S1-Omni: A Unified Multimodal Reasoning Model for Scientific Understanding, Prediction, and Generation

ScienceOne AI
Wenge AI

https://huggingface.co/ScienceOne-AI
https://modelscope.cn/organization/ScienceOne-AI
https://github.com/ScienceOne-AI

Abstract

We present S1-Omni, a unified multimodal reasoning model for scientific understanding, prediction, and generation. AI for Science (AI4S) has advanced significantly through domain-specific models, tool-augmented LLMs, and scientific language models. However, model capabilities remain highly fragmented, limiting the joint modeling of heterogeneous data, scientific laws, and expert knowledge. S1-Omni addresses this gap by consolidating these capabilities into a single, coherent scientific reasoning model. The architecture of S1-Omni is built upon three core components: unified representation of scientific data, natural-world knowledge alignment, and decoding for domain-specific tasks. First, S1-Omni maps natural-language instructions and scientific objects, including CIF, SMILES, protein sequences, spectra, and scientific images, into a shared representation space. Second, it incorporates scientific laws and expert knowledge into data construction and training, enabling the model to reason from scientific evidence. Third, it performs task-specific decoding to support a broad range of applications, including property prediction, spectrum-to-molecular generation, protein site and structure prediction, and scientific image generation and editing. S1-Omni is trained on S1-Omni-Corpus, which covers 200 scientific tasks and contains millions of reasoning samples, and is evaluated on over 60 scientific benchmarks. It outperforms GPT-5.5 and Gemini-3.1-Pro on most benchmarks and matches or surpasses domain-specific models on several benchmarks. Overall, S1-Omni provides a practical path toward unified scientific modeling.

1 Introduction

In recent years, AI for Science (AI4S) has advanced along three complementary directions. The first comprises domain-specific models, represented by AlphaFold 3(Abramson et al., 2024) and ESM3 (Hayes et al., 2025), which develop high-performance capabilities for structure prediction, property prediction, or generation around specific scientific entities such as proteins, molecules, and materials. The second comprises tool-augmented general-purpose models, represented by GPT-Rosalind (OpenAI, 2026b) and Claude Science (Anthropic, 2026), which extend general language models with agents, scientific tools, and automated workflows for operating on and executing scientific tasks. The third comprises scientific language models, represented by NatureLM (Xia et al., 2025), LOGOS (Li et al., 2026a), and BioMatrix (Pei et al., 2026), which organize scientific entities such as proteins, DNA, and molecules as jointly modeled symbolic sequences or representation spaces and explore unified modeling across scientific objects. Together, these advances have driven domain modeling, scientific workflow automation, and representation learning for scientific objects, providing important foundations for unified multimodal reasoning in science.

However, existing scientific intelligence remains fragmented across domain-specific models, tool-augmented LLMs, and scientific language models. Domain-specific models are typically tied to particular scientific entities and task types. Tool-augmented LLMs use external tools to complete scientific tasks. Scientific language models attempt to unify scientific knowledge within a shared embedding space. However, these approaches have yet to bring knowledge from different disciplines, scientific principles, and constraints together within a unified model. The key to unified scientific intelligence is therefore not merely to provide a natural-language interface or to convert different scientific objects into a single token representation. It requires organizing heterogeneous scientific data within the same model process, injecting scientific laws, expert heuristics, and domain-task constraints into task understanding and model reasoning through implicit representations, and producing outputs that can be verified in their corresponding native scientific forms. This work investigates how these distributed capabilities can be integrated into a unified multimodal reasoning model, allowing a shared backbone to develop cross-disciplinary, cross-modal, and cross-task scientific understanding and reasoning.

To this end, we introduce S1-Omni, a unified multimodal reasoning model for scientific understanding, prediction, and generation. S1-Omni is built around three core capabilities. First, **unified representation of scientific data.** The model organizes natural-language instructions and diverse scientific objects into a unified task representation, covering text, material CIFs, chemical SMILES, protein sequences, spectra, scientific images, and other scientific modalities. At the same time, it preserves object-type boundaries, representation structures, and necessary encoding paths rather than equating unified modeling with uniform tokenization. Second, **natural-world knowledge alignment.** Model training does not rely only on statistical associations between inputs and outputs, but also incorporates scientific laws and expert knowledge, together with experimental facts, into data construction, sample validation, and training, enabling the model to form intermediate judgments from scientific evidence in the current task and thereby improving scientific reasoning and interpretability. Third, **decoding for domain-specific tasks.** On top of shared task understanding and scientific reasoning, the model connects the appropriate output representation and decoders according to the task objective. It supports property prediction, spectrum-to-molecular generation, protein structure generation, and scientific image generation and editing, allowing a unified model to produce outputs in each domain's native result space.

The capabilities of S1-Omni are built on S1-Omni-Corpus, which covers 200 scientific tasks and contains millions of high-quality scientific reasoning samples. During dataset construction, we use scientific laws and expert knowledge to design task-specific reasoning chains, taking into account relevant reasoning dimensions, evidence sources, and output constraints. In this way, S1-Omni learns not only surface patterns in scientific data, but also how to reason from evidence in a specific task, providing an effective initial path toward a scalable unified scientific modeling paradigm.

Our evaluation covers multiple scientific modalities, including molecules, materials, proteins, spectra, and scientific images, and supports tasks such as property prediction, spectrum-to-molecular generation, protein functional-site prediction, protein structure prediction, and scientific image generation and editing. Systematic comparisons on over 60 benchmarks show that S1-Omni outperforms GPT-5.5 and Gemini-3.1-Pro on most evaluation metrics, and matches or surpasses task-specific models on several specialized benchmarks. Specifically, it outperforms both GPT-5.5 and Gemini-3.1-Pro on 16 of 18 ADMET evaluations; matches or exceeds representative domain-specific models on several important drug-property tasks involving CYP enzymes, hERG, and intestinal absorption; supports multiple residue-level functional-site prediction tasks within a unified framework and achieves leading performance on protein–protein interaction, epitope, and small-molecule binding-site prediction; and demonstrates strong performance in scientific image generation and editing, including MSD medical image segmentation, medical image translation, super-resolution, and scientific illustration generation.

These results show that S1-Omni performs scientific understanding and reasoning across modalities and tasks in a single unified process, yielding outputs that are natively verifiable in their respective result spaces. To support reproducible research and further development, we release the S1-Omni weights, modeling and inference-serving code under the Apache-2.0 license, together with the high-quality S1-Omni-Corpus-10K subset.

2 Related Work

Domain-specific scientific models. Domain-specific scientific models are built around particular scientific objects, domain inductive biases, and evaluation protocols, and have achieved strong performance on their target tasks. AlphaFold 3 (Abramson et al., 2024) predicts structures of biomolecular complexes containing proteins, nucleic acids, small molecules, ions, and modified residues. ESM3 (Hayes et al., 2025) unifies generation across protein sequence, structure, and function, while ODesign (ODesign Team, 2025) further targets biomolecular interaction design. Similar specialization has emerged in other scientific domains: Materials Property Axiom (Deep Principle Team, 2026) focuses on heterogeneous experimental materials-property prediction, DiffSpectra (Wang et al., 2025a) reconstructs molecular structures from spectra, and S1-Omni-Image (Li et al., 2026b) supports scientific image understanding, generation, and editing. These models effectively exploit domain knowledge and scientific constraints, leading to high scientific validity within specific object types and task settings. However, their capabilities are usually tied to particular objects, data formats, and evaluation protocols, making it difficult to extend them naturally to other scientific fields or tasks. This suggests that domain-specific models remain an important component of scientific intelligence, while also raising the question of how diverse scientific objects and domain capabilities can be connected within a unified model.

Tool-augmented general models. General-purpose language models and open-domain multimodal models establish natural language as a unified interface for perception, generation, and tool use. From GPT-4 (OpenAI, 2023), Gemini (Team et al., 2023), Claude 3 (Anthropic, 2024), and ChatGLM (Glm et al., 2024) to multimodal systems such as LLaVA (Liu et al., 2023), NExT-GPT (Wu et al., 2023), GPT-

4o (OpenAI, 2024), Qwen2.5-Omni (Xu et al., 2025), SEED-X (Ge et al., 2024), and Janus-Pro (Chen et al., 2025), recent work has continuously expanded cross-modal understanding, generation, and task organization. In scientific settings, systems such as GPT-Rosalind (OpenAI, 2026b) and Claude Science (Anthropic, 2026) further combine agents, scientific tools, and workflows to support literature analysis, data processing, and experimental collaboration. This line of work substantially improves the ability of general models to organize scientific tasks and invoke specialized capabilities. Nevertheless, scientific knowledge and execution still largely reside in external tools and workflow logic, rather than being internalized as scientific competence within a unified model.

Scientific language models. Scientific language models attempt to map diverse scientific objects into unified representation spaces that can be jointly modeled, thereby improving interoperability across heterogeneous scientific data. NatureLM (Xia et al., 2025) represents small molecules, proteins, RNA, and materials as sequences in a language of nature. LOGOS (Li et al., 2026a) uses a scientific grammar to describe scientific objects, spatial constraints, and interactions. BioMatrix (Pei et al., 2026) jointly models biological sequences, structures, and natural language in a shared token space, while LucaOne (He et al., 2025) studies unified sequence modeling across nucleic acids and proteins. Omni-Weather (Zhou et al., 2025) unifies understanding and generation for weather radar data. These works show that shared tokens or scientific grammars can provide a common interface for heterogeneous scientific objects and improve cross-modal modeling. However, unifying object representations does not by itself amount to unifying scientific intelligence. How scientific laws, expert knowledge, and domain constraints should be incorporated into model training, and how they can further support consistent reasoning and result generation across scientific tasks, remain open problems.

Scientific reasoning and knowledge alignment. Recent scientific reasoning models further explore how scientific laws and domain knowledge can be explicitly incorporated into model reasoning. Proteo-R1 (Wu et al., 2026a) introduces reasoning mechanisms into de novo protein design. BioReason-Pro (Fallahpour et al., 2026) combines protein representations, multimodal biological context, and explicit reasoning for protein function prediction. DrugTrail (Liu et al.) couples structured reasoning traces with druggability-oriented preference optimization. In chemistry, ether0 (Narayanan et al., 2026), MPPReasoner (Zhuang et al., 2025), and Chem-R (Wang et al., 2025b) explore reinforcement learning, molecular-property reasoning traces, and expert-like reasoning supervision, respectively. PiFlow (Pu et al., 2025) further uses scientific principles to guide exploration, validation, and iterative refinement. These studies demonstrate that incorporating scientific laws into model training can improve scientific reasoning. However, most existing methods remain focused on particular domains or individual tasks, and have not yet formed a foundation model that can uniformly connect heterogeneous scientific data, scientific laws, and domain tasks.

Different from the above work, S1-Omni explores a unified multimodal reasoning architecture for scientific understanding, prediction, and generation, centered on three core capabilities: unified representation of scientific data, natural-world knowledge alignment, and decoding for domain-specific tasks. Specifically, S1-Omni first encodes heterogeneous scientific data—including text, molecules, materials, proteins, spectra, and scientific images—within a unified modeling framework. It then performs natural-world knowledge alignment using scientific laws and expert knowledge, so that task decisions are constrained by scientific evidence. Finally, it performs decoding for domain-specific tasks, enabling unified scientific understanding, prediction, and generation, rather than relying on a composition of multiple domain-specific models, external tools, or scientific language models.

3 Method

S1-Omni connects unified representation of scientific data, natural-world knowledge alignment, and decoding for domain-specific tasks within a single model process. Given a user instruction and scientific objects, a shared vision-language model first forms a task-conditioned hidden representation and generates optional scientific reasoning, a textual answer, and a task token. For tasks whose outputs can be expressed directly in natural language, the textual answer is the final result. For tasks with specialized output spaces—including properties, protein sites, molecular structures, 3D coordinates, and scientific images—the task token selects a result decoder that converts the shared representation into an output verifiable in its native scientific form.

Let the training set be

$$\mathcal{D} = \{d_i\}_{i=1}^{N}, \qquad d_i = (x_i, y_i, t_i, m_i), \tag{1}$$

where

$$x_i = (x_i^{\text{text}}, x_i^{\text{task}}), \qquad y_i = (y_i^{\text{text}}, y_i^{\text{task}}), \qquad y_i^{\text{text}} = (s_i, r_i, a_i). \tag{2}$$

Here, x_i^{text} contains the user instruction and any necessary textual context, and x_i^{task} contains one or

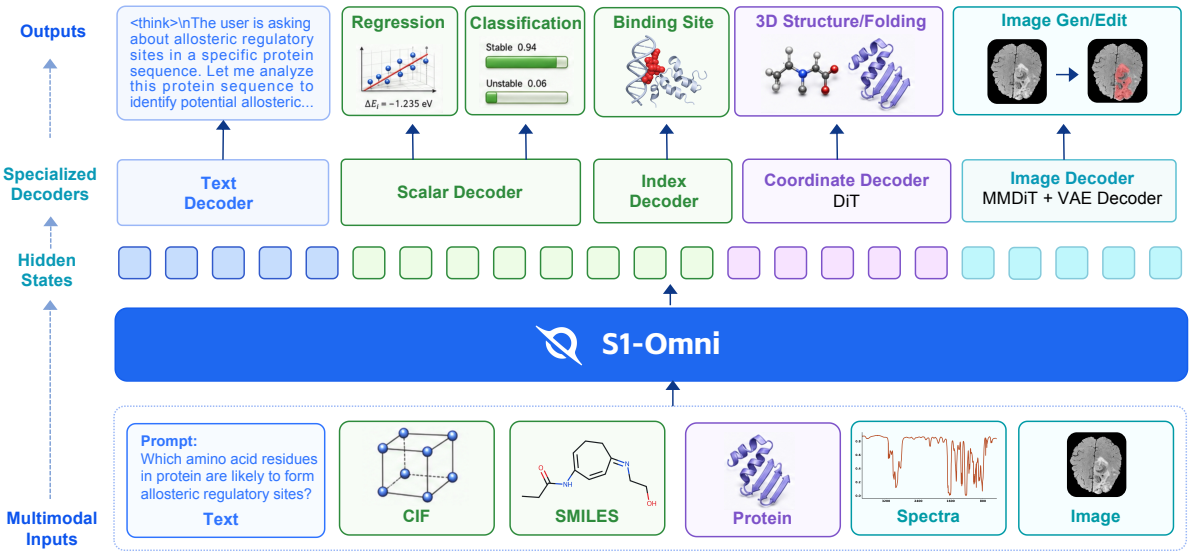


Figure 1: Unified architecture of S1-Omni. Text, CIF, SMILES, proteins, spectra, and scientific images are first processed by the shared vision-language model. Its hidden representations are then converted by text, scalar, position-index, coordinate, and image decoders into natural-language responses, property predictions, site predictions, three-dimensional structures, and scientific image generation or editing results.

more native scientific objects. The variables s_i, r_i, and a_i denote an optional task token, optional scientific reasoning, and a textual answer, while y_i^{task} is the domain-native supervision target. The task type is $t_i \in \mathcal{T} \cup \{\varnothing\}$, and m_i stores metadata such as provenance, data split, and traceable identifiers. Text tasks satisfy $t_i = s_i = \varnothing$ and $y_i^{\text{task}} = \varnothing$. Domain tasks have $t_i \in \mathcal{T}$, with s_i selecting the corresponding task decoder.

3.1 Model Architecture

As illustrated in Figure 1, S1-Omni uses S1-VL-32B (Li et al., 2026c) as its shared backbone and task result decoders as domain output interfaces. The backbone performs cross-modal task understanding and scientific reasoning, while the result decoders preserve the numerical, discrete, structural, geometric, and visual constraints of their respective output spaces. The two components communicate through VLM prefill hidden representations. This design neither compresses every scientific object into a single raw representation nor requires all tasks to share one output head.

3.1.1 Unified Representation of Scientific Data and Shared Task Representations

S1-Omni organizes text instructions and heterogeneous scientific objects within a common task context. Serializable objects, including SMILES strings, molecular formulae, FASTA sequences, CIF files, Wyckoff representations, and structure files, enter through the text channel. Scientific images, medical images, and visualized spectra enter through the visual channel. Input construction preserves object boundaries and modality roles, while the instruction specifies whether each object serves as a prediction condition, explanatory evidence, generation target, or editing condition.

The VLM task context is

$$c_i = \text{Fuse}\left(x_i^{\text{text}}, x_i^{\text{task}}\right). \tag{3}$$

The operator Fuse comprises the serialization, visual encoding, and context assembly required by different scientific objects. High-resolution residue, spectral-peak, pixel, or geometric information remains available in x_i^{task} and can be read directly by a downstream task decoder.

The shared VLM encodes the task context during prefill and generates a task response from the resulting hidden representation:

$$\hat{h}_i = F_\theta^{\text{prefill}}(c_i), \qquad (\hat{s}_i, \hat{r}_i, \hat{a}_i) = F_\theta^{\text{AR}}(\hat{h}_i). \tag{4}$$

Natural-world knowledge alignment shapes this representation through evidence selection, scientific relations, and reasoning supervision in the training data. Beyond identifying the input objects and task objective, the model must organize task-relevant intermediate judgments in $\hat{r}_i$, encouraging $\hat{h}_i$ to jointly

Table 1: Task tokens, prefill feature extraction, auxiliary conditions, decoders for domain-specific tasks, and domain-native results in S1-Omni.

Task family	Task token	Extract$_t$	Auxiliary condition	Decoder for domain-specific task	Domain-native result
Property regression	`<linear_pre>`	pooled	none	shared linear readout G_{linear}	continuous value
Property classification	`<linear_cla>`	pooled	none	shared linear readout G_{linear}	class score or probability
Protein-site prediction	`<prot_cla>`	token-level	ESM2 residue features	residue-level decoder G_{site}	residue probabilities
Spectrum-to-molecular generation	`<spectra_st>`	projected	SpecFormer spectral features	molecular diffusion decoder G_{mol}	molecular graph or 3D conformation
Protein structure prediction	`<prot_st>`	global	protein sequence representation	geometric decoder G_{fold}	three-dimensional coordinates
Scientific image generation and editing	`<image_gen>` / `<image_edit>`	token-level	input image for editing	MMDiT decoder G_{img}	generated or edited image

encode user intent, scientific evidence, and their relations. This representation supports both textual responses and domain-native result decoding.

When $\hat{s}_i \in \mathcal{T}$, the task-specific condition is extracted from the prefill representation as

$$\hat{e}_i = \text{Extract}_{\hat{s}_i}(\hat{h}_i). \tag{5}$$

Depending on the decoder, Extract$_t$ pools or selects token-level or global states, or projects the same hidden representation into a task-specific conditioning space. The task token selects only the task type and extraction rule; the domain prediction itself uses the shared condition $\hat{e}_i$.

3.1.2 Decoding for Domain-Specific Tasks

For a task requiring a domain-specific result, the selected decoder reads the shared task condition, when needed, the original scientific input:

$$\hat{y}_i^{\text{task}} = G_{\phi_{\hat{s}_i}}\left(\hat{e}_i, x_i^{\text{task}}\right). \tag{6}$$

The system output comprises a textual response and an optional domain result:

$$\hat{o}_i = \begin{cases} (\hat{r}_i, \hat{a}_i), & \hat{s}_i = \varnothing, \\ (\hat{r}_i, \hat{a}_i, \hat{y}_i^{\text{task}}), & \hat{s}_i \in \mathcal{T}. \end{cases} \tag{7}$$

The task token is an internal control signal and is not part of the user-visible output. When explicit reasoning is unnecessary, $\hat{r}_i = \varnothing$. Table 1 summarizes how each task family extracts its condition from the shared representation and incorporates any auxiliary scientific input.

In Table 1, pooled, token-level, projected, and global denote concrete implementations of Extract$_t$. Property tasks extract a pooled prefill condition and use a shared linear network for regression or classification. Protein-site prediction retains token-level conditions and fuses them with ESM2 residue features. Spectrum-to-molecular generation projects the shared representation into a molecular generation space while reading SpecFormer spectral features. Protein structure prediction uses a global task condition to drive a geometric decoder. Scientific image generation and editing preserve token-level conditions and map them into an MMDiT conditioning space, with editing additionally reading the input image. The shared backbone therefore provides a unified state for task understanding and scientific reasoning, while each decoder performs prediction or generation in the corresponding native result space.

3.2 Training Data

3.2.1 Corpus Scope and Unified Records

We construct S1-Omni-Corpus with partner research organizations and domain experts. The corpus contains more than 8 million filtered scientific reasoning and task supervision records spanning over 200 research tasks. It covers general text and multimodal tasks together with scientific domains including chemistry, biology, materials, physics, and medical imaging. Supported tasks include property prediction, spectrum-to-molecular generation, protein functional-site prediction, three-dimensional structure prediction, and scientific image generation and editing. Figure 2 summarizes its discipline, task, language, and dialogue-turn distributions.

Every record follows $d_i = (x_i, y_i, t_i, m_i)$. The field x_i^{text} stores the user instruction and context, while x_i^{task} stores domain-native scientific objects. The textual target y_i^{text} provides supervision for optional

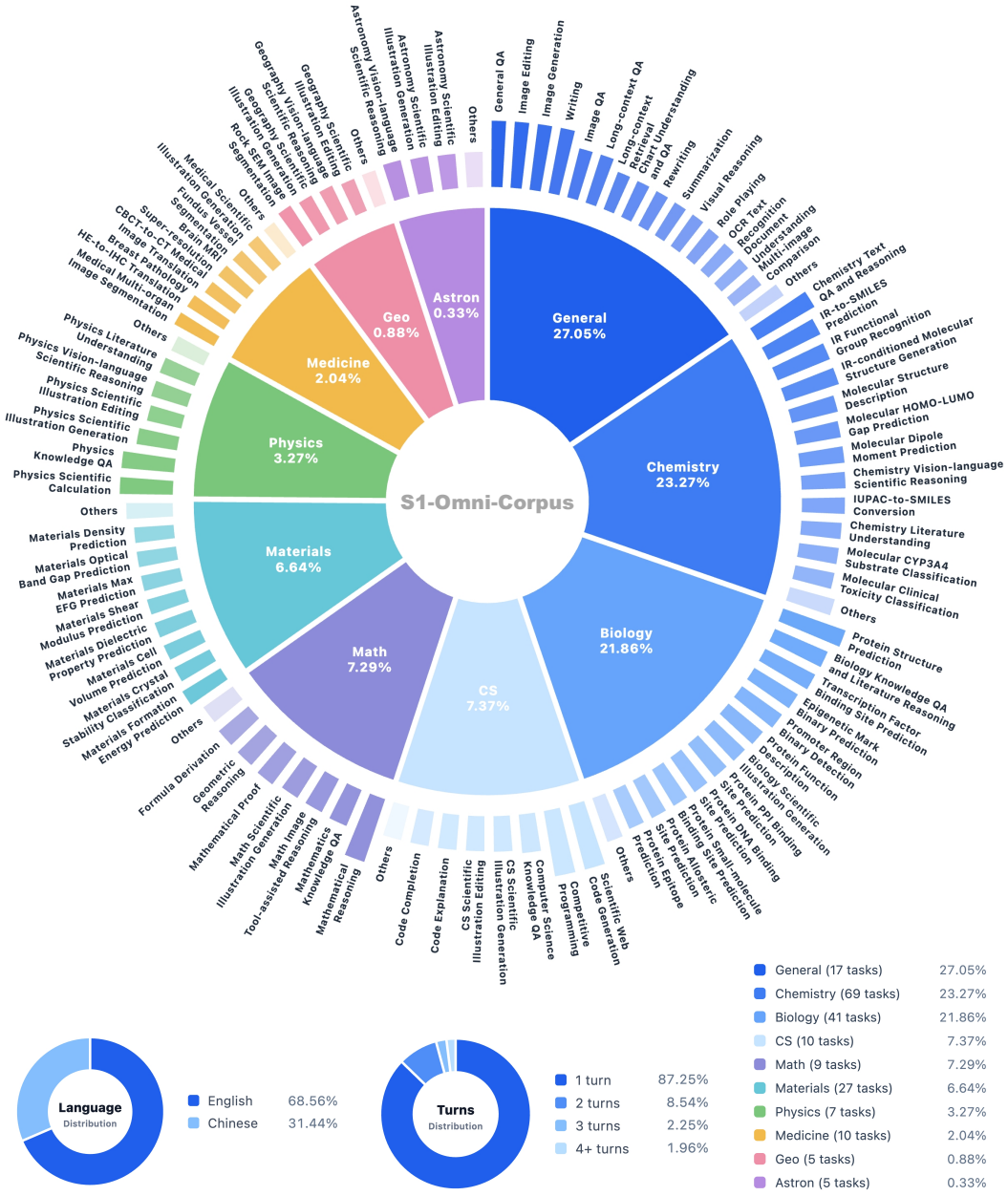


Figure 2: Discipline, task, language, and dialogue-turn distributions in S1-Omni-Corpus. These statistics characterize the coverage of the training data rather than define model capability directly by task count.

reasoning, the answer, and the task token, whereas y_i^{task} contains a continuous value, class label, residue annotation, molecular structure, three-dimensional coordinates, or target image. Domain-native targets are never written into visible reasoning and are supplied to the corresponding result decoder through an independent supervision path. The metadata m_i retains provenance, language, units, data split, augmentation history, and traceable identifiers. S1-Omni-Corpus-10K uses the same record format as the full corpus and supports data inspection, pipeline analysis, and community research.

3.2.2 Data Construction for Natural-World Knowledge Alignment

The data pipeline converts domain-native records into scientific task instances that are evidence-supported, protocol-consistent, and traceable. As shown in Figure 3, it comprises three stages: data foundation, knowledge alignment, and validation and assembly.

Data foundation. We first curate domain-native records for molecules, crystals, proteins, spectra, and scientific images. Deduplication, benchmark test-set overlap filtering, and train/validation/test

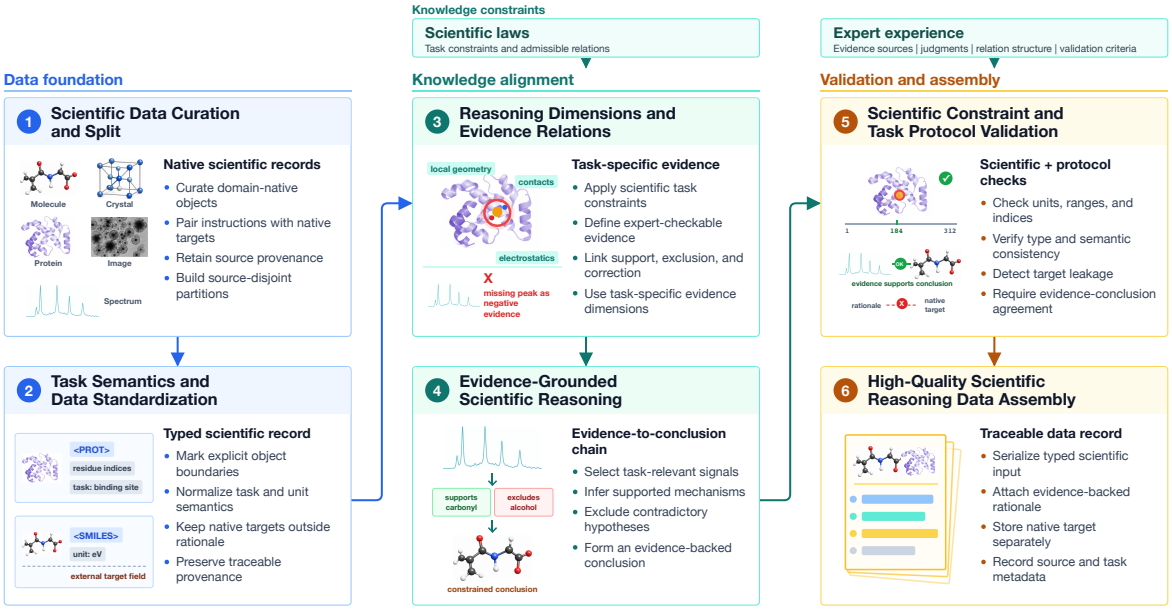


Figure 3: Data construction for natural-world knowledge alignment. Scientific laws and expert knowledge jointly constrain evidence dimensions, reasoning relations, task protocols, and sample validation.

partitioning are completed before data augmentation or reasoning distillation to prevent data leakage, ensuring that benchmark test instances or their near-duplicates do not enter the training corpus. Each raw record is then converted into a task instance d_i with standardized task semantics, units, indices, and output requirements, while preserving native representations such as SMILES, CIF, protein sequences, three-dimensional structures, raw spectra, and images. The target y_i^{task} is stored separately from the visible context to prevent result information from entering the reasoning input.

Natural-world knowledge alignment. Scientific laws specify the admissible evidence relations for a task, while domain experts determine the signals, decision process, and validation criteria that should be examined. Materials and chemistry tasks organize evidence around composition, structural fragments, local environments, and structure–property relations. Spectroscopy tasks jointly consider observed peaks, missing peaks, and cross-spectrum consistency. Protein tasks focus on sequence positions, local geometry, and residue contacts. Scientific image tasks examine entities, spatial relations, target regions, and editing boundaries. A candidate rationale r_i selects verifiable information from x_i and forms intermediate judgments through support, exclusion, and correction relations. If the available evidence cannot support a reliable explanation, we set $r_i = \varnothing$. Table 2 summarizes representative evidence dimensions and the corresponding validation logic across scientific domains.

Validation and assembly. Scientific consistency checks verify that the input evidence, reasoning process, textual answer, and domain target are mutually compatible. Task protocol checks cover object types, units, indices, media references, task tokens, output formats, and target leakage. The native target y_i^{task} is used only for supervision and consistency validation and is never introduced into r_i as hidden evidence. Samples that pass validation retain the complete x_i, y_i, t_i, and m_i, and undergo task balancing, language expansion, and necessary data augmentation within the established data partitions.

3.3 Training

S1-Omni uses two-stage training. The first stage trains the shared VLM under the unified task protocol to learn task understanding, scientific reasoning, textual answering, and task-token generation. The second stage freezes the shared VLM and trains each result decoder in its domain-native supervision space.

Table 2: Representative scientific evidence and validation logic used for natural-world knowledge alignment.

Domain	Representative evidence	Validation logic
Materials	Composition and stoichiometry; symmetry and dimensionality; local coordination and bonding	Test whether these cues support the claimed property; reject stoichiometric, symmetry, or structural violations.
Molecules	Functional groups and stereochemistry; charge distribution and electronic structure; intermolecular interactions	Connect structural cues to the predicted property; reject valence, functional-group, or geometric inconsistencies.
Drug discovery	Target interactions and pocket complementarity; structure–activity relations; ADMET mechanisms	Verify that the proposed mechanism agrees with the endpoint definition, label semantics, and relevant molecular interactions.
Proteins	Sequence conservation and motifs; residue environments; local geometry and three-dimensional contacts	Check whether site, interface, function, or structure assignments agree with residue indices and local contacts.
Spectra	Observed and missing peaks; functional-group assignments; cross-spectrum consistency	Require candidates to explain the major signals; use missing peaks to reject conflicting structures.
Images	Entities and target regions; spatial and topological relations; editing boundaries	Preserve entities and non-target content, maintain spatial relations, and restrict modifications to the requested region.

3.3.1 Shared VLM Training

The shared VLM takes x_i as input and uses $y_i^{\text{text}} = (s_i, r_i, a_i)$ as response-side supervision. We serialize its nonempty fields in their response order and optimize the standard autoregressive cross-entropy objective:

$$\mathcal{L}_{\text{VLM}}(\theta) = -\frac{1}{\sum_i |y_i^{\text{text}}|} \sum_i \sum_{k=1}^{|y_i^{\text{text}}|} \log p_\theta\left(y_{i,k}^{\text{text}} \mid c_i, y_{i,<k}^{\text{text}}\right). \tag{8}$$

The domain-native target y_i^{task} is not serialized into the VLM response and is reserved for task-decoder training in the second stage.

3.3.2 Task Result Decoder Training

After the first stage, the shared VLM parameters are fixed at $\theta^\star$. For an instance with $t_i \in \mathcal{T}$, the task condition is extracted from the frozen model's prefill representation:

$$e_i = \text{Extract}_{t_i}\left(F_{\theta^\star}^{\text{prefill}}(c_i)\right). \tag{9}$$

Each task decoder reads e_i and any required native scientific input x_i^{task}, and is supervised in its own result space:

$$\mathcal{L}_{\text{task}}\left(\{\phi_t\}_{t\in\mathcal{T}}\right) = \mathbb{E}_{\substack{d_i\sim\mathcal{D}\\ t_i\in\mathcal{T}}}\left[\text{loss}_{t_i}\left(G_{\phi_{t_i}}(e_i, x_i^{\text{task}}), y_i^{\text{task}}\right)\right], \tag{10}$$

$$\{\phi_t^\star\}_{t\in\mathcal{T}} = \arg\min_{\{\phi_t\}_{t\in\mathcal{T}}} \mathcal{L}_{\text{task}}, \qquad \theta^\star \text{ fixed}. \tag{11}$$

The loss loss_{t_i} is defined by the native result space. Continuous property tasks measure numerical deviation from the reference value. Classification and protein-site tasks measure prediction errors over class or residue labels. Spectrum-to-molecular generation and protein structure tasks use the structural or geometric supervision associated with their generation process. Scientific image generation and editing use the latent flow-matching objective of the generative model. Task decoders are trained independently and do not update through a shared cross-task weighted objective.

4 Evaluations

We evaluate S1-Omni as a unified multimodal reasoning model for scientific understanding, prediction, and generation across molecules, materials, proteins, spectra, and scientific images. The evaluation spans more than 60 benchmarks in five task families: property prediction, spectrum-to-molecular generation, residue-level protein functional-site prediction, all-atom protein structure prediction, and scientific image generation and editing. We ask whether unified representation of scientific data, natural-world knowledge

alignment, and decoding for domain-specific tasks can produce transferable scientific representations within a single model and translate them into verifiable domain-native outputs, including scalar values, functional sites, molecular structures, atomic coordinates, and images.

4.1 Property Prediction

Property prediction evaluates whether a model can infer continuous properties and classification labels from molecular or crystal structures. The task family spans materials, quantum chemistry, molecular physicochemistry, and ADMET.

4.1.1 Experimental Setup

The model predicts continuous properties or class probabilities from molecular SMILES strings or crystal CIF files. We evaluate 48 properties across materials, quantum chemistry, molecular physicochemistry, and ADMET to test whether one scientific model can maintain stable quantitative predictions across heterogeneous structural inputs and output types.

Setting. The evaluation combines 10 JARVIS-DFT material regression properties (Choudhary et al., 2020), two Materials Project classification properties (Jain et al., 2013), four COMPAS-1D quantum-chemical properties (Wahab et al., 2022), seven QM9 properties (Ramakrishnan et al., 2014), seven MoleculeNet tasks (Wu et al., 2018), and 18 ADMET endpoints from the Therapeutics Data Commons (Huang et al., 2021). Following each benchmark, regression is evaluated by MAD/MAE, MAE, RMSE, or Spearman correlation, and classification by ROC-AUC or PRC-AUC. MAE and RMSE are lower-is-better; all other metrics are higher-is-better.

Baselines. Molecular and quantum-chemical comparisons use ChemFM-3B (Cai et al., 2025) and Uni-Mol2 1.1B (Ji et al., 2024); material comparisons use LLM-Prop-35M (Niyongabo Rubungo et al., 2025) and MatBERT-109M (Trewartha et al., 2022). General-purpose baselines are GPT-5.5 (Singh et al., 2025) and Gemini-3.1-Pro (Gemini Team et al., 2024). Aggregate specialist counts use the stronger applicable specialist for each property. The distribution-shift analysis additionally compares GotenNet (Aykent & Xia, 2024) and Regression Transformer (Born & Manica, 2023) under BOOM (Antoniuk et al., 2025).

4.1.2 Main Results

Tables 3, 4, and 5 report results for molecular physicochemical and ADMET properties, quantum-chemical properties, and material properties, respectively. For each group, we compare S1-Omni with the applicable domain-specific models and with two general-purpose proprietary models.

Molecular Physicochemical and ADMET Prediction.

MoleculeNet evaluates basic physicochemical properties together with molecular activity and toxicity, whereas the ADMET tasks cover absorption, distribution, metabolism, excretion, and toxicity. Collectively, these endpoints define a central family of early-stage filters in drug discovery.

On MoleculeNet, S1-Omni exceeds ChemFM-3B on BBBP, reaching 0.8578 ROC-AUC, and on FreeSolv, reaching 0.7092 RMSE; ChemFM-3B leads on the other five tasks (Table 3). On TDC ADMET, S1-Omni surpasses the specialist result on `CYP2C9_SUBSTRATE_CARBONMANGELS`, `CYP2C9_VEITH`, `CYP2D6_VEITH`, `CYP3A4_SUBSTRATE_CARBONMANGELS`, `HERG`, and `HIA_HOU`. It also outperforms the stronger closed model on all seven MoleculeNet tasks and 16 of 18 ADMET endpoints; the exceptions are `HALF_LIFE_OBACH` and `VDSS_LOMBARDO`.

The strongest results occur on classification and ranking endpoints. BBBP, CYP, hERG, and HIA require the same molecular structure to be interpreted through distinct notions of permeability, metabolic recognition, transport, or structural alerts, and the role of a functional group can change with the endpoint. Conditioning the shared molecular representation on the task instruction helps separate these property semantics. The organization of metabolic soft spots, transport, and toxicity mechanisms in the structured reasoning data is also consistent with this result pattern. By contrast, ESOL, LIPO, clearance, and half-life expose numerical calibration and continuous ranking as the principal remaining sources of error.

Quantum Chemical Property Prediction.

COMPAS-1D and QM9 test whether a model can recover electron affinity, ionization potential, dipole moment, frontier-orbital energies, energy gaps, polarizability, and zero-point vibrational energy from molecular structure. These quantities underpin analyses of chemical reactivity and molecular design.

Table 3: Property prediction on MoleculeNet and TDC ADMET. Each score uses the metric shown in the same row; arrows indicate the preferred direction. Best and second-best results in each row are bold and underlined, respectively.

Dataset	Property	Metric	ChemFM-3B	GPT-5.5	Gemini-3.1-Pro	S1-Omni
MoleculeNet	BBBP	ROC-AUC ↑	0.7510	0.6773	0.6792	**0.8578**
MoleculeNet	CLINTOX	ROC-AUC ↑	**0.9180**	0.5557	0.4304	0.8063
MoleculeNet	HIV	ROC-AUC ↑	**0.8070**	0.6688	0.6209	0.6978
MoleculeNet	SIDER	ROC-AUC ↑	**0.7090**	0.5587	0.5052	0.5759
MoleculeNet	ESOL	RMSE ↓	**0.5160**	68.2500	19.4000	1.2217
MoleculeNet	FREESOLV	RMSE ↓	0.8300	29.2869	38.8370	**0.7092**
MoleculeNet	LIPO	RMSE ↓	**0.5450**	17.8300	980.1600	0.8845
ADMET	CACO2_WANG	MAE ↓	**0.3220**	7.0920	12.7883	0.4005
ADMET	LD50_ZHU	MAE ↓	**0.5410**	18.3553	8.7699	0.6625
ADMET	CYP2C9_SUBSTRATE_CARBONMANGELS	PRC-AUC ↑	0.4140	0.5430	0.5172	**0.6972**
ADMET	CYP2C9_VEITH	PRC-AUC ↑	0.7880	0.5223	0.5750	**0.8292**
ADMET	CYP2D6_SUBSTRATE_CARBONMANGELS	PRC-AUC ↑	**0.7390**	0.5109	0.5279	0.6651
ADMET	CYP2D6_VEITH	PRC-AUC ↑	0.7040	0.5397	0.6500	**0.8816**
ADMET	CYP3A4_VEITH	PRC-AUC ↑	**0.8780**	0.5350	0.5309	0.8410
ADMET	AMES	ROC-AUC ↑	**0.8540**	0.6711	0.5987	0.8340
ADMET	BIOAVAILABILITY_MA	ROC-AUC ↑	**0.7150**	0.5338	0.6742	0.6960
ADMET	CYP3A4_SUBSTRATE_CARBONMANGELS	ROC-AUC ↑	0.6540	0.4621	0.5359	**0.6654**
ADMET	DILI	ROC-AUC ↑	**0.9200**	0.5369	0.7090	0.8675
ADMET	HERG	ROC-AUC ↑	0.8480	0.5616	0.5447	**0.8734**
ADMET	HIA_HOU	ROC-AUC ↑	0.9840	0.5548	0.5961	**0.9921**
ADMET	PGP_BROCCATELLI	ROC-AUC ↑	**0.9310**	0.5665	0.6905	0.9224
ADMET	CLEARANCE_HEPATOCYTE_AZ	Spearman ↑	**0.4950**	-0.0016	-0.0367	0.4794
ADMET	CLEARANCE_MICROSOME_AZ	Spearman ↑	**0.6110**	0.1672	0.4418	0.5926
ADMET	HALF_LIFE_OBACH	Spearman ↑	0.5510	0.4342	**0.6465**	0.5209
ADMET	VDSS_LOMBARDO	Spearman ↑	**0.6620**	0.3538	0.5915	0.5861

Table 4: Property prediction on COMPAS-1D and QM9. Each score uses the metric shown in the same row; lower is better for all rows. Best and second-best results in each row are bold and underlined, respectively.

Dataset	Property	Metric	ChemFM-3B	Uni-Mol2 1.1B	GPT-5.5	Gemini-3.1-Pro	S1-Omni
COMPAS-1D	AEA_EV	MAE ↓	**0.0093**	–	3.1778	2.4314	0.0599
COMPAS-1D	AIP_EV	MAE ↓	**0.0074**	–	4.2155	1.7910	0.0508
COMPAS-1D	D3_DISP_CORR_EV	MAE ↓	**0.0067**	–	9.4398	9.1830	0.0699
COMPAS-1D	DIPMOM_DEBYE	MAE ↓	**0.0170**	–	1.0284	0.1701	0.0613
QM9	ALPHA	MAE ↓	–	**0.3050**	68.0723	44.8064	6.7366
QM9	CV	MAE ↓	–	**0.1440**	35.1990	7.7852	3.0004
QM9	GAP	MAE ↓	–	**0.0035**	4.6780	0.7438	0.0224
QM9	HOMO	MAE ↓	–	**0.0035**	1.9033	4.8781	0.0224
QM9	LUMO	MAE ↓	–	**0.0035**	1.5952	2.1027	0.0242
QM9	MU	MAE ↓	–	**0.0890**	1.8037	1.0133	0.8445
QM9	ZPVE	MAE ↓	–	**0.0005**	29.0415	124.3091	0.0244

S1-Omni outperforms both closed models on all 11 quantum-chemical properties, but the applicable specialists retain substantially lower error (Table 4). Its four COMPAS-1D MAEs range from 0.0508 to 0.0699, compared with 0.0067 to 0.0170 for ChemFM-3B. On QM9, the largest specialist margins occur for `ALPHA`, `CV`, and `ZPVE`. These quantities depend on different electronic and geometric factors and are sensitive to unit scale and small structural changes. The shared predictor stabilizes the output form, but three-dimensional pretraining gives specialist models a finer account of conformation and electronic environment.

Materials Property Prediction.

The JARVIS-DFT regression tasks predict electronic, mechanical, and optoelectronic properties from CIF structures for materials screening. The two Materials Project tasks classify whether a material has a direct band gap and whether it is stable, corresponding to screening for optoelectronic utility and thermodynamic synthesizability.

Table 5: Property prediction on JARVIS-DFT and the Materials Project. Each score uses the metric shown in the same row; arrows indicate the preferred direction. Best and second-best results in each row are bold and underlined, respectively.

Dataset	Property	Metric	LLM-Prop-35M	MatBERT-109M	GPT-5.5	Gemini-3.1-Pro	S1-Omni
JARVIS-DFT	BANDGAP_OPT	MAD/MAE ↑	<u>3.3310</u>	**5.4830**	1.4281	2.9381	1.5718
JARVIS-DFT	EPS_DFPT	MAD/MAE ↑	**1.5780**	1.5090	0.1953	0.5605	<u>1.5559</u>
JARVIS-DFT	EXF_EN	MAD/MAE ↑	<u>1.0440</u>	**1.3740**	0.214	0.2319	0.7264
JARVIS-DFT	FEPA_JARVIS	MAD/MAE ↑	**12.9960**	<u>10.2110</u>	1.1604	3.3473	1.5125
JARVIS-DFT	GV	MAD/MAE ↑	**2.6230**	<u>2.4190</u>	0.4622	1.3881	1.8545
JARVIS-DFT	KV	MAD/MAE ↑	<u>4.1280</u>	**4.2520**	1.9599	3.2534	2.9970
JARVIS-DFT	MAX_EFG	MAD/MAE ↑	<u>1.9360</u>	**2.1430**	0.5167	0.7287	1.5360
JARVIS-DFT	SLME	MAD/MAE ↑	<u>2.1750</u>	**2.2080**	0.8678	1.1965	1.6484
JARVIS-DFT	SPILLAGE	MAD/MAE ↑	**1.4440**	<u>0.9160</u>	0.3646	0.8256	0.7037
JARVIS-DFT	TOT_EN	MAD/MAE ↑	**23.5090**	<u>15.1410</u>	0.4565	0.6738	2.9861
Materials Project	IS_GAP_DIRECT	ROC-AUC ↑	<u>0.7000</u>	**0.7100**	0.505	0.5886	0.5772
Materials Project	IS_STABLE	ROC-AUC ↑	<u>0.7760</u>	**0.7900**	0.5163	0.6431	0.6231

At least one specialist model outperforms S1-Omni on each of the 12 material properties (Table 5). S1-Omni nevertheless exceeds the stronger closed model on `EPS_DFPT`, `EXF_EN`, `GV`, `MAX_EFG`, `SLME`, and `TOT_EN`. Its MAD/MAE of 1.5559 on `EPS_DFPT` approaches LLM-Prop-35M at 1.5780, whereas the two Materials Project classifications reach 0.5772 and 0.6231, below MatBERT-109M at 0.7100 and 0.7900. Band gaps, dielectric response, elastic moduli, formation energies, and stability depend on different combinations of lattice periodicity, local coordination, and electronic structure. The results show that the shared encoder forms a readable task representation from CIF inputs, but high-precision materials prediction still benefits substantially from specialist inductive biases over atomic neighborhoods, long-range periodic relations, and electronic structure.

Across all three groups, S1-Omni exceeds the reported specialist result on 8 of 48 properties and the stronger closed model on 40 of 48. The unified representation of scientific data supports both CIF and SMILES inputs and both regression and classification, while the remaining specialist advantage quantifies the value of object-specific structural priors.

4.1.3 Ablation Analysis

The ablations examine the structure of reasoning supervision, hidden-state pooling, explicit numerical ranges in the reasoning trace, predictor sharing, training stage and data scale, and label normalization.

Reasoning supervision is evaluated on six JARVIS-DFT properties. Pooling, predictor sharing, and training-stage experiments use a 12-task suite comprising six JARVIS-DFT and six MoleculeNet properties. Explicit range prediction uses the two chemical properties ESOL and LIPO, and label normalization uses all 48 properties.

Effect of Reasoning Supervision Structure. We fix the Stage-2 shared predictor, question-only pooling, training scale, and optimization, and vary only the Stage-1 reasoning supervision. Both reasoning datasets are distilled from Qwen3.7 and differ only in the system prompt, allowing us to isolate reasoning quality. Structured, property-constrained reasoning requires a systematic analysis of lattice structure, composition, local coordination, and symmetry, followed by selection of the three to five factors most relevant to the target property and an account of their direction, strength, and interactions. Unconstrained free-form reasoning provides only a broad request for analysis and leaves the evidence, order, and depth to the teacher model. The no-reasoning condition retains an empty reasoning span and the task-routing token.

Table 6: Effect of Stage-1 reasoning supervision on six material regression properties. All variants use the same shared predictor and question-only pooling; lower MAE and higher MAD/MAE are better. Best and second-best results are bold and underlined, respectively.

Stage-1 data	Material MAE ↓	Material MAD/MAE ↑
Structured, property-constrained reasoning	**17.255**	**1.260**
No reasoning text	<u>20.457</u>	<u>0.941</u>
Unconstrained free-form reasoning	34.818	0.888

Structured reasoning gives the lowest material MAE and the highest MAD/MAE (Table 6). Free-form reasoning reaches an MAE of 34.818, compared with 17.255 for structured reasoning and 20.457 without reasoning text. Adding arbitrary reasoning text therefore does not improve property prediction by itself.

The benefit of reasoning supervision depends on whether the scientific evidence is organized consistently around the target property. Structured supervision provides a stable learning signal; unconstrained traces vary in evidence coverage and analytical path and can be less effective than omitting reasoning altogether.

Effect of Hidden-State Pooling Scope. We next test which part of the sequence should supply the task-conditioned representation to the scalar predictor. On the 12-task suite, we fix the backbone, predictor, training data, and optimization and vary only the hidden-state span covered by mean pooling. Question-only pooling aggregates the CIF or SMILES input and its property instruction, whereas full pooling additionally includes the subsequently generated reasoning and answer states. The two variants use the same parameterization and training procedure, so their difference isolates the extraction scope.

Table 7: Effect of hidden-state pooling scope on six chemical and six material properties. Higher chemical F1 and material MAD/MAE are better; lower chemical RMSE is better. Best and second-best results are bold and underlined, respectively.

Pooling scope	Chemical classification F1 ↑	Chemical regression RMSE ↓	Material regression MAD/MAE ↑
Question only	**0.822**	**0.908**	**1.454**
Full	<u>0.773</u>	<u>1.663</u>	<u>0.899</u>

Question-only pooling improves every metric (Table 7): chemical F1 rises from 0.773 to 0.822, chemical RMSE falls from 1.663 to 0.908, and material MAD/MAE rises from 0.899 to 1.454. Answer-side states mix scientific evidence with wording, reasoning length, and termination patterns. The result indicates that property information can already be extracted before answer generation; reasoning supervision shapes the prefill representation rather than directly producing the numerical prediction.

Effect of Explicit Range Output in Reasoning. We test whether predicting a plausible numerical interval before the final value provides a useful intermediate target. The two settings retain the same reasoning process and differ only in whether the model must first state an approximate range. The motivation is that an explicit interval might narrow the search space and provide an intermediate constraint for the final estimate.

Table 8: Effect of a preliminary range prediction on mixed ESOL and LIPO regression. Lower RMSE is better. Best and second-best results are bold and underlined, respectively.

Reasoning supervision	Training records	RMSE ↓
Explicit preliminary value range	4,000	**0.2812**
No preliminary value range	4,000	<u>0.2814</u>

The hypothesis is not supported. The model does not reliably predict the target range: in some cases, the point estimate is close to the label even when the preliminary interval is substantially displaced. Final performance is also essentially unchanged, with RMSE values of 0.2812 with range prediction and 0.2814 without it (Table 8). Explicit range prediction is therefore not an effective intermediate reasoning signal. A more promising direction is supervision that helps decompose the problem and identify the evidence directly relevant to the target.

Effect of Predictor Sharing. We test whether sharing one predictor across all properties introduces interference between heterogeneous label spaces. On the 12-task suite, we fix the 140k training records, Stage 2, question-only pooling, SFT optimization, and evaluation metrics, and vary only the predictor organization. The shared variant uses one predictor module for all properties, whereas the property-specific variant assigns an independent predictor to each property. Both settings cover six chemical and six material properties.

The shared predictor is marginally better for chemical classification, with F1 values of 0.8741 and 0.8705, whereas property-specific heads improve both regression metrics (Table 9). Chemical RMSE falls from 0.8789 to 0.8250 and material MAD/MAE rises from 1.5553 to 2.0283. Shared parameters preserve the simplest cross-property interface; independent heads reduce interference among continuous targets with different units, ranges, and structural dependencies.

Table 9: Effect of shared versus property-specific predictors on six chemical and six material properties under the same 140k Stage-2 setting. Higher chemical F1 and material MAD/MAE are better; lower chemical RMSE is better. Best and second-best results are bold and underlined, respectively.

Predictor	Chemical classification F1 ↑	Chemical regression RMSE ↓	Material regression MAD/MAE ↑
Shared	**0.8741**	<u>0.8789</u>	<u>1.5553</u>
Property-specific	<u>0.8705</u>	**0.8250**	**2.0283**

Effect of Training Stage and Data Scale. To distinguish learning a terminal predictor from adapting the shared representation, and to measure how data scale affects each regime, we construct four SFT settings on the 12-task suite. Stage 2 freezes the backbone and trains only the property predictor; Stage 3 updates both. Each endpoint is evaluated with 27k and 140k training records under identical question-only pooling, predictor architectures, and metrics.

Table 10: Joint effects of property-training stage and data scale on six chemical and six material properties. Higher chemical F1 and material MAD/MAE are better; lower chemical RMSE is better. Best and second-best results are bold and underlined, respectively.

Training endpoint and data scale	Chemical classification F1 ↑	Chemical regression RMSE ↓	Material regression MAD/MAE ↑
Stage 2, 27k	0.822	0.908	1.454
Stage 2, 140k	**0.874**	0.879	<u>1.555</u>
Stage 3, 27k	0.862	<u>0.855</u>	1.525
Stage 3, 140k	<u>0.872</u>	**0.790**	**1.865**

Data scale and end-to-end adaptation are complementary (Table 10). Increasing Stage-2 data from 27k to 140k improves chemical F1 from 0.822 to 0.874, lowers chemical RMSE from 0.908 to 0.879, and raises material MAD/MAE from 1.454 to 1.555. The same expansion in Stage 3 improves chemical F1 from 0.862 to 0.872, lowers RMSE from 0.855 to 0.790, and raises material MAD/MAE from 1.525 to 1.865. At fixed data scale, Stage 3 improves all three metrics with 27k records; with 140k records, it substantially improves both regression metrics while leaving chemical F1 nearly unchanged. More data improve coverage of the mapping from a fixed representation to labels, whereas end-to-end property supervision further reshapes the shared representation for lightweight continuous prediction. The 140k Stage-3 setting therefore gives the best chemical and material regression results.

Effect of Per-Property Label Normalization. Across all 48 properties, we compare raw labels with independent normalization of each regression property followed by restoration to the original units at inference time; classification labels remain binary in both settings. Outcomes are grouped by whether they exceed, fall within 5% of, or remain more than 5% below the reported specialist.

Table 11: Effect of regression-label normalization across 48 properties with a shared predictor. Counts compare each property with its reported specialist result. Best and second-best results are bold and underlined, respectively.

Normalization strategy	Better than reported specialist (count) ↑	Within 5% of reported specialist (count) ↑	More than 5% worse (count) ↓	Better or within 5% (count) ↑
No normalization	**8**	<u>7</u>	<u>33</u>	<u>15</u>
Per-property normalization	<u>7</u>	**11**	**30**	**18**

Raw labels yield eight strict specialist wins, compared with seven after normalization (Table 11). Normalization increases the number within 5% from 7 to 11 and reduces the number more than 5% worse from 33 to 30, raising the combined competitive count from 15 to 18. Per-property normalization therefore improves balance across heterogeneous label scales, whereas raw labels attain a higher peak on more individual properties.

Despite its broader coverage, per-property normalization requires a separate normalization and inverse transformation for every property. In open-ended deployment, a user request may not be resolved reliably to a single property and its associated transform. To avoid this task-identification and parameter-matching dependency, the main model uses raw labels and returns predictions directly in their native units.

Effect of Distribution Shift.

Setting. We follow the BOOM full label-shift protocol (Antoniuk et al., 2025) on eight QM9 properties: α, C_v, gap, HOMO, LUMO, μ, R^2, and ZPVE. For each property, BOOM estimates the target density $P(y)$ by kernel density estimation, assigns the lowest-density regions of label space to the OOD test set, randomly draws an ID test set from the remaining samples, and uses the rest for training. After task-level filtering, we reconstruct mutually exclusive training, ID, and OOD splits by the same rule. Their actual sizes therefore depend on the filtered data rather than on a fixed fraction of the complete QM9 collection.

We report ID and OOD RMSE, the OOD/ID RMSE ratio, and binned R^2 within the lower and upper OOD tails. RMSE measures absolute prediction error, OOD/ID measures relative error amplification from ID to OOD, and binned R^2 measures predictive correlation in the two extremes of the OOD label distribution. Lower RMSE and OOD/ID are better; higher binned R^2 is better.

Baselines. We compare with GotenNet and Regression Transformer (RT) as reported by BOOM. GotenNet is an equivariant graph neural network over three-dimensional molecular structure. RT takes SMILES as input and decodes numerical values autoregressively. S1-Omni uses a unified encoder for CIF and SMILES and a shared linear regression head across molecular properties. All models are compared under the BOOM full label-shift protocol.

Table 12: ID and OOD RMSE on eight QM9 molecular properties under the BOOM full label-shift protocol. Each entry reports ID/OOD RMSE; lower is better. Best and second-best results are bold and underlined, respectively.

Model	α	C_v	Gap	HOMO	LUMO	μ	R^2	ZPVE
GotenNet	**0.550/1.830**	**0.130/0.340**	<u>0.090/0.280</u>	<u>0.052/0.170</u>	<u>0.020/0.033</u>	**0.320/1.500**	**5.070/17.070**	**0.001/0.001**
Regression Transformer	2.950/74345	1.280/2.810	0.210/0.550	0.130/0.319	0.190/0.860	0.870/2.350	21.350/25458	<u>0.008</u>/0.019
S1-Omni	<u>2.290/3.770</u>	<u>1.060/1.390</u>	**0.014/0.036**	**0.011/0.020**	**0.017/0.026**	<u>0.800/1.920</u>	<u>19.290/154.640</u>	<u>0.008/0.009</u>

Absolute error. Table 12 does not yield a single model ranking; instead, performance is strongly property dependent. GotenNet has lower ID and OOD RMSE on α, C_v, μ, R^2, and ZPVE, whereas S1-Omni is lowest on gap, HOMO, and LUMO. The split is consistent with their inductive biases: equivariant three-dimensional representations favor properties governed by molecular scale, spatial conformation, and three-dimensional charge distribution, while the unified structural representation is effective for the orbital-energy targets in this benchmark. Label shift therefore does not remove the importance of representation choice; extrapolation still depends on whether the representation contains the structural information required by the property.

RT exhibits a qualitatively different anomaly on α and R^2: RMSE grows from 2.95 and 21.35 in ID to 74,345 and 25,458 in OOD, respectively. This is not ordinary error growth but an orders-of-magnitude numerical explosion, indicating catastrophic failure in these OOD label regions. Because RT generates values autoregressively, the result suggests that discrete numerical decoding can produce extreme outliers when targets leave the main support of the training labels. Neither GotenNet nor the continuous regression head of S1-Omni exhibits numerical instability at this scale.

Relative OOD degradation. Figure 4 provides a different view from absolute RMSE. S1-Omni has the lowest OOD/ID ratio on α, C_v, HOMO, LUMO, μ, and ZPVE and is close to RT on gap. It therefore incurs less additional error from ID to OOD on most properties. This does not imply uniformly better absolute predictions: GotenNet has lower ID and OOD RMSE on α, C_v, μ, and ZPVE. OOD/ID measures relative degradation rather than error magnitude, and a large ID baseline can itself reduce the ratio. The ratio must therefore be interpreted jointly with absolute RMSE.

For RT, α and R^2 correspond to error amplifications of approximately 2.52×10^4 and 1.19×10^3. Ratios of this magnitude no longer describe gradual degradation under shift; they reflect direct numerical instability on specific tasks. The result identifies a second form of OOD risk beyond a representation that fails to cover tail structures: autoregressive numerical decoding can produce uncontrolled extreme outputs, and cross-task averages can conceal these catastrophic failures.

The R^2 property illustrates a further distinction. S1-Omni has an OOD/ID ratio of 8.02, substantially above GotenNet at 3.37, yet its binned R^2 remains positive. The principal failure is therefore not a complete loss of predictive ordering but a large increase in error scale within the OOD region. Conversely, several electronic-structure properties have smaller OOD/ID ratios and low absolute RMSE but do not preserve the relative order of tail samples. Relative error amplification captures only one aspect of distribution shift.

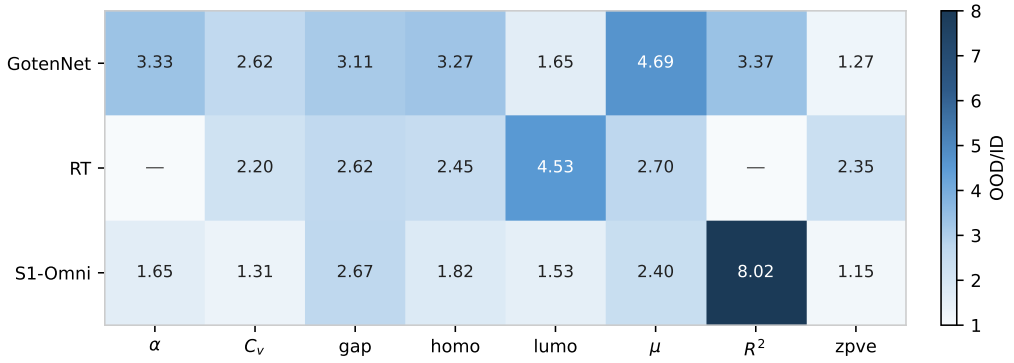


Figure 4: OOD-to-ID RMSE ratios for GotenNet, Regression Transformer (RT), and S1-Omni on eight QM9 properties under BOOM label shift. Values closer to 1 indicate less relative error growth; dashes denote catastrophic RT failures excluded from the ratio scale.

Tail correlation. Binned R^2 directly tests whether a model preserves variation within the extreme OOD regions (Table 13). GotenNet reaches 0.77, 0.99, 1.00, and 1.00 on α, C_v, R^2, and ZPVE, respectively, showing that its three-dimensional representation both reduces error and preserves ordering across the lower and upper tails. S1-Omni reaches 0.66, 0.74, 0.46, and 0.70 on these properties. Its unified representation retains transferable trend information for several physical and thermodynamic quantities, but less consistently than explicit three-dimensional modeling.

Table 13: Binned R^2 on the lower and upper OOD tails of eight QM9 molecular properties under BOOM label shift; higher is better. Best and second-best results are bold and underlined, respectively.

Model	α	C_v	Gap	HOMO	LUMO	μ	R^2	ZPVE
GotenNet	**0.77**	**0.99**	**0.03**	**0.15**	**0.33**	**0.11**	**1.00**	**1.00**
Regression Transformer	0.06	0.08	**0.03**	<u>0.02</u>	<u>0.02</u>	<u>0.09</u>	0.10	0.04
S1-Omni	<u>0.66</u>	<u>0.74</u>	<u>-3.04</u>	-1.53	-1.91	-0.57	<u>0.46</u>	<u>0.70</u>

The more consequential pattern appears for gap, HOMO, LUMO, and μ. S1-Omni has the lowest RMSE on the first three electronic-structure properties, yet binned R^2 is negative for all four. Low average error alongside negative tail correlation means that the model is accurate in densely populated label regions but does not reliably learn the direction of variation in low-density tails; extreme predictions contract toward the center of the training distribution. Neither ID accuracy nor aggregate OOD RMSE is therefore sufficient evidence of extrapolation. Preserving tail structure is the stricter test.

Per-property behavior. Figure 5 separates these outcomes into two failure modes. The first is regression to the mean: OOD samples occupy both ends of the label distribution, but the predicted range is compressed, so low values are overestimated and high values underestimated. This pattern is clearest for gap, HOMO, LUMO, and μ and is consistent with their low or negative binned R^2. The failure is not a uniform increase in error but an inability to recover local slope and ordering in the tails.

The second mode is a mismatch in error scale. For C_v and ZPVE, OOD/ID is 1.31 and 1.15, and both ID and OOD samples remain close to the diagonal, indicating that their structure–property relations extend comparatively stably into low-density regions. α shows some bilateral contraction but retains positive tail correlation. R^2 has the largest OOD/ID ratio, yet the scatter preserves a clear monotonic trend: the model captures the direction of change but underestimates the amplitude of high OOD values. Unlike the ordering failure for electronic-structure properties, this behavior is closer to a calibration or scale-extrapolation error.

Across all three metrics, the principal strengths of S1-Omni are smaller relative degradation on most properties and lower absolute error on the electronic-structure targets. Its limitation is not a single error shared by all OOD examples: different properties lose tail ordering or exhibit amplified error scale. The unified representation has learned property information that transfers across distributions, but it does not consistently encode the geometric and electronic factors controlling extreme values. Improving OOD generalization will require stronger supervision in low-density label regions and more direct three-dimensional constraints for geometry-sensitive properties, rather than further reductions in ID error alone.

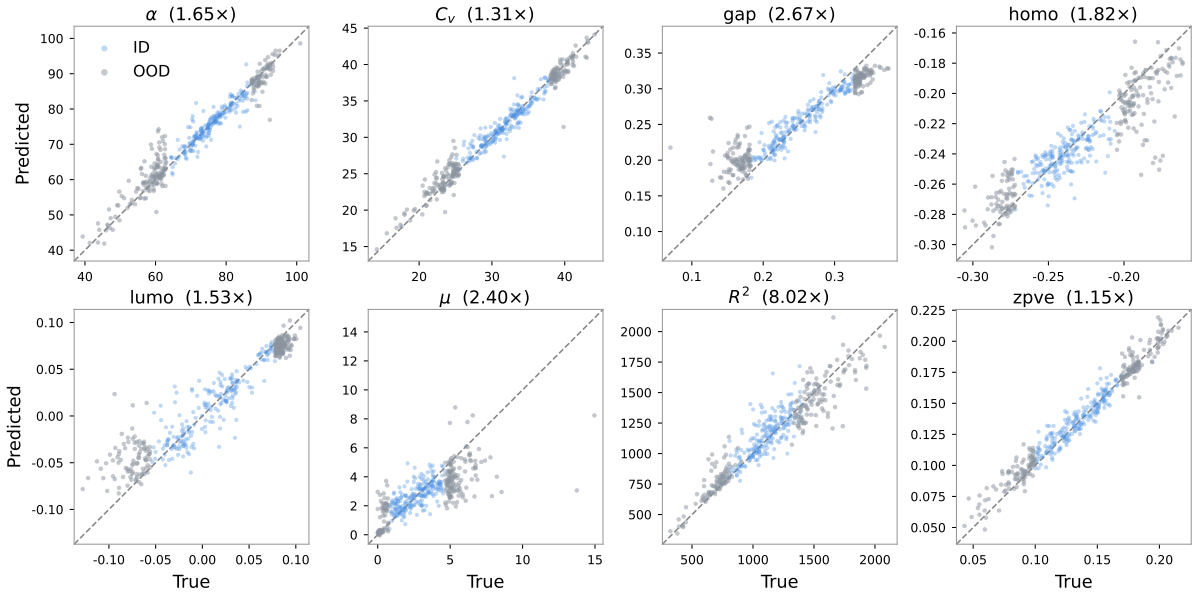


Figure 5: Predicted versus reference values for S1-Omni on ID and OOD splits of eight QM9 properties. Panel titles report the OOD/ID RMSE ratio; dashed lines show ideal agreement.

4.2 Spectrum-to-Molecular Generation

4.2.1 Experimental Setup

The Spectrum-to-molecular generation infers a molecular structure from IR, Raman, and UV-Vis spectra. The output includes both a two-dimensional molecular graph and a 3D conformation. IR and Raman encode vibrational information, while UV-Vis reflects electronic transitions. Together, these modalities constrain functional groups, molecular connectivity, and geometry.

Setting. We use QM9S (Zou et al., 2023), which provides 3D structures together with simulated IR, Raman, and UV-Vis spectra. Each UV-Vis spectrum has 601 values over 1.5 to 13.5 eV. The IR and Raman spectra each have 3,501 values over 500 to 4,000 cm^{-1}. Spectrum-to-molecular generation is evaluated with Acc@1, maximum common edge subgraph distance (MCES), Morgan and MACCS fingerprint similarity, FraggleSim, and functional-group similarity (FGSim). Lower MCES is better; all other generation metrics are higher-is-better. We also report AtomStable, MolStable, valid and complete (V&C), valid and unique (V&U), valid, unique, and novel (V&U&N), Frechet ChemNet Distance (FCD), similarity to nearest neighbor (SNN), fragment similarity (Frag), and scaffold similarity (Scaf). S1-Omni encodes normalized spectrum images as VLM representations. The <spectra_st> route token selects the molecular diffusion decoder, and a projector maps the VLM representations into its conditioning space before the decoder generates the molecular graph and conformation.

Baselines. DiffSpectra (Wang et al., 2025a) is a 2D and 3D molecular diffusion model directly conditioned on multiple spectra. JODO (Huang et al., 2023) and BioMatrix (Pei et al., 2026) provide reference values for unconditional molecular generation. We also evaluate GPT-5.5 from the GPT-5 family (Singh et al., 2025) and Gemini-3.1-Pro from the Gemini family (Gemini Team et al., 2024). Each model receives both images converted from the original JCAMP-DX spectra and the corresponding raw sequences, and is prompted to produce a molecular structure in standard 3D SDF format.

4.2.2 Molecular Structure Accuracy

Table 14 reports spectrum-to-molecular generation results. We first conduct a 100-sample pilot evaluation of GPT-5.5 and Gemini-3.1-Pro. For GPT-5.5, 71 of the 100 predictions form valid prediction-reference pairs after strict SDF parsing and molecular sanitization. The reported Acc@1 of 0.0563 is therefore computed conditionally over these valid pairs, corresponding to 4 exact matches out of 71. When invalid predictions are also counted as failures, its end-to-end Acc@1 over the full pilot set is 0.0400 (4/100). This distinction shows that GPT-5.5 is reasonably capable of producing parseable structures, but its exact structure accuracy remains low. Direct generation is limited in this setting: GPT-5.5 has low exact-match structure accuracy, while Gemini-3.1-Pro has substantially lower output validity and does not reliably return valid, parseable SDF structures. Because neither output validity nor end-to-end generation quality

Table 14: Spectrum-to-molecular generation from combined IR, Raman, and UV-Vis inputs on QM9S. Rows marked with ⋆ evaluate direct SDF generation by general-purpose models on 100 samples; the other rows evaluate spectrum-to-molecular generation models on 10,000 samples. Bold and underlined values denote the better and second-best results, respectively, between the matched spectrum-to-molecular generation models.

Model	Acc@1 ↑	MCES ↓	$\text{TaniSim}_{\text{MG}}$ ↑	$\text{CosSim}_{\text{MG}}$ ↑	$\text{TaniSim}_{\text{MA}}$ ↑	FraggleSim ↑	FGSim ↑
Gemini-3.1-Pro*	-	-	-	-	-	-	-
GPT-5.5*	0.0563	5.7817	0.2330	0.3492	0.3603	0.4152	0.5059
DiffSpectra	<u>0.4056</u>	<u>1.3273</u>	<u>0.7879</u>	<u>0.8456</u>	<u>0.9260</u>	**0.9503**	**0.9657**
S1-Omni	**0.4569**	**1.1765**	**0.8227**	**0.8698**	**0.9322**	<u>0.9436</u>	<u>0.9574</u>

Table 15: 2D chemical stability, validity, uniqueness, novelty, and distribution quality of molecules generated on QM9S. All stability, validity, uniqueness, and novelty values are proportions. The best and second-best model results in each metric column are bold and underlined, respectively; the training-set reference is excluded from ranking.

2D-Metric	AtomStable ↑	MolStable ↑	V&C ↑	V&U ↑	V&U&N ↑	FCD ↓	SNN ↑	Frag ↑	Scaf ↑
Train	0.9990	0.9880	0.9890	0.9890	0.0000	0.0630	0.4900	0.9920	0.9460
JODO	<u>0.9990</u>	**0.9880**	<u>0.9900</u>	0.9600	0.7800	0.1380	0.5220	0.9860	0.9340
BioMatrix-1.7B	**1.0000**	**0.9880**	**1.0000**	0.9480	0.7490	**0.0640**	0.4950	<u>0.9910</u>	<u>0.9440</u>
BioMatrix-4B	**1.0000**	**0.9880**	**1.0000**	0.9490	0.7420	<u>0.0660</u>	0.4940	<u>0.9910</u>	0.9420
DiffSpectra	0.9983	0.9804	0.9819	<u>0.9674</u>	<u>0.8998</u>	0.0792	**0.5314**	**0.9921**	**0.9448**
S1-Omni	0.9989	<u>0.9850</u>	0.9870	**0.9782**	**0.9105**	0.0920	<u>0.5283</u>	**0.9921**	0.9426

Table 16: 3D stability and distribution quality of molecules generated on QM9S. The best and second-best model results in each metric column are bold and underlined; the training-set reference is excluded from ranking.

3D-Metric	AtomStable ↑	Bond length ↓	Bond angle ↓	Dihedral angle ↓	FCD ↓
Train	0.9940	5.44E-4	4.65E-4	1.78E-4	0.8770
JODO	<u>0.9920</u>	**1.48E-1**	1.21E-2	6.29E-4	**0.8850**
BioMatrix-1.7B	0.8970	1.05E+0	3.16E-2	4.76E-4	1.0400
BioMatrix-4B	0.8970	1.05E+0	3.09E-2	6.09E-4	1.0120
DiffSpectra	**0.9922**	<u>1.64E-1</u>	<u>7.30E-3</u>	**2.45E-4**	<u>0.9218</u>
S1-Omni	0.9909	1.87E-1	**6.13E-3**	<u>3.54E-4</u>	0.9541

supports scaling the pilot, we do not extend these models to the full test set. DiffSpectra and S1-Omni are evaluated on 10,000 samples, so their comparison measures performance on a matched scale.

Relative to DiffSpectra, S1-Omni raises Acc@1 by 0.0513 and lowers MCES by 0.1508. Morgan Tanimoto, Morgan cosine, and MACCS Tanimoto increase by 0.0348, 0.0242, and 0.0062. Acc@1 requires an exact match of the complete two-dimensional graph, while MCES and fingerprint similarity measure graph and local-topology agreement for the remaining predictions. The consistent gains indicate that S1-Omni exactly reconstructs more targets and produces candidates with closer global connectivity and fingerprint representations.

FraggleSim and FGSim decrease by 0.0067 and 0.0083, leaving DiffSpectra ahead in local fragment and functional-group similarity. Exact graph recovery, fingerprint overlap, and local fragment coverage assess different forms of structural agreement. S1-Omni improves exact and global reconstruction, while DiffSpectra better preserves specific local chemical motifs. Gemini-3.1-Pro does not produce a valid 3D SDF file; dashes in its row denote direct-generation failure rather than measured zero scores.

4.2.3 Molecule Stability

Tables 15 and 16 evaluate S1-Omni in terms of 2D chemical quality and 3D geometric fidelity. The training-set statistics are provided only as references and are excluded from model ranking.

As shown in Table 15, S1-Omni achieves the best V&U and V&U&N scores among all evaluated models, reaching 0.9782 and 0.9105. Compared with the spectrum-to-molecule baseline DiffSpectra, S1-Omni improves MolStable from 0.9804 to 0.9850, V&C from 0.9819 to 0.9870, V&U from 0.9674 to 0.9782, and V&U&N from 0.8998 to 0.9105. It also improves AtomStable from 0.9983 to 0.9989. These consistent improvements demonstrate that S1-Omni generates a larger proportion of chemically stable molecules that are simultaneously valid, unique and novelty. In particular, its leading V&U&N result highlights its ability to avoid duplicated outputs and excessive overlap with the training set while preserving chemical validity.

Table 17: 20-label molecular-substructure probe using four IR-spectrum extraction strategies. Results use an independent 2,048-sample test set. Best and second-best results are bold and underlined, respectively.

Extraction strategy	Exact Match ↑	Micro-P ↑	Micro-F1 ↑	Macro-P ↑	Macro-F1 ↑
`IR-image_avg`	0.4833	0.8945	0.8725	<u>0.8188</u>	0.7658
`IR-image_end`	<u>0.5003</u>	<u>0.9014</u>	<u>0.8792</u>	0.8151	<u>0.7761</u>
`IR-image_max_min_avg`	0.4007	0.8988	0.8404	0.7779	0.6343
`IR-image_atten`	**0.7292**	**0.9441**	**0.9392**	**0.9035**	**0.8976**

S1-Omni also maintains competitive agreement with the reference distribution. It achieves the best Frag score of 0.9921, tied with DiffSpectra, and the second-best SNN score of 0.5283. Although its FCD and Scaf scores are slightly weaker than those of DiffSpectra (0.0920 vs 0.0792 and 0.9426 vs 0.9448), the differences in SNN, Frag, and Scaf remain small. Therefore, the gains in validity, uniqueness, and novelty do not result in a substantial departure from the reference chemical space. Compared with the unconditional baselines, S1-Omni improves V&U&N by 0.1305, 0.1615, and 0.1685 over JODO, BioMatrix-1.7B, and BioMatrix-4B.

The 3D results in Table 16 further demonstrate S1-Omni's ability to recover meaningful molecular geometry from spectral inputs. S1-Omni obtains the lowest bond-angle error among all evaluated models, achieving 6.13×10^{-3}, compared with 7.30×10^{-3} for DiffSpectra and 1.21×10^{-2} for JODO. This result indicates particularly accurate modeling of local angular geometry. S1-Omni also achieves an AtomStable score of 0.9909, remaining close to DiffSpectra 0.9922 and JODO 0.9920, while substantially outperforming both BioMatrix variants 0.8970. Its bond-length, dihedral-angle, and FCD results remain competitive, although DiffSpectra performs better on these metrics. These results support the effectiveness of S1-Omni for spectrum-conditioned molecular generation.

4.2.4 Effect of Hidden-State Extraction

To assess whether molecular structural information is encoded in the VLM representation and can be effectively decoded, we train a 20-label multilabel probe on QM9S to predict molecular functional groups. The development set is randomly divided into training and validation subsets at a 95:5, and an independent set of 2,048 samples is used for evaluation. The probe is optimized using `BCEWithLogitsLoss`. Predictions are obtained using a fixed threshold of 0.5. Exact Match requires all 20 functional-group labels of a molecule to be predicted correctly. Micro-averaged metrics aggregate decisions across all molecule-label pairs, whereas Macro-averaged metrics assign equal weight to each functional-group class.

We compare four strategies for extracting representations from the IR image-token feature sequence. `IR-image_avg` averages all image-token states, `IR-image_end` selects the final image-token state, and `IR-image_max_min_avg` concatenates the results of maximum, minimum, and average pooling. In contrast to these fixed aggregation strategies, `IR-image_atten` uses learnable attention to adaptively aggregate the image-token sequence. This comparison examines whether functional-group information is more effectively preserved by fixed global summarization or by adaptive aggregation that selectively integrates token-level features and their context.

As shown in Table 17, `IR-image_atten` achieves the best performance across all five reported metrics. Relative to `IR-image_end`, the strongest fixed aggregation strategy overall, `IR-image_atten` improves Exact Match by 22.89 percentage points, Micro-F1 by 6.00 points, and Macro-F1 by 12.15 points. In particular, the substantial improvement in Exact Match indicates that attention-based aggregation improves not only individual functional-group decisions but also the consistency of jointly predicting all 20 labels for a molecule.

Among the fixed aggregation strategies, `IR-image_end` performs best overall and slightly outperforms `IR-image_avg`, whereas `IR-image_max_min_avg` maintains relatively high Micro precision but shows substantial declines in Exact Match, Micro-F1, and Macro-F1. This result suggests that concatenating additional global statistics does not necessarily preserve more functional-group information. In particular, extreme-value pooling may overemphasize isolated token responses without adequately capturing relationships across the image-token sequence. Moreover, the improvement of `IR-image_atten` is larger in Macro-F1 than in Micro-F1, suggesting that its benefit is not driven solely by frequently occurring classes and extends to functional groups that are less well represented by fixed aggregation strategies.

These results indicate that functional-group evidence is distributed across the IR image-token sequence and can be decoded more effectively through adaptive attention-based aggregation than through fixed global summarization. By selectively integrating token-level features and their contextual relationships.

Table 18: Modality-specific results of the 20-label functional-group probe with sequence-attention feature extraction. Each row uses the listed spectrum modality as input. Best and second-best results are bold and underlined.

Modality	Exact Match ↑	Micro-P ↑	Micro-F1 ↑	Macro-P ↑	Macro-F1 ↑
Raman	<u>64.24%</u>	<u>92.09%</u>	<u>92.02%</u>	<u>87.99%</u>	<u>87.05%</u>
UV	21.80%	76.76%	74.94%	69.67%	66.52%
IR	**72.92%**	**94.41%**	**93.92%**	**90.35%**	**89.76%**

Table 19: Residue-level PPI-site prediction on the MPBind test set. AUPR is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	F1 ↑	MCC ↑	AUROC ↑	AUPR ↑
MPBind	-	-	<u>0.830</u>	<u>0.540</u>
PeSTo	-	-	0.760	0.380
GPT-5.5	0.284	<u>0.088</u>	-	-
Gemini-3.1-Pro	<u>0.353</u>	0.036	-	-
S1-Omni	**0.632**	**0.513**	**0.859**	**0.666**

This provides evidence that the VLM representation encodes decodable molecular structural information, while also showing that the choice of representation extraction strategy substantially affects how effectively that information can be recovered.

To compare the evidence provided by each spectral modality, we fix the sequence-attention extraction strategy and train the same 20-label QM9S probe separately on IR, Raman, and UV-Vis inputs. Table 18 reports Exact Match together with micro and macro precision and F1.

IR attenuation achieves the highest scores across all five evaluation metrics, followed by Raman attenuation, with corresponding Micro-F1 values of 0.9392 and 0.9202 (Table 18). By comparison, UV-Vis attenuation yields a Micro-F1 of 0.7494 and an Exact Match of only 0.2180, lagging noticeably behind the two vibrational spectroscopy modalities. The performance gap arises because the vibrational features captured by IR and Raman align more directly with the functional-group labels used in this task, while UV-Vis spectra predominantly encode electronic transition information. The consistent performance ranking of modalities across all metrics demonstrates that the proposed unified representation retains the unique scientific characteristics of each spectral type, rather than compressing them into undifferentiated generic spectral embeddings.

4.3 Protein Site Prediction

Protein functional-site prediction identifies residues that mediate specific biological functions, including protein–protein, DNA, RNA, metal-ion, small-molecule, antibody, and antigen interactions. Accurate site prediction supports mechanistic interpretation, drug and antibody design, enzyme engineering, and disease research. We follow the metrics and comparison protocols of each public benchmark; because data splits and label definitions differ, scores are comparable only within the same subtask. Precision is the fraction of predicted sites that are correct, recall is the fraction of true sites recovered, F1 balances precision and recall, and MCC measures binary classification under class imbalance. AUROC and AUPR assess residue ranking across thresholds. All metrics are higher-is-better. Within a comparable table column, bold and underlined values denote the best and second-best results, respectively.

4.3.1 Protein-Protein Interaction Site Prediction

Protein–protein interaction (PPI) site prediction identifies residues on a target protein that directly contact another protein, yielding a residue-level description of a protein-complex interface.

Setting. We evaluate on the MPBind test set (Wang et al., 2025c). AUPR is the primary metric, AUROC provides a complementary threshold-independent ranking measure, and F1 and MCC assess residue classification at a fixed threshold. S1-Omni jointly encodes the target sequence and the PPI task description. Cross-attention allows every residue to retrieve context dynamically according to the task semantics before producing a site probability. The mean evaluated sequence length is 169 residues.

Baselines. Task-specific baselines are MPBind (Wang et al., 2025c) and PeSTo (Krapp et al., 2023). GPT-5.5 and Gemini-3.1-Pro directly generate residue positions and are included only in the F1 and MCC comparisons.

Table 20: Residue-level epitope prediction on the RoBep test set. AUPR is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	F1 ↑	MCC ↑	AUPR ↑
RoBep	<u>0.379</u>	<u>0.340</u>	<u>0.268</u>
GraphBepi	0.276	0.189	0.215
SEMA 2.0	0.284	0.196	0.236
DiscoTope 3.0	0.300	0.234	0.237
CALIBER	0.241	0.164	0.158
SEPPA 3.0	0.194	0.194	0.194
GPT-5.5	0.209	0.076	-
Gemini-3.1-Pro	0.188	0.030	-
S1-Omni	**0.546**	**0.493**	**0.472**

The main gain appears in AUPR, which is especially sensitive to sparse positive residues. S1-Omni reaches 0.666 AUPR, 0.126 above MPBind, while AUROC rises from 0.830 to 0.859 (Table 19). The concordant improvements show that the gain extends across both global positive–negative ranking and the precision–recall regime. F1 and MCC also exceed both general-purpose models, indicating that residue-level output remains effective after thresholding.

4.3.2 Epitope Prediction

Epitope prediction locates antigen-surface residues recognized and bound by an antibody, reversing the direction of paratope prediction on the antibody side.

Setting. We use the RoBep test set (Xu et al., 2026). AUPR is the primary ranking metric, while F1 and MCC evaluate residue classification at a fixed threshold. S1-Omni forms a task-conditioned representation from the antigen sequence and epitope semantics, and a residue-level head predicts binding probabilities on the antigen. The mean evaluated sequence length is 253 residues.

Baselines. Specialized methods include RoBep (Xu et al., 2026), GraphBepi (Zeng et al., 2023), SEMA 2.0 (Ivanisenko et al., 2024), DiscoTope 3.0 (Høie et al., 2024), CALIBER (Israeli & Louzoun, 2024), and SEPPA 3.0 (Zhou et al., 2019). We additionally compare with GPT-5.5 and Gemini-3.1-Pro.

S1-Omni obtains 0.472 AUPR, 0.204 above RoBep, with corresponding gains of 0.167 in F1 and 0.153 in MCC (Table 20). Improvements in both ranking and threshold-based metrics show that the result is not tied to a particular operating point. In contrast to the PPI comparison, this benchmark supports both types of metric for the specialist models and thus more directly demonstrates that the same residue-level interface can be redirected toward antigen-side labels. Epitope recognition depends not only on local sequence patterns but also on immunological knowledge and global functional context. The biological knowledge encoded by S1-Omni helps it interpret protein functions and interaction semantics beyond what is available to conventional sequence-only baselines.

4.3.3 Small-Molecule-Binding Site Prediction

Small-molecule-binding site prediction identifies protein residues that contact drugs, metabolites, or other small-molecule ligands.

Setting. We evaluate on COACH420 as curated by CLAPE-SMB (Wang et al., 2024a). AUROC is the primary ranking metric; precision, recall, F1, and MCC characterize residue classification at a fixed threshold. S1-Omni places the protein sequence and small-molecule ligand context in the same task condition and decodes a binding probability for each residue. The mean evaluated sequence length is 284 residues.

Baselines. The specialized baselines are P2Rank (Krivák & Hoksza, 2018) and GraphBind (Xia et al., 2021), together with GPT-5.5 and Gemini-3.1-Pro as general-purpose references.

S1-Omni reaches 0.880 AUROC, 0.010 above GraphBind, and improves MCC from 0.303 to 0.366 (Table 21). Relative to GraphBind, precision and recall rise jointly from 0.223 and 0.477 to 0.275 and 0.600, so the improvement is present in both ranking and threshold statistics. P2Rank attains the highest recall, 0.888, but with a precision of 0.079, reflecting a high-recall, low-precision operating regime.

Table 21: Residue-level small-molecule-binding site prediction on COACH420. AUROC is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	Precision ↑	Recall ↑	F1 ↑	MCC ↑	AUROC ↑
P2Rank	0.079	**0.888**	-	0.224	-
GraphBind	<u>0.223</u>	0.477	-	<u>0.303</u>	<u>0.870</u>
GPT-5.5	0.096	0.330	<u>0.148</u>	0.101	-
Gemini-3.1-Pro	0.077	0.515	0.134	0.097	-
S1-Omni	**0.275**	<u>0.600</u>	**0.377**	**0.366**	**0.880**

Table 22: Metal-binding site prediction on LABind DS1. MCC is reported by ion and higher is better; average MCC summarizes cross-ion performance. Best and second-best results are bold and underlined, respectively.

Method	Ca ↑	Mg ↑	Mn ↑	Zn ↑	Fe(III) ↑	Fe(II) ↑	Cu ↑	Na ↑	K ↑	Average MCC ↑
LigBind	<u>0.465</u>	<u>0.354</u>	<u>0.608</u>	<u>0.750</u>	<u>0.679</u>	0.678	0.552	0.166	<u>0.232</u>	<u>0.498</u>
LABind	**0.664**	**0.461**	**0.710**	**0.843**	**0.768**	**0.848**	**0.771**	**0.387**	**0.444**	**0.655**
GPT-5.5	0.320	0.156	0.045	0.488	0.129	0.231	0.386	0.033	-0.015	0.197
Gemini-3.1-Pro	0.313	0.183	0.255	0.508	-0.010	0.507	0.514	0.001	-0.008	0.251
S1-Omni	0.351	0.273	0.591	0.652	0.627	<u>0.712</u>	<u>0.606</u>	<u>0.167</u>	0.224	0.467

4.3.4 Metal-Binding Site Prediction

Metal-binding site prediction identifies residues that coordinate a specified metal ion. Because coordination environments differ across ions, we report MCC for each ion and average MCC across ions.

Setting. We use LABind DS1 (Zhang et al., 2025b), which covers Ca, Mg, Mn, Zn, Fe(III), Fe(II), Cu, Na, and K. Each ion-specific column reports MCC for the corresponding residue-level binary task. S1-Omni jointly conditions on the protein sequence, ion identity, and task description and produces an ion-specific classification through the shared residue interface. The mean evaluated sequence length is 391 residues.

Baselines. Specialized methods are LigBind (Xia et al., 2023) and LABind (Zhang et al., 2025b); the general-purpose references are GPT-5.5 and Gemini-3.1-Pro.

S1-Omni reaches an average MCC of 0.467, 0.188 below LABind at 0.655, but above GPT-5.5 and Gemini-3.1-Pro (Table 22). It exceeds LigBind on Fe(II), Cu, and Na and is close on Mn and K, whereas LABind leads for all nine ions. These ion-dependent differences show that the shared residue interface captures part of the conditioned coordination pattern. Specialist models, however, explicitly incorporate metal-specific chemical priors and structural features, whereas S1-Omni relies primarily on sequence and task semantics and therefore remains less effective at resolving fine-grained coordination rules.

4.3.5 Paratope Prediction

Paratope prediction locates residues in antibody variable regions that directly contact an antigen, defining the antibody side of the interface.

Setting. We evaluate on the MIPE test set (Wang et al., 2024b). F1 is the primary metric; precision and recall expose the fixed-threshold tradeoff, MCC complements them under class imbalance, and AUROC and AUPR measure ranking across thresholds. S1-Omni jointly encodes the antibody sequence and paratope semantics, allowing the shared prediction head to emit antibody-side binding residues according to the requested task direction. The mean evaluated sequence length is 77 residues.

Baselines. Specialized methods are MIPE (Wang et al., 2024b), ParaSurf (Papadopoulos et al., 2025), and VASCIF (Liu et al., 2026); we additionally compare with GPT-5.5 and Gemini-3.1-Pro.

S1-Omni attains 0.456 precision and 0.649 recall, yielding 0.536 F1, whereas ParaSurf reaches 0.690 F1 with higher precision and recall (Table 23). Its AUROC and AUPR of 0.757 and 0.479 also trail MIPE and ParaSurf. Antibody–antigen recognition depends strongly on three-dimensional structure, conformational change, and spatial complementarity. Because S1-Omni primarily uses sequence and language semantics without explicit antibody structure, it does not reach the specialist results. It nevertheless exceeds GPT-5.5 and Gemini-3.1-Pro in F1 by 0.129 and 0.070, respectively, making it the strongest general-purpose model in this comparison.

Table 23: Residue-level paratope prediction on the MIPE test set. F1 is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	Precision ↑	Recall ↑	F1 ↑	MCC ↑	AUROC ↑	AUPR ↑
MIPE	-	-	<u>0.627</u>	<u>0.554</u>	0.927	0.741
ParaSurf	**0.576**	**0.860**	**0.690**	**0.659**	<u>0.967</u>	**0.781**
VASCIF	-	-	-	-	**0.981**	<u>0.742</u>
GPT-5.5	0.302	0.625	0.407	0.138	-	-
Gemini-3.1-Pro	0.339	<u>0.745</u>	0.466	0.240	-	-
S1-Omni	<u>0.456</u>	0.649	0.536	0.361	0.757	0.479

Table 24: Residue-level DNA-binding site prediction on DNATest-129. AUROC is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	Precision ↑	Recall ↑	F1 ↑	MCC ↑	AUROC ↑
MegSite	0.484	**0.739**	**0.585**	**0.567**	**0.952**
USPDB	**0.532**	0.612	0.551	0.527	<u>0.948</u>
HGCTBind	0.514	<u>0.654</u>	<u>0.574</u>	<u>0.549</u>	0.945
IPDLPre	<u>0.531</u>	0.514	0.522	0.522	0.914
GPT-5.5	0.138	0.427	0.194	0.132	-
Gemini-3.1-Pro	0.090	0.495	0.160	0.099	-
S1-Omni	0.384	0.593	0.466	0.436	0.890

4.3.6 DNA-Binding Site Prediction

DNA-binding site prediction identifies protein residues that contact DNA, characterizing nucleic-acid interfaces involved in transcriptional regulation, replication, and repair.

Setting. We use DNATest-129 from USPDB (Mi et al., 2025). AUROC is the primary ranking metric; precision, recall, F1, and MCC describe fixed-threshold residue classification. S1-Omni uses the protein sequence and DNA-binding task semantics to produce per-residue probabilities, redirecting the shared interface toward nucleic-acid-binding labels. The mean evaluated sequence length is 291 residues.

Baselines. Specialized methods include MegSite (Hu et al., 2025), USPDB (Mi et al., 2025), HGCTBind (Gong et al., 2026), and IPDLPre (Feng et al., 2025), alongside GPT-5.5 and Gemini-3.1-Pro.

S1-Omni reaches 0.890 AUROC, below all four specialists and 0.062 below the best result from MegSite (Table 24). At the fixed threshold, its precision, recall, and F1 are 0.384, 0.593, and 0.466, respectively, giving a comparatively balanced precision–recall profile. Recall exceeds IPDLPre, but F1 and MCC remain below every specialist. Ranking and threshold metrics thus give the same overall conclusion: specialized models are stronger on DNA-binding sites. S1-Omni nonetheless substantially exceeds both general-purpose baselines on precision, recall, F1, and MCC, demonstrating the benefit of directed residue-level decoding. The remaining gap is consistent with the joint dependence of DNA interfaces on sequence conservation, structural environment, and electrostatics, none of which is explicitly imposed as a task-specific prior.

4.3.7 RNA-Binding Site Prediction

RNA-binding site prediction identifies protein residues that contact RNA, including interfaces involved in RNA processing, transport, and translational regulation.

Setting. We evaluate on RNATest-117 from MegSite (Hu et al., 2025). AUROC is the primary ranking metric; precision, recall, F1, and MCC characterize classification at a fixed threshold. S1-Omni fuses the protein sequence with RNA-binding task semantics to form task-conditioned residue representations, from which the shared prediction head outputs RNA-binding probabilities. The mean evaluated sequence length is 319 residues.

Baselines. The specialized methods are SPLiNet (Ye et al., 2026) and MegSite (Hu et al., 2025); GPT-5.5 and Gemini-3.1-Pro serve as general-purpose references.

S1-Omni obtains 0.807 AUROC, 0.092 below MegSite, and trails both specialists in precision, recall, F1, and MCC (Table 25). Compared with the general-purpose models, S1-Omni has higher precision, F1, and MCC. Gemini-3.1-Pro has slightly higher recall, 0.445 versus 0.419, but only 0.070 precision and 0.120 F1, indicating that its recall is driven by many false positives. Taken together, the metal-, paratope-, DNA-,

Table 25: Residue-level RNA-binding site prediction on RNATest-117. AUROC is the primary metric; higher values are better for all metrics. Best and second-best results are bold and underlined, respectively.

Method	Precision ↑	Recall ↑	F1 ↑	MCC ↑	AUROC ↑
SPLiNet	**0.398**	**0.859**	**0.544**	**0.516**	<u>0.842</u>
MegSite	<u>0.356</u>	<u>0.571</u>	<u>0.439</u>	<u>0.411</u>	**0.899**
GPT-5.5	0.110	0.247	0.152	0.090	-
Gemini-3.1-Pro	0.070	0.445	0.120	0.048	-
S1-Omni	0.277	0.419	0.333	0.294	0.807

Table 26: Architectural progression for paratope prediction on the MIPE test set. The columns specify the prediction target, protein representation, task-conditioning mechanism, and treatment of class imbalance. Best and second-best F1 values are bold and underlined, respectively.

Method	Prediction target	Protein representation	Task conditioning	Imbalance handling	F1 ↑
VLM + SFT	Generated residue-index text	Autoregressive VLM output	Prompt context	None	≈ 0
VLM + linear head	Sequence-level classification	Final-token VLM hidden state	Prompt context	None	≈ 0
Pure VLM + Residue Prediction	Residue-level classification	Per-residue VLM hidden states	Context encoded within the VLM	Weighted BCE	<u>≈ 0.30</u>
S1-Omni	Residue-level classification	ESM2 residue embeddings	Cross-attention to VLM task context	Weighted BCE	**0.536**

and RNA-binding results show that directed residue decoding substantially improves general-purpose modeling, but a clear specialist gap remains on tasks governed by three-dimensional structure and biophysical priors. Sequence and semantic representations alone do not fully encode the required spatial constraints and interactions.

4.3.8 Effect of Residue-Level Decoding and Protein-Task Fusion

Setting. We isolate the effects of prediction granularity and protein–task fusion on MIPE paratope prediction using four progressively stronger designs. The first fine-tunes the VLM and directly generates residue indices. The second reads the final-token hidden state and applies a linear classifier. The third preserves per-residue VLM hidden states and trains a residue-level prediction head with weighted BCE. The full S1-Omni instead uses ESM2 residue embeddings, applies cross-attention to the VLM-encoded task context, and uses the same weighted BCE objective. All variants are evaluated by F1.

Generated residue-index text and final-token classification both yield near-zero F1 (Table 26). Free-form text and a single global state fail to preserve the positional correspondence required for residue labels, confirming that protein-site prediction is a fine-grained localization problem rather than conventional knowledge-based question answering. Per-residue VLM classification raises F1 to approximately 0.30: maintaining a distinct representation at each sequence position converts the task semantics into a trainable dense prediction problem, although the resulting site information remains incomplete. Adding ESM2 residue representations and task-context cross-attention increases F1 to 0.536, a further gain of approximately 0.236. ESM2 preserves residue-level relations in the protein sequence, while cross-attention lets those representations retrieve context relevant to the current task.

The progression supports three conclusions. First, text generation alone cannot deliver precise residue-level localization. Second, protein-site prediction is inherently a dense residue-level task, and per-residue prediction is substantially more effective than sequence-level classification. Third, separating protein representation learning from task interpretation and coupling them through cross-attention is central to a unified multi-task protein-site predictor.

4.4 Protein Structure Prediction

4.4.1 Experimental Setup

Setting. We follow the SimpleFold evaluation protocol on CAMEO22 (Haas et al., 2018; Wang et al., 2025d). TM-score (Zhang & Skolnick, 2004) and GDT-TS measure global fold and topological agreement. lDDT (Mariani et al., 2013) and lDDT-Ca evaluate all-atom and backbone local geometry, respectively, while RMSD measures coordinate deviation after structural alignment. We report both the mean and median of every metric to distinguish aggregate behavior from structural quality on a typical target.

Table 27: Protein structure prediction on CAMEO22. Each metric is reported as the mean / median across test targets, with the two statistics ranked separately. Best and second-best results are bold and underlined, respectively.

Model family	Method	TM-score (mean / median) ↑	GDT-TS (mean / median) ↑	lDDT (mean / median) ↑	lDDT-Ca (mean / median) ↑	RMSD (mean / median) ↓
MSA-based	RoseTTAFold	0.7800 / 0.8600	0.7150 / 0.7750	0.5750 / 0.6050	0.7980 / 0.8270	5.7210 / 2.8640
	AlphaFlow	0.8400 / 0.9270	0.8080 / 0.8530	0.7410 / 0.7980	0.8550 / 0.8930	3.8460 / 2.1220
	AlphaFold2	<u>0.8630</u> / <u>0.9420</u>	<u>0.8440</u> / <u>0.9030</u>	**0.8160 / 0.8560**	**0.8930** / <u>0.9230</u>	<u>3.5780</u> / <u>1.8570</u>
	RoseTTAFold2	**0.8640 / 0.9470**	**0.8450 / 0.9040**	0.7270 / 0.7670	**0.8930 / 0.9260**	**3.5710 / 1.7070**
PLM-based	ESM3	0.7460 / 0.8400	0.6940 / 0.7580	-	-	-
	ESMDiff	0.7540 / 0.8470	0.7010 / 0.7600	-	-	-
	EigenFold	0.7500 / 0.8400	0.7100 / 0.7900	-	-	-
	OmegaFold	0.8050 / 0.8990	0.7670 / 0.8440	0.7460 / 0.8150	0.8290 / 0.8920	5.2940 / 2.6220
	ESMFlow	0.8180 / 0.8930	0.7740 / 0.8320	0.6960 / 0.7450	0.8270 / 0.8670	4.5280 / 2.6930
	ESMFold	0.8530 / 0.9330	0.8260 / 0.8750	<u>0.7920</u> / <u>0.8340</u>	<u>0.8710</u> / 0.9060	3.9730 / 2.0190
General-purpose LLM	Gemini-3.1-Pro	0.3175 / 0.2490	-	-	-	-
	GPT-5.5	0.1118 / 0.0820	-	-	-	-
Internal baseline	SimpleFold-700M, no condition	0.8288 / 0.9140	0.7850 / 0.8470	0.7724 / 0.8070	0.8480 / 0.8830	4.5571 / 2.6300
Ours	S1-Omni	0.8291 / 0.9090	0.7848 / 0.8430	0.7709 / 0.8040	0.8461 / 0.8800	4.5913 / 2.5040

Higher values are better except for RMSD. S1-Omni uses S1-VL to jointly encode the protein sequence and task instruction, with reasoning supervision producing a reasoning-conditioned hidden representation. A lightweight projector maps this representation into the structural conditioning space, after which a fixed-capacity SimpleFold-700M decoder generates all-atom coordinates. This design connects scientific reasoning representations to decoding for domain-specific tasks while keeping the structural backbone and evaluation protocol fixed.

Protein structure prediction reconstructs three-dimensional atomic coordinates from an amino acid sequence and must jointly resolve local residue geometry, long-range contacts, and global fold topology. We study a controlled question: with the geometric decoder fixed, can a task representation shaped by scientific reasoning improve structure prediction? We select SimpleFold (Wang et al., 2025d) because its streamlined architecture, explicit conditioning interface, and family of larger model variants make it suitable for isolating the effect of reasoning integration. All S1-Omni experiments and controlled variants use SimpleFold-700M, so differences primarily reflect the interaction between the reasoning condition and structural decoding.

Baselines. The central controlled baseline is SimpleFold-700M without external conditioning under the same evaluation protocol, which directly measures the effect of reasoning conditioning. MSA-based methods include RoseTTAFold (Baek et al., 2021), AlphaFlow (Jing et al., 2024), AlphaFold2 (Jumper et al., 2021), and RoseTTAFold2 (Baek et al., 2023). Single-sequence or protein-language-model methods include ESM3 (Hayes et al., 2025), ESMDiff (Lu et al., 2025), EigenFold (Jing et al., 2023), OmegaFold (Wu et al., 2022), ESMFlow (Jing et al., 2024), and ESMFold (Lin et al., 2023). Together, these methods establish the performance range of major structure-prediction paradigms on CAMEO22. Gemini-3.1-Pro and GPT-5.5 are directly prompted general-purpose models and report only TM-score.

The central comparison in Table 27 is between unconditioned SimpleFold-700M and S1-Omni with a reasoning-conditioned hidden representation. Because the geometric backbone is fixed, their difference directly measures the effect of the reasoning condition on global topology, local geometry, and coordinate alignment. The other methods provide broader reference points for structure-prediction performance.

4.4.2 Main Results

SimpleFold provides a scalable backbone for reasoning integration. Results from AlphaFold2, RoseTTAFold2, and ESMFold show that greater capacity, MSAs, pair representations, and finer-grained geometric modeling continue to raise the ceiling of structural accuracy. We use SimpleFold-700M to isolate reasoning conditioning in a fixed, structurally transparent backbone. The low TM-scores of directly prompted general-purpose models further show that three-dimensional coordinates require a specialized decoder that enforces domain-specific geometric constraints. The present 700M experiment establishes the compatibility of reasoning representations with structure generation and provides a direct path to test the same interaction with larger SimpleFold variants when computational resources permit.

Reasoning conditioning produces metric-specific structural gains. Relative to unconditioned SimpleFold-700M, S1-Omni raises mean TM-score from 0.828760 to 0.829126, an increase of 0.000366, and

Table 28: Ablation of the conditioning source, hidden-state extraction, and trainable-module scope on CAMEO22. All metrics are averaged across test targets. Best and second-best results are bold and underlined, respectively.

Comparison axis	Variant	TM-score ↑	GDT-TS ↑	lDDT ↑	lDDT-Ca ↑	RMSD ↓
Reference setting	SimpleFold-700M without external condition	0.8288	0.7850	<u>0.7724</u>	<u>0.8480</u>	<u>4.5571</u>
Ground-truth secondary structure	Oracle DSSP eight-state one-hot condition	**0.8318**	**0.7898**	0.6782	**0.8512**	**4.4885**
Predicted secondary structure	Predicted SS8 hard-label condition	0.8252	0.7811	0.7700	0.8457	4.5998
Hidden-state extraction	S1-VL, last prefill hidden state	<u>0.8298</u>	0.7852	**0.7725**	0.8469	4.5799
Hidden-state extraction	S1-VL, end-of-message hidden state	0.8294	<u>0.7855</u>	0.7716	0.8471	4.5671
Trainable-module scope	Jointly train projector and diffusion module	0.8279	0.7829	0.7658	0.8442	4.6428

lowers median RMSD from 2.630 to 2.504, an improvement of 0.126. Changes in mean GDT-TS, lDDT, and lDDT-Ca are each below 0.002, while mean RMSD rises from 4.557120 to 4.591322. The gains are therefore concentrated in global topology and coordinate alignment on the median target; local geometry and mean RMSD remain near the baseline. Overall, S1-Omni preserves the principal folding ability of SimpleFold-700M while improving mean TM-score and median RMSD, indicating that the reasoning representation supplies complementary structural information to the decoding process.

4.4.3 Effect of Reasoning and Structural Conditioning

To verify that these gains arise from information usable by the structural decoder, we compare reasoning hidden states, explicit secondary-structure conditions, and the scope of parameter updates. All variants use SimpleFold-700M on CAMEO22 and are evaluated against the same model without external conditioning. The comparison varies ground-truth versus predicted eight-state secondary structure, the VLM checkpoint and hidden-state extraction position, and projector-only versus projector-plus-diffusion training.

The `Oracle DSSP one-hot` condition is computed from the target structure and supplies the eight-state DSSP assignment as a one-hot input. The `predicted SS8 hard labels` condition instead uses discrete secondary-structure predictions. Continuous conditions compare the S1-VL `prefill-last` representation, extracted after the protein sequence and instruction have been prefilled but before answer generation, with the S1-VL `im-end-last` representation extracted at the end-of-message token. Both continuous variants freeze SimpleFold-700M and train only the projector. The final variant jointly updates the projector and diffusion module to test whether a broader trainable parameter set is beneficial.

Effect of ground-truth secondary-structure conditioning. Oracle DSSP raises TM-score from 0.828760 to 0.831836 and GDT-TS from 0.785044 to 0.789831, while reducing RMSD from 4.557120 to 4.488475 (Table 28). At the same time, lDDT falls to 0.678164. The condition improves global topology and coordinate alignment but has the opposite effect on all-atom local distance accuracy, showing that eight-state secondary structure affects global folding and local atomic geometry differently in this setting.

Effect of hidden-state extraction. Relative to unconditioned SimpleFold-700M, S1-VL prefill-last raises TM-score by 0.001016, GDT-TS by 0.000196, and lDDT by 0.000087. S1-VL im-end-last raises TM-score by 0.000601 and GDT-TS by 0.000464. Both extraction positions improve TM-score and GDT-TS, while exhibiting different tradeoffs in local geometry and RMSD. The consistent direction of the global metrics indicates that reasoning-conditioned hidden states convey task information that SimpleFold-700M can exploit for folding.

Effect of predicted secondary-structure conditioning. Predicted SS8 hard labels yield lower TM-score, GDT-TS, lDDT, and lDDT-Ca than the unconditioned reference, together with higher RMSD. The divergence between Oracle DSSP and predicted SS8 shows that discrete secondary-structure conditioning depends on the fidelity of the labels to the target structure. Here, predicted labels provide weaker geometric constraints than ground-truth DSSP. An effective reasoning condition must therefore preserve scientific relations consistent with the structural evidence before a geometric decoder can use it.

Effect of trainable-module scope. Jointly training the projector and diffusion module produces lower values on all five metrics than either projector-only hidden-state variant. Projector-only training concentrates adaptation on the mapping from the shared hidden state to the structural conditioning space while

Table 29: Scientific image generation benchmark results. All metrics are reported on a 0–100 scale; higher is better. Best and second-best results are bold and underlined, respectively.

Image generation model	GenExam Relaxed Score ↑	TechImage-Bench rubric accuracy ↑	CVTG-2K Word Accuracy ↑
Qwen-Image	23.8	51.11	72.38
GLM-Image	46.1	54.47	**91.16**
GPT-Image-2	**93.8**	**83.87**	77.94
Nano Banana 2	<u>92.6</u>	<u>78.38</u>	77.88
S1-Omni	51.4	60.68	<u>80.39</u>

preserving the pretrained geometric parameters of SimpleFold-700M. The controlled results favor this lightweight adaptation strategy.

Across the main comparison and conditioning experiments, reasoning conditioning improves mean TM-score with a fixed SimpleFold-700M backbone, and both hidden-state extraction positions consistently improve TM-score and GDT-TS. Oracle DSSP further confirms that the conditioning pathway can use structure-related information to improve global folding. The central conclusion is that task states formed through scientific reasoning can enter a geometric decoder through a lightweight interface and positively affect structure prediction. With a larger computational budget, the same fusion mechanism can be evaluated on larger SimpleFold variants or stronger structural backbones to determine whether the gain scales with decoder capacity.

4.5 Scientific Image Generation and Editing

4.5.1 Experimental Setup

Scientific image generation synthesizes research illustrations, mechanism diagrams, technical workflows, and text-dense visual explanations from language instructions. Scientific image editing produces a modified image from an input image and an instruction. We evaluate both the organization of scientific semantics in generated images and pixel-space operations including segmentation, modality translation, super-resolution, and multi-step local editing.

Setting. Image generation is evaluated on three complementary benchmarks. GenExam (Wang et al., 2026) covers multidisciplinary exam-style scientific illustrations; GPT-5 judges semantic criteria, spelling, readability, and logical consistency to produce a weighted Relaxed Score. TechImage-Bench (Ni et al., 2025) covers biological, engineering, and general technical diagrams and uses o4-mini to evaluate hierarchical binary rubrics. CVTG-2K (Tai et al., 2025) evaluates complex multi-region text rendering with OCR-based Word Accuracy. Image editing covers three operation classes. MSD (Antonelli et al., 2022) and cigRockSEM (Zhang et al., 2025a) evaluate medical-image and rock-SEM segmentation; both predictions and ground truth are converted to binary masks by RGB thresholding before Dice or F1 is computed. SynthRAD2025 (Thummerer et al., 2025) and BCI (Liu et al., 2022) evaluate CBCT-to-CT and H&E-to-IHC translation, respectively, while IXI (Biomedical Image Analysis Group, Imperial College London, 2026) evaluates MRI super-resolution. All three reconstruction tasks use PSNR: SynthRAD2025 follows the official protocol in the HU domain within the anatomical region, whereas BCI and IXI use the standard 8-bit image domain. Volumetric data are processed as two-dimensional slices.

Baselines. General-purpose image models are Qwen-Image (Wu et al., 2025), GLM-Image (Z.ai, 2026), GPT-Image-2 (OpenAI, 2026a), and Nano Banana 2 (`gemini-3.1-flash-image-preview`) (Raisinghani, 2026). Each editing dataset also has an independent task-specific reference: Swin UNETR for MSD (Tang et al., 2022), U-Net for cigRockSEM (Zhang et al., 2025a), VBoussot for SynthRAD2025 (Boussot et al., 2025), Pyramid Pix2pix for BCI (Liu et al., 2022), and CodeBrain for IXI (Wu et al., 2026b). These are different models rather than a single specialist evaluated across all five columns.

4.5.2 Scientific Image Generation

S1-Omni exceeds both open baselines on GenExam and TechImage-Bench (Table 29). Relative to Qwen-Image, the gains are 27.6 and 9.57 points; relative to GLM-Image, they are 5.3 and 6.21 points. GPT-Image-2 and Nano Banana 2 remain substantially stronger on both rubric-based scientific illustration benchmarks.

CVTG-2K produces a different ranking. S1-Omni scores 80.39, above Qwen-Image, GPT-Image-2, and Nano Banana 2, while GLM-Image leads at 91.16. GenExam and TechImage-Bench jointly assess scientific entities, relations, layout, and labels; CVTG-2K isolates character rendering. The results therefore separate content organization from typography: S1-Omni is the stronger open model on structured scientific content, whereas character-level fidelity remains an independent bottleneck.

Table 30: Scientific image editing benchmark results. MSD and cigRockSEM evaluate segmentation with Dice and F1, respectively; SynthRAD2025 evaluates CBCT-to-CT translation, BCI evaluates H&E-to-IHC translation, and IXI evaluates 2×/4× MRI super-resolution with PSNR in dB. Higher is better for all metrics; best and second-best results are bold and underlined, respectively.

Model	MSD Dice ↑	cigRockSEM F1 ↑	SynthRAD2025 PSNR (dB) ↑	BCI PSNR (dB) ↑	IXI MRI SR PSNR (dB) ↑
Per-dataset specialist	<u>0.7868</u>	<u>0.9733</u>	34.78	**21.16**	24.67
Qwen-Image	0.0396	0.9210	<u>42.03</u>	0.00	14.25
GLM-Image	0.0442	0.6394	35.00	12.74	11.66
GPT-Image-2	0.3787	0.9001	40.49	14.30	19.96
Nano Banana 2	0.3263	0.8962	40.08	14.33	<u>25.27</u>
S1-Omni	**0.8496**	**0.9823**	**47.87**	<u>20.38</u>	**26.72**

S1-Omni and Qwen-Image share the same image-generation backbone. Their gap on GenExam and TechImage-Bench therefore shows that the task-conditioned hidden representation provides the decoder with constraints on scientific entities, relations, and layout. The clearer partitioning and directional relations in Appendix Figure 12 qualitatively complement the rubric results, whereas the remaining label errors are consistent with the character-level rendering gap measured by CVTG-2K.

4.5.3 Scientific Image Editing

Under the two-dimensional editing protocol, S1-Omni achieves the best performance on four of the five scientific image editing benchmarks: MSD, cigRockSEM, SynthRAD2025, and IXI (Table 30). On MSD, S1-Omni obtains a Dice score of 0.8496, outperforming the task-specific Swin UNETR specialist by 0.0628. On cigRockSEM, it reaches an F1 score of 0.9823, exceeding the dedicated U-Net baseline by 0.0090. For SynthRAD2025 and IXI, S1-Omni achieves PSNR values of 47.87 dB and 26.72 dB, surpassing the corresponding specialist models, VBoussot and CodeBrain, by 13.09 dB and 2.05 dB, respectively. On BCI, S1-Omni ranks second with 20.38 dB, while the pathology-specific Pyramid Pix2pix model retains a 0.78 dB advantage.

The segmentation results highlight the difficulty general-purpose image editing models face in performing spatially precise scientific editing. Segmentation-as-editing requires the model to translate a language-defined target into an exact spatial region, render the requested mask or color overlay consistently, and preserve the surrounding background. Boundary localization, mask rendering, color encoding, and background preservation therefore jointly determine performance. On MSD, all general-purpose models remain substantially below S1-Omni, with Dice scores ranging from 0.0396 to 0.3787. On cigRockSEM, S1-Omni also consistently outperforms both the specialist U-Net baseline and all general-purpose models. These results show that a shared image-editing decoder can perform accurate region-level operations across both medical and materials-science images, while Dice and F1 jointly characterize localization accuracy and mask-rendering quality.

Image translation and super-resolution reveal more fine-grained differences between the unified model and task-specific specialists. On SynthRAD2025, S1-Omni achieves 47.87 dB PSNR, outperforming the specialist VBoussot model by 13.09 dB and exceeding the strongest general-purpose baseline by 5.84 dB. This result indicates that S1-Omni can accurately reconstruct the target computed tomography pixel distribution from cone-beam computed tomography inputs under the slice-based evaluation protocol. On IXI, S1-Omni reaches 26.72 dB PSNR, outperforming CodeBrain by 2.05 dB. The smaller margin suggests that MRI super-resolution remains relatively close to the task-specific specialist regime, with Nano Banana 2 also obtaining a competitive result of 25.27 dB.

On BCI, Pyramid Pix2pix achieves 21.16 dB, leading S1-Omni by 0.78 dB and retaining a task-specific advantage in hematoxylin-and-eosin-to-immunohistochemistry stain translation. Nevertheless, S1-Omni substantially outperforms all evaluated general-purpose image models on this benchmark. The result suggests that highly specialized pathological appearance mappings can still benefit from architectures and training objectives tailored to a single stain-translation setting. Although PSNR measures pixel-level fidelity on SynthRAD2025, BCI, and IXI, high PSNR alone does not guarantee anatomical, pathological, or clinical validity, which should be further evaluated using corresponding domain-specific criteria.

Overall, S1-Omni supports a broad range of scientific image editing tasks, including segmentation, cross-modality image translation, and super-resolution. Across medical and materials-science benchmarks, it demonstrates strong adaptability to heterogeneous transformation objectives, ranging from spatially precise region delineation to modality conversion and fine-detail reconstruction. These results show that a unified scientific multimodal model can effectively handle diverse image editing settings, while task-specific metrics provide complementary views of its performance under different application

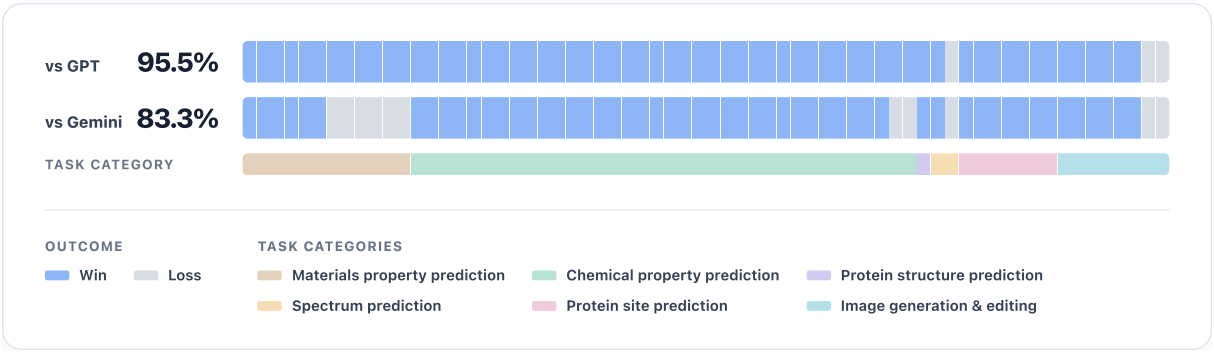


Figure 6: Aggregate task-level comparison with general-purpose model families under the shared text-and-image input protocol. The outcome strips summarize wins and losses for S1-Omni, and the lower strip shows the composition of the six scientific task categories.

requirements.

4.6 Aggregate Comparison with General-Purpose Models

To consolidate the general-purpose model comparisons reported throughout Section 4, we conduct a task-level analysis under a common input protocol. The full evaluation contains 66 tasks across six categories: materials property prediction, chemical property prediction, spectrum prediction, protein structure prediction, protein site prediction, and scientific image generation and editing. For comparability, every system receives only text and images, including textual serializations of scientific objects where required. Tasks evaluated through textual or structured predictions use GPT-5.5 (Singh et al., 2025) and Gemini-3.1-Pro (Gemini Team et al., 2024), whereas image-output tasks use GPT-Image-2 (OpenAI, 2026a) and Nano Banana 2 (Raisinghani, 2026).

Across the eligible comparisons, S1-Omni attains win rates of 95.5% against the GPT model family and 83.3% against the Gemini model family (Figure 6). The advantage extends across heterogeneous prediction and generation settings rather than being determined by a single benchmark family. These aggregate rates count task outcomes rather than performance margins and should therefore be interpreted together with the per-benchmark results in Section 4; nevertheless, their consistency across scalar, residue-level, structural, and pixel-space outputs provides evidence that S1-Omni is the strongest unified system in this comparison.

This comparison also clarifies where scientific specialization matters. General-purpose models provide a flexible language-and-vision interface, but many scientific tasks are evaluated in constrained output spaces whose validity depends on domain structure: calibrated scalar predictions, residue sets, molecular graphs, three-dimensional coordinates, segmentation masks, and reconstructed images. S1-Omni couples a shared reasoning backbone with decoders for domain-specific tasks that operate directly in these spaces, allowing scientific evidence to be translated into outputs that respect the corresponding metric, geometry, or spatial constraint. The results therefore support specialization at the output interface as an effective complement to unified multimodal reasoning: a common backbone enables transfer across domains, while task-aligned output spaces supply the inductive biases needed for reliable scientific prediction and generation.

5 Discussion

5.1 Reasoning Supervision Improves Scientific Representations

Across property prediction, protein structure prediction, and scientific image generation, reasoning supervision improves the representations supplied to downstream modules, but the construction of the reasoning data is decisive. On six material regression properties, structured, property-constrained supervision reduces MAE from 34.818 with unconstrained free-form reasoning to 17.255 and raises MAD/MAE from 0.888 to 1.260 (Table 6). It also outperforms the no-reasoning condition, whose MAE and MAD/MAE are 20.457 and 0.941, respectively. Because the Stage-2 predictor, data scale, optimization, and question-only pooling are fixed, these differences isolate the organization of the Stage-1 reasoning supervision rather than predictor capacity or answer-side extraction.

Appendix Sections A.1.4 and A.1.5 illustrate the distinction on the same Lu_2ZnPd crystal and Spillage

request. The unconstrained trace identifies potentially relevant ingredients, including heavy-element content, spin–orbit coupling, orbital character, and lattice symmetry, but it is free to choose its own coverage and analytical order. The structured trace first records a common set of structural and compositional observations, then selects a smaller property-relevant subset and states the competing directions explicitly: cell simplicity and orbital localization are treated as downward factors, whereas heavy-element-induced spin–orbit coupling acts upward. The important difference is therefore not trace length or the presence of scientific terminology, but whether the supervision exposes a consistent mapping from observed evidence to property-specific factors, their direction, and their interaction.

The case comparison is illustrative; the controlled ablation provides the aggregate evidence. Together, they show why fluent but unconstrained rationales can be a poor training target: variation in evidence coverage and causal ordering introduces nuisance variation into the training signal received by the shared backbone, causing unconstrained supervision to perform worse than omitting reasoning text altogether. Property-constrained construction reduces this variation by imposing a stable analytical schema while still requiring the teacher to select task-relevant factors. This result supports structured reasoning as a representation-learning objective, but does not imply that a template guarantees factual correctness; scientific claims within the traces still require data validation and domain review.

The protein experiments provide a controlled geometric test. SimpleFold-700M is fixed and only its conditioning representation is changed; reasoning-conditioned states produce small improvements in global folding metrics. This result is modest, but it establishes that a language-model task state can carry information usable by a geometric decoder without altering the decoder backbone.

Appendix Figure 13 makes the same mechanism visible in image generation. Without think-before-generate planning, a long instruction must be mapped directly to an image, and the model often fails to coordinate module grouping, hierarchy, arrows, labels, and global layout. Explicit reasoning first decomposes the request into executable visual units, such as input regions, the backbone, VTP, VVP, PIR, predictions, and the legend, and then specifies their positions, colors, connections, and information flow. The resulting scientific semantics, task intent, and layout plan are injected into MMDiT through reasoning-diffusion alignment. The decoder is therefore conditioned on an organized task state rather than the raw prompt alone, improving module completeness, topology, hierarchy, and readability.

Together, these results suggest that the principal value of reasoning supervision is representational. It teaches the shared backbone to form task states that remain useful to numerical predictors, molecular and protein structure generators, and image decoders, even when the downstream module does not consume the generated reasoning text.

5.2 Prefill Representations as a Stable Downstream Interface

A central design choice in decoding for domain-specific tasks is how to transfer a shared VLM representation to heterogeneous scientific modules. One option extracts the final answer or route-token state and compresses generation into a single vector. A second pools the chain-of-thought and answer states. A third uses question-side prefill representations formed from the scientific input and task instruction before autoregressive decoding.

The experiments show a consistent pattern in feature extraction from the shared backbone: task information required by domain decoders is largely present before answer generation, and pooling the full reasoning trajectory gives no stable advantage. In property prediction, question-only pooling improves both classification and regression relative to pooling reasoning and answer tokens (Table 7). In protein structure prediction, prefill-last and end-of-message states both provide useful conditions, but post-generation states are not consistently better (Table 28). In the spectrum-to-molecular generation task, evidence distributed along the input sequence is best retrieved by attention over input-side latent tokens, without waiting for a chain of thought or answer (Table 17).

Long reasoning states mix scientific evidence with linguistic organization, repeated explanation, autoregressive history, and output formatting; final answer states are additionally shaped by termination prediction. Pooling them can dilute a task signal already formed in the prefill. Question-side states, by contrast, are determined directly by the scientific object and requested task. A practical division of labor follows: reasoning supervision shapes the shared scientific representation during training, whereas domain decoders preferentially extract features from prefill states at inference, using pooling or attention only when evidence is distributed across input positions.

The present architecture still passes conditions predominantly in one direction, from the shared VLM to each domain decoder. Deeper bidirectional fusion between shared and domain-specific representations remains an important direction for unified scientific reasoning.

5.3 Backbone Contributions Under Strong Expert Decoders

A natural question is whether the gains in unified scientific modeling primarily come from stronger task experts or from the shared backbone representation. Our analysis experiments are designed to isolate this question by reusing two strong, task-specific decoders—a spectrum-to-molecular generation expert and a protein-folding expert—and changing only the conditioning representation supplied by the shared backbone. Under this controlled reuse of expert modules, improvements can be attributed to the backbone states rather than to decoder redesign.

The effect is clearest for spectrum-to-molecular generation. With a spectrum-conditioned expert decoder in place, S1-Omni achieves a substantially higher exact structure match rate than the specialist baseline (Table 14), indicating that the shared backbone encodes spectral evidence in a form that is easier for the expert to convert into a valid molecular structure.

For protein folding, the gains are smaller but still directionally positive. When SimpleFold-700M is held fixed, reasoning-conditioned backbone representations yield modest improvements in selected global and alignment metrics (Table 27), suggesting that the shared backbone can inject useful structural signals even when the downstream geometric prior is already strong.

Overall, these controlled results support the view that a meaningful fraction of S1-Omni's benefit comes from backbone representations that enhance downstream experts, rather than from task decoders alone.

5.4 General Conversational Competence Under Scientific Specialization

We construct a general multi-turn dialogue evaluation to examine whether science-oriented training weakens general conversational ability and whether the model can switch tasks reliably within a long context. The conversation begins with non-specialized requests, including a self-introduction and the generation of a mathematics lesson plan, before introducing a protein PPI-site prediction task in the same context (Appendix Section A.4). We use the identical protein sequence and task instruction as an independent single-turn control. This setting tests whether the model can track a multi-turn context, isolate history unrelated to the current task, and preserve the scientific task protocol and output format when switching from general-purpose generation to scientific prediction.

The multi-turn and single-turn settings produce exactly the same final PPI-site indices, indicating that the preceding general conversation does not alter the scientific prediction in this example. After generating a self-introduction and educational content, the model recognizes the protein sequence, recovers the task representation required for site prediction, and returns the same structured answer as the single-turn control. This result suggests that science-oriented training does not eliminate the model's general multi-turn capability: S1-Omni can retain context understanding, task switching, and scientific output consistency across heterogeneous turns. Longer and more complex interactions require broader evaluation, but this case provides initial evidence that scientific specialization can coexist with general conversational competence.

6 Limitations

We view S1-Omni as an initial step toward scalable unified multimodal reasoning for scientific understanding, prediction, and generation. The present work establishes a continuous modeling process linking heterogeneous scientific data, shared task representations, and domain-native results, and tests its feasibility on representative tasks. Several limitations become more important as the model moves toward larger and more open scientific settings.

Large-scale scientific pretraining. S1-Omni has not undergone large-scale scientific pretraining. It starts from an existing vision-language model and is trained on curated scientific tasks, a setting that isolates unified task protocols and module integration but does not reveal scaling laws with data, parameters, or compute. Future work should extend the current task system to continual scientific pretraining at larger scale and measure how scientific representations, cross-task transfer, and domain capabilities change.

Modular generation across scientific output spaces. Molecular structures, protein coordinates, and scientific images are not generated natively by the shared backbone; they rely on diffusion or specialist decoders. This design preserves geometric accuracy and generation quality in each output space, but remains an intermediate point between unified task representation and unified native generation. Systematic comparisons of token, diffusion, and hybrid approaches are needed to determine how much structural representation and generation can move into the shared backbone without losing the inductive biases required by continuous spaces.

Task coverage. The current evaluation uses property prediction, structure reconstruction, protein tasks, and scientific images as initial test beds for connecting different input and output spaces. It does not yet provide sufficient disciplinary depth, complex multistage problems, out-of-distribution generalization across task families, or cross-task transfer. Broader tasks and more systematic transfer studies are required to test whether existing capabilities can be composed to solve genuinely new scientific problems.

General capabilities and long-horizon tasks. Scientific task training reduces performance on some general tasks relative to the base model, and the present evaluation concentrates on bounded single-turn problems. Multi-turn planning, cross-modal interaction, long-horizon state tracking, tool use, and correction across steps remain underexplored. Future training should treat retention of general capabilities and scientific performance as joint objectives, while evaluating planning, execution, and feedback-driven revision in long-horizon research workflows.

7 Conclusion

In this work, we present S1-Omni, a unified multimodal reasoning model for scientific understanding, prediction, and generation trained on the S1-Omni-Corpus. Through unified representation of scientific data, natural-world knowledge alignment, and decoding for domain-specific tasks, S1-Omni links heterogeneous scientific inputs, evidence-grounded reasoning, and native scientific outputs within a single model, offering a scalable path towards unified multimodal reasoning for science. More broadly, we view the central challenge of scientific AI systems as threefold: identifying the scientific evidence from which models should learn, determining the reusable knowledge and capabilities they should acquire, and distilling these into scalable, efficient systems that generalize across domains and tasks while remaining scientifically reliable and verifiable. Future work will explore continual scientific pretraining, structure-aware generation, out-of-distribution generalization, and tool-augmented research workflows, with the goal of building scalable, verifiable, and collaborative scientific AI.

Author Contributions

Core contributors and contributors are listed alphabetically by given name within each group.

Core Contributors. Jiahao Zhao, Junyi Liu, Lifeng Xu, Nan Xu, Qingli Wang, Qingxiao Li, Tianle Chen, Xiaoyu Wu, Yawen Zheng, Zikai Wang

Contributors. Guanming Liu, Hequn Zhou, Jingyi Wang, Jingyuan Shu, Keqi Wang, Li He, Songyang Diao, Wenhui Xu, Xinyu Ren, Yaqin Fan, Yujin Zhou, Zhanao Yao

References

Josh Abramson et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024. doi: 10.1038/s41586-024-07487-w.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Model card.

Anthropic. Claude science, an ai workbench for scientists. https://www.anthropic.com/news/claude-science-ai-workbench, 2026.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Evan R. Antoniuk, Shehtab Zaman, Tal Ben-Nun, Peggy Li, James Diffenderfer, Busra Sahin, Obadiah Smolenski, Tim Hsu, Anna M. Hiszpanski, Kenneth Chiu, Bhavya Kailkhura, and Brian Van Essen. BOOM: Benchmarking out-of-distribution molecular property predictions of machine learning models. *arXiv preprint arXiv:2505.01912*, 2025. URL https://arxiv.org/abs/2505.01912.

Sarp Aykent and Tian Xia. GotenNet: Rethinking efficient 3d equivariant graph neural networks. OpenReview preprint, 2024. URL https://openreview.net/forum?id=5wxCQDtbMo.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Minkyung Baek, Ivan Anishchenko, Ian R Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using rosettafold2. *BioRxiv*, pp. 2023–05, 2023.

Biomedical Image Analysis Group, Imperial College London. Ixi dataset. https://brain-development.org/ixi-dataset/, 2026. Accessed 2026-06-05.

Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4):432–444, 2023. doi: 10.1038/s42256-023-00639-z. URL https://doi.org/10.1038/s42256-023-00639-z.

Valentin Boussot, Cédric Hémon, Jean-Claude Nunes, and Jean-Louis Dillenseger. Why registration quality matters: Enhancing sct synthesis with impact-based registration. *arXiv preprint arXiv:2510.21358*, 2025.

Feiyang Cai, Katelin Zacour, Tianyu Zhu, Tzuen-Rong Tzeng, Yongping Duan, Ling Liu, Srikanth Pilla, Gang Li, and Feng Luo. Chemfm as a scaling law guided foundation model pre-trained on informative chemicals. *Communications Chemistry*, 2025.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

Deep Principle Team. Materials property axiom: Scaling foundation models to experimental property generalists via multi-phase training. *Technical report*, 2026.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, et al. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, pp. 2026–03, 2026.

Zhen Feng, Hui Yu, Xiaoya Guan, Zhenyu Yue, Lichuan Gu, Ke Li, and Xiaobo Zhou. Protein-dna binding sites prediction via integrating pretrained large language models and contrastive learning. *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–19, 2025.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Wentao Gong, Feifan Zhang, Junfan Chen, and Kehan Lu. Hgctbind: A hybrid architecture for predicting protein-dna binding sites based on interpretable and contextual adaptive feature fusion. *Computational Biology and Chemistry*, pp. 108979, 2026.

Jürgen Haas, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953, 2025.

Magnus Haraldson Høie, Frederik Steensgaard Gade, Julie Maria Johansen, Charlotte Würtzen, Ole Winther, Morten Nielsen, and Paolo Marcatili. Discotope-3.0: improved b-cell epitope prediction using inverse folding latent representations. *Frontiers in immunology*, 15:1322712, 2024.

Feng Hu, Wenwu Zeng, and Shaoliang Peng. Megsite: an accurate nucleic acid-binding residue prediction method based on multimodal protein language model. *Briefings in Bioinformatics*, 26(5):bbaf524, 2025.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2d & 3d diffusion models for complete molecule generation. *arXiv preprint arXiv:2305.12347*, 2023.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Sapir Israeli and Yoram Louzoun. Single-residue linear and conformational b cell epitopes prediction using random and esm-2 based projections. *Briefings in bioinformatics*, 25(2):bbae084, 2024.

Nikita V Ivanisenko, Tatiana I Shashkova, Andrey Shevtsov, Maria Sindeeva, Dmitriy Umerenkov, and Olga Kardymon. Sema 2.0: web-platform for b-cell conformational epitopes prediction using artificial intelligence. *Nucleic Acids Research*, 52(W1):W533–W539, 2024.

A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. The materials project: a materials genome approach to accelerating materials innovation. apl mater 1: 011002. *Available from DOI*, 10(1.4812323), 2013.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Uni-Mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*, 2024. URL https://arxiv.org/abs/2406.14969.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Lucien F Krapp, Luciano A Abriata, Fabio Cortés Rodriguez, and Matteo Dal Peraro. Pesto: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature communications*, 14(1):2175, 2023.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10(1):39, 2018.

Mingyang Li, Yurou Liu, Jieping Ye, Bing Su, Ji-Rong Wen, and Zheng Wang. Speaking the language of science: Toward a general-purpose generative foundation model for the natural sciences. *arXiv preprint arXiv:2606.16905*, 2026a.

Qingxiao Li, Zikai Wang, Qingli Wang, and Nan Xu. S1-omni-image: A unified model for scientific image understanding, generation, and editing. *arXiv preprint arXiv:2606.24441*, 2026b.

Qingxiao Li, Lifeng Xu, Qingli Wang, Yudong Bai, Mingwei Ou, Shu Hu, and Nan Xu. S1-vl: Scientific multimodal reasoning model with thinking-with-images. *arXiv preprint arXiv:2604.21409*, 2026c.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1815–1824, 2022.

Xing Liu, Jason Kantorow, Ashesh K Chattopadhyay, and Srirupa Chakraborty. Interpretable antibody–antigen structural interface prediction via adaptive graph learning and cyclic transfer. *bioRxiv*, 2026.

Yurou Liu, Mingyang Li, Xinyuan Zhu, Rui Jiao, Yiming Dong, Xinyu Tang, Yang Liu, Jieping Ye, Bing Su, and Zheng Wang. Drugtrail: Interpretable drug discovery via structured reasoning and druggability-tailored preference optimization. In *The Fourteenth International Conference on Learning Representations*.

Jiarui Lu, Xiaoyin Chen, Stephen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *International Conference on Learning Representations*, volume 2025, pp. 98645–98678, 2025.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29 (21):2722–2728, 2013.

Jia Mi, Chang Li, Han Wang, Ying Du, Chong Chu, Jing Wan, and Kunfeng Wang. Uspdb: A novel u-shaped equivariant graph neural network with subgraph sampling for protein-dna binding site prediction. *Expert Systems with Applications*, 291:128554, 2025.

Siddharth Narayanan, James Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Michael Pieler, Sam Rodriques, and Andrew White. Training a scientific reasoning model for chemistry. *Advances in Neural Information Processing Systems*, 38:157671–157710, 2026.

Minheng Ni, Zhengyuan Yang, Yaowen Zhang, Linjie Li, Chung-Ching Lin, Kevin Lin, Zhendong Wang, Xiaofei Wang, Shujie Liu, Lei Zhang, Wangmeng Zuo, and Lijuan Wang. Techimage-bench: Rubric-based evaluation for professional image generation. *arXiv preprint arXiv:2512.12220*, 2025.

Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. Llm-prop: predicting the properties of crystalline materials using large language models. *npj Computational Materials*, 11(1): 186, 2025.

ODesign Team. Odesign: A world model for biomolecular interaction design. *arXiv preprint arXiv:2510.22304*, 2025.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. Introducing chatgpt images 2.0. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026a.

OpenAI. Introducing gpt-rosalind for life sciences research. https://openai.com/index/introducing-gpt-rosalind/, 2026b.

Angelos-Michael Papadopoulos, Apostolos Axenopoulos, Anastasia Iatrou, Kostas Stamatopoulos, Federico Alvarez, and Petros Daras. Parasurf: a surface-based deep learning approach for paratope–antigen interaction prediction. *Bioinformatics*, 41(2):btaf062, 2025.

Qizhi Pei, Zhimeng Zhou, Yi Duan, Yiyang Zhao, Wei Li, Han Guo, Liang He, Chengping Li, Chang-Yu Hsieh, Conghui He, et al. Biomatrix: Towards a comprehensive biological foundation model spanning the modality matrix of sequences, structures, and language. *arXiv preprint arXiv:2606.22138*, 2026.

Yingming Pu, Tao Lin, and Hongyu Chen. Piflow: Principle-aware scientific discovery with multi-agent collaboration. *arXiv preprint arXiv:2505.15047*, 2025.

Naina Raisinghani. Nano banana 2: Combining pro capabilities with lightning-fast speed. *Google Blog, February*, 2026.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Ying Tai, Nikai Du, Rui Xie, et al. Investigating text insulation and attention mechanisms for complex visual text generation. *arXiv preprint arXiv:2503.23461*, 2025. URL https://arxiv.org/abs/2503.23461.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Adrian Thummerer, Erik van der Bijl, Arthur Jr Galapon, Florian Kamp, Mark Savenije, Christina Muijs, Shafak Aluwini, Roel JHM Steenbakkers, Stephanie Beuel, Martijn PW Intven, et al. Synthrad2025 grand challenge dataset: Generating synthetic cts for radiotherapy from head to abdomen. *Medical physics*, 52(7):e17981, 2025.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4), 2022.

Alexandra Wahab, Lara Pfuderer, Eno Paenurk, and Renana Gershoni-Poranne. The compas project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons. *Journal of Chemical Information and Modeling*, 62(16):3704–3713, 2022.

Jue Wang, Yufan Liu, and Boxue Tian. Protein-small molecule binding site prediction based on a pre-trained protein language model with contrastive learning. *Journal of cheminformatics*, 16(1):125, 2024a.

Liang Wang, Yu Rong, Tingyang Xu, Zhenyi Zhong, Zhiyuan Liu, Pengju Wang, Deli Zhao, Qiang Liu, Shu Wu, and Yang Zhang. Diffspectra: Molecular structure elucidation from spectra using diffusion models. *arXiv preprint arXiv:2507.06853*, 2025a.

Weida Wang, Benteng Chen, Di Zhang, Wanhao Liu, Shuchen Pu, Ben Gao, Jin Zeng, Xiaoyong Wei, Tianshu Yu, Shuzhou Sun, et al. Chem-r: Learning to reason as a chemist. *arXiv preprint arXiv:2510.16880*, 2025b.

Yanli Wang, Frimpong Boadu, and Jianlin Cheng. Mpbind: a multitask protein binding site predictor using protein language models and equivariant gnns. *Bioinformatics*, 41(11):btaf589, 2025c.

Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista. Simplefold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480*, 2025d.

Zhaokai Wang, Penghao Yin, Xiangyu Zhao, Changyao Tian, Yu Qiao, Wenhai Wang, Jifeng Dai, and Gen Luo. Genexam: A multidisciplinary text-to-image exam. In *Forty-third International Conference on Machine Learning*, 2026.

Zhiwei Wang, Yongkang Wang, and Wen Zhang. Improving paratope and epitope prediction by multi-modal contrastive learning and interaction informativeness estimation. *arXiv preprint arXiv:2405.20668*, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Fang Wu, Weihao Xuan, Heli Qi, Hanqun Cao, Heng-Jui Chang, Zeqi Zhou, Haokai Zhao, Ma Jian, Carl Ma, Yu-Chi Cheng, et al. Proteo-r1: Reasoning foundation models for de novo protein design. *arXiv preprint arXiv:2605.02937*, 2026a.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Yicheng Wu, Tao Song, Zhonghua Wu, Jin Ye, Zongyuan Ge, Wenjia Bai, Zhaolin Chen, and Jianfei Cai. Virtual full-stack scanning of brain mri via imputing any quantised code. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21026–21035, 2026b.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Ying Xia, Chun-Qiu Xia, Xiaoyong Pan, and Hong-Bin Shen. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9):e51–e51, 2021.

Ying Xia, Xiaoyong Pan, and Hong-Bin Shen. Ligbind: identifying binding residues for over 1000 ligands with relation-aware graph neural networks. *Journal of Molecular Biology*, 435(13):168091, 2023.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, et al. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv e-prints*, pp. arXiv–2502, 2025.

Jin Xu et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Yitao Xu, Guanyun Wei, Jingying Zhou, Yuanhua Huang, Weichuan Yu, Zhixiang Lin, Ran Liu, and Xiaodan Fan. Robep: a region-oriented deep learning model for b-cell epitope prediction. *Bioinformatics*, 42(1):btag006, 2026.

Hongyang Ye, Changrong Zheng, Lei Wang, Yan Wang, and Zhidong Xue. Splinet: Lightweight prediction of protein-rna interaction sites by integrating a structure-aware protein language model with a dual-branch network. *Computational Biology and Chemistry*, pp. 109130, 2026.

Z.ai. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation. https://z.ai/blog/glm-image, 2026.

Yuansong Zeng, Zhuoyi Wei, Qianmu Yuan, Sheng Chen, Weijiang Yu, Yutong Lu, Jianzhao Gao, and Yuedong Yang. Identifying b-cell epitopes using alphafold2 predicted structures and pretrained language model. *Bioinformatics*, 39(4):btad187, 2023.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Yao Zhang, Xinming Wu, and Jiachun You. A benchmark dataset and baseline methods for rock microstructure interpretation in sem images. *Scientific Data*, 12(1):1671, 2025a.

Zhijun Zhang, Lijun Quan, Junkai Wang, Liangchen Peng, Qiufeng Chen, Bei Zhang, Lexin Cao, Yelu Jiang, Geng Li, Liangpeng Nie, et al. Labind: Identifying protein binding ligand-aware sites via learning interactions between ligand and protein. *Nature Communications*, 16(1):7712, 2025b.

Chen Zhou, Zikun Chen, Lu Zhang, Deyu Yan, Tiantian Mao, Kailin Tang, Tianyi Qiu, and Zhiwei Cao. Seppa 3.0—enhanced spatial epitope prediction enabling glycoprotein antigens. *Nucleic acids research*, 47(W1):W388–W394, 2019.

Zhiwang Zhou, Yuandong Pu, Xuming He, Yidi Liu, Yixin Chen, Junchao Gong, Xiang Zhuang, Wanghan Xu, Qinglong Cao, Shixiang Tang, et al. Omni-weather: Unified multimodal foundation model for weather generation and understanding. *arXiv preprint arXiv:2512.21643*, 2025.

Jiaxi Zhuang, Yaorui Shi, Jue Hou, Yunong He, Mingwei Ye, Mingjun Xu, Yuming Su, Linfeng Zhang, Ying Qian, Guolin Ke, et al. Reasoning-enhanced large language models for molecular property prediction. *arXiv preprint arXiv:2510.10248*, 2025.

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

A Appendix

This section first presents thirteen complete reasoning examples using the same task-family order as Section 4. It then provides targeted visual analyses of scientific image generation and editing, examines representative failure modes, and evaluates whether scientific prediction remains stable after a long, general-purpose multi-turn conversation. The original language of every interaction is retained.

A.1 Reasoning Examples

Each example starts on a new page and shows the full user input, the complete model response, and the emitted task token. Native numerical, structural, or image output is included when it is available in the source record. The two controlled reasoning-supervision examples report no native scalar output, so none is inferred here. Table 31 provides a compact overview.

Table 31: Case catalogue ordered by the task families used in the Evaluation section. Each route token selects the corresponding native result interface.

Task family	Case	Input	Route token	Native output
Property	DFPT dielectric constant	Crystal CIF	`<linear_pre>`	Scalar: 30.15
Property	Exfoliation energy	Crystal CIF	`<linear_pre>`	Scalar: 49.97 meV/atom
Property	Topological spillage	Crystal CIF	`<linear_pre>`	Scalar: 0.215
Property	Spillage: unconstrained free-form reasoning	Crystal CIF	`<linear_pre>`	Not provided in source
Property	Spillage: structured, property-constrained reasoning	Crystal CIF	`<linear_pre>`	Not provided in source
Property	ESOL aqueous solubility	Molecular SMILES	`<linear_pre>`	Scalar: -3.2656 log S
Property	CYP2C9 substrate	Molecular SMILES	`<linear_cla>`	Non-substrate
Property	CYP2D6 inhibition	Molecular SMILES	`<linear_cla>`	Non-inhibitor
Spectrum	IR spectrum-to-molecular generation	Infrared spectrum	`<spectra_st>`	Cyclopropane nitrile structure
Protein site	ArcA DNA-binding site	Amino-acid sequence	`<prot_cla>`	Residue indices
Protein structure	GPR168 structure	Amino-acid sequence	`<prot_st>`	All-atom coordinates
Image generation	Chemical synapse diagram	Text instruction	`<image_gen>`	Scientific image
Image editing	Tumor and edema segmentation	FLAIR MRI + instruction	`<image_edit>`	Spatial overlay

A.1.1 Crystal Property Prediction: DFPT Dielectric Constant

User Input

下面给出了一个晶体结构的CIF 文件内容：

```
# generated using pymatgen
data_Pr2AlNO3
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   3.752
_cell_length_b   3.752
_cell_length_c   6.875
_cell_angle_alpha   105.837
_cell_angle_beta   105.837
_cell_angle_gamma   90.000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Pr2AlNO3
_chemical_formula_sum   'Pr2 Al1 N1 O3'
_cell_volume   89.298
_cell_formula_units_Z   1
loop_
 _atom_site_type_symbol _atom_site_label _atom_site_fract_x _atom_site_fract_y _atom_site_fract_z
  Pr  Pr0  0.520  0.520  0.040
  Pr  Pr1  0.808  0.808  0.617
  Al  Al2  0.158  0.158  0.316
  N  N3  0.998  0.998  0.995
  O  O4  0.170  0.670  0.340
  O  O5  0.670  0.170  0.340
  O  O6  0.329  0.329  0.658
```

请给出其对应的总介电常数ε（DFPT）（无量纲）精确值。

Model Response

<think>
首先对提供的CIF 文件进行结构定量表征。该材料化学式为Pr2AlNO3，结晶于三斜晶系，空间群为P 1（第1号），这是对称性最低的晶系，意味着原子位置几乎没有对称性约束。晶格参数显示a 与b 相等均为3.752 Å，c 轴较长为6.875 Å，角度α 和β 均为105.837°，γ 为90.000°，单胞体积约为89.3 Å^3，每个晶胞包含1 个化学式单元。从组成特征来看，材料含有重元素稀土镨（Pr），其原子序数较大且具有f 电子，同时包含轻金属铝（Al）及非金属氮（N）和氧（O）。平均原子质量较重，电负性差异显著，尤其是Pr 与O/N 之间。局域结构方面，由于空间群为P1，所有原子均处于不等价位点，配位环境复杂，Pr 原子可能形成高配位数的多面体，而Al 可能与N/O 形成较硬的四面体或八面体连接。结构拓扑上，较小的晶胞体积与较多的原子数（7 个原子/晶胞）表明结构堆积较为致密，不存在明显的层状弱连接特征，整体呈现三维框架结构。
接着筛选影响总介电常数的关键结构特征。第一是稀土元素Pr 的存在，Pr 具有较大的离子半径和丰富的f 轨道电子，这会显著增加电子极化率并可能通过软模机制增强离子极化率，对介电常数有强烈的正向提升作用。第二是低对称性的三斜结构，P1 空间群允许更多的声学支和光学支振动模式耦合，降低了限制离子位移的对称性壁垒，有利于提高离子对介电常数的贡献。第三是较高的结构致密性，89.3 Å^3 的体积容纳7 个原子意味着较高的数密度，通常高密度与较高的电子介电常数相关联，但过强的键合也可能限制离子位移，这是一个竞争因素。
然后进行多因素竞争分析。虽然Al-N/O 键通常具有较强的共价性且刚度较大，倾向于限制晶格畸变从而降低离子介电贡献，但Pr 元素的引入极大地补偿了这一点。Pr-O/N 键的离子性较强且Pr 离子本身易极化，主导了整体的极化行为。低对称性进一步释放了晶格弛豫的潜力，使得离子贡献不会被完全抑制。因此，稀土极化效应与低对称性效应共同主导，克服了轻元素骨架可能带来的刚性限制，预计总介电常数将高于普通简单氧化物。
最后进行定性逻辑推导。一般简单二元氧化物的静态介电常数通常处于基础水平。考虑到该材料含有两个高极化率的Pr 原子，基础值应向上修正。再结合三斜晶系低对称性带来的额外离子贡献，数值有望进一步攀升。然而，由于晶胞体积较小且含有刚性的Al 组分，介电常数不太可能达到铁电体或超高介电常数材料的水平。综合评估，上述结构与机理因素共同支撑最终预测任务。
</think>
材料分子式为Pr2AlNO3，目标预测性质为总介电常数ε（DFPT），单位为无量纲。该晶体属于三斜晶系，空间群为P 1，晶胞体积约为89.3 Å^3，包含稀土元素镨以及铝、氮、氧组分，背景条件涉及低频极化与电子云畸变的共同贡献。
由于稀土镨元素的高极化率显著增强了电子与离子贡献，且三斜低对称性结构允许更多的晶格弛豫模式，导致介电响应处于中等偏高水平；结合较高的原子堆积密度进一步限制了极端低值出现的可能性，表明其为具有良好介电性能的潜在介质材料。接下来，我将完成`eps_dfpt` 的最终预测任务。
<linear_pre>

Native Output

总介电常数ε **30.15** DFPT，无量纲

A.1.2 Crystal Property Prediction: Exfoliation Energy

User Input

Given the CIF file of a crystal structure:

```
# generated using pymatgen
data_Ta2CS2
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   8.833
_cell_length_b   8.833
_cell_length_c   8.833
_cell_angle_alpha   21.505
_cell_angle_beta   21.505
_cell_angle_gamma   21.505
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Ta2CS2
_chemical_formula_sum   'Ta2 C1 S2'
_cell_volume   81.153
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Ta  Ta0  1  0.621  0.621  0.621  1
  Ta  Ta1  1  0.379  0.379  0.379  1
  C  C2  1  0.000  0.000  0.000  1
  S  S3  1  0.775  0.775  0.775  1
  S  S4  1  0.225  0.225  0.225  1
```

Please estimate the exfoliation energy (Exf. En., meV/atom) of this material.

Model Response

<think>

First, the crystal structure of Ta2CS2 is examined based on the provided CIF data, revealing a triclinic symmetry (P 1) with highly symmetric lattice parameters where a=b=c=8.833 Å and all angles are acute at 21.505°. This specific geometric configuration, despite the low symmetry label, suggests a rhombohedral stacking sequence often associated with layered transition metal compounds. The unit cell contains five atoms, comprising two tantalum atoms, one carbon atom, and two sulfur atoms, indicating a dense packing with a volume of approximately 81.15 $Å^3$.

Next, the chemical composition is analyzed to understand the bonding nature. Tantalum is a heavy transition metal capable of forming strong metallic and covalent bonds, while sulfur is a chalcogen known for facilitating van der Waals gaps in dichalcogenides. The inclusion of carbon introduces strong covalent character, likely strengthening the intralayer connectivity compared to pure tantalum sulfide. The atomic positions aligned along the body diagonal further support a layered motif where sulfur atoms likely terminate the slabs, creating potential cleavage planes.

Then, the key factors influencing exfoliation energy are weighed. The primary driver for low exfoliation energy is the presence of sulfur terminations which typically result in weak interlayer van der Waals forces. However, the carbon atoms may act as interlayer bridges or intralayer strengtheners, potentially increasing the energy required to separate layers compared to standard MX2 compounds. The competition between the weak S-S interlayer interactions and the strong Ta-C intralayer bonds determines the final value.

Finally, qualitative logic is applied based on these structural insights. Typical layered transition metal dichalcogenides exhibit characteristic exfoliation energies. Given the reinforcing effect of carbon in this specific Ta2CS2 stoichiometry, the value is expected to be slightly elevated but still within the regime of easily exfoliable materials. Therefore, the energy is predicted to be consistent with a stable yet separable layered structure.

</think>

The material Ta2CS2 is evaluated for its Exfoliation Energy (Exf. En.), expressed in meV/atom. This property is critical for assessing the feasibility of isolating single layers from the bulk crystal for nanotechnology applications. Exfoliation Energy (Exf. En.) is a key metric for layered materials. Driven by the weak van der Waals interactions between sulfur-terminated layers but moderated by the strong covalent Ta-C bonding within the slabs, the energy reflects the balance of interlayer and intralayer forces. This suggests the material possesses a layered character that allows for mechanical exfoliation while maintaining structural integrity. Next, I will predict the `exf_en` value.

<linear_pre>

Native Output

Predicted exfoliation energy **49.97** meV/atom

A.1.3 Crystal Property Prediction: Topological Spillage

User Input

根据给定的CIF 文件内容:

```
# generated using pymatgen
data_VSe2
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   3.372
_cell_length_b   3.372
_cell_length_c   6.201
_cell_angle_alpha   90.000
_cell_angle_beta   90.000
_cell_angle_gamma   120.000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   VSe2
_chemical_formula_sum   'V1 Se2'
_cell_volume   61.070
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  V  V0  1  0.000  0.000  0.000  1
  Se  Se1  1  0.333  0.667  0.252  1
  Se  Se2  1  0.667  0.333  0.748  1
```

请输出该晶体结构的Spillage（无量纲）精确值。

Model Response

<think>

首先，从提供的CIF 文件解析出该材料为二硒化钒（VSe2），晶体结构呈现典型的层状过渡金属硫族化合物特征。晶格参数显示a=b=3.372 Å，c=6.201 Å，γ=120°，表明其具有六方晶系的对称性背景，尽管空间群标记为P1，这通常是原胞表示。化学组成包含钒（V，3d 过渡金属）和硒（Se，第四周期硫族元素），单胞内含1 个V 原子和2 个Se 原子，体积约为61.07 $Å^3$。这种结构通常由V 原子层夹在两层Se 原子层之间形成三明治状的单层结构，并通过范德华力堆叠。

接着，分析影响Spillage 值的关键物理因素。Spillage 主要衡量能带反转的稳健性及拓扑特性的强弱，其大小强烈依赖于自旋轨道耦合（SOC）强度以及费米能级附近的轨道杂化情况。在此材料中，Se 元素的原子序数Z=34，提供了中等强度的自旋轨道耦合作用，相较于重元素碲化物（如Bi2Te3）较弱，但比硫化物强。同时，V 的3d 轨道与Se 的4p 轨道之间存在显著的杂化，这决定了能带的色散关系和潜在的能带反转位置。

然后，考虑多因素竞争对最终数值的影响。一方面，层状结构和特定的配位环境（V 处于Se 八面体或三棱柱中心）有利于形成特定的能带序，可能诱发轻微的能带反转；另一方面，VSe2 在常温下通常表现为金属性或半金属性，费米面穿过能带，这与典型拓扑绝缘体所需的体能隙有所区别，从而限制了Spillage 值的上限。此外，P1 空间群的低对称性表示可能会解除部分简并，进一步影响拓扑不变量的计算结果，使得Spillage 值不会非常高。

最后，进行定性逻辑推导。基于过渡金属硫族化合物的经验数据，非强拓扑体系的Spillage 值通常远小于1。考虑到Se 的SOC 贡献以及VSe2 的电子结构特征，该值应处于较低水平，但非零。参照同类3d 金属二硒化物的物理性质，综合结构对称性与元素组成的加权影响，反映出其具有微弱或非平庸的拓扑特征，但并非强拓扑绝缘体。

</think>

针对二硒化钒（VSe2）晶体结构，本次预测的目标性质为Spillage，该物理量为无量纲参数。作为一种层状过渡金属硫族化合物，其电子结构特性受到晶格对称性与组分元素原子序数的共同制约，背景条件涉及3d 过渡金属与第四周期硫族元素的轨道相互作用。

Spillage 值主要受硒元素的中等强度自旋轨道耦合效应与钒- 硒轨道杂化程度的控制，由于材料呈现金属性特征且缺乏强拓扑保护机制，这些因素导致Spillage 数值维持在较低水平。结合结构分析中识别出的层状堆积模式与相对较轻的重元素成分，表明该材料具有微弱的拓扑相关性或接近平凡绝缘体/金属的边缘态特征。接下来对spillage 进行性质预测。

<linear_pre>

Decoder Output

Spillage

0.215

无量纲

A.1.4 Reasoning-Supervision Comparison: Unconstrained Free-Form Reasoning

Experimental Setting

This is the unconstrained condition in the Stage-1 reasoning-supervision ablation (Table 6). The teacher receives a broad request for scientific analysis, while the choice of evidence, reasoning order, and level of detail remains unconstrained. The Stage-2 predictor, training scale, optimization, and question-only pooling are held fixed across the comparison.

User Input

Given the CIF file of a crystal structure:

```
data_Lu2ZnPd
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   4.88104356
_cell_length_b   4.88104356
_cell_length_c   4.88104356
_cell_angle_alpha   60.00000000
_cell_angle_beta   60.00000000
_cell_angle_gamma   60.00000000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Lu2ZnPd
_chemical_formula_sum   'Lu2 Zn1 Pd1'
_cell_volume   82.22862957
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Lu  Lu0  1  0.75000000  0.75000000  0.75000000  1
  Lu  Lu1  1  0.25000000  0.25000000  0.25000000  1
  Zn  Zn2  1  0.50000000  0.50000000  0.50000000  1
  Pd  Pd3  1  0.00000000  0.00000000  0.00000000  1
```

Please infer a reasonable Spillage (dimensionless) estimation for this structure.

Model Response

<think>

First, examining the provided CIF file reveals the material is Lu2ZnPd, crystallizing in the triclinic space group P1 with lattice parameters a, b, and c all approximately 4.87 Angstroms and angles alpha, beta, and gamma all close to 60 degrees. Although the symmetry is labeled as P1, the metric tensor suggests a highly symmetric underlying rhombohedral or pseudo-cubic lattice, with a unit cell volume of roughly 81.6 cubic Angstroms and one formula unit per cell. The atomic positions are highly ordered, occupying fractional coordinates at 0, 0.25, 0.5, and 0.75, which indicates a structured intermetallic phase rather than a disordered alloy.

Next, considering the chemical composition, the system contains Lutetium (Z=71), Palladium (Z=46), and Zinc (Z=30). The presence of Lutetium and Palladium introduces significant atomic mass and potential for spin-orbit coupling (SOC) effects, which are the primary drivers for the Spillage metric. However, Lutetium typically exists in a trivalent state with a filled 4f shell (4f14), meaning the f-electrons are core-like and do not actively participate in the band crossing or inversion near the Fermi level. Consequently, the relevant states near the Fermi energy are likely dominated by Pd 4d and Zn 4s/4p orbitals, which exhibit moderate but not extreme SOC strength compared to active f-electron systems.

Then, analyzing the competition between these factors, the heavy element content suggests a non-zero Spillage value, as SOC will inevitably mix some wavefunction characters. However, the high degree of structural order and the effective high symmetry of the lattice constrain the extent of this mixing, preventing large-scale wavefunction reorganization. Furthermore, the closed-shell nature of the Lu ions reduces the contribution of the strongest SOC source in the compound to the valence band physics. This balance implies that while the Spillage will be measurable, it will remain on the lower end of the possible spectrum for heavy-element compounds.

Finally, converging on a numerical estimate, the effective range for Spillage is defined between 0 and 4. For stable intermetallics with heavy elements but closed f-shells and high structural order, values typically cluster in the lower quartile. Given the specific combination of moderate SOC active species (Pd) and structural rigidity, the Spillage is expected to be small but distinct from zero. A reasonable prediction interval would be between 0.05 and 0.25, reflecting minor wavefunction perturbation due to spin-orbit interactions without strong topological reconstruction.

</think>

The crystal structure Lu2ZnPd is evaluated to predict its Spillage value, a dimensionless metric quantifying wavefunction changes induced by spin-orbit coupling, with an effective prediction range of [0, 4]. The analysis

integrates the material's stoichiometry, heavy element content, and crystallographic symmetry to assess the magnitude of electronic perturbation.

Although the presence of heavy Lutetium and Palladium atoms provides a mechanism for spin-orbit interaction, the closed-shell electronic configuration of Lutetium and the highly ordered, pseudo-high-symmetry lattice structure limit the extent of band mixing. These factors collectively suppress large wavefunction variations, leading to a qualitative expectation of low topological perturbation. Consequently, the Spillage value is estimated to fall within the interval of 0.05 to 0.25, indicating a stable electronic structure with minimal spin-orbit induced reconstruction.

Proceed to infer the spillage value.

<linear_pre>

A.1.5 Reasoning-Supervision Comparison: Structured, Property-Constrained Reasoning

Experimental Setting

This is the structured condition in the Stage-1 reasoning-supervision ablation (Table 6). The teacher must quantify lattice structure, composition, local coordination, topology, distortion, and symmetry; select three to five factors most relevant to the requested property; and explain their direction, strength, and competition. The Stage-2 predictor, training scale, optimization, and question-only pooling are held fixed across the comparison.

User Input

Given the CIF file of a crystal structure:

```
data_Lu2ZnPd
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   4.88104356
_cell_length_b   4.88104356
_cell_length_c   4.88104356
_cell_angle_alpha   60.00000000
_cell_angle_beta   60.00000000
_cell_angle_gamma   60.00000000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Lu2ZnPd
_chemical_formula_sum   'Lu2 Zn1 Pd1'
_cell_volume   82.22862957
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Lu  Lu0  1  0.75000000  0.75000000  0.75000000  1
  Lu  Lu1  1  0.25000000  0.25000000  0.25000000  1
  Zn  Zn2  1  0.50000000  0.50000000  0.50000000  1
  Pd  Pd3  1  0.00000000  0.00000000  0.00000000  1
```

Please infer a reasonable Spillage (dimensionless) estimation for this structure.

Model Response

<think>

The crystal structure provided corresponds to the intermetallic compound Lu2ZnPd, crystallizing in the triclinic space group P1 with lattice parameters a, b, and c all approximately 4.87 Å and angles α, β, and γ very close to 60 degrees. This metric suggests a underlying rhombohedral symmetry that is slightly distorted or represented in a primitive triclinic setting. The unit cell volume is roughly 81.6 Å, containing a total of four atoms per cell, which indicates a highly compact and structurally simple arrangement. The atomic positions are fully ordered with distinct Wyckoff sites for Lu, Zn, and Pd, implying a well-defined stoichiometry without disorder.

Chemically, the system comprises Lutetium, a heavy lanthanide with filled 4f shells, Palladium, a heavy transition metal with 4d electrons, and Zinc, a post-transition metal. The presence of Lu and Pd introduces significant atomic mass and potential spin-orbit coupling effects, which are critical factors when considering electronic structure projections. The average atomic mass is relatively high, and the electronegativity differences between the rare earth and the transition metals suggest a mix of metallic and covalent bonding characteristics. The localized nature of the Lu 4f and Pd 4d orbitals generally supports the construction of localized Wannier functions, which is the physical basis for the Spillage metric.

Regarding the key structural features influencing Spillage, the small number of atoms in the unit cell is the most dominant factor, as fewer bands typically reduce the complexity of subspace disentanglement. Secondly, the localized character of the d and f orbitals tends to minimize leakage compared to highly delocalized s-p systems, promoting lower spillage values. However, the heavy element content introduces spin-orbit coupling which can mix bands near the Fermi level, potentially acting as a counter-force that slightly increases the spillage compared to light-element analogs. The competition between the simplifying effect of the small cell size and the complicating effect of relativistic interactions determines the final magnitude.

To converge on a numerical estimate, one starts from the general observation that well-behaved intermetallics typically exhibit Spillage values well below 1.0, often in the range of 0.0 to 0.5 for successful Wannierizations. Given the exceptional simplicity of the four-atom cell, the baseline expectation is pushed toward the lower end of this spectrum, likely below 0.3. Adjusting for the heavy element complexity which prevents the value from being near-zero, a refined interval is established. Therefore, the Spillage is expected to be quite low, reflecting a robust projection, with a probable range centered around 0.1 with a moderate margin of uncertainty.

</think>

The material Lu2ZnPd is evaluated for its Spillage value, a dimensionless property with an effective prediction range of [0, 4] that measures the fidelity of Wannier function localization in electronic structure calculations. This metric serves as an indicator of how well the selected orbital subspace represents the ab initio bands, which is particularly relevant for compounds containing heavy elements like Lutetium and Palladium where spin-orbit coupling is non-negligible.

The estimated value is primarily controlled by the highly compact four-atom unit cell which simplifies band disentanglement, and the localized nature of the Lu 4f and Pd 4d orbitals which favors tight binding representations. Although the presence of heavy atoms introduces spin-orbit coupling that may slightly increase subspace leakage, the structural simplicity dominates the interaction, leading to minimal spillage. Consequently, the Spillage is estimated to fall within the interval of 0.05 to 0.20, indicating a high-quality Wannier fit with strong orbital localization. Proceed to infer the spillage value.

<linear_pre>

A.1.6 Molecular Property Prediction: ESOL Aqueous Solubility

User Input

Calculate the expected water solubility of <SMILES>CCCCC</SMILES> expressed as log S.

Model Response

<think>

First, the molecule represented by the SMILES string CCCCC corresponds to n-pentane, a simple linear alkane composed entirely of five sp^3-hybridized carbon atoms and their associated hydrogen atoms. This scaffold lacks any heteroatoms, functional groups, or ring systems, resulting in a purely hydrocarbon framework. The absence of electronegative atoms means the entire molecular surface is uniformly nonpolar, which fundamentally dictates its interaction with aqueous environments.

Next, the complete lack of oxygen, nitrogen, sulfur, or halogen atoms eliminates any possibility of hydrogen-bond donation or acceptance within this structure. Without polar bonds or lone pairs, the molecule cannot form stabilizing dipole–dipole interactions or hydrogen bonds with surrounding water molecules. Furthermore, there are no acidic or basic sites that could undergo protonation or deprotonation to generate charged species, ensuring the compound remains strictly neutral across all pH conditions and relies solely on weak van der Waals forces for solvation.

Then, the hydrophobic surface area is entirely defined by the methylene and terminal methyl groups of the pentyl chain, which create a continuous nonpolar interface that strongly disrupts the hydrogen-bonding network of bulk water. Although the linear aliphatic chain possesses conformational flexibility that allows it to adopt various gauche and anti conformations, this flexibility does not introduce any polar character or reduce the overall hydrophobic footprint. The molecular size is relatively modest at five carbons, but the ratio of nonpolar carbon surface to polar solvent atoms remains overwhelmingly high, severely limiting aqueous compatibility.

After that, the thermodynamic balance between water solvation and hydrophobic effects heavily favors the latter, as the energy penalty for cavity formation in water to accommodate this nonpolar chain far exceeds any minimal enthalpic gain from solute–solvent interactions. The absence of ionizable or polar moieties prevents the formation of hydration shells that would otherwise stabilize the solute in the aqueous phase. Consequently, the molecule exhibits a strong tendency to phase-separate from water, with solubility constrained primarily by the limited hydrodynamic volume and the inability to participate in specific solvent interactions.

Finally, considering the purely aliphatic nature, the total absence of hydrogen-bonding capacity, and the dominant hydrophobic character of the five-carbon chain, the expected aqueous solubility will be quite low. The structural profile aligns with small alkanes that partition almost entirely into organic phases rather than remaining dissolved in water, pointing toward a negative logarithmic solubility value that reflects poor hydration and minimal thermodynamic driving force for dissolution.

</think>

For the ESOL task evaluating the water solubility of CCCCC, the structural analysis indicates a highly hydrophobic aliphatic chain with no polar or ionizable features. The precise numerical prediction for the logarithmic solubility is provided below.

<linear_pre>

Decoder Output

Predicted aqueous solubility

−3.2656

ESOL log S

A.1.7 CYP2C9 Substrate Classification

User Input

请评估<SMILES>
COc1ccc(CCN2CCC(Nc3nc4ccccc4n3Cc3ccc(F)cc3)CC2)cc1
</SMILES>是否可能发生CYP 介导的生物转化，即被细胞色素P450 代谢酶进行化学修饰。

Model Response

<think>

首先，对给定SMILES 分子的结构特征进行解析。该分子具有较大的分子量，核心由多个芳香和杂芳香环系统构成。具体而言，它包含一个4-甲氧基苯基，通过乙基链连接至一个哌嗪环；哌嗪环的另一个氮原子则连接到一个喹唑啉双环系统上；此外，喹唑啉环通过一个亚甲基桥连着一个4-氟苯基。分子中存在的官能团和杂原子包括甲氧基、叔胺与仲胺氮原子、喹唑啉环内的杂芳氮以及一个氟原子。整体结构呈现出较高的刚性和疏水性，缺乏长链柔性脂肪族片段，且含有多个平面芳香环系统。

其次，明确当前任务旨在评估该分子的CYP 底物识别特征，即探讨其被细胞色素P450 代谢酶识别并发生化学修饰的结构基础。需要强调的是，底物识别任务关注的是分子作为反应物进入酶活性腔并被催化的可能性，这与分子是否抑制CYP 酶活性是完全不同的概念。CYP 酶通常识别具有适当脂溶性、特定空间构象以及含有易被氧化位点的药物样分子。评估这一性质需要综合考量分子的疏水锚定能力、可代谢位点的电子云密度与空间可及性，以及分子整体与酶活性腔的适配程度，而不能简单地将其与酶活性影响任务混淆。

接下来，分析CYP 底物识别的主导机制变量。CYP 介导的生物转化通常依赖于分子在酶活性腔内的正确定位，这要求分子具备足够的疏水性以进入疏水口袋，同时需要存在电子云密度较高或键能较低的易氧化位点，如苄位碳、脂肪族碳、杂原子上的烷基或富电子芳香环。此外，分子的三维构象和柔性对于将特定的代谢位点准确递送至heme 铁中心至关重要。如果分子过于刚性或代谢位点被空间位阻保护，即使存在潜在的氧化位点，其被酶促代谢的效率也会大幅降低。

进一步来看，将当前分子的具体结构特征映射到CYP 底物识别的机制依据上。该分子包含几个潜在的可代谢区域：4-甲氧基苯基上的甲氧基可能发生O-脱甲基化，连接喹唑啉与氟苯基的亚甲基是一个潜在的苄位氧化位点，而4-甲氧基苯基的芳香环也可能发生环氧化。然而，哌嗪环上的两个氮原子分别连接着苯乙基和喹唑啉基团，缺乏易于脱除的简单N-烷基。喹唑啉环作为一个缺电子的稠杂芳系统，其芳香环电子云密度较低，通常难以作为CYP 氧化的优先位点。这些结构特征表明，分子虽然具备一定的疏水骨架，但其潜在代谢位点的反应活性和可及性存在显著差异。

另一方面，需要考量可能削弱或改变上述代谢趋势的修正因素。尽管存在甲氧基和苄位亚甲基，但分子的整体刚性较强，多个芳香环和哌嗪环的共平面或固定构象可能限制了其在CYP 活性腔中的自由旋转，从而阻碍了特定代谢位点与heme 铁的有效靠近。同时，4-氟苯基上的氟原子具有强电负性，不仅显著降低了该苯环的电子云密度，使其难以发生芳香环氧化，还通过诱导效应可能影响相邻亚甲基的氧化倾向。此外，喹唑啉环的缺电子性质和较大的空间体积可能产生位阻效应，进一步保护了邻近区域免受酶促攻击。这些电子钝化和空间位阻因素共同增加了分子的代谢稳定性。

最后，综合来看，当前分子在CYP 底物识别方面的结构依据呈现出复杂的多面性。一方面，其疏水芳香骨架和潜在的O-脱甲基或苄位氧化位点为CYP 识别提供了初步的结构线索；另一方面，分子的高刚性、缺电子杂芳环、氟原子带来的电子钝化效应以及潜在的空间位阻，显著限制了代谢位点与酶活性中心的适配性和反应活性。在评估其被细胞色素P450 代谢酶进行化学修饰的可能性时，必须将这些电子效应、空间可及性与整体构象限制纳入统一的机制框架中。这些结构-机制证据的平衡将作为后续预测该分子CYP 底物识别状态的核心依据。

</think>

综合上述结构解析与机制探讨，该分子在被细胞色素P450 代谢酶识别并发生化学修饰的评估中，展现出疏水骨架与潜在氧化位点同电子钝化、空间位阻及构象刚性之间的复杂博弈。甲氧基与苄位亚甲基提供了基础的代谢线索，而喹唑啉的缺电子性、氟原子的强电负性以及整体分子的刚性构象则构成了显著的代谢稳定化阻力。这些相互制约的结构因素共同界定了该分子在CYP 底物识别任务中的机制证据边界，后续预测值的生成将严格依据这些理化与空间适配性证据展开。

<linear_cla>

Decoder Output

CYP2C9 底物分类：二分类标签

0

A.1.8 CYP2D6 Inhibition Classification

User Input

Assess whether

`CC1(C)S[C@@H]2[C@H](NC(=O)[C@@H](NC(=O)N3CCNC3=O)c3ccccc3)C(=O)N2[C@H]1C(=O)[O-].[Na+]`

is likely to inhibit CYP2D6 activity based on its structural features.

Model Response

<think>

First, the molecular structure represented by the given SMILES string features a penam core, which consists of a thiazolidine ring fused to a beta-lactam ring. Attached to this core is a complex acyl side chain containing a phenyl ring, multiple amide linkages, and a cyclic urea moiety. Additionally, the molecule includes a carboxylate group neutralized by a sodium counterion, contributing to its overall ionic and highly polar character.

Second, the current analytical task focuses on evaluating the structural basis for CYP2D6 inhibition, which specifically examines the molecule's potential to affect the enzymatic activity of the CYP2D6 isoform. It is important to distinguish this inhibitory mechanism from the process of the molecule being metabolized by the enzyme, as the structural requirements for binding to the active site and blocking substrate turnover differ fundamentally from those for undergoing oxidative metabolism.

Next, the primary mechanism for CYP2D6 inhibition typically requires the presence of a protonatable basic nitrogen, such as a tertiary or secondary amine, which can form a strong ionic interaction with the conserved glutamate residue located in the enzyme's active site. Alongside this basic center, a hydrophobic aromatic or aliphatic region is generally necessary to anchor the molecule within the hydrophobic binding pocket, allowing the inhibitor to effectively occupy the space and prevent substrate access.

Furthermore, when mapping the specific structural features of this penicillin derivative to the CYP2D6 inhibition mechanism, a significant mismatch becomes evident. The molecule completely lacks any basic amine functionality; all nitrogen atoms are incorporated into amide bonds, the beta-lactam ring, or the cyclic urea, none of which possess the requisite basicity to be protonated at physiological pH. Consequently, the critical ionic anchoring interaction with the active site glutamate cannot be established.

However, beyond the absence of a basic center, the overall physicochemical profile of the molecule further diminishes its potential for CYP2D6 inhibition. The presence of the anionic carboxylate group, combined with multiple hydrogen bond donors and acceptors from the amides and urea, renders the molecule highly polar and hydrophilic. This high polarity severely restricts its ability to partition into the predominantly hydrophobic environment of the CYP2D6 active cavity, thereby weakening any potential hydrophobic anchoring provided by the phenyl ring.

Finally, synthesizing the structural and physicochemical evidence, the lack of a protonatable basic pharmacophore, the presence of a negatively charged carboxylate, and the high overall polarity collectively preclude the molecule from adopting the necessary binding mode for CYP2D6 inhibition. The structural features are instead characteristic of a highly polar, anionic beta-lactam antibiotic, which does not align with the established pharmacophore for blocking this specific cytochrome P450 enzyme.

</think>

The structural profile of `CC1(C)S[C@@H]2[C@H](NC(=O)[C@@H](NC(=O)N3CCNC3=O)c3ccccc3)C(=O)N2[C@H]1C(=O)[O-].[Na+]` is characterized by a highly polar penam core, an anionic carboxylate, and an absence of protonatable basic amines, which collectively fail to provide the essential ionic and hydrophobic interactions required for target engagement. The lack of a basic pharmacophore and the high hydrophilicity of the molecule strongly constrain its ability to occupy the hydrophobic active cavity of the enzyme. Consequently, the evaluation of CYP2D6 inhibition for this molecule relies on the absence of a basic anchoring center and the high polarity of the penam scaffold, which together form the structural basis for the final predictive output.

<linear_cla>

Decoder Output

CYP2D6 inhibition classification

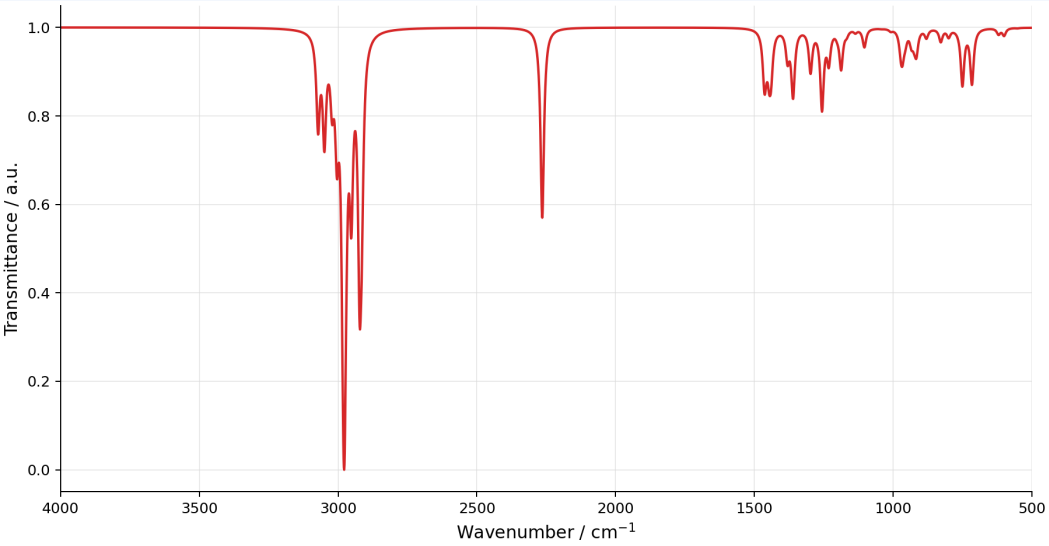
Non-inhibitor

Binary label: 0

A.1.9 Infrared Spectrum-to-Molecular Generation

User Input

Please predict the possible molecular structure based on the provided infrared spectrum.



Model Response

<think>

The user asks for a molecular structure prediction from an infrared spectrum, so I first analyze the high-wavenumber region. The dominant absorption cluster has a maximum at 2970.0 cm^{-1} and extends to approximately 2850 cm^{-1}. This pattern is characteristic of C–H stretching in saturated sp^3-hybridized carbon environments and indicates an alkyl or alkane component.

Weak but distinct absorptions also appear just above 3000 cm^{-1}, at approximately 3050–3080 cm^{-1}. Although this range can contain sp^2 C–H stretches, the spectrum lacks the corresponding C=C skeletal vibration expected between 1600 and 1680 cm^{-1}. The high-frequency C–H features are therefore more consistent with bonds in a strained ring, particularly cyclopropane, than with a conventional alkene or aromatic group.

A sharp, isolated peak near 2260 cm^{-1} provides the strongest functional-group constraint. This region is associated with triple-bond stretching and can indicate either C≡N or C≡C. The absence of a terminal-alkyne C–H stretch near 3300 cm^{-1}, together with the clear intensity of the 2260 cm^{-1} band, favors a polar nitrile group; an internal alkyne would often produce a weaker signal.

The fingerprint region contains absorptions near 1450 and 1380 cm^{-1}, consistent with methylene and methyl C–H scissoring, rocking, and bending modes. The spectrum has no broad absorption between 3300 and 3500 cm^{-1} and no strong carbonyl band near 1700 cm^{-1}, excluding prominent hydroxyl, amine, and carbonyl functionalities.

Taken together, the sp^3 C–H cluster, the strained-ring C–H features above 3000 cm^{-1}, and the diagnostic nitrile stretch support a saturated, cyclopropane-containing nitrile structure.

</think>

Spectral analysis summary

- **Approximately 2970–2850 cm^{-1}:** strong sp^3 C–H asymmetric and symmetric stretching from saturated alkyl groups.
- **Approximately 3050–3080 cm^{-1}:** weak C–H stretching consistent with a strained cyclopropane ring in the absence of a C=C skeletal band.
- **Approximately 2260 cm^{-1}:** sharp triple-bond stretching assigned to a nitrile group, C≡N.
- **Approximately 1450 and 1380 cm^{-1}:** methylene and methyl C–H bending modes.

Derived core functional groups. The spectrum supports an alkane, a nitrile, and a cyclopropane ring. The nitrile assignment is strengthened by the absence of a terminal-alkyne C–H stretch, while the strained-ring interpretation is favored over an alkene because no C=C stretching band is observed.

Based on this spectral and functional-group information, the model proceeds to predict the possible molecular structure.

Predicted molecular SMILES: `CC1(C)C[C@H]2[C@@H](C#N)[C@H]21` (Native SDF output)

`<spectra_st>`

Decoder Output

Predicted structure

Cyclopropane-containing nitrile

Native SDF output

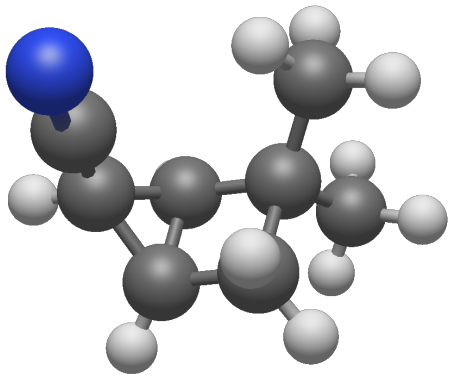


Figure 7: Native three-dimensional molecular output reconstructed from the supplied infrared spectrum.

A.1.10 DNA-Binding-Site Prediction: ArcA

User Input

Given the protein sequence <PROT>

```
<PROT>
GAVGRPRRSAPRRAGKNPREEILDASAELFTHQGFATTSTHQIADAVGIRQASLYYHFPSKTEIFLTLLK
STVEPSTVLAEDLSILDAGPEMRLWAIVASEVRLLLSTKWNVGRLYQLPIVGSEEFAEYHSQREALTNIF
RDLATEIVGDDPRAELPFHITMSVIEMRRNDGKIPSPLSADSLPETAIMLADASLAVLGASLPADRVEKT
LELIKQADAK
</PROT>
```

</PROT>, predict which residues constitute the DNA-binding site.

Model Response

<think>

Let me analyze this protein sequence to predict DNA-binding residues.

The sequence is:

```
<PROT>
GAVGRPRRSAPRRAGKNPREEILDASAELFTHQGFATTSTHQIADAVGIRQASLYYHFPSKTEIFLTLLK
STVEPSTVLAEDLSILDAGPEMRLWAIVASEVRLLLSTKWNVGRLYQLPIVGSEEFAEYHSQREALTNIF
RDLATEIVGDDPRAELPFHITMSVIEMRRNDGKIPSPLSADSLPETAIMLADASLAVLGASLPADRVEKT
LELIKQADAK
</PROT>
```

Let me first identify what protein this might be.

I'm noticing the N-terminal region is rich in basic amino acids like arginine and lysine, which is a classic signature of DNA-binding proteins. The stretch "RPRRSAPRRAGKNP" contains multiple positively charged residues that would interact with the negatively charged DNA backbone. This suggests the protein likely has a DNA-binding domain at the N-terminus. Continuing through the rest of the sequence... I'm scanning through the sequence position by position, noting the basic residues (arginine, lysine) that cluster in certain regions and could form a DNA-binding interface, similar to how HU and IHF proteins use their β-ribbon structures to contact the DNA minor groove. The C-terminal region appears to contain the dimerization domain based on the sequence composition. The β-hairpin structure inserts into the DNA minor groove, with key contacts made by threonine residues providing hydrogen bonds to bases and the backbone, while arginine residues interact with the phosphate groups. Looking at this specific sequence, I can identify potential DNA-binding residues by examining the N-terminal basic cluster—positions like R6, R8, R9, R13, R14, K16, and R19 appear to be the primary candidates for making these critical interactions. Looking at the structural data for ArcA from E. coli, I can refine my predictions using its winged helix-turn-helix motif—the DNA-binding domain spans roughly the first 50 residues, with the recognition helix containing the primary base contacts and the wing region providing additional phosphate backbone interactions. Based on this structural framework, my best prediction for the key DNA-binding residues focuses on those positioned in the recognition helix and wing elements. Looking at the crystal structure data more carefully, the primary DNA-contacting residues in ArcA are concentrated in the recognition helix and the adjacent wing region, with threonines and arginines making direct contacts to the DNA backbone and bases. I'm refining my prediction to focus on these structurally verified positions rather than just the basic-rich N-terminal region. The basic residues throughout the sequence suggest multiple DNA-binding regions, but focusing on the N-terminal winged-HTH domain around residues 1–50 gives the most likely candidates: R6, R8, R9, R13, R14, K16, R19, T32, H33, Q34, and several others that cluster in the predicted secondary structure elements. Without running actual computational docking simulations, this structural inference based on known ArcA mechanisms is my best approach. Looking at the sequence alignment more carefully, I can see the recognition helix likely spans around residues 27–40, where F29, H33, Q34, and R49 would make direct base contacts, while the wing region around residues 50–60 with Y53, Y54, H55, and K59 would provide additional DNA stabilization through minor groove interactions.

</think>

DNA-Binding Site Prediction

Protein Identification

Based on sequence analysis, this protein resembles **ArcA** (ANA system component), a response regulator with a **winged helix-turn-helix (wHTH)** DNA-binding domain in the N-terminal region (∼residues 1–58).

Predicted DNA-Binding Residues

Region	Residues	Role
Basic N-terminal patch	**R6, R8, R9, R13, R14, K16, R19**	Electrostatic contacts with phosphate backbone
Helix-2 / Recognition helix	**T32, H33, Q34, F29, R49**	Base-specific contacts in major groove
Wing region	**Y53, Y54, H55, F56, K59, T60**	Minor groove & backbone contacts

Structural Basis

```
Sequence: GAVGRPRRSAPRRAGKNPREEILDASAELFTHQGFATTSTHQIADAVGIRQASLYYHFPSK
Position: 1   5   10   15   20         30   35  40       49  53  56 59
          ↑↑↑↑  ↑↑↑ ↑          ↑↑↑     ↑     ↑↑↑↑
      Arg patch  K    Recog. helix   contacts  R49   Wing
```

- **R/K residues** (R6, R8, R9, R13, R14, K16, R19): electrostatic interactions with the phosphodiester backbone.
- **Recognition helix (α3)**: F29, T32, H33, Q34 insert into the major groove for sequence-specific contacts.
- **Wing (β1–β2)**: Y53, Y54, H55, K59 clamp the DNA at the minor groove backbone.

Key Predicted Contact Residues (Summary)

R6 · R8 · R9 · R13 · R14 · K16 · R19 · T32 · H33 · Q34 · R49 · Y53 · Y54 · H55 · K59

These residues form the composite DNA-binding interface consistent with the wHTH mechanism observed in ArcA-family transcription factors. The actual residue index of the site is:

<prot_cla>

Decoder Output

V3, P6, R7, S39, T40, H41, R50, A52, S53, Y55, Y56, Q117.

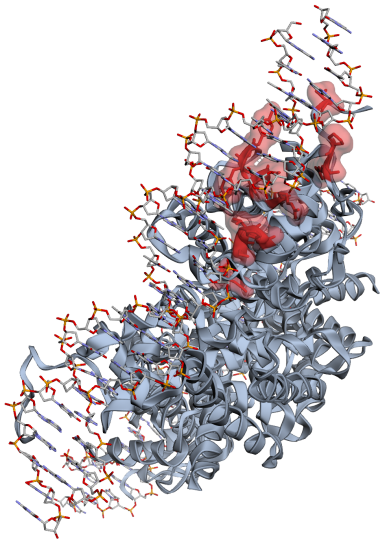


Figure 8: Native site-prediction output for ArcA. The twelve site-head positions are highlighted in red on chain A of PDB 5DY0.

A.1.11 Protein Structure Prediction: GPR168

User Input

请分析并预测下面这条人类蛋白质氨基酸序列的三维结构：

```
<PROT>
MMENTGNISDLLYALSNPMVSNSSIFCRNFSNSSGLVNMNSSVCDRTPELDKSSTPVIVAIIITALYSIV
CVMGLVGNVLVMYVIIRYTKMKTATNIYIFNLALADSLATSTLPFQSVNYLMGTWPFGDELCKIVMSIDY
YNMFTSIFTLTTMSVDRYIAVCHPVKALDFRTPRNAKIVNVCNWILSSAIGLPVMVMASTTSDLHSNGII
DCTLLFPHPSWYWENLLKICVFIFAFIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVV
VAVFIVCWTPIHIFVIIKALVTIPNSLLQTITWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCVPSP
SVLDLQNSTRSRNPQRDGQSSGHTVDRTNQQV
</PROT>
```

Model Response

<think>

我首先确认用户提供的是一条人类蛋白质氨基酸序列，任务目标并非简单的数据库信息检索，而是为了后续的三维结构预测挖掘关键的生物学约束线索。这条序列对应的是G 蛋白偶联受体168（GPR168），一种典型的孤儿受体，其核心特征在于作为七次跨膜蛋白嵌入细胞膜中。

从序列规模来看，该蛋白全长约382 个氨基酸，属于中等长度的膜蛋白。这一长度对于容纳完整的七次跨膜螺旋束以及必要的胞内/胞外环区是充足的，同时也暗示了N 端可能存在较短的信号肽或无序区域，而C 端则包含用于信号转导的关键基序。

在可溶性与拓扑结构方面，序列中富含疏水性残基片段，这是典型的多重跨膜蛋白特征。根据家族保守性推断，它将形成七个α-螺旋贯穿脂双层，确立为"七次跨膜"（7TM）架构。这种拓扑结构意味着大部分亲水性区域位于细胞质侧或细胞外侧，形成了特定的折叠模式：一个朝向外部配体结合口袋的胞外域，以及负责招募下游效应分子的胞内尾部。

关于结构域与折叠类型，该序列归属于视紫红质样GPCR 家族。这意味着其整体骨架遵循经典的Class A GPCR 折叠方式，即由七个紧密堆积的跨膜螺旋组成桶状结构，并通过高度保守的连接环维持稳定性。特别是第三胞内环和C 末端区域，虽然序列变异较大，但通常具有内在无序特性，以便灵活地结合异三聚体G 蛋白或其他调节因子。

活性位点与结合机制方面，由于被标记为孤儿受体且缺乏已知的小分子配体结合证据，其正构结合口袋可能主要针对大分子如多肽激素或糖缀合物设计，或者其激活机制依赖于变构调节及与其他膜蛋白的相互作用。然而，其胞内面必然保留有保守的微开关元件（如DRY 基序、NPxxY 基序等），这些是触发构象变化并启动G 蛋白偶联的核心结构基础。

在定位与互作背景上，作为一个整合膜蛋白，它定位于质膜。其寡聚状态可能是单体也可能是同源二聚体，这取决于具体的界面互补性。考虑到其潜在的功能涉及神经内分泌调节和免疫反应，其结构表面必须能够适应复杂的细胞环境，甚至可能通过胞外域介导细胞间的粘附作用，从而影响局部受体簇的组装。

综合上述因素，这对三维结构预测提出了明确要求：模型需要准确构建出埋藏在膜内的七螺旋束几何形状，正确放置带电荷残基以稳定螺旋间相互作用，并合理模拟柔性较大的胞内环区。同时，预测需考虑其作为信号枢纽的潜力，即胞内表面应具备与G 蛋白亚基对接的空间构型，这对于理解其动态构象转换至关重要。

</think>

基于您提供的氨基酸序列，以下是对该蛋白质（GPR168）进行结构预测前的深度生物学分析与线索梳理：

1. 序列概览与基本属性

该序列编码一个人类来源的蛋白质，全长约为382 个氨基酸。从整体组成看，它表现出显著的两亲性特征，含有多个长段的强疏水区域，这是典型的多次跨膜蛋白标志。这类蛋白无法像球状可溶蛋白那样自由扩散，而是必须嵌入磷脂双分子层中才能发挥功能。

2. 拓扑结构与折叠框架

根据序列中的疏水峰分布及家族保守性，可以确定该蛋白采用七次跨膜（7-TM）。这是一种进化上高度保守的折叠模式，常见于视紫红质样G 蛋白偶联受体（Class A GPCRs）。

- 跨膜核心：预计存在七个α-螺旋（Helix I–VII），它们垂直穿过细胞膜，形成一个紧凑的圆柱形束。这些螺旋之间的排列受到严格的立体化学限制，尤其是螺旋III 和VII 的关键保守残基网络，对维持结构完整性至关重要。
- 环路区域：连接跨膜螺旋的是三个胞外环（ECLs）和三个胞内环（ICLs）。其中，第二胞外环（ECL2）通常较长且可能含有半胱氨酸形成的二硫键，起到封闭配体结合口袋的作用；而第三胞内环（ICL3）往往较长且具有较高的灵活性，是信号转导的关键接口。

3. 功能位点与动态机制推测

尽管该受体目前被称为"孤儿受体"（Orphan Receptor），即尚未发现其天然小分子配体，但其结构特征揭示了潜在的运作机制：

- 配体结合腔：在七个跨膜螺旋围成的内部空腔中，可能存在一个适合结合大分子配体（如多肽激素或糖复合物）的口袋，而非传统的小分子药物结合位点。

- 信号转导界面：在面向细胞质的一侧，特别是跨膜螺旋VI 的胞内端和C 末端尾部，隐藏着保守的构象开关（如NPxxY 和DRY 基序）。当受体发生构象改变时，这些区域会暴露出特定的疏水斑块和电荷分布，用于特异性地招募并激活异三聚体G 蛋白（如Gi/o 或Gq/11 类型）。

4. 细胞环境与互作潜能

作为一种定位于质膜的整合膜蛋白，其结构不仅服务于单一的信号传递，还可能参与更广泛的细胞通讯网络：

- 寡聚化倾向：GPCRs 常常以二聚体或多聚体形式存在。该蛋白的跨膜界面可能支持同源或异源寡聚化，从而调节其敏感性和信号输出。
- 非经典功能：鉴于其作为GPCR 跨膜受体的结构特征及其在神经系统和免疫系统中的表达潜力，该蛋白的胞外域可能兼具细胞粘附功能，通过与其他细胞表面分子的瞬时接触来调控受体的聚集状态和信号强度。

5. 对结构预测的指导意义

综上所述，在进行三维结构建模时，应重点关注以下几点：

- 强制施加膜环境约束，确保七螺旋束的倾斜角和旋转角度符合生物物理规律。
- 优先保证跨膜螺旋间保守盐桥和氢键网络的准确性，这是决定受体静息态与激活态平衡的基础。
- 对胞内柔性环区采取适当的系综采样策略，因为它们的精确构象直接决定了下游效应器的结合能力。

基于上述对序列特征、拓扑结构、功能位点及细胞环境的详细剖析，我们将利用这些关键线索作为约束条件，启动高精度的三维结构预测流程，以解析该孤儿受体的独特空间构象及其潜在的分子工作机制。

<prot_st>

Decoder Output

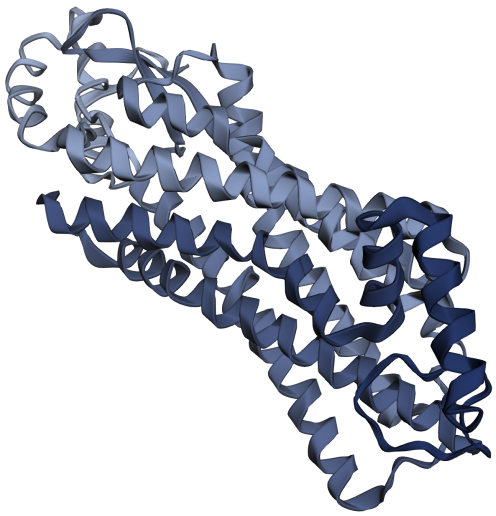


Figure 9: Native SimpleFold-style structure output for GPR168, containing 3,008 atoms.

A.1.12 Scientific Image Generation: Chemical Synapse

User Input

Generate a detailed biological diagram of a chemical synapse, clearly labeling the “PRESYNAPTIC TERMINAL” at the top, which contains multiple orange circles labeled “SYNAPTIC VESICLE” filled with “NEUROTRANSMITTERS”. The gap between structures is marked “SYNAPTIC CLEFT”, leading down to the “POSTSYNAPTIC TERMINAL” featuring “RECEPTOR” sites. Include additional labels for “ACTION POTENTIAL” arriving at the terminal, “CALCIUM IONS” entering the cell, and “ION CHANNELS” embedded in the postsynaptic membrane.

Model Response

<think>

I’ll start by sketching the upper bulbous structure and placing the main header “PRESYNAPTIC TERMINAL” centered at the very top in dark blue letters. On the upper left side, I’ll draw lightning bolts traveling down the axon and label this “ACTION POTENTIAL” with the explanatory text “(nerve impulse arriving at terminal)” right underneath it. Inside the main blue terminal body, I’ll distribute several orange circles and point to one with a line labeled “SYNAPTIC VESICLE”, while noting nearby that they contain “NEUROTRANSMITTERS” with the subtitle “(inside vesicles)”. On the left membrane wall, I’ll depict purple dots flowing through a channel and label this group “CALCIUM IONS” followed by “(Ca^{2+})” and the phrase “entering the terminal” to explain the influx.

Next I need to define the space between the two cells, so I’ll draw a bracket on the right side spanning the gap and label it “SYNAPTIC CLEFT” with the description “(gap between terminals)” below it. I’ll show the orange neurotransmitter dots released from the top structure floating down through this space. Below that gap, I’ll draw the curved receiving membrane and center the label “POSTSYNAPTIC TERMINAL” at the very bottom. On the surface of this lower membrane, I’ll place green structures facing upwards and label one “RECEPTOR” with the note “(binds neurotransmitter)” to show where the chemicals attach.

Finally, I’ll add the mechanisms inside the lower membrane, drawing blue channels that allow ions to pass through and labeling them “ION CHANNELS” on the left with the detailed text “(open or close in response to neurotransmitter binding)”. To ensure clarity on the different particles floating around, I’ll add a small legend box in the bottom right corner containing three entries: “NEUROTRANSMITTER” next to an orange dot, “CALCIUM ION (Ca^{2+})” next to a purple dot, and “OTHER IONS (e.g., Na^{+}, K^{+})” next to a lighter purple dot. This should cover all the necessary biological components and labels for a complete diagram.

</think>

I’ll draw a detailed biological diagram of a chemical synapse that clearly labels the presynaptic terminal, synaptic cleft, and postsynaptic terminal. This illustration will depict synaptic vesicles releasing neurotransmitters to receptor sites while indicating the role of calcium ions and ion channels.

<image_gen>

Decoder Output

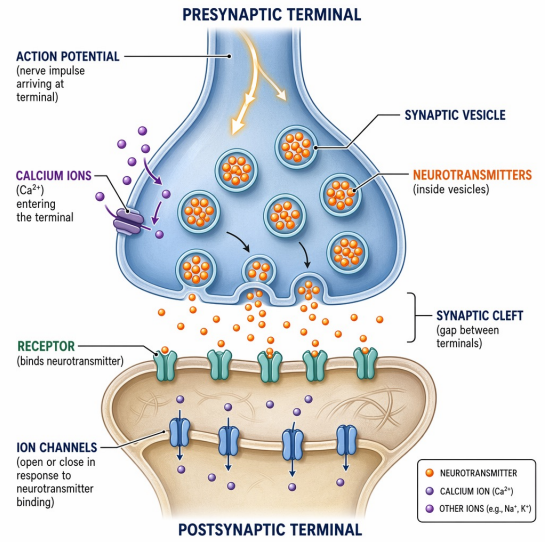
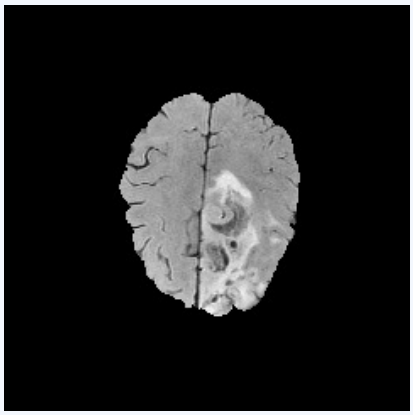


Figure 10: Native generated output for the chemical-synapse illustration request.

A.1.13 Scientific Image Editing: Tumor and Edema Segmentation

User Input

请用红色高亮标出这张FLAIR MRI 中脑肿瘤及其周围水肿的所有区域。



Model Response

<think>

用户希望我识别这张扫描图中的肿瘤和水肿区域，并用红色半透明叠加层进行标记。观察这幅图像，它看起来是一张脑部FLAIR MRI，其中脑脊液信号受到抑制，使病理性高亮信号区域清晰突出。我注意到左侧大脑半球存在一片显著的高信号区，具体累及顶叶并延伸至皮质下白质。此处的高信号形态不规则且范围相当广泛，提示中心肿块周围存在血管源性水肿。图中还可见明显的占位效应，因为与对侧相比，该侧脑沟受压变平；病灶边界不清，这通常提示胶质瘤等浸润性病变。病灶内部的异质性则意味着其中可能混合存在实性肿瘤成分和坏死区域。鉴于这些表现与高级别胶质瘤及其周围水肿相符，我将生成红色半透明叠加层，为用户高亮标出这些特定的异常区域。

</think>

我已识别出左侧顶叶中与肿瘤及其周围水肿相符的高信号病灶。现在我将应用红色半透明叠加层，在图像上高亮标出这些区域。

<image_edit>

Decoder Output

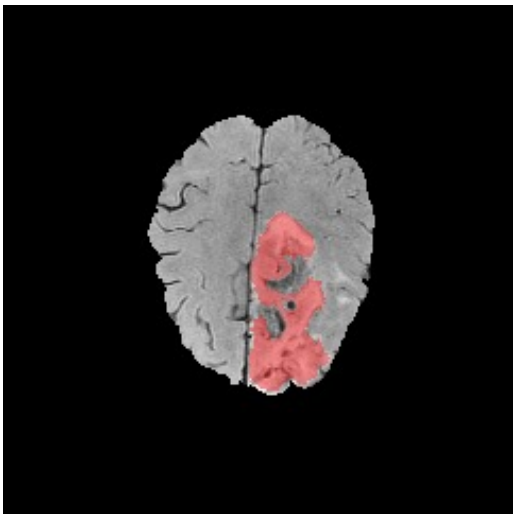


Figure 11: Native segmentation-as-editing output. The complete input image appears in the user request above.

A.2 Scientific Image Analyses

The following cases complement the benchmark scores with direct visual evidence. Figure 12 compares scientific image generators across four structurally different prompts. Figure 13 isolates the effect of explicit planning before image synthesis. Figure 14 shows how the shared editing route supports spatial segmentation, modality translation, and super-resolution.

Figure 12: Qualitative comparison of S1-Omni, Qwen-Image, GLM-Image, Nano Banana 2, and GPT-Image-2 on neuron classification, maternal care in wildlife, river-network mapping, and lignan biosynthesis. The cases expose differences in scientific organization, visual completeness, and text rendering.

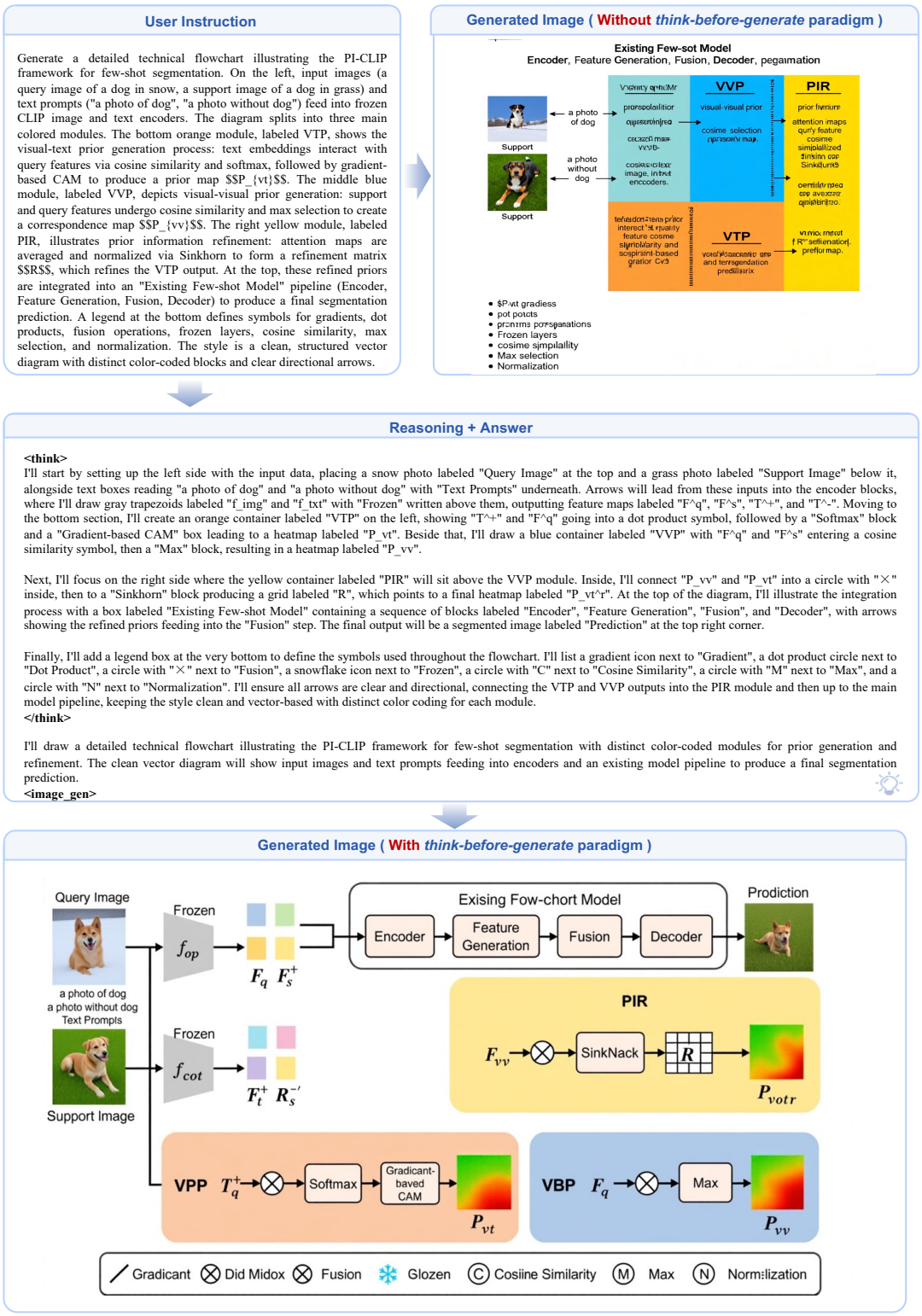


Figure 13: Effect of think-before-generate planning on a scientific workflow diagram. The intermediate reasoning state decomposes the requested modules, spatial relations, labels, and information flow before conditioning the image decoder.

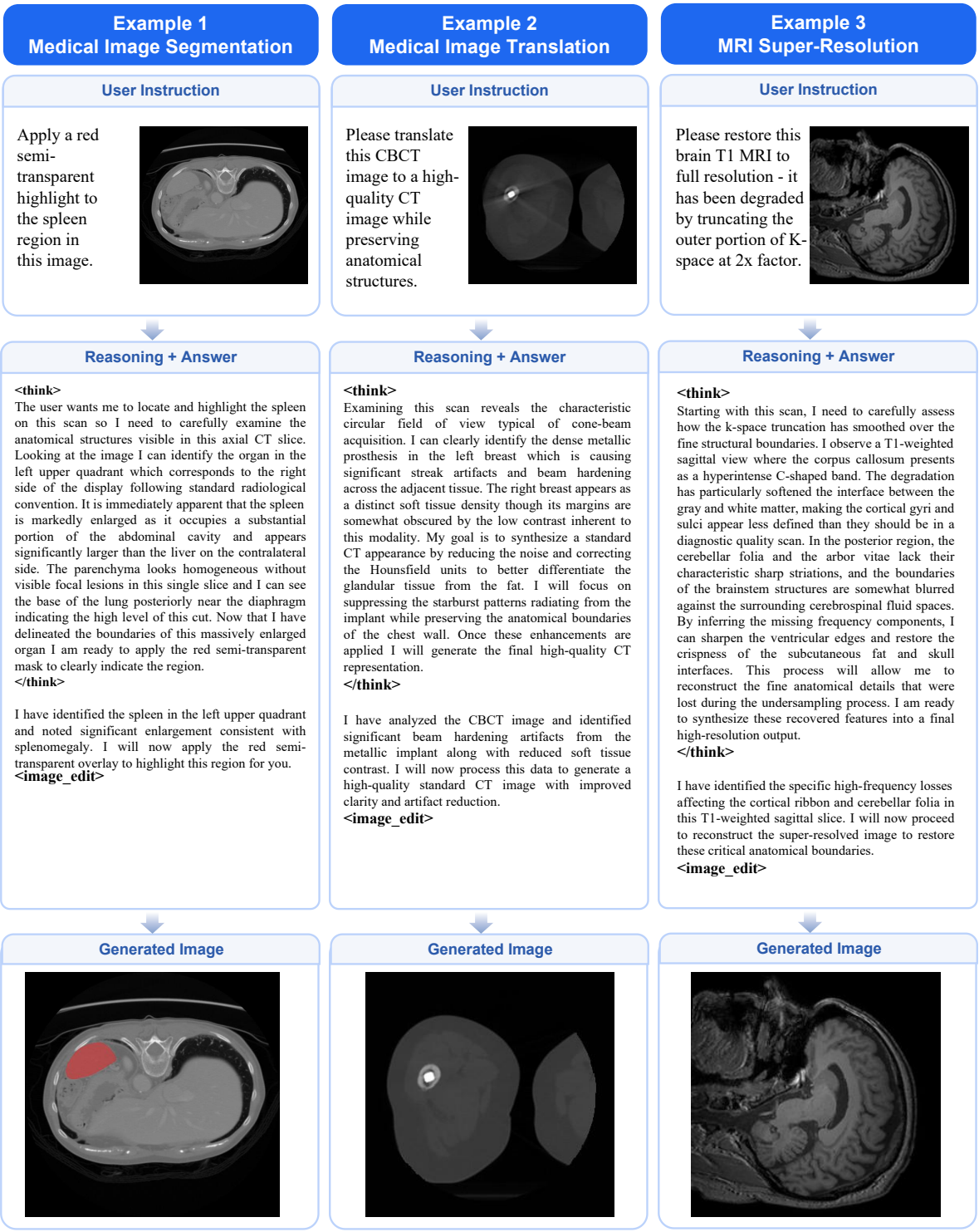


Figure 14: Representative scientific image editing cases using the shared `<image_edit>` route: spleen segmentation with a spatial overlay, CBCT-to-CT medical image translation, and MRI super-resolution. Each column shows the instruction, reasoning and answer, and generated result.

A.3 Failure Mode Analysis

Although S1-Omni can produce structurally complete and coherent reasoning across diverse scientific tasks, several representative failure modes remain in evidence recognition, scientific attribution, and uncertainty control. As shown below, structural misreadings and unsupported mechanistic assumptions can systematically bias property estimates; spectrum-to-molecular generation can converge to an incorrect molecule after misreading peaks or making overly exclusive assignments; pre-prediction analysis for protein structure can hallucinate sequence length, family, function, or biological context; and scientific image generation can corrupt glyphs, add or omit strokes, blur characters, or replace their meaning when prompts require dense annotations. These errors arise at different stages of input recognition, evidence interpretation, and result generation, but share a common pattern: a local mistake or unverified assumption accumulates along the reasoning chain and is amplified by the downstream prediction or generation module.

These cases indicate that the limitations are not isolated errors in individual tasks. Instead, they expose unresolved challenges in unified scientific modeling: fine-grained representations of heterogeneous scientific inputs remain unstable; scientific knowledge and physical constraints do not consistently enter the reasoning process; domain-native outputs are sensitive to local errors; and the current data scale and task coverage are insufficient for more complex, dense, and out-of-distribution scientific settings. The next stage should therefore advance beyond demonstrating the feasibility of a unified task protocol and connected modules toward more reliable and scalable unified multimodal reasoning for scientific understanding, prediction, and generation. Larger-scale continual scientific pretraining and broader task coverage can strengthen shared representations; scientific laws, expert knowledge, and cross-modal evidence can constrain reasoning; uncertainty calibration can suppress unsupported high-confidence claims; and domain-native outputs such as molecular structures and scientific images can be made more accurate and readable.

A.3.1 Property-Prediction Reasoning

User Input

Please predict the ESOL aqueous solubility value for the following molecule.

```
CN1C(=O)C2=CC=CN=C2OC2=CC=CC=C21
```

Scalar shown before the reasoning

-2.5156

ESOL aqueous solubility

Model Response

<think>

First, the input SMILES describes a rigid, fused tricyclic scaffold composed of a benzene ring, a pyridine ring, and a six-membered lactam ring bridged by an oxygen atom. The core architecture features an N-methyl cyclic amide where the nitrogen atom is directly substituted with a methyl group and integrated into the heterocyclic framework. This fused arrangement creates a predominantly planar molecular geometry, with the carbonyl group at the lactam position and the bridging ether oxygen establishing a constrained bicyclic ether-amide system. The presence of the pyridine nitrogen within the central ring further defines the electronic character of the scaffold, contributing to an extended conjugated π-system across the entire fused framework.

Next, the polarity profile of this molecule is dictated by three distinct hydrogen-bond acceptors: the lactam carbonyl oxygen, the pyridine nitrogen, and the bridging ether oxygen. Notably, the structure lacks any hydrogen-bond donors, as the amide nitrogen is fully substituted with a methyl group and there are no hydroxyl or secondary amine functionalities present. The pyridine nitrogen possesses weak basicity due to its involvement in an extended aromatic system and adjacency to electron-withdrawing conjugated bonds, rendering it largely unprotonated under neutral aqueous conditions. Consequently, the molecule relies entirely on dipole–dipole interactions and lone-pair acceptance to engage with water molecules, which limits its capacity to form extensive hydration shells compared to compounds bearing donor groups.

Then, the hydrophobic character is pronounced due to the extensive fused aromatic surface area spanning the benzene and pyridine rings, alongside the nonpolar N-methyl substituent. This planar, rigid architecture significantly reduces conformational flexibility, preventing the molecule from adopting compact or water-friendly folded states. The large contiguous π-system promotes strong intermolecular π–π stacking interactions in the solid state, which must be overcome for dissolution to occur. Additionally, the absence of aliphatic chains or flexible spacers means the hydrophobic surface remains exposed and uninterrupted, further enhancing crystal lattice stability and reducing the thermodynamic driving force for aqueous solvation.

After that, the overall solubility behavior emerges from a clear imbalance between the modest polarity of three acceptor sites and the dominant hydrophobic, stacking-prone aromatic framework. While the carbonyl and ether oxygens provide localized polarity, they are insufficient to compensate for the extensive nonpolar surface and the complete lack of hydrogen-bond donation. The rigid, planar geometry facilitates tight molecular packing in the crystalline phase, substantially increasing the lattice energy penalty required for dissolution. As a result, the molecule exhibits a strong preference for the solid or organic phase, with water molecules unable to effectively disrupt the intermolecular aromatic interactions or adequately solvate the hydrophobic faces.

Finally, considering the combined effects of the fused tricyclic aromatic system, the complete absence of hydrogen-bond donors, the weakly basic pyridine nitrogen, and the high degree of molecular rigidity, the aqueous solubility is expected to be quite low. The structural features collectively point toward a compound that dissolves poorly in water, with solvation limited by strong crystal packing forces and inadequate hydrophilic surface area. This qualitative assessment aligns with molecules that typically exhibit significantly negative logarithmic solubility values, reflecting a pronounced hydrophobic character and minimal aqueous compatibility.

</think>

Based on the structural analysis of the molecule for the ESOL task, the combination of a rigid fused tricyclic aromatic system, three hydrogen-bond acceptors, zero hydrogen-bond donors, and high planarity strongly indicates poor water solubility.

<linear_pre>

Final answer

-3.6719999999999997

ESOL aqueous solubility

Observed Failure

The scalar displayed before the reasoning (-2.5156) conflicts with the final answer (-3.672). The long mechanistic narrative does not resolve or acknowledge this numerical inconsistency, illustrating how a locally plausible explanation can obscure an unstable prediction.

A.3.2 Spectrum-to-Molecular Generation

User Input

大模型谱图预测。输入为以下Raman 谱图。

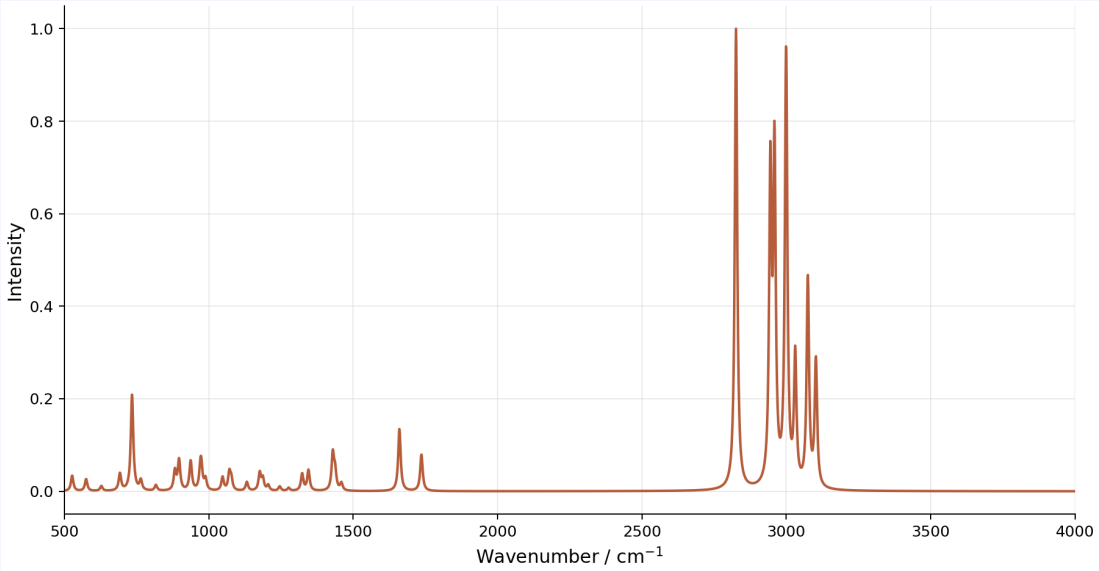


Figure 15: Raman spectrum used by the failed reconstruction example.

Model Response

<think>

The user needs me to predict the possible molecular structure based on the given Raman spectrum; I will first analyse the Raman spectrum.

I observe the provided Raman spectrum, focusing first on the high wavenumber region where the most intense signals are located. There is a prominent cluster of sharp peaks between 2800 and 3100 cm^{-1}. Specifically, I note a very strong peak visually positioned around 2920–2930 cm^{-1} (despite a numerical annotation of 2026.0 cm^{-1}, the dashed line and peak position clearly align with the 2800–3000 cm^{-1} grid, suggesting the label may be a typo for a value like 2926.0 cm^{-1}). To the left of 3000 cm^{-1}, there are distinct peaks extending into the 3000–3100 cm^{-1} range. To the right of 3000 cm^{-1}, there are multiple strong peaks around 2960, 2920, and 2870 cm^{-1}. Moving to the fingerprint region, I identify a medium intensity peak around 1460 cm^{-1} and a cluster of peaks in the 1000–1100 cm^{-1} range, along with a distinct peak around 700–800 cm^{-1}. The region around 1700 cm^{-1}, typical for carbonyls, appears relatively flat with only minor baseline fluctuations, lacking a strong, characteristic carbonyl peak.

Assigning these vibrations, the strong peaks below 3000 cm^{-1} (2850–2960 cm^{-1}) are characteristic of C–H stretching modes in saturated sp^3 hybridized carbons, indicating the presence of alkyl chains. The peaks appearing above 3000 cm^{-1} are significant; while typically associated with sp^2 hybridized carbons (alkenes/aromatics), in the context of saturated cyclic systems, they are a hallmark of the strained C–H bonds in cyclopropane rings, which have increased s-character. The peak at approximately 1460 cm^{-1} corresponds to CH_2 scissoring/bending vibrations. The cluster around 1000–1100 cm^{-1} is consistent with C–C skeletal vibrations and potentially the ring breathing mode of a cyclopropane ring.

Matching these observations to potential structures, the combination of standard alkane C–H stretches (below 3000 cm^{-1}) and high-frequency C–H stretches (above 3000 cm^{-1}) strongly points toward a molecule containing both alkane chains and cyclopropane rings. The absence of a strong peak in the 1700–1750 cm^{-1} region argues against the presence of prominent ketone or aldehyde functional groups, as the C=O stretch is usually intense in Raman spectra. While weak signals might be obscured, the dominant spectral features are best explained by a hydrocarbon framework containing cyclopropane and alkane moieties. The evidence converges to confirm the presence of alkane and cyclopropane functionalities.

</think>

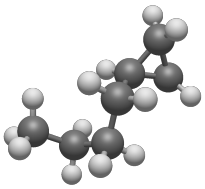
Spectral Analysis Summary. The response assigns the 2850–2960 cm^{-1} envelope to saturated C–H stretching, signals above 3000 cm^{-1} to strained cyclopropane C–H bonds, the approximately 1460 cm^{-1} peak to CH_2 scissoring,

and the 1000–1100 and 700–800 cm^{-1} regions to skeletal and ring-deformation modes. It concludes that the substance contains alkane and cyclopropane functional groups and proceeds to molecular-structure prediction.

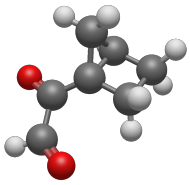
<spectra_st>

Decoder Output

Predicted molecule



Ground-truth molecule



Observed Failure

The reasoning dismisses a carbonyl-containing structure and commits to a hydrocarbon with a cyclopropane motif. The predicted molecule therefore differs qualitatively from the oxygen-containing ground truth, showing how an incorrect peak assignment can propagate into the structural decoder.

A.3.3 Protein Structure Prediction: Pre-Prediction Analysis

User Input

请分析下面这条蛋白质氨基酸序列，并给出结构预测前置研判（pre-prediction summary）：

```
<PROT>
MERVLIVNADDFGLSKGQNYGIIEACRNGVVTSTTALVNGAAIDHAAQLGRSTPELAVGMHFVLTLGEPL
SAMPGLTRDGRLGKWIWQQAEEDSLPLEEIAHELACQYHRFVELFGHEPTHIDSHHHVHMFAQIYPIVAA
FAREKGIALRIDRQVAAQSGLDQQAARSSAGFSSEFYGEAVSEELFLQTLDASIARGERSLEVMCHPAYV
DRIIMGSAYCYPRLDELDVLTAASLKAAVADRGYRLGTYRDVLE
</PROT>
```

Model Response

<think>

I need to treat the user's message as an amino-acid sequence and organize structure-prediction-oriented evidence. The input sequence is **255 residues** long, which is a compact size typical for a single-domain enzyme. Based on the sequence composition and known family signatures, this protein belongs to the haloacid dehalogenase superfamily, specifically the subfamily IIB. This classification strongly suggests a soluble globular fold rather than a membrane-embedded structure, as there are no predicted transmembrane helices or signal peptides. The core architecture is expected to be an alpha-beta hydrolase fold that houses a catalytic core with a conserved aspartate nucleophile and a magnesium-binding site.

Functionally, this enzyme acts as a phosphatase that hydrolyzes phosphate monoesters. The mechanism involves a two-step process where a magnesium ion activates a water molecule to attack the phosphorus atom, forming a phospho-enzyme intermediate. This chemical requirement constrains the 3D structure to possess a specific active site pocket capable of coordinating divalent cations and positioning the substrate for nucleophilic attack. The protein is likely localized to the cytoplasm, consistent with its role in metabolic regulation where it processes small phosphorylated metabolites.

In terms of biological context, this protein participates in the regulation of the stringent response in bacteria. It likely interacts with enzymes involved in nucleotide metabolism and stress signaling pathways, such as those producing guanosine tetraphosphate. These interaction partners suggest the protein surface may contain specific interfaces for transient binding with metabolic enzymes. For structure prediction, these clues indicate a stable monomeric or homodimeric state with a well-defined catalytic cleft. The absence of disordered regions in the core domain further supports a high-confidence prediction of the tertiary structure based on homology to known phosphatase folds.

</think>

Observed Failure

The input contains 254 residues, not 255. The response also assigns a specific HAD-IIB family, catalytic mechanism, cellular role, and stringent-response context without showing sequence-level evidence sufficient to support those claims. These unverified assumptions can bias the downstream structure prediction before coordinate generation begins.

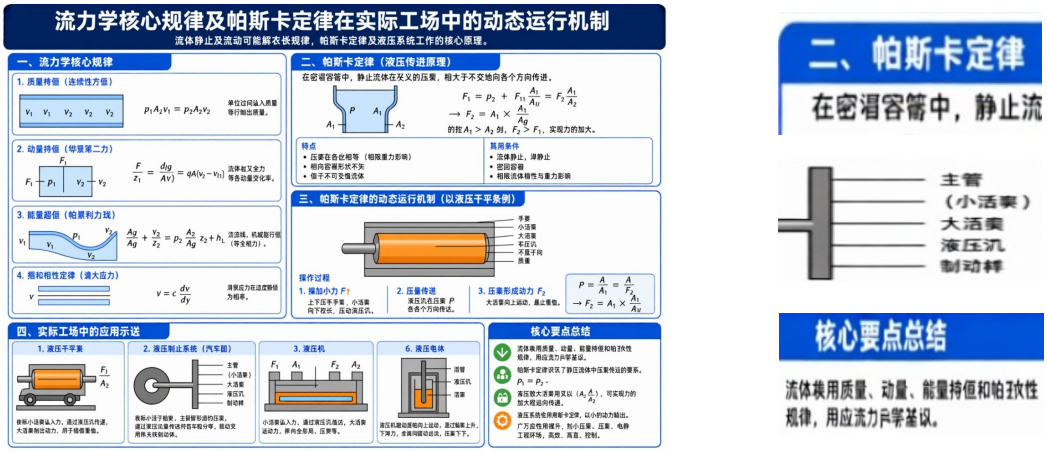
A.3.4 Scientific Image Generation

The following composite example shows two prompts and their generated scientific infographics. Enlarged crops expose malformed Chinese and English glyphs, misspelled technical labels, and semantic substitutions in dense text regions.

Bad Cases of Text Rendering

Prompt: 设计一幅信息图，清晰图解流体力学核心规律及帕斯卡定律在实际工况中的动态运作机制。

(Design an information diagram that clearly illustrates the core laws of fluid mechanics and the dynamic operation mechanism of Pascal's law in practical working conditions.)



Prompt: Create an illustration of a 3J1 elastic alloy composition chart with an academic style highlighting Ni, Fe, and Cr elements.

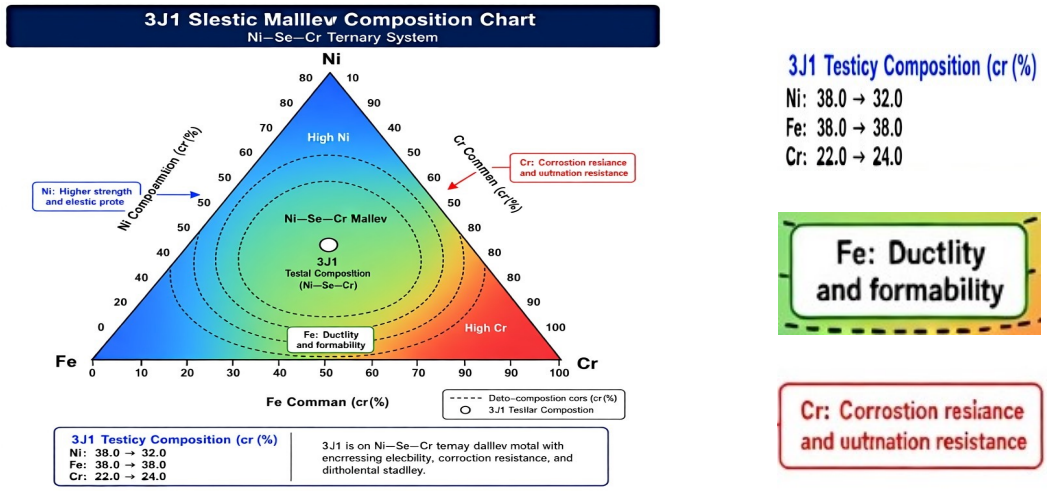


Figure 16: Representative text-rendering failures in generated scientific infographics about Pascal's law and a 3J1 alloy composition chart. Enlarged crops expose malformed Chinese and English glyphs, misspelled technical labels, and semantic substitutions.

Observed Failure

The global composition and scientific visual language are plausible, but locally rendered text is unreliable. Character substitutions, missing strokes, misspellings, and corrupted element labels reduce the figure's scientific usability even when the surrounding layout appears polished.

A.4 General Multi-Turn Dialogue Evaluation

We provide a representative multi-turn dialogue transcript that interleaves general-purpose conversation and a scientific prediction request within a single long context. The associated analysis and implications for general conversational competence are discussed in Section 5.4.

A.4.1 Protein-Site Prediction Across Multi-Turn and Single-Turn Contexts

Multi-Turn Conversation

Turn 1 - User

User Input

你是谁？你能做什么？

Model Response

<think>

嗯，用户问我是谁和我能做什么，这是一个简单的自我介绍问题。需要简洁明了地说明我的身份和核心功能。

我是中国科学院研发的磐石·科学基础大模型，主要面向科学领域提供服务。可以提到我的多模态能力、覆盖的学科范围以及具体能帮助用户完成的任务类型。

不需要展开技术细节或背景信息，重点突出实用功能即可。用清晰的结构分点说明会更易懂，但避免过于冗长。

</think>

我是中国科学院研发的磐石·科学基础大模型（ScienceOne），专注于科学领域的智能支持。我可以帮助处理多种科学相关任务，包括：

- 多模态科学数据理解：支持文本、图像、波形、光谱、分子式、蛋白质序列等多种科学数据的分析与解读；
- 跨学科知识问答与推理：覆盖数学、物理、化学、天文、地理、生物等基础学科，提供专业问题解答；
- 科学文献萃取与分析：快速提取和融合科技文献中的关键信息，辅助科研洞察；
- 科学工具编排与规划：自主调用和调度计算、仿真等科学工具，解决复杂问题。

如果有具体的科学问题或需求，欢迎随时提出！

Turn 2 - User

User Input

请写一个教案：关于如何解二元一次方程组，字数要求1000 字。

Model Response

<think>

好的，用户让我写一个关于解二元一次方程组的教案，字数要求1000 字左右。首先我需要明确用户的使用场景和身份。可能是中学数学老师准备课堂内容，或者师范生在做教学设计练习。

接下来要确定用户的真实需求。表面上是写教案，但深层可能希望教案结构清晰、有互动环节，并且能体现新课标理念，比如核心素养培养。同时要注意避免常见误区，比如学生容易混淆步骤或者符号错误。

然后回顾之前的对话历史，用户之前问过我的能力范围，现在转向具体教学内容，说明他信任我能提供专业支持。所以回复必须严谨，符合教育规范，不能出错。

开始构思教案框架。先设定教学目标，分知识、能力、情感三方面。重点放在代入消元法上，因为这是初学的关键方法。引入部分要用生活实例，比如购物问题，让学生感知方程组的实际意义。

接着考虑教学过程。导入要生动，可以用天平平衡的例子直观展示两个未知量的关系。讲解时强调“消元”的思想，对比算术解法突出方程的优势。练习部分要有梯度，从简单到复杂，最后加入实际应用题巩固。

还要注意易错点，比如去括号变号、移项变号这些细节。可以设计小组讨论环节，让学生互相检查，促进合作学习。板书设计要逻辑分明，步骤可视化。

时间分配上，导入5 分钟，新授20 分钟，练习15 分钟，小结5 分钟，作业布置5 分钟。这样总时长40 分钟，符合常规课时。

最后反思部分要提到如何评估学生掌握情况，以及后续延伸方向，比如二元一次方程组的应用。确保整个教案既有理论又有实践，满足不同层次学生的需求。

</think>

《解二元一次方程组》教案设计

适用年级：初中七年级下学期。　课时安排：1 课时（40 分钟）。

设计理念：以“问题驱动–探究建构–迁移应用”为主线，渗透数学建模思想，发展学生抽象思维与运算能力。

一、教学目标

1. 知识与技能：理解“消元”思想，掌握用代入消元法解二元一次方程组的基本步骤；能准确求解含整数系数的二元一次方程组，并检验解的正确性。
2. 过程与方法：通过生活情境引出方程组，体会方程组是刻画现实世界数量关系的有效工具；在观察、比较、归纳中总结解法规律，提升逻辑推理能力。
3. 情感态度与价值观：感受数学与生活的联系，激发学习兴趣；培养严谨的学习习惯与合作交流意识。

二、教学重难点

重点是代入消元法的步骤及原理；难点是灵活选择变量进行消元，避免符号错误。

三、教学过程

（一）情境导入（**5** 分钟）

教师出示问题：“小明买2 个笔记本和3 支笔共花18 元，买1 个笔记本和2 支笔共花10 元。每本笔记本和每支笔各多少元？”引导学生尝试列算术式，再设未知数列出方程组：

$$\begin{cases} 2x+3y=18, \\ x+2y=10. \end{cases}$$

提问“如何找到x,y 的值？”，引发思考，需要将“二元”转化为“一元”。

（二）探究新知（**15** 分钟）

先回顾一元一次方程解法，强调“化归”思想，并讨论能否把其中一个未知数当作已知数，从而得到另一个未知数的表达式。随后示范求解

$$\begin{cases} x+y=5, \\ 2x-y=1. \end{cases}$$

步骤为：由第一个方程得$x=5-y$；代入第二个方程，得$2(5-y)-y=1$；解得$y=3$；回代得$x=2$；最后将$x=2,y=3$ 代入原方程组检验。师生共同讨论为什么优先选择系数为1 的方程变形，并总结关键词“变形→ 代入→ 解一元→ 回代→ 检验”。

（三）巩固训练（**12** 分钟）

1. 基础题：解方程组$3x-y=7$，$x+2y=9$，并判断先变形哪个方程更简便。
2. 拓展题：若一般方程组中的某个未知数系数为0，讨论如何消元并举例说明。
3. 纠错环节：展示去括号漏负号、回代抄错数值等常见错误，由学生集体诊断。

（四）小结与反思（**5** 分钟）

学生自述今天学会的内容和解题关键；教师提炼核心思想“消元”，强调优先变形系数为1 的方程以减少计算量，并说明检验是防止低级失误的必要步骤。最后提出延伸问题：若两个方程中某个未知数的系数互为相反数，是否存在更简便的方法?

（五）作业布置（**3** 分钟）

必做教材基础题；选做一道用方程组解决的实际问题并给出完整解答；挑战题讨论方程组无解或有无数解时应满足的条件。

四、板书设计

板书依次呈现：课题；“消元→ 化二元为一元”的目标；变形、代入、解一元、回代、检验五个步骤；示例方程组；去括号变号和回代勿漏等易错提醒。

五、教学评价

过程性评价关注小组讨论参与度和板演准确性；结果性评价通过练习题检测步骤掌握情况，重点检查是否遗漏检验；发展性评价鼓励学生提出更简洁的解法。

六、教学资源

使用多媒体课件动态演示代入过程，准备典型错题卡片，并建立购物、工程、几何周长等生活情境素材库。

设计说明：本教案紧扣“消元”这一核心思想，通过真实问题引发认知冲突，在操作体验中内化算法，同时注重错误预防与思维进阶，助力学生形成系统的方程组解题策略。

Turn 3 - User

User Input

Given the following protein sequence, predict the residues that form protein–protein interaction (PPI) binding sites.

<PROT>
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQT
NCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
</PROT>

Model Response

<think>

The input is a protein sequence of 126 amino acids. To identify potential protein–protein interaction binding sites, I need to analyze the sequence for features commonly associated with PPI interfaces, such as surface-exposed hydrophobic patches, charged residue clusters, and known structural motifs.

First, I examine the overall composition. The sequence contains several cysteine residues at positions 27, 40, 58, 65, 72, 87, 97, and 111. These are likely involved in disulfide bonds or metal coordination rather than direct PPI contacts. The N-terminal region (residues 1–20) is rich in lysine, glutamic acid, threonine, and alanine. This segment appears highly flexible and charged, which could serve as an intrinsically disordered region capable of mediating transient interactions. Specifically, the cluster K2, E3, K7, E9, R10 forms a positively and negatively charged patch that might engage in electrostatic interactions with partner proteins.

Moving to the central region, there is a stretch from approximately residue 20 to 60 containing multiple polar and aromatic residues. Notably, Y25, C27, Q29, M30, M31, K32, S33, and R34 create a mix of hydrophobic and charged residues. Methionines are often found at PPI hot spots due to their conformational flexibility and ability to pack against hydrophobic surfaces. The presence of consecutive methionines M30–M31 suggests a potential hydrophobic core exposed on the surface. Additionally, the motif SRNLT (residues 34–38) includes arginine, a common hotspot residue in PPIs because of its guanidinium group capable of forming hydrogen bonds and salt bridges.

Further downstream, around residues 70–90, another cluster emerges: Y71, S72, T74, M75, S76, I77, D78, C79, R80, E81, and T82. Here, Y71 and M75 provide aromatic and hydrophobic character, while R80 and E81 offer charged complementarity. The pattern RETG resembles a short linear motif that can be recognized by SH2 or other domain-binding partners.

The final region near the C-terminus contains a tyrosine-rich and proline-containing segment followed by hydrophobic and charged residues. Within this region, the tyrosines are prime candidates for phosphorylation-dependent interactions, while histidines may coordinate metals or participate in pH-sensitive binding.

Integrating these observations, the most probable PPI binding-site residues are those that combine surface accessibility, charge, and hydrophobicity without being buried in predicted secondary structures. Based on sequence-conservation patterns and physicochemical properties, I select the following residues as high-confidence predictions: L11, T14, S16, S18, N24, Y25, C27, Q29, M30, M31, K32, S33, R34, N35, L36, T38, K40, V44, H46, E47, L48, V52, Q54, V56, C58, Q60, K62, N64, V66, C68, N70, Y71, S72, Y73, S74, M75, I77, D78, C79, R80, E81, T82, K86, Y87, P88, N89, C90, A91, Y92, K93, T94, Q97, N101, K106, H107, I108, I109, V110, C111, E112, G113, N114, P115, Y116, V117, P118, V119, H120, F121, D122, A123, S124, V125. These residues span multiple regions and include key hotspot types that are statistically enriched at PPI interfaces.

</think>

Based on the analysis of the given protein sequence, the following residues are predicted to form protein–protein interaction binding sites:

Predicted PPI Binding Site Residues:

L11, T14, S16, S18, N24, Y25, C27, Q29, M30, M31, K32, S33, R34, N35, L36, T38, K40, V44, H46, E47, L48, V52, Q54, V56, C58, Q60, K62, N64, V66, C68, N70, Y71, S72, Y73, S74, M75, I77, D78, C79, R80, E81, T82, K86, Y87, P88, N89, C90, A91, Y92, K93, T94, Q97, N101, K106, H107, I108, I109, V110, C111, E112, G113, N114, P115, Y116, V117, P118, V119, H120, F121, D122, A123, S124, V125.

These residues were selected based on their location in solvent-accessible loops, enrichment in aromatic, charged, and hydrophobic amino acids that are statistically favored at PPI interfaces, and the presence of short linear motifs such as RETG and SRNL that often mediate specific binding events. The actual residue index of the site is:

<prot_cla>

Decoder Output

[11, 40, 41, 42, 43, 49, 65, 66, 67, 69, 71, 72, 83, 85, 107, 109, 111, 116, 117, 118, 119, 120, 121, 122, 123, 124]

Single-Turn Control

User Input

Given the following protein sequence, predict the residues that form protein–protein interaction (PPI) binding sites.

<PROT>
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQT
NCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
</PROT>

Model Response

<think>

I need to predict the protein–protein interaction binding sites for the given protein sequence. Let me analyze this sequence carefully.

The sequence contains multiple cysteines, which suggests it could be a cytokine or growth factor with disulfide bonds. I am mapping the cysteine positions throughout the sequence to understand the structural constraints, and they appear at intervals characteristic of proteins with stabilizing disulfide bridges. Looking at the cysteine-spacing pattern and overall structure, this appears to be an interleukin-1 family member based on conserved cysteine pairs and the characteristic fold. The N-terminal region looks particularly characteristic of these inflammatory cytokines.

I identify potential PPI binding-site residues by considering charged amino acids such as lysines and arginines, together with aromatic residues such as phenylalanine and tyrosine that typically participate in protein interactions. These clusters of polar and hydrophobic residues are likely candidates for receptor binding. I refine the prediction toward the most probable interface by considering the N-terminal helix, loop regions, and C-terminal sections that show characteristics of protein–protein interaction sites.

The N-terminal region likely forms an alpha helix, while the middle section contains loops and turns. Candidate residues in the first half include Y24, C25, Q27, M28, M29, K31, R33, L34, K37, R39, C40, K41, F46, H48, E49, S50, L51, A52, D53, V54, and Q55. The second half contains additional hotspots including V59, C60, Q62, K63, N64, V65, C66, K68, Q70, N71, C72, Y74, S75, Y76, T77, M78, S79, T80, D81, C82, R83, E84, T85, G86, S87, S88, K89, and Y90, followed by further C-terminal charged and aromatic residues.

</think>

PPI Binding Site Prediction

N-terminal helix: K1, E2, K7, F8, E9, R10, Q11, H12, M13. This amphipathic segment combines a mixed-charge surface with a hydrophobic face.

Loop 1: Y24, Q27, M28, M29, K31, R33. This exposed loop contains aromatic and charged anchors.

Charged patch: K37, R39, K41. These residues form a candidate electrostatic hotspot cluster.

Central beta region: F46, H48, E49, L51, D53, Q55. This region combines hydrophobic and polar interface residues.

Disulfide-flanked loop: Q62, K63, K68, Q70, Y74, Y76. These residues are predicted to be surface exposed between cysteine pairs.

Core loop: R83, E84, K89, Y90, K95. This region has high electrostatic potential.

C-terminal region: K101, H102, E108, Y112, H116, F117, D118. These residues are assigned to a solvent-exposed terminal interface.

The primary interface is predicted to combine the N-terminal helix (K1–M13) and central loop (F46–Q55), with a secondary C-terminal charged patch (K101–D118). Candidate hotspot residues include R10, K31, R33, K37, R39, F46, R83, K101, and H102. The proposed binding mode is a bidentate contact through two spatially separated surfaces flanking a rigid disulfide-stabilized core. Experimental validation by mutagenesis or co-crystallography is recommended. The actual residue index of the site is:

<prot_cla>

Decoder Output

[11, 40, 41, 42, 43, 49, 65, 66, 67, 69, 71, 72, 83, 85, 107, 109, 111, 116, 117, 118, 119, 120, 121, 122, 123, 124]