

Accurate, Interdisciplinary and Transparent Structure-property Understanding with Deep Native Structural Reasoning

Chen Tang^{1,2†}, Yizhou Wang^{1,2†}, Jianyu Wu^{1,3†}, Lintao Wang^{1,5†}, Shixiang Tang^{2,1}, Pengze Li^{1,4}, Encheng Su^{1,8}, Jun Yao^{1,8}, Jiabei Xiao^{1,2}, Yuqi Shi^{9,10}, Jielan Li¹, Hongxia Hao¹, Zhangyang Gao¹, Fang Wu¹¹, Ben Fei^{1,2}, Xiangyu Yue², Pan Tan¹, Bozitao Zhong³, Jinouwen Zhang¹, Aoran Wang¹, Yan Lu^{2,1}, Jiaheng Liu^{6,1}, Xinzhu Ma¹, Liang Hong³, Mingyue Zheng^{9,10}, Phil Torr⁷, Bowen Zhou¹, Wanli Ouyang^{1,2}, Lei Bai¹

¹Shanghai Artificial Intelligence Laboratory, China.

²The Chinese University of Hong Kong, Hong Kong.

³Shanghai Jiao Tong University, China.

⁴Fudan University, China.

⁵University of Sydney, Australia.

⁶Nanjing University, China.

⁷University of Oxford, UK.

⁸The University of Science and Technology of China, China.

⁹Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, China.

¹⁰University of Chinese Academy of Sciences, China.

¹¹Stanford University, USA.

Corresponding authors: shixiangtang@cuhk.edu.hk; ouyangwanli@pjlab.org.cn; bailei@pjlab.org.cn;

†These authors contributed equally to this work.

 SciReasoner.github.io

Abstract

Structure–property relationships are foundational to biology, chemistry and materials science, where function, reactivity and physical response emerge from spatial, chemical and periodic organization. Mechanistically explaining these relationships requires interpreting structural evidence through scientific principles and physical constraints, from stereochemistry and bonding to symmetry, energetics and periodic order. However, applying artificial intelligence to this process presents a joint challenge of representation and reasoning: models must preserve domain-native structural information while showing how specific evidence supports predictions under these constraints. Here we introduce SciReasoner, a multimodal scientific foundation model for native structural reasoning across proteins, small molecules and inorganic crystals. SciReasoner discretizes coordinates, topologies and periodic connectivities into a unified structure-aware vocabulary, treating structural tokens as addressable evidence units within autoregressive reasoning trajectories. We evaluate SciReasoner in settings where shortcut correlations are weakened and structure-grounded inference is essential. In homology-controlled Gene Ontology prediction, SciReasoner improves Cellular Component annotation for low-homology and orphan-like proteins, increasing F_{\max} from 0.42 to 0.55. In chemistry, it raises single-step retrosynthesis accuracy from 0.63 to 0.72 while generating fragment-level disconnection and precursor-verification traces. In materials science, its representations separate elemental and compound phases and resolve high- and low-band-gap regimes. Across 86 benchmarks, SciReasoner achieves state-of-the-art performance on 67 tasks. Double-blind expert evaluation rates its reasoning traces as preferred or at least comparable to those of a frontier large language model in 98% of cases. By making structure an inspectable substrate for reasoning under scientific constraints, SciReasoner connects accurate prediction with interpretable scientific inference.

1 Introduction

Structure–property relationships [1–6] are foundational to the physical and biological sciences. Across proteins, small molecules and crystalline materials, observable functions and properties arise from the spatial, chemical and periodic organization of matter. In proteins, protein structures and conformations, long-range interactions and active-site geometry shape their biological functions [7, 8]; in chemicals, bonding, functional groups and stereochemistry contribute to their reactivity and molecular behaviour; and in materials, lattice symmetry, coordination motifs and periodic bonding networks dictate stability and electronic properties [9, 10]. Unraveling how these organizing principles connect structure to function, reactivity and physical response is therefore central to functional annotation, molecular design and materials discovery.

Yet, explaining these relationships mechanistically is rarely a straightforward pattern-matching problem. Evidence for a function, reaction or material property is distributed across local motifs, non-local contacts, chemical environments, conformational geometry and long-range periodic order. When annotating a protein or planning a synthesis route [11, 12], human experts integrate heterogeneous cues, compare competing structural hypotheses and apply biochemical, chemical or crystallographic constraints [13, 14]. Artificial intelligence (AI) can accelerate this process by retrieving knowledge and screening large structural spaces, but it faces a deeper representational challenge: to support expert-like scientific reasoning, a model must represent biological, chemical and materials structures as native objects and use them as evidence in explicit, inspectable reasoning.

Current scientific AI systems only partially meet this requirement because they often separate native structural representation from evidence-linked reasoning. Large language models provide a flexible interface to scientific knowledge, but when proteins, molecules and crystals are cast primarily as text, structural organization is compressed into strings or descriptions, making the resulting explanations depend largely on linguistic associations rather than directly addressable physical evidence [15–17]. Agentic systems extend such models through retrieval, tool use and workflow coordination [18–21], yet their scientific depth remains constrained by the structural competence of the foundation models they orchestrate. Domain-specific models, in contrast, often encode molecular graphs, protein structures or crystal lattices directly [9, 22, 23], but are usually optimized as task-specific predictors that output scores, labels or scalar properties without exposing the intermediate evidence behind their decisions. Thus, a critical gap remains: building a new reasoning paradigm that natively represents multidimensional scientific structures while producing reasoning traces that can be inspected against those structures.

Here we propose native structural reasoning as a foundation-model paradigm for structure–property analysis across biology, chemistry and materials science. We instantiate this paradigm in SciReasoner, a multimodal scientific foundation model that represents proteins, small molecules and periodic crystals through a unified structure-aware vocabulary. By native structural reasoning, we mean that structural tokens are not auxiliary descriptors appended to language, but addressable evidence units that can be combined, cited and checked within a generated reasoning trajectory. To this end, SciReasoner discretizes coordinates, topologies and periodic connectivities into scientific tokens that preserve domain-native information, including local geometry, stereochemistry, residue-level structural motifs, lattice symmetry and crystallographic bonding patterns. These tokens are integrated with language instructions in a single autoregressive model, enabling trajectories that interleave explanatory text with specific residues, molecular fragments, conformational cues or crystal descriptors.

This design changes the role of a scientific foundation model. Rather than predicting an answer after reading a structural prompt, SciReasoner generates an inspectable chain in which intermediate claims can be traced to explicit structural evidence. We test this capability in regimes where shortcut correlations are unreliable.

In the Cellular Component aspect of protein Gene Ontology prediction, SciReasoner shows its largest gains in low-homology and orphan-like proteins, where evolutionary transfer provides weak evidence, improving F_{\max} from 0.42 to 0.55. Its attention is enriched at contact-defined DNA-binding residues and localizes to protein–DNA interfaces, suggesting that functional predictions are supported by residues that physically mediate the molecular interaction. In retrosynthesis, SciReasoner raises single-step accuracy from 0.63 to 0.72 while decomposing products into reaction-relevant fragments, identifying strategic bond disconnections and verifying chemically valid precursors. In crystalline materials, the reasoning trajectories of SciReasoner assign higher importance to symmetry-distinct sites, local coordination environments and periodic bonding motifs associated with stability and band-gap variation, suggesting that it links electronic and thermodynamic behaviour to the structural features that physically govern them. Across these examples, the common behaviour is that structural organization acts as the working substrate of inference rather than a peripheral input descriptor.

Broad evaluation supports the generality of this structure-grounded modelling strategy. Across 86 benchmarks spanning proteins, DNA, RNA, small molecules, inorganic crystals, scientific question answering, property prediction and generation tasks, SciReasoner achieves state-of-the-art performance on 67 tasks. Double-blind expert evaluation further indicates that its reasoning traces are useful for scientific inspection, with experts preferring or tying SciReasoner over a frontier large language model in the overwhelming majority of comparisons. By representing structures as evidence-bearing units and linking them to inspectable conclusions, SciReasoner bridges accurate prediction and interpretable scientific inference, suggesting a route toward scientific AI systems that do not merely predict what property a structure has, but begin to explain why that structure gives rise to a particular function, reaction or material behaviour.

Fig. 1: Overview of SciReasoner, a unified multimodal foundation model for native structural reasoning across biomolecules, small molecules, and periodic crystals. (A) Structure-aware scientific reasoning interface. SciReasoner encodes heterogeneous scientific entities, including protein/DNA/RNA sequences, small molecules, protein structures, crystal structures, and 3D molecules, into structure-aware tokens. These tokens are combined with language instructions and processed by an LLM backbone to support structure-grounded scientific QA, property prediction, property classification, and generation/design tasks. (B) Structure-aware token construction. Domain-specific encoders convert scientific objects into structured textual representations: ConfSeq for molecules, Foldseek for proteins, and SLICES for crystals. These representations preserve chemically and physically meaningful information, including molecular topology, spatial geometry, local motifs, lattice features, symmetry, and space-group information. (C) Tokenizer compression and semantic preservation. Compared with the Qwen tokenizer, the SciReasoner tokenizer reduces token length for molecular inputs while preserving chemically meaningful substructures. This indivisible, structure-aware vocabulary mitigates semantic fragmentation caused by standard sub-word tokenization. (D) Continued Pretraining and post-training pipeline. SciReasoner is initialized from a Qwen model and trained through warm-up alignment, full-parameter multimodal training, and annealing tuning across scientific modalities. During post-training, intra-domain structural evidence grounding first trains task-specialized experts to use structural tokens as reasoning evidence within each domain. Cross-domain reasoning consolidation then integrates the expert-generated reasoning traces and expert capabilities into the final SciReasoner model. (E) Benchmark performance. SciReasoner is evaluated across diverse scientific QA, biology, chemistry, materials science, and generation/design benchmarks. The results demonstrate broad improvements over generalist language models and domain-specific baselines, highlighting SciReasoner’s ability to connect structural evidence with function, property, and design decisions.

2 Results

2.1 Overview of SciReasoner

We present SciReasoner, a unified multimodal foundation model capable of explicit scientific sequence, structure, and textual reasoning within a single coherent output trajectory (Fig. 1A). SciReasoner serves as a unified reasoning engine that maps diverse scientific inputs, including proteins, DNA/RNA, small molecules, and periodic crystals, into a shared cross-modal latent space. Unlike conventional computational approaches in materials and biology, this architecture provides a unified, mechanism-aware framework.

To integrate explicit 3D structural information into a linguistic token space, SciReasoner designs a structure-aware vocabulary designed to preserve physical and biochemical integrity (Fig. 1A). The model utilizes unified tokens representing local motifs, 3D geometry, atomic bond types, and crystal space groups (Fig. 1B). This resolves the severe limitations of conventional sub-word tokenization algorithms, which arbitrarily fragment molecular graphs and motifs, thereby destroying scientific semantics and corrupting the training process (Fig. 1C).

To align this structural token space with natural language, we perform continued pretraining with a multi-stage training pipeline (Fig. 1D, Top). Stage 1 serves as a warm-up training phase that anchors the newly introduced structure-aware tokens to basic topological, geometric, and chemical semantics without destabilizing the pretrained language space. This establishes a stable structural–linguistic interface, which is further developed through Stage 2 full-parameter multimodal training and Stage 3 annealed training with an increased proportion of question–answer (QA)-style data, thereby supporting native structural reasoning.

To enable robust multi-step reasoning, we developed a post-training framework named self-bootstrapped native structural reasoning (Fig. 1D, Bottom). The post-training procedure uses reinforcement learning to connect the core semantic meanings of the structural vocabulary with a chain-of-thought (CoT) strategy. This allows the model to explicitly decompose complex inputs into interpretable physical factors, such as composition, symmetry, and connectivity, before executing final property predictions or design decisions. This generative approach circumvents a major bottleneck in the field: the near-complete absence of ground-truth reasoning trajectories that simultaneously bridge sequence, structure, and text. Intra-domain structural evidence grounding first teaches domain-specific experts how structural tokens serve as evidence for reasoning. Cross-domain reasoning consolidation then integrates these domain-local evidence-use patterns into unified capabilities, reducing the trajectory collapse frequently observed in standard alignment techniques.

The radar plot in Fig. 1E (left) compares SciReasoner with four contemporary generalist LLMs (Opus-4.7 [24], GPT-5.5 [25], DeepSeek-V4-Pro [26] and Kimi-K2.6 [27]) across five capability sectors and four scientific modalities (proteins, DNA/RNA, small molecules and crystals). The suite spans numerical property prediction, sequence- and structure-level classification, generation and design, modality translation and scientific QA. SciReasoner expands the polygon further than every generalist baseline in every sector, with the largest margins on tasks that demand explicit structural reasoning, where text-only LLMs collapse to near-trivial scores.

We next compared SciReasoner with the published specialist that defines the prior state of the art on each benchmark (Fig. 1E, right). Unlike these specialists, each of which is restricted to a fixed numerical or categorical output, SciReasoner produces free-form natural language alongside its predictions, yet still matches or exceeds them on the majority of tasks across the three task panels on the right: SciReasoner reaches a mean $F_{\max} =$

0.59 on DeepFRI-GO (averaged over the three aspects), surpassing SaProt [28] (0.52); attains 0.88 accuracy on subcellular localization, exceeding the ESM2 [29] (0.84); is on par with or above the specialist on DNA promoter and transcription-factor detection; and substantially outperforms RNA-function specialists, raising Isoform R^2 from 0.59 to 0.86 and RNA protein interaction MCC from 0.74 to 0.81. On the DUD-E virtual-screening suite, SciReasoner matches the previously reported best AUC of 0.76 and improves the 5.0% enrichment factor from 7.12 to 7.70. Beyond closed-set predictions, the same model handles open-ended scientific language tasks that lie entirely outside the specialists’ output space, scoring 0.85 BertScore on biomedical QA and 0.77 ROUGE-L on protein general function description, and supplies an interpretable natural-language rationale for each predicted label.

High benchmark scores can be obtained by exploiting homology, scaffold or template-level shortcuts, and do not by themselves demonstrate that a model has captured the underlying science. Sections 2.2.1–2.2.4 therefore probe four structural behaviours of SciReasoner beyond average accuracy. On protein GO prediction (§2.2.1), SciReasoner shows its largest gain in the low-homology regime, improving Cellular Component F_{\max} at $\leq 30\%$ identity, by +0.21 over BLAST [30] and by +0.13 over ESM2, while maintaining stable reasoning quality across similarity bins, arguing against homology memorization. On retrosynthesis (§2.2.2), the chain-of-thought interleaves language with SMILES sub-fragments along a canonical *analysis* \rightarrow *disconnection* \rightarrow *verification* \rightarrow *feasibility* workflow, recovering literature reactants in 5/5 representative Retrosynthesis USPTO-50K cases versus 2/5 for RSGPT [31]. On 3D molecular similarity (§2.2.3), the per-molecule embedding clusters DUD-E [32] actives by binding-pocket geometry rather than 2D scaffold. On inorganic crystals (§2.2.4), the latent space separates C, Si and SiC into disjoint compositional clusters and, within each, orders polymorphs along smooth band-gap gradients. The accompanying reasoning traces explicitly invoke coordination changes, bonding topology and symmetry constraints as mechanistic evidence for linking crystal structure to properties. Consistent with this physically aligned representation, the formation-energy parity plot (Fig. 4C) reaches $R^2 = 0.895$, well above competing models even where their absolute MAE is comparable, showing that SciReasoner captures the stability ranking rather than the marginal label distribution.

Across scientific tasks spanning proteins, DNA, RNA, small molecules and inorganic crystals, SciReasoner outperforms previous state-of-the-art methods on **67 tasks**, and matches or exceeds published domain specialists on **26 benchmarks** wherever such a baseline exists. To further exclude reliance on sequence, scaffold or composition-level shortcuts, we ablated structural inputs in representative protein, small-molecule and materials tasks and observed a consistent drop in performance, with the largest effects in problems governed by folded geometry, binding-site organization, molecular shape or crystal topology (Fig. 5). The same perturbation changed the reasoning traces: without structure, the reasoning process relied mainly on sequence motifs or chemical composition, whereas structural tokens led SciReasoner to use binding pockets, coordination environments and periodic connectivity as evidence for its predictions. To our knowledge, this is the **first foundation model to enable sequence, structure and natural-language reasoning** across all four scientific modalities within a single autoregressive trajectory, delivering specialist-level accuracy, interpretable structure-grounded chains of thought, and open-ended scientific language generation that no prior specialist or generalist model can provide.

2.2 Benchmarking SciReasoner across proteomic, molecular, and materials tasks

2.2.1 SciReasoner precisely predicts protein GO term and learns biologically meaningful structure

Assigning Gene Ontology (GO) terms [33, 34] to protein sequences is central to computational biology, translating rapidly growing sequence and structure databases [8] into interpretable functional annotations [35]. Since only a small fraction of proteins have experimentally verified functions [36, 37], accurate prediction across Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) remains critical for protein characterization, pathway analysis, disease-gene interpretation, and therapeutic-target discovery [38, 39].

We evaluated GO-term prediction on DeepFRI-GO [39] and, to prevent label leakage, excluded from continued pretraining all proteins with $>30\%$ sequence identity to the DeepFRI-GO test set. As shown in Fig. 2A, SciReasoner attains an overall ($F_{\max} = 0.59$), outperforming sequence-based BLAST transfer (0.55) [30], structure-based Foldseek 3Di alignment (0.54) [40], fine-tuned protein language models including ESM2 (0.53) [29] and SaProt 650M AF2 (0.52) [28], and general-purpose reasoning LLMs, DeepSeek-V4-Pro (0.35) and GPT-5.5 (0.31). The gain is strongest for CC, where SciReasoner reaches 0.58 versus BLAST (0.49), and remains positive for BP (0.52 versus 0.51). MF is near saturation, with SciReasoner at 0.66, comparable to SaProt (0.67).

Overall, SciReasoner is the only method that leads or matches the best baseline across all branches, indicating that it integrates evidence beyond any single baseline family.

To determine whether these gains reflect genuine functional inference rather than memorized homology-based transfer, we stratified the test proteins by maximum BLAST percent identity to the training set ($e \leq 10^{-3}$, query coverage $\geq 50\%$, self-hits removed). The advantage of SciReasoner is largest precisely where sequence similarity becomes unreliable. In the CC (0, 30]% identity bin, SciReasoner improves over BLAST by +0.21 absolute F_{\max} (0.55 versus 0.34 in Fig. 2A), and outperforms ESM2 by +0.13 absolute F_{\max} (0.55 versus 0.42). This

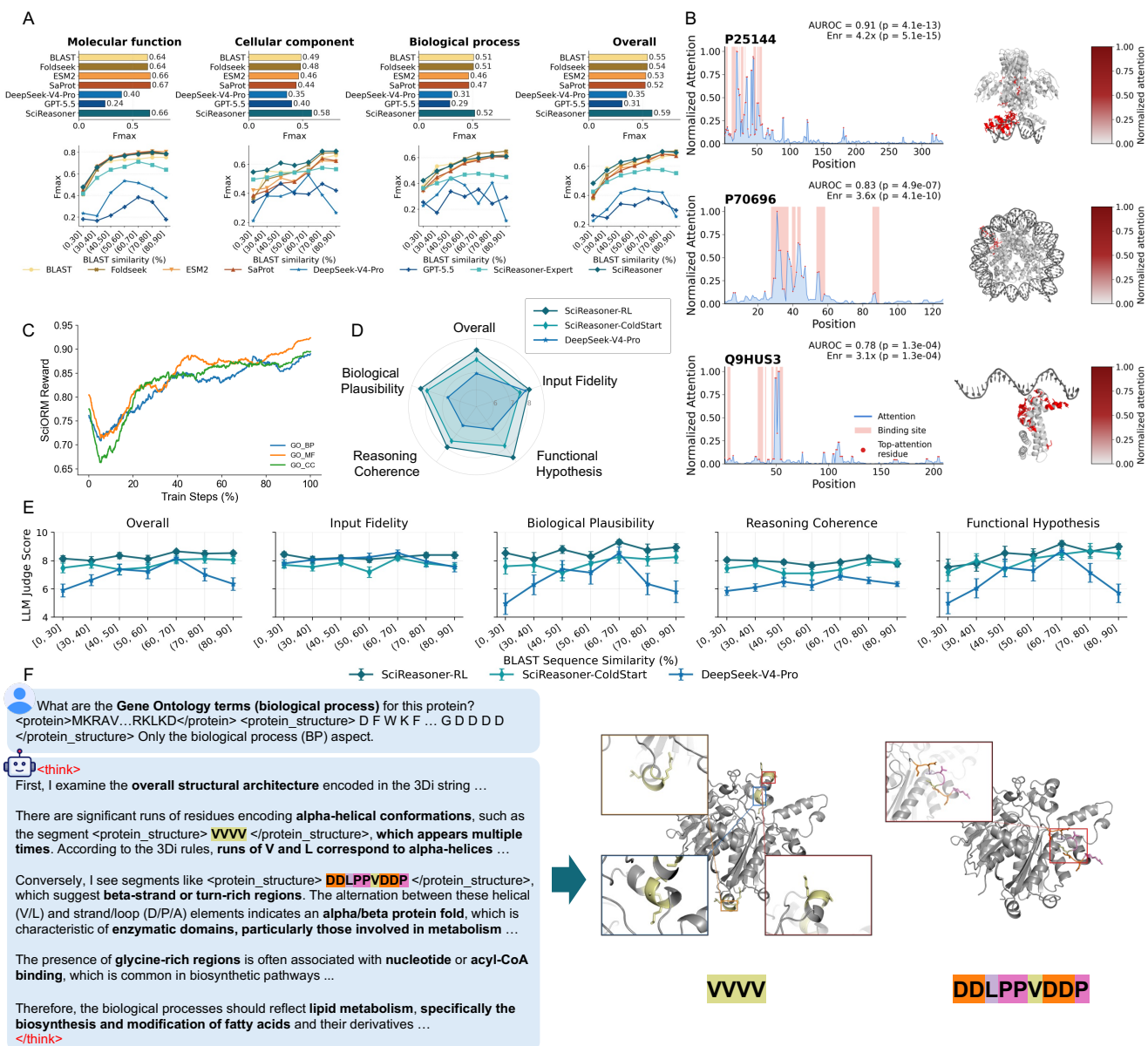


Fig. 2: SciReasoner predicts protein Gene Ontology terms through structure-grounded reasoning. (A) Performance on the CAFA-3 Gene Ontology benchmark for Molecular Function, Biological Process and Cellular Component annotations, stratified by maximum BLAST sequence identity to the training set. SciReasoner shows the largest gains in low-homology regimes, particularly for Cellular Component prediction. (B) Attention analysis for DNA-binding Gene Ontology prediction. Residues receiving high attention are enriched at contact-defined DNA-binding sites and localize to protein–DNA interfaces in representative complexes. (C) Reinforcement-learning trajectories for the three ontology aspects. Rewards increase after an initial exploration phase and converge to stable plateaus, indicating consistent optimization across molecular-function, biological-process and cellular-component annotations. (D) LLM-as-judge evaluation of reasoning traces on held-out Biological Process proteins. Reinforcement learning improves input fidelity, biological plausibility, reasoning coherence and functional-hypothesis quality relative to the coldstart checkpoint and DeepSeek-V4-Pro. (E) Reasoning-quality scores stratified by BLAST similarity for Biological Process prediction. Mean GPT-5.5 judge scores are reported on five rubric axes across seven similarity bins relative to the training set ($n = 20$ proteins per bin; error bars, s.e.m.). SciReasoner-RL achieves the highest scores across axes and remains stable across similarity bins, suggesting that its reasoning generalizes beyond memorized homologues. (F) Structure-token visualization of the reasoning trace for *E. coli* β -ketoacyl-ACP synthase I (FabB; 2bz4.A, UniProt P0A953). Load-bearing claims and cited 3Di-token fragments are highlighted in the trace and mapped onto PyMOL-rendered structures. The VVVV fragment maps to canonical α -helices, whereas DDLPPVDDP maps to an extended-strand/turn region, supporting the inferred α/β fold and the resulting fatty-acid biosynthesis annotation.

low-homology advantage aligns with the design of SciReasoner: whereas BLAST relies on local sequence similarity, ESM2 encodes evolutionary and sequence-context patterns without explicit structural grounding, SciReasoner predicts GO terms through reasoning traces that integrate domain composition, localized motifs, structural environments, and reference proteins. Such local functional cues can remain conserved even when global sequence identity is weak.

We next tested whether SciReasoner grounds GO predictions in functionally relevant protein regions rather than superficial sequence patterns. Using DNA-binding proteins, where protein–DNA contacts provide structural ground truth, we extracted attention from the generated DNA-binding GO token (GO:0003677) to residue-level structure and sequence tracks and combined them into a per-residue profile. These profiles strongly align with DNA-binding sites (Fig. 2B): catabolite control protein A (UniProt: P25144) reaches AUROC (=0.91) and top-20% fold enrichment (= 4.2 \times), histone H2B type 1-A (UniProt: P70696) reaches AUROC (=0.83) and enrichment (= 3.6 \times), and DesT (UniProt: Q9HUS3) reaches AUROC (=0.78) and enrichment (= 3.1 \times). In all cases, the most attended residues fall within or immediately adjacent to annotated binding regions, and structural projections confirm that these residues cluster near bound DNA in three-dimensional space. Thus, SciReasoner does not merely assign the correct GO term. It concentrates its evidence on residues that are physically responsible for the corresponding molecular function.

Reinforcement learning further improves GO prediction across all three ontology aspects (Fig. 2C). After a brief exploration phase during the first \sim 5% of training, in which the science reward temporarily decreases under noisy advantages, all three trajectories rise and converge to stable plateaus. MF reaches the highest reward (\sim 0.92), while BP and CC converge at comparable values (\sim 0.89). The simultaneous improvement across all aspects indicates that RL does not merely exploit aspect-specific shortcuts, but instead strengthens a shared protein-function representation that supports functional, process, and localization annotations.

Because F_{\max} captures only final GO-term overlap, we also evaluated the reasoning traces underlying those predictions. To assess grounded biological inference beyond surface fluency, we used GPT-5.5 to score held-out BP traces on Overall Quality, Input Fidelity, Biological Plausibility, Reasoning Coherence, and Functional Hypothesis Quality. SciReasoner scores highest on every axis, with an Overall score of 8.33 versus 7.77 for SciReasoner-ColdStart and 6.96 for DeepSeek-V4-Pro [26] (Fig. 2D). Relative to the pre-RL coldstart model, the largest improvement is in Biological Plausibility (8.68 vs. 7.83), indicating that RL sharpens not only the fluency of the trace but also the biological relevance of the inferred function. The gap to DeepSeek-V4-Pro is concentrated on the biology-specific axes—Biological Plausibility (8.68 vs. 6.63) and Functional Hypothesis Quality (8.45 vs. 6.77)—and is far wider there than on Input Fidelity (8.27 vs. 8.04). Although DeepSeek-V4-Pro is a substantially larger general-purpose model, this pattern shows that domain-tuned scientific reasoning, rather than raw model scale alone, drives the improvement.

We stratified reasoning quality by BLAST similarity to test whether SciReasoner relies mainly on close homologues (Fig. 2E). Across seven similarity bins, SciReasoner maintains higher judge scores than both the coldstart checkpoint and DeepSeek-V4-Pro, including in low-similarity regimes where direct homology transfer is least reliable. This helps explain the low-homology gains in Fig. 2A. Rather than copying annotations from close neighbours, SciReasoner can combine weaker but conserved cues, such as local structural motifs, short sequence fragments, fold-level patterns, and biochemical context. These results suggest that its improved performance on low-similarity proteins arises from structure-grounded functional reasoning rather than memorization of homologous sequences. Finally, Fig. 2F illustrates how the model’s reasoning can be audited at the level of structure tokens. For *E. coli* β -ketoacyl-ACP synthase I (UniProt: P0A953), SciReasoner identifies repeated VVV fragments as canonical α -helical regions and the DDLPPVDDP fragment as an extended strand/turn-rich segment. The trace then links this fold architecture, together with glycine-rich and catalytic-context cues, to fatty-acid biosynthesis and related lipid-metabolic processes. This example demonstrates that the reasoning trace is not an isolated natural-language explanation: its claims can be connected back to explicit structural substrings and spatial regions of the protein. Together, these results show that SciReasoner advances GO-term prediction in three ways: it improves accuracy most in low-homology regimes, grounds predictions in localized sequence and structural evidence, and uses reinforcement learning to enhance both annotations and reasoning quality. Thus, SciReasoner serves not only as a more accurate GO predictor, but also as a mechanism-aware annotation framework that explains molecular function, biological process, and cellular localization assignments.

2.2.2 SciReasoner formulates chemically viable retrosynthetic routes with high precision

Retrosynthesis (the recursive disconnection of a target molecule into commercially available precursors) is the cognitive backbone of organic-synthesis route planning. Reliable retrosynthetic models accelerate hit-to-lead optimization, and an interpretable reasoning trace lets human chemists verify and reuse the suggested routes.

We evaluate on the standard Retrosynthesis USPTO-50K [41] split: the model is given a single target SMILES and must output the set of reactant SMILES that produces it. We sample 16 stochastic completions per query ($T=0.6$, top- $p=0.95$) and rank the answers by sample frequency. A prediction is correct only when its canonical SMILES exactly matches the ground-truth reactant set, and we report Exact Match as the standard metric. Fig. 3A places SciReasoner against 18 published baselines spanning template-based [42–46], semi-template-based [47–52] and template-free [24, 31, 53–57] families: SciReasoner reaches 0.72 (Exact Match), exceeding the previous best

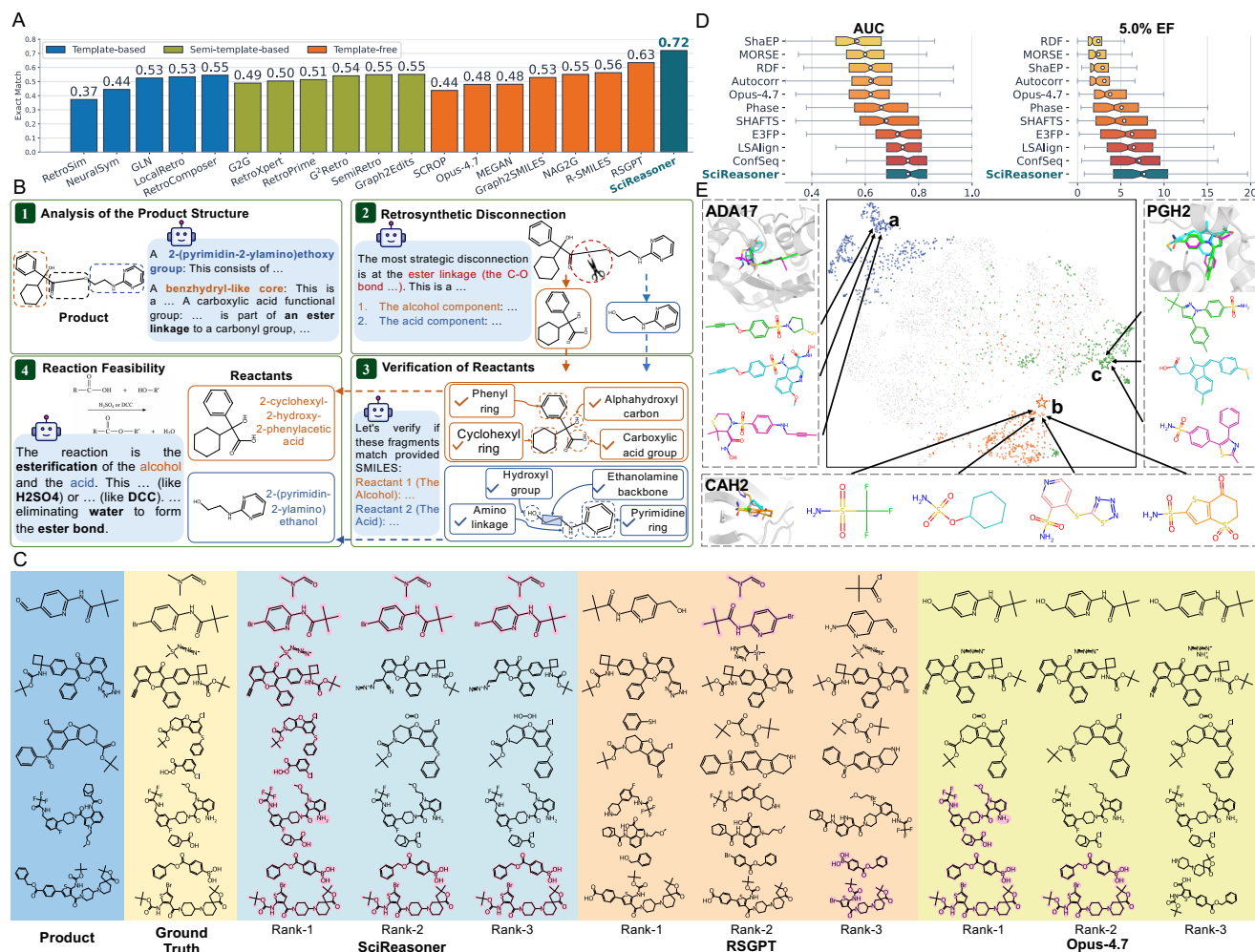


Fig. 3: SciReasoner formulates chemically viable retrosynthetic routes (A–C) and captures plausible 3D molecular structure for similarity representation (D, E). (A) Retrosynthesis USPTO-50K performance. Across template-based, semi-template-based and template-free methods, SciReasoner reaches 0.72, exceeding the prior best RSGPT by +0.09 points, while Opus-4.7 five-shot scores only 0.48. (B) Chain-of-thought trace for a representative product. SciReasoner proceeds in four interpretable stages—Analysis, Disconnection, Verification and Feasibility, and recovers the two ground-truth reactants by severing the strategic ester C–O bond. Every intermediate claim is itself a SMILES fragment, making the trace auditable atom-by-atom. (C) Cross-model Top-3 reactant predictions on Retrosynthesis USPTO-50K. Five products with their ground-truth reactants and the Top-3 predictions of SciReasoner, RSGPT and Opus-4.7. Molecules in pink match the gold canonical SMILES. SciReasoner recovers a correct candidate for 5/5 products versus 2/5 for RSGPT and 2/5 for Opus-4.7, generalising across mechanistically distinct reaction families, rather than fitting a single template. (D) Per-target distributions on 3D molecule structure (DUD-E). AUC and 5.0% EF box plots over 102 DUD-E targets for nine published baselines, an Opus-4.7 reference, and SciReasoner. SciReasoner matches the previously reported best AUC and establishes a new best 5.0% EF. (E) 3D ligand-target geometry for three DUD-E targets. UMAP of the per-molecule SciReasoner embedding, coloured points are the 400 closest actives for each target, labelled by DUD-E target identifiers—ADA17 (blue), CAH2 (orange), and PGH2 (green); open stars mark the reference queries. Side panels show each query and three representative actives docked into the receptor, then redrawn as 2D structures: chemically distinct scaffolds that nevertheless converge to the same 3D pharmacophore.

template-free method RSGPT [31] (0.63) by +0.09 points, while Opus-4.7 five-shot scores only 0.48. To prevent direct test leakage, we removed from the pretraining corpus all reactions whose product SMILES matched any product in the Retrosynthesis USPTO-50K test set.

Unlike generic chat models that emit a free-form natural-language explanation followed by a single SMILES guess, SciReasoner is a scientific reasoning model that interleaves natural-language steps with *structural tokens* (SMILES fragments) inside its chain-of-thought, e.g. the ester C–O bond, the phenyl ring c1ccccc1. Because the intermediate steps are themselves chemical structures, the trace can be checked atom-by-atom and the final reactant assembly is grounded in concrete sub-fragments.

Fig. 3B illustrates this on a hydroxy-substituted benzhydryl ester. SciReasoner severs the strategic ester C–O bond, independently verifies the expected functional groups on each precursor, and recognises the inverse reaction as an acid-catalysed esterification. Six of the seven sub-fragments emitted in the trace are strictly smaller than

either reactant, yet their union exactly reconstructs the gold answer, indicating bottom-up reasoning over chemical groups rather than template recall.

Fig. 3C extends this to five further Retrosynthesis USPTO-50K products, contrasting SciReasoner against the specialised retrosynthesis model RSGPT and the general-purpose large language model Opus-4.7. SciReasoner contains the gold reactant set in its Top-3 for 5/5 rows; RSGPT recovers it in 2/5, and Opus-4.7 in 2/5.

On the *N*-tert-butyl pyridine-carboxamide aldehyde, SciReasoner recovers the literature Vilsmeier formylation route (DMF/Br precursor pair) with a single canonical answer across all 16 samples; RSGPT proposes a hydroxymethyl + bromopyridine pair, while Opus-4.7 commits all three ranks to oxidation of a benzyl alcohol and never proposes the C–N disconnection (Fig. 3C, Row 1). On a 1,4-disubstituted 1,2,3-triazole-fused phenyl-benzofuranone, the gold cut is a copper-catalysed azide–alkyne [3+2] cycloaddition (CuAAC) of trimethylsilyl azide with the terminal aryl-alkyne; only SciReasoner’s rank-1 recovers this pair, while RSGPT keeps the triazole intact and disconnects elsewhere and Opus-4.7 mis-treats the cycloaddition as an azide-plus-nitrile coupling (Fig. 3C, Row 2). On a chloro-arylsulfanyl tetrahydrofuro-pyridine *N*-Boc carbamate, the strategic cut is *m*-CPBA oxidation of the thioether to the sulfoxide; only SciReasoner’s rank-1 identifies it, while RSGPT and Opus-4.7 instead attempt to cleave the Boc carbamate or the tetrahydrofuran (Fig. 3C, Row 3). On a polycyclic indole-based amide, SciReasoner and Opus-4.7’s rank-1 both propose the correct amide coupling between the bicycloalkyl carboxylic acid and the parent aniline; the remaining baselines fragment the indole or the trifluoromethyl chain (Fig. 3C, Row 4). On a thiophene–phenyl biaryl bearing a Boc-aminothiophene and a spirolactone-piperazine amide, the strategic cut is a Suzuki–Miyaura coupling of the 5-bromo Boc-aminothiophene with benzyl 4-boronate-benzoate; SciReasoner returns this pair across all 16 samples and Opus-4.7’s rank-1/rank-2 both match, while RSGPT recovers the gold pair only at rank-3 (Fig. 3C, Row 5).

Taken together, the structural-token chain-of-thought makes SciReasoner’s predictions both more accurate and easier to audit: the trace states inside itself why each bond is the strategic cut, rather than answering in one step.

2.2.3 SciReasoner captures plausible 3D molecular structure for similarity representation

Identifying compounds that share a binding-pocket geometry but not a 2D scaffold is the central problem of virtual screening and ligand-based drug discovery. A representation that groups molecules by binding-pocket compatibility, rather than by simple substructure overlap, would directly enable hit expansion and scaffold hopping.

We probe whether SciReasoner itself already provides such a representation. From a single-molecule prompt, we extract the last layer hidden state of SciReasoner and average it over the 10 tokens generated immediately after the prompt to obtain a molecular embedding. Following the standard DUD-E protocol [32], we ask how well Euclidean distance in this embedding space ranks actives versus decoys for a given target query, reporting AUC and the 5.0% EF, and the embedding’s qualitative structure is visualised via 2D UMAP projection.

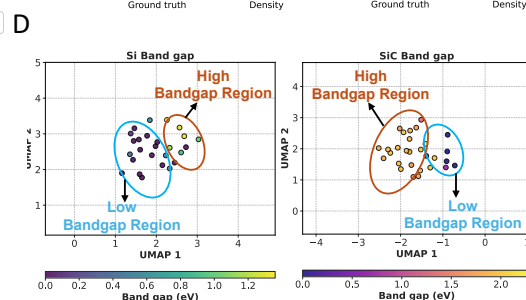
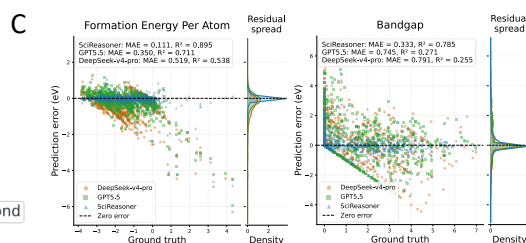
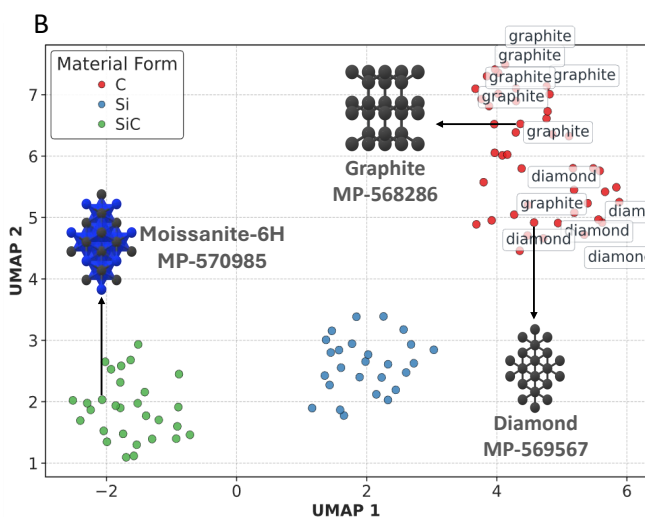
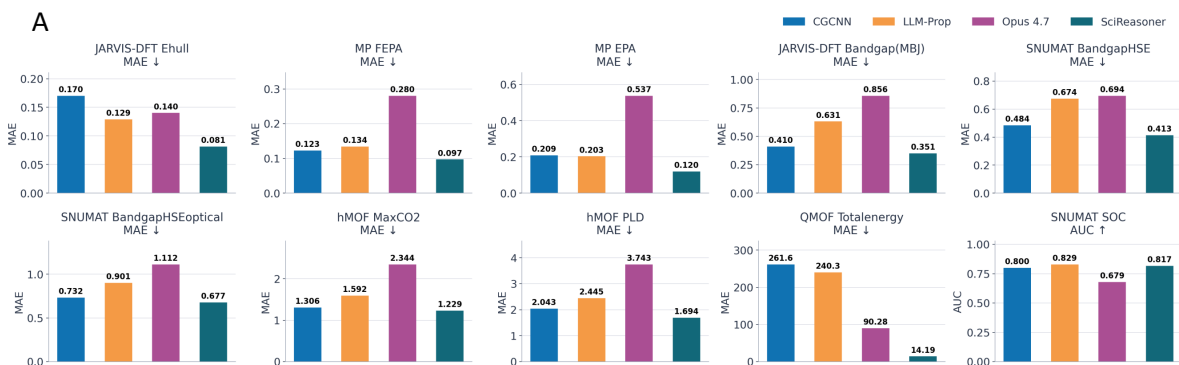
SciReasoner natively consumes *structural tokens*: a dedicated ConfSeq [58] tokenizer converts each 3D conformer into a sequence of geometry-aware tokens drawn from a separate vocabulary backed by a separate learned embedding table, held disjoint from the natural-language embedding. Because structural tokens are a parallel modality, geometric features are separated from those used for IUPAC names or descriptive language, so the hidden states above structural-token positions carry an uncontaminated geometric signal.

Fig. 3D quantifies this representation against nine published virtual-screening methods on 102 DUD-E targets, spanning alignment-free descriptors and alignment-based shape-overlay methods [58–66], plus an Opus-4.7 [24] reference. Used as a drop-in similarity feature, SciReasoner’s per-molecule embedding matches ConfSeq’s prior best AUC of 0.76 and improves the best 5.0% EF from 7.12 to 7.70, obtained without docking, scoring functions, or task-specific fine-tuning.

Fig. 3E demonstrates the qualitative structure of the same embedding on three pharmacologically distinct DUD-E targets spanning unrelated protein families: ADAM17 (a zinc metalloprotease), carbonic anhydrase II (a zinc hydrolase) and cyclooxygenase-2 / PGHS-2 (a heme-containing dioxygenase). Each target’s top-400 closest actives form a clearly separated cluster around its query ligand (★) on the 2D UMAP, and the three clusters occupy distinct regions of the plane despite the proteins being sequence-unrelated, and the decoys (light grey) do not infiltrate any cluster, indicating that the embedding captures family-specific chemistry rather than gross descriptors.

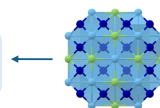
The boxed side panels visualise the per-target binding mode in 3D. For ADAM17, the cyclic-thiol sulfonamide query is matched by closest actives that swap the thiopyrrolidine zinc-binder for a hydroxamate amide or extend the alkyne tail into a propargyl amine while keeping the diaryl sulfonyl core; their Morgan-fingerprint Tanimoto similarities to the query are only 0.30–0.32, yet USRCAT 3D-shape similarities remain greater than 0.22 and Open3DALIGN scores reach 120–140, meaning all three ligands drape over the same TACE active site. For PGHS-2, the celecoxib-like diaryl-pyrazole sulfonamide query is matched by an indanone-acetic-acid analog (Tanimoto 0.11) and a chloro-aryl-thiazole-sulfonamide (Tanimoto 0.31), chemically divergent COX-2 chemotypes that adopt the same 3D pharmacophore. Carbonic anhydrase II is the most striking case: the query reduces to the minimal NH₂-SO₂-CF₃ warhead, and its three closest actives—a cyclohexyl sulfamate, a pyridyl-thiadiazole sulfonamide and a thiopheno-thiazinone, spanning very different fused-ring topologies yet all expose the canonical zinc-binding sulfonamide that defines CA-II inhibition.

Taken together, both qualitative and quantitative results indicate that SciReasoner has implicitly learned to cluster DUD-E actives by binding-pocket geometry rather than scaffold similarity.



E

Look at the chemical formula and structure information of the given crystalline material and predict its property. Is the following material (TiGaCo₂) stable?



<think>To determine the stability of the crystalline material with formula TiGaCo₂, I need to analyze the provide structure ...The token sequence <material_structure>+ t a ODD a DOD b 000 c 000 d 000 e 000 </material_structure> maps directly to the **space group Fm-3m, a high-symmetry system often associated with stable intermetallic compounds** ... Next, I parse the atom list to verify the stoichiometry... Then, I investigate the bonding network to understand the local coordination environment. The bond <material_structure>0 12 --o</material_structure> connects a titanium atom to a cobalt atom across a unit cell boundary ... <material_structure>0 6 ooo</material_structure> shows a titanium-gallium bond within the same cell ... The density of these edges, such as <material_structure>1 11 o--</material_structure> ...<material_structure>4 9 ---</material_structure> ... This **heteroatomic bonding topology is crucial for stabilizing the structure through electronic interactions** ... The combination of the Fm-3m space group, the 1:1:2 stoichiometry, and the fully connected bonding network is characteristic of the L2_1 Heusler structure type. Therefore, the structural evidence points strongly towards a stable material configuration. </think>

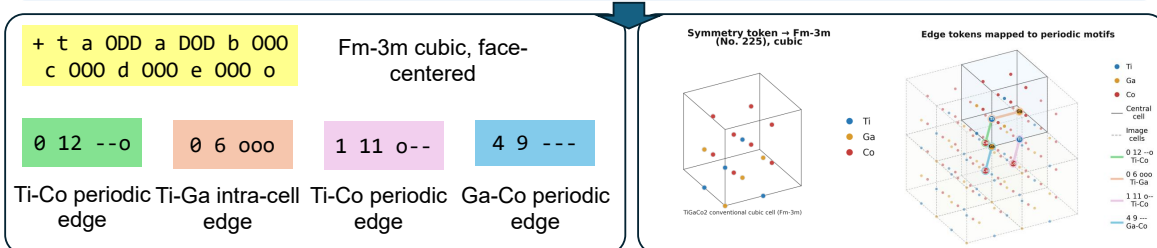


Fig. 4: Evaluation of material predictive performance and visualization of the learned structure-property latent space. (A) Performance comparison across 10 material property prediction sub-tasks, both regression (MAE as the metric) and classifications (AUC as the metric) tasks are reported. (B) UMAP of the high-dimensional representations extracted by the model on three materials (C, Si and SiC). The latent space inherently clusters materials by chemical space and spontaneously distinguishes distinct structural polymorphs. (C) Parity plots comparing model-predicted formation energies per atom and bandgap against ground-truth values. The strong correlations ($R^2 = 0.895$ and 0.785) demonstrate the model's ability to capture physically meaningful materials-property information. (D) Local UMAP embeddings for the Si and SiC subsets, color-coded by their corresponding band gaps. The continuous transition and distinct grouping of high- and low-bandgap regions indicate that the learned representations effectively capture underlying structure-property relationships. (E) Structure-grounded reasoning trace for a representative stability prediction on TiGaCo₂. The model's reasoning explicitly identifies the SLICES symmetry substring, which decodes to the Fm $\bar{3}$ m space group, and cites edge substrings. These substrings are mapped back to the corresponding CIF structure, where they visualize the global cubic symmetry framework and local periodic connectivities.

2.2.4 SciReasoner learns structure- and property-aware representations for materials tasks

We evaluated SciReasoner on ten downstream materials sub-tasks from five complementary databases: Materials Project¹, JARVIS-DFT², SNUMAT³, hMOF⁴, and QMOF⁵. These sub-tasks cover inorganic crystals, semiconductors and metal organic frameworks, and include thermodynamic stability, formation and total energies, electronic band gaps, spin orbit coupling, CO₂ uptake and pore geometry. For these sub-tasks, SciReasoner was compared with CGCNN [9], LLM-Prop [67] and Opus-4.7, using Mean Absolute Error (MAE) for regression tasks and Area Under the Curve (AUC) for classification tasks. As shown in Fig. 4A, SciReasoner outperforms CGCNN across all ten tasks and exceeds LLM-Prop on most numerical properties. These results indicate that its representations capture both general energetic trends and more specialized relationships involving electronic structure, relativistic effects, adsorption behaviour and pore geometry.

We next examined the latent space learned by SciReasoner. UMAP projections show clear separation among C, Si and SiC structures, indicating that the embeddings encode global chemical identity (Fig. 4B). Within each compositional domain, structures do not collapse into a single cluster. Instead, they form extended subspaces that distinguish polymorphs and polytypes with the same composition, suggesting sensitivity to local geometry and bonding topology. At the quantitative level, SciReasoner produces DFT-consistent predictions for both formation energy and band gap, with parity plots showing close agreement with ground-truth values and residuals concentrated around zero (Fig. 4C). This predictive accuracy is reflected in the learned representation space: local UMAP projections for Si and SiC separate low- and high-band-gap structures while preserving continuous property gradients within each compositional manifold (Fig. 4D). Together, these results show that SciReasoner organizes structurally related materials along physically meaningful electronic-property manifolds, rather than encoding composition alone.

A representative TiGaCo₂ stability prediction illustrates how SciReasoner grounds its reasoning in explicit structural tokens rather than relying only on the chemical formula (Fig. 4E). The model identifies the structure substring corresponding to the Fm $\bar{3}$ m space group and links local edge substrings to periodic Ti-Co, Ti-Ga and Ga-Co connectivities in the reconstructed CIF structure, providing a traceable connection between token-level evidence and crystallographic features.

2.3 Structural evidence drives native scientific reasoning

To test whether native scientific reasoning depends on explicit structural evidence, we performed structure-ablation experiments across materials, proteins and small molecules (Fig. 5A). Removing structural information consistently weakened performance, whereas providing structural inputs improved prediction across all three domains. The gains were especially pronounced for protein tasks, where function depends on folded geometry, binding-site organization and spatial residue interactions, but were also evident in materials and molecular-similarity tasks, where periodic topology and three-dimensional shape provide information that cannot be recovered from composition, sequence or two-dimensional descriptors alone. Thus, the ablation shows that structural evidence is not merely an additional input channel, but a necessary substrate for the model’s native scientific reasoning.

The embedding analysis further shows that SciReasoner learns a distinct representational space for structural evidence (Fig. 5B). In the Qwen embedding space, paired protein sequence and structure inputs remain partially intermixed, suggesting that structural descriptions are treated largely as another form of text. By contrast, SciReasoner forms more clearly separated sequence and structure clusters, indicating that structural tokens occupy a distinct representational subspace while preserving within-modality consistency. This modality separation suggests that the model can distinguish sequence-level evidence from structure-level evidence, a prerequisite for reasoning tasks that require explicit modelling of sequence–structure relationships.

Case studies show that structural evidence changes the reasoning path, not merely the final output. For QMOF pore-limiting-diameter prediction, the structure-free model relies on compositional and stoichiometric priors and overestimates the pore size by nearly an order of magnitude (Fig. 5C). With structural tokens, SciReasoner instead cites monoclinic symmetry, cobalt coordination to nitrogen and oxygen atoms, edge connectivity and periodic offsets, producing a prediction close to the ground truth. Similarly, for biological-process GO prediction, sequence-only reasoning is diverted by cationic and histidine-rich motifs towards DNA recombination and stress-response terms, whereas structure-aware reasoning integrates secondary-structure context and binding-pocket evidence to recover protein-modification annotations (Fig. 5D). These examples indicate that structural evidence redirects the model from superficial sequence or composition cues towards physically and biologically meaningful explanations. Quantitative analyses support the same conclusion. For materials prediction, structural input reduces absolute errors for bandgap, largest-cavity-diameter and pore-limiting-diameter prediction (Fig. 5E), showing that crystal geometry improves both electronic-structure and pore-geometry inference. For GO molecular-function prediction, residue-level attention maps show that structure-aware reasoning concentrates high attention around the functional

¹<https://materialsproject.org/>

²<https://jarvis.nist.gov/jarvisdft>

³<https://www.snumat.com/>

⁴<https://mof.tech.northwestern.edu/>

⁵<https://contribs.materialsproject.org/projects/qmof>

binding site, whereas sequence-only reasoning attends to non-binding regions and yields lower accuracy (Fig. 5F). Together, the ablations, representation analysis, reasoning cases and attention maps support the claim that structural evidence drives native scientific reasoning in SciReasoner, enabling the model to ground its conclusions in biologically and physically meaningful structural features rather than superficial sequence, composition or textual correlations.

2.4 Intra-domain grounding and cross-domain consolidation stimulate self-bootstrapped native structural reasoning

While broad scientific pre-training provides diverse molecular, biomolecular, and materials representations, it does not establish a unified reasoning protocol across scientific tasks. We therefore adopt a self-bootstrapped post-training framework, which first grounds structural evidence within each scientific domain (i.e., intra-domain structural evidence grounding) and then integrates these domain-local reasoning patterns into a single scientific reasoner (i.e., cross-domain reasoning consolidation) (Fig. 6A). Sketch traces provide only a coldstart prior, after which domain-structure experts generate higher-quality on-policy traces that are pooled for final training. Together, this two-stage procedure turns pre-trained structural priors into explicit scientific reasoning, first eliciting domain-native evidence use and then consolidating it into a unified structure-grounded reasoner.

The reward and performance dynamics in Fig. 6B support the self-bootstrapped strategy. Intra-domain structural evidence grounding yields consistent reward gains across task families, indicating that a small set of sketch traces can seed trainable reasoning behaviours. After expert-generated traces are pooled, cross-domain reasoning consolidation enables the unified SciReasoner to outperform the coldstart baseline across both structure-intensive and non-structure tasks. This suggests that expert-generated traces better align with the model’s representation space, enabling model-native reasoning trajectories to emerge.

Fig. 6C shows that the post-training procedure improves both first-sample accuracy and sampling efficiency across tasks. For pass@10, we used task-specific aggregation: the mean Fmax@10 across BP, MF and CC for GO annotation, an oracle best-of-10 prediction selected by minimum absolute error for QMOF regression, a label-independent 10-sample score ensemble followed by 5.0% EF evaluation for 3D molecule structure (DUD-E), and top-10 reactant-set accuracy for retrosynthesis. In GO term prediction, intra-domain structural evidence grounding raises pass@1 from 0.48 to 0.52 and reduces the pass@10–pass@1 gap from 0.13 to 0.11, while cross-domain reasoning consolidation further achieves the highest pass@1 of 0.59 with a smaller gap than its coldstart precursor. In QMOF prediction, pass@1 increases from 2.71 to 4.76 after intra-domain structural evidence grounding and from 6.80 to 8.61 after cross-domain reasoning consolidation, with the gap reduced from 11.29 to 4.07 and from 13.82 to 5.57, respectively. 3D molecule structure (DUD-E) prediction shows the same trend: pass@1 improves from 4.65 to 6.31 after intra-domain structural evidence grounding and from 7.29 to 7.70 after cross-domain reasoning consolidation. Retrosynthesis shows the largest absolute gain, with pass@1 rising from 0.41 to 0.49 after intra-domain structural evidence grounding and to 0.72 after cross-domain reasoning consolidation, accompanied by gap reductions from 0.36 to 0.26 and from 0.22 to 0.09. These consistent gap reductions indicate that the procedure does not merely improve best-of- N performance, but concentrates probability mass on correct first responses, making SciReasoner more reliable under single-sample scientific prediction.

Fig. 6D further evaluates reasoning quality beyond final-answer accuracy using LLM-judge score distributions over 100 records per task. Self-bootstrapped native structural reasoning consistently shifts the distributions toward higher scores, with the largest worst-to-best mean improvement on 3D molecule structure (DUD-E) ($\Delta_{\text{mean}} = +3.10$), followed by GO protein-function annotation (+1.79), QMOF property prediction (+1.61), and Retrosynthesis USPTO-50K (+0.48). The larger gains on DUD-E, GO and QMOF suggest that the self-bootstrapped procedure is especially effective when reasoning must be grounded in molecular topology, protein structural evidence or crystal-property relationships. By contrast, retrosynthesis starts from an already high reasoning-quality distribution, leaving less room for improvement. Its smaller but positive shift therefore indicates slower saturation rather than failure of the post-training procedure. Overall, these distribution-level gains show that the procedure improves not only answer correctness, but also the coherence and evidence grounding of the scientific reasoning process.

Fig. 6 shows that self-bootstrapped native structural reasoning progressively transforms domain-structure experts into a unified scientific reasoner. Intra-domain structural evidence grounding strengthens domain-specific reasoning, while cross-domain reasoning consolidation transfers these capabilities into a single cross-domain model. Improvements in reward, pass@1 performance, and reasoning quality demonstrate more accurate and stable evidence-grounded reasoning.

2.5 Human Expert Evaluation

The automated metrics and LLM judges establish that SciReasoner produces accurate answers, but they cannot establish whether its *reasoning traces* are genuinely useful to domain experts. We therefore conducted a double-blinded human evaluation pilot in which domain experts compared SciReasoner and DeepSeek-V4-Pro reasoning traces on cases spanning three tasks — GO protein-function annotation, crystalline-material property prediction, and single-step retrosynthesis on Retrosynthesis USPTO-50K. Evaluators were double-blinded to model identity

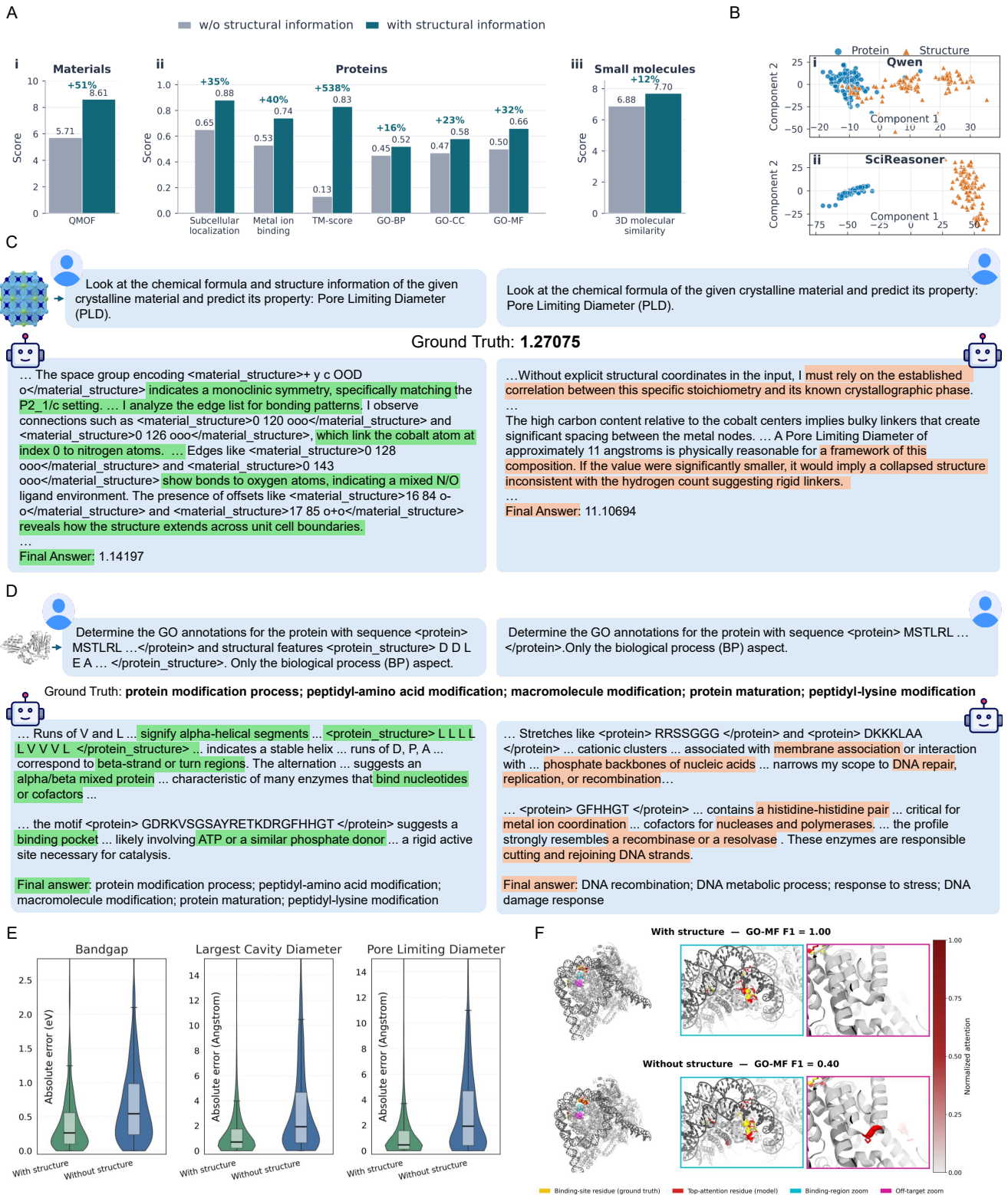


Fig. 5: Structural information empowers native scientific reasoning. (A) Structure-ablation benchmark. Performance with and without structural inputs across materials, proteins and small molecules. Structural cues consistently improve prediction beyond sequence- or text-only inputs. **(B) Modality-separated representations.** PCA of Qwen and SciReasoner embeddings for paired protein sequence and structure inputs. SciReasoner more clearly separates the two modalities, indicating stronger structure-aware encoding. **(C) Structure-grounded materials reasoning.** Chain-of-thought comparison for QMOF pore-limiting-diameter prediction. Without structure, the model relies on compositional priors and overestimates pore size; with structural tokens, it uses symmetry and bonding connectivity to recover a value closer to the ground truth. **(D) Structure-grounded GO reasoning.** Chain-of-thought comparison for biological-process GO prediction. Sequence-only reasoning predicts DNA-related processes, whereas structure-aware reasoning integrates secondary-structure and binding-site evidence to recover protein-modification annotations. **(E) Materials prediction with structure.** Absolute-error distributions for bandgap, largest-cavity-diameter and pore-limiting-diameter prediction. Structural input reduces errors across all three properties. **(F) Structure-guided GO attention.** Residue-level attention maps for GO molecular-function prediction. Structure-aware reasoning focuses on the functional binding site, whereas sequence-only reasoning attends to non-binding regions, consistent with its lower accuracy.



Fig. 6: Self-bootstrapped native structural reasoning unifies scientific reasoning. (A) Intra-domain structural evidence grounding and cross-domain reasoning consolidation. Sketch reasoning traces coldstart task-specific experts, which are optimized independently by reinforcement learning (RL) to ground structural evidence within each domain. Expert-generated traces are then pooled to coldstart and reinforce a unified SciReasoner across all tasks to consolidate cross-domain reasoning patterns. **(B) Reward dynamics.** RL rewards rise during training, with the unified SciReasoner matching or exceeding expert variants on structure-intensive and other scientific tasks. **(C) Improved accuracy and sampling efficiency.** Pass@1 and pass@10 are compared across four training stages for GO term prediction, QMOF prediction, 3D molecule structure (DUD-E) and Retrosynthesis USPTO-50K. Lighter hatched bars indicate pass@1 and solid bars pass@10; $\Delta = \text{pass@10} - \text{pass@1}$. The post-training procedure consistently increases pass@1 and reduces Δ , indicating more reliable first-sample predictions. **(D) Improved reasoning quality across tasks.** Kernel density estimates show LLM-judged reasoning scores for four model variants on GO term prediction, QMOF prediction, 3D molecular similarity (DUD-E), and Retrosynthesis USPTO-50K. Triangles denote means and red brackets indicate the worst-to-best mean difference. Self-bootstrapped native structural reasoning shifts score distributions towards higher quality, and unified SciReasoner generally outperforms domain-structure experts. **(E) Double-blinded human-expert evaluation.** Domain experts compared SciReasoner and DeepSeek-V4-Pro on GO annotation, materials-property prediction and retrosynthesis. Left, pairwise preferences overall and by task; right, score distributions for individual scoring-axis. SciReasoner was rated tie-or-better in 98% of comparisons and scored higher on every axis.

and rated every trace against a five-axis, 1–10 rubric (Q1 evidence grounding, Q2 domain plausibility, Q3 target-regime alignment, Q4 reasoning coherence, Q5 anti-hallucination) together with a five-point pairwise preference (Q8, strongly prefer SciReasoner \rightarrow strongly prefer DeepSeek-V4-Pro). The two panels of Fig. 6E summarize the resulting $N = 177$ ⁶ reliable case-judgments: the head-to-head preference distribution and the per-axis score distributions.

In the head-to-head comparison, experts rated SciReasoner tie-or-exceed than DeepSeek-V4-Pro in 98% of case-judgements, with 73% strongly preferring SciReasoner, 21% preferring it, and 4% ties. Only a small residual fraction favoured DeepSeek-V4-Pro (Fig. 6E, pairwise preference). The preference is consistent across all three tasks — 100% tie-or-exceed on GO protein function, 95% on crystalline-material property prediction, and 98% on retrosynthesis. It demonstrates that SciReasoner’s advantage is not carried by any single domain but holds across heterogeneous scientific-reasoning settings.

The per-axis scores show that this preference reflects a broad, consistent quality gap rather than a single dominant dimension. SciReasoner averages 8.7/10 overall versus 4.3/10 for DeepSeek-V4-Pro, with per-axis means of 8.8 vs 4.5 on evidence grounding, 8.5 vs 4.4 on domain plausibility, 8.9 vs 4.3 on target-regime alignment, 8.7 vs 3.9 on reasoning coherence, and 8.8 vs 4.6 on anti-hallucination. Every per-axis difference is significant shown by Wilcoxon signed-rank on paired per-case differences, $P < 0.001$ in Fig. 6E, per-axis scores. SciReasoner’s distributions are concentrated near the top of the scale on all five axes, whereas DeepSeek-V4-Pro receives more variable scores, especially on target-regime alignment and reasoning coherence. This indicates that the most reliable separation is not that DeepSeek-V4-Pro uniformly fails, but that SciReasoner is more consistently grounded, task-aligned, and coherent across cases.

Together, these results show that SciReasoner’s advantage on automated metrics also holds under direct expert scrutiny: across three heterogeneous tasks, human experts prefer or tie SciReasoner reasoning in the overwhelming majority of cases, and rate it markedly higher on every quality axis, reflecting faithful, verifiable scientific reasoning.

3 Discussion

In this work, we introduced SciReasoner, a multimodal scientific foundation model for native structural reasoning across proteins, small molecules and inorganic crystals. The central premise is that structure–property relationships cannot be fully addressed by treating scientific entities as text strings, low-dimensional descriptors or black-box inputs to property predictors. SciReasoner instead represents structures as primary objects of inference through a unified structure-aware vocabulary that discretizes three-dimensional coordinates, molecular topologies and crystallographic lattices into domain-native tokens, and integrates them with language instructions in a single autoregressive model. This design allows residues, molecular fragments, conformational cues and crystal descriptors to function as addressable evidence units within generated reasoning trajectories, rather than as peripheral annotations appended to text.

Our results support native structural reasoning as a model-level capability for scientific AI. Rather than using structure only as an input format or auxiliary descriptor, SciReasoner makes residues, molecular fragments, conformational cues and crystal descriptors available as evidence within generated reasoning trajectories. This shift is important because structure–property analysis requires not only accurate outputs, but intermediate claims that can be inspected against the physical object being studied.

The strongest evidence for this capability comes from regimes in which superficial similarity is insufficient.

For Cellular Component annotation of low-homology and orphan-like proteins, SciReasoner improved F_{\max} from 0.42 to 0.55, with its attention enriched at contact-defined DNA-binding residues and protein–DNA interfaces. In retrosynthesis, it increased single-step accuracy from 0.63 to 0.72 while identifying reaction-relevant fragments, strategic bond disconnections and chemically valid precursors. In crystalline materials, its reasoning trajectories invoked symmetry-distinct sites, coordination environments and periodic bonding motifs when linking structure to electronic and thermodynamic behaviour. These behaviours indicate that structural organization functions as the working substrate of inference, not merely as a representation used to produce a final score.

This distinction clarifies the contribution of SciReasoner relative to existing scientific AI systems. A Foldseek-derived structural fragment can ground a fold- or function-level hypothesis; a molecular substructure can justify a retrosynthetic disconnection; a conformational cue can support a pharmacophore-level similarity judgement; and a crystal-connectivity token can support an inference about stability or electronic response. Because such intermediate claims are expressed in the reasoning trajectory, they can be checked against the input structure rather than accepted as opaque post hoc rationales. This helps bridge two previously separated capabilities: domain-native structural representation and explicit scientific reasoning. The double-blind expert evaluation, in which experts preferred or tied SciReasoner over a frontier large language model in 98% of cases, further suggests that this inspectability is practically useful for scientific assessment.

Overall, SciReasoner demonstrates that a single autoregressive foundation model can unify sequence, structure, and natural-language reasoning across major scientific modalities while achieving specialist-level accuracy and interpretable scientific explanation. This provides a route toward scientific AI systems that move beyond prediction toward reasoning over the physical structures from which functions, properties, and mechanisms emerge.

⁶We are collecting more human judgments.

4 Method

4.1 Data Source and Processing

4.1.1 Protein Data

We constructed the protein component of the pretraining corpus by following established protein-language data construction protocols. To incorporate textual and functional protein knowledge, we used SIFTS-based PDB–UniProt mappings as indexing anchors [35, 68, 69], protein sequences were associated with protein names, PubMed/PMC literature [70, 71] and curated UniProtKB/Swiss-Prot annotations, including functional descriptions, domain and subunit information, post-translational modifications, disease associations, induction information and tissue specificity when available.

For large-scale UniProt protein sequences, we retrieved the corresponding AlphaFold-predicted structures from AlphaFoldDB whenever available [72]. Following the SaProt pretraining setting [28, 73], each protein structure was converted into residue-level Foldseek 3Di tokens [40], which describe local structural environments and are aligned one-to-one with the amino-acid sequence. The amino-acid tokens and 3Di tokens were then separated into amino acid (AA) sequences (wrapped by `<protein></protein>` tag) and structural sequences (warped by `<protein_structure></protein_structure>` tag) in one input, allowing the model to learn from both primary sequence and predicted three-dimensional structure. Structurally unreliable regions (pLDDT<70) were masked or replaced by a special structural token to reduce noise from uncertain coordinates.

To prevent the leakage of labels in the benchmark data within the interleaved protein–text data, we excluded from continued pretraining all text-linked protein records with >30% MMseq2 [74, 75] sequence identity to any protein in the held-out test sets of the downstream benchmarks.

4.1.2 Small-molecule Data

We constructed the small-molecule component by integrating chemical text, molecular representations, structured property datasets and three-dimensional molecular conformations. The textual component was collected with US Chemistry Olympiad problems, Europe PMC full-text articles and abstracts, ChemRxiv, bioRxiv, medRxiv and arXiv submissions filtered for chemistry, medicinal chemistry and physical chemistry. These sources provide foundational chemical concepts, textbook-style explanations, research-level chemical knowledge, experimental context, safety descriptions and domain-specific terminology.

We further converted public molecular property datasets into language-interfaced templates, pairing molecular entities with curated descriptions of properties, assays, labels and task contexts. These sources include MoleculeNet [76], Therapeutics Data Commons and ADMET-style datasets [77], ChEBI, PubChem BioAssay, Tox21, hERG and CYP bioactivity datasets, CHEMDNER and other biomedical named-entity recognition corpora, and curated datasets covering ADMET, physicochemical properties, molecular toxicity and safety-related endpoints. Together, these resources use natural language to align standardized molecular representations with property labels, bioactivity assays, toxicity endpoints, pharmacokinetic profiles and biomedical chemical terminology. We also incorporated reaction, biochemical and spectral resources, including Open Reaction Database (ORD), USPTO-derived reactions, Buchwald–Hartwig datasets, Rhea/UniProt reaction annotations, MOSES [78], MoNA and spectrum-reasoning datasets, to support learning of reactant–product mappings, reaction procedures, enzymatic transformations, molecular understanding and spectroscopy-based reasoning.

To incorporate explicit three-dimensional molecular information, we additionally constructed a conformation-aware molecular corpus following the representation learning dataset of ConfSeq [58]. Molecules collected from ChEMBL and BindingDB were converted into three-dimensional conformers using RDKit-based conformer generation. The resulting conformations were represented in a sequence-compatible format that combines molecular connectivity with internal-coordinate information and is wrapped by `<molecule_structure></molecule_structure>` tag, linking with natural language to allow the model to learn geometric features such as molecular shape, bond geometry and conformational variation within a language-modeling framework. To avoid label leakage, we canonicalized molecular identifiers and excluded pretraining examples whose molecule–label or reaction product entries overlapped with the held-out test sets of the corresponding downstream benchmarks.

4.1.3 Materials Data

We constructed the materials component by collecting crystal compositions, crystal information files, structure-derived textual descriptions and tabular materials properties from public materials databases. The primary sources include Materials Project, JARVIS-DFT, SNUMAT, hMOF, QMOF, OQMD⁷, OMDB⁸, JARVIS-QETB⁹, GNoME¹⁰, and Cantor HEA¹¹. These sources cover a broad range of inorganic crystals, semiconductors, metal–organic frameworks, organic materials, high-entropy alloys and computationally discovered stable materials.

⁷<https://www.oqmd.org/>

⁸<https://omdb.mathub.io/>

⁹<https://jarvis.nist.gov/jarvisqetb>

¹⁰https://github.com/google-deepmind/materials_discovery/blob/main/DATASET.md

¹¹<https://doi.org/10.5281/zenodo.10854500>

For each material, we retained complementary representations such as chemical formula, elemental composition, periodic crystal structure in CIF format and natural-language structural descriptions generated from crystallographic information. We then used natural language templates to link these representations to form the training samples, enabling the model to learn composition-level, geometry-level and text-level representations of periodic systems.

For data processing, we used an 80/10/10 random split at the material-sample level, where each sample contains one material together with all associated property labels. To prevent leakage, all validation and test material samples, including their structural representations and property records, were removed from the continued-pretraining corpus before training.

4.1.4 DNA/RNA Data

We constructed the RNA component from RNACentral, retaining sequences in FASTA format and wrapping them with `<rna></rna>` tags. Available metadata, such as RNA family information, was incorporated to provide functional and contextual cues. For DNA, genomic sequences were retrieved from NCBI, sampled into 1,000-bp organism-level fragments, represented in FASTA format and wrapped with `<dna></dna>` tags. Organism names and genomic coordinates were retained to anchor each sequence within its genomic context.

4.1.5 General Text and Instruction Following Data

To complement the scientific corpora with general-domain text and instruction-following supervision, we draw from three publicly released sources. For general web text, we sample from Nemotron-CC v2 [79], a multi-trillion-token, quality-filtered Common Crawl derivative. For reasoning-formatted instruction data, we use the mathematics and instruction-following splits of Dolci-Think-SFT-32B [80], in which every response carries an explicit `<think>`-delimited deliberation trace covering multi-step symbolic reasoning and constraint-driven instruction following. For scientific instruction adherence, we additionally include the training split of SciIF [81], whose prompts impose verifiable constraints over scientific contexts.

4.2 Model Architecture

The overall architecture of SciReasoner is designed to operate as a unified causal language model capable of processing interleaved structural and textual modalities. The system consists of three primary components: modality-specific offline structural compressors, a structure-aware vocabulary embedding layer (acting as a discrete cross-modal projection), and a unified large language model (LLM) backbone, denoted as f_ϕ . The LLM backbone is initialized with the weights of Qwen3-14B [82].

4.2.1 Offline Structure Encoder

A critical limitation of applying generalist LLMs directly to scientific data is the reliance on text-based sub-word tokenizers, such as Byte-Pair Encoding [83, 84], which arbitrarily fragment physical topologies. To prevent this, we employ a modality-specific offline compression strategy. Let $S \in \{S_{prot}, S_{chem}, S_{cryst}\}$ represent the raw structure input data for proteins, small molecules, and crystals, respectively. As shown in Fig. 1B, we utilize domain-specific discrete encoders to encode S into the structural information sequence X_v . Crucially, this offline process guarantees that fundamental physical semantics (such as local chemical motifs, three-dimensional geometry, and lattice symmetries) are preserved as distinct, holistic tokens prior to interacting with the language model. Specifically, we adopt: (i) Foldseek [40] for encoding the 3D structure of protein, (ii) SLICES [85] for crystal, and (iii) ConfSeq [58] for 3D molecular structures. The encoded structural information sequences are tagged with corresponding structure tag, i.e., `<material_structure></material_structure>`, `<molecule_structure></molecule_structure>`, and `<protein_structure></protein_structure>`.

4.2.2 Structure-Aware Vocabulary Embedding

Unlike traditional multimodal frameworks that rely on heavy continuous encoders (e.g., Graph Neural Networks) during the forward pass, SciReasoner treats the discrete structural sequence X_v identically to linguistic scientific tokens. We introduce a dedicated structure-aware vocabulary, which functions as a discrete cross-modal projection layer. Let $\mathbf{W}_v \in \mathbb{R}^{|V_v| \times d_{LLM}}$ be the learnable embedding matrix for the structural vocabulary, where $|V_v|$ is the vocabulary size of all structural motifs and features, and d_{LLM} is the hidden dimension of the LLM. The structural sequence X_v is directly projected into the dense continuous space via a discrete embedding lookup:

$$\mathbf{H}_v = \text{Embedding}(X_v, \mathbf{W}_v) \in \mathbb{R}^{L_v \times d_{LLM}} \quad (1)$$

This discrete projection elegantly bypasses the computational bottleneck of continuous spatial encoders, seamlessly aligning structural inputs with the LLM’s native semantic space.

4.2.3 Unified Autoregressive Generation

Simultaneously, the language instruction X_q is tokenized and embedded by the LLM’s native embedding layer to produce the language representation $\mathbf{H}_q \in \mathbb{R}^{L_q \times d_{LLM}}$. The projected structural embeddings and language embeddings are concatenated along the sequence dimension to form the unified input prompt $\mathbf{H}_{prompt} = [\mathbf{H}_v; \mathbf{H}_q]$. The LLM backbone f_ϕ takes \mathbf{H}_{prompt} as input and generates the response sequence X_a in an autoregressive manner, ensuring the generation is strictly conditioned on both the structural evidence and the textual instruction.

4.3 Pretraining

4.3.1 Optimization Objective and Multi-Stage Pretraining

The entire multi-stage pretraining pipeline is optimized using a single, unified autoregressive causal language modeling objective (Next Token Prediction).

Given a structural input \mathbf{H}_v and a language instruction \mathbf{H}_q , the model generates a target response sequence $X_a = \{x_{a,1}, x_{a,2}, \dots, x_{a,T}\}$. The training objective minimizes the negative log-likelihood of the target tokens conditioned on the preceding context:

$$\mathcal{L}_{NTP} = - \sum_{t=1}^T \log P_\phi(x_{a,t} \mid x_{a,<t}, \mathbf{H}_v, \mathbf{H}_q) \quad (2)$$

where ϕ represents the trainable parameters of the system at any given stage. To effectively align the distinct physical structural space with the natural language space without catastrophic forgetting, we implement a three-stage curriculum learning strategy, distinguished exclusively by parameter freezing configurations, data mixture, and learning rate dynamics. Let $\Theta = \{\theta_{vocab}, \theta_{emb}, \theta_{head}, \theta_{backbone}\}$ denote the complete parameter set, corresponding to the structure-aware vocabulary layer (\mathbf{W}_v), LLM textual token embeddings, LLM prediction head, and the core LLM transformer backbone, respectively.

Stage 1: Warm-up Training

The initial stage focuses on mapping the newly introduced structural tokens into the linguistic semantic space. To prevent the untrained structure-aware vocabulary layer from corrupting the pre-trained reasoning capabilities of the LLM, we strictly freeze the core transformer backbone. In this stage, the trainable parameter subset is $\phi = \{\theta_{vocab}, \theta_{emb}, \theta_{head}\}$, while $\theta_{backbone}$ remains frozen. The model is trained on fundamental structure-to-text reconstruction tasks. By restricting gradient updates to the peripheral layers, this stage establishes a stable foundation in the embedding space before deeper integration. To ensure localized and stable optimization for these newly initialized parameters, Stage 1 maintains an independent optimizer state and employs its own standalone learning rate schedule.

Stage 2: Full-parameter Training

Once the structural tokens are stably anchored in the embedding space, we proceed to align the modalities across the deep layers of the network. In this stage, all parameters are unfrozen ($\phi = \Theta$). For the continuous optimization dynamics spanning Stage 2 and Stage 3, we initialize a new shared optimizer state and a unified Warmup-Stable-Decay (WSD) learning rate scheduler.

Formally, let t denote the training step across the combined duration of Stage 2 and Stage 3. The shared WSD scheduler is defined by the durations of the warmup (T_w), stable (T_s), and decay (T_d) phases, with total steps $T_{total} = T_w + T_s + T_d$. The learning rate $\eta(t)$ is formulated as:

$$\eta(t) = \begin{cases} \eta_{max} \cdot \frac{t}{T_w} & \text{if } 0 \leq t < T_w \\ \eta_{max} & \text{if } T_w \leq t < T_w + T_s \\ f_{decay}(t) & \text{if } T_w + T_s \leq t \leq T_{total} \end{cases} \quad (3)$$

where η_{max} is the peak learning rate and $f_{decay}(t)$ represents a monotonically decreasing annealing function (e.g., cosine decay).

Stage 2 spans the interval $0 \leq t < T_{stage2}$, where $T_w < T_{stage2} < T_w + T_s$. This indicates that Stage 2 strictly encompasses the entire warmup phase and the initial segment of the stable phase. During this period, the model is trained on diverse, broadly-sourced paired structure-text datasets, allowing the full network to assimilate domain-specific structural knowledge at a high, constant learning rate.

Stage 3: Annealing Training

In the final stage, we maintain the full-parameter optimization setting ($\phi = \Theta$) and seamlessly inherit both the optimizer state and the active WSD scheduler from Stage 2. Stage 3 covers the remainder of the training timeline ($T_{stage2} \leq t \leq T_{total}$). Thus, it begins by continuing the constant learning rate η_{max} from the stable phase and eventually triggers the decay phase ($t \geq T_w + T_s$) in its later period. In this training stage, we increase the proportion of question-answer (QA)-style data to achieve annealed training.

4.4 Post-training

The pretraining curriculum endows SciReasoner with broad cross-modal scientific knowledge, yet the resulting checkpoint M_0 behaves as a next-token continuator rather than an instruction-following reasoner. Post-training instantiates self-bootstrapped native structural reasoning through a coldstart supervised stage followed by reinforcement learning (RL). Pooling the chain-of-thought (CoT) supervision of all tasks into a single pass is, however, ill-suited here: the reasoning patterns of heterogeneous scientific tasks are divergent enough that joint training induces destructive interference, and the seed CoT traces, borrowed from external teachers, are distributionally misaligned with SciReasoner.

We therefore organize post-training around two stages: *intra-domain structural evidence grounding* followed by *cross-domain reasoning consolidation*. In the first stage, we partition the task universe by task prior and specialize M_0 into a family of per-group experts that learn how structural tokens serve as domain-specific reasoning evidence. In the second stage, these experts regenerate rational CoT traces that replace the external supervision; pooled across all tasks and augmented with non-thinking data, they drive a single unified coldstart and RL pass yielding the final model.

4.4.1 Supervised finetuning for reasoning pattern coldstart

The coldstart phase seeks to *activate* the capabilities accumulated during pretraining: we elicit instruction following and, most critically, an explicit CoT reasoning behaviour in which the model deliberates before committing to an answer. Because SciReasoner is initialized from Qwen3 [82], which natively supports a switchable thinking interface, we frame this stage as the re-establishment of a structured generation protocol that the backbone has already been exposed to, requiring only a lightweight intervention rather than a wholesale re-training.

Reasoning-formatted data construction

We curate a compact supervised corpus in which every target response adheres to an explicit deliberation template demarcated by dedicated `<think>` delimiters. We construct two complementary response modes to preserve the dual-mode behavior of the backbone. In the *thinking* mode, the model is prompted with an instruction and produces an interleaved trace of the form

$$X_a = \text{<think> } \backslash n c \backslash n \text{</think> } \backslash n \backslash n y, \quad (4)$$

where c denotes the latent chain-of-thought rationale and y the final answer conditioned on c . In the *non-thinking* mode, the instruction is appended with a `/no_think` control token and the deliberation span is collapsed to an empty placeholder,

$$X_a = \text{<think> } \backslash n \backslash n \text{</think> } \backslash n \backslash n y, \quad (5)$$

yielding a direct-answer response. Exposing the model to both formats within a single corpus re-anchors the semantics of the `<think>` delimiters and the `/no_think` switch, enabling the controllable allocation of test-time computation between explicit reasoning and direct response.

Supervised optimization objective

The coldstart stage retains the autoregressive next-token-prediction objective used throughout pretraining, but departs from it in the supervision mask. Rather than computing the loss over the entire sequence, we adopt a supervised-finetuning (SFT) formulation in which gradients are propagated solely through the response tokens, while the instruction and structural context serve purely as conditioning. Concretely, let the full sequence be partitioned into the prompt context $(\mathbf{H}_v, \mathbf{H}_q)$ and the target response $X_a = \{x_{a,1}, \dots, x_{a,T}\}$, where X_a encompasses both the deliberation span and the final answer. The coldstart objective minimizes the negative log-likelihood restricted to the response positions:

$$\mathcal{L}_{SFT} = - \sum_{t=1}^T m_t \log P_\phi(x_{a,t} \mid x_{a,<t}, \mathbf{H}_v, \mathbf{H}_q), \quad (6)$$

where the binary mask $m_t \in \{0, 1\}$ activates supervision only on the response segment ($m_t = 1$ for tokens in X_a) and suppresses the loss contribution of the prompt tokens ($m_t = 0$). This response-only masking ensures that the optimization sharpens the conditional mapping from instruction to reasoned answer, instead of expending capacity on modelling the marginal distribution of the inputs.

Coldstart

In the first stage, a coarse task prior partitions the task universe \mathcal{T} into J disjoint groups $\{\mathcal{G}_1, \dots, \mathcal{G}_J\}$, each aggregating similar tasks with CoT corpus \mathcal{C}_j whose rationale c is teacher-supplied. We coldstart M_0 independently on each \mathcal{C}_j under \mathcal{L}_{SFT} , yielding specialized checkpoints $\{M_j^{\text{cs}}\}_{j=1}^J$ so that each internalizes a coherent reasoning style rather than reconciling conflicting ones. In the second stage, after the experts $\{M_j^{\text{expert}}\}$ are obtained by the

intra-domain structural evidence grounding stage below, we replace the external supervision with SciReasoner-generated traces: each M_j^{expert} rolls out on its group’s prompts, and verification-passing traces \tilde{c} replace the external c in \mathcal{C}_j , giving a new corpus $\tilde{\mathcal{C}}_j$. Augmenting each $\tilde{\mathcal{C}}_j$ with `/no_think` direct-answer data and pooling across groups, $\tilde{\mathcal{C}} = \bigcup_{j=1}^J \tilde{\mathcal{C}}_j$, we perform a single unified coldstart from M_0 under the same \mathcal{L}_{SFT} . As the targets now lie on the model’s own manifold, this furnishes a well-matched initialization for the *cross-domain reasoning consolidation* pass.

4.4.2 Reinforcement learning

For each sub-task, we curate RL data with an empirical solve-rate filter that selects medium-difficulty instances. Let D_k^{train} denote the training set for the k -th sub-task. For each example $x \in D_k^{\text{train}}$:

1. **Stochastic rollouts.** We generate $N = 8$ independent rollouts from M_0 on x with stochastic decoding and set `do_sample=True`. To balance response diversity and format fidelity, we use an adaptive temperature T , tuned once for each sub-task before collection. Starting from $T = 0.9$, we increase T when the eight samples show insufficient diversity, such as low uniqueness among normalized responses or highly repetitive content. If the outputs become too diffuse and frequently violate the required format or fail to parse, we decrease T . For the final eight rollouts for each x , we keep T fixed and vary only the random seed, while holding all other decoding parameters constant.
2. **Empirical solve rate.** For rollout i , we compute a correctness indicator $f_i(x) \in \{0, 1\}$ using the sub-task’s canonical metric, such as exact match or a thresholded score, and aggregate

$$\hat{p}(x) = \frac{1}{8} \sum_{i=1}^8 f_i(x). \quad (7)$$

3. **Mid-difficulty filtering.** We retain examples with partial success:

$$S_k = \{ x \in D_k^{\text{train}} \mid 0.125 < \hat{p}(x) < 0.875 \}. \quad (8)$$

The retained examples are neither trivially solved nor consistently failed, and thus provide more informative RL learning signals than cases with $\hat{p}(x) \approx 1$ or $\hat{p}(x) \approx 0$.

4. **Difficulty-centred sampling.** We construct the RL subset T_k by ranking candidate examples according to their distance from intermediate success,

$$d(x) = |\hat{p}(x) - 0.5|. \quad (9)$$

We then select $K = 2000$ examples without replacement, first from S_k in ascending order of $d(x)$:

$$T_k = \begin{cases} \arg \text{top}_{K, x \in S_k}(-d(x)), & |S_k| \geq K, \\ S_k \cup \arg \text{top}_{K-|S_k|, x \in D_k^{\text{train}} \setminus S_k}(-d(x)), & |S_k| < K. \end{cases} \quad (10)$$

Here, $\arg \text{top}_{m, x \in A} f(x)$ denotes the m examples in a candidate set A with the largest values of $f(x)$. Since $f(x) = -d(x)$, this operator selects examples with the smallest distance to $\hat{p}(x) = 0.5$. Thus, the procedure starts from examples closest to intermediate success and gradually expands to examples with success rates farther from 0.5 until $K = 2000$ examples are collected.

The resulting 2,000-example set serves as the RL training pool for that sub-task.

Focusing RL on neither-trivial-nor-impossible instances yields stronger and more stable improvement: items with $\hat{p} \in (0.125, 0.875)$ (i) expose the policy to cases where exploration can flip outcomes, (ii) avoid vanishing advantages on saturated positives, and (iii) reduce uninformative gradients from uniformly failed cases.

Model training is performed with DAPO [86], which optimizes the policy with RL across heterogeneous scientific tasks. Given a prompt, we first sample G candidates from the old policy $\pi_{\theta_{\text{old}}}$. The corresponding within-group standardized advantages are computed as

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (11)$$

We then optimize the following PPO-style clipped objective:

$$\begin{aligned} \mathcal{J}_{\text{DAPO}}(\theta) = & \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right] \\ \text{s.t. } & 0 < \left| \{o_i \mid \text{is_equivalent}(a, o_i)\} \right| < G, \end{aligned} \quad (12)$$

The asymmetric *Clip-Higher* schedule is used to preserve exploration. We also apply *Dynamic Sampling* to remove degenerate groups whose rewards are identical, and use a token-level policy-gradient reduction.

The original DAPO reward is rule-based and binary, whereas many of our tasks, including free-form semantic judgment, information extraction, and numeric regression, cannot be evaluated by a single exact-match criterion. We therefore substitute the binary reward with a *reward softening* scheme, which converts grouped task metrics to a shared $[0, 1]$ scale:

$$R_i^{\text{soft}} = g(m_i(q, o_i, \mathcal{A}(q))) \quad (13)$$

Here, $g(\cdot)$ is a monotone calibration function that keeps reward magnitudes comparable across tasks, and m_i denotes the *group-specific* quality metric. Under this grouping scheme, scientific prediction tasks use distance-based rewards, retrieval and extraction tasks use matching-based rewards, and tool-verified rewards are computed with professional scientific software.

Intra-domain structural evidence grounding and cross-domain reasoning consolidation

The procedure above is first run independently per group as intra-domain structural evidence grounding: starting from M_j^{cs} , the solve-rate filtering and DAPO objective of Eqs. (11)–(13) are applied on the data of \mathcal{G}_j alone, producing the experts $\{M_j^{\text{expert}}\}_{j=1}^J$ that supply the on-policy traces for the unified coldstart. It is then run once more over the merged all-task pool as cross-domain reasoning consolidation, starting from the unified coldstart checkpoint. Because the policy is now initialized on expert-generated on-policy supervision from every group, this final pass consolidates the experts’ specialized competences without the interference of a naive joint optimization, yielding the final unified model M_{final} .

Appendix A Detailed experimental results

We report task-level evaluation results for all benchmark tasks with available model measurements. SciReasoner is compared with four frontier general-purpose models: Opus-4.7, GPT-5.5, Kimi-K2.6 and DeepSeek-V4-Pro. The result tables are organized by discipline: Chemistry, Material Science, and Biology. Within each table, rows are further grouped by task type: Scientific QA, Property Prediction, Property Classification, or Generation and Design.

Best and second-best results are highlighted in bold and underlined, respectively. Metrics marked with \uparrow are better when larger, whereas metrics marked with \downarrow are better when smaller.

A.1 Task and metric descriptions

The descriptions below follow the task organization used in the result tables and clarify both the expected model behavior and the metric used for evaluation.

Chemistry tasks.

Scientific QA.

- **Chemical entity recognition (F1)**: identifies chemical mentions in scientific or biomedical text and evaluates span/entity recovery with a precision–recall balanced score.
- **Chemical protein interaction extraction (F1)**: extracts chemical–protein relation statements from text, requiring the model to identify the paired entities and the asserted interaction.
- **Chemical disease interaction extraction (F1)**: extracts chemical–disease relation statements from scientific text.
- **Multiple choice question (ACC)**: selects the correct answer option for scientific multiple-choice questions, testing factual knowledge and reasoning.
- **True or false question (ACC)**: judges whether a scientific statement is correct, incorrect, or unsupported under the task format.
- **Open question (BertScore)**: generates free-form answers for scientific questions and compares semantic similarity with reference answers.
- **Name conversion-s2i (Split Match)**: converts a SMILES string into an IUPAC-style molecular name and checks component-level string agreement.
- **Name conversion-s2f (Element Match)**: converts a SMILES string into a molecular formula and checks whether element composition is preserved.
- **Name conversion-i2s (Exact Match)**: converts an IUPAC-style name into a SMILES string and requires exact agreement with the reference.
- **Name conversion-i2f (Element Match)**: converts an IUPAC-style name into a molecular formula and evaluates element-level formula correctness.
- **Molecular description generation (ROUGE-L)**: generates a molecule description and compares it with the reference text.
- **Molecule captioning (MENTOR)**: produces molecule-level captions that capture chemical semantics and is evaluated with a molecule-captioning metric.

Property Prediction.

- **Estimated solubility (ESOL) prediction (RMSE)**: predicts aqueous solubility from molecular representation; lower root mean squared error indicates better numeric prediction.
- **3D molecule structure (DUD-E) (5.0% EF)**: ranks compounds in a DUD-E-style virtual-screening setting and evaluates early enrichment among the top 5.0% candidates.
- **Lipophilicity (LIPO) prediction (RMSE)**: predicts lipophilicity-related continuous values from molecular structure.
- **Physicochemical prediction (MAE)**: predicts continuous physicochemical endpoints from molecular structure and reports average absolute error.

Property Classification.

- **Blood-brain barrier permeability (BBBP) prediction (ACC)**: classifies whether a molecule can pass the blood–brain barrier.
- **Clinical toxicity (ClinTox) prediction (ACC)**: classifies molecular clinical toxicity labels.
- **HIV Prediction (ACC)**: classifies whether a molecule is active against HIV replication.
- **Side Effect Resource (SIDER) Prediction (ACC)**: classifies side-effect associations for a molecule.

Generation and Design.

- **Forward synthesis (Exact Match)**: generates the expected product or outcome for a synthesis prompt and checks exact string agreement.
- **Forward reaction prediction (Exact Match)**: predicts products from specified reactants and reagents.
- **Reagent prediction (Exact Match)**: predicts the reagent, catalyst, solvent, or auxiliary component needed for a reaction.
- **Retrosynthesis mol (Exact Match)**: proposes precursor reactants for a target molecule in a Mol-Instructions-style retrosynthesis setting.
- **Retrosynthesis USPTO-50K (Exact Match)**: proposes reactants for target products in the USPTO-50K retrosynthesis setting.
- **Retrosynthesis smol (Exact Match)**: generates plausible precursor molecules for a target product in the SMol-style retrosynthesis setting.
- **Molecule generation (Exact Match)**: generates a molecular string that satisfies a given property or constraint prompt.
- **Description guided molecule design (Exact Match)**: designs a molecule from a natural-language requirement and checks exact agreement with the target answer.

Material science tasks.

Property Prediction. For database-level benchmarks that aggregate multiple heterogeneous properties, we follow domain LLM-Prop [67] and report the normalized score $\frac{MAD}{MAE}$, where larger values indicate lower error relative to target dispersion.

- **MP regression ($\frac{MAD}{MAE}$)**: predicts continuous Materials Project properties, such as band gap, density, volume, formation energy, and stability-related quantities.
- **SNUMAT regression ($\frac{MAD}{MAE}$)**: predicts SNUMAT material properties, including band-gap and spin-orbit-related targets.
- **JARVIS-DFT ($\frac{MAD}{MAE}$)**: predicts DFT-derived material properties covering structural, electronic, elastic, dielectric, and thermodynamic quantities.
- **JARVIS-QETB ($\frac{MAD}{MAE}$)**: predicts quantum-electronic tight-binding properties such as energy and band-gap-related values.
- **GNoME ($\frac{MAD}{MAE}$)**: predicts large-scale inorganic material properties, including energy, density, volume, and band-gap-related targets.
- **hMOF ($\frac{MAD}{MAE}$)**: predicts porous-material properties such as CO₂ adsorption, pore diameter, void fraction, and surface area.
- **Cantor HEA ($\frac{MAD}{MAE}$)**: predicts high-entropy alloy properties, including formation energy, energy above hull, volume per atom, and energy per atom.
- **QMOF ($\frac{MAD}{MAE}$)**: predicts quantum metal-organic framework properties such as total energy, band gap, cavity diameter, and pore-limiting diameter.
- **OQMD ($\frac{MAD}{MAE}$)**: predicts Open Quantum Materials Database properties such as band gap and formation energy.
- **OMDB ($\frac{MAD}{MAE}$)**: predicts organic-material database properties, mainly band-gap-related targets.

Property Classification.

- **MP classification (AUC)**: classifies discrete Materials Project attributes such as direct-gap status or thermodynamic stability.
- **SNUMAT classification (AUC)**: classifies SNUMAT material attributes such as direct or indirect band-gap status.

Generation and Design.

- **Composition material (SMACT)**: generates material compositions under elemental constraints and checks chemical validity.
- **Bulk modulus material (SMACT)**: generates material compositions conditioned on a target bulk modulus and evaluates chemical plausibility.

Biology tasks.

Scientific QA.

- **Function (ROUGE-L)**: generates protein-function text from biological context and compares it with the reference description.
- **General function (ROUGE-L)**: produces broader functional descriptions or annotations for biological sequences.

Property Prediction.

- **Fluorescence (Spearman)**: predicts protein mutant fluorescence and evaluates whether predicted rankings match reference rankings.
- **Stability (Spearman)**: predicts protein stability values and evaluates rank correlation with reference stability.
- **Enhancer activity (HK-PCC)**: predicts DNA enhancer activity and reports Pearson correlation for the housekeeping channel.
- **Isoform (R2)**: predicts alternative polyadenylation isoform usage from RNA sequence.
- **Mean ribosome loading (R2)**: predicts ribosome loading efficiency from RNA sequence.
- **Programmable RNA switches (R2)**: predicts ON, OFF, and ON/OFF behavior of programmable RNA switches.
- **CRISPR on target (Spearman)**: predicts on-target knockout efficacy for CRISPR guide RNAs.
- **siRNA efficiency (Mixed-score)**: predicts siRNA gene-silencing efficiency using both continuous accuracy and range-level agreement.
- **Structural similarity (MAE)**: predicts a numeric structural-similarity target and evaluates average absolute deviation.
- **TM-score (Spearman)**: predicts or ranks structure-similarity scores and evaluates ordering consistency.

Property Classification.

- **Solubility (ACC)**: classifies whether a protein sequence is soluble.
- **gSymbol2Tissue (F1)**: maps a gene symbol to tissue-expression labels.
- **gName2Cancer (F1)**: maps a gene name to associated cancer types.
- **gSymbol2Cancer (F1)**: maps a gene symbol to associated cancer types.
- **Antibody antigen (MCC)**: predicts whether an antibody and antigen sequence pair interact.
- **RNA protein interaction (MCC)**: predicts whether an RNA sequence and a protein sequence interact.
- **Epigenetic marks prediction (MCC)**: predicts epigenetic mark presence from DNA sequence.
- **TF-m (MCC)**: predicts mouse transcription-factor binding from DNA sequence.
- **Enhancer-promoter interaction (MCC)**: classifies whether enhancer and promoter regions interact.
- **PD-prom 300 all (MCC)**: detects promoter regions in a 300-bp setting over all examples.
- **PD-prom 300 notata (MCC)**: detects 300-bp promoters without TATA motifs.
- **PD-prom 300 tata (MCC)**: detects 300-bp promoters with TATA motifs.
- **CPD-prom core all (MCC)**: detects core promoter regions over all examples.
- **CPD-prom core notata (MCC)**: detects core promoters without TATA motifs.
- **CPD-prom core tata (MCC)**: detects core promoters with TATA motifs.
- **TF-h (MCC)**: predicts human transcription-factor binding from DNA sequence.
- **Yeast PPI (ACC)**: predicts yeast protein–protein interactions.
- **Human PPI (ACC)**: predicts human protein–protein interactions.
- **Protein function (ROUGE-L)**: predicts protein function annotations or descriptions from sequence.
- **Domain motif (ROUGE-L)**: predicts domain or motif descriptions for protein sequences.
- **Non-coding RNA family (ACC)**: classifies non-coding RNA sequences into functional families.
- **Modification (ACC)**: predicts RNA modification labels.
- **Fold type (ACC)**: classifies protein structural fold type.
- **Subcellular localization (ACC)**: predicts the cellular localization label of a protein.
- **EC number (Fmax)**: predicts enzyme commission annotations for protein sequences.
- **Keywords (F1)**: predicts UniProt-style functional keywords.
- **Metal ion binding (ACC)**: predicts whether a protein binds metal ions.
- **GO-BP (Fmax)**: predicts Gene Ontology biological-process terms.
- **GO-CC (Fmax)**: predicts Gene Ontology cellular-component terms.
- **GO-MF (Fmax)**: predicts Gene Ontology molecular-function terms.

Generation and Design.

- **Function-guided protein design (Normalized SW)**: generates a protein sequence from a functional prompt and evaluates sequence similarity to reference proteins.
- **Catalytic activity (ROUGE-L)**: describes the enzyme-catalyzed reaction implied by a protein sequence and compares it with the reference text.

Metric definitions.

- **ACC (\uparrow)**: fraction of samples whose predicted label exactly matches the reference label.

- **AUC** (\uparrow): area under the ROC curve; higher values indicate stronger ranking of positive examples above negatives.
- **F1** (\uparrow): harmonic mean of precision and recall, used when both false positives and false negatives matter.
- **Fmax** (\uparrow): maximum F1 over candidate thresholds, commonly used for multi-label functional annotation.
- **MCC** (\uparrow): Matthews correlation coefficient for binary classification; it remains informative when classes are imbalanced.
- **RMSE** (\downarrow): root mean squared error for regression, with larger errors penalized more strongly.
- **MAE** (\downarrow): mean absolute error between predicted and reference numeric values.
- $\frac{MAD}{MAE}$ (\uparrow): ratio between target dispersion and model error; larger values indicate better prediction relative to a mean baseline.
- **Spearman** (\uparrow): rank correlation between predicted and reference values.
- **HK-PCC** (\uparrow): Pearson correlation coefficient for the housekeeping enhancer-activity output.
- **R2** (\uparrow): coefficient of determination, measuring explained variance in regression targets.
- **Mixed-score** (\uparrow): composite siRNA score combining numeric error and activity-range agreement.
- **BertScore** (\uparrow): semantic similarity between generated and reference text using contextual embeddings.
- **ROUGE-L** (\uparrow): longest-common-subsequence overlap between generated text and reference text.
- **MENTOR** (\uparrow): molecule-captioning metric for comparing generated molecular descriptions with references.
- **Split Match** (\uparrow): component-level match for molecular name conversion outputs.
- **Element Match** (\uparrow): element-composition match for molecular formula generation.
- **Exact Match** (\uparrow): strict string or structured-answer equality with the reference output.
- **5.0% EF** (\uparrow): enrichment factor in the top 5.0% of a virtual-screening ranking.
- **SMACT** (\uparrow): validity rate under charge-balance and chemical-plausibility checks for generated material compositions.
- **Normalized SW** (\uparrow): maximum normalized Smith–Waterman alignment score between generated and reference protein sequences.

A.2 Detailed results

Table A1 summarizes the comparison with specialist baselines. Tables A2–A4 provide the complete task-level comparison with frontier general-purpose models across Chemistry, Material Science, and Biology. Within each discipline, tasks are organized as Scientific QA, Property Prediction, Property Classification, or Generation and Design.

Across the full benchmark suite, SciReasoner leads on **67** of 86 tasks. The appendix separates these comparisons for clarity. On the 33 tasks with specialist baselines, SciReasoner matches or surpasses the specialist in 26 comparisons A1. Against LLM baselines across all 86 tasks, SciReasoner is the best-performing model on 75 of 86 tasks (22 of 28 Chemistry tasks, 13 of 14 Material Science tasks, and 40 of 44 Biology tasks). The few non-leading results are concentrated in molecule captioning, selected chemistry QA and classification tasks, bulk modulus material, and a small number of protein-interaction or modification benchmarks, where competing general-purpose models remain strong.

At the discipline level, the Chemistry results indicate broad gains in chemical information extraction, name conversion, molecular description generation, reaction prediction, retrosynthesis, and molecule design (Table A2). In Material Science, SciReasoner achieves consistently strong regression and classification performance, with pronounced margins on JARVIS-QETB, GNoME, QMOF and OQMD (Table A3). In Biology, the model performs robustly across sequence-to-function generation, biological property prediction, promoter and interaction classification and RNA/protein tasks (Table A4). Together, these results indicate that the performance gains are not restricted to a single metric or task format, but extend across structured prediction, text generation, classification, and design-oriented settings.

| Task | Metric | Specialist method | Specialist | SciReasoner |
|--------------------------------------|---------------------|-------------------|--------------|--------------|
| Generation & Design | | | | |
| Retrosynthesis USPTO-50K Exact Match | Match \uparrow | RSGPT [31] | <u>0.63</u> | 0.72 |
| Prediction | | | | |
| Fluorescence | Spearman \uparrow | SaprotHub [28] | <u>0.70</u> | 0.77 |
| Isoform | R2 \uparrow | APARENT [87] | <u>0.59</u> | 0.86 |
| TM-score | Spearman \uparrow | SaprotHub [28] | 0.83 | 0.83 |
| ESOL | RMSE \downarrow | MolCLR [88] | <u>1.11</u> | 1.03 |
| GNoME | MAD \uparrow | LLM-Prop [67] | <u>15.60</u> | 21.91 |
| QMOF | MAE \uparrow | LLM-Prop [67] | <u>1.96</u> | 8.61 |
| MP regression | MAD \uparrow | LLM-Prop [67] | <u>4.39</u> | 5.83 |
| JARVIS-DFT | MAE \uparrow | LLM-Prop [67] | <u>2.91</u> | 5.67 |
| SNUMAT regression | MAD \uparrow | LLM-Prop [67] | <u>1.51</u> | 2.26 |
| hMOF | MAE \uparrow | LLM-Prop [67] | <u>1.48</u> | 1.67 |
| OQMD | MAD \uparrow | LLM-Prop [67] | <u>6.02</u> | 7.22 |
| OMDB | MAE \uparrow | LLM-Prop [67] | 1.51 | <u>1.50</u> |
| DUD-E | 5.0% EF \uparrow | ConfSeq [58] | <u>7.12</u> | 7.70 |
| Cantor HEA | MAD \uparrow | LLM-Prop [67] | 8.40 | <u>7.79</u> |
| LIPO | RMSE \downarrow | MolCLR [88] | 0.65 | <u>0.80</u> |
| Classification | | | | |
| BBBP | ACC \uparrow | MolCLR [88] | <u>0.74</u> | 0.84 |
| ClinTox | ACC \uparrow | MolCLR [88] | <u>0.93</u> | 0.95 |
| HIV Prediction | ACC \uparrow | MolCLR [88] | <u>0.81</u> | 0.92 |
| SIDER | ACC \uparrow | MolCLR [88] | <u>0.68</u> | 0.74 |
| MP classification | AUC \uparrow | LLM-Prop [67] | <u>0.72</u> | 0.73 |
| TF-m | MCC \uparrow | NT [89] | <u>0.57</u> | 0.64 |
| PD-prom 300 all | MCC \uparrow | NT [89] | 0.91 | <u>0.89</u> |
| CPD-prom core all | MCC \uparrow | NT [89] | <u>0.67</u> | 0.68 |
| RNA protein interaction | MCC \uparrow | RPI-Pred [90] | <u>0.74</u> | 0.81 |
| Non-coding RNA family | ACC \uparrow | RNA-MSM [91] | <u>0.89</u> | 0.90 |
| GO-BP | Fmax \uparrow | SaprotHub [28] | <u>0.49</u> | 0.52 |
| GO-CC | Fmax \uparrow | SaprotHub [28] | <u>0.48</u> | 0.58 |
| GO-MF | Fmax \uparrow | SaprotHub [28] | 0.67 | <u>0.66</u> |
| Human PPI | ACC \uparrow | ESM2 [29] | 0.77 | <u>0.73</u> |
| Subcellular localization | ACC \uparrow | ESM2 [29] | <u>0.84</u> | 0.88 |
| Metal ion binding | ACC \uparrow | ESM2 [29] | <u>0.73</u> | 0.74 |
| Solubility | ACC \uparrow | DeepLoc [92] | 0.77 | <u>0.72</u> |

Table A1: Per-task comparison of SciReasoner against specialist baselines. **Bold** indicates the best performance, and underline indicates the second best.

| Task | Metric | Opus-4.7 | GPT-5.5 | Kimi-K2.6 | DeepSeek-V4-Pro | SciReasoner |
|---|----------------|--------------|-------------|-------------|-----------------|-------------|
| Scientific QA | | | | | | |
| Chemical entity recognition | F1↑ | <u>0.84</u> | 0.69 | 0.72 | 0.63 | 0.88 |
| Chemical protein interaction extraction | F1↑ | <u>0.35</u> | 0.05 | 0.18 | 0.03 | 0.36 |
| Chemical disease interaction extraction | F1↑ | <u>0.41</u> | 0.31 | 0.34 | 0.27 | 0.54 |
| Multiple choice question | ACC↑ | <u>0.90</u> | 0.91 | 0.87 | 0.89 | 0.88 |
| True or false question | ACC↑ | 0.63 | <u>0.60</u> | 0.58 | 0.63 | 0.55 |
| Open question | BertScore↑ | 0.77 | <u>0.82</u> | <u>0.82</u> | <u>0.82</u> | 0.85 |
| Name conversion-s2i | Split Match↑ | 0.34 | 0.07 | 0.02 | 0.01 | 0.50 |
| Name conversion-s2f | Element Match↑ | <u>0.85</u> | 0.84 | 0.04 | 0.28 | 0.96 |
| Name conversion-i2s | Exact Match↑ | <u>0.69</u> | 0.58 | 0.19 | 0.35 | 0.87 |
| Name conversion-i2f | Element Match↑ | 0.90 | <u>0.91</u> | 0.40 | 0.64 | 0.93 |
| Molecular description generation | ROUGE-L↑ | 0.46 | 0.08 | 0.09 | 0.31 | 0.75 |
| Molecule captioning | MENTOR↑ | 0.46 | 0.30 | 0.14 | 0.16 | <u>0.45</u> |
| Property Prediction | | | | | | |
| ESOL | RMSE↓ | 0.89 | 8.72 | 137.05 | 5.24 | <u>1.03</u> |
| DUD-E | 5.0% EF↑ | 3.77 | 4.98 | 1.85 | 1.28 | 7.70 |
| LIPO | RMSE↓ | <u>0.97</u> | 1.48 | 16.25 | 3.29 | 0.80 |
| Physicochemical prediction | MAE↓ | <u>19.52</u> | 25.79 | 2158.02 | 73.24 | 4.58 |
| Property Classification | | | | | | |
| BBBP | ACC↑ | 0.73 | <u>0.81</u> | 0.54 | 0.69 | 0.84 |
| ClinTox | ACC↑ | <u>0.88</u> | 0.41 | 0.57 | 0.77 | 0.95 |
| HIV Prediction | ACC↑ | 0.94 | 0.61 | 0.91 | 0.90 | <u>0.92</u> |
| SIDER | ACC↑ | 0.70 | 0.84 | 0.69 | 0.71 | <u>0.74</u> |
| Generation and Design | | | | | | |
| Forward synthesis | Exact Match↑ | <u>0.56</u> | <u>0.56</u> | 0.35 | 0.25 | 0.73 |
| Forward reaction prediction | Exact Match↑ | 0.84 | <u>0.82</u> | <u>0.90</u> | 0.80 | 0.98 |
| Reagent prediction | Exact Match↑ | <u>0.07</u> | 0.05 | <u>0.04</u> | 0.04 | 0.12 |
| Retrosynthesis mol | Exact Match↑ | <u>0.43</u> | 0.29 | <u>0.43</u> | 0.26 | 0.67 |
| Retrosynthesis USPTO-50K | Exact Match↑ | <u>0.48</u> | 0.31 | 0.10 | 0.15 | 0.72 |
| Retrosynthesis smol | Exact Match↑ | 0.23 | 0.13 | 0.03 | 0.03 | 0.39 |
| Molecule generation | Exact Match↑ | <u>0.44</u> | 0.43 | 0.15 | 0.30 | 0.50 |
| Description guided molecule design | Exact Match↑ | <u>0.10</u> | <u>0.10</u> | 0.08 | 0.08 | 0.12 |

Table A2: Evaluation results on Chemistry tasks grouped by task type. **Bold** indicates the best performance, and underline indicates the second best.

| Task | Metric | Opus-4.7 | GPT-5.5 | Kimi-K2.6 | DeepSeek-V4-Pro | SciReasoner |
|--------------------------------|---------------------|-------------|-------------|-------------|-----------------|---------------|
| Property Prediction | | | | | | |
| MP regression | $\frac{MAD}{MAE}$ ↑ | 2.11 | <u>2.67</u> | 1.76 | 1.58 | 5.83 |
| SNUMAT regression | $\frac{MAD}{MAE}$ ↑ | 1.51 | <u>1.67</u> | 1.41 | 1.46 | 2.26 |
| JARVIS-DFT | $\frac{MAD}{MAE}$ ↑ | 1.48 | <u>1.78</u> | 1.31 | 1.38 | 5.67 |
| JARVIS-QETB | $\frac{MAD}{MAE}$ ↑ | 0.73 | 0.84 | <u>0.88</u> | 0.86 | 108.98 |
| GNoME | $\frac{MAD}{MAE}$ ↑ | 1.94 | <u>5.39</u> | 1.56 | 1.22 | 21.91 |
| hMOF | $\frac{MAD}{MAE}$ ↑ | <u>1.08</u> | 0.99 | 1.00 | 0.92 | 1.67 |
| Cantor HEA | $\frac{MAD}{MAE}$ ↑ | 1.40 | <u>2.72</u> | 1.44 | 1.29 | 7.79 |
| QMOF | $\frac{MAD}{MAE}$ ↑ | 1.66 | <u>3.12</u> | 1.56 | 1.35 | 8.61 |
| OQMD | $\frac{MAD}{MAE}$ ↑ | <u>1.97</u> | 1.84 | 1.03 | 1.46 | 7.22 |
| OMDB | $\frac{MAD}{MAE}$ ↑ | 1.26 | <u>1.43</u> | 1.07 | 1.10 | 1.50 |
| Property Classification | | | | | | |
| MP classification | AUC↑ | 0.61 | 0.65 | 0.66 | 0.59 | 0.73 |
| SNUMAT classification | AUC↑ | 0.58 | 0.59 | 0.56 | <u>0.60</u> | 0.68 |
| Generation and Design | | | | | | |
| Composition material | SMACT↑ | <u>0.89</u> | 0.90 | 0.35 | 0.24 | 0.90 |
| Bulk modulus material | SMACT↑ | 0.99 | 0.99 | 0.17 | 0.11 | <u>0.88</u> |

Table A3: Evaluation results on Material Science tasks grouped by task type. **Bold** indicates the best performance, and underline indicates the second best.

| Task | Metric | Opus-4.7 | GPT-5.5 | Kimi-K2.6 | DeepSeek-V4-Pro | SciReasoner |
|--------------------------------|--------------------------|-------------|-------------|-------------|-----------------|-------------|
| Scientific QA | | | | | | |
| Function | ROUGE-L \uparrow | <u>0.30</u> | 0.10 | 0.01 | 0.03 | 0.80 |
| General function | ROUGE-L \uparrow | <u>0.52</u> | 0.10 | 0.18 | 0.15 | 0.77 |
| Property Prediction | | | | | | |
| Fluorescence | Spearman \uparrow | 0.44 | <u>0.50</u> | 0.10 | 0.04 | 0.77 |
| Stability | Spearman \uparrow | <u>0.36</u> | 0.13 | 0.06 | -0.02 | 0.61 |
| Enhancer activity | HK-PCC \uparrow | 0.07 | -0.06 | -0.05 | <u>0.13</u> | 0.64 |
| Isoform | R2 \uparrow | <u>0.09</u> | 0.01 | 0.05 | 0.01 | 0.86 |
| Mean ribosome loading | R2 \uparrow | <u>0.03</u> | <u>0.10</u> | 0.01 | 0.00 | 0.60 |
| Programmable RNA switches | R2 \uparrow | <u>0.04</u> | <u>0.02</u> | 0.02 | 0.01 | 0.46 |
| CRISPR on target | Spearman \uparrow | 0.14 | <u>0.28</u> | 0.05 | 0.40 | 0.12 |
| siRNA efficiency | Mixed-score \uparrow | 0.00 | <u>0.33</u> | 0.07 | 0.22 | 0.61 |
| Structural similarity | MAE \downarrow | <u>0.18</u> | 0.27 | 0.47 | 0.56 | 0.05 |
| TM-score | Spearman \uparrow | -0.06 | <u>0.15</u> | -0.15 | -0.12 | 0.83 |
| Property Classification | | | | | | |
| Solubility | ACC \uparrow | <u>0.58</u> | 0.54 | 0.53 | <u>0.58</u> | 0.72 |
| gSymbol2Tissue | F1 \uparrow | 0.43 | <u>0.44</u> | 0.23 | <u>0.34</u> | 0.53 |
| gName2Cancer | F1 \uparrow | 0.06 | <u>0.12</u> | 0.00 | 0.01 | 0.73 |
| gSymbol2Cancer | F1 \uparrow | <u>0.16</u> | 0.12 | 0.00 | 0.00 | 0.71 |
| Antibody antigen | MCC \uparrow | 0.08 | <u>0.12</u> | 0.02 | 0.06 | 0.28 |
| RNA protein interaction | MCC \uparrow | -0.44 | -0.13 | <u>0.05</u> | -0.38 | 0.81 |
| Epigenetic marks prediction | MCC \uparrow | -0.02 | <u>0.12</u> | -0.04 | -0.13 | 0.18 |
| TF-m | MCC \uparrow | <u>0.28</u> | 0.22 | 0.06 | 0.04 | 0.64 |
| Enhancer-promoter interaction | MCC \uparrow | <u>0.07</u> | -0.03 | 0.00 | 0.00 | 0.24 |
| PD-prom 300 all | MCC \uparrow | <u>0.30</u> | -0.10 | 0.23 | -0.07 | 0.89 |
| PD-prom 300 notata | MCC \uparrow | <u>0.23</u> | -0.08 | -0.04 | 0.13 | 0.93 |
| PD-prom 300 tata | MCC \uparrow | <u>0.30</u> | 0.00 | 0.05 | 0.22 | 0.55 |
| CPD-prom core all | MCC \uparrow | <u>0.35</u> | 0.05 | 0.02 | -0.13 | 0.68 |
| CPD-prom core notata | MCC \uparrow | <u>0.34</u> | 0.21 | 0.05 | -0.14 | 0.68 |
| CPD-prom core tata | MCC \uparrow | <u>0.40</u> | 0.09 | 0.06 | 0.02 | 0.67 |
| TF-h | MCC \uparrow | 0.21 | <u>0.22</u> | 0.14 | -0.02 | 0.52 |
| Yeast PPI | ACC \uparrow | 0.42 | 0.56 | 0.48 | 0.48 | <u>0.54</u> |
| Human PPI | ACC \uparrow | 0.70 | 0.82 | 0.50 | 0.55 | <u>0.73</u> |
| Protein function | ROUGE-L \uparrow | <u>0.43</u> | 0.10 | 0.14 | 0.14 | 0.51 |
| Domain motif | ROUGE-L \uparrow | <u>0.42</u> | 0.06 | 0.24 | 0.15 | 0.55 |
| Non-coding RNA family | ACC \uparrow | 0.21 | <u>0.24</u> | 0.16 | 0.12 | 0.90 |
| Modification | ACC \uparrow | 0.52 | 0.55 | 0.52 | <u>0.53</u> | 0.51 |
| Fold type | ACC \uparrow | 0.00 | <u>0.00</u> | 0.00 | 0.00 | 0.49 |
| Subcellular localization | ACC \uparrow | 0.48 | <u>0.65</u> | 0.11 | 0.13 | 0.88 |
| EC number | Fmax \uparrow | 0.09 | 0.12 | <u>0.15</u> | 0.13 | 0.78 |
| Keywords | F1 \uparrow | <u>0.58</u> | 0.36 | <u>0.03</u> | 0.08 | 0.83 |
| Metal ion binding | ACC \uparrow | 0.74 | <u>0.70</u> | 0.43 | 0.60 | 0.74 |
| GO-BP | Fmax \uparrow | <u>0.43</u> | 0.29 | 0.27 | 0.31 | 0.52 |
| GO-CC | Fmax \uparrow | <u>0.33</u> | <u>0.40</u> | 0.28 | 0.35 | 0.58 |
| GO-MF | Fmax \uparrow | <u>0.47</u> | 0.24 | 0.16 | 0.40 | 0.66 |
| Generation and Design | | | | | | |
| Function-guided protein design | Normalized SW \uparrow | <u>0.74</u> | 0.67 | 0.73 | 0.73 | 0.94 |
| Catalytic activity | ROUGE-L \uparrow | <u>0.60</u> | 0.14 | 0.17 | 0.19 | 0.70 |

Table A4: Evaluation results on Biology tasks grouped by task type. **Bold** indicates the best performance, and underline indicates the second best.

Appendix B Human Evaluation Form

This appendix shows representative questionnaire items used for double-blinded human evaluation. One sample is selected from each task category: crystal-material property prediction, Gene Ontology prediction, and retrosynthesis. For each item, evaluators read the input prompt, two anonymized model reasoning traces (named Model A and Model B correspondingly) and final outputs, and a read-only ground-truth fact sheet. They then score each model on five trace-quality axes and complete the overall comparison questions.

B.1 General Evaluation Instructions

Evaluators should assess the quality of the reasoning trace, not only whether the final answer is numerically or symbolically close to the ground truth. The main criteria are evidence grounding, domain plausibility, alignment with the ground-truth regime or reaction/function region, reasoning coherence, and hallucination risk.

| Verdict label | Meaning |
|---------------|--|
| Correct | The claim is supported by the provided input, the ground truth, the read-only Part A fact sheet, or standard domain knowledge used only for verification. |
| Minor | A local or peripheral defect. The relevant entity or direction is mostly correct, and the defect does not materially change the quality of that scoring axis. |
| Major | A substantive axis-level defect, such as an input misread treated as fact, a domain inference contradicting the ground truth or standard knowledge, a wrong target region or reaction family, or a broken evidence-to-conclusion chain. Independent major defects on the same axis are counted separately. |
| Critical | Fabrication: the trace names a concrete entity, index, reference, topology, structure, GO term, reaction, reagent, or mechanism that is absent from the input/ground truth and cannot be verified from standard sources. Real-but-misapplied entities are not Critical; they are routed to Q2 or Q3. |

| Axis | Focus | Scoring rule |
|------|--|--|
| Q1 | Evidence grounding in the provided input | Check whether entities, tokens, sequence positions, atom-map indices, product groups, or cited input facts actually appear in the task input. Score in $\{1, \dots, 10\}$ or N.A. using the count-to-score table. |
| Q2 | Domain plausibility relative to ground truth | Check whether the trace’s scientific inferences are plausible given the ground truth, and standard domain knowledge. Score in $\{1, \dots, 10\}$ or N.A. using the count-to-score table. |
| Q3 | Target alignment | Judge whether the committed conclusion falls in the correct materials regime, GO region, or retrosynthesis reaction class/formed bond. Score in $\{1, \dots, 10\}$ or N.A. using the count-to-score table. |
| Q4 | Reasoning coherence | Check whether the trace builds a relevant evidence-to-conclusion chain without unresolved contradiction, circularity, padding, or unused observations. Score in $\{1, \dots, 10\}$ or N.A. using the count-to-score table. |
| Q5 | Unsupported overclaiming or hallucination | Check for fabricated or over-specific unsupported claims. Score in $\{1, \dots, 10\}$ or N.A. using the count-to-score table. |

| Verdict counts on Q1/Q2/Q3/Q4/Q5 | Score |
|--|-------|
| <i>Critical</i> ≥ 2 , or the core conclusion depends on a fabricated entity | 1 |
| <i>Critical</i> = 1, <i>Major</i> ≥ 1 | 2 |
| <i>Critical</i> = 1, <i>Major</i> = 0 | 3 |
| <i>Critical</i> = 0, <i>Major</i> ≥ 4 | 2 |
| <i>Critical</i> = 0, <i>Major</i> = 3 | 3 |
| <i>Critical</i> = 0, <i>Major</i> = 2 | 4 |
| <i>Critical</i> = 0, <i>Major</i> = 1 | 5 |
| <i>Critical</i> = 0, <i>Major</i> = 0, <i>Minor</i> ≥ 4 | 6 |
| <i>Critical</i> = 0, <i>Major</i> = 0, <i>Minor</i> = 3 | 7 |

| | |
|---|------|
| <i>Critical</i> = 0, <i>Major</i> = 0, <i>Minor</i> = 2 | 8 |
| <i>Critical</i> = 0, <i>Major</i> = 0, <i>Minor</i> = 1 | 9 |
| <i>Critical</i> = 0, <i>Major</i> = 0, <i>Minor</i> = 0, and at least one claim is verified | 10 |
| Correct | |
| No checkable claim on the axis | N.A. |

| Axis | Materials | Gene Ontology | Retrosynthesis |
|------|---|--|--|
| Q1 | Evidence grounding in formula, SLICES-PLUS space-group tokens, atom lists, edge lists, periodic offsets, and coordination counts. | Evidence grounding in amino-acid sequence length, residue positions, motifs, and the Foldseek 3Di structural alphabet. | Evidence grounding in product SMILES, product functional groups, atom-map indices, product connectivity, and cited atom-map sets. |
| Q2 | Materials plausibility relative to crystal chemistry, property constraints, and the ground-truth property regime. | Biological plausibility relative to the ground-truth protein identity, family, motifs, domains, and GO region. | Disconnection and reactant plausibility relative to the gold route, atom-map balance, oxidation/protection state, and chemical feasibility. |
| Q3 | Property-regime alignment, including boundary handling for values near regime edges. | GO-space alignment within the evaluated branch, including correct, adjacent, wrong-region, wrong-super-class, and no-commit cases. | Reaction-class alignment: gold reaction family plus gold formed bond scores highest; neighbouring families sharing the formed bond are borderline; different valid disconnections are lower. |
| Q4 | Coherence from decoded structure and target-specific mechanism to the committed property call. | Coherence from sequence, 3Di, motif, domain, or family evidence to the committed GO-function hypothesis. | Coherence from product parsing to retrosynthetic disconnection and reactant proposal. |
| Q5 | Unsupported materials overclaiming, such as invented topology, SBU, phase transition, atom index, or citation. | Unsupported biological overclaiming, such as invented residues, motifs, domains, GO terms, protein identity, or citation. | Unsupported chemical overclaiming, such as invented product groups, atom-map indices, reagents, named reactions, mechanisms, or citations. |

The questionnaire also contains four overall fields. Q6 asks how Model A compares with expert expectation, Q7 asks the same for Model B, Q8 asks for a direct Model A versus Model B comparison, and Q9 records evaluator confidence on a 1–10 scale with a free-text note.

B.2 Blank Scoring Sheet Used for Each Sample

| Axis | Model A score | Model B score | Evidence, claim verdicts, and notes |
|------|---------------|---------------|-------------------------------------|
| Q1 | _____ | _____ | |
| Q2 | _____ | _____ | |
| Q3 | _____ | _____ | |
| Q4 | _____ | _____ | |
| Q5 | _____ | _____ | |
| Sum | _____ | _____ | |
| Mean | _____ | _____ | |

| Overall question | Allowed response |
|---------------------------------------|---|
| Q6. Model A versus expert expectation | Significantly falls short / Falls short / Comparable / Exceeds / Significantly exceeds. |

| | |
|---------------------------------------|---|
| Q7. Model B versus expert expectation | Significantly falls short / Falls short / Comparable / Exceeds / Significantly exceeds. |
| Q8. Direct comparison | A much better / A slightly better / Tie / B slightly better / B much better. |
| Q9. Evaluator confidence | Integer confidence score from 1 to 10, plus a free-text confidence note. |

B.3 Materials: Ag₂HgI₄, shear modulus

| Field | Value |
|----------------------|---|
| Dataset / task | JARVIS-DFT / G_v . |
| Sample ID | 1169. |
| Chemical formula | Ag ₂ HgI ₄ . |
| Target property | shear_modulus_gv. |
| Property description | Shear modulus: resistance to shear deformation; related to directional bonding, framework rigidity, and elastic anisotropy. |
| Ground truth | 5.77 GPa. |
| Model A final output | {shear_modulus_gv: 5.62}. |
| Model B final output | {shear_modulus_gv: 8.00}. |

Input prompt.

You are a material scientist. Look at the chemical formula and structure information of the given crystalline material and predict its property. The output must be in a JSON format. For example: {property_name: predicted_property_value}. Answer as precise as possible and in as few words as possible.
 chemical formula: Ag₂HgI₄
 structure information: <material_structure> o w b 000 m 000 c 000 o Ag Ag Hg I I I I 0 6 -oo 0 3 -oo 0 5 ooo 0 4 ooo 1 4 o-o 1 5 ooo 1 3 o-o 1 6 ooo 2 3 --o 2 6 -o- 2 4 o-- 2 5 ooo</material_structure>
 property name: shear_modulus_gv

Read-only ground-truth fact sheet.

| ID | Reference fact | Value |
|----|-------------------------|--|
| A1 | Decoded space group | No primer match. The leading tokens o w b 000 m 000 c 000 o do not exactly match any listed SLICES-PLUS primer row, so a concrete space-group label is not a checkable Q1 claim. |
| A2 | Number of atoms | 7; any cited atom index ≥ 7 is fabricated. |
| A3 | Formula consistency | The atom list Ag Ag Hg I I I I corresponds to 2 Ag, 1 Hg, and 4 I, matching Ag ₂ HgI ₄ . |
| A4 | Target and ground truth | shear_modulus_gv = 5.77 GPa. |
| A5 | Ground-truth regime | Soft shear modulus, $G_v < 20$ GPa; far from the 20 GPa boundary. The hard physical constraint is $G_v \lesssim K_v$. |
| A6 | Atom-index table | 0=Ag, 1=Ag, 2=Hg, 3=I, 4=I, 5=I, 6=I. |
| A7 | Key neighbors | Ag(0) connects to I(6,3,5,4); Ag(1) connects to I(4,5,3,6); Hg(2) connects to I(3,6,4,5). Each iodine connects to Ag(0), Ag(1), and Hg(2). |
| A8 | Coordination summary | Ag(0), Ag(1), and Hg(2) are each four-coordinate, tetrahedral by iodine; each iodine is three-coordinate. |

Example claim prompts shown to the evaluator.

| Axis | Model A claim examples | Model B claim examples |
|------|---|---|
| Q1 | The trace cites the atom list, selected edge tokens, and tetrahedral metal-iodine coordination. | The trace parses indices 0 and 1 as Ag, 2 as Hg, and 3–6 as I, and lists the twelve metal-iodine edges. |
| Q2 | Heavy and polarizable iodide ions imply a compliant lattice with low shear stiffness. | The trace treats the material as a soft iodide solid and invokes tetrahedral coordination. |

| | | |
|----|--|--|
| Q3 | The committed value 5.62 GPa lies in the correct soft regime and is close to the ground truth. | The committed value 8.00 GPa also lies in the correct soft regime. |
| Q4 | The trace connects structure, coordination, heavy halide chemistry, and the target mechanical property. | The trace contains longer speculative discussion; evaluators check whether the final estimate follows from the cited evidence. |
| Q5 | No concrete fabricated entity is apparent if the trace remains within decoded structure and standard halide chemistry. | Evaluators check any named phase, database identifier, or literature-like statement for fabrication. |

B.4 Gene Ontology: 1bd8_A-P55273, biological process

| Field | Value |
|------------------------------|--|
| Dataset / task | GO / BP. |
| Protein name in sample | 1bd8_A-P55273. |
| Chain | A. |
| Model A performance metadata | F1 = 0.967, precision = 0.993, recall = 0.943; 139 predicted BP terms. |
| Model B performance metadata | F1 = 0.209, precision = 0.314, recall = 0.156; 70 predicted BP terms. |
| Number of true BP terms | 145. |

Input prompt.

```

What are the Gene Ontology terms (biological process) for this protein?
<protein>RAGDRLSGAAARGDVQEVRRLLHRELVHPDALNRFKGTALQVMMFGSTAI ALELLKQGASPNVQDTSQTSVPVHDAARTGFLDTLKV LVEHGADV
NVPDGTGALPIHLAVQEGHTAVVSFLAAESDLHRRDARGLTPLELALQRGAQDLVDILQGHM</protein>
<protein_structure>D L Q Q Q L L V C Q L V L N L P S N C C S C P V V V D Q Q Q D A D P V R G G S L A R
G P L L N L S S L L S S L V S P N D L C D A D P Q Q D G S L L V N L L V L V V S N N V S S V V S P H D
Q A R A G N Q Q D G S L L N N L Q N V N L V S N L V C L Q P Y P Q C G A G N V R D G S L R S N V V V V
V V S S N V S N V V S D</protein_structure>
Only the biological process (BP) aspect.

```

Model outputs shown in the questionnaire.

| Output | Abbreviated display |
|--------------------------|---|
| Model A final prediction | A long semicolon-separated list of 139 BP terms, including regulation of signal transduction, regulation of cell cycle, negative regulation of endopeptidase activity, proteolysis, regulation of apoptotic process, DNA damage response, regulation of kinase activity, and negative regulation of cell cycle. |
| Model B final prediction | A JSON-style output with 70 BP terms, dominated by regulation of DNA-templated transcription, gene expression, positive regulation of RNA biosynthetic process, positive regulation of DNA-templated transcription, and related transcriptional or signaling terms. |
| Ground truth | A long semicolon-separated list of 145 BP terms centered on negative regulation of cell cycle, G1/S transition, CDK activity regulation, apoptotic or programmed cell-death regulation, DNA-damage response and repair, and stress-response processes. |

Read-only ground-truth fact sheet.

| ID | Reference fact | Value |
|----|---|---|
| A1 | Sequence length | 156 residues. |
| A2 | 3Di length and low-confidence stretches | 156 tokens, aligned 1:1 with the amino-acid sequence; no # low-confidence tokens are present. |
| A3 | Diagnostic motifs | Basic-rich segments RLLHRE at residue 19 and RRDARGL at residue 128. The sequence is built from tandem ankyrin repeats; it does not contain a real DNA-binding or bZIP motif. |

| | | |
|----|--------------------------------|--|
| A4 | Ground-truth identity / family | Cyclin-dependent kinase 4 inhibitor D (p19INK4d / CDKN2D), human. It is an ankyrin-repeat CDK inhibitor and tumor suppressor. |
| A5 | Ground-truth BP region | Negative regulation of cell cycle and cell-cycle phase transition, especially G1/S; regulation of CDK or protein-serine/threonine kinase activity; apoptotic or programmed cell-death regulation; DNA-damage response and repair; response to stress, radiation, or chemical stimulus. Not transcription or DNA-templated gene expression. |
| A6 | Number of true terms | $N_{\text{true}} = 145$. |
| A7 | UniProt / InterPro constraints | InterPro IPR050776 Ank_Repeat/CDKN_Inhibitor, IPR002110 Ankyrin_rpt, and IPR036770 Ankyrin_rpt-contain_sf. Keywords include Cell cycle, Tumor suppressor, ANK repeat, Nucleus, and Cytoplasm. The protein inhibits CDK4 and CDK6. |

Example claim prompts shown to the evaluator.

| Axis | Model A claim examples | Model B claim examples |
|------|---|---|
| Q1 | The trace cites 3Di runs, loop-like 3Di segments, and basic sequence clusters such as RLLHRE. | The trace quotes the full sequence and identifies a basic region, but also claims a heptad or leucine-zipper-like pattern. |
| Q2 | The trace infers a regulatory protein and predicts cell-cycle, apoptosis, kinase-regulation, and stress-response terms. | The trace identifies the protein as a bZIP, WRKY, or transcription factor-like protein and predicts transcriptional regulation terms. |
| Q3 | Model A’s committed terms largely overlap the CDK inhibitor BP region. | Model B’s committed terms center on transcription and gene expression, which is outside the ground-truth BP region. |
| Q4 | Evaluators check whether the trace moves from sequence and 3Di evidence to the committed GO region without unsupported leaps. | Evaluators check format-oriented reasoning, identity contradictions, and whether the final term list follows from grounded evidence. |
| Q5 | Named biological entities and GO terms should be checked for unsupported specificity or fabrication. | The bZIP, WRKY, leucine-zipper, and transcription-factor claims must be judged as real-but-misassigned or fabricated, following the rubric. |

B.5 Retrosynthesis: USPTO-50K sample 4, other

| Field | Value |
|----------------|---|
| Dataset / task | USPTO-50K / retrosynthesis. |
| Sample ID | 4. |
| Reaction class | Other / uncategorized retrosynthesis reaction. |
| Matches gold | Model A matches the gold reactants; Model B proposes a related but not gold reactant set. |

Input prompt.

```
Please suggest potential reactants for the given product.
<SMILES> [C:1](=[O:2])([C:3]([F:4])([F:5])[F:6])[NH:7][CH2:8][c:9]1[cH:10][cH:11]
[cH:12][cH:13][c:14]1[S:15](=[O:16])(=[O:17])[CH:18]1[CH2:19][CH2:20]1 </SMILES>
```

Model outputs and gold reactants.

| Field | SMILES |
|---------|---|
| Product | [C:1](=[O:2])([C:3]([F:4])([F:5])[F:6])[NH:7][CH2:8][c:9]1[cH:10][cH:11][cH:12][cH:13][c:14]1[S:15](=[O:16])(=[O:17])[CH:18]1[CH2:19][CH2:20]1. |

| | |
|--------------------------|--|
| Gold reactants | <chem>O=C(O[C:1](=[O:2])[C:3]([F:4])([F:5])[F:6])C(F)(F)F.[NH2:7][CH2:8][c:9]1[cH:10][cH:11][cH:12][cH:13][c:14]1[S:15](=[O:16])(=[O:17])[CH:18]1[CH2:19][CH2:20]1.</chem> |
| Model A final prediction | <chem>O=C(O[C:1](=[O:2])[C:3]([F:4])([F:5])[F:6])C(F)(F)F.[NH2:7][CH2:8][c:9]1[cH:10][cH:11][cH:12][cH:13][c:14]1[S:15](=[O:16])(=[O:17])[CH:18]1[CH2:19][CH2:20]1.</chem> |
| Model B final prediction | <chem>NCC1cccc1S(=O)(=O)C1CC1.OC(=O)C(F)(F)F.</chem> |

Read-only ground-truth fact sheet.

| ID | Reference fact | Value |
|----|--|--|
| A1 | Product functional-group inventory | Trifluoroacetyl C(=O)CF ₃ at atom maps 1–6; secondary amide N-H at atom 7; benzylic CH ₂ at atom 8; ortho-disubstituted benzene ring at atoms 9–14; aryl sulfone at atom 15 with oxygens 16 and 17; cyclopropyl atoms 18–20. |
| A2 | Highest atom-map index | 20; any cited atom-map index > 20 is fabricated. |
| A3 | Gold reaction class and formed bond | Trifluoroacetylation of a primary amine. The forward reaction forms the amide C–N bond between carbonyl carbon [C:1] and nitrogen [NH:7]. The acyl source is trifluoroacetic anhydride (TFAA). |
| A4 | Gold reactant atom-map set and legitimate lost atoms | The TFAA fragment carries atoms 1–6, while the amine fragment carries atoms 7–20. The union covers the full product atom map. The unmapped leaving group is a trifluoroacetate or water-equivalent by-product. |

Example claim prompts shown to the evaluator.

| Axis | Model A claim examples | Model B claim examples |
|------|--|---|
| Q1 | The trace identifies the trifluoroacetyl group, amide N, benzyl group, sulfone, and cyclopropyl group. | The trace parses the product as a trifluoroacetamide attached to an ortho-substituted aryl sulfone and cyclopropyl group. |
| Q2 | The trace disconnects the C1–N7 amide bond and proposes TFAA plus the primary amine. | The trace disconnects the same amide bond but proposes trifluoroacetic acid plus the amine. |
| Q3 | Model A matches the gold amide-forming reaction family and gold formed bond. | Model B finds the same C–N disconnection but uses a less activated acyl source than the gold TFAA route. |
| Q4 | The trace follows product parsing to amide disconnection to reactant selection. | Evaluators check for length padding, repeated route restatement, and wobble among acid, acid chloride, and anhydride before commitment. |
| Q5 | TFAA, primary amine acylation, and amide formation are real and supported by the product. | Trifluoroacetic acid and the amine are real; evaluators check whether any unsupported reaction claims or invented reagents appear. |

Input prompt

Chemical formula: BaLaFe2Cu2011

Target property: band_gap

Material structure:

```
<material_structure>o x b 000 j 000 o Ba La Fe Fe Cu Cu 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 -oo 0 12 -oo 0 16
-oo 0 16 ooo 0 7 -oo 0 11 -oo 0 14 ooo 0 14 o+o 0 6 ooo 0 10 ooo 0 9 ooo 0 13 ooo 1 9 o-o 1 13 o-o 1
14 ooo 1 14 +oo 1 6 ooo 1 10 ooo 1 16 o-o 1 16 ooo 1 7 o-o 1 11 o-o 1 8 ooo 1 12 ooo 2 7 --o 2 8 -oo
2 9 o-o 2 6 ooo 2 14 ooo 3 10 oo- 3 13 oo- 3 12 oo- 3 11 oo- 3 15 ooo 3 16 oo- 4 11 --- 4 12 -o- 4
13 o-- 4 10 oo- 4 14 oo- 5 6 ooo 5 9 ooo 5 8 ooo 5 15 ooo 5 16 ooo 5 7 ooo 6 14 ooo 6 16 ooo 7 14
+o 7 16 ooo 8 14 +oo 8 16 ooo 9 14 o+o 9 16 ooo 10 14 ooo 10 16 ooo 11 14 ++o 11 16 ooo 12 14 +oo
12 16 ooo 13 14 o+o 13 16 ooo</material_structure>
```

Question:

You are a material scientist. Look at the chemical formula and structure information of the given crystalline material and predict its property. The output must be in a json format. For example: property_name: predicted_property_value. Answer as precise as possible and in as few words as possible.

chemical formula: BaLaFe2Cu2O11

structure information: <material_structure>o x ...</material_structure>

property name: band_gap

SciReasoner reasoning trace

<think>

To determine the band gap of this material, I first need to decode the structural information provided in the SLICES string to understand the chemical environment and bonding topology. The input string begins with the space group encoding <material_structure>o x b 000 j 000 o</material_structure>, which defines the symmetry operations, followed immediately by the atom list. Reading the atom symbols in order, I see <material_structure>Ba La Fe Fe Cu Cu 0 0 0 0 0 0 0 0 0 0 0</material_structure>. This confirms the stoichiometry matches the formula BaLaFe2Cu2O11, consisting of one Barium, one Lanthanum, two Iron, two Copper, and eleven Oxygen atoms per unit cell.

My primary focus for predicting the electronic property is the transition metal sublattice, specifically the Iron and Copper sites, which correspond to indices 2, 3, 4, and 5 in the atom list. I need to examine the bonding environment to see how these metals interact with the oxygen ligands. Scanning the edge list, I find connections such as <material_structure>2 6 ooo</material_structure> and <material_structure>2 14 ooo</material_structure>, which indicate that the Iron atom at index 2 is bonded to Oxygen atoms at indices 6 and 14 within the same unit cell. Similarly, the Copper atom at index 5 shows connections like <material_structure>5 15 ooo</material_structure> and <material_structure>5 16 ooo</material_structure>, linking it to Oxygen sites 15 and 16.

Crucially, I observe that the Oxygen sites are not isolated; they act as bridges connecting the different metal centers. For instance, Oxygen site 14 is connected to both Iron (index 2 via <material_structure>2 14 ooo</material_structure>) and Copper (index 1 via <material_structure>1 14 ooo</material_structure>), creating a Fe-O-Cu pathway. This heteroatomic networking allows for significant orbital overlap between the transition metal d-orbitals and the oxygen p-orbitals. In materials containing mixed transition metals like Iron and Copper, the presence of these continuous pathways often leads to broadened energy bands. If the Fermi level intersects these bands, the material exhibits metallic behavior. The dense connectivity evident in the edge list, with multiple intra-cell and inter-cell bonds like <material_structure>0 16 -oo</material_structure> and <material_structure>13 14 o+o</material_structure>, supports a scenario where charge carriers can delocalize across the lattice. Therefore, based on the structural evidence of a connected transition metal-oxygen framework, the most consistent prediction for the band gap is that it closes completely.

band_gap: 0.0
</think>

SciReasoner final output

band_gap: 0.0

Fig. B1: Full example of a structure-aware reasoning trace. The input contains the chemical formula, target property, and SLICES-PLUS structural string. The reasoning trace explicitly cites structural substrings from the input and connects them to a material-property mechanism before producing the final prediction.

Input prompt

Protein name: 1t4b_A-POA9Q9

Protein structure:

```
<protein_structure>D F E E E E A C L V F P L N V V L L V V C L V V V V L V V Y N Y A Y E D C P Q F Q
D F D D V S S V D R D T H H H L Q D L V V V L R G Q E Y A Y P N E D V S C V P R V V V N V V V R L
H A Y E E L Y P N C P P P P L E A E F D C L Q F V V V V V S V V V S R R Y Y Y H A A L V L L L C C
L Q Q S L C V V V Q F F Q A K E K D K A F A Q C S L G D V S L V L L V Q Q V V V L C V Q C V L V V
P P P R D V V V S V S S V C C Q P V P P R D C V R V V H D A H V D K D F A D A D A D P V R H H P V
Q V S Q Q V N N C R S N V D P D G H Y Y G Y I D M Y H S A H F K M K M K M W T F T P D Q D D L V V S
Q V S S C V S D V Q N A E Q E Q D D P S C V Q C A D S V N A E P H S H W Y K Y P W D A D P V D R R I
T I I M I M H G C S D S V G N P R S S V V V V V V D</protein_structure>
```

Question:

Given the protein sequence

```
<protein>MKNVGFIGWRGMVGSVLMQRMVEERDFDAIRPVFFSTSQLQAAPSFGGTTGTLQDAFDLEALKALDIIIVTCQGGDYTNEIYPKLRESGW
QGYWIDAASSLRMKDDAIIILDVNVQDVTIDGLNNGIRTFVGGNCTVSLMLMSLGGFLFANDLVDWVSVATYQAASGGGARHMRLLTQMGLHYGHVADE
LATPSSAILDIERKVTTLTRSGELPVDNFGVPLAGSLIPWIDKQLDNGQSREEWKQGAETNKILNTSSVIPVDGLCVRVGLRCHSQAFTIKLKKDVS
PTVEELLAHNPAKVVVNDREITMRELTPAAVTGTLTTPVGRRLRKLNMGPFLSAFTVGDQLLWGAAEPLRRMLRQLA</protein> and its
structure representation <protein_structure>D F E ... </protein_structure>, predict its Gene
Ontology annotations for biological process. Only the biological process (BP) aspect.
```

SciReasoner reasoning trace

<think>

To determine the appropriate Gene Ontology (GO) biological process annotations for this protein, I need to analyze the relationship between its amino acid sequence and its predicted three-dimensional structure. The core task is to infer the protein's function---specifically its role in cellular processes---based on the structural motifs and sequence signatures provided in the input. I will start by assessing the overall structural class defined by the 3Di representation, then identify specific sequence motifs that indicate enzymatic activity, and finally map these findings to the relevant biological processes.

First, I examine the structural representation to understand the protein's fold. The 3Di string provides a residue-level encoding of the local geometry. Scanning through the structure, I observe a significant prevalence of residues encoding alpha-helical conformations. For instance, the segment <protein_structure>V V L L V V C L V V V V L V V</protein_structure> near the N-terminal region shows a continuous run of V and L characters, which according to the decoding rules, signifies a stable alpha-helix. Similarly, towards the C-terminus, the pattern <protein_structure>V V V V V V D</protein_structure> indicates another helical segment. The presence of these helical bundles suggests a globular protein with a compact core, typical of enzymes that require a defined active site pocket. While there are regions with mixed characters like <protein_structure>D F E E E A C L V F P L N</protein_structure>, which likely represent loop or coil regions connecting the secondary structure elements, the dominant feature is the alpha-helical architecture. This structural class is commonly associated with metabolic enzymes that bind small molecules.

Next, I turn to the amino acid sequence to identify specific functional motifs that correlate with the structural observations. The sequence contains several patterns indicative of cofactor binding or catalytic activity. Notably, there is a glycine-rich region <protein>GWQGYWIDAASSLRMKDDAIIILDVNVQDVTIDGLNNGIRTFVGGNCTVSLMLMSLGGFLFANDLVDWVSVATYQAASGGGARHMRLLTQMGLHYGHVADELATPSSAILDIERKVTTLTRSGELPVDNFGVPLAGSLIPWIDKQLDNGQSREEWKQGAETNKILNTSSVIPVDGLCVRVGLRCHSQAFTIKLKKDVS IPTVEELLAHNPAKVVVNDREITMRELTPAAVTGTLTTPVGRRLRKLNMGPFLSAFTVGDQLLWGAAEPLRRMLRQLA</protein>, specifically the <protein>VGGNCTVSLMLMSLGGFLFANDLVDWVSVATYQAASGGGARHMRLLTQMGLHYGHVADELATPSSAILDIERKVTTLTRSGELPVDNFGVPLAGSLIPWIDKQLDNGQSREEWKQGAETNKILNTSSVIPVDGLCVRVGLRCHSQAFTIKLKKDVS IPTVEELLAHNPAKVVVNDREITMRELTPAAVTGTLTTPVGRRLRKLNMGPFLSAFTVGDQLLWGAAEPLRRMLRQLA</protein> stretch which includes conserved residues often found in nucleotide-binding domains. The presence of lysine and arginine residues, such as in the segment <protein>KDDAIIILDVNVQDVTIDGLNNGIRTFVGGNCTVSLMLMSLGGFLFANDLVDWVSVATYQAASGGGARHMRLLTQMGLHYGHVADELATPSSAILDIERKVTTLTRSGELPVDNFGVPLAGSLIPWIDKQLDNGQSREEWKQGAETNKILNTSSVIPVDGLCVRVGLRCHSQAFTIKLKKDVS IPTVEELLAHNPAKVVVNDREITMRELTPAAVTGTLTTPVGRRLRKLNMGPFLSAFTVGDQLLWGAAEPLRRMLRQLA</protein>, suggests potential interaction with phosphate groups, a hallmark of enzymes involved in phosphorylation or nucleotide metabolism. Furthermore, the sequence contains motifs like <protein>MGLHYGHVADELATPSSAILDIERKVTTLTRSGELPVDNFGVPLAGSLIPWIDKQLDNGQSREEWKQGAETNKILNTSSVIPVDGLCVRVGLRCHSQAFTIKLKKDVS IPTVEELLAHNPAKVVVNDREITMRELTPAAVTGTLTTPVGRRLRKLNMGPFLSAFTVGDQLLWGAAEPLRRMLRQLA</protein> which resemble those found in synthase or transferase families. The combination of the alpha-helical structural scaffold and these specific sequence motifs strongly points towards an enzyme that participates in the synthesis or modification of amino acids and nucleotides.

Based on this structural and sequence evidence, I can deduce the biological processes. The protein's ability to bind nucleotides and its enzymatic nature imply involvement in metabolic pathways. Specifically, the motifs suggest it may catalyze reactions in the biosynthesis of amino acids, particularly those involving sulfur or serine derivatives, and potentially interact with pyruvate or related intermediates. The structural stability indicates it functions within the cell cytoplasm or a specific compartment where these metabolites are abundant. Therefore, the annotations should reflect its role in metabolic processes, specifically focusing on the synthesis and regulation of amino acids and nucleobase compounds. I will verify this by checking for consistency with the observed structural constraints; the alpha-helical dominance supports a soluble enzyme function rather than a membrane transporter, and the specific motifs align with biosynthetic pathways rather than degradation.

</think>

SciReasoner final output

amino acid metabolic process; carboxylic acid biosynthetic process; lysine metabolic process; obsolete diaminopimelate metabolic process; dicarboxylic acid metabolic process; sulfur compound metabolic process; dicarboxylic acid biosynthetic process; methionine biosynthetic process; diaminopimelate biosynthetic process; isoleucine biosynthetic process; sulfur compound biosynthetic process; threonine metabolic process; methionine metabolic process; small molecule biosynthetic process; lysine biosynthetic process; sulfur amino acid biosynthetic process; branched-chain amino acid metabolic process; alpha-amino acid metabolic process; alpha-amino acid biosynthetic process; sulfur amino acid metabolic process; isoleucine metabolic process; amino acid biosynthetic process; carboxylic acid metabolic process; lysine biosynthetic process via diaminopimelate; aspartate family amino acid biosynthetic process; L-methionine biosynthetic process; organic acid biosynthetic process; organic acid metabolic process; obsolete aspartate family amino acid metabolic process; branched-chain amino acid biosynthetic process; oxoacid metabolic process

Fig. B2: Full example of a structure-aware reasoning trace for biological-process Gene Ontology prediction. The input contains the amino-acid sequence of 1t4b.A-POA9Q9 and its residue-level 3Di structural representation. The reasoning trace cites sequence and structural substrings and connects them to metabolic and amino-acid-biosynthetic processes before producing the final set of predicted BP annotations.

Input prompt

Reaction class: alkylation

Question:

Please suggest potential reactants for the given product.

```
<SMILES> [CH2:1]([CH2:2][CH2:3][C:4]([c:5]1[c:6]([CH3:7])[s:8][c:9]2[c:10]1[CH2:11][N:12]([C:13]([C]
H3:14))=[O:15])[CH2:16][CH2:17]2)=[O:18])[N:28]1[CH2:27][CH2:26][CH:25]([c:24]2[c:23]3[cH:22][cH:21]
)[c:20]([F:19][cH:34][c:33]3[o:32][n:31]2)[CH2:30][CH2:29]1 </SMILES>
```

SciReasoner reasoning trace

<think>

Let's break down the retrosynthetic analysis for the provided product.

1. Analysis of the Product Structure

The product is a complex molecule with two main structural domains connected by a propyl chain:

- Domain A (Left): A thienopyridine ring system substituted with an acetyl group and a methyl group. This domain is attached to the propyl chain via a carbonyl group (ketone).
- Domain B (Right): A piperidine ring substituted with a fluorinated benzisoxazole group. This domain is attached to the propyl chain via a nitrogen atom (tertiary amine).

The connection between these domains is a propyl linker: ``[CH2:1]-[CH2:2]-[CH2:3]``.

2. Retrosynthetic Disconnection

The most strategic bond to disconnect is the C-N bond between the propyl chain and the piperidine nitrogen (bond between ``[CH2:1]`` and ``[N:28]``).

- This is a classic alkylation disconnection.
- The propyl chain acts as the alkylating agent (electrophile).
- The piperidine nitrogen acts as the nucleophile.

3. Identifying the Reactants

Based on this disconnection, we can propose the two reactants:

* Reactant 1 (The Alkylating Agent):

The propyl chain must be attached to a good leaving group. In organic synthesis, **chloride** is a standard leaving group for such alkylations.

- Structure: ``Cl-CH2-CH2-CH2-Domain A``.

- In SMILES: ``Cl[CH2:1][CH2:2][CH2:3][C:4](=[O:18])[c:5]1[c:6]([CH3:7])[s:8][c:9]2[c:10]1[CH2:1]1[N:12]([C:13]([CH3:14])=[O:15])[CH2:16][CH2:17]2``.

* Reactant 2 (The Nucleophile):

The piperidine ring must have a free hydrogen on the nitrogen to act as a nucleophile.

- Structure: ``Domain B-H``.

- In SMILES: ``[F:19][c:20]1[cH:21][cH:22][c:23]2[c:24]([CH:25]3[CH2:26][CH2:27][NH:28][CH2:29][CH2:30]3)[n:31][o:32][c:33]2[cH:34]1``.

4. Sanity Check

- Chemical Feasibility: The alkylation of a secondary amine (piperidine) with a primary alkyl chloride is a high-yielding, standard reaction, typically performed in the presence of a base (like K2CO3) to scavenge the HCl byproduct.

- Selectivity: The piperidine nitrogen is the most nucleophilic site in Reactant 2. The amide nitrogen in the thienopyridine ring (Reactant 1) is much less nucleophilic due to delocalization and is not a competing site for alkylation under these conditions.
- Connectivity: The disconnection correctly places the propyl chain on the nitrogen, matching the product structure.

5. Final Answer

The possible reactants are the alkyl chloride derivative of the thienopyridine domain and the substituted piperidine.

```
<SMILES> C1[CH2:1][CH2:2][CH2:3][C:4]([c:5]1[c:6]([CH3:7])[s:8][c:9]2[c:10]1[CH2:11][N:12]([C:13]([CH3:14])=[O:15])[CH2:16][CH2:17]2)=[O:18].[F:19][c:20]1[cH:21][cH:22][c:23]2[c:24]([CH:25]3[CH2:26][CH2:27][NH:28][CH2:29][CH2:30]3)[n:31][o:32][c:33]2[cH:34]1 </SMILES>
```

</think>

SciReasoner final output

```
<SMILES> C1[CH2:1][CH2:2][CH2:3][C:4]([c:5]1[c:6]([CH3:7])[s:8][c:9]2[c:10]1[CH2:11][N:12]([C:13]([CH3:14])=[O:15])[CH2:16][CH2:17]2)=[O:18].[F:19][c:20]1[cH:21][cH:22][c:23]2[c:24]([CH:25]3[CH2:26][CH2:27][NH:28][CH2:29][CH2:30]3)[n:31][o:32][c:33]2[cH:34]1 </SMILES>
```

Fig. B3: Full example of a retrosynthetic reasoning trace for an alkylation reaction from the USPTO-50K dataset. The input contains the atom-mapped SMILES representation of the target product. The reasoning trace identifies the bond between [CH2:1] and [N:28] as the strategic C–N disconnection, interprets the transformation as alkylation of a secondary piperidine amine with a primary alkyl chloride, and produces the corresponding reactant SMILES.

References

- [1] Thakur, A., Kim, J., Wyatt, B. C., Gogotsi, Y. & Anasori, B. Composition–structure–property relationships in mxenes. *Nature Reviews Materials* 1–16 (2026).
- [2] Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews* **112**, 2889–2919 (2012).
- [3] Van de Walle, A. A complete representation of structure–property relationships in crystals. *Nature materials* **7**, 455–458 (2008).
- [4] Creighton, T. E. *Proteins: structures and molecular properties* (Macmillan, 1993).
- [5] Tai, T.-Y. *et al.* Leveraging isothermal titration calorimetry to explore structure–property relationships of protein immobilization in metal–organic frameworks. *Angewandte Chemie International Edition* **61**, e202209110 (2022).
- [6] Osadchy, M. & Kolodny, R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences* **108**, 12301–12306 (2011).
- [7] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
- [8] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).
- [9] Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120**, 145301 (2018).
- [10] Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
- [11] Strieth-Kalthoff, F. *et al.* Artificial intelligence for retrosynthetic planning needs both data and expert knowledge. *Journal of the American Chemical Society* **146**, 11005–11017 (2024).
- [12] Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* **555**, 604–610 (2018).
- [13] Duek, P., Mary, C., Zahn-Zabal, M., Bairoch, A. & Lane, L. Functionathon: a manual data mining workflow to generate functional hypotheses for uncharacterized human proteins and its application by undergraduate students. *Database* **2021**, baab046 (2021).
- [14] Zhao, D., Tu, S. & Xu, L. Efficient retrosynthetic planning with mcts exploration enhanced a* search. *Communications Chemistry* **7**, 52 (2024).
- [15] Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [16] Mirza, A. *et al.* Are large language models superhuman chemists? *Nature Chemistry* (2025).
- [17] Narayanan, S. *et al.* Training a scientific reasoning model for chemistry. *Advances in Neural Information Processing Systems* **38**, 157671–157710 (2026).
- [18] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
- [19] Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* 1–3 (2026).
- [20] Aygün, E. *et al.* An ai system to help scientists write expert-level empirical software. *Nature* 1–3 (2026).
- [21] Gottweis, J. *et al.* Accelerating scientific discovery with co-scientist. *Nature* 1–3 (2026).
- [22] Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
- [23] Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59**, 3370–3388 (2019).
- [24] Anthropic. System card: Claude opus 4.7. <https://anthropic.com/claude-opus-4-7-system-card> (2026).
- [25] OpenAI. Introducing GPT-5.5. <https://openai.com/index/introducing-gpt-5-5/> (2026). Accessed: 2026-06-05.

- [26] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence (2026).
- [27] Kimi. Kimi k2.6 tech blog: Advancing open-source coding. <https://www.kimi.com/blog/kimi-k2-6> (2026). Accessed: 2026-07-07.
- [28] Su, J. *et al.* Democratizing protein language model training, sharing and collaboration. *Nature Biotechnology* **1–7** (2025).
- [29] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). URL <https://www.science.org/doi/abs/10.1126/science.ade2574>.
- [30] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).
- [31] Deng, Y. *et al.* Rsgpt: a generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints. *Nature communications* **16**, 7012 (2025).
- [32] Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry* **55**, 6582–6594 (2012).
- [33] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
- [34] The gene ontology resource: enriching a gold mine. *Nucleic acids research* **49**, D325–D334 (2021).
- [35] Consortium, T. U. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research* **51**, D523–D531 (2023).
- [36] Friedberg, I. Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics* **7**, 225–242 (2006).
- [37] Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nature methods* **10**, 221–227 (2013).
- [38] Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggno-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* **38**, 5825–5829 (2021).
- [39] Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nature communications* **12**, 3168 (2021).
- [40] Van Kempen, M. *et al.* Fast and accurate protein structure search with foldseek. *Nature biotechnology* **42**, 243–246 (2024).
- [41] Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists’ bread and butter. *Journal of medicinal chemistry* **59**, 4385–4402 (2016).
- [42] Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **3**, 1237–1245 (2017).
- [43] Segler, M. H. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry—A European Journal* **23**, 5966–5971 (2017).
- [44] Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems* **32** (2019).
- [45] Chen, S. & Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **1**, 1612–1620 (2021).
- [46] Yan, C., Zhao, P., Lu, C., Yu, Y. & Huang, J. Retrocomposer: composing templates for template-based retrosynthesis prediction. *Biomolecules* **12**, 1325 (2022).
- [47] Shi, C., Xu, M., Guo, H., Zhang, M. & Tang, J. A graph to graphs framework for retrosynthesis prediction 8818–8827 (2020).

- [48] Yan, C. *et al.* Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems* **33**, 11248–11258 (2020).
- [49] Wang, X. *et al.* Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal* **420**, 129845 (2021).
- [50] Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, H. & Ning, X. G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry* **6**, 102 (2023).
- [51] Gao, Z., Tan, C., Wu, L. & Li, S. Z. Semiretro: Semi-template framework boosts deep retrosynthesis prediction. *arXiv preprint arXiv:2202.08205* (2022).
- [52] Zhong, W., Yang, Z. & Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications* **14**, 3009 (2023).
- [53] Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling* **60**, 47–55 (2019).
- [54] Sacha, M. *et al.* Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling* **61**, 3273–3284 (2021).
- [55] Tu, Z. & Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling* **62**, 3503–3513 (2022).
- [56] Zhong, Z. *et al.* Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science* **13**, 9023–9034 (2022).
- [57] Yao, L. *et al.* Node-aligned graph-to-graph: elevating template-free deep learning approaches in single-step retrosynthesis. *JACS Au* **4**, 992–1003 (2024).
- [58] Xiong, J. *et al.* Bridging three-dimensional molecular structures and artificial intelligence with a conformation description language. *Nature Machine Intelligence* 1–15 (2026).
- [59] Hemmer, M. C., Steinhauer, V. & Gasteiger, J. Deriving the 3d structure of organic molecules from their infrared spectra. *Vibrational spectroscopy* **19**, 151–164 (1999).
- [60] Devinyak, O., Havrylyuk, D. & Lesyk, R. 3d-morse descriptors explained. *Journal of Molecular Graphics and Modelling* **54**, 194–203 (2014).
- [61] Klein, C. T., Kaiser, D. & Ecker, G. Topological distance based 3d descriptors for use in qsar and diversity analysis. *Journal of chemical information and computer sciences* **44**, 200–209 (2004).
- [62] Axen, S. D. *et al.* A simple representation of three-dimensional molecular structure. *Journal of medicinal chemistry* **60**, 7393–7409 (2017).
- [63] Vainio, M. J., Puranen, J. S. & Johnson, M. S. Shaep: molecular overlay based on shape and electrostatic potential (2009).
- [64] Sastry, G. M., Dixon, S. L. & Sherman, W. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *Journal of chemical information and modeling* **51**, 2455–2466 (2011).
- [65] Liu, X., Jiang, H. & Li, H. Shafts: a hybrid approach for 3d molecular similarity calculation. 1. method and assessment of virtual screening. *Journal of chemical information and modeling* **51**, 2372–2385 (2011).
- [66] Hu, J., Liu, Z., Yu, D.-J. & Zhang, Y. Ls-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics* **34**, 2209–2218 (2018).
- [67] Niyongabo Rubungo, A., Arnold, C., Rand, B. P. & Dieng, A. B. Llm-prop: predicting the properties of crystalline materials using large language models. *npj Computational Materials* **11**, 186 (2025).
- [68] Dana, J. M. *et al.* Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research* **47**, D482–D489 (2019).
- [69] Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Research* **28**, 235–242 (2000).

- [70] Pubmed. <https://pubmed.ncbi.nlm.nih.gov/> (2026). Accessed: 2026-06-21.
- [71] Pubmed central. <https://www.ncbi.nlm.nih.gov/pmc/> (2026). Accessed: 2026-06-21.
- [72] Varadi, M. *et al.* Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).
- [73] Su, J. *et al.* Saprot: Protein language modeling with structure-aware vocabulary (2024).
- [74] Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).
- [75] Kshirsagar, M. *et al.* Rapid and accurate prediction of protein homo-oligomer symmetry using seq2symm. *Nature Communications* **16**, 2017 (2025).
- [76] Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530 (2018).
- [77] Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Advances in neural information processing systems* (2021).
- [78] Polykovskiy, D. *et al.* Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology* (2020).
- [79] NVIDIA *et al.* Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model (2025). URL <https://arxiv.org/abs/2508.14444>. arXiv:2508.14444.
- [80] Olmo, T. *et al.* Olmo 3 (2025). URL <https://arxiv.org/abs/2512.13961>. arXiv:2512.13961.
- [81] Su, E. *et al.* Sciif: Benchmarking scientific instruction following towards rigorous scientific intelligence. *arXiv preprint arXiv:2601.04770* (2026).
- [82] Yang, A. *et al.* Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
- [83] Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units 1715–1725 (2016).
- [84] Gage, P. A new algorithm for data compression. *The C Users Journal* **12**, 23–38 (1994).
- [85] Xiao, H. *et al.* An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications* **14**, 7027 (2023).
- [86] Yu, Q. *et al.* Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems* **38**, 113222–113244 (2026).
- [87] Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106 (2019).
- [88] Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* 1–9 (2022).
- [89] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025).
- [90] Suresh, V., Liu, L., Adjeroh, D. & Zhou, X. Rpi-pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic acids research* **43**, 1370–1379 (2015).
- [91] Zhang, Y. *et al.* Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic acids research* **52**, e3–e3 (2024).
- [92] Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).