O=COC(CO)CO` 1: `O=CC(=O)C(CO)CO`	
Task A3734A8D4C629ED7 Source: `CC(=O)C(C)(C#N)C#N` Scaffold: `[*:1](C)C(C)(C#N)C#N`	0*: `CC(=O)C(C)(C#N)C#N` 1: `CNC(C)(C#N)C#N`	
Task A5EF11CDF9E14552 Source: `CCC(=O)N(C=O)C=O` Scaffold: `[*:1](CC)N(C=O)C=O`	0: `CCCN(C=O)C=O` 1*: `CCC(=O)N(C=O)C=O`	
Task A7C3C4816799E0C2 Source: `OCC(O)CO` Scaffold: `[*:1]C(CO)CO`	0: `N#CC(CO)CO` 1: `CC(=O)C(CO)CO` 2: `COC(CO)CO` 3*: `OCC(O)CO`	
Task A87FD55731782952 Source: `CNC(C)(C)C#N` Scaffold: `[*:1]NC(C)(C)C#N`	0: `CC(C)(C#N)NC(N)=O` 1: `CC(=O)NC(C)(C)C#N` 2*: `CNC(C)(C)C#N` 3: `CC(C)(C#N)NC=O`	
Task AA24AB4408560E71 Source: `CCC(F)(F)F` Scaffold: `[*:1]CC(F)(F)F`	0: `O=CCC(F)(F)F` 1*: `CCC(F)(F)F` 2: `OCC(F)(F)F` 3: `CC(=O)CC(F)(F)F`	

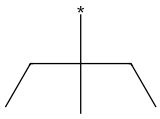
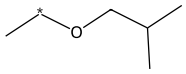
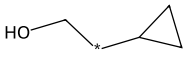
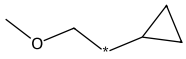
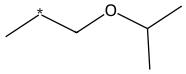
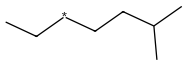
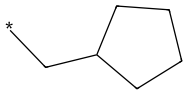
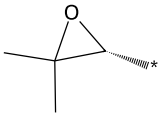
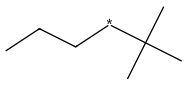
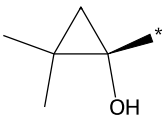
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task AACEF21B5C0B518C Source: CN(C)C(=O)OCC#N Scaffold: [*:1](CC#N)N(C)C	0*: CN(C)C(=O)OCC#N 1: CN(C)C(=O)CC#N	
Task AB494D4F810488A4 Source: CC(C)OC(C)C Scaffold: [*:1](C(C)C)C(C)C	0: CC(C)OC(=O)C(C)C 1: CC(C)C(=O)C(C)C 2*: CC(C)OC(C)C	
Task AE2202D59DBD5246 Source: CN(C=O)C=O Scaffold: [*:1]N(C=O)C=O	0: O=CN(C=O)C=O 1: NC(=O)N(C=O)C=O 2: CC(=O)N(C=O)C=O 3*: CN(C=O)C=O	
Task AF7B02587CACF57F Source: COC(=O)[C@@H](O)C1CC1 Scaffold: [*:1](C)[C@@H](O)C1CC1	0: CC[C@@H](O)C1CC1 1*: COC(=O)[C@@H](O)C1CC1	
Task B0B5353EAB912566 Source: N#CC(=O)C1(O)CC1 Scaffold: [*:1]C(=O)C1(O)CC1	0: CC(=O)C(=O)C1(O)CC1 1*: N#CC(=O)C1(O)CC1 2: CC(=O)C1(O)CC1 3: COC(=O)C1(O)CC1	
Task B2D636889CB3A680 Source: N#CCC1CCC1 Scaffold: [*:1]CC1CCC1	0: OCC1CCC1 1*: N#CCC1CCC1 2: NC(=O)CC1CCC1 3: CC(=O)CC1CCC1	
Task BA45AF3D599B3BC8 Source: CC(C)O[C@H](C#N)CN Scaffold: [*:1](C(C)C)[C@H](C#N)CN	0*: CC(C)O[C@H](C#N)CN 1: CC(C)N[C@H](C#N)CN	
Task BFC8332E939DA9F3 Source: CC(=O)C1COC1 Scaffold: [*:1]C(=O)C1COC1	0*: CC(=O)C1COC1 1: O=CC(=O)C1COC1 2: NC(=O)C1COC1 3: CC(=O)C(=O)C1COC1	
Task C0AF48D3C3C39A0B Source: C/N=C(\C#N)N(C)C Scaffold: [*:1]C(=NC)N(C)C	0: C/N=C(\C(N)=O)N(C)C 1*: C/N=C(\C#N)N(C)C	
Task C0B54784822A1928 Source: CCCOC1CC1 Scaffold: [*:1](C)COC1CC1	0*: CCCOC1CC1 1: CC(=O)COC1CC1 2: COC(=O)COC1CC1	

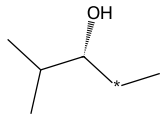
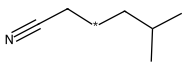
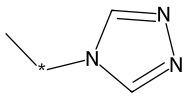
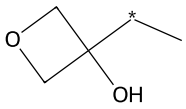
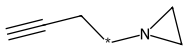
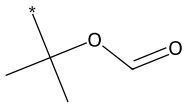
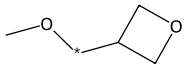
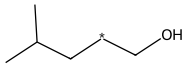
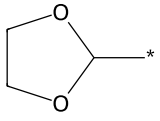
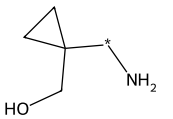
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task C386BB77321A4F80 Source: `CCC(C)(C#N)CC` Scaffold: `[*:1]C(C)(CC)CC`	0: `CCC(C)(C=O)CC` 1*: `CCC(C)(C#N)CC` 2: `CCC(C)(CC)C(N)=O` 3: `CCC(C)(O)CC`	
Task C52FF4DABD2EF5F4 Source: `CCOCC(C)C` Scaffold: `[*:1](C)OCC(C)C`	0*: `CCOCC(C)C` 1: `CC(=O)OCC(C)C`	
Task C6F354240B9C0145 Source: `OCCC1CC1` Scaffold: `[*:1](CO)C1CC1`	0*: `OCCC1CC1` 1: `O=C(CO)C1CC1` 2: `O=C(CO)OC1CC1`	
Task C9454DE19FDBD111 Source: `COCC(=O)C1CC1` Scaffold: `[*:1](COC)C1CC1`	0*: `COCC(=O)C1CC1` 1: `COCCC1CC1` 2: `COCC(=O)OC1CC1`	
Task CC3D1FF1B192C8E7 Source: `COC(=O)COC(C)C` Scaffold: `[*:1](C)COC(C)C`	0: `CCCOC(C)C` 1: `CC(=O)COC(C)C` 2*: `COC(=O)COC(C)C`	
Task CD67FD0D0C427DA5 Source: `CCC(=O)CCC(C)C` Scaffold: `[*:1](CC)CCC(C)C`	0*: `CCC(=O)CCC(C)C` 1: `CCOCCC(C)C`	
Task CDCFF67DEEFC762A Source: `CC(=O)CC1CCCC1` Scaffold: `[*:1]CC1CCCC1`	0: `OCC1CCCC1` 1: `N#CCC1CCCC1` 2*: `CC(=O)CC1CCCC1` 3: `NC(=O)CC1CCCC1`	
Task D22AF6CCC7C5A672 Source: `CC1(C)O[C@H]1C(N)=O` Scaffold: `[*:1][C@H]1OC1(C)C`	0*: `CC1(C)O[C@H]1C(N)=O` 1: `CC1(C)O[C@H]1C=O` 2: `C[C@H]1OC1(C)C` 3: `CC1(C)O[C@H]1C#N`	
Task D316E49E71315C17 Source: `CCCOC(C)(C)C` Scaffold: `[*:1](CCC)C(C)(C)C`	0*: `CCCOC(C)(C)C` 1: `CCCC(=O)C(C)(C)C`	
Task D3AC12168083A92D Source: `CC1(C)C[C@]1(O)C#N` Scaffold: `[*:1][C@]1(O)CC1(C)C`	0: `CC1(C)C[C@]1(O)C(N)=O` 1: `CC1(C)C[C@]1(C)O` 2*: `CC1(C)C[C@]1(O)C#N` 3: `CC(=O)[C@@]1(O)CC1(C)C`	

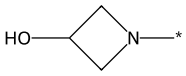
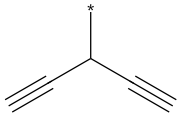
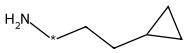
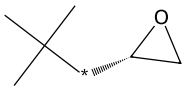
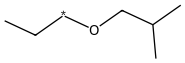
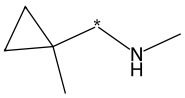
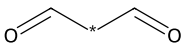
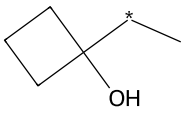
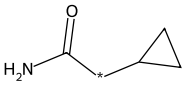
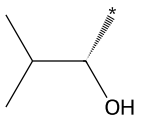
Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task D4899C2A8A512A89 Source: `CC[C@@H](O)C(C)C` Scaffold: `[*:1](C)[C@@H](O)C(C)C`	0: `COC(=O)[C@@H](O)C(C)C` 1*: `CC[C@@H](O)C(C)C`	
Task D6E26BE8568B816E Source: `CC(C)CC(=O)CC#N` Scaffold: `[*:1](CC#N)CC(C)C`	0: `CC(C)CNCC#N` 1*: `CC(C)CC(=O)CC#N`	
Task D775DD7E29DC79AD Source: `CC(=O)n1cnnc1` Scaffold: `[*:1](C)n1cnnc1`	0: `CCn1cnnc1` 1*: `CC(=O)n1cnnc1` 2: `COC(=O)n1cnnc1`	
Task D98AA4EA6458E7F6 Source: `CCC1(O)COC1` Scaffold: `[*:1](C)C1(O)COC1`	0: `CC(=O)C1(O)COC1` 1: `COC(=O)C1(O)COC1` 2*: `CCC1(O)COC1`	
Task DD4C95D9F30BB38C Source: `C#CCC(=O)N1CC1` Scaffold: `[*:1](CC#C)N1CC1`	0*: `C#CCC(=O)N1CC1` 1: `C#CCOC(=O)N1CC1`	
Task DF6B46BD1891D2D1 Source: `CC(C)(C)OC=O` Scaffold: `[*:1]C(C)(C)OC=O`	0: `CC(C)(OC=O)C(N)=O` 1*: `CC(C)(C)OC=O` 2: `CC(C)(C=O)OC=O` 3: `CC(=O)C(C)(C)OC=O`	
Task E1262DF703B497E5 Source: `COCC1COC1` Scaffold: `[*:1](OC)C1COC1`	0: `COC(=O)C1COC1` 1*: `COCC1COC1`	
Task E1413048E6C74A5A Source: `CC(C)COC(=O)CO` Scaffold: `[*:1](CO)CC(C)C`	0*: `CC(C)COC(=O)CO` 1: `CC(C)CCCO`	
Task E4B8BF5C0922DBF7 Source: `N#CC1OCCO1` Scaffold: `[*:1]C1OCCO1`	0*: `N#CC1OCCO1` 1: `CC1OCCO1` 2: `NC(=O)C1OCCO1`	
Task E4F05C2D54C8C77F Source: `NC(=O)C1(CO)CC1` Scaffold: `[*:1](N)C1(CO)CC1`	0: `NC(=O)OC1(CO)CC1` 1*: `NC(=O)C1(CO)CC1`	

Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task E6C7C7741498105B Source: `O=CN1CC(O)C1` Scaffold: `[*:1]N1CC(O)C1`	0*: `O=CN1CC(O)C1` 1: `CC(=O)N1CC(O)C1`	
Task E8A55EE38951E20F Source: `C#CC(C#C)C#N` Scaffold: `[*:1]C(C#C)C#C`	0: `C#CC(C#C)C(C)=O` 1*: `C#CC(C#C)C#N` 2: `C#CC(O)C#C`	
Task E9CAE26E473183AA Source: `NC(=O)CCC1CC1` Scaffold: `[*:1](N)CCC1CC1`	0: `NC(=O)OCCC1CC1` 1*: `NC(=O)CCC1CC1`	
Task ED106620D1D1314C Source: `CC(C)(C)C(=O)[C@H]1CO1` Scaffold: `[*:1]([C@H]1CO1)C(C)(C)C`	0: `CC(C)(C)C[C@H]1CO1` 1*: `CC(C)(C)C(=O)[C@H]1CO1`	
Task EFA40744DABDF89B Source: `CCCOCC(C)C` Scaffold: `[*:1](CC)OCC(C)C`	0*: `CCCOCC(C)C` 1: `CCC(=O)OCC(C)C`	
Task F0E40A98082FD2D6 Source: `CNC(=O)C1(C)CC1` Scaffold: `[*:1](NC)C1(C)CC1`	0*: `CNC(=O)C1(C)CC1` 1: `CNC(=O)OC1(C)CC1`	
Task F185637AA03BDE72 Source: `O=CNC=O` Scaffold: `[*:1](C=O)C=O`	0*: `O=CNC=O` 1: `O=CC(=O)C=O`	
Task F40E5886AB249191 Source: `COC(=O)C1(O)CCC1` Scaffold: `[*:1](C)C1(O)CCC1`	0: `CC(=O)C1(O)CCC1` 1: `CCC1(O)CCC1` 2*: `COC(=O)C1(O)CCC1`	
Task FAB8ED1AC4A795AB Source: `NC(=O)NC1CC1` Scaffold: `[*:1](C(N)=O)C1CC1`	0: `NC(=O)C(=O)C1CC1` 1*: `NC(=O)NC1CC1` 2: `NC(=O)OC1CC1` 3: `NC(=O)C(=O)OC1CC1`	
Task FC2588AEFF1C037A Source: `CC(C)[C@@H](O)C(N)=O` Scaffold: `[*:1][C@H](O)C(C)C`	0: `CC(C)[C@@H](O)C#N` 1: `CC(C)[C@@H](C)O` 2*: `CC(C)[C@@H](O)C(N)=O` 3: `CC(=O)[C@H](O)C(C)C`	

Continued on next page

Complete task families (continued)

Task and SMILES	Candidate list	Scaffold
Task FCF4A1D9FEF6387E Source: `CC(C)[C@@H](O)C(=O)C=O` Scaffold: `[*:1](C=O)[C@H](O)C(C)C`	0: `CC(C)[C@@H](O)CC=O` 1*: `CC(C)[C@@H](O)C(=O)C=O`	