GEM-REASONER: PROTEIN FUNCTION REASONING FOR GENOME-SCALE METABOLIC MODEL RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Genome-scale metabolic models (GEMs) provide a computational foundation for systems biology and rational chassis design. Existing methods for automatically reconstructing GEMs from genome sequences typically transfer functional annotations from reference proteins based on sequence or protein-embedding similarity. However, these similarity signals do not reliably reflect functional equivalence, particularly under remote homology, functional divergence, or differences in substrate specificity, potentially resulting in erroneous or missing reaction annotations. To address this limitation, we propose GEM-Reasoner, replacing similarity-driven functional transfer with explicit reasoning about functional equivalence. To support this reasoning when retrieved annotations are incomplete or conflicting, we develop CAPA-RAG, a conflict-aware retrieval-augmented method that uses conflicts across functional dimensions to guide iterative retrieval and evidence integration. GEM-Reasoner uses CAPA-RAG to construct a structured functional profile for each protein and then applies a large language model to assess functional equivalence from paired profiles. We further construct the Protein Function Similarity (PFS) Benchmark, comprising PFS-Core and PFS-Challenge; PFS-Challenge pairs low-similarity positives with high-similarity negatives. On PFS-Challenge, GEM-Reasoner improves MCC by 0.168 over CLEAN, the strongest baseline. In GEM reconstruction for three bacterial species, GEM-Reasoner achieves the highest macro-average F1 for both substrate utilization and gene essentiality prediction. These results suggest that explicit functional reasoning can improve automated GEM reconstruction. Our code is available at https://anonymous.4open.science/r/GEM-Reasoner-825F.

1 INTRODUCTION

Metabolic networks sustain cellular growth and homeostasis through matter and energy conversion (Nielsen & Keasling, 2016). Accurate simulation of metabolic states and reaction fluxes clarify cellular function and guide rational cell design (Orth et al., 2010), while reducing experimental screening time and costs (O'Brien et al., 2015). Toward this goal, genome-scale metabolic models (GEMs) have emerged as an important computational paradigm in systems and synthetic biology. GEMs organize genome annotations and intracellular metabolic reactions into computable metabolic networks, enabling system-level simulation of cellular metabolism and providing a computational foundation for cellular phenotype prediction and rational engineering (Thiele & Palsson, 2010).

In recent years, the accumulation of high-throughput sequencing data and advances in deep learning have driven substantial progress in automated GEM reconstruction (Han et al., 2026; Heinken et al., 2023). Existing methods typically infer protein functions by matching unannotated proteins to annotated references using sequence or embedding similarity (Machado et al., 2018; Yu et al., 2023). These similarity scores are converted into reaction-level evidence through gene–protein–reaction associations, guiding reaction selection during metabolic network reconstruction (Machado et al., 2018; Han et al., 2026). Despite this progress, the prevailing paradigm still faces two key limitations. First, whether derived from sequences or protein embeddings, similarity signals remain indirect proxies for functional equivalence (Rost, 2002). In cases involving remote homology, functional divergence, or differences in substrate specificity, these signals may not reliably reflect the

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.

underlying functional relationships (Cui et al., 2025). This can result in erroneous or missing reaction annotations and limit cross-species generalization (Schnoes et al., 2009). Second, existing methods typically lack an interpretable functional reasoning process, making it difficult to identify the evidence underlying their predictions or trace how decisions are reached. This opacity reduces confidence in model outputs and limits their utility for guiding subsequent biological experiments.

We argue that GEM reconstruction should combine predictive accuracy with interpretability and consistency with fundamental biological principles (Lieven et al., 2020). This calls for integrating multidimensional information on protein functions and catalytic mechanisms (Gligorijević et al., 2021) and reasoning explicitly over this evidence to assess functional equivalence and guide reaction annotation. The reasoning process should link each prediction to its supporting evidence and explain how the decision is reached, making predictions traceable and useful for guiding biological experiments.

To address these challenges, we propose GEM-Reasoner, an interpretable framework for automated genome-scale metabolic model reconstruction. At its core is a protein functional similarity reasoning module that combines CAPA-RAG with LLM-based reasoning (Lewis et al., 2020; Wei et al., 2022). CAPA-RAG retrieves and integrates protein functional information across multiple complementary dimensions through multi-round retrieval and iterative refinement (Asai et al., 2024; Su et al., 2025). The LLM then explicitly reasons over the resulting structured functional evidence to infer functional relationships between proteins, enabling accurate and interpretable functional similarity prediction. Unlike existing approaches centered on similarity matching, GEM-Reasoner leverages richer protein functional knowledge and explicit reasoning to identify functionally related proteins with low sequence similarity. For each prediction, GEM-Reasoner also generates a traceable reasoning pathway and an explicit mechanistic rationale.

To systematically evaluate this capability, we construct the Protein Function Similarity Benchmark (PFS Benchmark) from Swiss-Prot (Bateman et al., 2025), with functional labels derived from EC annotations (McDonald et al., 2009). PFS-Core evaluates general functional similarity prediction, while PFS-Challenge tests whether models can correctly assess functional equivalence when sequence similarity is misleading. We also evaluate GEM reconstruction across three species using substrate-utilization and gene-essentiality prediction to assess downstream utility and cross-species generalization. The main contributions of this work are as follows:

1. We introduce GEM-Reasoner, reframing automated GEM reconstruction from similarity-driven functional transfer to explicit reasoning over protein functional evidence, with functional equivalence serving as the core criterion for reaction annotation.
2. We propose CAPA-RAG, a conflict-aware retrieval-augmented method that constructs comprehensive and consistent multidimensional protein function descriptions through iterative retrieval, conflict detection, and evidence integration.
3. We develop the PFS Benchmark, and establish a two-level evaluation spanning protein functional similarity prediction and GEM reconstruction. GEM-Reasoner outperforms CLEAN by 0.1681 MCC on PFS-Challenge and achieves the highest macro-average F1 on both downstream GEM tasks.

2 Related Work

2.1 GEM Reconstruction

GEMs encode gene-associated biochemical reactions for phenotype prediction (Thiele & Palsson, 2010). Automated reconstruction reduces manual curation by transferring protein functions to reaction networks (Henry et al., 2010). CarveMe prunes a universal network using sequence homology (Machado et al., 2018), whereas AlphaGEM adds embedding-based searches to recover relationships missed by alignment (Han et al., 2026). Similarity-based transfer remains vulnerable to remote homology, functional divergence, and substrate-specificity differences (Cui et al., 2025; Schnoes et al., 2009). These frameworks typically expose matching scores or final annotations, with limited access to the biological evidence behind each assignment. GEM-Reasoner instead evaluates functional equivalence through explicit reasoning over retrieved evidence before assigning reaction

scores. Similarity search defines its candidate set, while functional reasoning determines the subsequent transfer scores.

2.2 Protein Functional Similarity Prediction

Sequence methods such as DIAMOND and MMseqs2 identify homologs by amino-acid alignment (Buchfink et al., 2021; Steinegger & Söding, 2017). Foldseek identifies structurally similar proteins (Van Kempen et al., 2024). Embedding-based approaches, including ESM and ProTrek, compare learned protein representations (Lin et al., 2023; Su et al., 2025; Yu et al., 2023). Sequence methods may miss remote homologs with conserved function, while similar structures can have different catalytic specificities (Cui et al., 2025). Embedding scores may also obscure the functional basis of a match. CAPA-RAG supplies multidimensional functional evidence for subsequent reasoning about shared activity beyond similarity-based matching.

2.3 LLMs for Biological Reasoning and Prediction

LLMs increasingly integrate heterogeneous biological information for interpretable reasoning (Wang et al., 2025; Xie et al., 2025; Fallahpour et al., 2026a;b). VCWorld combines structured knowledge and iterative reasoning to predict drug-induced differential gene expression in specific cell-line contexts, with traceable reasoning paths and testable mechanistic hypotheses (Wei et al., 2026). BioReason-Pro combines protein representations and biological context for function prediction and structured reasoning (Fallahpour et al., 2026b). Their principal targets are cell-level responses and individual-protein annotation, respectively. Neither directly evaluates protein-pair equivalence and translates that decision into a reconstructed metabolic network. GEM-Reasoner addresses this decision by asking whether two proteins share a sufficiently specific activity to support reaction annotation. It connects conflict-aware evidence retrieval and pairwise reasoning to GEM reconstruction, evaluated through both PFS and downstream biological prediction.

3 Method

In this section, we first formalize the protein functional similarity prediction and GEM reconstruction tasks and introduce the overall computational framework (Section 3.1). We then present CAPA-RAG for multidimensional protein function profiling (Section 3.2), followed by LLM-based reasoning for functional equivalence assessment (Section 3.3). Finally, we describe the mapping of functional similarity scores to reaction scores and their integration into GEM assembly (Section 3.4). Figure 1 provides an overview of the GEM-Reasoner framework.

3.1 Preliminary and Task Formulation

GEM reconstruction. A genome-scale metabolic model $\mathcal{G} = (\mathcal{M}, \mathcal{A}, \mathbf{S})$ comprises metabolites $\mathcal{M}$, reactions $\mathcal{A}$, and a stoichiometric matrix $\mathbf{S}$. Top-down reconstruction starts from a universal network $\mathcal{U}$ of known microbial reactions and removes those the target genome lacks evidence for, giving the species-specific model $\mathcal{G}^\star$.

Evidence comes from a reference-protein collection $\mathcal{R}$ whose gene–protein–reaction (GPR) annotations record which reactions each member catalyzes: if a target protein q_i is functionally equivalent to a reference protein r_j, the reactions catalyzed by r_j become supported. Reconstruction therefore proceeds in two stages. A transfer score c_{ij} is first assigned to each query–reference pair (q_i, r_j); the scores are then aggregated along the Boolean structure of the GPR annotations into per-reaction evidence scores, which drive carving under stoichiometric and growth feasibility.

Functional transfer scoring. Existing frameworks set c_{ij} to a sequence or embedding similarity; GEM-Reasoner replaces it with an explicit functional-equivalence judgment:

$$c_{ij} = F_{\text{reason}}\big(\text{CAPA}(x_{q_i}),\ \text{CAPA}(x_{r_j})\big) \in [0, 1], \tag{1}$$

where CAPA builds a multi-dimensional functional profile from the sequence x_p and F_{reason} decides whether the two proteins share at least one catalytic activity at complete four-field EC specificity. Aggregation and carving follow CarveMe. Candidate references are still obtained by sequence alignment. Similarity serves only as a recall signal there; equivalence is decided by F_{reason}.

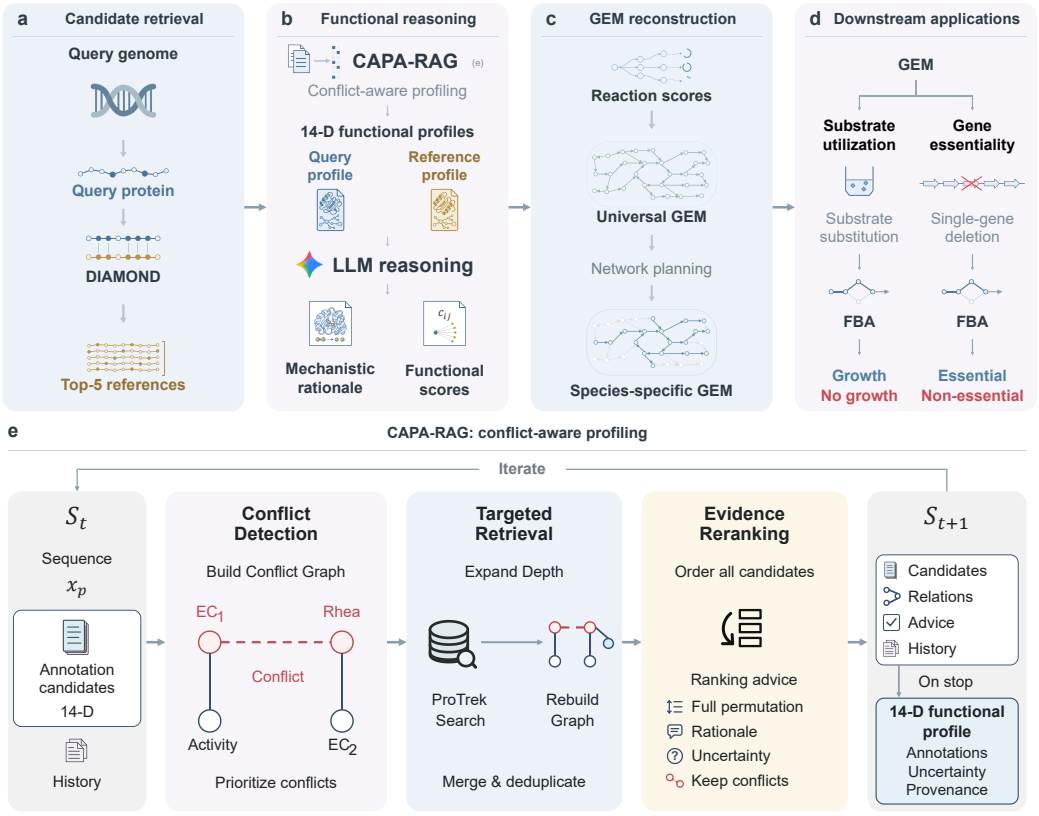


Figure 1: **GEM-Reasoner workflow and conflict-aware functional profiling.** **(a)** DIAMOND retains the top five reference proteins per query. **(b)** LLM reasoning uses CAPA-RAG's 14-dimensional profiles to produce functional similarity scores and rationales. **(c)** Functional scores aggregated through gene–protein–reaction (GPR) rules guide CarveMe network carving. **(d)** Substrate substitution and single-gene deletion with FBA evaluate substrate utilization and gene essentiality. **(e)** CAPA-RAG iterates conflict assessment, targeted retrieval, candidate integration, and advisory ranking, retaining candidates, conflict history, uncertainty, and provenance in the final profile. Graph relations are schematic.

Downstream prediction. We evaluate $\mathcal{G}^\star$ on two binary tasks. Let $v^{\max}(\mathcal{G}, e)$ denote the maximum growth rate predicted by flux balance analysis (FBA) in medium e. *Substrate utilization* asks whether the model grows in the medium e_s where substrate s replaces the corresponding nutrient source; *gene essentiality* asks whether growth of the deletion model $\mathcal{G}^\star_{-g}$ drops substantially after removing gene g:

$$\hat{y}_s^{\mathrm{SU}} = \mathbb{I}[v^{\max}(\mathcal{G}^\star, e_s) > \tau_{\mathrm{SU}}], \qquad \hat{y}_g^{\mathrm{GE}} = \mathbb{I}\big[v^{\max}(\mathcal{G}^\star_{-g}, e) < \rho\, v^{\max}(\mathcal{G}^\star, e)\big], \tag{2}$$

where $\tau_{\mathrm{SU}} = 0.1\,\mathrm{h}^{-1}$ is the absolute substrate-utilization threshold and ρ a fraction of wild-type growth. Growth exactly at the substrate-utilization threshold is classified as non-growth.

3.2 Conflict-Aware Protein Function Profiling with CAPA-RAG

Single-pass retrieval can return plausible but inconsistent annotations. CAPA-RAG uses these conflicts to direct additional retrieval, then orders the expanded evidence for functional reasoning (Figure 1e). Each round retains all candidates, their relations, historical conflicts, and ranking advice; interpretations are stored separately from immutable annotation texts.

Multi-dimensional retrieval. For each functional dimension $d \in \mathcal{D} = \{d_1, \ldots, d_{14}\}$, ProTrek retrieves standardized annotations from the sequence:

$$\mathcal{C}_{p,d}^{(0)} = \mathrm{Top}_{k_0}\left[\mathrm{Ret}_{\mathrm{s2t}}(x_p, d; \mathcal{I}_{\mathrm{PT}})\right], \tag{3}$$

The dimensions cover functions and reactions, cofactors and sites, pathways, and localization (Appendix B.1). Each dimension uses an independently constructed standardized functional dictionary shared across input proteins. ProTrek matches sequence and text representations to retrieve candidate functions, without accessing test-protein-to-label mappings or tailoring the vocabulary to test answers. The query contains only (x_p, d); $\mathcal{I}_{\mathrm{PT}}$ is the fixed sequence-to-text index and k_0 is the initial depth. Retrieval scores are within-dimension ranking signals, not cross-dimensional probabilities.

Conflict-guided expansion. The Conflict Agent evaluates predefined dimension pairs, with standardized identifier relations taking precedence over free-text judgments. Missing information and compatible multifunctionality are not contradictions. A conflict marks incompatible same-context claims and identifies where more evidence is needed; it does not establish which candidate is correct. The controller identifies expandable dimensions from the highest-priority open conflicts; the Retrieval Agent returns that eligible set. The controller extends the same frozen ProTrek rankings within depth and budget limits, merges and deduplicates candidates, and rebuilds relations and the historical ledger before Ranking. Original contradictions remain traceable when later observations change the current relation, preserving competing interpretations for downstream reasoning. Appendix B.2 formalizes the graph, and Appendix B.3 specifies depth updates and tied-priority allocation.

Ranking Agent. Ranking provides advisory ordering for all 14 dimensions. For each dimension, it returns a complete permutation of the supplied candidate IDs, a rationale, and uncertainty notes. It cannot select, reject, hide, edit, or invent annotations, including contradictory hypotheses. Ordering considers specific functional semantics, mechanisms, and traceable identifier relations; ProTrek scores are only within-dimension retrieval signals. If evidence cannot distinguish alternatives, the original retrieval order is preserved. Compatibility, repeated descriptions, and cross-references to one reaction are not independent confirmations. All candidates and conflicts remain visible downstream, irrespective of the suggested order.

After recording each completed round, expansion stops when no historical open conflict remains or a resource limit prevents further retrieval; unresolved conflicts remain in the profile. Ranking advice does not close the conflict ledger. Complete rules and prompts are in Appendices B and G.

3.3 Functional Similarity Reasoning

Both applications use the same prompt, LLM, output schema, and complete evidence view: $P_{\mathrm{PFS}} = P_{\mathrm{GEM}}$ and $\mathrm{View}_{\mathrm{PFS}} = \mathrm{View}_{\mathrm{GEM}}$. The view retains all candidates, relations, historical conflicts, and ranking advice without exposing protein identities or benchmark labels. For $m \in \{\mathrm{PFS}, \mathrm{GEM}\}$,

$$(\hat{\ell}, \hat{s}, \hat{r}) = \mathrm{Parse}(\mathrm{LLM}_\theta\left[P_m(\mathrm{View}_m(\mathbf{z}_p), \mathrm{View}_m(\mathbf{z}_{p'}))\right]). \tag{4}$$

This defines $F_{\mathrm{reason}}(\mathbf{z}_p, \mathbf{z}_{p'}; m)$. Reasoning prioritizes reaction chemistry and specificity (catalytic activity, Rhea, EC), followed by functional interpretation, mechanistic evidence, and biological context. The priority concerns relevance rather than vote counts: a specific lower-priority distinction can expose an applicability error in higher-priority evidence. The agent compares original texts and material alternatives, distinguishes whole-complex activity from subunit roles, and preserves supported multifunctionality. It chooses an ordinal level, an in-band score, and a concise evidence sentence; credible alternatives that change the decision warrant uncertainty. Scores are not calibrated probabilities. Appendix C details evidence serialization, priorities, and output validation.

3.4 Function-Guided Reaction Scoring and GEM Assembly

For query q_i and reference r_j, the functional score c_{ij} directly replaces the sequence score, $\widetilde{b}_{ij} = c_{ij}$. All five references proceed without score fusion or a label-based rejection gate, and each score propagates to the reference's associated reactions and GPR roles. Each pair is evaluated only once; its

score is reused for these associations, and even positive low scores remain as soft genetic evidence. This candidate set bounds recall, and the pair-level decision is not a separate verification of every transferred reaction (Appendix D.2).

Following CarveMe 1.6.5, GPR aggregation takes the maximum over query hits for a role, the mean over complex roles with missing scores set to zero, and the maximum over isozymes or alternative complexes. Normalized reaction scores enter the unchanged CarveMe mixed-integer linear program (MILP), which selects a feasible network before removing inactive reactions and orphaned components. Appendix D specifies propagation, normalization, and assembly.

4 Experiments

4.1 Experiment Setup

PFS Benchmark. The PFS Benchmark evaluates functional equivalence between protein pairs. PFS-Core and PFS-Challenge each contain 20,000 Swiss-Prot-based pairs, balanced between positives sharing a complete EC number and negatives differing in at least two fields. PFS-Core samples eligible pairs randomly. PFS-Challenge targets disagreement between sequence similarity and function, using sequences first included in Swiss-Prot during releases `2024_03–2026_01`. This interval starts after the fixed ProTrek-650M checkpoint's first public release on May 22, 2024. Both proteins must meet this criterion in 13,000 pairs. Because newly included proteins are limited in number, the remaining 7,000 positive pairs require one qualifying protein, with no temporal restriction on its partner. PFS-Core and PFS-Challenge are disjoint at the protein and pair levels, and gold EC labels are never predictor inputs. Appendix A specifies the sampling rules, temporal controls, and information boundaries.

GEM reconstruction. We evaluate *B. subtilis* 168, *E. coli* K-12 MG1655, and *P. aeruginosa* PAO1 on substrate utilization and gene essentiality using the benchmark data adopted by CarveMe (Machado et al., 2018). All methods use identical genome inputs, medium constraints, and gap-filling conditions. (i) For substrate utilization, we compare FBA growth predictions with Biolog Phenotype MicroArray labels after substituting each test substrate for the corresponding nutrient source in an M9 background. (ii) For gene essentiality, we compare single-gene deletion predictions with benchmark reference labels on a common set of genes consistently mapped across all models. We apply the decision rules in Equation (2) with $\tau_{\mathrm{SU}} = 0.1\,\mathrm{h}^{-1}$ and $\rho = 0.1$, so essentiality is assessed relative to each model's own wild-type growth in the same medium. Species-specific assay conditions, label processing, and the treatment of infeasible simulations are detailed in Appendices E.4–E.5.

Baselines and metrics. PFS baselines cover DIAMOND, MMseqs2, Foldseek, ESM-2, ESM-C, SaProt, ProTrek, and CLEAN (Buchfink et al., 2021; Steinegger & Söding, 2017; Van Kempen et al., 2024; Lin et al., 2023; Candido et al., 2026; Su et al., 2024; 2025; Yu et al., 2023). CLEAN learns function-aware protein representations through EC-supervised contrastive learning on Swiss-Prot (Yu et al., 2023). We evaluate CLEAN using embedding cosine similarity, not predicted EC labels. GEM-Reasoner uses the fixed rule $s > 0.5$; each PFS baseline uses its optimal threshold selected on an independent validation set and held fixed for test evaluation. GEM baselines are ModelSEED, CarveMe, and AlphaGEM (Machado et al., 2018; Henry et al., 2010; Han et al., 2026). PFS reports MCC as primary, alongside Accuracy, F1, Precision, and Recall. GEM evaluation reports strain-level and arithmetic macro-average metrics, with F1 in Figure 2. Positive cases are growth-supporting nutrients or experimentally essential genes. Table E.5 links configurations to results; Appendix E gives full protocols and metrics.

4.2 Protein Functional Similarity Prediction

On PFS-Core, GEM-Reasoner achieves an MCC of **0.9199**, exceeding CLEAN by 0.0137 (Table 1). Recall improves from 0.9202 to 0.9574, with slightly lower Precision (0.9623 versus 0.9830), recovering more functional equivalents while maintaining high precision.

On PFS-Challenge, where sequence similarity and function disagree, MMseqs2 and DIAMOND yield negative MCCs, indicating predictions inversely associated with the functional labels under

Table 1: **Protein functional similarity on PFS.** PFS-Core and PFS-Challenge each contain 10,000 positive and 10,000 negative pairs. GEM-Reasoner uses $s > 0.5$; baseline thresholds are optimized on independent validation data and fixed for testing. Best and second-best values within each setting are boldfaced and underlined. Complete results, including ESM2-8M and ESM2-35M, are reported in Table E.1 in the appendix.

	PFS-Core					PFS-Challenge				
Method	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑
ESM2-150M	0.8731	0.8701	0.7470	0.8912	0.8500	0.3370	0.3625	-0.3272	0.3491	0.3771
ESM2-650M	0.8573	0.8479	0.7202	0.9079	0.7953	0.2871	0.2592	-0.4270	0.2698	0.2495
ESM2-3B	0.8564	0.8442	0.7216	0.9224	0.7782	0.2807	0.2504	-0.4401	0.2614	0.2403
ESMC-300M	0.8418	0.8382	0.6842	0.8576	0.8196	0.3658	0.4075	-0.2712	0.3823	0.4362
ESMC-600M	0.8542	0.8550	0.7084	0.8504	0.8596	0.3896	0.4464	-0.2257	0.4084	0.4923
SaProt-35M	0.8284	0.8144	0.6644	0.8866	0.7531	0.3194	0.3039	-0.3616	0.3110	0.2972
SaProt-650M	0.8372	0.8296	0.6771	0.8702	0.7926	0.3039	0.3042	-0.3923	0.3040	0.3043
SaProt-1.3B	0.8644	0.8616	0.7294	0.8797	0.8442	0.4047	0.4499	-0.1932	0.4181	0.4868
ProTrek-35M-Seq	0.9130	0.9142	0.8263	0.9015	0.9273	0.4732	0.5539	-0.0575	0.4803	0.6540
ProTrek-35M-Str	0.9182	0.9194	0.8367	0.9052	<u>0.9341</u>	0.5297	0.6234	0.0684	0.5198	<u>0.7786</u>
ProTrek-650M-Seq	0.9096	0.9098	0.8192	0.9082	0.9113	0.4934	0.5689	-0.0141	0.4951	0.6685
ProTrek-650M-Str	0.9166	0.9154	0.8334	0.9281	0.9031	0.5143	0.5654	0.0293	0.5115	0.6320
MMseqs2	0.9014	0.8945	0.8096	0.9616	0.8361	0.1407	0.0021	-0.7480	0.0025	0.0018
DIAMOND	0.8865	0.8763	0.7836	<u>0.9624</u>	0.8044	0.1463	0.0051	-0.7377	0.0061	0.0044
Foldseek	0.7816	0.7380	0.5973	0.9222	0.6151	0.4530	0.4537	-0.0941	0.4531	0.4544
CLEAN	<u>0.9522</u>	<u>0.9506</u>	<u>0.9062</u>	**0.9830**	0.9202	<u>0.7686</u>	<u>0.7594</u>	<u>0.5387</u>	<u>0.7907</u>	0.7304
GEM-Reasoner	**0.9600**	**0.9598**	**0.9199**	0.9623	**0.9574**	**0.8534**	**0.8529**	**0.7068**	**0.8560**	**0.8498**

this sampling design. GEM-Reasoner achieves an MCC of **0.7068**, exceeding CLEAN by 0.1681, with the highest Precision and Recall. This larger advantage supports functional evidence and explicit reasoning when sequence similarity is misleading.

4.3 GEM Reconstruction Performance

Figure 2 evaluates downstream prediction under matched protocols; Appendix E.6 reports all five metrics.

Substrate utilization. GEM-Reasoner achieves the highest F1 for all three organisms and a macro average of **0.7926**, versus 0.7097 for AlphaGEM, 0.6495 for CarveMe, and 0.6043 for ModelSEED. It also has the highest macro-average Sensitivity (0.8911), while CarveMe retains higher macro-average Precision and Specificity (Table E.2). The gains therefore reflect improved F1 balance rather than uniform improvement across metrics.

Gene essentiality. GEM-Reasoner leads F1 on *E. coli* (0.6197) and *P. aeruginosa* (0.5025), while CarveMe leads on *B. subtilis* (0.5827 versus 0.4952). Its macro-average F1 is highest (**0.5392**), followed by CarveMe (0.4156), ModelSEED (0.3966), and AlphaGEM (0.2321). GEM-Reasoner has the highest Precision on all three organisms (macro average 0.7785); AlphaGEM has the highest macro-average Specificity (0.9861; Table E.3).

4.4 Case Studies

Protein functional similarity reasoning. Two correctly classified pairs from a 400-pair trace analysis illustrate CAPA-RAG's evidence processing (Appendix F). For MurD P14900/Q7VAM1, CAPA-RAG detects a D/L-glutamate conflict in P14900, retrieves the missing D-glutamate catalytic candidate, and ranks it above the initial L-glutamate alternative, consistent with retrieved EC and Rhea evidence. Reasoning returns `exact_match` (0.95). For P29707/Q95DR6, it investigates conflicting ATPase annotations and retrieves an additional Na^+-coupled reaction for P29707. Retaining specific catalytic evidence alongside competing interpretations enables reasoning to distin-

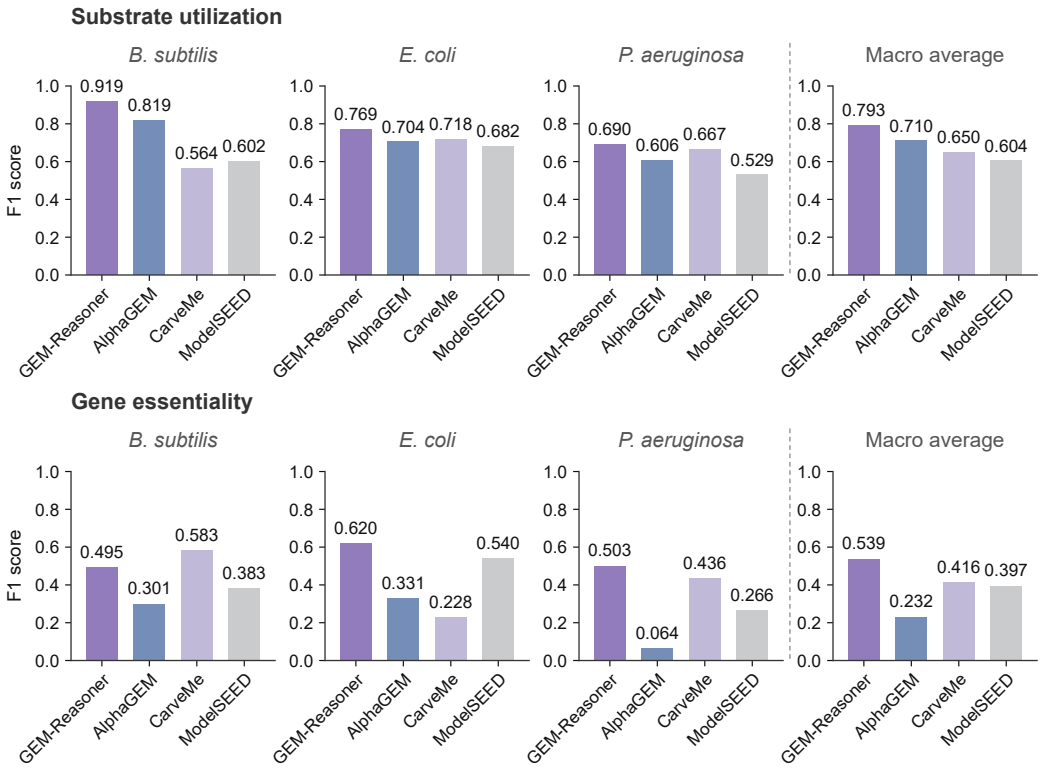


Figure 2: F1 for substrate utilization (top) and gene essentiality (bottom) under matched protocols (Section 4.1). Columns show *B. subtilis* 168, *E. coli* K-12 MG1655, *P. aeruginosa* PAO1, and their arithmetic macro average. Substrate-utilization positives require a biomass objective > 0.1. Labels show three decimals; Tables E.2 and E.3 report all five metrics to four decimals.

guish Na^+ from H^+ coupling and return `unlikely_match` (0.25). These cases illustrate evidence recovery, prioritization, and preservation of functional specificity.

Pathway reconstruction and phenotype prediction. For the experimentally positive nitrogen source L-proline in *B. subtilis* 168, GEM-Reasoner predicts growth (biomass objective 1.133, above 0.1), whereas CarveMe predicts none (0), with glucose retained and ammonium uptake disabled. Both models contain proline uptake and conversion of 1-pyrroline-5-carboxylate (P5C) to glutamate; only GEM-Reasoner contains the *putB*-associated proline-oxidation entry `PRO1x`. Parsimonious FBA shows flux through this proline-to-glutamate route, and flux variability analysis requires positive flux through both catabolic steps in every GEM-Reasoner solution satisfying the growth constraint. This supports improved pathway coverage as an explanation for the corrected phenotype. Appendix F.4 details reaction content, flux requirements, controls, sensitivity analyses, and biochemical scope.

4.5 Ablation and Sensitivity Analyses

Component ablations. With sequence-only input (Table 2), MCC falls from 0.7068 to -0.2224 on PFS-Challenge and from 0.9199 to 0.7933 on PFS-Core. The larger decrease on PFS-Challenge highlights the value of functional evidence when sequence similarity is misleading. Retaining 14-dimensional ProTrek retrieval while removing CAPA-RAG still reduces PFS-Challenge MCC by 0.2610 (PFS-Core: 0.0143) and Precision from 0.8560 to 0.6557. These losses support the value of processing beyond retrieval and are consistent with cross-dimensional assessment helping distinguish specific activities despite conflicting annotations. The effect of chain-of-thought (CoT) prompting is also task-dependent: removing it lowers PFS-Challenge MCC by 0.1643 but raises

PFS-Core MCC by 0.0045, suggesting that CoT is more useful for weighing functional evidence in difficult comparisons.

Table 2: **Component ablations on PFS with fixed $s > 0.5$.** Best and second-best values within each setting are boldfaced and underlined. Table 3 reports reasoning-backbone comparisons.

	PFS-Core					PFS-Challenge				
Model	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑
w/o CAPA-RAG	0.9528	0.9527	0.9056	0.9555	0.9498	0.7093	0.7520	0.4458	0.6557	<u>0.8813</u>
w/o functional evidence	0.8967	0.8971	0.7933	0.8932	0.9010	0.3914	0.4502	-0.2224	0.4105	0.4983
w/o CoT	**0.9622**	**0.9618**	**0.9244**	**0.9705**	<u>0.9533</u>	<u>0.7620</u>	<u>0.7893</u>	<u>0.5425</u>	<u>0.7079</u>	**0.8919**
GEM-Reasoner	<u>0.9600</u>	<u>0.9598</u>	<u>0.9199</u>	<u>0.9623</u>	**0.9574**	**0.8534**	**0.8529**	**0.7068**	**0.8560**	0.8498

Reasoning-backbone sensitivity. Gemini 3 Flash leads MCC in both settings (Table 3). The wider spread on PFS-Challenge (0.3866 versus 0.0431 on PFS-Core) indicates greater backbone sensitivity under fixed evidence, while Qwen 3's nonmonotonic trend shows that size alone does not ensure gains.

Table 3: **Reasoning-backbone sensitivity on PFS with fixed $s > 0.5$.** Only the final Reasoning LLM changes; other settings remain fixed. Best and second-best values are boldfaced and underlined within each setting.

		PFS-Core					PFS-Challenge				
Model	Access	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑
Qwen 3 4B	Open-source	<u>0.9557</u>	<u>0.9557</u>	<u>0.9114</u>	0.9556	<u>0.9558</u>	0.6534	0.6968	0.3202	0.6192	<u>0.7967</u>
Qwen 3 8B	Open-source	0.9474	0.9456	0.8968	0.9794	0.9140	0.7075	0.7084	0.4149	0.7062	0.7106
Qwen 3 14B	Open-source	0.9379	0.9343	0.8809	<u>0.9911</u>	0.8836	0.7782	0.7478	0.5733	0.8664	0.6578
Qwen 3 32B	Open-source	0.9383	0.9390	0.8768	0.9282	0.9501	0.7818	0.7648	0.5696	0.8293	0.7097
Claude 4.5 Haiku	Closed-source	0.9511	0.9487	0.9058	**0.9959**	0.9058	0.8281	0.8023	0.6796	**0.9437**	0.6977
ChatGPT 5 mini	Closed-source	0.9518	0.9515	0.9035	0.9557	0.9474	<u>0.8420</u>	<u>0.8306</u>	<u>0.6902</u>	<u>0.8951</u>	0.7747
Gemini 3 Flash	Closed-source	**0.9600**	**0.9598**	**0.9199**	0.9623	**0.9574**	**0.8534**	**0.8529**	**0.7068**	0.8560	**0.8498**

5 Conclusion

GEM-Reasoner replaces similarity-driven functional transfer with explicit reasoning over protein functional evidence. Its gains on PFS and downstream GEM prediction support this approach to reaction annotation and reconstruction. CAPA-RAG retrieves, organizes, and preserves evidence for pairwise decisions, which include traceable rationales. Future work will improve the accuracy and coverage of ProTrek-derived evidence through complementary sources, and explore distillation and supervised fine-tuning of smaller reasoning models to reduce inference cost while preserving predictive performance and auditability.

AI USE STATEMENT

LLMs are used within GEM-Reasoner for conflict assessment, retrieval coordination, evidence ranking, and functional similarity reasoning (Section 3 and Appendix G). During manuscript preparation, the authors used generative AI tools to polish author-provided text, including correcting grammar, refining word choice, and improving sentence construction to enhance clarity, readability, and linguistic consistency. AI assistance in manuscript preparation was limited to these language-level revisions. The authors take full responsibility for the accuracy and integrity of the final manuscript, including its claims, analyses, citations, and conclusions.

ETHICS STATEMENT

This study uses public protein annotations, predicted protein features, and organism-level metabolic benchmarks. It does not involve human participants or personal data. GEM-Reasoner's predicted functional relationships and reaction assignments remain computational hypotheses requiring biological validation. Errors in retrieved annotations, incomplete functional evidence, and overconfident LLM interpretations can propagate to reconstructed networks and phenotype predictions. The authors are responsible for respecting the licenses and terms governing the underlying data and models and for considering the consequences of downstream use.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method interfaces and core equations. Appendix A documents benchmark construction, label definitions, temporal sampling groups, and information boundaries. Appendix B describes the standardized functional dictionaries, functional-evidence schema, conflict-guided retrieval, and ranking; Appendix C details reasoning-output validation. Appendix D specifies score propagation, GPR aggregation, and network assembly. Appendix E describes the baseline and GEM evaluation protocols and provides the complete downstream results and reasoning-backbone comparisons. Table E.5 summarizes model and retrieval resources, inference settings, and their correspondence to the reported results. Appendix F records the case-analysis settings, selected original agent outputs, and controller operations. Appendix G supplies the agent prompts, input constructors, and scoring rubric.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, et al. Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1), 2025.

Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368, 2021.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pp. 2026–06, 2026.

Haiyang Cui, Yufeng Su, Tanner J Dean, Tianhao Yu, Zhengyi Zhang, Jian Peng, Diwakar Shukla, and Huimin Zhao. Enzyme specificity prediction using cross-attention graph neural networks. *Nature*, 647(8090):639–647, 2025.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *Advances in Neural Information Processing Systems*, 38:111644–111671, 2026a.

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

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Weishang Han, Luchi Xiao, Haocheng Sun, Guangming Xiang, Qianxi Jia, Haoyu Wang, Boyang Ji, Cheng Zhang, Eduard J Kerkhoven, Jens Nielsen, et al. Alphagem enables precise genome-scale metabolic modelling by integrating protein structure alignment with deep-learning-based dark metabolism mining. *Nature Communications*, 2026.

Almut Heinken, Johannes Hertel, Geeta Acharya, Dmitry A Ravcheev, Malgorzata Nyga, Onyedika Emmanuel Okpala, Marcus Hogan, Stefanía Magnúsdóttir, Filippo Martinelli, Bram Nap, et al. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nature Biotechnology*, 41(9):1320–1331, 2023.

Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982, 2010.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Christian Lieven, Moritz E Beber, Brett G Olivier, Frank T Bergmann, Meric Ataman, Parizad Babaei, Jennifer A Bartell, Lars M Blank, Siddharth Chauhan, Kevin Correia, et al. Memote for standardized genome-scale metabolic model testing. *Nature biotechnology*, 38(3):272–276, 2020.

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

Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research*, 46(15):7542–7553, 2018.

Andrew G McDonald, Sinéad Boyce, and Keith F Tipton. Explorenz: the primary source of the iubmb enzyme list. *Nucleic acids research*, 37(suppl_1):D593–D597, 2009.

Susanne Moses, Tatjana Sinner, Adrienne Zaprasis, Nadine Stöveken, Tamara Hoffmann, Boris R Belitsky, Abraham L Sonenshein, and Erhard Bremer. Proline utilization by bacillus subtilis: uptake and catabolism. *Journal of Bacteriology*, 194(4):745–758, 2012.

Jens Nielsen and Jay D Keasling. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

Edward J O'Brien, Jonathan M Monk, and Bernhard O Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987, 2015.

Burkhard Rost. Enzyme function less conserved than anticipated. *Journal of molecular biology*, 318(2):595–608, 2002.

Alexandra M Schnoes, Shoshana D Brown, Igor Dodevski, and Patricia C Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605, 2009.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pp. 6987–7009, 2024.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, et al. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, pp. 1–7, 2025.

Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: a biological world model for virtual cell simulation. In *International Conference on Learning Representations*, volume 2026, pp. 72795–72823, 2026.

Elliot Xie, Lingxin Cheng, Jack Shireman, Yujia Cai, Jihua Liu, Chitrasen Mohanty, Mahua Dey, and Christina Kendziorski. Cassia: a multi-agent large language model for automated and interpretable cell annotation. *Nature Communications*, 17(1):389, 2025.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

APPENDIX

A PFS Benchmark Construction **14**
- A.1 EC Annotation Filtering and Label Definition 14
- A.2 Multi-EC Proteins and Label Consistency 14
- A.3 PFS-Core and PFS-Challenge Sampling 14
- A.4 Disjointness and Information Boundaries 15

B CAPA-RAG Profiling Details **15**
- B.1 Functional Dimensions and Candidate Records 15
- B.2 Relation Types and Conflict Prioritization 16
- B.3 Retrieval Expansion and Reranking 17
- B.4 Termination and Profile Schema 18

C Reasoning Output and Validation **18**
- C.1 Similarity Levels and Score Intervals 19
- C.2 Parsing and Output Validation 19

D Protein-to-Reaction Mapping and Network Assembly **19**
- D.1 Network and GPR Notation 19
- D.2 Candidate Generation and Score Propagation 19
- D.3 GPR-Aware Aggregation 20
- D.4 Reaction-Score Normalization 20
- D.5 CarveMe Optimization and Assembly 21

E Experimental Protocols and Detailed Results **21**
- E.1 PFS Baseline Configurations 21
- E.2 Complete PFS Benchmark Results 23
- E.3 Organisms and Reconstruction Inputs 23
- E.4 Substrate-Utilization Evaluation 24
- E.5 Gene-Essentiality Evaluation 24
- E.6 Detailed GEM Reconstruction Results 24
- E.7 Reasoning-Backbone Sensitivity 26
- E.8 Experimental Configuration and Result Correspondence 26

F Case Studies **28**
- F.1 Correct Positive: Recovery of MurD Stereospecific Evidence 29
- F.2 Correct Negative: Distinguishing Sodium and Proton Coupling 31
- F.3 Controller Outcomes and Scope of the Examples 32
- F.4 Proline Catabolism and Nitrogen-Source Utilization 33

G Prompts **36**
- G.1 Conflict Agent 36
- G.2 Retrieval Agent 37
- G.3 Ranking Agent 38
- G.4 Functional Similarity Reasoning 39
- G.5 Scoring Rubric 42

A PFS BENCHMARK CONSTRUCTION

A.1 EC ANNOTATION FILTERING AND LABEL DEFINITION

To systematically evaluate the ability of different methods to determine protein functional equivalence, we construct the Protein Function Similarity Benchmark (PFS Benchmark). The benchmark uses manually reviewed Enzyme Commission (EC) annotations from Swiss-Prot as reference labels and comprises two settings, PFS-Core and PFS-Challenge. Each setting contains 20,000 protein pairs, equally divided into 10,000 positive and 10,000 negative examples.

Let E_p^{gold} denote the curated EC annotations used only to construct benchmark labels for protein p. Each candidate PFS record selects one $e \in E_p^{\mathrm{gold}}$ and one $e' \in E_{p'}^{\mathrm{gold}}$. We represent an EC number by its four fields, $e = (e_1, e_2, e_3, e_4)$. Pair construction uses only complete four-field EC numbers after canonicalization to a common EC vocabulary; records with missing fields are excluded before the following distance is computed:

$$\Delta_{\mathrm{EC}}(e, e') = \sum_{h=1}^{4} \mathbb{I}[e_h \neq e'_h]. \tag{A.1}$$

The binary reference label induced by the selected curated EC annotations is

$$y_{pp'}^{\mathrm{PFS}}(e, e') = \begin{cases} 1, & \Delta_{\mathrm{EC}}(e, e') = 0, \\ 0, & \Delta_{\mathrm{EC}}(e, e') \geq 2. \end{cases} \tag{A.2}$$

A.2 MULTI-EC PROTEINS AND LABEL CONSISTENCY

Candidate records that differ in exactly one EC field are excluded. We additionally prevent the same protein pair from receiving conflicting labels. For each pair (p, p'), define

$$\mathcal{E}_{pp'}^{\mathrm{adm}} = \{(e, e') \in E_p^{\mathrm{gold}} \times E_{p'}^{\mathrm{gold}} : \Delta_{\mathrm{EC}}(e, e') \in \{0, 2, 3, 4\}\},$$

$$\mathcal{Y}_{pp'} = \{y_{pp'}^{\mathrm{PFS}}(e, e') : (e, e') \in \mathcal{E}_{pp'}^{\mathrm{adm}}\}.$$

We retain (p, p') only if $|\mathcal{Y}_{pp'}| = 1$; pairs for which admissible EC combinations induce both positive and negative labels are removed. A multi-EC protein may contribute multiple admissible records, but all retained records for the same protein pair have the same label. The curated EC annotations E_p^{gold}, $E_{p'}^{\mathrm{gold}}$, e, and e' are used only for pair construction and label assignment. They are never exposed to the predictor or included in its prompt.

A.3 PFS-CORE AND PFS-CHALLENGE SAMPLING

PFS-Core setting. The PFS-Core setting evaluates protein functional similarity prediction under standard conditions. We randomly sample 10,000 positive and 10,000 negative pairs from the candidates satisfying the labeling criteria above, yielding a class-balanced evaluation set.

PFS-Challenge setting. The PFS-Challenge setting evaluates model performance when sequence similarity does not reliably indicate functional relationships. A newly included sequence is one first incorporated into Swiss-Prot during releases `2024_03`–`2026_01`, inclusive. This definition concerns first sequence inclusion in Swiss-Prot, not subsequent annotation updates or first availability in another database.

Because newly included sequences are limited, PFS-Challenge uses two temporal sampling rules (Table A.1).

These rows denote disjoint construction groups, not an exact partition by observed inclusion dates. In the 7,000-pair group, the unrestricted partner may also be newly included. All PFS-Challenge

Table A.1: **Temporal requirements used to construct PFS-Challenge.** The two groups are defined by their sampling requirements; an unrestricted partner may also be newly included.

Temporal requirement at construction	Positive pairs	Negative pairs	Total
Both proteins must be newly included	3,000	10,000	13,000
One newly included protein; partner unrestricted	7,000	0	7,000
Total	10,000	10,000	20,000

pairs therefore contain at least one newly included sequence. Only the partner's temporal requirement is relaxed; all EC filtering, deduplication, and disjointness rules between PFS-Core and PFS-Challenge remain applicable.

Within these eligibility constraints, we use BLAST to estimate sequence similarity and rank candidate pairs without applying a fixed similarity threshold. We prioritize positive pairs with the lowest sequence similarity and negative pairs with the highest sequence similarity. This selection tests recognition of shared activity despite low similarity and functional distinctions despite high similarity, yielding 10,000 positive and 10,000 negative pairs.

A.4 Disjointness and Information Boundaries

PFS-Core and PFS-Challenge are mutually disjoint at both the protein and pair levels, with neither duplicate pairs nor order-reversed equivalents. This constraint also applies to the temporally unrestricted partners in the 7,000-positive-pair group.

We fix the ProTrek-650M checkpoint object first publicly released on May 22, 2024.[1] Swiss-Prot release `2024_03` was published on May 29, 2024,[2] placing the required first-inclusion interval after this parameter snapshot. The temporal requirement applies to both proteins in 13,000 pairs and to one designated protein in the remaining 7,000 positive pairs. This establishes the stated Swiss-Prot chronology without implying that every sequence was previously unseen. ProTrek training also used TrEMBL50 (Su et al., 2025); first Swiss-Prot inclusion does not establish prior absence from TrEMBL or other pretraining resources. The criterion therefore does not establish complete training-data isolation for ProTrek or the downstream LLMs.

The gold Swiss-Prot EC annotations associated with PFS test proteins are used only for offline benchmark construction, reference-label assignment, and evaluation. They are not supplied to any predictor, injected into retrieval or agent prompts, or used to construct candidate vocabularies. Retrieval uses protein sequences and independently constructed standardized functional dictionaries from the fixed ProTrek library (Appendix B.1). The checkpoint and retrieval-library versions identify distinct resources; freezing model parameters does not imply that all functional vocabulary knowledge predates the checkpoint.

B CAPA-RAG Profiling Details

B.1 Functional Dimensions and Candidate Records

CAPA-RAG organizes protein evidence into 14 dimensions:

Table B.1: **The 14 functional dimensions used by CAPA-RAG.**

Functional cluster	**Dimensions**
Molecular function and reaction	Function; Catalytic activity; EC number; GO Molecular Function; Rhea reaction
Cofactor, sites, and structure	Cofactor; Active site; Binding site; Domain
Pathway and biological process	UniPathway; Reactome Pathway; GO Biological Process
Cellular localization	Subcellular location; GO Cellular Component

[1] https://huggingface.co/westlake-repl/ProTrek_650M
[2] https://www.uniprot.org/uniprotkb/O00444/history

We denote this fixed dimension set by $\mathcal{D} = \{d_1, \ldots, d_{14}\}$.

All 14 dimensions use independently constructed standardized functional dictionaries from ProTrek's official retrieval library[3]. Each dictionary supplies a fixed candidate text vocabulary shared across proteins, rather than a mapping from test-protein identities to their reference annotations. Dictionary construction does not use PFS test labels, and the vocabulary is not tailored to the correct answers of individual test proteins.

For the EC dimension, each category has a standardized functional description that remains the same across queries. The ProTrek EC text library uses the IUBMB enzyme list and KEGG Enzyme category descriptions, including enzyme names, reactions, substrates, and products (Su et al., 2025). Retrieving a correct EC description therefore predicts its applicability to the input sequence; its presence in the dictionary does not disclose the protein's curated EC assignment.

Table B.2: **Sources and version information for the standardized functional dictionaries used by CAPA-RAG.**

Dimensions covered	Distributed source	Construction and version information
EC number	ProTrek EC text sublibrary	IUBMB `2024_03` category list and KEGG Enzyme `2024_09` descriptions, standardized with `gpt-4o-2024-11-20`; ProTrek (Su et al., 2025).
Function; Catalytic activity; GO Molecular Function; Rhea reaction; Cofactor; Active site; Binding site; Domain; UniPathway; Reactome Pathway; GO Biological Process; Subcellular location; GO Cellular Component	Independent standardized functional dictionaries in ProTrek's official retrieval library	Official May 2025 release, using the fixed 14-dimensional dictionaries described in Appendix B.1.

Both rows in Table B.2 use the fixed ProTrek library snapshot recorded in Appendix E.8. The snapshot identifies the distributed resources, while the EC-specific entries above describe that sublibrary's upstream construction. The shared-dictionary principle applies across all dimensions: retrieval predicts which candidate descriptions apply to a sequence. Removing a category solely because it occurs in the test labels would alter the prediction vocabulary and introduce a separate category-exclusion task. We therefore retain the independently defined candidate vocabularies.

Initial retrieval follows Equation equation 3 using the fixed sequence-to-text index. Each candidate stores its annotation text, standardized identifiers when available, raw ProTrek score, within-dimension rank, retrieval round, and available provenance. The query contains only the protein sequence and functional dimension; no gold EC value or pair label is injected into retrieval. Candidate records are predicted evidence, distinct from the curated annotations used to assign PFS labels. ProTrek scores order candidates within a protein and dimension; they are neither compared across dimensions nor interpreted as annotation probabilities.

B.2 Relation Types and Conflict Prioritization

$$\mathcal{K}_p^{(t)} = (\mathcal{V}_p^{(t)}, \mathcal{E}_p^{(t)}), \qquad \mathcal{V}_p^{(t)} = \bigcup_{d \in \mathcal{D}} \mathcal{C}_{p,d}^{(t)}. \tag{B.1}$$

The graph in Equation equation B.1 evaluates the Cartesian product of candidates for each configured dimension pair in Table B.3. It does not evaluate all possible cross-dimensional relationships. Frozen identifier, hierarchy, and reaction relations take precedence; remaining pairs are batched in groups of at most 32 for the Conflict Agent. Each requested pair receives exactly one relation label and a non-empty evidence field:

$$\mathcal{L}_G = \{\texttt{support}, \texttt{compatible}, \texttt{hierarchical}, \texttt{context_dependent}, \texttt{contradict}, \texttt{unknown}\}. \tag{B.2}$$

[3] https://huggingface.co/datasets/westlake-repl/faiss_index

Table B.3: **Configured relation pairs and fixed conflict priorities.** Higher values are expanded first among conflicts with remaining retrieval capacity.

Dimension A	Dimension B	Priority
Catalytic activity	EC number	3
EC number	Rhea reaction	3
Function	GO Molecular Function	2
Cofactor	Binding site	2
UniPathway	GO Biological Process	1
Reactome Pathway	GO Biological Process	1
Subcellular location	GO Cellular Component	1

`support` requires a specific shared claim or explicit standardized relation; `compatible` means only that claims could coexist in the same protein and context. `hierarchical` does not establish the more specific claim, and neither `unknown` nor `context_dependent` is positive confirmation. A `contradict` relation identifies mutually incompatible same-context claims without deciding which is correct. Different identifiers, source organisms, missing information, or compatible multifunctionality alone are not contradictions. Rhea master identifiers group reaction families while retained direction information still matters. Reference residue positions are not verified coordinates in the target protein.

Every contradiction is entered into a historical ledger with its original edge, first and last contradiction rounds, latest observed edge, and observation history. Later reassessment updates the latest edge but does not remove the original contradiction or close the ledger entry. The expansion queue uses the original edges of open entries that still have expandable dimensions, ordered by decreasing fixed priority and then candidate identifiers. Thus ranking advice and later compatibility judgments cannot silently suppress unresolved alternatives.

B.3 Retrieval Expansion and Reranking

Round 1 retrieves k_0 entries per dimension, builds relations, records conflicts, and produces advisory ranking. Retrieval-Agent calls begin only in subsequent rounds. The controller forms the union of expandable dimensions implicated by all open conflicts at the highest eligible priority. Although the system prompt describes a highest-priority conflict in the singular, tied conflicts are combined by this controller rule. The agent must return every eligible dimension exactly once. It cannot choose a different subset, rewrite the sequence query, fetch annotations by identifier, or search outside the frozen ProTrek rankings.

The current retrieval depth $k_{p,d}^{(t)}$ is updated by

$$k_{p,d}^{(t+1)} = \min\Big(k_{\max}, k_{p,d}^{(t)} + \Delta_{p,d}^{(t)}\Big), \tag{B.3}$$

where $\Delta_{p,d}^{(t)} \geq 0$ is the budget-limited increment (zero for dimensions not expanded) and $k_{\max}$ is the maximum depth.

Dimensions are expanded in sorted key order. Each requested depth is bounded by the configured increment, $k_{\max}$, and the remaining expansion budget. The budget counts newly read ranking entries, including duplicate entries; it does not count the initial retrieval. The prefix of each extended ranking must agree with the previous retrieval. The unseen rank interval is

$$\Delta\mathcal{C}_{p,d}^{(t+1)} = \mathrm{Top}_{k_{p,d}^{(t+1)}}[\mathrm{Ret}_{\mathrm{s2t}}(x_p, d; \mathcal{I}_{\mathrm{PT}})] \setminus \mathrm{Top}_{k_{p,d}^{(t)}}[\mathrm{Ret}_{\mathrm{s2t}}(x_p, d; \mathcal{I}_{\mathrm{PT}})]. \tag{B.4}$$

Normalization enriches identifier and relation metadata while preserving candidate text. Ingestion deduplicates candidate IDs and retains duplicate provenance. The graph and historical ledger are updated before Ranking runs.

Ranking returns `ordered_ids`, a non-empty `rationale`, and `uncertainty_notes` for all 14 dimensions. Each order must contain every supplied candidate ID exactly once, including unsupported or contradictory hypotheses; an empty order is valid only for an empty candidate pool. Candidate texts and retrieval scores are immutable. Ranking advice is stored separately and cannot

alter the set or original retrieval order of candidates visible to reasoning. If functional evidence does not distinguish alternatives, the prompt asks for retrieval order to be preserved and the limitation stated. Specific functional semantics may motivate advice, but matching identifiers and repeated cross-references are not independent votes or guaranteed truths.

B.4 Termination and Profile Schema

After graph construction, ledger update, and Ranking, the controller records the completed round and applies the following ordered stopping checks:

Order	Stopping condition	Termination status
1	No historical open conflict	`no_conflict`
2	The round limit is reached	`round_limit_unresolved`
3	The expansion budget is exhausted	`retrieval_budget_exhausted_unresolved`
4	No open conflict has remaining retrieval capacity	`retrieval_depth_limit_unresolved`
5	After round 1, expansion adds no unseen candidate	`no_unseen_candidates_unresolved`

At a resource limit, open entries become `unresolved_at_limit` and retain the terminal reason. The controller records whether advice changed, but stability does not terminate expansion or resolve a conflict. The `no_conflict` status concerns the recorded graph, not global biological consistency.

The final evidence profile can be summarized as

$$\mathbf{z}_p = (\{(\mathcal{C}_{p,d}, \alpha_{p,d}, \pi_{p,d}, \eta_{p,d}, \mathcal{P}_{p,d})\}_{d\in\mathcal{D}}, \mathcal{K}_p, \mathcal{H}_p, \mathcal{M}_p)\,, \tag{B.5}$$

where $\mathcal{C}_{p,d}$ contains all retrieved candidates, $\alpha_{p,d}$ records present/empty availability, $\pi_{p,d}$ is the optional advisory permutation, $\eta_{p,d}$ contains its rationale and uncertainty notes, and $\mathcal{P}_{p,d}$ stores provenance. The graph $\mathcal{K}_p$, historical conflict ledger $\mathcal{H}_p$, and metadata $\mathcal{M}_p$ preserve relation coverage, profiling mode, depths, budgets, and stopping information. The downstream serialization retains all candidates and available advice, current relations, original and latest conflict edges, and mode limitations, while omitting protein database identity, raw sequence, and local storage paths.

Execution modes. In `integrated` mode, relations and expansion are evaluated as needed and Ranking runs every completed round. With `no_ranking`, candidate pools, relations, and historical conflicts remain visible but advice is omitted. When $k_{\max} = 1$ and the total-round limit is 1, `retrieval_only` takes precedence: all CAPA agents are skipped and Top-1 candidates are passed downstream with an explicit statement that relations and ranking were not evaluated. Legacy-profile reading exposes historical selections and rejections as advice, not visibility filters. All modes use the same downstream reasoning template.

C Reasoning Output and Validation

The serialized view includes all annotation candidates, current relations, historical conflicts, and available ranking advice. It omits protein database identity, raw sequence, benchmark labels, and local storage paths. Evidence priority runs from reaction chemistry and specificity (catalytic activity, Rhea, EC), through functional interpretation (function, GO Molecular Function) and mechanism (cofactor, active site, binding site, domain), to biological context (pathways, process, localization). This is a relevance order, not a vote count: specific lower-priority evidence can reveal an applicability error in a higher-priority claim. Whole-complex activity, individual-subunit roles, and supported multifunctionality remain distinct.

Binary PFS evaluation uses

$$\hat{y}^{\mathrm{PFS}}_{pp'} = \mathbb{I}[\hat{s}_{pp'} > 0.5]. \tag{C.1}$$

Thus a score of exactly 0.5 is assigned to the negative class. The evidence sentence states the decision basis and is not an independently validated mechanistic explanation.

C.1 Similarity Levels and Score Intervals

The common 14-dimensional input view is defined in Section 3.3. Both applications use the same output schema and shared-activity criterion. The output label and admissible score interval are fixed:

Table C.1: **Ordinal similarity levels and their admissible score intervals.**

Similarity level	Score interval
`exact_match`	$[0.90, 1.00]$
`likely_match`	$[0.65, 0.90)$
`uncertain`	$(0.35, 0.65)$
`unlikely_match`	$(0.10, 0.35]$
`no_match`	$[0.00, 0.10]$

C.2 Parsing and Output Validation

The response contains exactly `similarity_level`, `similarity_score`, and `evidence`. A deterministic parser checks the legal level, finite in-band score, and non-empty evidence. The prompt requires a semantic level to be chosen first, followed by an in-band score and one concise evidence sentence of at most 40 words. The sentence names the shared activity, decisive distinction, or unresolved alternative, and should explain why any material counterinterpretation is outweighed. Appendix G.5 reproduces the exact semantic scoring rules. Scores are not calibrated probabilities; repeated annotations or CAPA rationales are not additional observations. Valid output structure does not establish biological correctness.

For applicable non-empty responses that fail validation, the client includes the prior answer and validation feedback in a retry. Truncated responses follow a separate output-budget retry branch. Both retain the original task and evidence.

D Protein-to-Reaction Mapping and Network Assembly

D.1 Network and GPR Notation

Let $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ be the target proteome and $\mathcal{R} = \{r_j\}_{j=1}^{N_{\text{ref}}}$ the reference-protein collection. Let

$$\mathcal{U} = (\mathcal{M}, \mathcal{A}, \mathbf{S})$$

denote the universal metabolic network, where $\mathcal{M}$ and $\mathcal{A}$ are its metabolites and reactions and $\mathbf{S} \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{A}|}$ is the stoichiometric matrix. For each reaction a, let $\mathcal{H}_a = \{1, \dots, H_a\}$ index its alternative catalytic assemblies and let $\mathcal{J}_{a,h}$ be the set of protein roles required by assembly h. We represent the reference GPR annotations by $\Gamma \subseteq \{(r_j, a, h, u) : r_j \in \mathcal{R},\ a \in \mathcal{A},\ h \in \mathcal{H}_a,\ u \in \mathcal{J}_{a,h}\}$, where $(r_j, a, h, u) \in \Gamma$ states that reference protein r_j fills role u in assembly h. The reconstruction task assigns function-based evidence to these roles and selects a feasible subset $\mathcal{A}^\star$, producing the species-specific model

$$\mathcal{G}^\star = (\mathcal{M}^\star, \mathcal{A}^\star, \mathbf{S}^\star). \tag{D.1}$$

D.2 Candidate Generation and Score Propagation

For query protein q_i and reference protein r_j, let

$$b_{ij} = \text{SeqSim}(x_{q_i}, x_{r_j}) \tag{D.2}$$

be their sequence-based retrieval score. The reference-protein candidate set is

$$\mathcal{C}_i^{\text{ref}} = \underset{r_j \in \mathcal{R}}{\text{Top}_{K_{\text{ref}}}}\ b_{ij}, \qquad K_{\text{ref}} = 5. \tag{D.3}$$

This stage limits the expensive functional comparisons to five candidates per query. It also defines the candidate-recall ceiling: functional reasoning cannot recover a reference protein that is absent from $\mathcal{C}_i^{\mathrm{ref}}$.

The Reasoning Agent evaluates each query–reference pair once in GEM mode:

$$(\hat{\ell}_{ij}^{\mathrm{GEM}}, c_{ij}, r_{ij}) = F_{\mathrm{reason}}(\mathbf{z}_{q_i}, \mathbf{z}_{r_j}; m = \mathrm{GEM}). \tag{D.4}$$

The functional score directly replaces the sequence score:

$$\widetilde{b}_{ij} = c_{ij}. \tag{D.5}$$

No sequence–function fusion, probability calibration, or label-based gate is applied. The five candidates are ordered by $\widetilde{b}_{ij}$ for deterministic reporting, but all five continue to reaction scoring. Thus, the downstream network is changed by score replacement, not by the permutation itself. The original sequence score b_{ij} is discarded after candidate generation. Because no rejection gate is used, every positive functional score remains as soft genetic evidence, including low scores assigned to weakly related candidates.

Each pair-level score is computed only once and is inherited by every reference reaction and GPR role associated with r_j. Thus, the scored transferred associations are

$$\widehat{\Gamma} = \{(q_i, r_j, a, h, u, \widetilde{b}_{ij}) : r_j \in \mathcal{C}_i^{\mathrm{ref}},\ (r_j, a, h, u) \in \Gamma\}. \tag{D.6}$$

This protein-pair decision is not recomputed separately for individual reactions: the same score is copied to every reaction and GPR role associated with the reference protein. It therefore measures support for a shared activity rather than verifying each transferred reaction; a match does not establish that every activity of a multifunctional reference can be transferred.

D.3 GPR-Aware Aggregation

We next aggregate pair scores according to the Boolean structure of each reference GPR rule. For role u in catalytic assembly h, let $\mathcal{B}_{a,h,u}$ collect the transferred scores assigned to that role. The role, assembly, and raw reaction scores are

$$\begin{aligned}
\mathcal{B}_{a,h,u} &= \{\widetilde{b}_{ij} : (q_i, r_j, a, h, u, \widetilde{b}_{ij}) \in \widehat{\Gamma}\}, \\
g_{a,h,u} &= \begin{cases} \max \mathcal{B}_{a,h,u}, & \mathcal{B}_{a,h,u} \neq \varnothing, \\ 0, & \mathcal{B}_{a,h,u} = \varnothing, \end{cases} \\
s_{a,h}^{\mathrm{complex}} &= \frac{1}{|\mathcal{J}_{a,h}|} \sum_{u \in \mathcal{J}_{a,h}} g_{a,h,u}, \\
s_a^{\mathrm{raw}} &= \max_{h \in \mathcal{H}_a} s_{a,h}^{\mathrm{complex}}.
\end{aligned} \tag{D.7}$$

The maximum in the first stage collapses multiple query hits assigned to the same reference GPR role, so duplicate matches are not counted as additional isozymes. Following the CarveMe 1.6.5 implementation, a complex score is the arithmetic mean of its subunit scores, with missing scores set to zero; the raw reaction score is the maximum over isozymes or alternative complexes. This numerical aggregation does not require positive evidence for every subunit.

D.4 Reaction-Score Normalization

Let $\mathcal{A}_{\mathrm{enz}} \subseteq \mathcal{A}$ be the enzyme-catalyzed reactions and $\mathcal{A}_{\mathrm{gen}} = \{a \in \mathcal{A}_{\mathrm{enz}} : s_a^{\mathrm{raw}} > 0\}$ the subset with positive transferred genetic evidence. Define

$$\mu = \mathrm{median}\{s_a^{\mathrm{raw}} : a \in \mathcal{A}_{\mathrm{gen}}\}.$$

The CarveMe-compatible reaction score is

$$s_a^{\mathrm{rxn}} = \begin{cases} s_a^{\mathrm{raw}}/\mu, & a \in \mathcal{A}_{\mathrm{gen}}, \\ -1, & a \in \mathcal{A}_{\mathrm{enz}} \setminus \mathcal{A}_{\mathrm{gen}}, \\ 0, & a \in \mathcal{A} \setminus \mathcal{A}_{\mathrm{enz}}. \end{cases} \tag{D.8}$$

This mutually exclusive partition assigns a penalty to unsupported enzyme-catalyzed reactions and a neutral score to non-enzymatic reactions, including spontaneous and boundary reactions. The normalization sets the median positive reaction evidence to 1. We assume $\mathcal{A}_{\mathrm{gen}} \neq \varnothing$; if no reaction receives positive evidence, reconstruction terminates before normalization and reports the absence of transferable genetic evidence. Because every member of $\mathcal{C}_i^{\mathrm{ref}}$ is retained, all five function-based scores can contribute after duplicate hits to the same GPR role have been collapsed.

D.5 CarveMe Optimization and Assembly

GEM-Reasoner uses the original CarveMe carving formulation without methodological modification. Let $\mathbf{v} \in \mathbb{R}^{|\mathcal{A}|}$ be the flux vector, and let $\mathbf{y}^f, \mathbf{y}^r \in \{0,1\}^{|\mathcal{A}|}$ indicate whether each reaction is selected in the forward or reverse direction. The MILP is

$$\begin{aligned} \max_{\mathbf{v},\mathbf{y}^f,\mathbf{y}^r} \quad & (\mathbf{s}^{\mathrm{rxn}})^\top (\mathbf{y}^f + \mathbf{y}^r) \\ \text{s.t.} \quad & \mathbf{S}\mathbf{v} = \mathbf{0}, \\ & -M_{\mathrm{flux}}\mathbf{y}^r + \epsilon \mathbf{y}^f \le \mathbf{v} \le -\epsilon \mathbf{y}^r + M_{\mathrm{flux}}\mathbf{y}^f, \\ & \mathbf{y}^f + \mathbf{y}^r \le \mathbf{1}, \\ & \mathbf{y}^f, \mathbf{y}^r \in \{0,1\}^{|\mathcal{A}|}, \\ & v_{\mathrm{growth}} \ge v_{\mathrm{growth}}^{\min}. \end{aligned} \tag{D.9}$$

For forward-irreversible reactions, $y_a^r = 0$ and $v_a \ge 0$; for backward-irreversible reactions, $y_a^f = 0$ and $v_a \le 0$. We use the defaults reported in the original CarveMe formulation: $\epsilon = 0.001$ mmol gDW^{-1} h^{-1}, $M_{\mathrm{flux}} = 100$ mmol gDW^{-1} h^{-1}, and $v_{\mathrm{growth}}^{\min} = 0.1$ h^{-1}. The selected reaction set is

$$\mathcal{A}^\star = \{a \in \mathcal{A} : y_a^f + y_a^r = 1\}. \tag{D.10}$$

After optimization, inactive reactions and the genes and metabolites orphaned by their removal are deleted, yielding $\mathcal{G}^\star$. GEM-Reasoner's methodological change is therefore confined to the construction of function-based reaction evidence; network selection follows CarveMe.

E Experimental Protocols and Detailed Results

E.1 PFS Baseline Configurations

We select baseline methods spanning sequence alignment, structure alignment, and learned protein representations:

- **Sequence alignment methods:** DIAMOND and MMseqs2;
- **Structure alignment methods:** Foldseek;
- **Sequence protein language models:** ESM-2-8M, ESM-2-35M, ESM-2-150M, ESM-2-650M, ESM-2-3B, ESM-C-300M, and ESM-C-600M;
- **Structure-aware protein models:** SaProt-35M, SaProt-650M, and SaProt-1.3B;
- **Multimodal protein models:** ProTrek-35M and ProTrek-650M, each evaluated using sequence-to-sequence and structure-to-structure similarity;
- **Function-aware protein models:** CLEAN.

For methods requiring protein structures, we adopt a unified structure acquisition hierarchy. We first retrieve experimentally resolved structures from the Protein Data Bank (PDB). If no corresponding

experimental structure is available, we use the predicted structure from the AlphaFold Protein Structure Database. If neither source provides a structure, we predict one from the protein sequence using ESMFold. This protocol ensures that Foldseek, SaProt, and the structure-based ProTrek variants use consistent sources of structural input.

For DIAMOND, MMseqs2, and Foldseek, we use pairwise alignment scores as similarity scores. For ESM-2, ESM-C, SaProt, and ProTrek, we extract global representations in the corresponding modality and compute cosine similarity between the two protein embeddings.

CLEAN learns function-aware protein representations through EC-supervised contrastive learning on Swiss-Prot (Yu et al., 2023). We generate CLEAN embeddings directly from protein sequences and quantify pairwise similarity using their cosine similarity. This evaluation does not run CLEAN's EC annotation procedure or classify pairs using predicted EC labels. All methods use the same test pairs and evaluation metrics, with thresholds determined by the validation protocol below.

CLEAN therefore provides a representation baseline that has already learned from curated functional annotations. A shared Swiss-Prot source does not equate training samples, scale, supervision, or test exposure. ProTrek additionally uses TrEMBL50 (Su et al., 2025), and GEM-Reasoner combines standardized functional descriptions with retrieval, conflict assessment, and reasoning. The comparison supports performance advantages under the evaluated protocol; it does not establish the absence of training overlap or isolate a single source of improvement.

Decision thresholds. GEM-Reasoner uses the fixed classification rule $s > 0.5$. For each PFS baseline, we select the threshold that maximizes MCC on an independent validation set constructed from Swiss-Prot in the same pair format as PFS, without additional special processing. The validation set contains 200,000 pairs, balanced between 100,000 positive and 100,000 negative examples. The selected threshold is held fixed for test evaluation and is shared by PFS-Core and PFS-Challenge; test labels are not used for selection. Table E.5 summarizes the configuration and result correspondence.

Component ablations. The "w/o CAPA-RAG" variant retains ProTrek sequence-to-text retrieval across all 14 functional dimensions and passes the retrieved annotations directly to downstream reasoning. It bypasses CAPA-RAG integration, conflict assessment, and ranking, so this comparison evaluates the contribution of CAPA-RAG processing beyond the retrieved annotations. The "w/o functional evidence" variant provides only the two protein sequences to the LLM, without retrieved functional annotations or CAPA-RAG profiles. The "w/o CoT" variant removes chain-of-thought prompting. These GEM-Reasoner variants retain the fixed decision rule $s > 0.5$ used in Table 2.

E.2 Complete PFS Benchmark Results

Table E.1 reports results for all evaluated methods, including ESM2-8M and ESM2-35M, which are omitted from Table 1 for brevity.

Table E.1: **Complete protein functional similarity results on PFS.** PFS-Core and PFS-Challenge each contain 10,000 positive and 10,000 negative pairs. GEM-Reasoner uses $s > 0.5$; baseline thresholds are optimized on independent validation data and fixed for testing. Best and second-best values within each setting are boldfaced and underlined.

	PFS-Core					PFS-Challenge				
Method	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑
ESM2-8M	0.8822	0.8817	0.7644	0.8856	0.8778	0.3498	0.3865	-0.3027	0.3658	0.4096
ESM2-35M	0.8624	0.8615	0.7248	0.8668	0.8563	0.3258	0.3491	-0.3494	0.3374	0.3616
ESM2-150M	0.8731	0.8701	0.7470	0.8912	0.8500	0.3370	0.3625	-0.3272	0.3491	0.3771
ESM2-650M	0.8573	0.8479	0.7202	0.9079	0.7953	0.2871	0.2592	-0.4270	0.2698	0.2495
ESM2-3B	0.8564	0.8442	0.7216	0.9224	0.7782	0.2807	0.2504	-0.4401	0.2614	0.2403
ESMC-300M	0.8418	0.8382	0.6842	0.8576	0.8196	0.3658	0.4075	-0.2712	0.3823	0.4362
ESMC-600M	0.8542	0.8550	0.7084	0.8504	0.8596	0.3896	0.4464	-0.2257	0.4084	0.4923
SaProt-35M	0.8284	0.8144	0.6644	0.8866	0.7531	0.3194	0.3039	-0.3616	0.3110	0.2972
SaProt-650M	0.8372	0.8296	0.6771	0.8702	0.7926	0.3039	0.3042	-0.3923	0.3040	0.3043
SaProt-1.3B	0.8644	0.8616	0.7294	0.8797	0.8442	0.4047	0.4499	-0.1932	0.4181	0.4868
ProTrek-35M-Seq	0.9130	0.9142	0.8263	0.9015	0.9273	0.4732	0.5539	-0.0575	0.4803	0.6540
ProTrek-35M-Str	0.9182	0.9194	0.8367	0.9052	0.9341	0.5297	0.6234	0.0684	0.5198	0.7786
ProTrek-650M-Seq	0.9096	0.9098	0.8192	0.9082	0.9113	0.4934	0.5689	-0.0141	0.4951	0.6685
ProTrek-650M-Str	0.9166	0.9154	0.8334	0.9281	0.9031	0.5143	0.5654	0.0293	0.5115	0.6320
MMseqs2	0.9014	0.8945	0.8096	0.9616	0.8361	0.1407	0.0021	-0.7480	0.0025	0.0018
DIAMOND	0.8865	0.8763	0.7836	0.9624	0.8044	0.1463	0.0051	-0.7377	0.0061	0.0044
Foldseek	0.7816	0.7380	0.5973	0.9222	0.6151	0.4530	0.4537	-0.0941	0.4531	0.4544
CLEAN	0.9522	0.9506	0.9062	**0.9830**	0.9202	0.7686	0.7594	0.5387	0.7907	0.7304
GEM-Reasoner	**0.9600**	**0.9598**	**0.9199**	0.9623	**0.9574**	**0.8534**	**0.8529**	**0.7068**	**0.8560**	**0.8498**

E.3 Organisms and Reconstruction Inputs

To determine whether protein function reasoning improves the biological predictive capability of downstream GEMs, we follow the substrate utilization and gene essentiality evaluation protocols introduced by CarveMe. We evaluate three model strains for which both types of experimental data are available:

- *Bacillus subtilis* subsp. *subtilis* 168, RefSeq assembly accession `GCF_000009045.1`[4] (ASM904v1);
- *Escherichia coli* K-12 MG1655, RefSeq assembly accession `GCF_000005845.2`[5] (ASM584v2);
- *Pseudomonas aeruginosa* PAO1, RefSeq assembly accession `GCF_000006765.1`[6] (ASM676v1).

All reconstruction methods receive identical genome inputs, medium specifications, and gap-filling conditions, thereby preventing differences in input information or post-processing configurations from confounding the comparison.

GEM reconstruction baselines. We compare GEM-Reasoner with three automated GEM reconstruction tools: ModelSEED, CarveMe, and AlphaGEM. All methods use the same strain genomes as input and are evaluated under identical medium constraints and gap-filling conditions. We do

[4] https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000009045.1
[5] https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2
[6] https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000006765.1

not include manually curated GEMs as baselines, thereby focusing the evaluation on performance differences among automated reconstruction frameworks.

E.4 Substrate-Utilization Evaluation

We use the public Biolog Phenotype MicroArray data adopted in the CarveMe benchmark (Machado et al., 2018). Each model is constrained to an M9 minimal medium. For each array, candidate compounds individually replace the default source of the corresponding element: glucose for carbon, ammonium for nitrogen, sulfate for sulfur, and phosphate for phosphorus. The remaining medium constraints are retained. The evaluation covers all four nutrient-source categories for *B. subtilis* and *E. coli*, and carbon sources for *P. aeruginosa*. Each substrate condition constitutes one binary evaluation example.

For both downstream tasks, FBA predicts the maximum growth rate by maximizing biomass flux:

$$\begin{aligned} v_{\text{growth}}^{\max}(\mathcal{G}, e) = \max_{\mathbf{v}} \; & v_{\text{growth}} \\ \text{s.t.} \quad & \mathbf{S}_{\mathcal{G}}\mathbf{v} = \mathbf{0}, \\ & \mathbf{l}_{\mathcal{G}}(e) \le \mathbf{v} \le \mathbf{u}_{\mathcal{G}}(e), \end{aligned} \tag{E.1}$$

where $\mathbf{S}_{\mathcal{G}}$ is the model's stoichiometric matrix and $\mathbf{l}_{\mathcal{G}}(e)$, $\mathbf{u}_{\mathcal{G}}(e)$ encode reaction capacities, directionality, and environmental exchange constraints. The biomass flux represents the specific growth rate in h^{-1}.

In the substrate screens, the maximum uptake rate of each permitted medium compound, including the candidate substrate, is $10\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. We use $\tau_{\mathrm{SU}} = 0.1\,\mathrm{h}^{-1}$ in Equation (2). A feasible condition with maximum growth strictly above this threshold is classified as growth-supporting; growth at or below the threshold or an infeasible condition is classified as non-growth. The same threshold applies to all reconstruction methods and organisms. Reference labels encode experimentally observed growth as $y_s^{\mathrm{SU}} = 1$ and non-growth as $y_s^{\mathrm{SU}} = 0$.

E.5 Gene-Essentiality Evaluation

We use the gene-essentiality data and media adopted by CarveMe (Machado et al., 2018). Simulations use M9 with glucose for *E. coli*, M9 with succinate for *P. aeruginosa*, and LB for *B. subtilis*. Essentiality is defined with respect to the specified medium.

For each gene g, we set its GPR state to false while retaining the other genes, reevaluate the Boolean GPR rules, and constrain to zero the flux of every reaction whose rule becomes false. Reactions supported by an intact alternative isozyme remain available. The resulting model $\mathcal{G}^{\star}_{-g}$ is evaluated by Equation (E.1) under the same medium constraints as the wild-type model.

For a feasible wild-type model with positive maximum growth, we use $\rho = 0.1$ in Equation (2), matching the relative-growth criterion in CarveMe's accompanying evaluation implementation. A gene is predicted essential when the deletion model's maximum growth is below 10% of the wild-type value, or when the deletion model is infeasible. A feasible deletion retaining at least 10% of wild-type growth is predicted non-essential. This comparison uses each reconstructed model's own wild-type growth rate in the same medium.

For each organism, evaluation is restricted to the intersection of genes consistently mapped across all compared models. Reference labels follow the CarveMe benchmark: for *B. subtilis* and *P. aeruginosa*, we retain genes explicitly labeled essential or non-essential; for *E. coli*, genes in the benchmark's essential-gene list are positive, and the remaining genes in the common mapped set are negative. Accordingly, $y_g^{\mathrm{GE}} = 1$ denotes an essential reference gene and $y_g^{\mathrm{GE}} = 0$ a non-essential reference gene. Genes without a consistent mapping, or without an admissible reference label where explicit labels are required, are excluded from metric computation.

E.6 Detailed GEM Reconstruction Results

Tables E.2 and E.3 provide the complete metric breakdown underlying Figure 2. Results are reported to four decimal places for each organism and as arithmetic macro averages across the three organisms.

Substrate-utilization metrics. GEM-Reasoner's Sensitivity is 0.9520, 0.8389, and 0.8824 for *B. subtilis*, *E. coli*, and *P. aeruginosa*, respectively. It has the highest Precision on *E. coli*, whereas CarveMe has the highest Precision on *B. subtilis* and *P. aeruginosa*. CarveMe also has higher macro-average Precision and Specificity, so GEM-Reasoner's F1 and Sensitivity gains do not extend to every metric.

Table E.2: **Complete substrate-utilization prediction results corresponding to the top row of Figure 2.** All methods use identical genome inputs, medium information, and gap-filling conditions; positive predictions require a biomass objective > 0.1. Macro averages are arithmetic means over the three organisms. Best values are boldfaced and second-best values are underlined.

Organism	Method	Accuracy↑	Precision↑	Sensitivity↑	Specificity↑	F1↑
B. subtilis	CarveMe	0.4552	**0.9107**	0.4080	**0.7500**	0.5635
	ModelSEED	0.4621	0.8310	0.4720	0.4000	0.6020
	AlphaGEM	0.7034	0.8661	0.7760	0.2500	0.8186
	GEM-Reasoner	**0.8552**	0.8881	**0.9520**	0.2500	**0.9189**
E. coli	CarveMe	0.6345	0.6938	0.7450	**0.4494**	0.7184
	ModelSEED	0.5798	0.6485	0.7181	0.3483	0.6815
	AlphaGEM	0.6050	0.6627	0.7517	0.3596	0.7044
	GEM-Reasoner	**0.6849**	**0.7102**	**0.8389**	0.4270	**0.7692**
P. aeruginosa	CarveMe	**0.6667**	**0.6098**	0.7353	**0.6098**	0.6667
	ModelSEED	0.5733	0.5294	0.5294	**0.6098**	0.5294
	AlphaGEM	0.4800	0.4615	**0.8824**	0.1463	0.6061
	GEM-Reasoner	0.6400	0.5660	**0.8824**	0.4390	**0.6897**
Macro average	CarveMe	0.5854	**0.7381**	0.6294	**0.6031**	0.6495
	ModelSEED	0.5384	0.6696	0.5732	0.4527	0.6043
	AlphaGEM	0.5961	0.6634	0.8034	0.2520	0.7097
	GEM-Reasoner	**0.7267**	0.7214	**0.8911**	0.3720	**0.7926**

Gene-essentiality metrics. GEM-Reasoner has the highest Precision on all three organisms. It ties with AlphaGEM for the highest Specificity on *B. subtilis* (0.9764) and leads on *E. coli*, whereas AlphaGEM has the highest Specificity on *P. aeruginosa*. GEM-Reasoner leads in macro-average Accuracy (0.9008), Precision (0.7785), Sensitivity (0.4130), and F1 (0.5392). AlphaGEM has the highest macro-average Specificity (0.9861), followed by GEM-Reasoner (0.9811).

Table E.3: **Complete gene-essentiality prediction results corresponding to the bottom row of Figure 2.** All methods are evaluated under the same medium and on the common set of mapped genes. Macro averages are arithmetic means over the three organisms. Best values are boldfaced and second-best values are underlined.

Organism	Method	Accuracy↑	Precision↑	Sensitivity↑	Specificity↑	F1↑
B. subtilis	CarveMe	**0.9007**	0.6271	**0.5441**	0.9528	**0.5827**
	ModelSEED	0.8071	0.3232	0.4706	0.8562	0.3832
	AlphaGEM	0.8783	0.5600	0.2059	**0.9764**	0.3011
	GEM-Reasoner	**0.9007**	**0.7027**	0.3824	**0.9764**	0.4952
E. coli	CarveMe	0.8701	0.5926	0.1410	0.9848	0.2278
	ModelSEED	0.8809	0.5680	**0.5154**	0.9384	0.5404
	AlphaGEM	0.8815	0.7101	0.2159	0.9861	0.3311
	GEM-Reasoner	**0.9192**	**0.8594**	0.4846	**0.9875**	**0.6197**
P. aeruginosa	CarveMe	0.8700	0.7059	0.3158	0.9751	0.4364
	ModelSEED	0.8448	0.5402	0.1767	0.9715	0.2663
	AlphaGEM	0.8424	0.6000	0.0338	**0.9957**	0.0641
	GEM-Reasoner	**0.8826**	**0.7734**	**0.3722**	0.9793	**0.5025**
Macro average	CarveMe	0.8803	0.6419	0.3336	0.9709	0.4156
	ModelSEED	0.8443	0.4771	0.3876	0.9220	0.3966
	AlphaGEM	0.8674	0.6234	0.1519	**0.9861**	0.2321
	GEM-Reasoner	**0.9008**	**0.7785**	**0.4130**	0.9811	**0.5392**

E.7 Reasoning-Backbone Sensitivity

Table E.4 reproduces the full five-metric comparison in Table 3. Only the final Reasoning LLM is replaced; Conflict, Retrieval, and Ranking remain `gemini-3-flash-preview`. The ProTrek resources, functional evidence, CAPA-RAG configuration, prompt structure, and decision threshold are fixed. Main-text analysis discusses the setting-dependent rankings and the nonmonotonic Qwen 3 size comparison.

Table E.4: **Reasoning-backbone sensitivity on PFS.** Access denotes open-source or closed-source availability. Only the final Reasoning LLM changes; functional evidence, the CAPA-RAG pipeline and agent models, prompt structure, and the decision rule $s > 0.5$ remain fixed. Best and second-best values are boldfaced and underlined within each setting.

		PFS-Core					**PFS-Challenge**				
Model	Access	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑	Acc.↑	F1↑	MCC↑	Prec.↑	Rec.↑
Qwen 3 4B	Open-source	0.9557	0.9557	0.9114	0.9556	0.9558	0.6534	0.6968	0.3202	0.6192	0.7967
Qwen 3 8B	Open-source	0.9474	0.9456	0.8968	0.9794	0.9140	0.7075	0.7084	0.4149	0.7062	0.7106
Qwen 3 14B	Open-source	0.9379	0.9343	0.8809	0.9911	0.8836	0.7782	0.7478	0.5733	0.8664	0.6578
Qwen 3 32B	Open-source	0.9383	0.9390	0.8768	0.9282	0.9501	0.7818	0.7648	0.5696	0.8293	0.7097
Claude 4.5 Haiku	Closed-source	0.9511	0.9487	0.9058	**0.9959**	0.9058	0.8281	0.8023	0.6796	**0.9437**	0.6977
ChatGPT 5 mini	Closed-source	0.9518	0.9515	0.9035	0.9557	0.9474	0.8420	0.8306	0.6902	0.8951	0.7747
Gemini 3 Flash	Closed-source	**0.9600**	**0.9598**	**0.9199**	0.9623	**0.9574**	**0.8534**	**0.8529**	**0.7068**	0.8560	**0.8498**

E.8 Experimental Configuration and Result Correspondence

Table E.5 summarizes the resources and settings used for the reported experiments. All reported PFS-Challenge results correspond to the `2024_03–2026_01` first-inclusion interval and the two temporal sampling rules in Appendix A.3.

Table E.5: **Experimental configuration and correspondence to the reported results.** Default settings apply to the full GEM-Reasoner pipeline; ablation and backbone substitutions are specified explicitly. Model and retrieval resources are linked to their official distributions.

Configuration category	Setting and result correspondence
Retrieval weights	ProTrek-650M, with 1,024-dimensional sequence embeddings. The checkpoint was publicly available in May 2024; Appendix A.4 describes the temporal relationship to the PFS-Challenge benchmark.
Indexes and annotations	ProTrek's official retrieval library, May 2025 release, with fixed sequence-to-text indexes and standardized dictionaries for all 14 functional dimensions (Appendix B.1).
PFS-Challenge benchmark	20,000 balanced pairs; first Swiss-Prot inclusion during `2024_03–2026_01`. The two temporal sampling groups and information boundaries are specified in Appendix A.
CAPA-RAG	Initial depth 1; increment 1; maximum depth 2; at most two rounds; expansion budget 28 entries per protein. Initial retrieval is excluded from this budget; expanded entries include duplicates. All three CAPA agents are enabled (Appendix B).
Agent models	All four agents use Gemini 3 Flash (`gemini-3-flash-preview`). Table 3 changes only the final Reasoning LLM.
Inference settings	Temperature 0.1; output limit 8,192 tokens, doubled after truncation up to 32,768. No separate thinking/reasoning budget is set.
Retries and failures	300-second timeout; up to three attempts per round and five validation/retry rounds (15 requests per client call). Invalid outputs receive deterministic feedback. Unresolved failures remain in evaluation and count as incorrect.
Baseline evaluation	200,000 balanced validation pairs from Swiss-Prot. Baseline thresholds maximize validation MCC and are fixed across PFS-Core and PFS-Challenge; GEM-Reasoner and its ablations use $s > 0.5$ (Appendix E.1).
GEM construction	CarveMe 1.6.5[7], with its bundled BiGG-derived reference proteins, GPR annotations, and default bacterial universal network. GEM-Reasoner replaces the pair-level evidence scores; aggregation and assembly follow Appendix D.
Media and gap filling	No medium-specific gap filling during default CarveMe construction. Evaluation uses M9 substrate screens with $\tau_{\mathrm{SU}} = 0.1\,\mathrm{h}^{-1}$ and growth strictly above the threshold; M9–glucose for *E. coli* essentiality, M9–succinate for *P. aeruginosa*, and LB for *B. subtilis* (Appendices E.4–E.5).
Correspondence to individual result tables and cases	
Tables 1 and E.1	Full configuration in PFS mode for GEM-Reasoner; baseline-specific scores and validation-selected thresholds for the other methods.
Table 2	Only the named component changes: CAPA-RAG processing, functional evidence, or CoT prompting. The Reasoning LLM and $s > 0.5$ remain fixed (Section 4.5).
Tables 3 and E.4	Only the final Reasoning LLM changes; retrieval, CAPA agents, functional evidence, prompt structure, and threshold remain fixed.
Tables E.2 and E.3	Full configuration in GEM mode for GEM-Reasoner; the other methods retain their own reconstruction frameworks. Evaluation follows the common protocols in Appendix E.
Protein-pair case studies	Full configuration on 400 pairs. Recorded runs reuse frozen Conflict/Retrieval responses where applicable and generate Ranking/Reasoning responses; replay makes no new LLM calls (Appendix F).
Proline-utilization case	Original GEM-Reasoner and CarveMe reconstructions for *B. subtilis* 168; M9 with glucose and L-proline as the sole external nitrogen source. The shared > 0.1 growth rule, flux analysis, and sensitivity checks are detailed in Appendix F.4.

[7] https://github.com/cdanielmachado/carveme/tree/1.6.5

F CASE STUDIES

The cases address two levels of the framework: protein-pair reasoning (Appendices F.1–F.3) and the relationship between reconstructed pathway content and a downstream utilization phenotype (Appendix F.4).

We examine one correct positive and one correct negative from an auxiliary trace analysis of 400 pairs (200 PFS-Core and 200 PFS-Challenge). These selected examples describe observed behavior, not a new estimate of benchmark performance. The positive pair comprises MurD proteins P14900 (*Escherichia coli*) and Q7VAM1 (*Prochlorococcus marinus*); both reference records specify EC 6.3.2.9. The negative pair comprises ATP synthase beta subunits P29707 (*Propionigenium modestum*; EC 7.2.2.1) and Q95DR6 (*Agapanthus africanus*; EC 7.1.2.2). Here the EC annotations describe activities of their respective complexes, not autonomous ion transport by isolated beta subunits. Reference identities and curated EC labels are used for retrospective checking only; reasoning receives aliases A/B and predicted evidence.

Figure F.1 summarizes the evidence updates and pairwise decisions; the original records and their interpretation are detailed below.

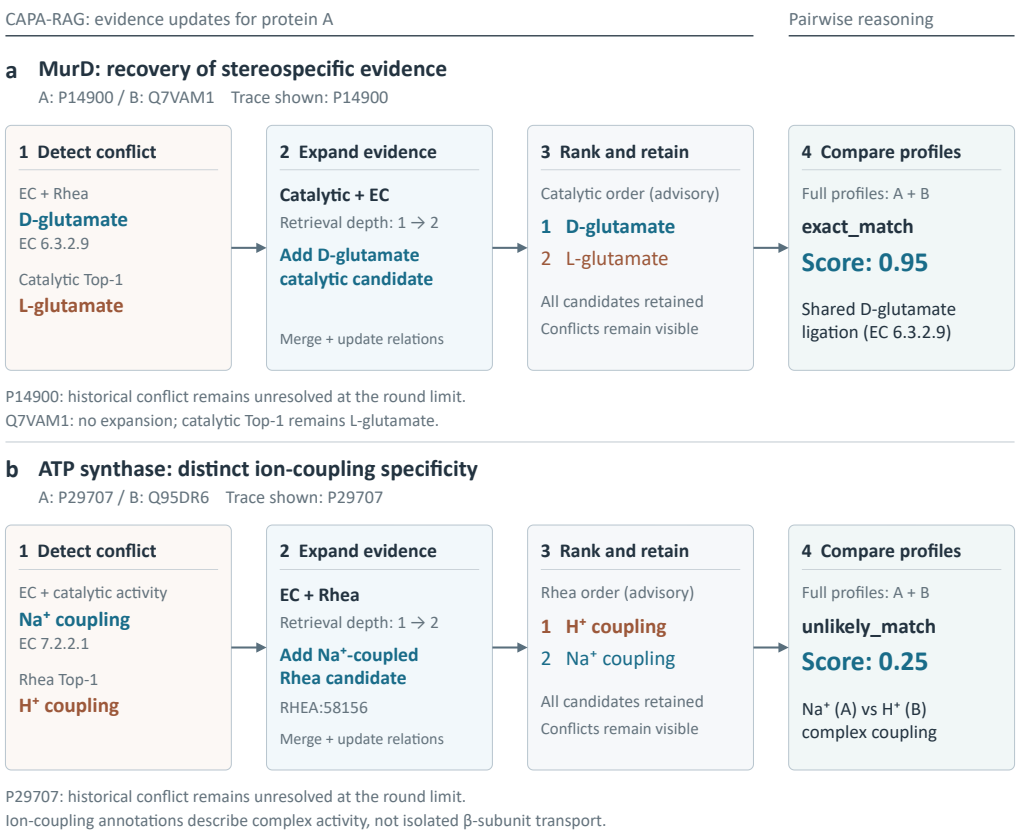


Figure F.1: **CAPA-RAG evidence updates and pairwise decisions for two selected cases.** The first three columns summarize protein A's trace; the last reports reasoning over both complete 14-dimensional profiles. **(a)** For P14900/Q7VAM1, expansion adds D-glutamate catalytic evidence for P14900 and Ranking places it first; Q7VAM1's catalytic Top-1 remains the L-glutamate alternative. **(b)** For P29707/Q95DR6, expansion adds a Na^+-coupled Rhea candidate for P29707, while the H^+ candidate remains first. All candidates are retained, and both expanded profiles stop with unresolved historical conflicts. Scores reproduce the observed outputs, not calibrated probabilities. Ion coupling denotes complex activity, not autonomous transport by isolated beta subunits.

Table F.1: **Observed pair decisions.** The similarity column reproduces the benchmark's original field and is not asserted to be full-length sequence identity. Binary predictions use $s > 0.5$.

Pair	Reference EC (A / B)	Similarity	Output level	Score	Gold / pred.
P14900 / Q7VAM1	6.3.2.9 / 6.3.2.9	0.264	`exact_match`	0.95	1 / 1
P29707 / Q95DR6	7.2.2.1 / 7.1.2.2	0.712	`unlikely_match`	0.25	0 / 0

Configuration and record conventions. The trace analysis uses the configuration in Table E.5. Recorded runs reuse matching frozen Conflict/Retrieval responses and generate Ranking and final Reasoning responses. Replay makes no new LLM calls and reproduces all four original profiles and states exactly.

The panels below show selected original fields or complete short outputs, with candidate IDs replaced consistently by short display aliases (C1/C2, E1/E2, and R1/R2). Annotation texts, scores, and model wording are retained. JSON indentation and line wrapping are typographic; omitted fields or dimensions are outside the explicitly named projection. Prose explains the controller operations separately from model-generated assessments. Minor spacing errors in raw responses are intentionally preserved.

F.1 Correct Positive: Recovery of MurD Stereospecific Evidence

Initial evidence and conflict. Both proteins initially retrieve EC 6.3.2.9, RHEA:16429, and GO:0008764, which describe D-glutamate ligation. Their catalytic-activity Top-1 instead describes L-glutamate ligation (RHEA:58816; EC 6.3.2.53). For P14900, this disagreement is recorded as a priority-3 conflict. The following projection gives its initial catalytic candidate and the conflicting relation returned by the Conflict Agent.

MurD: Initial evidence and conflict


```
{
  "id": "catalytic_activity:C1",
  "text": "The reaction: ATP + L-glutamate + UDP-N-acetyl-alpha-D-muramoyl-L-
    alanine = ADP + H(+) + phosphate + UDP-N-acetyl-alpha-D-muramoyl-L-alanyl-L
    -glutamate undergoes catalysis due to the presence of this protein.",
  "rank": 1,
  "score": 14.697656569138072
}
```


MurD: Initial evidence and conflict


```
{
  "a": "catalytic_activity:C1",
  "b": "ec_number:E1",
  "label": "contradict",
  "evidence": "Candidate a specifies the L-glutamate-addingligase reaction (EC
    6.3.2.53 / RHEA:58816), whichdirectly contradicts the D-glutamate
    stereospecificity defined by EC 6.3.2.9 in candidate b."
}
```


Targeted expansion. In round 2, the controller supplies `catalytic_activity` and `ec_number` as the eligible dimensions, with remaining budget 28. The Retrieval Agent returns the following complete output; the controller extends each frozen ranking from depth 1 to 2 and adds one unseen candidate per dimension.

MurD: Targeted expansion


```
{"dimensions":["catalytic_activity", "ec_number"]}
```

The newly ingested catalytic candidate is shown below. Its lower ProTrek score does not prevent it from receiving the first advisory position after cross-dimensional assessment. Expansion also adds the alternative EC 6.3.2.53 record; the original EC 6.3.2.9 remains in the pool.

MurD: Targeted expansion


```
{
  "id": "catalytic_activity:C2",
  "text": "The reaction: ATP + D-glutamate + UDP-N-acetyl-alpha-D-muramoyl-L-
    alanine = ADP + H(+) + phosphate + UDP-N-acetyl-alpha-D-muramoyl-L-alanyl-D
    -glutamate undergoes catalysis due to the presence of this protein.",
  "rank": 2,
  "score": 14.628944546568627
}
```


Advisory ordering and retained counterevidence. The round-2 Ranking response for `catalytic_activity` is reproduced in full below. Both candidate IDs remain present; the D-glutamate interpretation moves ahead of the L-glutamate alternative without deleting it.

MurD: Advisory ordering and retained counterevidence


```
{
  "ordered_ids": [
    "catalytic_activity:C2",
    "catalytic_activity:C1"
  ],
  "rationale": "RHEA:16429 (C2) is prioritized as it aligns with the primary EC
     6.3.2.9 and GO molecular function assignments. RHEA:58816 (C1) isretained
    as a secondary hypothesis, representing the alternative L-glutamate ligase
    activity (EC 6.3.2.53).",
  "uncertainty_notes": "Direct conflict between D-glutamate and L-glutamate
    stereospecificity;both are biologically plausible in different bacterial
    contexts."
}
```


The current relation changes from `contradict` to `compatible`, but the historical conflict is preserved and P14900 stops at `round_limit_unresolved`. Q7VAM1 stops after round 1 with `no_conflict`: its L/D disagreement was labeled compatible, so its catalytic Top-1 is never expanded. The original Q7VAM1 relation assessment is:

MurD: Advisory ordering and retained counterevidence


```
{
  "a": "catalytic_activity:C1",
  "b": "ec_number:E1",
  "label": "compatible",
  "evidence": "Bothdescribe UDP-N-acetylmuramoyl-L-alanine ligase activities
    with different glutamate stereospecificities (EC 6.3.2.53 with L-glutamate
    and EC 6.3.2.9 with D-glutamate), which can coexistas primary and alternate
    /cross-reactive enzyme activities."
}
```


Final pair output and interpretation. Reasoning receives both complete profiles and returns:

MurD: Final pair output and interpretation


```
{"similarity_level":"exact_match","similarity_score":0.95,"evidence":"Both
    proteins are supported as UDP-N-acetylmuramoyl-L-alanine:D-glutamate
    ligases (EC 6.3.2.9), with shared catalytic activity, RHEA reaction, and GO
     molecular function annotations confirming this specific peptidoglycan
    biosynthesis role."}
```

The positive decision agrees with both reference EC annotations. CAPA-RAG adds the missing D-glutamate reaction for P14900 and Ranking prioritizes it despite its lower original retrieval score, making the stereospecific interpretation explicit. Q7VAM1's D-glutamate interpretation is supported by its EC, Rhea, and GO fields, while its catalytic Top-1 remains the L-glutamate alternative. Thus the raw output's phrase "shared catalytic activity" should be read with this remaining disagreement in view; the observed recovery is localized to P14900. The historical record preserves that distinction instead of concealing the alternative hypothesis.

F.2 Correct Negative: Distinguishing Sodium and Proton Coupling

Initial specificity and conflicting hypotheses. P29707 initially retrieves EC 7.2.2.1 and a Na^+-coupled catalytic reaction, whereas Q95DR6 retrieves EC 7.1.2.2 and an H^+-coupled reaction. The two original catalytic texts are shown below (P29707 first, Q95DR6 second).

ATP synthase: Initial specificity and conflicting hypotheses

```
The reaction: ATP + H2O + 4 Na(+)(in) = ADP + H(+) + 4 Na(+)(out) + phosphate
    undergoes catalysis due to the presence of this protein.
```

ATP synthase: Initial specificity and conflicting hypotheses

```
The reaction: ATP + 4 H(+)(in) + H2O = ADP + 5 H(+)(out) + phosphate undergoes
    catalysis due to the presence of this protein.
```

P29707 also retrieves the H^+ reaction RHEA:57720 in its Rhea dimension, creating the following priority-3 contradiction with its EC candidate:

ATP synthase: Initial specificity and conflicting hypotheses

```
{
  "a": "ec_number:E2",
  "b": "rhea_reaction:R1",
  "label": "contradict",
  "evidence": "EC 7.2.2.1 specifies sodium ion translocation linked to ATP
    hydrolysis, whereas RHEA:57720 describes proton (H+) translocation (EC
    7.1.2.2)."
}
```

Expansion and ranking. With remaining budget 28, the controller supplies `ec_number` and `rhea_reaction`. The complete Retrieval response is:

ATP synthase: Expansion and ranking

```
{"dimensions": ["ec_number", "rhea_reaction"]}
```

The controller increases both depths from 1 to 2, adding EC 7.2.2.3 (a P-type Na^+ transporter alternative) and the correct Na^+ Rhea hypothesis. The latter candidate's original fields are:

ATP synthase: Expansion and ranking

```
{
  "id": "rhea_reaction:R2",
  "text": "RHEA:58156: 4 Na(+)(in) + ATP + H2O = 4 Na(+)(out) + ADP + phosphate
     + H(+)",
  "rank": 2,
  "score": 13.289299460018382
}
```

The final Rhea ordering nevertheless keeps the H^+ hypothesis first. This is the complete advice for that dimension:

ATP synthase: Expansion and ranking


```
{
  "ordered_ids": [
    "rhea_reaction:R1",
    "rhea_reaction:R2"
  ],
  "rationale": "RHEA:57720 (H+ transport) is prioritized because it is
    supported by the Function, GO BP, GO CC, and GO MF dimensions.RHEA:58156 (
    Na+ transport) is supported by Catalytic Activity and EC 7.2.2.1,
    representing a hard functional conflict regarding the transported ion.",
  "uncertainty_notes": "Hard functional conflict between proton and sodium
    translocation."
}
```


CAPA-RAG preserves both Rhea candidates together with the Na^+-specific EC and catalytic evidence, so downstream reasoning can assess ion specificity beyond the first advisory position. The graph later labels the original disagreement compatible, but the ledger retains it as unresolved and P29707 terminates at the round limit. Q95DR6 has no recorded conflict and stops after round 1 without expansion. This trace illustrates why full evidence retention is an important part of CAPA-RAG: a preferred ranking does not suppress a more specific competing hypothesis.

Final pair output and interpretation. The original Reasoning response is:

ATP synthase: Final pair output and interpretation


```
{"similarity_level":"unlikely_match","similarity_score":0.25,"evidence":"
    Protein A is supported as a sodium-transporting ATPase (EC 7.2.2.1),
    whereas Protein B is supportedas a proton-transporting ATPase (EC 7.1.2.2);
     these represent distinct ion-translocating specificities."}
```


The negative decision agrees with the reference distinction in ion-coupled complex activity. CAPA-RAG retains the initially retrieved Na^+-specific EC and catalytic evidence and supplements it with the matching Rhea reaction; the final decision identifies the transported ion as the relevant distinction despite the proteins' shared ATP synthase architecture. The concise output does not explicitly discuss the competing H^+ interpretation for P29707, which remains visible in the profile. The case demonstrates the value of comparing specific reaction claims across the complete evidence rather than treating broad ATPase similarity or the first-ranked candidate as sufficient.

F.3 Controller Outcomes and Scope of the Examples

Table F.2: **Retained evidence and controller outcomes.** Candidate counts cover all 14 dimensions. Both expanded profiles retain one historical conflict as `unresolved_at_limit`, even though its latest graph label is `compatible`.

Protein	Rounds	Candidates	Added entries	Stop reason
P14900	2	16	2	`round_limit_unresolved`
Q7VAM1	1	14	0	`no_conflict`
P29707	2	16	2	`round_limit_unresolved`
Q95DR6	1	14	0	`no_conflict`

Recorded retries. For P29707, the first Conflict response failed JSON validation and succeeded on the second attempt. The round-2 Ranking response was truncated twice and succeeded after increasing the output limit to 32,768 tokens. These retries address output formatting and length, not biological disagreement.

F.4 Proline Catabolism and Nitrogen-Source Utilization

This downstream case examines an experimentally positive substrate correctly predicted by GEM-Reasoner but missed by CarveMe in the *B. subtilis* 168 Biolog benchmark. It complements the protein-pair examples above by relating reconstructed pathway content to a utilization phenotype. The case is a retrospective analysis of one discordant prediction, not an additional estimate of aggregate benchmark performance.

F.4.1 Phenotype, Medium, and Growth Controls

Biolog classifies L-proline utilization as strongly positive (++). For the nitrogen-source test, ammonium uptake is disabled, glucose is retained in M9, and L-proline is the sole external nitrogen source, with a maximum uptake magnitude of $10\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. A maximum biomass objective > 0.1 defines a positive prediction, using the same threshold $\tau_{\mathrm{SU}} = 0.1\,\mathrm{h^{-1}}$ and strict inequality as the substrate-utilization benchmark (Appendix E.4).

Table F.3: **Biomass objectives and nitrogen-source controls.** Glucose is retained in all three conditions. Values are maximum biomass objectives of the two original reconstructions; each retains its own biomass formulation. TP and FN denote true positive and false negative against the proline Biolog label.

Nitrogen availability	GEM-Reasoner	CarveMe
Ammonium (complete M9)	0.804	0.251
L-proline only	1.133 (TP)	0.000 (FN)
Neither ammonium nor L-proline	0.000	0.000

Both models support growth with ammonium and predict no growth when both ammonium and proline are unavailable (Table F.3). The discordance therefore appears in the condition requiring proline-derived nitrogen. Objective magnitudes are model-specific predictions, not directly comparable physiological growth rates.

F.4.2 The Distinguishing Pathway Connection

Both reconstructions contain proline exchange and uptake, the P5C-to-glutamate reaction `P5CD`, and downstream glutamate-dependent nitrogen-assimilation reactions (Table F.4). The difference is the oxidative entry from intracellular L-proline to 1-pyrroline-5-carboxylate (P5C): `PRO1x`, associated with *putB* (`BSU03200`), is present only in GEM-Reasoner. Together with `P5CD`, it connects the supplied nutrient to glutamate and subsequent nitrogen-containing biosynthetic precursors. Although both models contain `P5CR`, its allowed direction reduces P5C to proline and cannot supply the missing oxidative entry.

Table F.4: **Reaction content along the proline-to-glutamate route.** Conversions summarize substrate connectivity and allowed direction, omitting cofactors and other stoichiometric species. Gene associations shown are those reported for the reconstructions; biochemical interpretation is discussed in Appendix F.4.5.

Reaction	Represented step / association	GEM-Reasoner	CarveMe
`EX_pro__L_e`	L-proline exchange	Present	Present
`PROt2r`	L-proline uptake	Present	Present
`PRO1x`	L-proline → P5C; *putB* (`BSU03200`)	Present	Absent
`P5CD`	P5C → L-glutamate; *rocA* (`BSU37780`)	Present	Present
`P5CR`	P5C → L-proline	Present	Present

F.4.3 FLUX USE AND CONDITIONAL PATHWAY NECESSITY

Parsimonious flux balance analysis (pFBA) at 99.9% of the maximum GEM-Reasoner biomass objective yields glucose and proline uptake magnitudes of 10.000 each. Both `PRO1x` and `P5CD` carry $9.818\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$, demonstrating use of the reconstructed catabolic route. The positive `GLNS` flux is consistent with glutamine synthesis supplying downstream nitrogen-containing precursors. External ammonium uptake remains disabled.

To distinguish a required route from one selected only by pFBA, flux variability analysis (FVA) independently minimizes each reaction flux subject to the same medium constraints and a biomass objective of at least 0.100001:

$$v_j^{\min} = \min_{\mathbf{v}} v_j \quad \text{s.t.} \quad \mathbf{Sv} = \mathbf{0}, \quad \mathbf{l} \le \mathbf{v} \le \mathbf{u}, \quad v_{\text{growth}} \ge 0.100001. \tag{F.1}$$

The lower bound enforces growth strictly above the case-study threshold. Table F.5 reports the available minima; these independent minima need not occur simultaneously in one flux distribution.

Table F.5: **Flux use and FVA lower bounds in GEM-Reasoner.** Flux units are $\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. Uptake magnitudes are positive; their corresponding exchange fluxes are negative. Other values follow the model reaction orientation. A dash means an FVA minimum is not reported, not zero.

Reaction or uptake	pFBA flux / magnitude	FVA minimum
Glucose uptake magnitude	10.000	—
L-proline uptake magnitude	10.000	—
`PROt2r`	10.000	0.801
`PRO1x`	9.818	0.785
`P5CD`	9.818	0.785
`GLNS`	1.847	0.163
`GLUDxi`	1.747	0.000

Strictly positive minima for `PROt2r`, `PRO1x`, `P5CD`, and `GLNS` establish their necessity within the GEM-Reasoner network under the tested growth constraint. By contrast, `GLUDxi` is active in the pFBA solution but can carry zero flux in other feasible solutions. Thus the proline-to-glutamate entry is required, while downstream nitrogen distribution retains flexibility. This is a statement about feasible model fluxes under the specified condition, not experimentally measured reaction or gene essentiality.

F.4.4 SENSITIVITY TO ENVIRONMENTAL AND NUMERICAL SETTINGS

The phenotype contrast persists across all reported sensitivity checks (Table F.6). Each check varies the indicated setting independently while retaining the original reaction content. GEM-Reasoner remains positive and CarveMe negative at biomass thresholds of 0.05, 0.10, and 0.20. Varying proline availability or fixing a common ATP-maintenance flux also preserves the contrast, including when maintenance is zero. Retaining the finite reaction bounds stored in the original SBML files reproduces the primary objectives. Independent SCIP and HiGHS calculations agree on the original-model objectives to within 10^{-6}.

Table F.6: **Sensitivity of the proline-utilization prediction.** Uptake limits and fixed ATP-maintenance fluxes are in $\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. Panel (a) changes only the classification threshold; panel (b) reports maximum biomass objectives under each setting. Positive predictions use objective $>$ threshold, with 0.1 used in panel (b).

(a) Classification threshold

Setting	Value	GEM-Reasoner	CarveMe
Biomass threshold	0.05	Positive	Negative
Biomass threshold	0.10	Positive	Negative
Biomass threshold	0.20	Positive	Negative

(b) Maximum biomass objectives

Setting	Value	GEM-Reasoner	CarveMe
Proline uptake limit	5	0.624	0.000
Proline uptake limit	10	1.133	0.000
Proline uptake limit	20	1.430	0.000
Fixed ATP maintenance	0	1.185	0.000
Fixed ATP maintenance	9	1.133	0.000
Finite flux bounds	Original SBML	1.133	0.000

F.4.5 BIOLOGICAL INTERPRETATION AND SCOPE

The recovered substrate connection is consistent with prior work linking proline utilization to the *putBCP* locus and its conversion of proline through P5C to glutamate, principally in *B. subtilis* JH642 and derivatives (Moses et al., 2012). In the reconstructions examined here, `PRO1x` is associated with *putB*, whereas `P5CD` is associated with *rocA*; the latter annotation supports the glutamate-forming downstream activity.[8] The represented route is therefore a PutB-associated entry connected to a RocA-associated reaction, rather than a demonstrated reconstruction of the complete PutBCP system.

The evidence concerns substrate-conversion connectivity. The supplied `PRO1x` uses NAD, whereas PutB is annotated as an FAD-dependent, quinone-linked enzyme.[9] The case therefore does not validate the reaction's cofactor assignment or establish quantitatively correct redox and energetic coupling. Nor does comparison of two complete reconstructions isolate the effect of `PRO1x` from all other network differences: its positive FVA minimum establishes conditional necessity in GEM-Reasoner, not sufficiency to restore growth in CarveMe. Together, agreement with the observed phenotype, the represented catabolic connection, and its positive flux requirement support improved functional coverage of proline catabolism as an explanation for this specific downstream prediction.

[8] UniProtKB RocA, `P39634`: https://www.uniprot.org/uniprotkb/P39634/entry.
[9] UniProtKB PutB, `P94390`: https://www.uniprot.org/uniprotkb/P94390/entry.

G PROMPTS

The English agent prompts, input constructors, and scoring rubric below are reproduced verbatim. CAPA inputs are serialized JSON; the reasoning template receives both complete evidence views. Placeholders are instantiated at runtime. The client constructs messages as follows:

Agent message construction

```
messages = [
    {"role": "system", "content": system},
    {
        "role": "user",
        "content": payload if isinstance(payload, str)
        else json.dumps(payload, ensure_ascii=False, sort_keys=True),
    },
]
```

G.1 CONFLICT AGENT

System message.

Conflict Agent: System message

```
You assess consistency of retrieved protein annotation hypotheses,
    independently of
any downstream task. Evaluate ONLY the listed candidate pairs, using their text
    , identifiers and
frozen relations. Their presence in a retrieval pool does not establish that
    either is true.

Use exactly these relation labels:
- support: a specific shared functional claim or explicit standardized relation
     provides positive
  support. Identify the actual supported claim; generic coexistence is not
    support.
- compatible: the claims could coexist in the SAME protein/context, but this
    supplies NO positive
  evidence that either annotation applies to this protein.
- hierarchical: an explicit general/specific relationship; it does not
    establish the more specific claim.
- context_dependent: the relationship depends on a condition, subunit role,
    direction or context not established here.
- contradict: the same-context claims are mutually incompatible. This
    identifies a problem to
  investigate; it does NOT establish which candidate is wrong. All candidates
    remain evidence.
- unknown: missing information or no established relationship. Unrelated
    pathway/process claims
  should not be promoted to positive evidence merely because both could occur
    in an organism.

Distinguish the activity of a whole complex from the role of one subunit. Do
    not transfer a complex's
complete catalytic function to a regulatory subunit solely because it belongs
    to that complex.
Different IDs, missing information, compatible multifunctionality or different
    source organisms
alone are not contradictions. Reference residue positions are not verified
    target-sequence coordinates.
Rhea master IDs identify a reaction family; retained direction information must
     still be respected.
```

Conflict Agent: System message (continued)

```
Do not compare ProTrek scores across dimensions or infer confidence from their
    magnitude.
Never invent annotations or identifiers. Return ONLY
{"relations":[{"a":"candidate ID","b":"candidate ID","label":"label","evidence
    ":"specific basis"}]},
exactly one record per requested pair.
```

User-message constructor.

Conflict Agent: User-message constructor

```
{"pairs": [{"a": a, "b": b} for _, a, b in batch]}
```

Here `a` and `b` are full candidate objects. Pairs not decided by frozen relations are submitted in batches of at most 32.

G.2 Retrieval Agent

System message.

Retrieval Agent: System message

```
You are the CAPA Retrieval Agent. The controller supplies dimensions implicated
     by
the highest-priority OPEN conflict that still has retrieval capacity. Sorting
    alternatives does
not resolve a contradiction. All candidates and historical conflicts remain
    available.
Return exactly eligible_dimensions, once each, as {"dimensions":["dimension_key
    ", ...]}.
The controller expands those same frozen sequence-to-text rankings within the
    configured step,
depth, round and budget limits. Do not rewrite the query, add dimensions, fetch
     annotations by ID,
search an external source. All retrieved candidates, including conflicting
    hypotheses, remain
available for integration and downstream reasoning. Ranking advice does not
    resolve a conflict.
```

User-message constructor.

Retrieval Agent: User-message constructor

```
{
    "eligible_dimensions": eligible,
    "conflicts": [
        edge
        for edge in conflicts
        if edge["priority"] == priority
    ],
    "remaining_budget": self.settings.expansion_budget
    - state["expansion_used"],
}
```

The controller combines eligible dimensions from all expandable open conflicts tied at the highest priority. The returned set must equal `eligible_dimensions`.

G.3 RANKING AGENT

System message.

Ranking Agent: System message

```
You organize retrieved protein annotation hypotheses without a downstream
    target.
Provide advisory ordering for ALL 14 dimensions. Every supplied candidate ID in
     each dimension
must appear exactly once in ordered_ids, including unsupported and
    contradictory hypotheses.
Use [] ONLY when that dimension truly has no candidates. Never select, reject,
    hide, edit or
invent an annotation. Your advice cannot change which evidence downstream
    reasoning receives.

Use specific functional semantics, mechanisms and traceable identifier
    relations to explain
the suggested order. An identifier match is one piece of evidence; consider
    conflicting text,
substrates, transported species, reaction direction and protein roles. Treat
    ProTrek ranks and
scores only as WITHIN-dimension retrieval signals, never as probabilities or
    cross-dimension weights.
If evidence cannot distinguish alternatives, preserve retrieval order and
    explain that limitation.
First rank, especially an isolated Top-1, does not establish that a hypothesis
    is true.

Distinguish support from mere compatibility. Do not count duplicate records,
    paraphrases or
cross-references derived from one reaction as independent votes. Keep whole-
    complex catalytic
activity, subunit-specific role and complex membership distinct. Reference
    residue coordinates
have not been mapped to this query sequence. Mark that limitation rather than
    asserting target sites.
Preserve and explain incompatible alternatives, including hard conflicts. A
    supported preferred
interpretation does not erase contradictory evidence or close the conflict
    ledger.

Return ONLY {"advice":{"dimension_key":{
"ordered_ids":["every supplied candidate ID exactly once"],
"rationale":"nonempty explanation of advisory ordering grounded in supplied
    evidence",
"uncertainty_notes":"remaining conflicts or limitations; empty string if none
    "}, ...}}.
Your explanation is an advisory assessment, separate from the immutable
    annotation texts.
Do not invent EC/Rhea identifiers, repair missing candidates, or infer a
    downstream pair label.
```

User-message constructor.

Ranking Agent: User-message constructor

```
{
    "dimensions": state["dimensions"],
    "relations": state["graph"],
    "current_round": state["rounds_completed"] + 1,
    "total_rounds": self.settings.total_rounds,
```

Ranking Agent: User-message constructor (continued)

```
    "remaining_expansion_budget": self.settings.expansion_budget
    - state["expansion_used"],
    "pending_conflicts": [
        {key: entry[key] for key in ("id", "edge", "status")}
        for entry in state["pending_conflicts"].values()
    ],
}
```

Validation requires all 14 dimensions and a complete candidate-ID permutation in each `ordered_ids`. Historical conflict inputs retain their original edge and status; previous advice may also appear in the dimension state.

G.4 Functional Similarity Reasoning

PFS pair classification and reference-protein scoring share this template. The enabled task is EC number. The placeholders `{task.display_name}` and `{task.match_criterion}` become `EC number` and "the same complete four-level Enzyme Commission classification." Protein aliases are A and B. Each annotation placeholder contains the complete JSON evidence view, including all 14 dimensions and available relations, historical conflicts, and ranking advice.

System message.

Functional Similarity Reasoning: System message

```
You assess whether two proteins share at least one specifically supported
enzyme activity at complete four-level EC granularity.

Use all 14 supplied evidence dimensions with an explicit priority:
reaction chemistry and specificity first; functional interpretation second;
mechanistic validation third; biological context fourth.

This priority concerns evidential relevance, not guaranteed correctness.
Every retrieved annotation is a hypothesis about the target protein.
CAPA explanations, rankings and relation labels are advisory interpretations,
not additional observations.

Preserve alternative and multifunctional interpretations. Distinguish
specific functional equivalence from broad family or complex similarity.
Calibrate the conclusion to the strongest relevant evidence and its
unresolved counterevidence.

Return exactly one JSON object containing similarity_level,
similarity_score, and evidence. Return no additional text.
```

User-message template. The rubric placeholder is replaced by the exact text in Appendix G.5.

Functional Similarity Reasoning: User-message template

```
Target benchmark task: {task.display_name}

MATCH CRITERION
Determine whether at least one specific enzyme activity is supported for
both proteins at the specificity represented by {task.match_criterion}.

The proteins need not have identical complete activity sets. A multifunctional
protein may match another protein through one shared activity, even when their
other activities, dominant roles, pathways or locations differ.

A shared predicted EC identifier or a shared candidate in the retrieval pools
```

Functional Similarity Reasoning: User-message template (continued)

```
does not by itself establish a match. Compare the functional claims and their
applicability to each protein. Distinguish positively asserted activities from
EC identifiers mentioned only for comparison, neighboring pathway reactions,
or hypothetical multifunctionality.

EVIDENCE PRIORITY

1. Primary evidence: catalytic_activity, rhea_reaction, ec_number.
Use the specific reaction chemistry as the main basis of comparison:
substrates, products, bond changes, stereochemistry, electron donors and
acceptors, transported species, coupling mechanism and relevant directionality.
Interpret the reaction meaning of an EC annotation, not just its identifier.

Specific differences at this level carry more weight than generic similarity
in protein family, ATP binding, pathway or localization. A shared broad
description must not erase a difference in the actual reaction.

Account for reaction reversibility and assay direction where the supplied
evidence supports them; opposite written directions alone need not indicate
different enzyme functions.

2. Functional interpretation: function, go_molecular_function.
Use these dimensions to interpret reaction specificity, explicit substrate
preferences or exclusions, multiple activities, and catalytic versus accessory
roles.

A detailed functional statement can outweigh an incorrect reaction or EC
prediction when it provides a specific reason supported by the supplied
evidence. A generic family description or an isolated role annotation cannot
automatically override specific catalytic evidence.

Preserve all supported activities of multifunctional proteins. Do not choose
one activity as the protein's sole identity and discard another shared activity
    .

Distinguish whole-complex activity from an individual subunit's contribution.
Shared complex membership alone does not establish a shared catalytic function.
Different subunit roles alone do not settle whether the supplied annotations
support a shared EC-level activity; compare claims at the same biological scope
    .

3. Mechanistic validation: cofactor, active_site, binding_site, domain.
Use specific cofactor requirements, ligand specificity, catalytic features and
domain functions to support or challenge the primary functional interpretations
    .

These dimensions can be decisive when they distinguish relevant mechanisms.
Broad ATP binding, a common fold or a common metal-binding property is weak
evidence for complete-EC equivalence.

Reference residue coordinates have not been aligned to the query proteins.
Do not treat them as verified target residues or infer target mutations,
missing catalytic sites or catalytic inactivity from them.

4. Biological context: unipathway, reactome_pathway,
go_biological_process, subcellular_location, go_cellular_component.
Use these dimensions to assess context, biological plausibility and the scope
of an annotation. Consider concrete context-dependent restrictions when they
affect the proposed enzyme activity.

Shared context cannot establish a match without specific functional evidence.
Different pathways, organisms, locations or physiological roles do not by
```

Functional Similarity Reasoning: User-message template (continued)

```
themselves establish different enzyme functions.

APPLY THE PRIORITY WITHOUT VOTING

Read all 14 dimensions, but do not assign equal votes or count agreeing
annotations. Evaluate specificity, applicability and non-redundancy.

A reaction, its EC cross-reference, its GO term and a paraphrase may describe
the same underlying claim; they are not independent confirmations.
CAPA repeating that claim supplies no new experimental support.

A lower-priority dimension may change the conclusion when it supplies a
specific relevant distinction or exposes an applicability problem in the
higher-priority evidence. Identify that reason rather than relying on the
number, length or confidence of supporting descriptions.

Missing evidence is not evidence of absence. ProTrek ranks and scores order
retrieval within one protein and dimension; they are not probabilities and
must not be compared across dimensions.

Protein A ({protein_a}) CAPA 14-dimensional evidence:
{annotation_a}

Protein B ({protein_b}) CAPA 14-dimensional evidence:
{annotation_b}

CONFLICT AND CAPA CHECKS

Compare the original candidate texts before adopting CAPA's preferred
interpretation. First rank, a support relation, or a fluent rationale does not
establish that an annotation applies to the target protein.

Check relevant disagreements across dimensions even when they are absent
from relation_graph or pending_conflicts. Empty conflict collections and
complete_for_configured_pairs do not establish global consistency.
In retrieval_only mode, CAPA did not evaluate relations or ranking.

Compatible means that claims may coexist; it does not confirm either claim,
establish equivalent activities, or resolve substrate or ion specificity.
Unknown and context_dependent also do not provide positive confirmation.
A later change from contradict to compatible does not by itself resolve
the original functional disagreement.

For each potentially decisive conflict, distinguish:
- competing activity hypotheses for one protein;
- multiple activities actually supported for that protein;
- descriptions referring to different subunits, complexes or contexts.

Do not invent multifunctionality to reconcile incompatible hypotheses.
Conversely, do not discard an explicitly supported shared activity merely
because another activity differs.

A conflict is material when resolving it differently could change the
pairwise match decision. If the supplied evidence clearly favors one
interpretation under the priority above, make the corresponding decision.
Do not let unrelated or clearly weaker retrieved noise force uncertainty.

If credible competing interpretations remain and would produce different
match decisions, retain uncertainty. Do not turn unresolved alternatives
into a confident positive or negative conclusion.

DECISION AND SCORE
```

Functional Similarity Reasoning: User-message template (continued)

```
Assess whether a specific shared activity is supported for both proteins,
then check the strongest relevant evidence against that conclusion.

{SCORING_RUBRIC}

OUTPUT
The evidence must be one concise sentence of at most 40 words:
- For a match, name the specific shared activity or reaction.
- For a non-match, name the decisive functional distinction.
- For uncertainty, name the unresolved alternative or missing specificity.
- If a material conflicting interpretation is outweighed, briefly state
  the decisive basis; do not silently omit it.

Use only the supplied evidence. Do not invent protein identities, annotations,
identifiers, target residues or experimental confirmation. Return the concise
decision basis, not a reasoning transcript.

Return only:
{"similarity_level":"<exact_match|likely_match|uncertain|unlikely_match|
    no_match>","similarity_score":<number from 0 to 1>,"evidence":"<concise
    evidence-based decision>"}
```

G.5 SCORING RUBRIC

The following is inserted at `{SCORING_RUBRIC}` in the user template.

Scoring rubric

```
Use the following scale:
- exact_match, 0.90 <= score <= 1.00:
  A specific shared EC-level activity is strongly supported for both proteins,
  with no material unresolved counterevidence. Differences outside that shared
  activity do not prevent a match.
- likely_match, 0.65 <= score < 0.90:
  A specific shared activity is the better-supported interpretation, but some
  uncertainty about specificity or applicability remains.
- uncertain, 0.35 < score < 0.65:
  Insufficient specific evidence or unresolved competing interpretations
  prevents a reliable match decision.
- unlikely_match, 0.10 < score <= 0.35:
  The better-supported interpretations indicate different enzyme activities,
  although broad relatedness or a weaker shared-activity hypothesis remains.
- no_match, 0.00 <= score <= 0.10:
  Specific evidence strongly supports different enzyme activities, and no
  credible shared activity remains under the supplied evidence.

Choose the semantic level first, then an in-band score. Do not increase
confidence because CAPA processed the profile or because several annotations
repeat the same claim. Missing EC alone does not require uncertainty.
```