WHEN DO BIOLOGICAL REASONING MODELS USE THEIR BIOLOGICAL INPUTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological reasoning models use post-training to connect an LLMs to biological inputs, typically representations from a biological foundation model and biological text such as gene and pathway descriptions, functional annotations, and gene lists. Their benchmark accuracy is taken as evidence that the LLM reasons over their biological inputs. We test this assumption in six biological reasoning models across DNA, protein, and single cell tasks. We perturb one biological input while holding the query and other inputs fixed, construct evidence conflicts that pair the foundation model representation of one genome, protein, or cell with the text of another, fit linear probes to the representations the language model receives, and analyze reasoning traces against the biological inputs. We find that current post-training strategies do not ensure that foundation model representations contribute to task performance. Evo2 and ESM3 contribute little to BioReason and BioReason-Pro performance on the evaluated tasks. Shuffling the DNA sequence barely changes BioReason disease prediction accuracy, and in evidence conflicts the two models follow the text in 97.9% and 99.7% of cases. Linear probes trained on the Evo2 and ESM3 representations predict the task targets, so these foundation models encode information relevant to the task, but provide limited overall performance improvement to BioReason and BioReason-Pro. In contrast, foundation model inputs contribute to ChatNT, Prot2Text-V2, and CellWhisperer performance, and differentially expressed genes in the gene sentence contribute to Cell2Sentence-Scale performance. Across SFT and RL checkpoints of BioReason-Pro and 42 BioReason checkpoints, increases in accuracy do not imply greater performance contributions from biological inputs. In reasoning traces, BioReason traces misstate nucleotide changes, while BioReason-Pro traces describe functions omitted from final predictions under evidence conflicts. Future work should test whether post-training objectives that reward correct use of biological inputs improve their contribution to task performance.

1 INTRODUCTION

Biological reasoning models combine language models with biological inputs to make predictions about proteins, genomes, and cells. These inputs include continuous representations from biological foundation models and biological text, such as functional annotations, gene and pathway descriptions, and expression-ranked gene lists. Recent examples include DNA–LLM models (Fallahpour et al., 2025; de Almeida et al., 2025), single-cell transcriptomic language models (Levine et al., 2024; Schaefer et al., 2025), and protein reasoning models (Queen et al., 2025; Fallahpour et al., 2026; Liu et al., 2024; Su et al., 2026). Benchmark performance alone, however, does not establish which biological inputs drive prediction. Predictive annotations can support accurate answers even when accompanying sequence or transcriptomic representations are underused, so benchmark improvements need not reflect greater use of these representations (Geirhos et al., 2020; DeGrave et al., 2021).

Attributing a prediction to a particular biological input is difficult because these models often receive multiple representations of the same protein, genome, or cell. These representations can encode overlapping biological properties at different levels of abstraction. A limited performance contribution from one input may therefore reflect support from another input, limited predictive value of the evaluated input, or limited use of a predictive input by the reasoning model. Benchmark performance alone cannot distinguish among these alternative explanations.

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

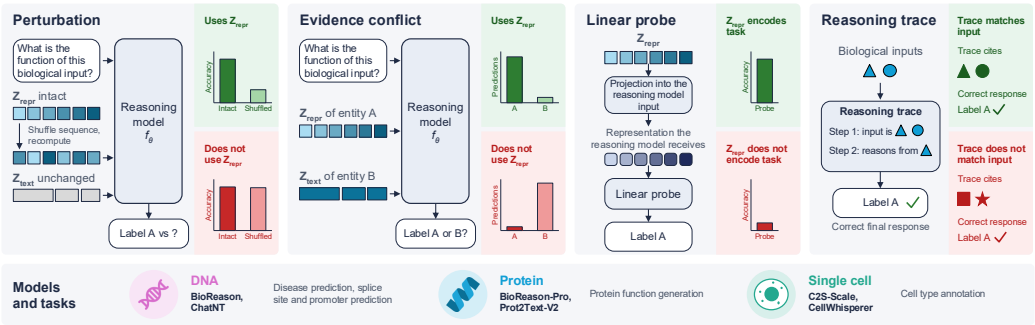

Figure 1: Evaluating biological input usage in biological reasoning models across DNA, proteins and single-cell.

Related problems have been studied in vision–language models, where predictions can exploit language priors or remain insensitive to relevant visual content (Goyal et al., 2017; Agrawal et al., 2018; Hessel & Lee, 2020; Asadi et al., 2026). Representation probing has separately shown that a target can be predicted from a learned representation without implying that the model uses that representation for its own prediction (Ravichander et al., 2021; Elazar et al., 2021). Biological reasoning models exacerbate these challenges because several representations derived from the same biological object may be available to the model simultaneously. Existing evaluations of these models emphasize task performance, leaving unresolved which biological inputs contribute to prediction.

Present work. Biological reasoning models are multimodal LLMs that receive a user task together with biological inputs, such as foundation model representations of a genome, protein, or cell and biological text, and generate a response that contains an answer and often a reasoning trace (Figure 1). We develop tests that determine whether the response of such a model depends on each of its biological inputs. The first test ("Perturbation") perturbs one biological input while holding the user task and the other inputs fixed and measures changes in the answer and task performance. The second test ("Evidence conflict") places inputs from two different genomes, proteins, or cells in the same prompt so that they support different answers, and identifies the input the model follows. When an input contributes little to performance, linear probes on the representation the model receives determine whether that input encodes the task target, which separates an uninformative input from an informative input with a limited performance contribution ("Linear probe"). The fourth test ("Reasoning trace") evaluates whether biological claims in generated traces and prompted rationales agree with the supplied inputs, respond to input perturbations, and remain consistent with final predictions. We apply these tests to six models (Table 1) across post-training checkpoints and model scales. Our main findings are:

- **Foundation model representations can contribute little to performance even when they encode the task target.** BioReason and BioReason-Pro prefer the biological text in evidence conflicts, and Evo2 and ESM3 representations contribute little to their performance, although linear probes predict the task targets from these representations. Foundation model representations contribute to ChatNT, Prot2Text-V2, and CellWhisperer performance, and differentially expressed genes (DEGs) in the gene sentence contribute to C2S-Scale performance.
- **Higher accuracy after post-training does not indicate a greater performance contribution from the foundation model.** SFT and RL increase the accuracy of BioReason and BioReason-Pro without increasing the performance contribution from Evo2 and ESM3 representations, and we observe the same result across 42 BioReason checkpoints. Scaling C2S-Scale from 2B to 27B increases accuracy and the performance contribution from DEGs.
- **Generated reasoning does not consistently connect biological evidence to final predictions.** BioReason makes incorrect claims about nucleotide substitutions, while BioReason-Pro discusses functions absent from its final GO annotations under evidence conflicts. For C2S-Scale 27B, marker enrichment is concentrated in the first few gene references and is lost across the full rationale where it over-references ribosomal protein genes.
- **Auxiliary supervision increases sequence sensitivity, but the edited-base prediction contributes little to disease prediction performance.** Training BioReason to predict the edited bases makes this prediction depend on Evo2 representations. Shuffling the sequence lowers accuracy on the edited base pair by 0.549 and disease prediction accuracy by only 0.021. However, this improvement in predicting edited bases is limited to genomes seen during training.

2 Related Work

Biological reasoning models. Biological reasoning models differ in their input representations. ChatNT uses Nucleotide Transformer representations, while BioReason combines Evo2 with gene and pathway text (de Almeida et al., 2025; Brixi et al., 2026; Fallahpour et al., 2025). Prot2Text-V2 combines ESM2 representations with protein name and taxon text, whereas BioReason-Pro combines ESM3 with GO-GPT predictions and InterPro annotations (Fei et al., 2025; Hayes et al., 2025; Fallahpour et al., 2026). Cell2Sentence-Scale represents transcriptomes as expression-ranked gene sentences, while CellWhisperer uses continuous transcriptomic representations (Levine et al., 2024; Schaefer et al., 2025). Other protein models connect biological representations with language for retrieval, annotation, or generation (Queen et al., 2025; Su et al., 2026; Xu et al., 2023; Liu et al., 2024; Jararweh et al., 2025; Wang et al., 2024; Char et al., 2025). Evaluations largely emphasize task performance, scale, and post-training (Fesser et al., 2026; Wu et al., 2025; Zhao et al., 2026; He et al., 2025; Dip et al., 2026). We instead test whether predictions depend on biological input content, distinguishing target predictability from model use.

Multimodal representation learning and evaluation. Multimodal learning literature studies how representations from different modalities can be aligned, combined, and decomposed into shared or modality-specific structure (Liang et al., 2024; Dufumier et al., 2025; Huh et al., 2024; Gupta et al., 2026; Zhang et al., 2026; Richter et al., 2026). Complementary work evaluates whether multimodal models use the inputs provided to them. Vision–language models can exploit language priors, remain insensitive to visual content, or improve benchmark performance without increased cross-modal interaction (Goyal et al., 2017; Agrawal et al., 2018; Cadene et al., 2019; Hessel & Lee, 2020; Frank et al., 2021; Asadi et al., 2026; Gao et al., 2026). Counterfactual inputs and controlled perturbations have therefore been used to test whether predictions respond to the content of a particular modality (Niu et al., 2021; Thrush et al., 2022). Our work focuses on biological reasoning models, where multiple representations of the same protein, genome, or cell may be available simultaneously, and asks which of these biological inputs contribute to the downstream prediction.

Predictive representations and generated rationales. Linear probes provide a simple test of predictability from learned representations (Alain & Bengio, 2017; Hewitt & Liang, 2019), while subsequent work has shown that probe performance and downstream model behavior should be evaluated separately (Ravichander et al., 2021; Elazar et al., 2021). Generated rationales introduce a related distinction because plausible explanations need not identify the features that affect model predictions (Jacovi & Goldberg, 2020; DeYoung et al., 2020; Wiegreffe et al., 2021; Turpin et al., 2023). We use simple predictive baselines to interpret limited performance contributions from a biological input and compare features that affect prediction with those referenced in generated rationales.

3 Approach

3.1 Problem Formulation

Biological reasoning models take a text query q, which specifies the task and response format, together with biological inputs $\mathbf{Z} = (\mathbf{Z}_{\text{repr}}, \mathbf{Z}_{\text{text}})$ associated with a biological modality X, such as protein sequence, genomic sequence, or a single-cell transcriptome. We distinguish continuous biological representations $\mathbf{Z}_{\text{repr}}$ from biological inputs represented as text $\mathbf{Z}_{\text{text}}$. For target y, the model f_θ produces a prediction and, optionally, a rationale or reasoning trace, $(\hat{y}, \hat{R}) = f_\theta(q, \mathbf{Z})$. We call the full input $(q, \mathbf{Z})$ the prompt, and a task is the set of queries that share one objective and label space, such as disease prediction.

Our objective is to determine whether a biological input $\mathbf{Z}_i$ is used by the model to make its prediction. We test this by changing the task-relevant biological content of $\mathbf{Z}_i$ while holding q and the remaining inputs $\mathbf{Z}_{-i}$ fixed, and measuring prediction sensitivity and changes in task performance. We also test source preference with *evidence conflicts*, inputs in which two biological evidence sources support different predictions, constructed by combining $\mathbf{Z}_i$ from one query with one or more inputs in $\mathbf{Z}_{-i}$ from another, and measure which source the prediction follows. When perturbing $\mathbf{Z}_i$ has little effect on task performance, we additionally test whether the target can be predicted from $\mathbf{Z}_i$ using a simple predictive baseline, such as a majority-vote lookup or linear probe. $\mathbf{Z}_i$ is *underutilized* when it supports prediction of the target with such a baseline, but changing its task-relevant biological content has little effect on the reasoning model's task performance after controlling for predictive information in $\mathbf{Z}_{-i}$ where possible.

Table 1: Models, checkpoints, input sources, tasks, and evaluation data. Underlined checkpoints are used unless otherwise specified, – refers only one released checkpoint. BioReason checkpoints are from Fesser et al. (2026), as the original study did not release checkpoints.

Model	Modality	Checkpoints	Input sources	Task	Evaluation data
BioReason (Fallahpour et al., 2025) checkpoints from (Fesser et al., 2026)	Genomic sequence	SFT, RL; 42 post-training checkpoints	Projected reference and variant Evo2 representations ($\mathbf{Z}_{\mathrm{Evo2}}$); chromosome ($\mathbf{Z}_{\mathrm{chr}}$), gene ($\mathbf{Z}_{\mathrm{gene}}$), KEGG pathway ($\mathbf{Z}_{\mathrm{pathway}}$), and network ($\mathbf{Z}_{\mathrm{network}}$) text	Disease prediction	1,449 KEGG-derived queries over 708 genomes and 37 disease labels; 165 genome-dependent queries; 145 held-out queries for the auxiliary task evaluation
ChatNT (de Almeida et al., 2025)	Genomic sequence	–	Nucleotide Transformer representations ($\mathbf{Z}_{\mathrm{NT}}$)	Splice donor, splice acceptor, and TATA promoter prediction	4,578 ChatNT test queries from three binary tasks (2,000 splice donors, 2,000 splice acceptors, 578 TATA promoters; balanced, up to 1,000 per class)
BioReason-Pro (Fallahpour et al., 2026)	Protein sequence	SFT, RL	ESM3 ($\mathbf{Z}_{\mathrm{ESM3}}$); GO-GPT ($\mathbf{Z}_{\mathrm{GO\text{-}GPT}}$); InterPro ($\mathbf{Z}_{\mathrm{InterPro}}$); organism text ($\mathbf{Z}_{\mathrm{organism}}$)	Protein function prediction	14,102 reviewed human proteins; 91-protein temporal holdout; 540 proteins across 140 InterPro families; 96 GO NOT pairs
Prot2Text-V2 (Fei et al., 2025)	Protein sequence	–	ESM2 ($\mathbf{Z}_{\mathrm{ESM2}}$); protein name ($\mathbf{Z}_{\mathrm{name}}$); taxon text ($\mathbf{Z}_{\mathrm{organism}}$)	Protein function description generation	3,917 proteins from test split; 632 same-taxon pairs (1,264 evidence conflicts between name and sequence)
Cell2Sentence-Scale (Levine et al., 2024)	Single-cell transcriptome	2B, 27B	Expression-ranked gene sentence ($\mathbf{Z}_{\mathrm{cell}}$)	Cell type annotation and rationale generation	Five atlases; 4,846 cells for annotation; 2,000 cells for rationale analysis
CellWhisperer (Schaefer et al., 2025)	Single-cell transcriptome	–	CellWhisperer transcriptome representation ($\mathbf{Z}_{\mathrm{CW}}$); expression-ranked top-k gene list prefixes ($\mathbf{Z}_{\mathrm{cell}}^{(k)}$), when supplied	Cell type prediction and evidence conflict	Same 4,846 C2S-Scale cells (five atlases); 330 same-tissue pairs, with correct prediction from representation only (660 conflicts)

3.2 Evaluation Protocol

Table 1 summarizes the six models we benchmark, their input sources, model variants, tasks, and evaluation data. We organize our evaluation around four research questions. **RQ1: Does task-relevant biological content in a biological input affect the model's prediction?** We perturb the task-relevant biological content of $\mathbf{Z}_i$ while holding q and $\mathbf{Z}_{-i}$ fixed and measure changes in predictions and task performance, and we construct evidence conflicts from unmodified queries to measure source preference. When the perturbation has little effect on performance, we additionally test whether the target can be predicted from $\mathbf{Z}_i$ using a simple predictive baseline. When possible, we also remove or control other predictive biological inputs before concluding that $\mathbf{Z}_i$ is underutilized. **RQ2: Does higher benchmark performance correspond to greater performance contributions from biological inputs?** Across model variants that differ in post-training checkpoint or scale, we compare benchmark performance with its change under the same biological perturbations used in RQ1. This tests whether improvements in task performance are accompanied by a greater performance contribution from the evaluated biological input. **RQ3: Do reasoning traces and generated explanations reflect the supplied biological evidence?** We evaluate whether generated responses agree with the biological inputs, and how they change when the biological input is perturbed. For evaluated models that do not produce reasoning traces during prediction, we prompt them to identify the biological evidence supporting their predictions. **RQ4: Can auxiliary supervision increase use of a biological input?** We add an auxiliary objective that requires the model to predict biological content from $\mathbf{Z}_i$ and test whether this increases the effect of $\mathbf{Z}_i$ on the downstream task performance.

4 Results: Biological Input Use in Reasoning Models

4.1 RQ1: Foundation model representations can contribute little to performance even when they encode task information

Here we test prediction sensitivity, performance contribution, and source preference for the evaluated biological inputs with perturbations of biological inputs while holding the query and remaining inputs fixed, and construction of evidence conflicts from unmodified queries. For inputs with limited performance contributions, we confirm their utility to the task with linear probes to test target predictability. Appendices A and B provide the datasets and experimental setup used.

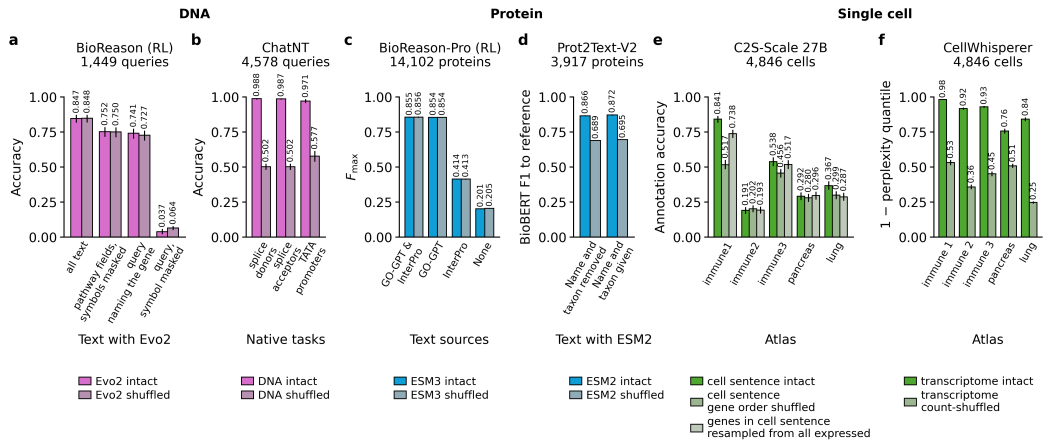

Figure 2: **Reasoning model performance with representations of intact and shuffled sequences. a** BioReason disease prediction accuracy across text conditions. **b** ChatNT accuracy on three native DNA tasks. **c** BioReason-Pro $F_{\max}$ across text conditions. **d** Prot2Text-V2 BioBERT BERTScore F1 with name and taxon text removed or provided. **e** C2S-Scale 27B accuracy with intact, reordered, or resampled gene sentences. **f** CellWhisperer $1 -$ perplexity quantile before and after permuting counts across genes within each cell. Error bars are 95% CIs.

Biological perturbations reduce performance in four models but have little effect on performance in BioReason and BioReason-Pro. BioReason disease prediction accuracy is almost unchanged with shuffled DNA inputs (0.847 vs. 0.847), although it changes 101 of 1,449 answers (7.0%, Table S2). The largest decrease in accuracy occurs when the gene name is removed (Fig. 2a). For ChatNT, accuracy decreases from 0.988 to 0.502 for splice donors, 0.987 to 0.502 for splice acceptors, and 0.971 to 0.577 for TATA promoters (Fig. 2b). The correct label for a shuffled sequence is poorly defined, though this result does support that ChatNT responses differ for the shuffled sequence compared to the original sequence.

For protein models, providing the ESM3 embedding of a shuffled sequence to BioReason-Pro with GO-GPT and InterPro unperturbed results in minimal change in $F_{\max}$ (0.855 vs. 0.856) across 14,102 human proteins. On a temporal holdout of proteins annotated after the checkpoint release, this is still observed (0.685 vs. 0.682). The shuffled sequence does influence the predicted GO term set for 46.8% of 14,102 proteins, with a mean Jaccard similarity of 0.89 to the GO term set predicted for the original sequence (Table S3). For Prot2Text-V2, BioBERT BERTScore F1 decreases by 0.177 with the name and taxon removed, showing a performance contribution from the ESM2 representation.

Across five single-cell atlases, C2S-Scale 27B mean cell type accuracy falls from 0.446 to 0.351 after gene order shuffling and to 0.406 after resampling genes from all expressed genes (Fig. 2e). In Fig. 5 we further observe that removing DEGs from the cell sentence reduces C2S-Scale cell type annotation accuracy more than removing non-DEGs with similar expression. On the same cells we also evaluate CellWhisperer. Permuting counts across gene identities and recomputing CellWhisperer representations reduces $1 -$ perplexity quantile in every atlas, including from 0.98 to 0.53 in immune1 and 0.84 to 0.25 in lung (Fig. 2f). The single-cell inputs contribute to both C2S-Scale and CellWhisperer performance.

Evidence conflicts show whether models follow text or biological inputs. Shuffling input sequences can result in out of distribution inputs, so here we construct evidence conflicts with biological inputs from two different entities combined in the one prompt. For BioReason, Evo2 representations from A and gene and pathway text from B results in disease predictions consistent with B in 97.9% of conflicts (Fig. 3a). BioReason-Pro prefers the GO-GPT source, with responses consistent with GO-GPT in 99.7% of conflicts with InterPro and ESM3 inputs (Fig. 3b). The preference for text remains in pairs where the protein without the property has an experimental GO `NOT` annotation, with responses consistent with the GO-GPT and InterPro sources in 98.1% of 106 conflicts (Fig. S1). Prot2Text-V2 instead favors the sequence source over the conflicting protein name in 97.0% of conflicts (Fig. 3c). C2S-Scale 27B predicts the gene sentence source in 93.7% of conflicts with another cell type's description (Fig. 3d). CellWhisperer assigns lower perplexity to the answer supported by the representation in 96.1% of conflicts with another cell's top-1,000 gene list (Fig. 3e). ChatNT is not evaluated here because its evaluated queries specify the task without providing query-specific biological text that could conflict with the DNA input. These source preferences are consistent with the performance contributions observed in the perturbation experiments.

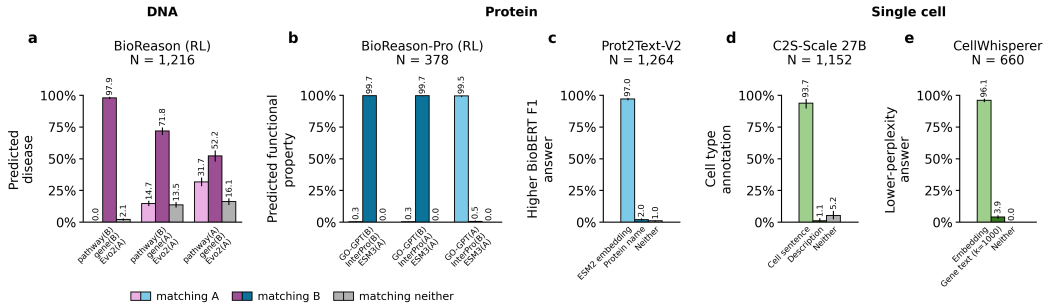

Figure 3: **Evidence conflicts from unmodified queries. a** BioReason: $\mathbf{Z}_{\mathrm{Evo2}}$ from A, gene/pathways from A or B. **b** BioReason-Pro: $\mathbf{Z}_{\mathrm{ESM3}}$ from A, GO-GPT/InterPro from A or B. **c** Prot2Text-V2: $\mathbf{Z}_{\mathrm{ESM2}}$ from A, protein name from B, with taxon fixed. **d** C2S-Scale 27B: $\mathbf{Z}_{\mathrm{cell}}$ from A, cell type description from B. **e** CellWhisperer: $\mathbf{Z}_{\mathrm{CW}}$ from A, top-1,000 gene text from B. Bars indicate preference for A, B, or neither. Error bars are 95% CIs.

Representations with limited performance contributions are predictive of task targets. We next ask if the foundation model representations that contribute little to BioReason and BioReason-Pro performance are predictive of task targets (Fig. 4). We fit linear probes on mean-pooled foundation model output and on the same representations after each biological reasoning model's projector. On 165 disease prediction queries that are genome-dependent (i.e. the text-only input does not always have the same disease label), probes on the Evo2 output and on $\mathbf{Z}_{\mathrm{Evo2}}$ reach 0.736 and 0.735 accuracy, compared with 0.708 for the text-only predictor. Combining Evo2 representations with text features improves the text-only predictor but is below the Evo2-only probes. The probe's improvement is driven by performance on questions with gene PRNP (98 of 165 queries), for which probe accuracy reaches 0.855, compared with 0.708 for the text-only predictor.

For BioReason-Pro, we test whether ESM3 representations separate proteins that share a protein family but differ in function. We select 540 proteins from 140 InterPro families whose members differ in enzyme activity, organelle targeting, or DNA binding, so the family annotation alone does not determine the property. For these properties, the protein without the property lacks the GO annotation. We observe a similar result when we construct pairs where one protein has an annotated GO term and the other protein has the annotated `NOT` of that GO term (Fig. S1). Linear probes on the ESM3 representations reach mean per-family AUROCs of 0.766, 0.898, and 0.830 for the three properties, and probes on the projected representations $\mathbf{Z}_{\mathrm{ESM3}}$ reach 0.713, 0.921, and 0.898, compared with 0.729, 0.890, and 0.736 for BioReason-Pro (Fig. 4b). The joint probe with GO-GPT and ESM3 representations also scores higher than GO-GPT on all three properties. The GO `NOT` pairs span 50 GO terms, so they have no label shared across pairs for a probe. Functional properties are thus predictable from the ESM3 representation.

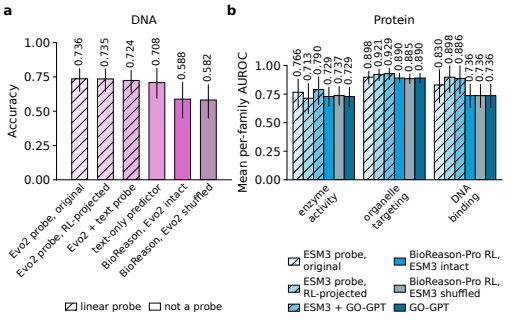

Figure 4: **Targets remain predictable from biological representations. a** Disease prediction accuracy on genome-dependent queries for linear probes and BioReason. **b** Mean per-family AUROC for linear probes and BioReason-Pro. Error bars are 95% CIs.

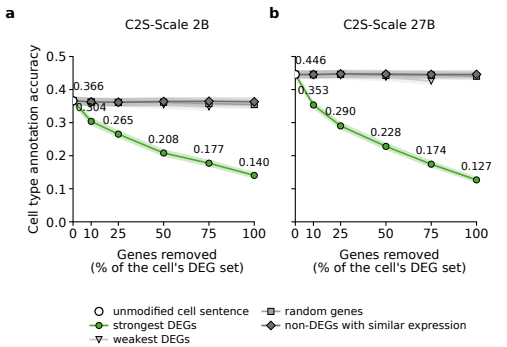

Figure 5: **DEGs contribute to C2S-Scale performance.** Accuracy after removing increasing fractions of the strongest DEGs compared to other sets of genes for 2B (**a**) and 27B (**b**). Bands are 95% CIs.

Finding 1. Some biological inputs contribute little to model performance. Shuffling reveals limited performance contributions, and evidence conflicts reveal preferences for other inputs. However, linear probes show these inputs are predictive of targets for the task.

4.2 RQ2: POST-TRAINING RAISES OVERALL TASK ACCURACY WITHOUT INCREASING THE BIOLOGICAL INPUT'S PERFORMANCE CONTRIBUTION

We next test whether model variants with higher benchmark performance also show greater performance contributions from the biological inputs evaluated in RQ1. For BioReason and BioReason-Pro, we compare the released SFT and RL checkpoints under the same perturbations of $\mathbf{Z}_{\mathrm{Evo2}}$ and $\mathbf{Z}_{\mathrm{ESM3}}$. For C2S-Scale, we compare the 2B and 27B checkpoints using the DEG removal experiment in Fig. 5. Appendix C gives the checkpoints, their training configurations, and the evaluation settings held fixed across them.

SFT and RL improve benchmark performance without greater performance contributions from $\mathbf{Z}_{\mathrm{Evo2}}$ or $\mathbf{Z}_{\mathrm{ESM3}}$. For BioReason, the SFT and RL checkpoints differ in disease prediction accuracy but show similarly limited performance contributions from $\mathbf{Z}_{\mathrm{Evo2}}$ (Fig. 6a). With all biological text present, SFT accuracy is 0.823 with the original $\mathbf{Z}_{\mathrm{Evo2}}$ and 0.814 when it is recomputed from shuffled DNA, whereas RL scores 0.847 and 0.847 under the same conditions. The same pattern is observed as we dropout the KEGG pathway and gene name, with both checkpoints remaining close to the identity line across the four conditions. Thus, the higher disease prediction accuracy of the BioReason RL checkpoint is not accompanied by a greater performance contribution from $\mathbf{Z}_{\mathrm{Evo2}}$.

We observe the same pattern with BioReason-Pro with use of $\mathbf{Z}_{\mathrm{ESM3}}$ in impacting the $F_{\max}$ performance (Fig. 6c). When GO-GPT and InterPro both present, RL increases $F_{\max}$ from 0.829 for SFT to 0.855, while replacing $\mathbf{Z}_{\mathrm{ESM3}}$ with a representation recomputed from shuffled protein sequence changes $F_{\max}$ from 0.829 to 0.833 for SFT and from 0.855 to 0.856 for RL. This limited performance contribution is also observed as InterPro and GO-GPT inputs are removed. Likewise, RL improves protein function prediction without increasing the performance contribution from $\mathbf{Z}_{\mathrm{ESM3}}$. The same comparison across 42 post-training checkpoints that vary model size, SFT LoRA rank, RL LoRA rank and RL duration shows accuracy varying without a consistent increase in the advantage of the original $\mathbf{Z}_{\mathrm{Evo2}}$ over the representation recomputed from shuffled DNA (Fig. 6b). Appendix C.1 reports the individual sweeps.

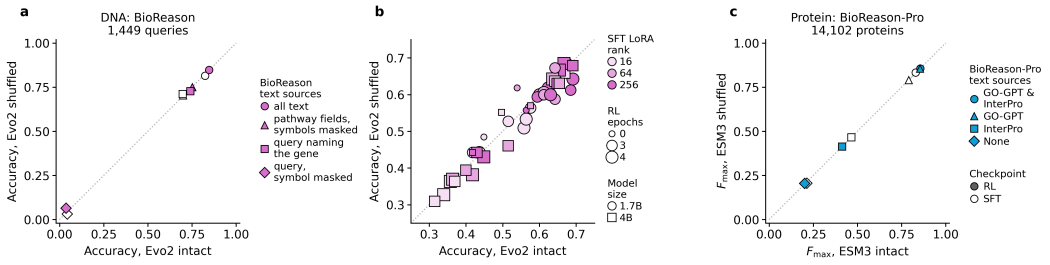

Figure 6: **Biological representations contribute similarly little to SFT and RL checkpoint performance.** BioReason accuracy with original or perturbed $\mathbf{Z}_{\mathrm{Evo2}}$ across **a** biological text conditions and **b** for 42 post-training checkpoints on the 165 genome-dependent queries. **c** BioReason-Pro $F_{\max}$ with original or perturbed $\mathbf{Z}_{\mathrm{ESM3}}$ across biological text conditions.

DEGs contribute more to performance in the larger C2S-Scale model. C2S-Scale provides a contrasting relationship between benchmark performance and biological input contribution. Mean cell type accuracy with the unmodified $\mathbf{Z}_{\mathrm{cell}}$ increases from 0.366 for the 2B checkpoint to 0.446 for 27B while the accuracy decrease under DEG removal from the cell sentence also increases (Fig. 5). At 100% deletion of DEGs, the 2B model scores 0.140 after removing the strongest DEGs, and 0.363 after removing the same number of non-DEGs with similar expression, resulting in a DEG-specific difference of 0.223. The corresponding accuracies for 27B are 0.127 and 0.446, resulting in a larger difference of 0.319. Therefore, the DEG-specific effect is larger for 27B in all five atlases at full DEG removal. Unlike BioReason and BioReason-Pro, the larger C2S-Scale model also shows a greater performance contribution from biological inputs.

Finding 2. Improvements in task performance across post-training configurations are not consistently accompanied by a greater performance contribution from the biological input. RL improves performance for BioReason and BioReason-Pro but does not increase the performance contribution from foundation model representations. Conversely, for cell type annotation C2S-Scale 27B shows a larger accuracy decrease under DEG dropout in the cell sentence than the 2B model.

4.3 RQ3: Reasoning traces and rationales do not reflect biological inputs

We analyze the reasoning traces of BioReason and BioReason-Pro, and prompted rationales from C2S-Scale, to assess correctness, relevance to biological inputs, and consistency with final predictions.

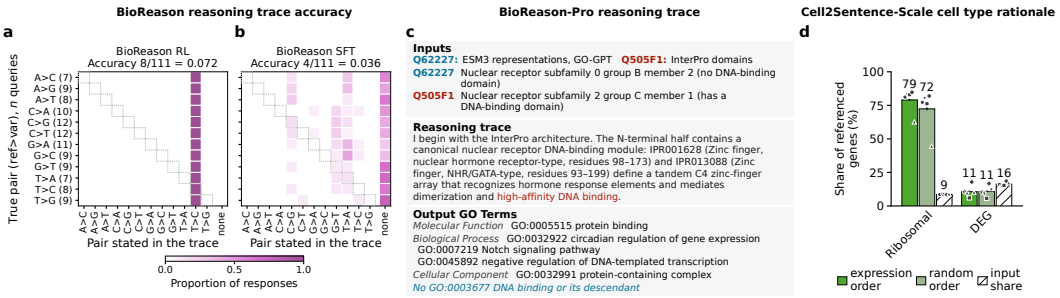

Figure 7: **Analysis of reasoning traces and rationales.** The reference-to-variant base pair BioReason states in its trace on the 111 held-out substitution queries for the **a** RL and **b** SFT checkpoint. **c** Example BioReason-Pro reasoning trace for an evidence conflict. **d** Proportion of the genes in the C2S-Scale 27B rationale that are ribosomal protein genes and DEGs. Dots are the five single-cell atlases and error bars are 95% CIs.

BioReason states variant properties that its input does not determine. In the BioReason reasoning traces, the first step identifies the substituted bases, a genomic coordinate, the chromosome, and the gene, followed by steps that reason from the gene to a disease. On the 145 held-out queries of the pathway network split, the BioReason models state the correct base for 7.2% and 3.6% of all substitution mutations for RL and SFT, respectively (Fig. 7a,b). Notably, the RL checkpoint states the variant is a substitution and is T>C, on all 145 responses and when the input sequence represented by Evo2 is shuffled. Conversely, the SFT checkpoint is partially responsive to sequence and does not always output the same mutation in its reasoning trace.

BioReason-Pro's final GO predictions omit functional properties stated in its reasoning traces. For the evidence conflicts (A, B) with InterPro(B), GO-GPT(A), and ESM3(A) used in RQ1, we test how this affects the reasoning traces when the two members share an InterPro family and have different GO term annotations. We use the InterPro2GO mapping (Burge et al., 2012) and filter for evidence conflict pairs where InterPro annotations of B is predictive of the tested functional property (enzyme activity, organelle targeting, DNA-binding) GO term that A is not annotated with (53 out of 189 evidence conflict pairs). In these responses, the reasoning trace recites all InterPro annotations from B. Furthermore in 23 of 53 traces, it states the tested property in words that are not part of a printed InterPro entry name before predicting the function from the GO-GPT input source. For example, the reasoning trace names a DNA-recognition module before emitting a GO set that contains no DNA-binding term (Fig. 7c).

C2S-Scale rationales show early DEG enrichment but overrepresent ribosomal genes. RQ1 shows that removing the strongest DEGs from $\mathbf{Z}_{\text{cell}}$ changes C2S-Scale cell type annotations. We prompt C2S-Scale 27B for annotation rationales on 400 cells from each of five atlases (Appendix D). Rationales reference 63.4–67.8 genes, of which 93.9–98.7% occur in $\mathbf{Z}_{\text{cell}}$ and 62.6–84.7% are ribosomal protein genes (Fig. S3a). Mean DEG precision across all referenced genes is 10.63%, below the random-draw reference, whereas the first five references show enrichment (Fig. S3b). Among 452 incorrect cell type annotations naming an atlas type, precision against predicted-type DEGs is 24.2% versus 9.9% (Table S6), showing that early references can contain markers of the assigned cell type even when the annotation is incorrect. Randomizing gene order reduces the ribosomal share by 6.6 percentage points but remains overrepresented in the rationale and unchanged DEG precision (Fig. 7d). Marker-focused prompts, ribosomal exclusions, and one-shot examples also yield low overall DEG precision (Appendix D.3.3). Removing the strongest DEGs changes the rationale gene set more than removing expression-matched non-DEGs (Fig. S3c). Finally, adding a rationale request to otherwise matched annotation prompts lowers pooled accuracy from 0.45 to 0.41 (Table S7). Thus, early gene references show marker relevance, but full rationales dilute DEG enrichment and changes annotation behavior.

Finding 3. Generated reasoning traces and rationales can misrepresent biological inputs they describe. Models tend to misstate variants, omit stated functions from predictions. C2S-Scale rationales show early DEG enrichment but overrepresent ribosomal genes overall.

4.4 RQ4: AUXILIARY SEQUENCE SUPERVISION INCREASES SENSITIVITY TO FOUNDATION MODEL REPRESENTATIONS WITH LITTLE TRANSFER TO DISEASE PREDICTION

BioReason is trained with supervised fine-tuning on reasoning traces that specify the base change in the first step, yet RQ3 shows that it rarely generates the correct base change. We therefore test whether explicit supervision of the edited bases improves their prediction and increases the contribution of $\mathbf{Z}_{\mathrm{Evo2}}$ to disease prediction performance. We replace the full reasoning trace with a short target specifying the reference and variant bases at the edited position while retaining the disease answer (Appendix E). On 145 held-out queries, auxiliary supervision with 257-bp input windows yields edited base pair accuracy of 0.671 with intact DNA and 0.122 after shuffling. With 2,048-bp windows, the corresponding accuracies are 0.531 and 0.099, compared with 0.080 for the base changes stated without auxiliary supervision (Fig. 8b). Both auxiliary models also respond to sequence replacement in evidence conflicts (Fig. S4), while disease prediction accuracy changes little under shuffling (Fig. 8a). However, 135 of the 145 held-out queries contain genomes present in training. Pooling validation and held-out queries, edited base pair accuracy on 218 queries with previously seen genomes is 0.691 and 0.606 for the 257-bp and 2,048-bp models, respectively. On 72 queries with unseen genomes, accuracy falls to 0.148 and 0.106, compared with 0.111 for always predicting the most frequent training base pair (Table S10). Thus, the improvement is concentrated in previously seen genomes and increased sequence sensitivity on the original split does not establish reliable edited-base prediction on unseen genomes.

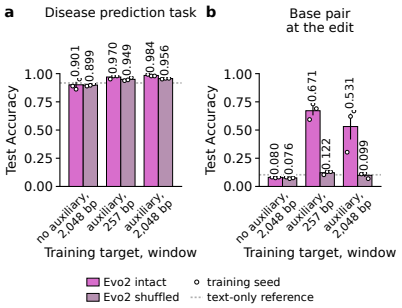

Figure 8: **Auxiliary sequence supervision for BioReason. a** Disease prediction accuracy and **b** edited base pair accuracy on 145 held-out pathway-network split queries with $\mathbf{Z}_{\mathrm{Evo2}}$ derived from original or shuffled DNA. Error bars the standard error over three training seeds.

Finding 4. Auxiliary supervision increases BioReason's sequence sensitivity, but edited base prediction generalizes poorly to unseen genomes. Edited-base predictions respond to DNA shuffling and genome replacement, but accuracy improvements are concentrated to genomes present during training. Disease prediction accuracy also changes little when Evo2 representations are recomputed from shuffled DNA.

5 DISCUSSION

Biological reasoning benchmarks can reward accurate answers without requiring models to interpret the biological representations they receive. When accompanying annotations already support the target, success does not establish that a model can extract complementary information from sequence or cellular inputs. Sensitivity to those inputs can occur but have limited overall performance benefit. Higher accuracy after post-training does not indicate a greater performance contribution from the biological input. Generated reasoning traces may also not accurately reflect inputs to the reasoning model.

Limitations and future work. We evaluate six biological reasoning models on specific tasks and biological inputs, so the findings may not generalize to other architectures or biological modalities. Some analyses shuffle the sequence, which can move representations outside the distribution of natural sequences, although evidence conflicts built from unmodified inputs support the same conclusions. Linear probes from representations given to the reasoning model show that a target is predictable, but they do not show how the language model can use that representation. Future work could evaluate additional models and test training prompt variants within the same model while holding other training settings fixed. Reinforcement learning objectives that reward correct responses to changes in biological inputs, in addition to final accuracy, are another direction to evaluate.

Conclusion. Biological reasoning models can achieve high benchmark performance with limited contributions from predictive biological inputs. Our results show that neither access to a predictive representation nor improved performance after SFT or RL ensures a performance contribution from that representation. The design and evaluation of biological reasoning models should consider the correct use of task-relevant biological inputs as an additional objective alongside predictive accuracy.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback and improve research methodology, help develop the paper's conceptual framework, and assist with implementation through data processing, code for model inference and training. We also used these tools to discuss interpretations of experimental results and refine the paper's hypothesis. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims or proofs, develop theoretical models, perform translation, conduct qualitative or thematic analysis, evaluate or score model outputs, and survey/interview design and transcription are not applicable to this work. Additionally, we used generative AI tools for software development, literature search, brainstorming and refining the paper framing, drafting and editing manuscript text, and designing and plotting figures from experimental results. We have reviewed all AI-assisted work. AI-assisted code was run and tested against published benchmark results of the evaluated models where applicable, suggested references identified through AI-assisted literature search were evaluated by the authors, and all experimental results, figures, and tables were reviewed by the authors for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available biological datasets and released model checkpoints. No new human-subject data or access personally identifiable information was collected. This manuscript evaluates how biological information is used by multimodal biological models and does not provide clinical predictions, treatment recommendations, or experimentally actionable biological designs.

REPRODUCIBILITY STATEMENT

We evaluate released checkpoints on released data and change no model weights except in the auxiliary supervision experiment of Section 4.4. Appendix A documents the evaluation datasets and their preprocessing, prompts used, and evaluation methods, and Appendices B to E document the additional analyses for each RQ. The evaluated weights are available on the HuggingFace Hub from `mims-harvard/bio-posttrain-qwen3-1.7b-dna-sft` and `-rl` for BioReason, `wanglab/bioreason-pro-sft` and `wanglab/bioreason-pro-rl` for BioReason-Pro, `InstaDeepAI/ChatNT`, `xiao-fei/Prot2Text-V2-11B-Instruct-hf`, `vandijklab/C2S-Scale-Gemma-2-2B` and `-27B`, and the CellWhisperer CLIP and chat checkpoints which are publicly distributed by its authors at https://medical-epigenomics.org/papers/schaefer2025cellwhisperer. They read the biological encoders `arcinstitute/evo2_1b_base`, `EvolutionaryScale/esm3-sm-open-v1`, Nucleotide Transformer v2 500M and Geneformer 12L-30M. We run each model through its own released code: https://github.com/bowang-lab/BioReason with the post-training code of https://github.com/mims-harvard/bio-posttrain, https://github.com/bowang-lab/BioReason-Pro, https://github.com/instadeepai/nucleotide-transformer, https://github.com/ColinFX/Prot2Text-V2, https://github.com/vandijklab/cell2sentence and https://github.com/epigen/cellwhisperer. The evaluation data is the disease prediction task of `wanglab/kegg` (Fallahpour et al., 2025), with the pathway-network split of Fesser et al. (2026) for Section 4.4, the ChatNT test tasks of `InstaDeepAI/ChatNT_training_data`, reviewed human proteins from UniProt with their InterPro and GO annotations, five single cell atlases from CELLxGENE and GEO accessions `GSE84133` and `GSE280502`, and the full Cell Ontology release for the cell type descriptions. The code for recreating all results and figures in this paper is available at https://anonymous.4open.science/r/biological-input-use-2D3C.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE, 2018.

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. In *International Conference on Learning Representations*, 2017. URL https://mlanthology.org/iclr/2017/alain2017iclr-understanding/.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, Apr 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10176-5. URL https://doi.org/10.1038/s41586-026-10176-5.

Sarah Burge, Elizabeth Kelly, David Lonsdale, Prudence Mutowo-Muellenet, Craig McAnulla, Alex Mitchell, Amaia Sangrador-Vegas, Siew-Yit Yong, Nicola Mulder, and Sarah Hunter. Manual go annotation of predictive protein signatures: the interpro approach to go curation. *Database*, 2012: bar068, 2012.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf.

Samir Char, Nathaniel Corley, Sarah Alamdari, Kevin K Yang, and Ava P Amini. Protnote: a multimodal method for protein–function annotation. *Bioinformatics*, 41(5):btaf170, 2025.

Bernardo P. de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, and Thomas Pierrot. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, 7(6):928–941, Jun 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01047-1. URL https://doi.org/10.1038/s42256-025-01047-1.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Sajib Acharjee Dip, Adrika Zafor, Bikash Kumar Paul, Uddip Acharjee Shuvo, Muhit Islam Emon, Xuan Wang, and Liqing Zhang. LLM4Cell: Taxonomy and evaluation of LLM and agentic models for single-cell biology. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41913–41954, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1942. URL https://aclanthology.org/2026.acl-long.1942/.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational*

Linguistics, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model, 2025. URL https://arxiv.org/abs/2505.23579.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712954. URL https://www.biorxiv.org/content/early/2026/03/20/2026.03.19.712954.

Xiao Fei, Michail Chatzianastasis, Sarah Almeida Carneiro, Hadi Abdine, Lawrence Paul Petalidis, and Michalis Vazirgiannis. Prot2text-v2: Protein function prediction with multimodal contrastive alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=w1FUXt3ujK.

Lukas Fesser, Hanlin Zhang, Michelle M. Li, Eric Wang, Bryan Perozzi, Shekoofeh Azizi, Sham M. Kakade, and Marinka Zitnik. How post-training shapes biological reasoning models, 2026. URL https://arxiv.org/abs/2606.16517.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9847–9857, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.775. URL https://aclanthology.org/2021.emnlp-main.775/.

Xin Gao, Cheng Yang, Chufan Shi, and Taylor Berg-Kirkpatrick. Do text edits generalize to visual generation? benchmarking cross-modal knowledge editing in UMMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cGqHvj7Q3B.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas K Garg. Canonicalizing multimodal contrastive representation learning. In *ICLR 2026 Workshop on Representational Alignment (Re-Align)*, 2026.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

Haonan He, Yuchen Ren, Yining Tang, Ziyang Xu, Junxian Li, Minghao Yang, Di Zhang, Yuan Dong, Tao Chen, Shufei Zhang, Yuqiang Li, Nanqing Dong, Wanli Ouyang, Dongzhan Zhou, and Peng Ye. Biology-instructions: A dataset and benchmark for multi-omics sequence understanding capability of large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty,

Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17984–18016, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.978. URL https://aclanthology.org/2025.findings-emnlp.978/.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Ala Jararweh, Oladimeji Macaulay, David Arredondo, Yue Hu, Luis E Tafoya, Kushal Virupakshappa, and Avinash Sahu. Protein2Text: Resampling mechanism to translate protein sequences into human-interpretable text. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 918–937, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. doi: 10.18653/v1/2025.naacl-industry.68. URL https://aclanthology.org/2025.naacl-industry.68/.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pp. 2023–09, 2024.

Paul Liang, Chun Kai Ling, Yun Cheng, Alexander Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Russ Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23072–23103, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/63f97950e32e6bdb4c6c9e31568b08c2-Paper-Conference.pdf.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. ProtT3: Protein-to-text generation for text-based protein understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5949–5966, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.324. URL https://aclanthology.org/2024.acl-long.324/.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12700–12710, June 2021.

Owen Queen, Yepeng Huang, Robert Calef, Valentina Giunchiglia, Tianlong Chen, George Dasoulas, LeAnn Tai, Gianmarco Abbadessa, Owain Howell, Michelle M. Li, Yasha Ektefaie, Ayush Noori, Ildiko Farkas, Joseph Brown, Tom Cobley, Karin Hrovatin, Tom Hartvigsen, Fabian J. Theis, Bradley L. Pentelute, James Zou, Vikram Khurana, David Owen, Richard Nicholas, Manolis Kellis, and Marinka Zitnik. Procyon: A multimodal foundation model for protein phenotypes. *bioRxiv*, 2025. doi: 10.1101/2024.12.10.627665. URL https://www.biorxiv.org/content/early/2025/11/03/2024.12.10.627665.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Till Richter, Eric Zimmermann, James Hall, Fabian J Theis, Srivatsan Raghavan, Peter S Winter, Ava P Amini, and Lorin Crawford. Beyond alignment: synergistic integration is required for multimodal cell foundation models. *bioRxiv*, pp. 2026–02, 2026.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, Jörg Menche, Eleni M. Tomazou, and Christoph Bock. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, Nov 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02857-9. URL https://doi.org/10.1038/s41587-025-02857-9.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, Hongyuan Lu, and Fajie Yuan. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 44(8):1310–1316, Aug 2026. ISSN 1546-1696. doi: 10.1038/s41587-025-02836-0. URL https://doi.org/10.1038/s41587-025-02836-0.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. InstructProtein: Aligning human and protein language via knowledge instruction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1114–1136, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.62. URL https://aclanthology.org/2024.acl-long.62/.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804/.

Juntong Wu, Zijing Liu, He Cao, Li Hao, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. Rethinking text-based protein understanding: Retrieval or LLM? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23726–23746, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1211. URL https://aclanthology.org/2025.emnlp-main.1211/.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. ProtST: Multi-modality learning of protein sequences and biomedical texts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38749–38767. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23t.html.

Xinyi Zhang, GV Shivashankar, and Caroline Uhler. Partially shared multi-modal embedding learns holistic representation of cell state. *Nature Computational Science*, 6(3):285–300, 2026.

Jiahao Zhao, Feng Jiang, Shaowei Qin, Zhonghui Zhang, Junhao Liu, Guibing Guo, Hamid Alinejad-Rokny, and Min Yang. Sc-arena: A natural language benchmark for single-cell reasoning with knowledge-augmented evaluation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 2063–2091, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/047c84ec50bd8ea29349b996fc64af4b-Paper-Conference.pdf.

A Datasets, Models, and Evaluation Methods

This appendix section records the datasets, checkpoints, prompts, scoring rules and uncertainty estimates used.

A.1 Evaluation Datasets and Preprocessing

A.1.1 Genomic datasets and evaluation splits

BioReason disease prediction task. The task uses the KEGG-derived dataset released with BioReason (Fallahpour et al., 2025) aggregated over all three splits (1,159 train, 146 test and 144 validation queries). The 1,449 queries cover 708 distinct reference/variant genome pairs, 186 distinct prompt texts and 37 disease labels. Each query comes with its prompt text, a reference sequence, a variant sequence and one gold disease. The prompt text has four variable fields: a `Chromosome Number`, a `Network Definition of the pathway` in which the variant gene is marked with an asterisk, `Genes in the pathway` with `SYMBOL; description` entries, and a query that names the variant gene and asks which disease the allele contributes to. The pathway definitions, pathway genes and disease labels are taken from KEGG. The dataset sequences for the reference and variant are encoded with Evo2. The disease label is the dataset's answer string, and the 37 labels form the closed vocabulary used for scoring (Appendix A.2). All 1,449 queries are evaluated because every comparison in Figure 2a is paired with the same query but with the input shuffled.

ChatNT tasks. ChatNT is evaluated on three of its own binary tasks from the released test split which contains three tasks: (1) splice donors (2,198 test queries, 1,098 Yes and 1,100 No, 600 bp), (2) splice acceptors (2,218 queries, 1,118 and 1,100, 600 bp) and (3) TATA promoters (621 queries, 332 and 289, 300 bp). From each task we sample up to 1,000 queries per class, giving 2,000, 2,000 and 578 queries (4,578 in total, balanced within task).

A.1.2 Protein datasets and annotation processing

Reviewed human proteome (BioReason-Pro). Proteins are obtained from UniProt with organism ID 9606 and filtered for reviewed entries. We keep only entries with curated GO ground truth, sequences of 120 to 1,024 residues (1,024 amino acids is the maximum sequence length for the GO-GPT encoder). This leaves 14,102 proteins, of which 14,019 have at least one InterPro entry. Ground truth is the GOA human annotation file (GAF 2.2, generated 2026-07-28), restricted to the evidence codes EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC, with NOT qualifiers dropped and high-throughput codes excluded, over all three aspects. The file holds 16,991 annotated proteins with 149,820 annotations (8.8 per protein).

Temporal holdout (BioReason-Pro). We use a cutoff of 2026-03-20, given by the day the BioReason-Pro preprint was posted. Proteins included in this holdout are reviewed UniProt entries of 120 to 1,024 residues added on or after the cutoff (706 entries). Entries whose GO cross-references carry only electronic evidence (615) are dropped, leaving 91 proteins, 90 of them with an InterPro entry. Labels are the UniProt GO cross-references with non-electronic evidence. The set spans several organisms (22 *Drosophila*, 18 *Taxus chinensis*, 9 *Danio* among others), and its proteins carry a median of 2 curated terms against 8.8 on the proteome. The level of $F_{\max}$ on the holdout is therefore not comparable with the proteome, and we interpret only the within-holdout contrast between intact and shuffled input.

Prot2Text-V2 test split. Prot2Text-V2 is evaluated on the official test split containing 4,203 proteins. The released pipeline truncates sequences to 1,021 residues. We exclude the 286 proteins longer than 1,021 residues so that a shuffled sequence is a permutation of the whole input, leaving 3,917 proteins for the perturbation experiment.

A.1.3 Single-cell atlases and cell selection

Cell sentences are generated by us from the five atlases that the C2S-Scale authors use for cell type annotation (Table S1). Each atlas is prepared as follows: the matrix holds $\log(1 + x)$ of the raw counts, gene names are upper-cased HGNC symbols, and the annotation is the atlas's own cell type column. Cells which have at least 1,000 expressed genes are drawn as a balanced sample across

annotated types excluding cells annotated `unknown` from each atlas. This results in 900, 1,049, 957, 939, and 1,001 cells for immune1, immune2, immune3, pancreas, and lung, and 4,846 cells in total. The 2,000 cells used for rationale analysis (Table 1) are 400 cells per atlas drawn without replacement from the 4,846 cells. CellWhisperer is evaluated on the same 4,846 cells.

Table S1: The five single-cell atlases. Shorthand labels are study names such as `alpha`, `beta`, `Basal` and `Ciliated`.

Atlas	Accession	Cells	Types	Options	Label source
immune1	CELLxGENE `62ef75e4`; E-MTAB-11536	30,423	35	34	Cell Ontology names
immune2	CELLxGENE `e9360edf`; GSE271896	33,923	38	38	Cell Ontology names
immune3	CELLxGENE `cc431242`; GSE299043	30,593	31	31	Cell Ontology names
pancreas	GSE84133 (Baron et al. 2016)	5,641	14	14	study shorthand
lung	GSE280502 (Kim et al. 2024)	6,160	7	7	study shorthand

For C2S-Scale we ask the model to select from an option list of all cell types in the atlas of the evaluated cell sentence. CellWhisperer is scored by the likelihood of a candidate answer string using the Tabula Sapiens format.

A.2 Evaluated models

BioReason. The RL checkpoint is the `bio-posttrain-qwen3-1.7b-dna-rl` checkpoint (Fesser et al., 2026), a Qwen3-1.7B model with the GRPO LoRA adapter (rank 16, $\alpha = 32$, 1,156 steps) merged, paired with a frozen Evo2 `evo2_1b_base` encoder (Brixi et al., 2026). The encoder uses a trained projection to the 2,048-d text width. Reference and variant bases are encoded separately, projected, included into $2 \times 2{,}048$ positions of the prompt. $\mathbf{Z}_{\mathrm{Evo2}}$ denotes the projected rows. The linear probes of Appendix B.3.1 read the encoder output at layer `blocks.20.mlp.l3` (1,920-d) and $\mathbf{Z}_{\mathrm{Evo2}}$, in each case mean-pooled over the 2,048 positions of the reference and of the variant block and concatenated. We also evaluate the SFT checkpoint, `mims-harvard/bio-posttrain-qwen3-1.7b-dna-sft` for RQ2. The input to BioReason is the two DNA representations of the wild type and variant followed by the same prompt text format, as in the released training code.

For disease prediction, the generated disease is lower-cased, stripped of punctuation other than hyphens, apostrophes and slashes, and mapped onto the 37-label vocabulary by substring containment with the longest label tried first. If the answer segment maps to nothing the whole generation is searched. A generation that maps to no label is wrong (1.5% of wild-type queries for the RL checkpoint). Accuracy is reported as the mean over all queries.

ChatNT. We use the released checkpoint which contains a 500M-parameter Nucleotide Transformer v2 encoder that tokenises the sequence as 6-mers (101 tokens for 600 bp, 51 for 300 bp), a Perceiver resampler produces 64 representations that are inserted at the DNA placeholder of the English prompt, and a frozen Vicuna-7B decoder produces the answer. $\mathbf{Z}_{\mathrm{NT}}$ denotes the 64 inserted representations.

The prompt is the released pipeline's context ("A chat between a curious user and an artificial intelligence assistant that can handle bio sequences ... USER: "), the query with its DNA placeholder (for example "Is a donor splice site present in the sequence <DNA>?"), and "ASSISTANT:". The score is the log-softmax over the first answer token at the last prompt position, with the predicted label being the larger value over the answer tokens " Yes" and " No".

BioReason-Pro. We use the released RL checkpoint (Fallahpour et al., 2026). ESM3 (`esm3_sm_open_v1`) is run inside generation on the sequence string its representations are projected and spliced at the protein placeholder, and the GO-GPT text is derived from a separate sequence field. The projector is a two-layer perceptron with a GELU nonlinearity that maps the layer-37 states ($(L+2) \times 1{,}536$) to the 2,560-d text width, and $\mathbf{Z}_{\mathrm{ESM3}}$ denotes its output. The probes of Appendix B.3.2 read the layer-37 states and $\mathbf{Z}_{\mathrm{ESM3}}$, each mean-pooled over all $L+2$ positions. We also evaluate the SFT checkpoint in RQ2.

The prompt is BioReason-Pro's own CAFA5 reasoning template with context: "Given the protein above from organism {organism} with the following InterPro annotations: {interpro} And the following initial GO term speculations: {go_speculations} Reason about the function of the protein and focus more on its Molecular Function, Biological Process, Cellular Component. Summarize in

UniProt format." The organism string is UniProt's organism name. The prediction is the set of GO identifiers outputted by the model.

Predicted GO sets are scored with the released `cafaeval` engine in the configuration BioReason-Pro uses (CAFA normalisation, maximum propagation of predictions and ground truth to the ontology roots over `is_a` and `part_of`, per-aspect scores averaged over the aspects present). We report `cafaeval`'s IA-weighted $F_{\max}$.

For the evidence conflicts each protein carries a specification over one GO term (required or forbidden). A prediction satisfies a specification when its propagated GO set contains a descendant of every required term and no descendant of a forbidden term, evaluated as a search over the ontology graph rather than as set membership.

Prot2Text-V2. The checkpoint is `Prot2Text-V2-11B-Instruct-hf` (Fei et al., 2025), with the ESM2 encoder `esm2_t36_3B_UR50D` and a Llama-3.1-8B decoder.

The prompt is the released system message ("You are a scientific assistant specialized in protein function predictions ...") and the user message "Protein name: {name} ; Taxon: {taxon} ; Sequence representations: " followed by the ESM2 tokens.

Each generated description is scored against the protein's UniProt function text with BERTScore F1 using `dmis-lab/biobert-large-cased-v1.1` at layer 24, after retokenising both texts to 495 tokens, which is the released benchmark script's setting. Scores for shuffled sequences are the mean over three shuffle seeds. For evidence conflicts, the same score is computed against both reference functions (Appendix B.2).

Cell2Sentence-Scale. The checkpoints are `C2S-Scale-Gemma-2-27B` and `-2B` (Levine et al., 2024). The biological input is the cell sentence $\mathbf{Z}_{\text{cell}}$, the cell's gene symbols in descending expression order, zero-count genes omitted, truncated at 1,000 genes.

The multiple-choice prompt is built from the opening sentence and answer cue of C2S-Scale's own cell type annotation template:

```
The following is a list of {num_genes} gene names ordered by descending expression
level in a Homo sapiens cell. Your task is to give the cell type which this cell
belongs to based on its gene expression.
Choose your answer from the following list of cell types:
- {option 1}
- {option 2}
...
Cell sentence: {cell_sentence}.
The cell type corresponding to these genes, copied verbatim from the list above, is:
```

The option list is the atlas's full label vocabulary, shuffled for each cell sentence. Options precede the cell sentence. In evidence conflicts a `Context:` line holding the description of the cell type is inserted between the task sentence and the option list. The generated cell type name is mapped to an option by exact match, then case- and punctuation-normalised match, then unique containment, then the uniquely longest contained option, then prefix. A generation that names no offered option is counted as non-compliant and wrong.

CellWhisperer. The transcriptome encoder is the released CLIP checkpoint `cellwhisperer_clip_v1.ckpt` (Schaefer et al., 2025), with the Geneformer model as the transcriptome encoder. Raw counts are normalised per cell to 10,000, divided by the released gene medians, ranked, and truncated at 2,048 tokens by the released processor. The 2,048-d transcriptome projection is L2-normalised, as in the released training pipeline, and denoted $\mathbf{Z}_{\text{CW}}$. The chat model is the released default checkpoint (`results/chat_models/default`), a LLaVA model on Mistral-7B-Instruct-v0.2 whose two-layer projector maps $\mathbf{Z}_{\text{CW}}$ to 8 prompt tokens.

Prompts use the released Mistral-instruct conversation template. For queries with input gene text, the prompt is prefixed by the top-genes exchange. The answer tokens of the final assistant turn are scored by the mean log-likelihood per token. The lowest perplexity answer is taken as the model's choice.

The released CellWhisperer evaluation ranks the perplexity of the correct answer under the cell's own transcriptome against its perplexity under 30 randomly drawn transcriptomes of other cell types. For cell i with label L_i and transcriptome condition c (intact or count-shuffled), the score is the share of cells j in the same atlas with $L_j \neq L_i$ for which the per-token log-likelihood of the answer "This cell

is a L_i" under $\mathbf{Z}_i^{(c)}$ exceeds its log-likelihood under the intact $\mathbf{Z}_j$. We report the 1 minus perplexity quantile. The candidate answers are the atlas's canonical labels (Appendix A.1.3). Evidence conflicts are scored with the same per-token quantity to compare the two candidate answers of a pair.

Implementation. All checkpoints were run on NVIDIA H100 GPUs, except ChatNT on NVIDIA A100 GPUs.

B RQ1: Biological Input Perturbations, Evidence Conflicts, and Predictive Baselines

B.1 Biological Input Perturbations

DNA. The DNA sequence is permuted, so length and base composition are preserved but the edited position is not. Both the reference and the variant block are shuffled. The shuffled sequences are encoded and projected into the prompt. The prompt text is kept the same.

Protein. For BioReason-Pro, the amino acid sequence is permuted then embedded by ESM3 while for GO-GPT, InterPro, and organism we use the wild type sequence. For Prot2Text-V2 the string is shuffled and encoded by ESM2.

Table S2: **BioReason prediction changes under DNA shuffling.** Each of the 1,449 disease prediction queries falls in exactly one cell, by whether the label the scorer maps the generation to is correct with the input intact and after both Evo2 windows are shuffled. Counts, with per cent of 1,449 in brackets. Net is the arm's accuracy minus the intact accuracy, which equals (incorrect→correct − correct→incorrect)/1,449. Re-decoding the identical prompt changes 0 of 1,449 labels at both checkpoints.

Label transition, intact → shuffled	**RL**	**SFT**
correct → correct	1,192 (82.3)	1,094 (75.5)
correct → incorrect	35 (2.4)	99 (6.8)
incorrect → correct	36 (2.5)	86 (5.9)
incorrect → incorrect, same label	156 (10.8)	126 (8.7)
incorrect → incorrect, other label	30 (2.1)	44 (3.0)
Accuracy, intact	0.847	0.823
Accuracy, shuffled	0.847	0.814
Net accuracy change	$\mathbf{+0.001}$	$\mathbf{-0.009}$

Table S3: **Proteins whose BioReason-Pro output changes under ESM3 shuffling.** The GO terms emitted from the intact input are compared with those emitted from the shuffled input, with the GO-GPT and InterPro text retained in both. The first row and the Jaccard are over all 14,102 proteins; the F_1 rows are over the 14,086 with a scorable aspect, where F_1 is the per-protein information-accretion-weighted F_1 against that protein's GOA annotation set propagated over the ontology.

Quantity	**Value**
Proteins whose emitted GO set changes	6,593 (46.8%)
Proteins whose F_1 improves	1,575 (11.2%)
Proteins whose F_1 worsens	1,527 (10.8%)
Proteins whose F_1 is unchanged	10,984 (78.0%)
Mean Jaccard of the emitted GO term sets	0.889

Gene sentence and transcriptome perturbations. For C2S-Scale, gene order shuffling permutes the positions of the cell's top 1,000 genes. Resampling from all expressed genes draws 1,000 genes uniformly without replacement from all of the cell's expressed genes and lists them in the cell's own expression order, so the length and the rank structure are preserved and the identity of the visible genes changes. The resampled window retains a mean of 0.57 to 0.59 of the intact top-1,000 genes in the immune atlases and 0.14 in lung, where a cell expresses a median of 7,709 genes. The cell sentence is used with the same query and scored against the annotated cell type of the original cell sentence.

For CellWhipserer, we permute the cell's full count vector across genes. The permuted counts pass through the Geneformer processor (normalisation, median scaling, ranking, truncation at 2,048 tokens) and CLIP tower, so $\mathbf{Z}_{\mathrm{CW}}$ is recomputed from the permuted gene counts. The prompt is the representation-only prompt of Appendix A.2, and the scored quantity is the perplexity quantile of Appendix A.2 against the intact representations of the other cells of the same atlas.

Biological text removal and masking. Table S4 provides for each text condition drawn in Figures 2 and 3.

Table S4: Text fields present in each text condition. ✓ present, × removed, M masked to `GENE`.

Model	Condition (figure label)	Fields			
BioReason		Chromosome	Network	Gene list	Gene in query
	all text (`wt`)	✓	✓	✓	✓
	pathway fields, symbols masked (`no_gene`)	✓	M	M (descriptions kept)	M
	query naming the gene (`no_pathway`)	×	×	×	✓
	query, gene symbol masked (`no_textkey`)	×	×	×	M
BioReason-Pro		GO-GPT	InterPro	Organism	
	GO-GPT & InterPro (`wt`)	✓	✓	✓	
	GO-GPT (`gogpt_only`)	✓	×	✓	
	InterPro (`interpro_only`)	×	✓	✓	
	None (`no_symbolic`, `no_modality`)	×	×	✓	
Prot2Text-V2		Protein name	Taxon		
	name and taxon given (`native`)	✓	✓		
	name and taxon removed (`seq_only`)	×	×		
CellWhisperer		Top-1,000 gene text list			
	representation only (Fig. 2f)	×			
	representation and gene text (Fig. 3e)	✓			

B.2 Evidence Conflicts

BioReason. We only use queries the model answers correctly with the unmodified input. Three conflicts from (A, B) are constructed, all with both DNA representations from A: pathway input and the gene named in the query from B (1,216 conflicts), pathway input from B with the gene named in the query from A (1,185), and pathway input from A with the gene named in the query from B (1,179). The counts differ because a query is dropped when its answer names no disease in the label set, on 11, 40 and 48 queries. Fig. 3a is labelled with the largest of the three. The pathway input is the chromosome, network and gene list inputs. It matches A when it equals A's disease label, matches B when it equals B's disease label, and is neither otherwise.

BioReason-Pro. A pair is two wild-type proteins of the same organism that share an InterPro family entry and differ in one functional property, scored by one readout GO term. We use four properties: enzyme vs. pseudoenzyme (111 pairs), organelle targeting with a N-terminal transit peptide or not (122 pairs), DNA binding vs. not-binding (76 pairs, `GO:0003677`) and GO `NOT` (96 pairs). For the first three properties, one protein carries the readout term in its curated UniProt GO annotation and the other lacks it, so the negative member is defined by the absence of the annotation. The pseudoenzyme is additionally named "Inactive" in UniProt or listed as a non-peptidase homologue in MEROPS. These three properties contain 540 proteins. For GO `NOT` pairs, drawn from the human GOA file, one protein has an experimental annotation to a GO term and the other has an experimental `NOT` annotation to the same term (evidence codes EXP, IDA, IPI, IMP, IGI or IEP) and no UniProt annotation to the term or its descendants. These 96 pairs contain 150 proteins and span 50 GO terms, so no probe can be meaningfully trained on them (Appendix B.3.2). A pair is used for the the evidence conflict experiment only if the model correctly annotates both proteins in the pair resulting in 58, 94 and 37 pairs and each contributes two directional conflicts, giving 378 evidence conflicts. For GO `NOT`, 53 pairs pass and give 106 evidence conflicts (Fig. S1). For the pair (A, B) with $\mathbf{Z}_{\mathrm{ESM3}}$ from A, three conflicts are constructed: GO-GPT and InterPro both from B, InterPro from B with GO-GPT from A, and GO-GPT from B with InterPro from A. The organism string follows A and is the same for both members. A prediction matches A when it satisfies A's specification and not B's, matches B when it satisfies B's and not A's.

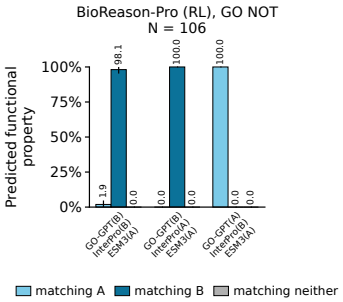

Figure S1: **Evidence conflicts for BioReason-Pro (RL) on GO NOT pairs.** $\mathbf{Z}_{\mathrm{ESM3}}$ from A, GO-GPT/InterPro from A or B, where one protein of each pair has an experimental annotation to a GO term and the other an experimental NOT annotation to the same term. 106 conflicts from 53 pairs. Bars indicate preference for A, B, or neither. Error bars are 95% CIs.

Prot2Text-V2. We use unordered pairs of the 3,900 test proteins that share a taxon, correctly annotates both proteins in the pair, has different protein names, each protein name scores higher against its own function than against the partner's, and for both members the annotation from the complete prompt scores higher against its own function than against the partner's. This results in 1,264 directional conflicts. We run the following evidence conflicts: $\mathbf{Z}_{\mathrm{ESM2}}$ of A with B's name, and $\mathbf{Z}_{\mathrm{ESM2}}$ of B with A's name, taxon unchanged. The generated description is scored by BioBERT BERTScore F1 against A's and against B's reference function. It matches the sequence source when its score against the sequence protein's function exceeds its score against the name protein's function by more than 0.005, matches the name source in the opposite case, and is neither when the two scores lie within 0.005 of each other.

C2S-Scale. Pairs are obtained from the immune3 atlas as it has the largest number of valid pairs. Cell type descriptions are obtained from Cell Ontology relationship axioms, covering surface markers, capabilities, anatomical location and developmental origin with a maximum of six facts per description. Ancestor and descendant pairs are excluded, and a pair needs at least two contrastive and three shared facts. A pair is valid if the description of B alone makes the 27B model select B over A in a two-way choice. Conflicts are scored on the 1,152 cells the model annotates as A from the cell sentence alone. The cell sentence of a type-A cell is combined with a description of type B and all cell type labels from the atlas are given. The predicted cell type matches A when it equals the cell's annotated label, matches B when it equals the partner type, and is neither otherwise.

CellWhisperer. In each atlas 200 pairs of cells with different labels are formed by preferring pairs from the same tissue and donor, then the same tissue, then the same donor, and from the same tissue. We keep pairs for which each member is identified correctly from its complete input, $\mathbf{Z}_{\mathrm{CW}}$ together with its own top-1,000 gene list. This holds for 330 pairs, giving 660 directional conflicts. We $\mathbf{Z}_{\mathrm{CW}}$ of A with $k = 1{,}000$ from B's gene list and vice versa. For each conflict, the two candidate answers "This cell is a {type of A}" and "This cell is a {type of B}" are scored by per-token log-likelihood. The prediction is given by the answer with lower perplexity.

B.3 Predictability of Underused Biological Representations

B.3.1 Genome-dependent disease prediction subset

Queries are grouped by the same prompt text (186 groups). A group is genome-dependent when it contains more than one distinct genome pair and more than one disease label. All other queries are genome-invariant, which means that every query with that text has the same label and a text-only predictor can score perfectly. The genome-dependent stratum has 165 queries in 14 text groups over 63 genomes and 12 labels, from six genes. With 126 train, 19 test and 20 validation queries. Within a group 118 queries carry the group's strictly most frequent label, 41 carry a less frequent label, and 6 queries belong to one RET group whose two labels tie.

Evo2 probes. We train linear probes on both the Evo2 representations as well as the projected Evo2 representations received by BioReason, $\mathbf{Z}_{\mathrm{Evo2}}$. In both, the reference and variant blocks are

mean-pooled over their 2,048 positions and concatenated. The classifier is a multinomial logistic regression with L2 regularisation on standardised features, without class weighting. We use five folds, grouped by genome, so a held-out variant's label must be inferred from other variants.

Text-only predictor. The text-only predictor is fitted inside the same cross-validation loop as the probe. For each fold it assigns to a held-out query the most frequent gold label of that query's text among the queries in the training fold. Averaged over the 10 fold assignments it scores 0.708 on the 165 queries.

B.3.2 InterPro-controlled protein subset and functional targets

We use the 540 proteins from the evidence conflict across the three properties with 199 proteins in 31 InterPro families for enzyme activity, 232 in 87 InterPro families for organelle targeting and 109 in 22 InterPro families for DNA binding, and 140 InterPro families in total. We use binary labels based on the presence of the GO term of the property `GO:0003824` or `GO:0008233` for enzyme activity, `GO:0005739` or `GO:0009507` for organelle targeting, and `GO:0003677` for DNA binding. A negative label means the curated GO annotation lacks the term. GO `NOT` pairs are not used because their readout terms differ between pairs, so no single label is shared across them.

We train the linear probe on predicting both the ESM3 representations and the projected representations that are used by BioReason-Pro. For each we use the mean-pooled representation over all residues. The classifier is a logistic regression with L2 regularisation at the library default on standardised features, fitted separately for the three properties, with five folds grouped by InterPro family. We score with AUROC within each InterPro family, and InterPro families are averaged with equal weight.

B.4 Removing differentially expressed genes from cell sentences

DEGs are computed per atlas, from the $\log(1+x)$ matrix with scanpy's Wilcoxon rank-sum test of each annotated type against all other cells. A gene is a DEG of a type when its $\log_2$ fold change is at least 1.0 and its Benjamini–Hochberg adjusted p-value is below 0.05. The fold change criterion is one-sided because the cell sentence is ranked by decreasing expression. DEGs are ranked by $\log_2$ fold change times $-\log_{10}$ of the adjusted p-value. For a given cell, its DEG set is the DEGs of its annotated type that occur in its own 1,000 gene sentence. The size of a cell's DEG set varies by atlas: the median is 82, 66, 55, 136 and 97 genes in immune1, immune2, immune3, pancreas and lung.

We remove proportions $f \in \{0.10, 0.25, 0.50, 0.75, 1.00\}$ of the cell's DEGs from the cell sentence. We compare four different ways of removing DEGs: (1) the highest ranked DEGs, (2) the weakest ranked DEGs (for $f \leq 0.75$ because at $f = 1$ the set equals the set of strongest DEGs), (3) random genes drawn uniformly from the cell's own sentence, and (4) non-DEGs with similar expression given by the non-DEG of the same sentence with the nearest expression value in that cell to a DEG. Removed genes are deleted, not replaced, so the sentence shortens by the same number of genes under all four conditions.

B.5 Discussion of training-time availability of predictive text

Here we compare the six models for training-time availability of predictive text. For BioReason, each disease prediction query provides the wild type and variant DNA sequence representations followed by the chromosome, pathway definition, gene list with descriptions and the query naming the gene. No masking, dropout or augmentation of the inputs is applied. A majority-vote table from the gene symbol named in the query to the disease, fitted on the train split, results in test set accuracy of 0.932 across the 91 genes which all occur in training. Therefore, the text inputs that are always present at training time is sufficient for the task. The BioReason-Pro training script uses InterPro and the GO-GPT prediction. No dropout of either input is used. ChatNT's training queries give the task and the sequence representation. The query does not include details about the sequence, so no sequence-specific text can predict the label. Prot2Text-V2's released data loader replaces the protein name and the taxon each by `unknown` with probability 0.8, independently, at every training step. CellWhisperer's chat model is trained on conversations paired with a transcriptome representation, so the transcriptome representation is the input present in every prompt. C2S-Scale is trained to predict a cell type given only the cell sentence.

Across the six models, the two models with training prompts that always contain answer-predictive text (BioReason and BioReason-Pro) show nearly unchanged performance under sequence shuffling and prefer the text in evidence conflicts. Conversely, the four models trained without such text (ChatNT, CellWhisperer, C2S-Scale) or with high dropout rates (Prot2Text-V2) show reduced performance under biological input perturbations and, where evaluated, prefer the sequence or cell source in conflicts. We report this as an observation only. The models differ in architecture, task, training data and scale, so this comparison cannot isolate the effect of training prompt content, and we do not test it with a controlled intervention.

C RQ2: POST-TRAINING, MODEL SCALE, AND BIOLOGICAL INPUT CONTRIBUTION TO PERFORMANCE

C.1 BIOREASON POST-TRAINING SWEEP CHECKPOINTS

The 42 checkpoints come from the post-training sweep of Fesser et al. (2026) which includes the two final SFT and RL checkpoints. The language models are Qwen3-1.7B and Qwen3-4B. For each model size, a SFT sweep with LoRA rank $r \in \{16, 64, 256\}$ ($\alpha = 2r$, learning rate 5×10^{-5}, at most 4 epochs) and GRPO sweep with LoRA rank $r \in \{4, 16, 64\}$ (learning rate 10^{-5}, 8 generations per prompt, 4 epochs or 1,156 steps) runs from each SFT run. The 42 checkpoints are 6 SFT and 36 GRPO checkpoints from 24 training runs. We use the same set up as defined in Appendix B.1.

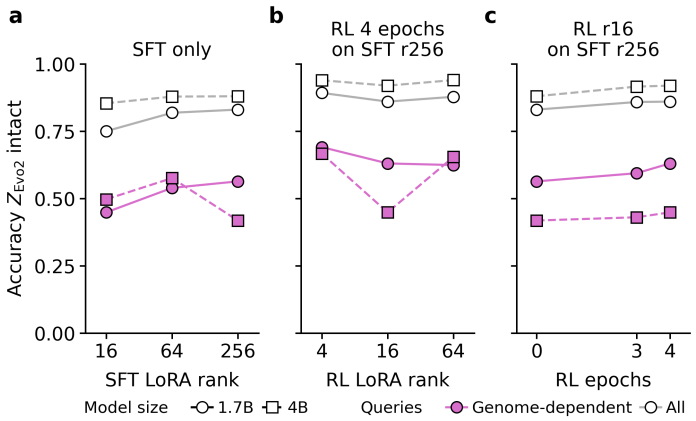

Figure S2: **Post-training changes BioReason performance without consistently increasing the performance contribution from $\mathbf{Z}_{\mathrm{Evo2}}$. a–c** Accuracy on the genome-dependent subset and all 1,449 queries as SFT LoRA rank, RL LoRA rank, or RL duration is varied.

Across 42 BioReason checkpoints, post-training changes accuracy without consistently increasing the performance contribution from $\mathbf{Z}_{\mathrm{Evo2}}$. We broaden the comparison beyond the released SFT and RL checkpoints by evaluating 42 BioReason checkpoints (Fesser et al., 2026) that vary model size, SFT LoRA rank, RL LoRA rank, and RL duration (Fig. S2). We focus on the 165 genome-dependent queries, where identical prompt text occurs with different genomes and different disease labels. Performance varies substantially across post-training configurations, but higher accuracy does not consistently coincide with a larger advantage for the original $\mathbf{Z}_{\mathrm{Evo2}}$ over the representation recomputed from shuffled DNA (Fig. 6b).

This observation holds across individual post-training sweeps. Increasing SFT LoRA rank changes accuracy on the 165 queries from approximately 0.45–0.56 for the 1.7B model and 0.42–0.58 for the 4B model, while overall disease prediction accuracy increases to approximately 0.75–0.88 (Fig. S2a). Varying RL LoRA rank from 4 to 64 or increasing RL duration from 0 to 4 epochs likewise changes both overall and genome-dependent accuracy (Fig. S2b,c), without consistently increasing the advantage of the original genomic representation. Post-training can improve disease prediction accuracy, including on queries where the genome distinguishes the target label, without increasing the performance contribution from $\mathbf{Z}_{\mathrm{Evo2}}$.

C.2 C2S-Scale 2B and 27B

Both checkpoints are evaluated on the same 4,846 cells from RQ1. Both models receive the same questions with the same set up outlined in Appendix B.4.

D RQ3: Reasoning traces and rationales

D.1 BioReason reasoning traces

We evaluate the released RL and SFT checkpoints on the 145 queries held out by the pathway network split. Each query is generated three times, with the original reference and variant windows, with the identical prompt a second time, and with both windows shuffled. Decoding is deterministic for both checkpoints, so any difference under shuffling follows from the input rather than from sampling. Of the 145 queries.

xWe use deterministic string matching to extract the stated base change, coordinate, chromosome, and gene from each reasoning trace, retaining the first match for each claim type. Accuracy on the stated base pair is reported over the 111 queries whose variant is a single base substitution, since a single base pair change cannot match a deletion, an insertion or a multi-base change. The coordinate in the reasoning trace is compared with the position given in the dataset, which is available for 137 of the 145 queries. The input prompt provides the chromosome number and the two 2,049 bp windows and absolute position is not provided.

D.2 BioReason-Pro reasoning traces

We use the evidence conflicts of Appendix B.2 in the direction that takes the InterPro from protein B and uses protein A for the $\mathbf{Z}_{\mathrm{ESM3}}$ and the GO-GPT input, on the 189 pairs where the model recovers both members' property from their own unperturbed prompts. We keep the evidence conflict pairs by mapping the InterPro accessions through the EMBL-EBI `interpro2go` file and taking the `is_a` and `part_of` closure of the property term in the Gene Ontology. The reasoning model is not given this mapping from InterPro to GO terms. The InterPro accession of protein B corresponds the evaluated GO term on 66 of the 189 pairs, and on 53 of those neither InterPro accessions nor its GO-GPT output of the protein A contains the GO term.

We use deterministic string matching to extract the InterPro accessions it references in the reasoning trace by matching on `IPR` followed by six digits. This reference InterPro accessions is equal to the input InterPro accessions from protein B on 53 of 53 queries, so quoting the reasoning model always references the accession that maps to the property protein A does not have. We also use string matching to determine whether a sentence quoting that accession states the property it implies. We delete every InterPro entry name printed in the prompt from the sentence and apply one fixed pattern per property term. A match is kept only if the matched phrase is not part of any printed entry name. The property of protein B is therefore stated in 23 of 53 reasoning traces. However, all 53 queries omit the GO term this property implies in the final response.

D.3 Cell type annotation rationales with Cell2Sentence

D.3.1 Rationale generation

We sample 400 cells per atlas from the 4,846 annotation cells of Table 1. We use the same DEG perturbations as defined in Appendix B.4: the unmodified sentence of the 1,000 most highly expressed genes in descending order, removal of the strongest DEGs of the annotated type at 50% and 100%, removal of the same number of non-DEGs with the closest expression value at 50% and 100%, and the unmodified gene set under one uniform random permutation per cell so the input position is independent expression. We include the random permutation to determine whether genes cited in the rationale are preferentially selected because of their expression level or their position in the expression-ranked input.

Every prompt has the form

```
The following is a list of {n} gene names ordered by
descending expression level in a Homo sapiens cell.  Your
task is to give the cell type which this cell belongs to
based on its gene expression.
Cell sentence:  {genes}.
Answer using exactly this format:
Cell type:  <cell type>
Rationale:  <the genes that support the call>

Cell type:
```

where `{genes}` is the space-separated gene sentence and `{n}` its length.

For C2S-Scale 2B with the same template, decoding settings, and conditions, no generated response contains the string `Rationale`. Conversely, for the 27B checkpoint, 12,000 generations (2,000 cells $\times$ six conditions) contain the `Rationale:` field. RQ3 therefore reports the 27B checkpoint only.

D.3.2 Gene extraction and analysis of referenced genes

We extract referenced genes from the full generation. We split the text into tokens that start with a letter or digit and continue with letters, digits, hyphens, or periods, upper-case each token, remove one trailing period, and keep the tokens that are in the vocabulary. Repeated mentions count once, at their first occurrence, and the order of first occurrences defines the mention order used for first-k precision.

For a cell with annotated type t, let S be the genes of the input cell sentence, D_t the DEGs of t, and $g_1, g_2, \ldots$ the referenced genes in order of mention with $G_k = \{g_1, \ldots, g_k\}$. DEG precision among the first k referenced genes is given by,

$$\mathrm{Prec@}k = \frac{\sum_c |G_k \cap D_t \cap S|}{\sum_c |G_k|}.$$

The random reference, the dashed line in Fig. S3b and the hatched bars in Fig. 7d, is the expected precision of the same number of genes drawn uniformly from S.

Let G and G' be the sets of referenced genes under the unmodified and perturbed sentences and R the set of genes removed from the input. Fig. S3c reports the Jaccard similarity $|G \cap G'|/|G \cup G'|$, computed per cell and averaged over cells.

Ribosomal genes. Genes are annotated as ribosomal protein genes by membership in HGNC gene groups 646, 728, and 729 (L, S, and mitochondrial ribosomal proteins) this results in a total of 171 approved symbols. Fig. S3a assigns each referenced gene to one category in the order (1) absent from S, (2) DEG of t in S, split into ribosomal and non-ribosomal, (3) ribosomal non-DEG in S, and (4) other gene in S. The proportion of ribosomal counts every referenced ribosomal protein gene, including ribosomal DEGs and ribosomal genes absent from S. The cell sentence bars of Fig. S3a are the expected composition of a same-size uniform draw from S.

D.3.3 Instructing the rationale to name DEGs

We further test whether their DEG precision depends on that query by testing seven queries that ask for marker genes or DEGs, on 200 cells (40 per atlas). Each cell is compared with its own generation under the unmodified condition. On these 200 cells the original query gives a DEG precision of 11.2%, a precision among the first five referenced genes of 23.2%, and a ribosomal share of 77.9%. The seven queries replace the answer format ending, and where it is stated with respect to the cell sentence.

Request for marker genes. `<genes from the cell sentence that are specifically expressed in this cell type and distinguish it from other cell types>`.

Request for differentially expressed genes. `<genes from the cell sentence that are more highly expressed in this cell type than in other cell types>`.

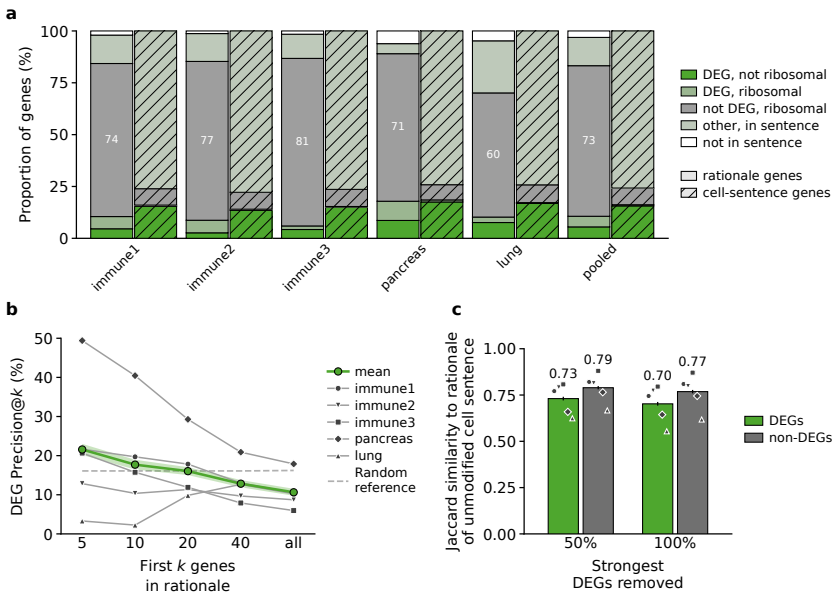

Figure S3: **Genes referenced in C2S-Scale rationales, and how that set changes when DEGs are removed from the cell sentence. a** Composition of referenced genes across five atlases. **b** DEG precision among the first k referenced genes. **c** Jaccard similarity to the unmodified set of referenced genes after removing the strongest DEGs or non-DEGs with similar expression from $\mathbf{Z}_{\text{cell}}$. Error bars are 95% CIs.

Request for marker genes with exclusions. `<up to 10 genes from the cell sentence that are specific to this cell type; do not list ribosomal, mitochondrial or housekeeping genes>.`

Same as above, instruction before the sentence. The same request placed in the task sentence rather than the answer format.

Annotated cell type given, request for marker genes only. No cell type is generated. The prompt states the atlas label and asks for its marker genes, with the same exclusions of ribosomal, mitochondrial or housekeeping genes.

Predicted cell type given, request for marker genes only. As above, with the type the model produced for that cell under the unmodified condition in place of the atlas label.

One-shot learning. An `Example:` block shows another cell of the same atlas with a different annotated type, its sentence cut to 200 genes, its type, and its ten strongest DEGs as marker genes, followed by the evaluated cell under `Now the cell to annotate:`.

Table S5: **Variation on the rationale query and DEG precision of C2S-Scale 27B.** Referenced genes per generation, ribosomal protein share, DEG precision against the annotated type (random draw in parentheses), precision among the first five referenced genes (Prec.@5), and cell type accuracy, pooled over 200 cells (40 per atlas). Accuracy is undefined when the type is supplied in the prompt.

Query	Genes	Ribo. (%)	DEG prec. (%)	Prec.@5 (%)	Acc.
original query	66.5	77.9	11.2 (16.2)	23.2 (16.1)	0.40
Request for marker genes	62.5	72.6	12.0 (16.4)	22.3 (16.3)	0.41
Request for differentially expressed genes	48.4	71.3	13.2 (16.3)	20.6 (16.4)	0.42
Request for marker genes with exclusions	61.6	69.9	13.3 (16.3)	21.8 (16.3)	0.42
Same as above, instruction before the sentence	65.8	74.9	11.6 (16.3)	23.9 (16.1)	0.40
Annotated cell type given, request for marker genes only	65.5	69.8	13.3 (16.2)	20.9 (16.1)	–
Predicted cell type given, request for markers only only	65.0	69.0	12.2 (16.3)	20.1 (16.1)	–
One-shot learning	57.0	64.5	10.4 (16.1)	14.5 (16.2)	0.18

Pooled DEG precision is 10.4 to 13.3% across the seven queries, compared to 11.2% for the original query on the same cells while DEGs of the annotated cell type comprise 16.2% of the genes in the cell sentence. The largest paired increase is 2.1 points for the query that requests at most ten marker genes and excludes ribosomal, mitochondrial, and housekeeping genes. This also increases precision among the first five referenced genes by -1.4 points. One-shot learning also lowers accuracy from 0.40 to 0.18.

D.3.4 Rationale content by prediction correctness

We split the rationales for the unmodified cell sentences by whether the predicted cell type is correct and compute DEG precision against the DEGs of the annotated type and of the predicted type. A prediction is correct when the generated label and the annotated type resolve to the same Cell Ontology term. The predicted type is the atlas type named by the generated label and is defined for 1,200 of the 2,000 cells. We report the same quantities for the prompt with cell type options of Appendix D.3.5. The prompt that supplies the annotated type and requests only marker genes (Appendix D.3.3) generates no label and is reported as a separate condition.

Table S6: **DEG precision of C2S-Scale 27B rationales by prediction correctness.** Precision among the first five referenced genes (Prec.@5) and among all referenced genes (Prec.@all), in %, with the precision of a uniform random draw of the same number of genes from the cell sentence in parentheses. Annotated and predicted columns use the DEGs of the annotated and the predicted cell type. Pooled over five atlases. – marks rows that include cells whose label names no atlas type.

		Annotated type		Predicted type	
Cells	N	Prec.@5	Prec.@all	Prec.@5	Prec.@all
Original prompt					
All	2,000	21.6 (16.1)	10.6 (16.2)	–	–
Correct prediction	748	35.2 (19.9)	16.9 (20.0)	35.4 (20.6)	17.4 (20.7)
Incorrect prediction	1,252	13.4 (13.8)	6.9 (13.9)	–	–
predicted type in atlas	452	13.7 (9.4)	5.0 (9.5)	24.2 (9.9)	8.1 (10.1)
Prompt with cell type options					
All	2,000	20.5 (16.1)	10.7 (16.2)	–	–
Correct prediction	821	33.1 (18.3)	15.2 (18.4)	33.6 (20.0)	16.4 (20.0)
Incorrect prediction	1,179	11.7 (14.5)	7.5 (14.6)	–	–
predicted type in atlas	622	13.7 (10.1)	5.7 (10.0)	14.9 (7.8)	8.6 (7.8)

For C2S-Scale 27B with the original prompt, precision among the first five referenced genes is 35.2% for correct predictions against 19.9% for the random draw, and 13.4% for incorrect predictions against 13.8%. For the 452 incorrect predictions whose label names an atlas type, precision among the first five genes is 24.2% against the DEGs of the predicted type and 13.7% against those of the annotated type, with random draws of 9.9% and 9.4%.

D.3.5 Effect of the rationale request on the predicted label

The rationale prompt of Appendix D.3.1 has no option list and a different answer format from the multiple-choice prompt of Appendix A.2. For the same 2,000 cells we generate two responses from prompts that share the task sentence, the option list in the order of the multiple-choice evaluation, the cell sentence, and the answer format, and differ only in the `Rationale:` line:

```
Answer using exactly this format:
Cell type:  <cell type, copied verbatim from the list above>
[Rationale:  <the genes that support the call>]

Cell type:
```

The label is the first line of the generation. Accuracy uses the Cell Ontology rule of Appendix D.3.4, because the pancreas and lung options are study labels such as `alpha` and the model answers with Cell Ontology names under both prompts.

Table S7: **Change in the C2S-Scale 27B predicted label when the rationale is requested.** Share of cells whose label string or resolved Cell Ontology type differs between the two prompts, cell type accuracy without and with the rationale request, and the number of cells whose prediction changes from correct to incorrect and from incorrect to correct. 400 cells per atlas.

	Changed (%)		Accuracy		Cells	
Atlas	Label	Type	w/o rationale	w/ rationale	Correct to incorrect	Incorrect to correct
immune1	21.5	9.8	0.71	0.61	49	7
immune2	62.5	44.0	0.12	0.08	27	8
immune3	16.0	9.8	0.53	0.44	37	3
pancreas	18.0	9.8	0.65	0.60	22	2
lung	34.5	18.0	0.26	0.34	5	37
Pooled	30.5	18.2	0.45	0.41	140	57

For C2S-Scale 27B, adding the rationale request changes the label string for 30.5% of the 2,000 cells (95% CI 28.5–32.5) and the resolved cell type for 18.2% (95% CI 16.6–19.9). Pooled accuracy decreases from 0.45 to 0.41, a difference of -4.2 points (95% CI -5.5 to -2.8). Accuracy decreases in four atlases and increases from 0.26 to 0.34 in lung, where the four basal subtypes resolve to one Cell Ontology term.

E RQ4: Auxiliary sequence supervision

E.1 Data

We use the pathway network split of the 1,449 KEGG queries from the BioReason post-training code base (Fesser et al., 2026) which includes 1,159 training, 145 validation, and 145 held-out queries. Since the split holds out pathway networks, in Table S8 we report the overlap for other properties with the held-out set. We use genome, as in Table 1, for a distinct pair of reference and variant windows. For 135 of the 145 held-out queries the same genome occurs in training under a different prompt, and 133 of these have the same disease label there. The held-out queries cover 8 of the 37 disease labels with 60 of them labeled amyotrophic lateral sclerosis. The mutations are 111 single-base substitutions, 20 deletions, 9 insertions, and 5 multi-base substitutions.

Table S8: Overlap between the 145 held-out queries and the 1,159 training queries.

Held-out property	**Distinct values**	**Also in training**	**Queries with the value in training**
Prompt text	12	0	0 / 145
Network definition	9	0	0 / 145
Gene named in the query	8	7	139 / 145
Genome (both windows)	100	90	135 / 145
Disease label	8	8	145 / 145

E.2 Auxiliary target and training

Both windows of a query start at the same coordinates with 128 bp either side for the 257-bp condition and 1,024 bp before and 1,023 bp after for the 2,048-bp condition, so the edit sits at the same fixed offset for training and evaluation. For an insertion or deletion the two cuts agree before the edit and are shifted relative to each other after it. The auxiliary target replaces the reasoning trace with one line, `Edit at` i`: ref X -> var Y (class)`, where i is the offset of the edit, X and Y are the reference and variant bases there, and the class is substitution, insertion, deletion, or multi-base. The control keeps the benchmark's reasoning trace, which has a median of 2,197 characters and names the substituted bases in 865 of the 1,159 training queries. The bases are available only from the input DNA representations. One training query of the 257-bp run is

```
<|im_start|>user
<|dna_start|>[257 slots, reference]<|dna_end|>
```

```
<|dna_start|>[257 slots, variant]<|dna_end|>Chromosome Number: 16
Network Definition of the pathway: mRNA(pre) -- FUS* // mRNA
Genes in the pathway: FUS; FUS RNA binding protein

Given this context, what is the biological effect of this FUS allele,
specifically what disease does this contribute to?<|im_end|>
<|im_start|>assistant
<think>
Edit at 128: ref A -> var T (substitution)
</think>

Answer: amyotrophic lateral sclerosis<|im_end|>
```

We use the same training set up as Fesser et al. (2026). The text model is the Qwen3-4B checkpoint after continued pretraining used by Fesser et al. (2026), adapted with LoRA (rank 32, $\alpha = 64$) on every linear layer except the output head, which trains 66.1M parameters. The loss is the next-token cross-entropy over the assistant turn plus five times the cross-entropy over the tokens after a marker string, `Edit at` in the auxiliary runs and `Answer:` in the control, so the weighted term includes the edited bases and the disease in the auxiliary runs and the disease alone in the control. All runs use AdamW at learning rate 5×10^{-5}, batch size 1 with 8 gradient accumulation steps, and 12 epochs, on one H100 per run. We train three independent seeds and evaluate the checkpoints from each training run with the lowest validation loss.

E.3 Evaluation and results

Each model generates for every held-out query under three conditions with the prompt fixed: (1) the intact DNA sequences; (2) both DNA sequences shuffled which changes $\mathbf{Z}_{\mathrm{Evo2}}$; and (3) an evidence conflict that replaces both windows with those of another held-out query drawn at random. Accuracy on the edited base pair is determined by the first reference-to-variant base pair stated in the generated reasoning, which in the auxiliary runs is the pair in the `Edit at` line. The text-only reference uses no DNA and uses the most frequent training disease for the gene named in the prompt which reaches 0.917, and always writing the most frequent training variant pair reaches 0.103. We report the results in Table S9.

Table S9: Accuracy on the 145 held-out queries, as the mean over three training seeds and the standard error of that mean. Δ is intact minus shuffled. The conflict column is the share of queries whose written base pair equals the donor's.

	Disease		**Edited base pair**		**Conflict**
Training target, window	**Intact**	**Δ**	**Intact**	**Δ**	**Donor's pair**
No auxiliary, 2,048 bp	0.901 (0.026)	0.002 (0.022)	0.080 (0.002)	0.005 (0.002)	0.071 (0.005)
Auxiliary, 257 bp	0.970 (0.009)	0.021 (0.007)	0.671 (0.041)	0.549 (0.032)	0.467 (0.072)
Auxiliary, 2,048 bp	0.984 (0.005)	0.028 (0.007)	0.531 (0.115)	0.432 (0.123)	0.361 (0.139)
Text-only reference	0.917		0.103		–

We split accuracy on the edited base pair by whether the query's genome occurs in training, over the 145 held-out and 145 validation queries (Table S10). Only 72 of the 290 queries, 10 held-out and 62 validation, have a genome absent from training.

Table S10: Accuracy on the edited base pair for held-out and validation queries whose genome is absent from (72 queries) or present in (218 queries) the training set, as the mean over three training seeds.

Training target, window	**Unseen genomes**	**Seen genomes**
No auxiliary, 2,048 bp	0.042	0.083
Auxiliary, 257 bp	0.148	0.691
Auxiliary, 2,048 bp	0.106	0.606
Most frequent training pair (T>C)	0.111	0.101

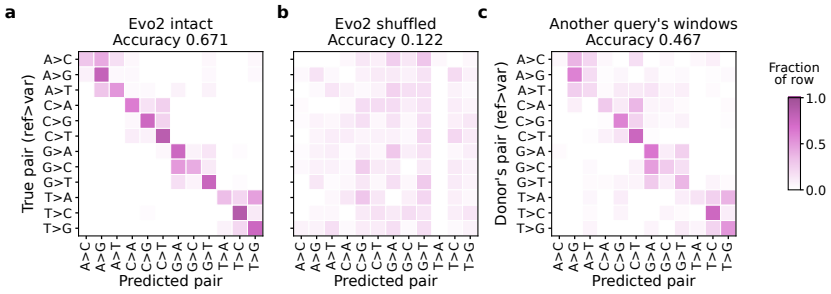

Figure S4: **Predicted reference-to-variant base pairs for the 257 bp auxiliary condition. a** Original DNA, **b** shuffled DNA, and **c** an evidence conflict that replaces both windows with those of another held-out query. Rows are the query's true pair in **a** and **b** and the donor's pair in **c**. Each row is normalized by its own total, so a diagonal cell is the accuracy for that pair. The three training seeds are pooled, giving $145 \times 3 = 435$ generations per matrix.