PROCYON: A multimodal foundation model for protein phenotypes

Owen Queen[1,*,+], Yepeng Huang[1,*], Robert Calef[1,2,*], Valentina Giunchiglia[1,3,4], Tianlong Chen[1,2,◇], George Dasoulas[1], LeAnn Tai[2], Yasha Ektefaie[1], Ayush Noori[1], Joseph Brown[5,†], Tom Cobley[2,6], Karin Hrovatin[7,8], Tom Hartvigsen[9], Fabian J. Theis[7,10], Bradley Pentelute[5,14], Vikram Khurana[11,12,14], Manolis Kellis[2,14], Marinka Zitnik[1,13,14,15,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[3]Department of Brain Sciences, Imperial College London, London, UK
[4]Centre for Neuroimaging Sciences, King's College London, London, UK
[5]Department of Chemistry, MIT, Cambridge, MA, USA
[6]Department of Computing, Imperial College London, London, UK
[7]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Munich, Germany
[8]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
[9]School of Data Science, University of Virginia, VA, USA
[10]School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[11]Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA
[12]Harvard Stem Cell Institute, Cambridge, MA, USA
[13]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA
[14]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[15]Harvard Data Science Initiative, Cambridge, MA, USA
[*]Co-first authors; [+]Present address: Department of Computer Science, Stanford University, Stanford, CA, USA; [◇]Present address: Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [†]Present address: Acceleration Consortium, University of Toronto, Toronto, ON, Canada; [‡]Corresponding author: marinka@hms.harvard.edu

Understanding the roles of human proteins remains a major challenge, with approximately 20% of human proteins lacking known functions and more than 40% missing context-specific functional insights. Even well-annotated proteins are often poorly characterized in diverse biological contexts, disease states, and perturbations. We present PROCYON, a foundation model for modeling, generating, and predicting protein phenotypes across five interrelated knowledge domains: molecular functions, therapeutic mechanisms, disease associations, functional protein domains, and molecular interactions. To support this, we created PROCYON-INSTRUCT, a dataset of 33 million protein phenotype instructions, representing a comprehensive resource for multiscale protein phenotypes. By co-training a large language model with multimodal molecular encoders, PROCYON integrates phenotypic and protein data. A novel architecture and instruction tuning strategy allow PROCYON to process arbitrarily interleaved protein-and-phenotype inputs, achieve zero-shot task transfer, and generate free-form text phenotypes interleaved with retrieved protein sequence, structure, and drug modalities in a single unified model. PROCYON achieves strong performance against single-modality models, multimodal models such as ESM3, as well as text-only LLMs on dozens of benchmarking tasks such as contextual protein retrieval and question answering. We extensively evaluate PROCYON for biological applications, including identifying protein domains that bind small molecule drugs, predicting peptide binding with enzymes, and assessing the functional impact of Alzheimer's disease mutations. PROCYON enables conditional retrieval of proteins linked to small molecules through complementary mechanisms of action. It generates candidate phenotypes for under-characterized proteins recently implicated in Parkinson's disease, facilitating hypothesis generation for poorly understood proteins and biological processes. PROCYON paves the way toward an effective, general solution for functional protein biology that can enable new insights into the human proteome.

Main

Decoding molecular phenotypes remains a fundamental challenge in biology, essential to understanding how living systems operate at the molecular level. Approximately 20% of human proteins lack well-defined functions [1, 2], and even well-annotated proteins often miss functional insights across various biological contexts and disease states [3]. Evidence from structural and functional annotation studies reveals that approximately 40% of human proteins are missing context-specific functional insights [1, 3]. This knowledge gap is further exacerbated by a stark bias in protein research: 95% of life science publications focus on only 5,000 well-studied proteins [4], leaving a vast portion of the human proteome unexplored.

Protein phenotypes—the observable characteristics resulting from a protein's function—are multiscale and encompass molecular functions, therapeutic interactions, cellular pathways, organismal traits, and diseases [5–7]. These phenotypes are context-dependent and can vary under different biological conditions, making their characterization complex [8–10]. Although valuable [11, 12], current approaches to predicting protein function often do not capture the full spectrum of protein phenotypes or predict functions that are too vague to provide meaningful phenotypic information (e.g., ATP binding, protein binding, nucleic acid binding) [13]. Addressing this challenge is a central focus of initiatives such as the Human Proteome Project Grand Challenge, "A Function for Each Protein" [13, 14], which aims to determine phenotypes of every protein—both individually and within networks and pathways—in the context of human health and disease.

While advances in computational biology, catalyzed by AlphaFold3 [15] and ESMFold [16], have enabled the prediction of atomic protein structure from sequence, predicting protein phenotypes remains an unsolved problem. Protein language models (PLMs) [17, 18] and representation learning methods [19–21] capture function through learned embeddings but lack the explicit modeling required to link proteins to diverse and multiscale phenotypes. These limitations are compounded by many models focusing on pre-defined functional categories, which restricts their ability to generalize to new phenotypes or uncharacterized proteins.

Multimodal protein-text models [22–24] attempt to integrate proteins and phenotype data to expand the scope of protein phenotype prediction. However, they often fall short of exhaustively covering phenotypes, either by focusing only on molecular functions in training or by limiting to contrastive learning setups with limited textual flexibility. This becomes evident for poorly characterized proteins, where the absence of functional annotations and limited structural data leave current models with insufficient context to make predictions. Furthermore, while capable of

generating human-readable output, text-only large language models (LLMs) [25] face limitations when applied to protein phenotype prediction. Text-only models depend on structured naming systems, such as UniProt IDs [26] or HGNC names [27], and cannot generalize to proteins without explicit naming references. They are heavily based on text-based representations, making them vulnerable to bias [28] and inconsistencies in the scientific literature [29]. LLMs also struggle with proteins that lack established nomenclature or text descriptions [30].

Protein phenotypes are multiscale [6, 31–33] and span knowledge domains: molecular function, therapeutic mechanisms, disease associations, functional protein domains, and molecular interactions. Sequence and structure models alone cannot capture all these phenotypes. While sequence information captures evolutionary relationships and conserved regions, it can fail to account for convergent evolution, where unrelated sequences adopt similar phenotypes or vice versa [34, 35]. For example, enzymes with different sequences but similar catalytic roles (e.g., serine proteases) cannot be fully characterized using sequence information alone. Similarly, phenotypes often emerge from interactions between compatible protein structures [31, 36], as seen in molecular machines like the ribosome or spliceosome, where individual protein structures do not provide a complete functional context.

Here, we introduce PROCYON, a multimodal foundation model for modeling, generating, and predicting protein phenotypes across multiscale knowledge domains. PROCYON models phenotypes by integrating proteins with interleaved phenotype descriptions, unifying these domains through its novel instruction tuning dataset (PROCYON-INSTRUCT) containing over 33 million human protein phenotypes spanning molecular to organismal scales. PROCYON's architecture specifies co-training a large language model with dedicated encoders for protein sequences, structures, and drugs. This integration enables the model to process arbitrarily interleaved inputs of text, protein sequences, protein domains, peptide data, and small molecule structures. PROCYON uses instruction tuning on PROCYON-INSTRUCT to learn to generate free-form text interleaved with retrieved protein-related entities, creating a unified latent space across diverse molecular modalities. The model incorporates advanced capabilities, such as interleaved phenotype-context modeling, multimodal fusion, and autoregressive generation, enabling precise phenotype predictions and contextual protein retrieval. By combining multimodal data with an open-end natural language interface, PROCYON supports flexible task adaptation and achieves zero-shot generalization to new tasks and phenotypes. This advance addresses limitations in existing approaches that rely on implicit representations, controlled vocabularies of phenotypes, or static task-specific architectures,

offering unprecedented flexibility in exploring and characterizing protein phenotypes.

One of PROCYON's key advances is its ability to perform zero-shot task transfer, enabling generalization not only to new phenotypes and phenotype descriptions but also to entirely new kinds of tasks (not only new phenotypic label sets). For example, PROCYON can prioritize protein domains that are likely to bind specific drugs or generate phenotype descriptions for poorly characterized proteins. This capability empowers researchers to explore and formulate hypotheses for underexplored proteins. To address the complexity of multiscale phenotypes and pleiotropic proteins, PROCYON models the many-to-many relationships between proteins and phenotypes—where each protein can be linked to many semantically distinct phenotypes [7, 10]. This algorithmic challenge is critical for molecular phenotypes but is not addressed by existing multimodal foundation models, which predominantly focus on language-vision modalities [37,38] and assume one-to-one relationships between views of the same entity in different modalities. By capturing these context-specific protein roles, PROCYON provides a general model of complex protein phenotypes.

PROCYON achieves substantial performance gains in contextual protein retrieval and question answering when compared to single-modality and multimodal models, including state-of-the-art protein representation and protein-text learning models. In benchmarking across 14 tasks spanning multiple knowledge domains, PROCYON consistently outperforms single-modality models in 10 tasks, demonstrating its ability to harmonize multimodal sequence-structure representations. PROCYON achieves an average $F_{\max}$ of 0.743 in protein retrieval (compared to 0.618 for the next best model) and a question-answering accuracy of 72.7% (vs. 67.8% for the best-performing other model). PROCYON improves upon multimodal protein-text models in 13 of 14 retrieval tasks with an average $F_{\max}$ improvement of 0.301, and similarly outperforms other models in 13 of 14 zero-shot retrieval tasks with an $F_{\max}$ improvement of 0.270 on average. PROCYON excels in contextual protein retrieval via natural language prompts that retrieve proteins conditioned on diverse contexts, such as disease phenotypes and therapeutic mechanisms. PROCYON allows for compositional prompts describing complex, pleiotropic phenotypes; for example, in the case of pathway crosstalk phenotype, it outperforms manual aggregation methods by 3.97%. Furthermore, it generates detailed and accurate phenotype descriptions for poorly characterized proteins, validated against text-based LLMs and experimental data, including phenotypes absent from its training set. These results highlight PROCYON's ability to integrate multimodal data flexibly, enabling robust performance across predictive and generative tasks.

Beyond model benchmarking, we applied PROCYON to predict protein phenotypes in numer-

ous use cases, including identifying drug-binding domains for MGAM, predicting peptide binders for the ACE2 enzyme, retrieving multifunctional proteins associated with pleiotropic phenotypes, and evaluating the functional impact of Alzheimer's disease protein variants. PROCYON also supports conditional retrieval of proteins for small molecules with complementary mechanisms of action and generates candidate phenotypes for under-characterized genes linked to Parkinson's disease. In zero-shot task transfer, PROCYON accurately identifies drug-binding protein domains, ranking the correct domains at the top in 68.5% of test cases without explicit task training. It adapts to protein-peptide binding tasks, achieving statistically significant differentiation between binders and non-binders, even in challenging retrieval scenarios. For modeling phenotypes of mutated proteins, PROCYON distinguishes between benign and pathogenic variants, generating descriptions aligned with known disease mechanisms in Alzheimer's. It also bridges drug and disease knowledge domains, retrieving proteins targeted by small molecules in a disease-specific manner, effectively prioritizing therapeutic candidates based on context. PROCYON also elucidates functions of poorly characterized proteins, as demonstrated through proteins implicated in Parkinson's.

Our approach paves the way toward a general and effective multimodal foundation model for molecular phenotypes. Our training datasets, training and inference code, and pretrained models are publicly available to encourage future work and exploration.

Results

PROCYON multimodal model for protein retrieval and phenotype generation

PROCYON is a multimodal foundation model that bridges the spaces of proteins and phenotypes using natural language. PROCYON interleaves multiple modalities in its input—including protein sequence, protein structure, and natural language—while generating phenotype descriptions and retrieving proteins. This multimodal representation allows PROCYON to perform a wide range of tasks previously achievable only by specialized models. PROCYON is not constrained by controlled vocabularies or pre-defined phenotype terminologies, such as those traditionally used for gene annotation, enabling it to generate open-ended, free-text phenotypes beyond what is seen during training. PROCYON uses natural language to enable **multi-knowledge domain inference** that connects phenotypes across scales, from molecular characteristics to organism-wide traits such as diseases and therapeutic interactions.

PROCYON has three core capabilities: retrieval, question answering, and free-text phenotype generation. In retrieval, PROCYON transforms a given free-text phenotype description into a ranked list of proteins that match the user's prompt (Figure 1a). In question answering (QA), PROCYON determines whether a given protein has the phenotype specified in the input prompt and can do so based on a multimodal input of protein sequence, structure or both (Figure 1b). Finally, in phenotype generation, PROCYON generates free-text descriptions of protein phenotypes conditioned on user-defined context. For example, given a protein and information about a drug, PROCYON can describe the mechanism of action of the drug with the protein (Figure 1c). PROCYON unifies these capabilities into a single model (Extended Data Figure 1).

PROCYON is an 11-billion parameter unified protein-text model. Its core is based on the transformer architecture of the Llama-3 language model [39]. PROCYON integrates advanced encoders for protein sequence (ESM models [16]) and protein structure (GearNet [21]) and supports multimodal interleaved input within textual prompts. PROCYON (Figure 1d) comprises a **Protein Retrieval module**, where the model receives an input prompt and a protein library, which may include proteins, domains, peptides, or other amino acid sequences. Users can provide custom sequences or use PROCYON's default of 18,174 UniProtKB/SwissProt human proteins. The protein library is processed by a protein language model and projected into a unified latent space via a retrieval-specific projector. The prompt is also processed by a language model and projected into the unified latent space by the query projector. PROCYON then searches against cached protein sequence embeddings, retrieving a ranked list of proteins with scores. In the **Phenotype Genera-**

tion module, PROCYON processes interleaved multimodal inputs to generate output descriptions. A phenotype prompt is split into multimodal components, where text is tokenized using a standard text tokenizer, and protein sequences and structures are encoded into token embeddings. These multimodal tokens form a unified prompt, processed by the same LLM used in the Protein Retrieval module, with shared weights. PROCYON then generates phenotypes through standard autoregressive sampling. Providing varying context in interleaved multimodal prompts enables PROCYON to model shifts in protein phenotypes that occur under different biological conditions, while generating many phenotype descriptions from a single prompt supports cases where proteins have multiple phenotypes within the same context (i.e., modeling many-to-many protein-phenotype maps, as described above).

PROCYON-INSTRUCT dataset with 33 million protein-phenotype instructions

To enable PROCYON to capture the context-dependent roles that proteins play across cellular environments, biological pathways, and disease states, we developed PROCYON-INSTRUCT, a multi-knowledge domain dataset of 677,154 protein-phenotype pairs (372,439 protein functions, 118,702 protein domains, 81,645 disease phenotypes, 57,106 interaction phenotypes, and 47,262 therapeutic phenotypes) between 48,920 unique phenotypes and 56,753 proteins, domains, and peptides (1). PROCYON-INSTRUCT captures protein phenotypes from 11 data sources and five knowledge domains: **function**, **disease**, **therapeutics**, **protein domains**, and **interaction** (Figure 2c; Supplementary Tables 1 and 2). We also use a data augmentation strategy to produce the final dataset PROCYON-INSTRUCT that contains 33,899,528 total protein phenotype instructions used to train PROCYON (Supplementary Table 3).

In the **function** knowledge domain, we include databases that capture molecular and cellular phenotypes of proteins, including pathways, subcellular localizations, and molecular functions. The **disease** knowledge domain includes protein-disease associations—based on genetic studies, clinical annotations, manually curated literature review, and expert evaluations. **Therapeutics** encompasses information about drug-protein associations, including drug-target data as well as other protein interactions with drugs, including carriers, transporters, and enzymes. In the **protein domains** knowledge domain, we include data about protein domains, i.e., evolutionarily-derived sub-units of proteins that are conserved across many proteins. The **interaction** knowledge domain includes data expressing biomolecular interactions, including protein-protein interactions and protein-peptide interactions, from STRING [40] and PDB [41]; such interactions are inferred

from a spectrum of data, including coexpression information, homology, experimental interaction assays, and physical binding data.

We transform protein-phenotype pairs into a multimodal instruction tuning dataset across these five knowledge domains, expressing protein phenotypes gathered from each database as a natural language instruction that emulates the instructions given by human scientists [42–46] (Figure 2a). Many of the databases used to create PROCYON-INSTRUCT contain standardized language that is templated and not representative of the diversity of expertise for users that might interact with PROCYON. In addition, this standardized language harms the diversity of text available for training the model, thus hindering the generalization of the model to out-of-distribution input texts. Therefore, we design a system that leverages an external LLM (OpenAI's GPT models [25]) to rephrase the descriptions in PROCYON-INSTRUCT (Figure 2b). We provide three critical components to the LLM—task definitions, example descriptions, and phenotype descriptions—and prompt the model to generate rephrased versions of these components (all described in Methods Sec. 3.1). Additionally, we employ input prompts that instruct the LLM to generate these rephrasings with two axes of variation: 1) rephrasing style, which includes direct rephrasing, summarization, or simplification; and 2) expertise level, which expresses the specificity and granularity expected by the user. Our expertise levels emulate junior scientists, mid-level scientists, and senior scientists, defined in Methods Sec. 1.3. This rephrasing strategy increases the number of text descriptions from 75,290 to 380,438, and PROCYON-INSTRUCT grows from 1,855,548 instructions to 33,899,528 instructions.

Multi-knowledge domain instruction-tuning enables zero-shot task transfer

PROCYON is trained on PROCYON-INSTRUCT instruction inputs, with retrieval using a contrastive learning mechanism (Methods Sec. 3.4), and both QA and phenotype generation using a causal language modeling loss (Methods Sec. 3.4). After training, the PROCYON model has three core capabilities: protein retrieval, question answering, and phenotype generation.

In molecular biology, the complexity and uniqueness of most problems make it essential to move beyond predefined task sets. Researchers of varying expertise often approach problems differently, requiring a model capable of accommodating diverse user levels and adapting to compositional queries that span multiple knowledge domains. Addressing this critical need, PROCYON introduces the ability to define tasks through natural language prompts, enabling the model to generalize to entirely new tasks at inference time. This innovation allows researchers

to dynamically specify tasks in intuitive ways, tailoring predictions and analyses to their specific challenges. Through this open-ended task specification, PROCYON empowers users to explore complex, context-dependent protein phenotypes. Since PROCYON is instructed on tasks in five knowledge domains during training, it can perform new tasks based on changing the task description provided to the model, even if the task was not seen during training. This **zero-shot task transfer** is a new capability enabled by encoding tasks and descriptions in natural language instead of fixed deep learning architectures [12]. This formulation is critical for protein phenotype modeling since the space of potential queries about protein phenotypes is prohibitively large (Extended Data Figure 1), making it impossible to *a priori* define all tasks a general tool could encounter.

A major challenge in modeling protein phenotypes is addressing pleiotropy, the phenomenon where a single protein is involved in multiple distinct phenotypes. Pleiotropy manifests in diverse ways, such as proteins participating in multiple molecular functions, influencing various diseases, or being implicated in several biological pathways. For instance, over 50% of proteins localize to multiple cellular compartments, indicating functional diversity across pathways [47], and 12% of disease-causing proteins are implicated in multiple diseases [48]. This complexity breaks the assumption of a one-to-one correspondence between proteins and phenotypes. Within PROCYON-INSTRUCT, this complexity is evident, with proteins linked to a median of 15 phenotypes and some databases showing a mean of 21 phenotypes per protein (Gene Ontology; Supplementary Figures 1 and 2). Addressing this one-to-many protein-to-phenotype mapping is critical but remains unaddressed in existing biological foundation models. These approaches typically assume a single output y for a given input x, thereby failing to capture the diversity of phenotypes a protein may exhibit or the multiscale nature of phenotypes.

In contrast, PROCYON addresses this challenge of complex phenotypes through a multifaceted approach. By training on multiple knowledge domains, PROCYON achieves multiscale phenotypic resolution, enabling precise querying across molecular functions, diseases, pathways, and therapeutics. To further enhance its accuracy, a novel context augmentation (CA) strategy integrates additional information about phenotypes into the model, significantly improving predictive performance (Methods Sec. 3.2). Furthermore, PROCYON supports compositional queries through natural language prompts, allowing users to dynamically define phenotypes by combining multiple traits into a single, precise query – offering flexibility that stands in stark contrast to the static vocabularies employed by previous models (Subheading "Prompting PROCYON with complex and pleiotropic protein phenotypes"). Finally, PROCYON's inference mechanisms handle one-to-many

relationships through contextual protein retrieval and conditional generation of phenotypes that capture diverse roles that proteins play in different contexts (Methods Sec. 2.3).

Benchmarking PROCYON against single-modality models

We construct fourteen benchmarks from PROCYON-INSTRUCT. Our benchmarking suite consists of seven tasks, framed as both question answering and protein retrieval problems, resulting in fourteen distinct types of tasks. These tasks span 75,310 protein-phenotype queries over 9,403 unique phenotypes across three knowledge domains (functions, diseases, and therapeutics).

To benchmark PROCYON's performance on phenotypes unseen during model training, we design training and evaluation splits by leveraging the hierarchical or networked structure of each database. For instance, in Gene Ontology, we hold out entire sub-trees associated with high-level terms, ensuring that related terms (e.g., a term and its regulation or parallel terms) are not split between the training and evaluation sets (Figure 2d). This approach minimizes information leakage and avoids overly proximal relationships between training and test phenotypes. Similarly, we use Pfam clans, disease code hierarchies, and ATC classifications to create splits for domain, disease, and drug datasets, respectively (Supplementary Note 1, Methods Sec. 1). By structuring the splits this way, we ensure meaningful and robust evaluations of model generalizability. To rigorously evaluate zero-shot capabilities of PROCYON, we train and evaluate a version of the model, named PROCYON-SPLIT, on the split version of PROCYON-INSTRUCT. More information on the different versions of the model is available in Supplementary Note 2.

We first compare PROCYON against seven single modality models that operate on protein sequence or structure alone. We represent the canonical approach of transferring annotations via sequence-similarity by using a k-nearest neighbor baseline from BLAST [49] alignment scores (Methods). To compare against deep representation learning methods, we construct baselines using transformations of state-of-the-art sequence or structure representation learning methods [18, 21]. Capabilities for single-modality models are summarized in Figure 2e, stratified by input modalities (protein sequence, protein structure, and free text) as well as zero-shot and dynamic task specification capabilities. For brevity, we show three representative single modality models in Figure 2f and Figure 2g, with full results shown in Extended Data Figure 2.

Compared to single-modality models, we see that PROCYON shows consistently strong performance in our benchmark of tasks across knowledge domains. PROCYON outperforms all single-modality models in 10 / 14 tasks and is amongst the top three performing approaches in all tasks.

Looking across knowledge domains, PROCYON has the highest performance in both types of problems in our benchmark with an average question answering accuracy of 72.7% (runner-up: ESM3-MLP, 67.8%) and an average protein retrieval F_{max} of 0.743 (runner-up: ESM3-MLP, 0.618). Although we observe that ESM3-MLP and BlastKNN can outperform PROCYON in specific tasks, PROCYON is the only model that consistently outperforms other models in all tasks. In contrast, the runner-up models show less consistency across tasks (ESM3-MLP, 6/10 wins; BlastKNN, 3/10 wins; GearNet-MLP, 1/10 wins). We also note that sequence or structure-only models have been demonstrated to outperform previous multimodal models in specific tasks [50], but rarely across all tasks. As PROCYON consistently improves upon models based on sequence or structure embeddings alone, we can conclude that PROCYON has learned to harmonize these embedding spaces through the use of PROCYON-INSTRUCT.

Benchmarking PROCYON against multimodal models and zero-shot setups

We next compare PROCYON to three multimodal models that integrate protein representations with free-text descriptions [22–24]. However, these models have limited capabilities compared to PROCYON, with some focusing solely on retrieval tasks [22, 23] and others restricted to the generation of phenotypes [24], without the ability to perform both. In all benchmarks, PROCYON demonstrates superior performance in 13 of 14 tasks. Analyzing performance across knowledge domains, PROCYON achieves an average question-answering accuracy of 72.7%, substantially outperforming the runner-up ProtLLM [24], which achieves a near-random accuracy of 49.9%. For protein retrieval, PROCYON achieves an average F_{max} of 0.743, significantly higher than the runner-up ProtST [23] with an F_{max} of 0.570. Importantly, all multimodal baseline methods are pretrained on large protein-text pair datasets. Thus, controlling for test set contamination to evaluate these baseline methods is extremely challenging. Despite this, PROCYON demonstrates robust performance using a rigorously split training dataset. The pretraining datasets of each baseline model are focused on molecular protein functions, such as Gene Ontology and UniProt, thereby limiting the models to molecular phenotypes. However, this may explain why other methods outperform PROCYON in protein retrieval for EC numbers, whose descriptions appear commonly in functional databases, where the leakage of the test set could favor other methods. However, PROCYON ranks as the runner-up model in this specific task.

We next test PROCYON in zero-shot settings. These evaluations utilize text descriptions of phenotypes that were not encountered during training, representing entirely new concepts for

PROCYON. Leveraging our dataset splitting, we evaluate PROCYON on 17,492 protein-phenotype pairs involving 3,250 previously unseen phenotypes spanning the function, therapeutic, and disease knowledge domains. PROCYON demonstrates strong performance in zero-shot settings, achieving results nearly equivalent to those in less challenging few-shot settings, with an average percentage change of only 0.85% for question answering tasks and 1.22% for protein retrieval tasks.

PROCYON retrieves proteins using flexible, natural language prompts

While evaluation on zero-shot phenotypes demonstrates that PROCYON generalizes to unseen phenotypes during testing, we also assess its ability to extend protein annotations beyond standardized ontologies and vocabularies. Unlike models constrained by controlled vocabularies, PROCYON allows users to specify compound phenotypes, varying levels of phenotypic specificity, and novel phenotypes not encountered during training. Using a natural language prompt, PROCYON retrieves proteins from a library of sequences and structures, adapting to user-defined contexts and queries.

As an example (Figure 3a), we examine PROCYON's output for a natural language prompt related to Parkinson's disease (PD). The model retrieves PD-related proteins and ranks them relative to 18,174 human proteins. The ranking quantile indicates how highly PROCYON associates a protein with the phenotype, with higher quantiles corresponding to stronger retrieval (better performance). This retrieval capability is distinct from text-only language models, which cannot ground outputs in protein libraries and have fragile mechanisms for selection [51], and from protein representation models [17, 18], which cannot interpret natural language prompts.

Next, we test PROCYON on a range of disease phenotypes using natural language prompts (Figure 3b). Using PROCYON-FULL, the version of the model trained on PROCYON-INSTRUCT, we evaluate performance across different lines of evidence linking proteins to diseases. The model achieves high average quantile rankings for disease phenotypes across evidence types: $78.11_{\pm 20.89}$ for affected pathways, $79.89_{\pm 17.53}$ for animal models, $81.53_{\pm 21.38}$ for genetic associations, $70.79_{\pm 24.31}$ for drug mechanisms of action, and $81.79_{\pm 17.67}$ for somatic mutations.

We also test PROCYON's ability to leverage increasingly detailed prompts to retrieve disease-associated proteins (Figure 3c). Using descriptions from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [52], we incrementally expand prompts from the disease name alone ("Disease"), to include "Diagnostic Features," and then "Associated Features Supporting Diagnosis." In the latter two categories, we exclude disease names to test whether the model can generalize to phenotype descriptions rather than simply matching the names of diseases. These additional

descriptions capture clinical presentations and associated phenotypes. PROCYON consistently improves retrieval performance as the level of detail in the prompt increases. When only the disease name is provided, the average quantile ranking is $86.99_{\pm 19.21}$. Adding "Diagnostic Features" increases performance to $90.11_{\pm 14.47}$, while "Associated Features Supporting Diagnosis" yields $90.13_{\pm 12.08}$. Combining all fields further improves performance, reaching $91.69_{\pm 10.15}$. This trend highlights PROCYON's ability to interpret and utilize phenotypes described with varying levels of granularity in user prompts.

We further test PROCYON on 22 monogenic diseases, where causal genes have been identified through cloning, linkage analysis or other types of experimental validation (Supplementary Table 4), serving as positive controls (Extended Data Figure 6). As with broader disease phenotypes, excluding disease names from the input prompts only slightly decreases performance, suggesting that PROCYON captures relationships between proteins and disease phenotypes beyond simple name matching. Finally, we visualize PROCYON's retrieval space (Figure 3d) both at initialization (top; before phenotype learning) and after training (bottom; after phenotype learning). The results show that contrastive retrieval training (Methods Sec. 3.4) leads to the spatial clustering of proteins with shared phenotypes in the retrieval space. This organization underpins PROCYON's performance on retrieval benchmarks, providing quantitative and qualitative evidence of its ability to associate proteins with phenotypes (Figure 2f,g).

Prompting PROCYON with complex and pleiotropic protein phenotypes

The flexibility of natural language prompts enables PROCYON to handle complex queries beyond individual, predefined concepts in controlled vocabularies. To illustrate this capability, we task PROCYON with retrieving potentially pleiotropic proteins involved in multiple biological pathways (Methods Sec. 4.6 for details on procedure and definition of pleiotropy). PROCYON retrieves proteins participating in up to four pathways (as an example, GAPDH is annotated with GO:0006096 glycoltic process, GO:0051402 neuron apoptotic process, GO:0050821 protein stabilization, and GO:0017148 negative regulation of translation) using either a single compositional prompt or by aggregating retrieved lists from individual pathway prompts (Figure 3e). The performance of both approaches is significantly better than random compositions (all p-values $< 10^{-4}$; Wilcoxon signed rank test). The composite prompt outperforms the aggregation of individual retrieval lists (p-value $< 10^{-4}$; Wilcoxon signed rank test), demonstrating PROCYON's ability to process multi-phenotype prompts directly, eliminating the need for users to manually engineer multi-task predictors.

Next, we prompt PROCYON to retrieve proteins involved in crosstalk between signaling pathways [53]. Crosstalk occurs when the activation of one pathway influences another pathway, indicating shared protein involvement across multiple biological processes. Without flexible natural language prompts, prioritizing such proteins would require querying each pathway separately within a controlled vocabulary and aggregating ranked lists. PROCYON successfully prioritizes crosstalk proteins using both compositional prompts and aggregated lists from individual pathway prompts, with both methods significantly outperforming random compositions (all p-values $< 10^{-4}$; Wilcoxon signed rank test). The compositional prompt slightly outperforms the aggregated lists (p-value $= 0.026$; Wilcoxon signed rank test), highlighting the utility of integrated prompts for resolving complex pathway relationships (Figure 3f; Methods Sec. 4.6). This result shows that PROCYON effectively combines user-provided information, even outside of the distribution of training samples, to retrieve proteins leveraging compositional phenotype descriptions.

We further challenge PROCYON with functional prompts (Supplementary Note 3) based on findings from a recent study [54] published in July 2024, after PROCYON's knowledge cut-off date (latest accession date of any database used for training: GtoP, Oct 2023). Specifically, we use prompts about neuronal inflammatory stress response (NISR) to evaluate whether PROCYON can prioritize the STING protein (UniProt: Q86WV6). The involvement of STING in NISR was only recently established [54], where it was shown that neuronal STING is activated by excitotoxicity, leading to autophagy-dependent cell death. Such involvement is absent from all our training data, where the closest phenotype to NISR in PROCYON-INSTRUCT, "inflammatory response" (from Gene Ontology, GO:0006954), neither annotates STING nor captures the specific neuronal context established in [54]. For this experiment, we compare the ranking of STING against a control group of proteins derived from all proteins expressed in adult central nervous system tissues, further subsetting to the top 10% ranked by PROCYON in the initial broadest prompt. This provides a stringent control group with inherently high biological relevance (Methods Sec. 4.6).

As we refine the functional prompt with increasing precision, PROCYON prioritizes STING more effectively. For a general phenomenon prompt (Figure 3g), STING ranks 27.3 quantile points below the control group. Including mechanistic details in the prompt (Figure 3g) reduces the gap to 20.3 quantile points. Finally, incorporating precise functional information from the recent study [54] elevates STING to 7.1 quantile points above the control group. Such a result follows an expected biological pattern, as the "basic phenomenon" field has many proteins potentially involved in the pathway, but as more information is added, the phenotype is more specific to STING

and the pathway elucidated by [54]. This progressive improvement demonstrates PROCYON's ability to incorporate novel functional insights described in natural language, linking them to relevant protein phenotypes (Figure 3g). These results highlight PROCYON's potential to leverage natural language prompts of emerging biological phenomena by processing detailed and context-specific prompts.

Generating phenotype descriptions using PROCYON

PROCYON performs Phenotype Generation, producing descriptive phenotypes conditioned on user instructions. Unlike retrieval, which outputs a ranked list of proteins for a given query, phenotype generation processes a multimodal interleaved input—consisting of a protein and a user-provided prompt—and generates proposed phenotypes for the protein. This is achieved through PROCYON's ability to process multimodal inputs using autoregressive language generation to produce textual descriptions of proteins (Methods Sec. 2.3). Unlike text-only LLMs, PROCYON generalizes to new proteins using its protein sequence and structure encoding mechanisms, allowing it to provide meaningful insights even for proteins that are poorly represented in the biomedical literature. In contrast, text-only LLMs depend on predefined naming systems, such as HGNC gene names [27], which, as we show in this section, significantly limits their utility for poorly studied or entirely novel proteins. To further evaluate PROCYON's capabilities, we compare its outputs with those of both open-source and proprietary LLMs.

Figure 4a highlights an example where PROCYON generates phenotype descriptions for a poorly characterized protein, AKNAD1 (UniProt: Q5T1N1). AKNAD1 is absent from PROCYON's training data, with no phenotype associations in PROCYON-INSTRUCT, and is broadly undercharacterized. UniProt assigns AKNAD1 an annotation score of 3/5 (where 5 represents a functionally well-characterized protein), indicating limited knowledge relative to other human proteins. Additionally, AKNAD1 shares low sequence similarity with other human proteins in SwissProt (BLASTp: most similar human protein: 27% sequence identity with AKNA), making functional inference via sequence similarity challenging.

PROCYON generates three molecular function descriptions for AKNAD1. Independent validation using the Human Protein Atlas (HPA) [8] confirms that two of these functions are supported by experimental evidence (data from https://www.proteinatlas.org/ENSG00000130590-AKNAD1). The first description, suggesting cytoskeletal localization, is backed by immunofluorescence data. The second description, involving spermatogenesis, aligns with gene expression evidence show-

ing overexpression in spermatids. The third description, positing regulation of cAMP-dependent protein kinase activity, requires further experimental validation. Details on validation of these phenotypes is provided in Methods Sec. 4.7. These results demonstrate PROCYON's potential to uncover functional insights for poorly understood proteins, addressing the gap in knowledge for the approximately 20% of human proteins with unknown functions [55, 56].

To enhance PROCYON's generation accuracy, we implement a QA filtering mechanism. Generated descriptions are re-evaluated by PROCYON using a QA prompt to predict whether each generated description matches the protein (Methods Sec. 2.3). Outputs falling below a confidence threshold (a "yes" score below 0.6) are excluded from the final results. This self-checking mechanism leverages PROCYON's free-text modeling capabilities to improve the quality of phenotype descriptions. Extended Data Figure 7 provides examples of this process, while the Extended Data Figure 8 shows improved accuracy as the filtering threshold becomes more stringent.

Next, we evaluate PROCYON's text generation capabilities against text-only LLMs of varying sizes across both QA and free-text generation tasks. PROCYON matches or outperforms larger text-only LLMs in QA (Figure 4b). For text generation, measured via n-gram similarity to reference texts, PROCYON also outperforms state-of-the-art models such as GPT-4o (Figure 4b).

However, text-only LLMs exhibit strong sensitivity to protein representation methods. When a protein is provided to a text-only LLM, a text encoding must be chosen to represent the protein. We find that text-only LLMs exhibit a strong sensitivity to this choice of encoding method. Using GPT-4o, we test three encoding formats for proteins: amino acid sequence, UniProt ID, and HGNC gene name. The highest quality responses are achieved with the HGNC names, while the performance is significantly reduced for amino acid sequences or UniProt IDs (Figure 4c). Note that to ensure consistency with PROCYON evaluation, we prompt GPT-4o with one example of the style of the output text to maximally benefit GPT-4o's in-context learning abilities. The variability in performance with protein representation reflects the limitation of text-only LLMs since their training data primarily reference proteins by gene names.

While using the amino acid sequence as a representation may be suboptimal for text-only LLMs, it is the only text-based method that allows for the specification of arbitrary, potentially novel proteins that lack a UniProt ID or HGNC gene name. In contrast, PROCYON's approach of encoding proteins through embeddings generalizes seamlessly to such novel proteins without relying on predefined nomenclature. This disparity in performance across different vocabulary representations points to a fundamental limitation of text-only LLMs: their fragility to variations

in naming conventions for proteins and other biomolecules. Relying solely on text-based representations ties performance to the completeness and consistency of training data, which reflects biases in scientific literature. PROCYON bypasses these constraints and provides a robust solution for handling proteins.

To holistically evaluate text generation, we employ an LLM-as-a-judge approach [57], using Claude-3.5-Sonnet [58], a state-of-the-art production LLM, to rank outputs from PROCYON and other models. This method captures broader aspects of correctness and relevance beyond n-gram similarity. PROCYON's open-ended text generation surpasses similarly sized text-only LLMs but lags behind significantly larger models when these are provided with optimal protein representations (Figure 4d). These results suggest that scaling PROCYON's base LLM could further enhance its generative capabilities [59].

Finally, we explore PROCYON's ability to tailor phenotype descriptions to user expertise using system prompts. Figure 4e illustrates this with two descriptions of Amyloid-beta precursor protein (APP, UniProt: P05067). For a junior scientist, PROCYON generates a concise, straightforward description, while for a human expert, it provides a more detailed and nuanced output. This customization can be used for any phenotype generation task [60].

Zero-shot task transfer: Identifying drug-binding protein domains

PROCYON performs predictive and generative tasks, including the ability to generalize to entirely new tasks, not just new label sets within predefined task definitions or single knowledge domains (Figure 2). We refer to this capability as **dynamic task specification**, where users can define novel tasks at test time that were not encountered during model training. By representing tasks through natural language, PROCYON leverages LLMs to handle previously unseen tasks, whether they involve annotating phenotypes at different scales or operating on new molecular modalities such as peptides. The following sections demonstrate this capability in various biological applications.

We first investigate the ability of PROCYON to prioritize protein domains that mediate drug binding. Using a protein domain-drug interaction dataset [61], we perform drug-target retrieval queries for several drugs. PROCYON employs its retrieval mechanism to retrieve candidate protein domains from a library that encompasses all domains of the target protein, evaluating where the correct binding domain is ranked. Figure 5a illustrates an example of the digestive enzyme Maltase-glucoamylase (MGAM) and the small molecule drug miglitol which binds MGAM and is indicated for type 2 diabetes (T2D). This multimodal prompt includes MGAM as the target protein,

miglitol as the drug (given as both a description of the mechanisms of action and the molecular structure), and a natural language question. Unlike existing models restricted to text-only or single-modality inputs, PROCYON can interleave multiple data modalities in inputs, including proteins and small molecule drugs. The figure also shows the inference setup, showing the decomposition of MGAM into domains and the retrieval scores generated by PROCYON. PROCYON achieves this without explicit training in this specific task, leveraging its capacity to adapt through natural language instructions.

MGAM consists of 2,753 amino acids and nine domains, including two main catalytic domains: NtMGAM at the N-terminus and CtMGAM at the C-terminus. Although miglitol binds to both domains, its inhibitory effects are stronger for NtMGAM [62]. PROCYON correctly ranks the NtMGAM catalytic domain (`PF01055_349`) as the top binding domain, assigning it a similarity score of 0.0269 units higher than the CtMGAM catalytic domain (`PF01055_2111`). This result is notable given the subtle but important differences in catalytic efficiency between these homologous domains, influenced by the length of oligosaccharides they process [63].

Figure 5b extends these findings to protein-drug complexes [61]. In 24 out of 35 examples of multi-domain proteins (68.6%), PROCYON ranks the correct drug-binding protein domain at the top of the list. On average, this ranking is 2.5 positions higher than expected if the domains were ranked randomly, considering all domains present in each protein. Performance on protein domain retrieval task highlights PROCYON's ability to represent phenotypic information at protein substructure level beyond whole-protein representations. This capability is likely due to the model's training in protein domain-specific phenotypic data, enabling it to achieve robust and accurate predictions. Even when the drug-binding protein domain is not ranked first, it is consistently ranked near the top: at position two in 5 multi-domain proteins (14.29%, average 1.3 positions higher than reference ranking), position three in 4 multi-domain proteins (11.43%, no significant difference with reference ranking), and position four in 2 cases (5.71%, average 1.25 positions higher than reference ranking). In none of the examples does the correct protein domain rank lower than five, even though the median of protein domains in 35 proteins is 5, with the maximum number of 12 domains in protein O14727 (Apoptotic protease-activating factor 1). These results underscore the robustness of PROCYON in operating at molecular scales such as protein domains.

Zero-shot task transfer: Identifying peptides that bind to target proteins

Next, we demonstrate PROCYON's ability to dissect complex molecular relationships between proteins and peptides. Peptides, short chains of amino acids, comprise 16% of FDA approved drugs as of 2023 [64]. For example, the widely known drug *semaglutide* (marketed as *Ozempic*, *Wegovy*, and *Rybelsus*) is a peptide therapy that mimics the glucagon-like peptide 1 (GLP-1) and is used for chronic weight management and type 2 diabetes [65]. We evaluate PROCYON's performance in predicting protein-peptide binding, a critical step in identifying promising peptide therapies (Figure 5c). Although PROCYON was not trained on physical binding data, we fine-tuned it using protein-peptide binding data from PDBBind [66], resulting in a specialized version called PROCYON-BIND.

We tested PROCYON-BIND on a synthetic peptide dataset with experimentally validated binding data for angiotensin-converting enzyme 2 (ACE2) [67]. ACE2 acts as the receptor for the spike S1 protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [68, 69], and peptide binders targeting ACE2 have been proposed as a mechanism to inhibit this interaction, potentially increasing resistance to SARS-CoV-2 [70]. The experimental dataset includes 58 detected ACE2 peptide binders and 5,072 non-binders, derived from a randomized library of synthetic peptides [71] Given the heavy imbalance in binders and non-binders, this creates a challenging retrieval setting that more closely mimics lead discovery in practice.

Despite the small fine-tuning dataset, PROCYON-BIND can be adapted to peptide binding prediction tasks. With just over 1,000 fine-tuning samples, PROCYON-BIND achieves a statistically significant separation between binders and non-binders ($p = 1.04 \times 10^{-4}$; Mann-Whitney U test), yielding an AUROC of 0.6480 on peptide binders (Figure 5c). We also run an inference based on ESM-2 that achieves 0.3923 AUROC (Supplementary Note 5). Although complete score-based separation of the peptide binder and non-binder classes is not achieved, no clear peptide sequence motif was experimentally identified to drive binding, and discerning binders from non-binders was experimentally challenging [67]. However, the results (Figure 5c) mimic ligand discovery techniques, which seek to enrich a distribution of binders above the library background for their affinity for the target. PROCYON-BIND's application to peptide-based predictions through fine-tuning, despite being pretrained solely on proteins and protein domains, illustrates its versatility.

Zero-shot task transfer: Modeling phenotypes linked to genetic variants

We next use PROCYON to (1) describe the functional impacts of genetic variants and (2) correlate these changes with phenotype descriptions generated for the mutated protein representations. Given a protein, we evaluate PROCYON to distinguish between mutations with varying functional consequences. This capability arises from PROCYON's integration of a protein language model, which captures sequence variants, with a large language model that generates detailed phenotypic descriptions of these alterations. We test this capability using benign and pathogenic mutations in the Alzheimer's disease (AD) genes PSEN1 and PSEN2.

PSEN2 (presenilin 2), an aspartyl protease and subunit of the γ-secretase complex, cleaves APP (amyloid precursor protein) into amyloid-beta peptides [72]. Additionally, PSEN2 is essential for calcium signaling and transport [73]. Phenotype descriptions generated by PROCYON for pathogenic PSEN2 variants omit references to calcium signaling, suggesting that these mutations may disrupt calcium homeostasis (Figure 5d). This disruption has been implicated in AD pathogenesis [74]. These findings illustrate PROCYON's ability to distinguish between benign and pathogenic variants and generate explanations for these differences, providing hypotheses that could guide future experimental investigations into disease mechanisms.

Using PSEN1 (presenilin 1), another AD protein, we inject single nucleotide polymorphisms associated with AD into the wild-type PSEN1 sequence and compare PROCYON's generated phenotypes for the mutated and wild-type protein representations (Figure 5e). For clarity, we focus on functional keywords extracted from the generated descriptions rather than narrative text. Structural data and AD-relevant PSEN1 variants were sourced from Bai et al. [75], which detail the protein's organization into nine transmembrane (TM) sections (Methods Sec. 4.11). PROCYON generated phenotypes that reveal phenotypic changes that vary based on the mutation's location within PSEN1's structure. Although the biological roles of PSEN1 and TM sections remain poorly understood, PROCYON-generated phenotypes align with known functions. For example, keywords related to cell-cell interaction and communication reflect PSEN1's involvement in the Notch signaling pathway, whose disruption is associated with AD [76, 77]. Similarly, keywords related to ion transport align with PSEN1's role in this process [78]. By correlating sequence variations with generated phenotypes, PROCYON can generate hypotheses about phenotypes linked to genetic variants.

Zero-shot task transfer: Retrieving proteins against small molecule drugs

PROCYON can handle phenotype queries that interleave drug and disease knowledge domains. One practical application of this capability is the identification of disease-specific targets, where PROCYON can retrieve proteins targeted by a small-molecule drug for a given disease. To illustrate this, we use bupropion as an example (Figure 5f). Bupropion is a small molecule drug approved by the US FDA to treat major depressive disorder (MDD) and aid in smoking cessation [79, 80]. Its therapeutic effects are mediated through distinct protein targets: the norepinephrine transporter (NET, UniProt: P23975), which is primarily involved in treating MDD [81,82], and the cholinergic receptor nicotinic alpha 3 subunit (AChR, CHRNA3, UniProt: P32297), which plays a role in smoking cessation [82]. Bupropion also targets the dopamine transporter (DAT, UniProt: Q01959), which contributes to the therapeutic mechanisms of both MDD and smoking cessation [82].

We constructed a novel task (not specified during model training) where PROCYON was instructed to retrieve proteins targeted by bupropion that are involved in the therapeutic mechanisms of either MDD or nicotine addiction (associated with smoking cessation). Disease descriptions were extracted from StatPearls [83, 84], and a compound retrieval prompt was defined to include both drug and disease context.

PROCYON's performance highlights its ability to discern between therapeutic mechanisms in a disease-specific manner. For the smoking cessation prompt, AChR ranks above NET ($p = 0.1153$, Wilcoxon signed-rank test), consistent with AChR's role in nicotine addiction. Conversely, for the MDD prompt, NET ranks far above AChR ($p = 1.974 \times 10^{-4}$, Wilcoxon signed-rank test), aligning with NET's involvement in MDD therapy. Further, all three proteins (NET, AChR, and DAT) are ranked highly among all human proteins considered as query candidates (all within the top 40 out of 18,174 proteins). These results showcase PROCYON's ability to bridge distinct knowledge domains, leveraging natural language to address novel, cross-domain queries about complex protein phenotypes. The differential ranking of AChR above NET for smoking cessation and the inversion of their rankings for MDD highlights PROCYON's sensitivity to composite queries that combine disease-specific information with drug context. This task could not be performed by traditional protein classification models, especially given that PROCYON was not explicitly trained for such tasks. By allowing conditional information to be incorporated into query prompts, PROCYON enables more precise drug-target discovery. This flexibility and contextual adaptability underscore PROCYON's potential to refine and advance drug discovery pipelines, enabling more targeted therapeutic insights.

PROCYON can elucidate functions of poorly-characterized proteins

PROCYON's architecture and training are designed to maximize transferability across phenotype descriptions. By integrating phenotypes with protein sequences and structures, PROCYON generates transferable representations, enabling zero-shot learning of phenotypes for sequences and structures PROCYON never encountered during training. To evaluate this zero-shot capability, we use PROCYON to predict phenotypes for poorly characterized proteins associated with Parkinson's disease (PD).

We first examined PROCYON-generated phenotypes for a subset of functionally poorly characterized proteins: ANKDD1A, SVOP, TCEAL5, and ZNF654, whose relevance for α-synucleinopathy [85] was experimentally confirmed [86] (Extended Data Figure 9). These proteins represent zero-shot predictions since PROCYON was not trained on their phenotypes. Furthermore, all four proteins have low sequence similarity to other human proteins in SwissProt (BLASTp, most similar human proteins: ANKDD1A, 35.5% sequence identity with ANKDD1B; SVOP, 39.6% sequence identity with SVOPL; ZNF654, 29.7% sequence identity with ZNF292), and TCEAL5 has low sequence similarity with proteins other than TCEAL3, TCEAL6, TCEAL2, and TCEAL4 (all poorly characterized with no more than 3/5 annotation score on UniProt; most similar other human protein: 30.5% sequence identity with SLC24A1), making functional inference via standard sequence similarity approaches unsuccessful. For ANKDD1A, PROCYON generated descriptions of endoplasmic reticulum (ER) to Golgi vesicle-mediated transport (Expert score: 3, Figure 6), a function disrupted by α-synuclein in PD [87]. For SVOP, PROCYON described sodium-dependent transporters, which are implicated in α-synuclein dysfunction, particularly through dopamine transporters (which are sodium-dependent) that can lead to the degeneration of dopaminergic neurons [88, 89]. However, PROCYON incorrectly generated renal tissue expression phenotype for SVOP, for which little evidence exists in the Human Protein Atlas [8]. For TCEAL5 and ZNF654, PROCYON predicted transcriptional regulation, a function broadly implicated in PD [90].

We then selected 35 proteins from Open Targets [91] with reported associations to PD but no presence in PROCYON's training data, referred to as the "PD Uncharacterized" set. Using PROCYON, we generate phenotype descriptions for these proteins. PROCYON-generated descriptions of biological processes for these proteins were subjected to QA filtering and grouped into pathways using a combination of semantic similarity-based clustering and manual curation. A neurologist specializing in PD reviewed these pathways and scored them on a scale of 0-3, where 0 indicated no association with PD, and 3 indicated a strong, well-known association (Figure 6a).

Next, we examine the enrichment of ProCyon-generated pathways for uncharacterized PD proteins compared to two control groups. The first control group ("Neuro") includes proteins expressed in the nervous system but not associated with PD or neurodegenerative diseases. The second ("General") includes proteins not expressed in neurological tissues. Using Open Targets, we compile a list of 962 well-characterized PD-associated proteins ("PD Associated") (Methods Sec. 4.13). We assess the proportion of predicted pathways for the PD Uncharacterized set that overlapped with the PD Associated list compared to the controls. We group pathways by expert-provided association scores and use a one-sided proportion z-test to determine significance.

Results in Figure 6b show that pathways identified by ProCyon for the PD Uncharacterized set are significantly enriched in the PD Associated list compared to both controls. Of the 31 predicted pathways, 21 show statistically significant enrichment against the Neuro control, and 21 against the General control. This enrichment highlights ProCyon 's ability to identify pathways specific to PD rather than general neurodegenerative mechanisms, such as protein aggregation or neuroinflammation [92, 93]. Interestingly, pathway enrichment was observed across all expert-provided scores, including pathways ranked lower by the neurologist. For example, 8/13 pathways scored 3, 7/9 scored 2, and 6/8 scored 1 were significant compared to the Neuro control. Compared to the General control, 9 out of 13 pathways with a score of 3 were significant, along with 7 out of 9 pathways scored 2, and 5 out of 8 pathways scored 1. None of the pathways scored 0 showed significant enrichment. This suggests that pathways considered understudied or less emphasized in PD literature may still hold relevance. While expert assessments are naturally biased toward well-established and extensively studied pathways in PD literature, some understudied pathways (those ranked lower by the expert) may still play a meaningful role in PD pathogenesis. For instance, while DNA transcription is not a primary focus of PD research, ProCyon identified it as significantly enriched in PD-associated proteins, suggesting potential roles in disease mechanisms despite its lesser emphasis in the literature [94]. Conversely, pathways ranked high by the expert were frequently validated, indicating ProCyon's capacity to derive specific functions for poorly annotated proteins. These results demonstrate ProCyon's potential for generating phenotype descriptions for proteins poorly understood in complex diseases like PD.

Discussion

PROCYON is a multimodal foundation model for modeling, generating, and predicting complex protein phenotypes in five knowledge domains: molecular functions, therapeutic mechanisms, disease associations, functional protein domains, and molecular interactions. It represents a shift from modeling protein sequences and structures to modeling complex protein phenotypes, which are inherently multiscale and encompass molecular-level interactions, cellular pathways, organismal traits, and disease associations. By analyzing millions of human protein phenotypes, PROCYON bridges phenotype descriptions with molecular modalities—protein sequences, domains, peptides, and structures—enabling the generation of novel phenotypic insights. Its architecture and training allow for zero-shot task transfer, which enables the model to generalize to entirely new phenotype tasks not encountered during training. Zero-shot task transfer enables PROCYON to go beyond pre-defined ontologies and explore phenotypes in an open-ended, hypothesis-driven manner.

Protein phenotypes are diverse and context-dependent, ranging from single-molecule functions to their roles in pathways and diseases. PROCYON models these phenotypes in five distinct but interrelated knowledge domains to predict and generate phenotypes on multiple scales. For example, PROCYON accurately generates molecular phenotype descriptions for poorly characterized proteins. It identifies pathways and processes relevant to disease biology, such as endoplasmic reticulum-to-Golgi transport, a process known to be disrupted in α-synuclein pathologies. PROCYON integrates sequence and structural data to predict protein-binding domains with high precision. For example, it ranked the correct binding domain for MGAM in its interaction with miglitol, demonstrating its ability to resolve subtle distinctions between homologous protein domains. These new and biologically relevant applications leverage PROCYON's zero-shot task transfer capability and show PROCYON's ability to provide mechanistic insights into underexplored proteins, retrieve proteins conditioned on small-molecule drugs by accounting for target differences in drug mechanisms of action, identify peptides that bind target proteins, and bridge molecular-scale phenotypes with broader disease-specific phenotypes.

Unlike models that focus on the integration of sequence and structure, PROCYON prioritizes the modeling and generation of complex protein phenotypes. By framing the predictions through a natural language interface, PROCYON democratizes access to protein phenotypes and allows researchers of various backgrounds (Figure 4e) to explore proteins in a contextualized way. For instance, PROCYON's ability to process composite queries involving drugs and diseases demonstrates its utility in tasks such as drug-target discovery that require reasoning across

different knowledge domains. An example is bupropion, for which PROCYON predicted protein targets based on their therapeutic functions in smoking cessation and major depressive disorder. This retrieval highlights PROCYON's ability to resolve disease contexts in input prompts.

PROCYON's limitations highlight areas for improvement and future research. While PROCYON's language modeling capabilities enable it to generate textual descriptions, these outputs may occasionally lack coherence or fail to capture nuanced biological contexts for highly specialized prompts. Evaluating novel phenotypes remains inherently challenging, as ground truth data may be scarce or absent, complicating validation efforts. The static nature of PROCYON's instruction tuning data can limit its ability to adapt to evolving biological research. Further, PROCYON does not yet incorporate conversational capabilities, which could enable iterative refinement of prompts in multi-turn interactions with scientists. Another limitation is the lack of uncertainty quantification in PROCYON's predictions. PROCYON does not explicitly convey the uncertainty associated with its generated descriptions or retrieval results, which could be critical to researchers when interpreting predictions. PROCYON currently lacks integration with multi-omics datasets. Although proteomics data remain relatively scarce, future integration of such modalities, as more data become available, could enhance PROCYON's ability to link protein phenotypes with cellular contexts. PROCYON's modular design opens opportunities to incorporate additional biological modalities that could expand its applicability to other species.

We envision models like PROCYON that unify multiple biological modalities and scales to address increasingly complex biological questions. The modular design of PROCYON allows for incorporating additional encoders to integrate modalities such as genomics, transcriptomics, proteomics, and imaging data. Future iterations could leverage these multimodal integrations to drive hypothesis generation, experimental design, and exploration of phenotypic complexity across diverse biological systems. Beyond function querying, PROCYON's protein-text representations hold promise for advancing other applications. For instance, these representations could be adapted for generative protein design, such as guiding protein generative diffusion models to create novel functional protein sequences. Additionally, PROCYON's embeddings, enriched with contextual information, could improve the generation of conformational states that are challenging for current models, potentially enabling better interpretation of dynamic conformations for the same protein sequence under varying conditions.

PROCYON is a multimodal foundation model designed to address two fundamental aspects of human protein biology: the multimodal nature of protein representation and the complexity of

protein phenotypes. By integrating protein phenotypes across multiple scales and knowledge domains, PROCYON bridges the gap between molecular data and human-readable phenotype descriptions, enabling insights that go beyond pre-defined phenotypic categories. This feature positions PROCYON as a foundational tool for improving our understanding of protein phenotypes.

Acknowledgements. We thank Jessika Baral, Michelle M. Li, Shanghua Gao, Thomas Wood, and Manuel A. Friese for their helpful discussions on the manuscript and their input on the applications of the model. We are grateful to Max Shad and Ben Sabath from the Research Engineering team at the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University for their guidance on multi-node GPU training and large-scale computational analyses. We thank Nathan Palmer and Francis Yu from Harvard Medical School AI Development Central. We also thank Kirill E. Medvedev, Dustin Schaeffer, and Nick V. Grishin for their help in processing data from DrugDomain on drugs and small molecules bound to domains. We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, ARPA-H BDF program, awards from Chan Zuckerberg Initiative, Bill & Melinda Gates Foundation INV-079038, Aligning Science Across Parkinson's (ASAP) Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, Harvard Data Science Initiative, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. V.G. was supported by the UK Medical Research Council, MR/W00710X/1. K.H. was supported by Joachim Herz Stiftung via Add-on Fellowships for Interdisciplinary Life Science and the Helmholtz Association under the joint research school "Munich School for Data Science." Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Data availability. Project website for PROCYON is at https://zitniklab.hms.harvard.edu/ProCyon. All processed data supporting the findings of this study are available at https://huggingface.co/datasets/mims-harvard/ProCyon-Instruct. Source data for Figures 1-5 are available in this manuscript.

Code availability. We fully release all of PROCYON's model weights and codebase (https://github.com/mims-harvard/ProCyon) as well as an online demonstration of PROCYON such that the research community can build on and explore our model's capabilities.

Authors contributions. O.Q., Y.H., M.K., and M.Z. conceived the study. O.Q., Y.H., R.C., V.G., Y.E., M.K., and M.Z. designed the experiments. O.Q., Y.H., R.C., G.D., L.T., T.Co., and K.H. collected and processed the data. O.Q., Y.H., R.C., T.Ch., G.D., and T.Co. developed the model and training codebase. O.Q., Y.H., R.C., V.G., L.T., Y.E., and A.N. conducted the experimental

analyses and interpreted the results. T.Ch., G.D., A.N., J.B., T.H., B.P., F.J.T, and V.K. provided feedback on the analyses. O.Q., Y.H., R.C., V.G., and M.Z. wrote the manuscript with input from all co-authors. M.K. and M.Z. supervised the research, with M.Z. leading the study.

Competing interests. F.J.T. consults for Immunai Inc., CytoReason Ltd, Cellarity, BioTuring Inc., and Genbio.AI Inc., and has an ownership interest in Dermagnostix GmbH and Cellarity. Other authors declare no competing interests.

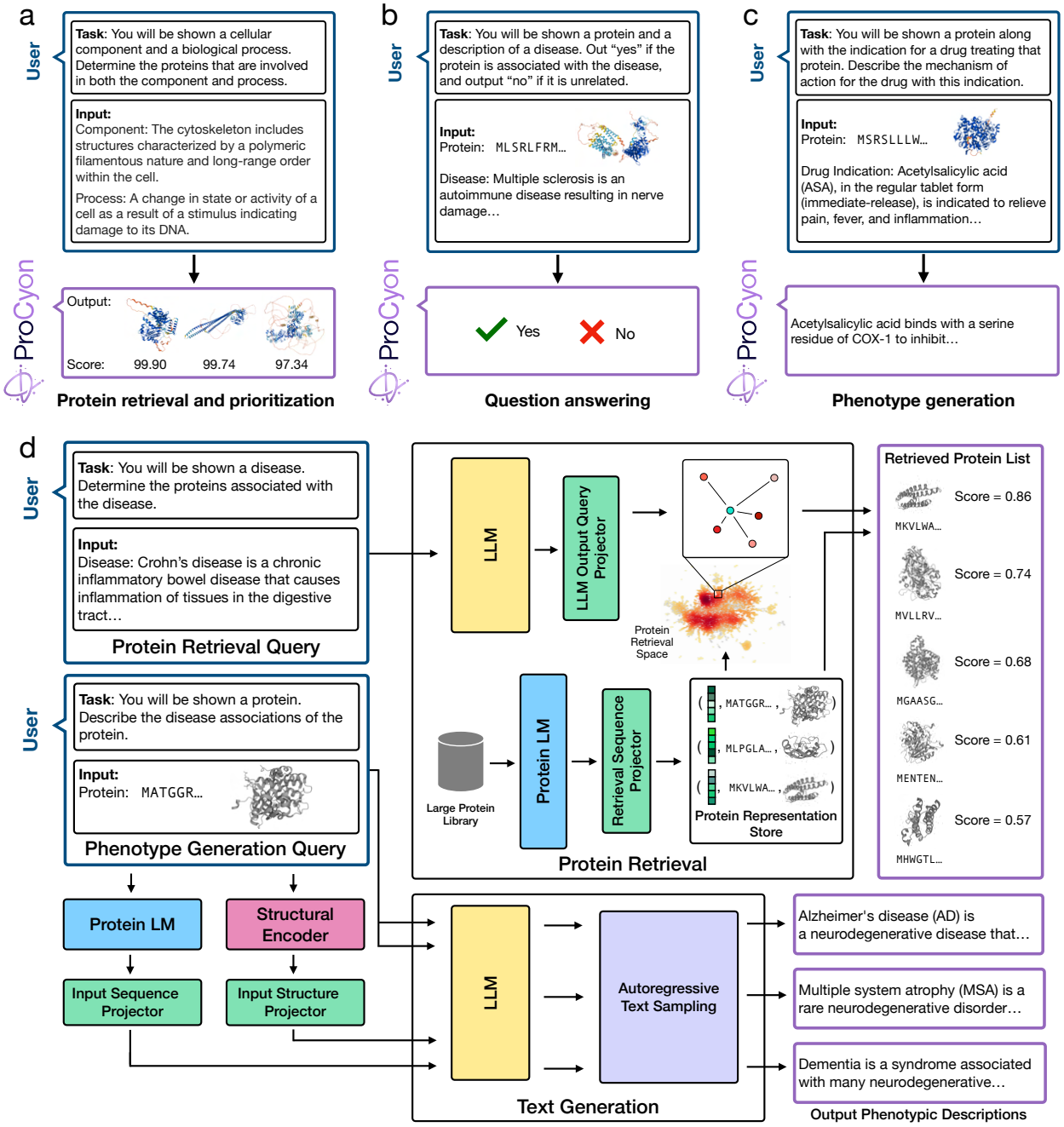

a
User
Task: You will be shown a cellular component and a biological process. Determine the proteins that are involved in both the component and process.
Input:
Component: The cytoskeleton includes structures characterized by a polymeric filamentous nature and long-range order within the cell.
Process: A change in state or activity of a cell as a result of a stimulus indicating damage to its DNA.
ProCyon
Output:
Score: 99.90 99.74 97.34
Protein retrieval and prioritization
b
User
Task: You will be shown a protein and a description of a disease. Out "yes" if the protein is associated with the disease, and output "no" if it is unrelated.
Input:
Protein: MLSRLFRM...
Disease: Multiple sclerosis is an autoimmune disease resulting in nerve damage...
ProCyon
Yes
No
Question answering
c
User
Task: You will be shown a protein along with the indication for a drug treating that protein. Describe the mechanism of action for the drug with this indication.
Input:
Protein: MSRSLLLW...
Drug Indication: Acetylsalicylic acid (ASA), in the regular tablet form (immediate-release), is indicated to relieve pain, fever, and inflammation...
ProCyon
Acetylsalicylic acid binds with a serine residue of COX-1 to inhibit...
Phenotype generation
d
User
Task: You will be shown a disease. Determine the proteins associated with the disease.
Input:
Disease: Crohn's disease is a chronic inflammatory bowel disease that causes inflammation of tissues in the digestive tract...
Protein Retrieval Query
User
Task: You will be shown a protein. Describe the disease associations of the protein.
Input:
Protein: MATGGR...
Phenotype Generation Query
Protein LM
Structural Encoder
Input Sequence Projector
Input Structure Projector
LLM
LLM Output Query Projector
Protein Retrieval Space
Large Protein Library
Protein LM
Retrieval Sequence Projector
MATGGR...
MLPGLA...
MKVLWA...
Protein Representation Store
Protein Retrieval
Retrieved Protein List
Score = 0.86
MKVLWA...
Score = 0.74
MVLLRV...
Score = 0.68
MGAASG...
Score = 0.61
MENTEN...
Score = 0.57
MHWGTL...
LLM
Autoregressive Text Sampling
Text Generation
Alzheimer's disease (AD) is a neurodegenerative disease that...
Multiple system atrophy (MSA) is a rare neurodegenerative disorder...
Dementia is a syndrome associated with many neurodegenerative...
Output Phenotypic Descriptions

Figure 1

Figure 1: PROCYON overview and architecture. (a) A demonstration of PROCYON protein retrieval and prioritization, whereby the output of the model is a ranked list of proteins, domains, peptides, or polypeptides more broadly. Shown is an example of querying for DNA damage response biological process as well as cytoskeleton cellular component. **(b)** PROCYON performs multimodal question answering where the input is a protein and a phenotype description (e.g., disease, description of drug, etc.); and the output consists of a yes or no answer. **(c)** Leveraging generative capabilities, PROCYON can generate protein phenotypes and more free-form answers to user prompts. **(d)** Architecture of PROCYON: For protein retrieval, a prompt is processed into the LLM and compared against a library of proteins. The output is a ranked list of proteins. In text generation (which includes QA and Phenotype generation), the input text and multimodal elements are processed to the LLM via multimodal token decomposition; an autoregressive text generation algorithm is then used to sample PROCYON for multiple output phenotypes.

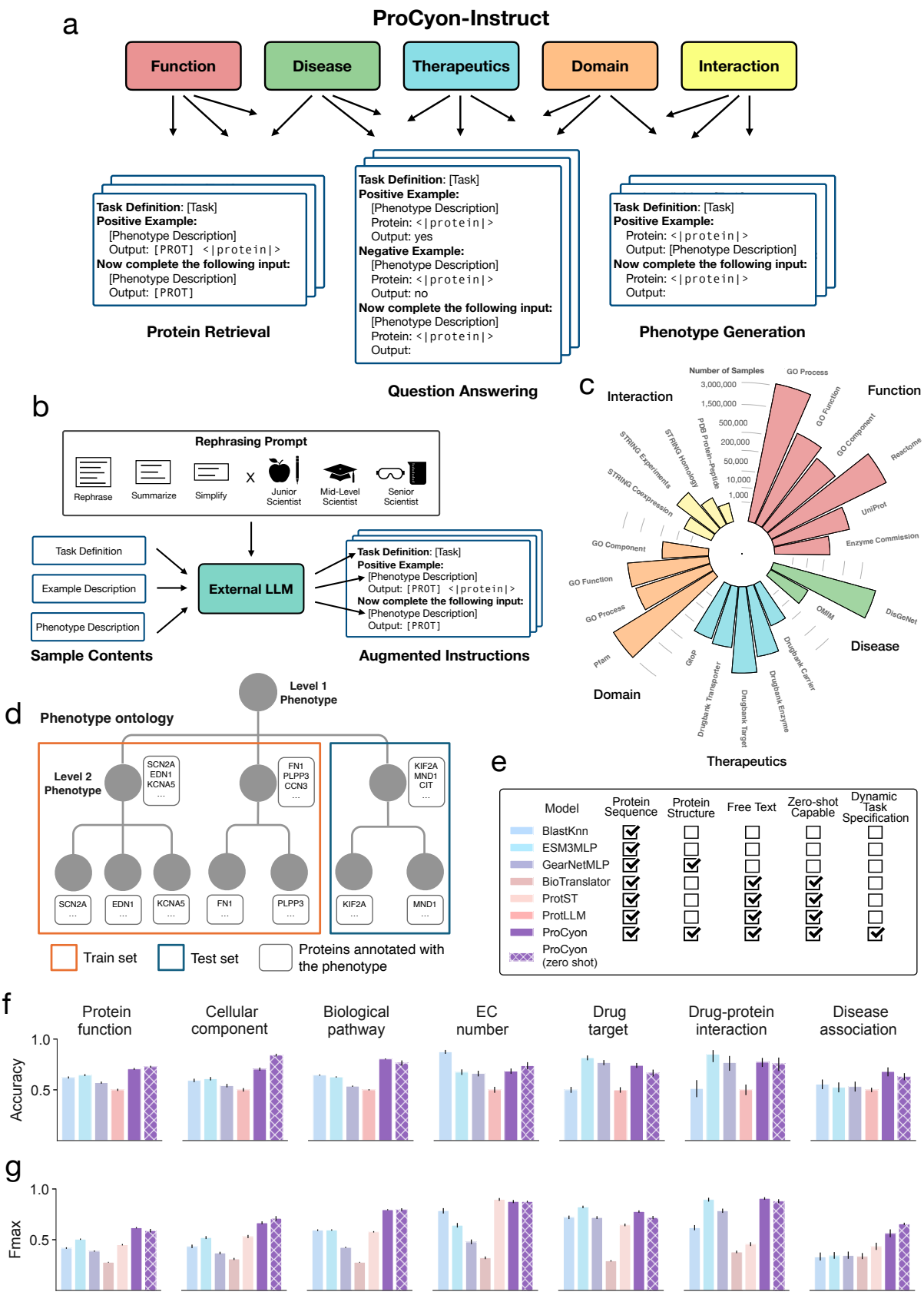

a
ProCyon-Instruct
Function
Disease
Therapeutics
Domain
Interaction
Task Definition: [Task]
Positive Example:
[Phenotype Description]
Output: [PROT] <|protein|>
Now complete the following input:
[Phenotype Description]
Output: [PROT]
Protein Retrieval
Task Definition: [Task]
Positive Example:
[Phenotype Description]
Protein: <|protein|>
Output: yes
Negative Example:
[Phenotype Description]
Protein: <|protein|>
Output: no
Now complete the following input:
[Phenotype Description]
Protein: <|protein|>
Output:
Question Answering
Task Definition: [Task]
Positive Example:
Protein: <|protein|>
Output: [Phenotype Description]
Now complete the following input:
Protein: <|protein|>
Output:
Phenotype Generation
b
Rephrasing Prompt
Rephrase
Summarize
Simplify
Junior Scientist
Mid-Level Scientist
Senior Scientist
Task Definition
Example Description
Phenotype Description
Sample Contents
External LLM
Augmented Instructions
c
Number of Samples
3,000,000
1,500,000
500,000
200,000
50,000
10,000
1,000
GO Process
GO Function
GO Component
Reactome
UniProt
Enzyme Commission
OMIM
DisGeNet
Drugbank Carrier
Drugbank Enzyme
Drugbank Target
Drugbank Transporter
GloP
Pfam
GO Process
GO Function
GO Component
STRING Coexpression
STRING Experiments
STRING Homology
PDB Protein-Peptide
Interaction
Function
Disease
Domain
Therapeutics
d
Phenotype ontology
Level 1 Phenotype
Level 2 Phenotype
SCN2A EDN1 KCNA5 ...
FN1 PLPP3 CCN3 ...
KIF2A MND1 CIT ...
SCN2A ...
EDN1 ...
KCNA5 ...
FN1 ...
PLPP3 ...
KIF2A ...
MND1 ...
Train set
Test set
Proteins annotated with the phenotype
e
Model
Protein Sequence
Protein Structure
Free Text
Zero-shot Capable
Dynamic Task Specification
BlastKnn
ESM3MLP
GearNetMLP
BioTranslator
ProtST
ProtLLM
ProCyon
ProCyon (zero shot)
f
Protein function
Cellular component
Biological pathway
EC number
Drug target
Drug-protein interaction
Disease association
Accuracy
1.0
0.5
g
Fmax
1.0
0.5

Figure 2

Figure 2: PROCYON-INSTRUCT dataset and benchmarking PROCYON a) Instruction tuning templates for protein-text pairs from each database in PROCYON-INSTRUCT. Samples in each database can be transformed into an instruction input via the presented instruction templates that formulate protein-text pairs into either a retrieval prompt, a question answering prompt, or a phenotype generation prompt. **b)** Pipeline for rephrasing descriptions and definitions. Using the templates shown in (a), we take the Task Definition, Example Description ("phenotype description" after the positive and negative examples), and Phenotype Description, input them to an external, proprietary LLM (GPT-4o), and then insert these rephrased versions into the instructions. Note that only the retrieval instruction is shown, but the same processed is repeated for all three types of tasks. We prompt the LLM with a rephrasing level and expertise level as discussed in Methods Sec. 3.1. **c)** After rephrasing augmentation, the PROCYON-INSTRUCT dataset has 33,899,528 total instructions. Here we show sample numbers for each database, where one sample consists of a protein-phenotype pair that constitutes an instruction. Note that these samples can be expanded into instructions (a), which is a different metric. We include data from five main knowledge domains: function, disease, therapeutics, domains, and interaction. Shown are the number of samples across each knowledge domain for each individual database in that knowledge domain. **d)** Illustration of our dataset splitting strategy for each database. To minimize the direct leakage of information, we utilize the hierarchical ontology of phenotypes (e.g., Gene Ontology, ATC code, Enzyme Commissions numbers), and derive the test set by holding out subtrees of phenotypes, along with all protein associations involving these phenotypes. For diseases (DisGeNet) specifically, we utilize the molecular similarity between diseases (defined as the Jaccard similarity between their respective associated proteins) and hold out all diseases with high molecular similarity to certain "seed" diseases (see Methods Sec. 1 for details). **e)** Comparison of PROCYON and baseline models across input modalities and capabilities. "Zero-shot capable" refers to the ability to generate predictions for novel phenotypes never seen in the training set. "Dynamic task specification" refers to the ability of the model to perform tasks beyond the tasks it was explicitly trained on. **f, g)** Performance of models across a variety of biologically relevant tasks when framed as question answering (**f**) or retrieval (**g**) problems. Performance measured by binary accuracy and Fmax, respectively. Bars indicate 95% confidence intervals.

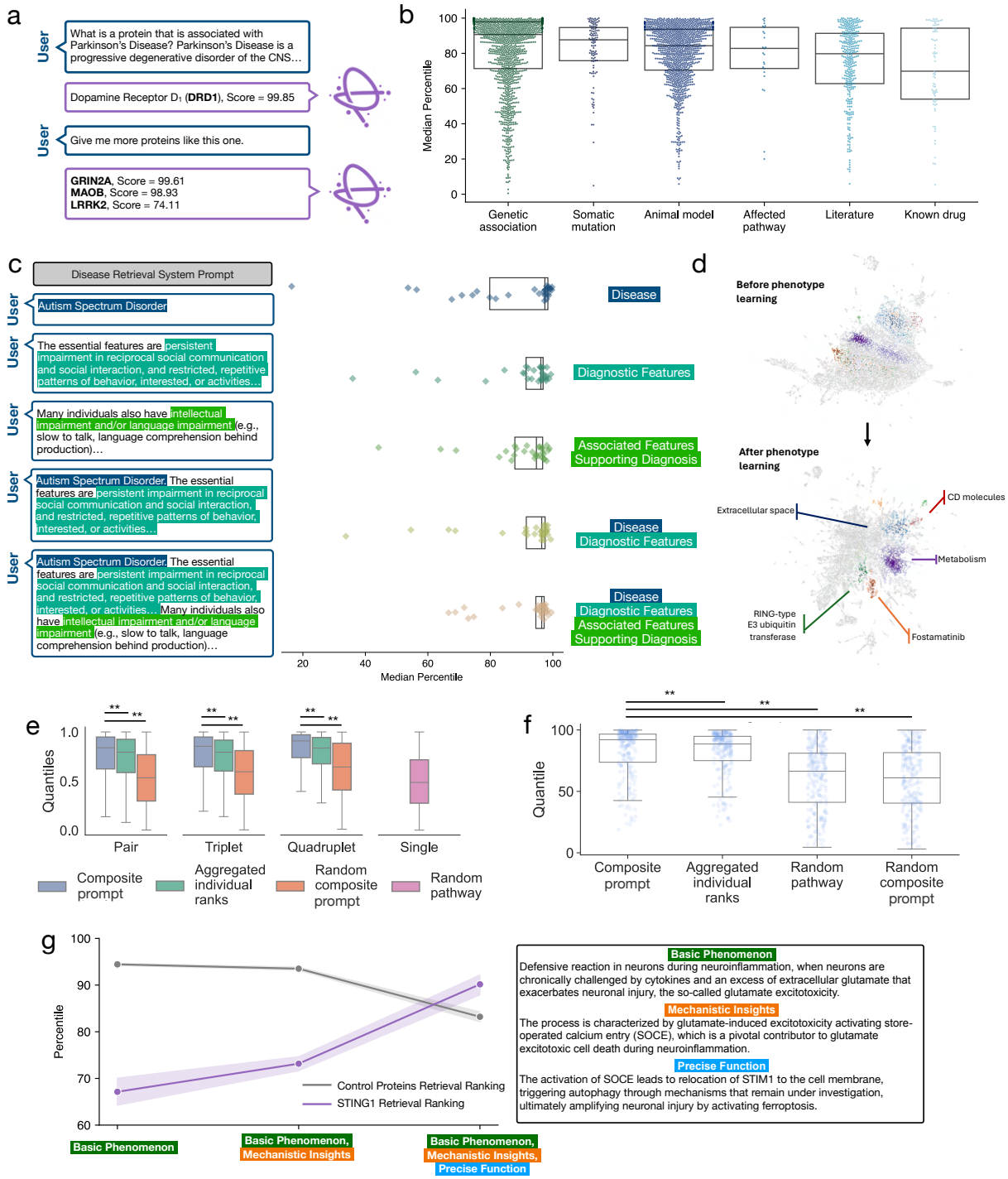

a
User
What is a protein that is associated with Parkinson's Disease? Parkinson's Disease is a progressive degenerative disorder of the CNS…
Dopamine Receptor D1 (DRD1), Score = 99.85
User
Give me more proteins like this one.
GRIN2A, Score = 99.61
MAOB, Score = 98.93
LRRK2, Score = 74.11
b
Median Percentile
Genetic association
Somatic mutation
Animal model
Affected pathway
Literature
Known drug
c
Disease Retrieval System Prompt
Autism Spectrum Disorder
The essential features are persistent impairment in reciprocal social communication and social interaction, and restricted, repetitive patterns of behavior, interested, or activities…
Many individuals also have intellectual impairment and/or language impairment (e.g., slow to talk, language comprehension behind production)…
Autism Spectrum Disorder. The essential features are persistent impairment in reciprocal social communication and social interaction, and restricted, repetitive patterns of behavior, interested, or activities…
Autism Spectrum Disorder. The essential features are persistent impairment in reciprocal social communication and social interaction, and restricted, repetitive patterns of behavior, interested, or activities… Many individuals also have intellectual impairment and/or language impairment (e.g., slow to talk, language comprehension behind production)…
Disease
Diagnostic Features
Associated Features Supporting Diagnosis
Disease Diagnostic Features
Disease Diagnostic Features Associated Features Supporting Diagnosis
Median Percentile
d
Before phenotype learning
After phenotype learning
Extracellular space
CD molecules
Metabolism
RING-type E3 ubiquitin transferase
Fostamatinib
e
Quantiles
Pair
Triplet
Quadruplet
Single
Composite prompt
Aggregated individual ranks
Random composite prompt
Random pathway
f
Quantile
Composite prompt
Aggregated individual ranks
Random pathway
Random composite prompt
g
Percentile
Control Proteins Retrieval Ranking
STING1 Retrieval Ranking
Basic Phenomenon
Basic Phenomenon, Mechanistic Insights
Basic Phenomenon, Mechanistic Insights, Precise Function
Basic Phenomenon
Defensive reaction in neurons during neuroinflammation, when neurons are chronically challenged by cytokines and an excess of extracellular glutamate that exacerbates neuronal injury, the so-called glutamate excitotoxicity.
Mechanistic Insights
The process is characterized by glutamate-induced excitotoxicity activating store-operated calcium entry (SOCE), which is a pivotal contributor to glutamate excitotoxic cell death during neuroinflammation.
Precise Function
The activation of SOCE leads to relocation of STIM1 to the cell membrane, triggering autophagy through mechanisms that remain under investigation, ultimately amplifying neuronal injury by activating ferroptosis.

Figure 3

Figure 3: PROCYON retrieval capabilities. a) Example of a retrieval prompt to PROCYON, where the user requests proteins implicated in Parkinson's disease (PD). The model retrieves relevant PD proteins, including Dopamine Receptor D_1 (DRD1), Glutamate [NMDA] receptor subunit epsilon-1 (GRIN2A), Monoamine oxidase B (MAOB), and Leucine-rich repeat kinase 2 (LRRK2). Scores shown are quantiles out of all human proteins available based on the ranking generated by PROCYON. **b)** External validation across different sources. The figure shows the median percentile retrieval scores derived for each disease across different ground truth sources (e.g., genetic associations, somatic mutations, animal models, affected pathway, literature and known drug). The boxplot shows the median score across all diseases per source. **c)** Protein retrieval with different disease descriptions. The right side of the panel shows the median percentile retrieval scores derived for each disease across different input prompts where the information is gradually increased; in the left side, an example text is provided for each type of prompt (Disease, Diagnostics, Associated Features supporting Diagnosis and their combinations) in case of Autistic Spectrum Disorder. **d)** Density of proteins with the same annotation across multiple databases before and after multimodal tuning. The annotations with the highest or second highest number of proteins per database are visualized in the figure. The databases included are: reactome (purple), EC (green), GTOP (red), GO (blue) and drugbank (orange). The annotations visualized are: Fostamatinib (Drugbank), Metabolism (Reactome), CD molecules (GTOP), Extracellular space (GO), and RING-type E3 ubiquitin transferase (EC). The top embedding figure was generated using ESM-2 embeddings and corresponds to a sequence-only embedding space (also the initial embedding space of PROCYON). The bottom embedding figure was generated using the PROCYON embeddings and corresponds to the multimodal tuned embedding space. **e)** Quantiles of potentially pleiotropic proteins ranked by PROCYON using compositional prompts (composite prompt) composed of two (pair), three (triplet), and four (quadruplet) pathways, in comparison with rank aggregation of prompts of individual pathways (aggregated individual ranks). "Random composite prompt" ranks the same set of potentially pleotropic proteins but samples random combinations of pathways. "Random pathway" ranks the same set of proteins but with a prompt of one randomly sampled pathway. * p-value < 0.05, ** p-value < 0.001, Wilcoxon signed-rank test. **f)** Quantiles of proteins involved in signaling pathway crosstalk ranked by PROCYON with compositional prompts (composite prompt) composed of the two signaling pathways, in comparison with rank aggregation of prompts of individual signaling pathways (aggregated individual ranks). "Random composite prompt" ranks the same set of proteins but with prompts composed of two randomly sampled signaling pathways. "Random pathway" ranks the same set of proteins but with prompts of a randomly sampled signaling pathway. * p-value < 0.05, ** p-value < 0.001, Wilcoxon signed-rank test. **g)** Quantiles of STING1 (UniProt: Q86WV6) when querying PROCYON with different levels of evidence about the potential involvement of STING1 in neuronal inflammatory stress response, in comparison to the quantiles of control proteins that are expressed in the central nervous system (CNS) and rank top 10% when querying PROCYON with the basic phenomenon level of evidence. The evidence descriptions are manually curated based on [54] (see Methods Sec. 4.6).

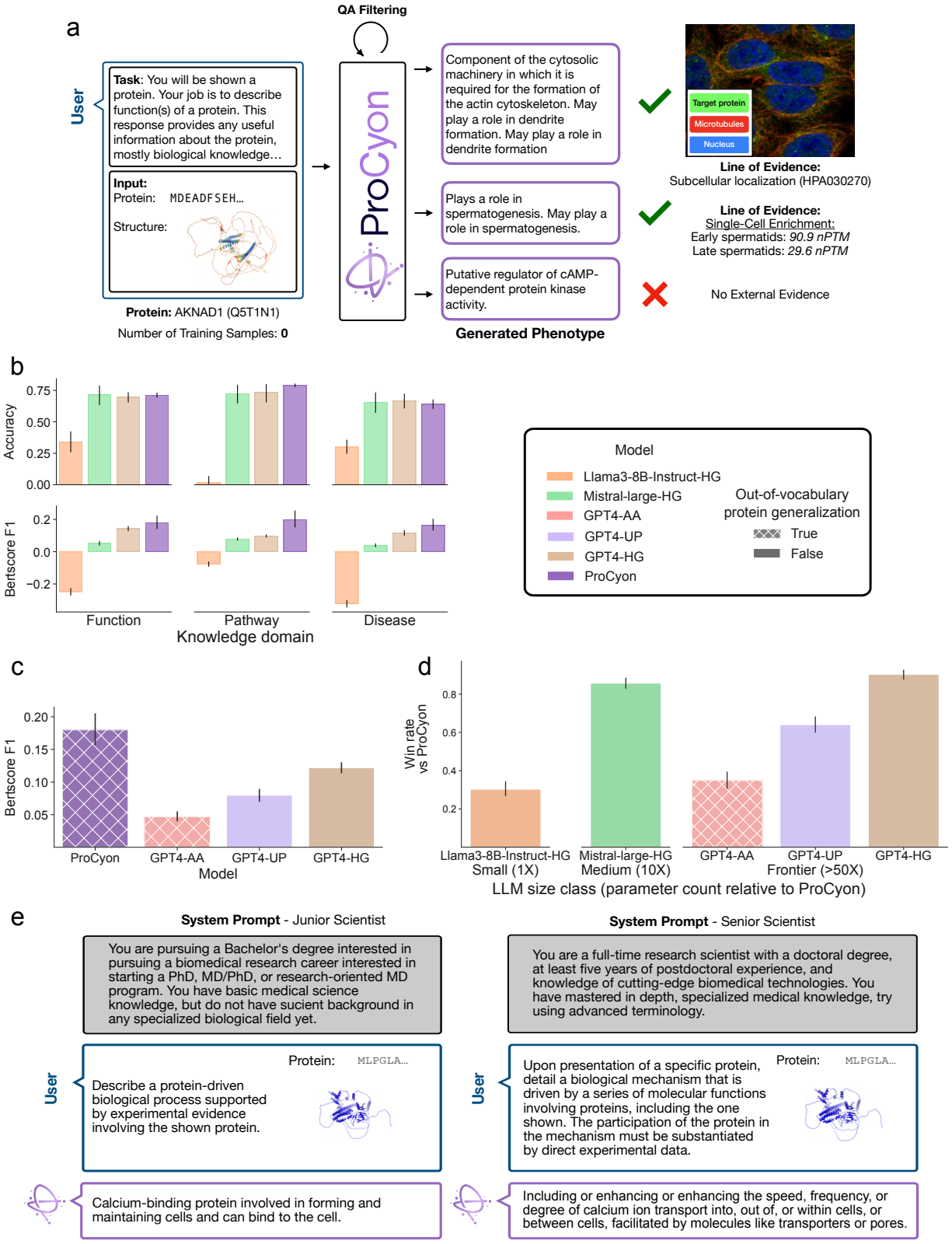

a
QA Filtering
User
Task: You will be shown a protein. Your job is to describe function(s) of a protein. This response provides any useful information about the protein, mostly biological knowledge…
Input:
Protein: MDEADFSEH…
Structure:
Protein: AKNAD1 (Q5T1N1)
Number of Training Samples: 0
ProCyon
Component of the cytosolic machinery in which it is required for the formation of the actin cytoskeleton. May play a role in dendrite formation. May play a role in dendrite formation
Plays a role in spermatogenesis. May play a role in spermatogenesis.
Putative regulator of cAMP-dependent protein kinase activity.
Generated Phenotype
Target protein
Microtubules
Nucleus
Line of Evidence:
Subcellular localization (HPA030270)
Line of Evidence:
Single-Cell Enrichment:
Early spermatids: 90.9 nPTM
Late spermatids: 29.6 nPTM
No External Evidence
b
Accuracy
Bertscore F1
Function
Pathway
Disease
Knowledge domain
Model
Llama3-8B-Instruct-HG
Mistral-large-HG
GPT4-AA
GPT4-UP
GPT4-HG
ProCyon
Out-of-vocabulary protein generalization
True
False
c
Bertscore F1
ProCyon
GPT4-AA
GPT4-UP
GPT4-HG
Model
d
Win rate vs ProCyon
Llama3-8B-Instruct-HG
Small (1X)
Mistral-large-HG
Medium (10X)
GPT4-AA
GPT4-UP
GPT4-HG
Frontier (>50X)
LLM size class (parameter count relative to ProCyon)
e
System Prompt - Junior Scientist
You are pursuing a Bachelor's degree interested in pursuing a biomedical research career interested in starting a PhD, MD/PhD, or research-oriented MD program. You have basic medical science knowledge, but do not have sucient background in any specialized biological field yet.
User
Describe a protein-driven biological process supported by experimental evidence involving the shown protein.
Protein: MLPGLA...
Calcium-binding protein involved in forming and maintaining cells and can bind to the cell.
System Prompt - Senior Scientist
You are a full-time research scientist with a doctoral degree, at least five years of postdoctoral experience, and knowledge of cutting-edge biomedical technologies. You have mastered in depth, specialized medical knowledge, try using advanced terminology.
User
Upon presentation of a specific protein, detail a biological mechanism that is driven by a series of molecular functions involving proteins, including the one shown. The participation of the protein in the mechanism must be substantiated by direct experimental data.
Protein: MLPGLA...
Including or enhancing or enhancing the speed, frequency, or degree of calcium ion transport into, out of, or within cells, or between cells, facilitated by molecules like transporters or pores.

Figure 4

Figure 4: PROCYON phenotype generation capabilities. a) For this experiment, we prompt PROCYON to generate a UniProt-style description for AKNAD1, a poorly-characterized protein that does not appear in PROCYON training data. After inputting the sequence and structure of this protein, we perform QA filtering to derive the top-3 generated phenotypes for this protein. Two of the three final phenotypes are supported by evidence in the Human Protein Atlas through subcellular localization assays and scRNA-seq studies (see Methods Sec. 4.7 for more information). **b)** PROCYON'S performance in question answering and text generation is greater than or equal to that of state-of-the-art text-only LLMs at a range of model sizes and across knowledge domains. Text generation performance measured using n-gram semantic similarity compared to reference texts. **c)** Text-only LLM performance strongly depends on how proteins are represented (-AA, amino acid sequence; -UP, UniProt ID; -HG, HGNC gene name). **d)** Holistic evaluation of generated text measured by LLM (Claude-3.5-Sonnet) preference. PROCYON outperforms similarly sized models, but strong open-ended text generation requires large-scale models. **e)** Expertise rephrasing by PROCYON using a system prompt to specify the background of the user. The shown example controls for the protein, type of prompt, and content of the output. The junior scientist output (left) shows a simpler, more straightforward output while the senior scientist output (right) is more detailed and nuanced.

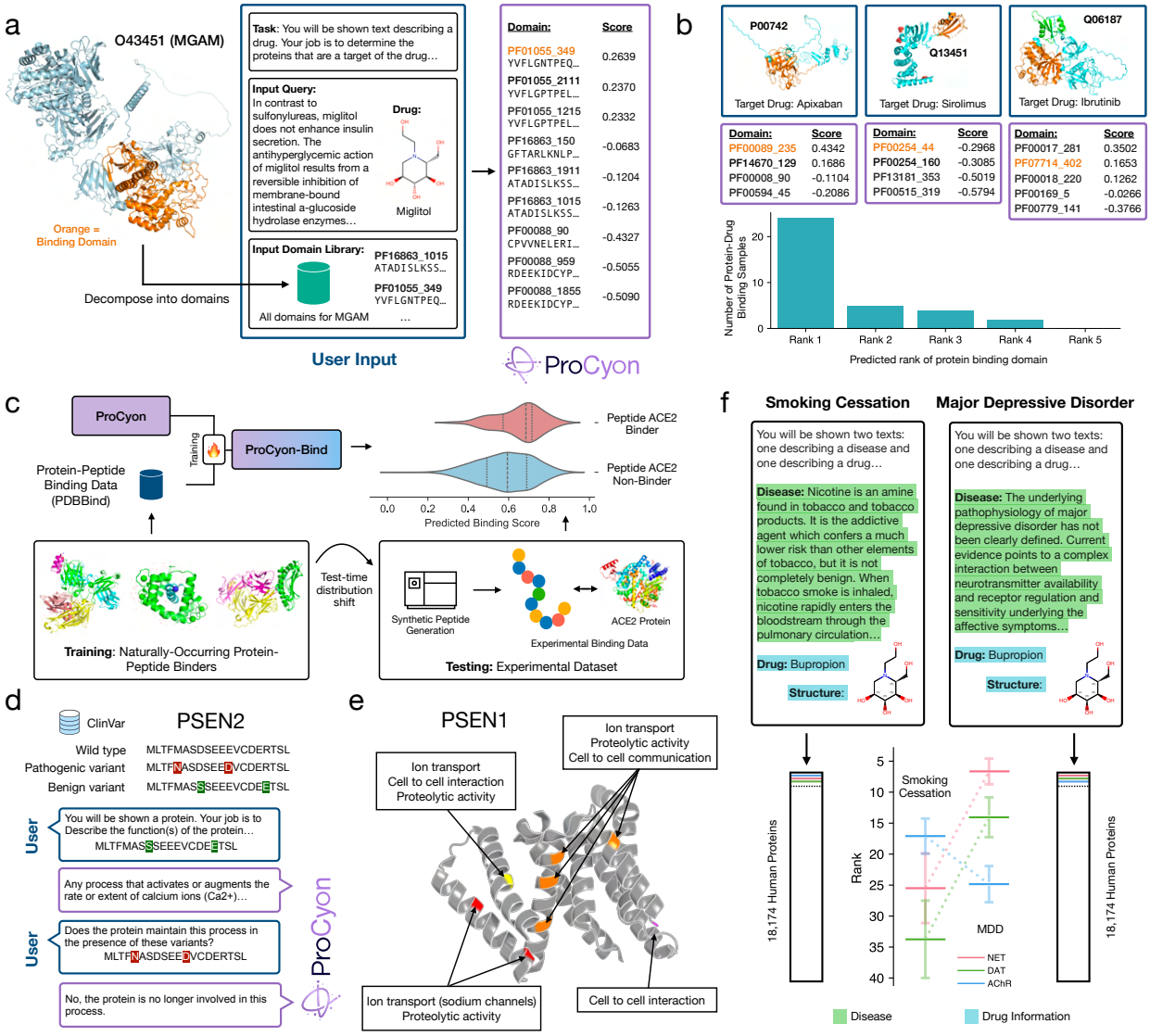

a
O43451 (MGAM)
Orange = Binding Domain
Decompose into domains
Task: You will be shown text describing a drug. Your job is to determine the proteins that are a target of the drug…
Input Query:
In contrast to sulfonylureas, miglitol does not enhance insulin secretion. The antihyperglycemic action of miglitol results from a reversible inhibition of membrane-bound intestinal a-glucoside hydrolase enzymes…
Drug:
Miglitol
Input Domain Library:
PF16863_1015 ATADISLKSS…
PF01055_349 YVFLGNTPEQ…
All domains for MGAM …
Domain: Score
PF01055_349 YVFLGNTPEQ… 0.2639
PF01055_2111 YVFLGPTPEL… 0.2370
PF01055_1215 YVFLGPTPEL… 0.2332
PF16863_150 GFTARLKNLP… -0.0683
PF16863_1911 ATADISLKSS… -0.1204
PF16863_1015 ATADISLKSS… -0.1263
PF00088_90 CPVVNELERI… -0.4327
PF00088_959 RDEEKIDCYP… -0.5055
PF00088_1855 RDEEKIDCYP… -0.5090
User Input
ProCyon
b
P00742
Target Drug: Apixaban
Q13451
Target Drug: Sirolimus
Q06187
Target Drug: Ibrutinib
Domain: Score
PF00089_235 0.4342
PF14670_129 0.1686
PF00008_90 -0.1104
PF00594_45 -0.2086
Domain: Score
PF00254_44 -0.2968
PF00254_160 -0.3085
PF13181_353 -0.5019
PF00515_319 -0.5794
Domain: Score
PF00017_281 0.3502
PF07714_402 0.1653
PF00018_220 0.1262
PF00169_5 -0.0266
PF00779_141 -0.3766
Number of Protein-Drug Binding Samples
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
Predicted rank of protein binding domain
c
ProCyon
Training
ProCyon-Bind
Protein-Peptide Binding Data (PDBBind)
Peptide ACE2 Binder
Peptide ACE2 Non-Binder
Predicted Binding Score
Training: Naturally-Occurring Protein-Peptide Binders
Test-time distribution shift
Synthetic Peptide Generation
Experimental Binding Data
ACE2 Protein
Testing: Experimental Dataset
d
ClinVar
PSEN2
Wild type MLTFMASDSEEEVCDERTSL
Pathogenic variant MLTFNASDSEEDVCDERTSL
Benign variant MLTFMASSSEEEVCDEETSL
User
You will be shown a protein. Your job is to Describe the function(s) of the protein…
MLTFMASSSEEEVCDEETSL
Any process that activates or augments the rate or extent of calcium ions (Ca2+)…
User
Does the protein maintain this process in the presence of these variants?
MLTFNASDSEEDVCDERTSL
No, the protein is no longer involved in this process.
ProCyon
e
PSEN1
Ion transport
Cell to cell interaction
Proteolytic activity
Ion transport
Proteolytic activity
Cell to cell communication
Ion transport (sodium channels)
Proteolytic activity
Cell to cell interaction
f
Smoking Cessation
Major Depressive Disorder
You will be shown two texts: one describing a disease and one describing a drug…
Disease: Nicotine is an amine found in tobacco and tobacco products. It is the addictive agent which confers a much lower risk than other elements of tobacco, but it is not completely benign. When tobacco smoke is inhaled, nicotine rapidly enters the bloodstream through the pulmonary circulation…
Drug: Bupropion
Structure:
You will be shown two texts: one describing a disease and one describing a drug…
Disease: The underlying pathophysiology of major depressive disorder has not been clearly defined. Current evidence points to a complex interaction between neurotransmitter availability and receptor regulation and sensitivity underlying the affective symptoms…
Drug: Bupropion
Structure:
18,174 Human Proteins
Rank
Smoking Cessation
MDD
NET
DAT
AChR
18,174 Human Proteins
Disease
Drug Information

Figure 5

Figure 5: Zero-shot task transfer with PROCYON. a) An example of prompting PROCYON to retrieve domains of O43451 (MGAM) to which the drug Miglitol might bind. We first decompose the protein into its domains and use these domains as reference library against which PROCYON can prioritize. Then, we instruct PROCYON to complete a drug target prioritization task, as a result of which PROCYON prioritizes the domains of the protein present in the provided reference library. PROCYON correctly prioritized as number 1 out of 9 potential domains the binding domain for Miglitol. **b)** We use a drug binding dataset that specifies the residues with which the small molecule interacts with the protein. We decompose into domains each of the 35 protein-drug pairs and record the ranked list of the binding domains. Shown are three examples of this type of prompt, where the orange highlight denotes the correct binding domain in the ranked list. PROCYON ranks as number 1 the correct binding domain for P00742 and Q13451 while for Q06187, PROCYON ranks it second out of 5. The barplot shows the number of samples for which PROCYON ranks the correct binding domain at each given rank; in 24 / 35 (68.6%) samples, PROCYON correctly identifies the binding domain. **c)** In the top-left corner, we show a schematic of the finetuning process for PROCYON-BIND, which is a finetuned version of PROCYON-FULL on naturally-occurring protein-peptide binders from the Protein Data Bank (PDB), of which a schematic is shown in the bottom-left corner. We then test PROCYON-BIND on an experimental dataset of synthetic peptides which have been screened for binding to the ACE2 protein, which is shown in the bottom-right corner. In the top-right corner, we see that there is clear separation between the predicted binding score for binders versus non-binders; this difference is statistically significant under a Mann-Whitney U test ($p = 1.04\text{e-}4$). **d)** PROCYON generates a phenotype description for a sequence of PSEN2 with a benign variant for Alzheimer's Disease derived from ClinVar, where it is correctly described how the variant is involved in calcium transport. However, PROCYON does not generate the same phenotype when a pathogenic variant is provided to the model. **e)** The PSEN1 structure is shown with mutated positions highlighted in red, yellow, orange, and pink, indicating mutation groups based on PROCYON-generated functional phenotypes. Each color corresponds to a group of mutations with functional keywords describing biological activities such as ion transport, cell-to-cell interaction, proteolytic activity, and sodium channel transport described in the captions common to the mutations in each group. **f)** Leveraging PROCYON to perform indication-specific protein target retrieval. The top of the figure presents two inputs to the model. On the left, the input consists of a novel task definition, the description of nicotine addiction, and the name and structure of Bupropion; on the right, the same task definition is provided but with the description of major depressive disorder (MDD) and the name and structure of Bupropion. After retrieval of all protein targets, PROCYON ranks the appropriate targets in the top 40/18,174 human proteins. For smoking cessation, the appropriate target, AChR, is ranked at 17.1 on average while NET is ranked at 25.5 (p =0.1153 for Wilcoxon signed rank test). For MDD, the appropriate target, NET, is ranked as 6.65 on average while AChR is ranked at 24.85 ($p = 1.974\text{e-}4$ for Wilcoxon signed-rank test). Shown are 95% confidence intervals derived by perturbing the words in the input prompt.

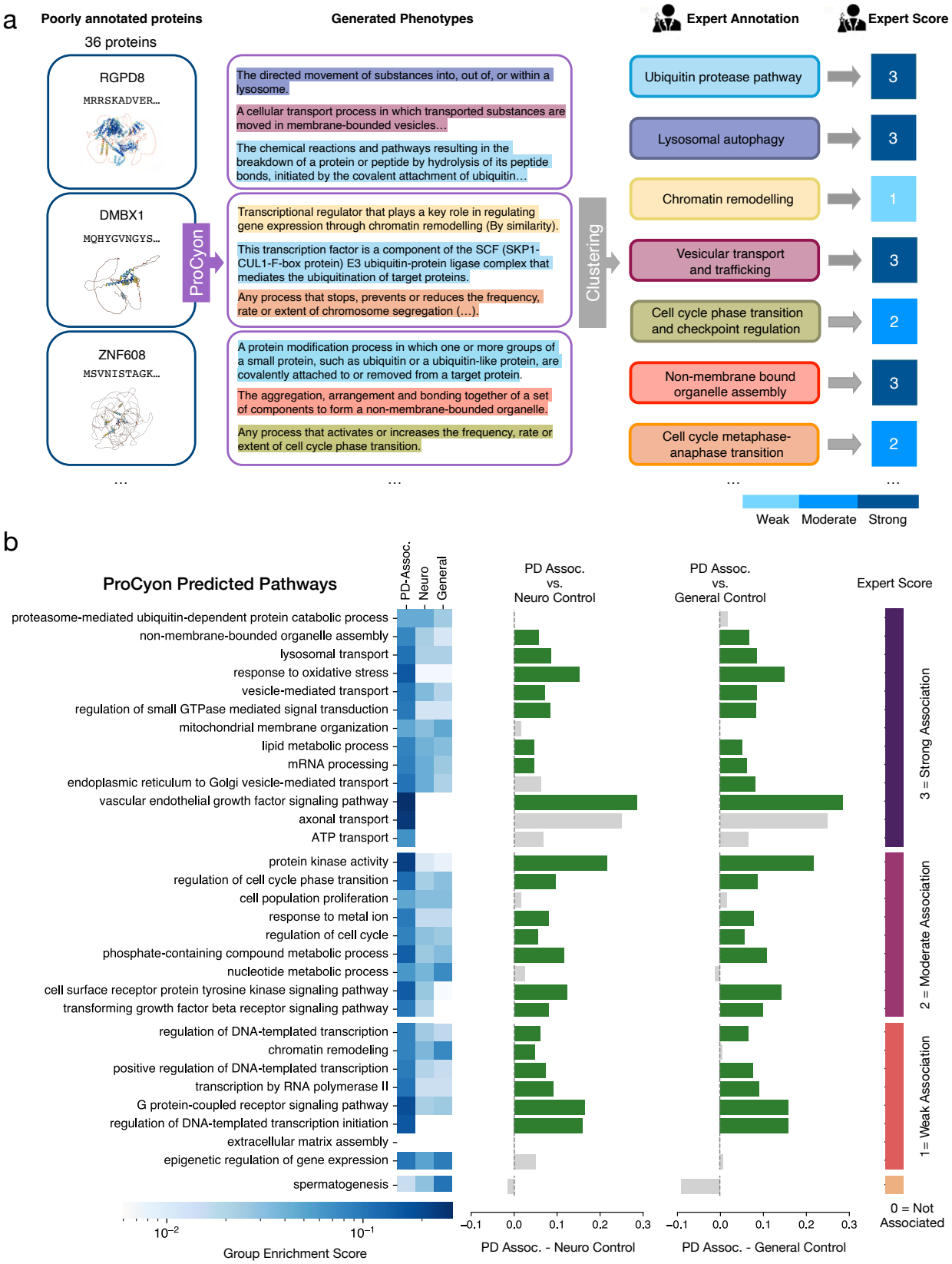
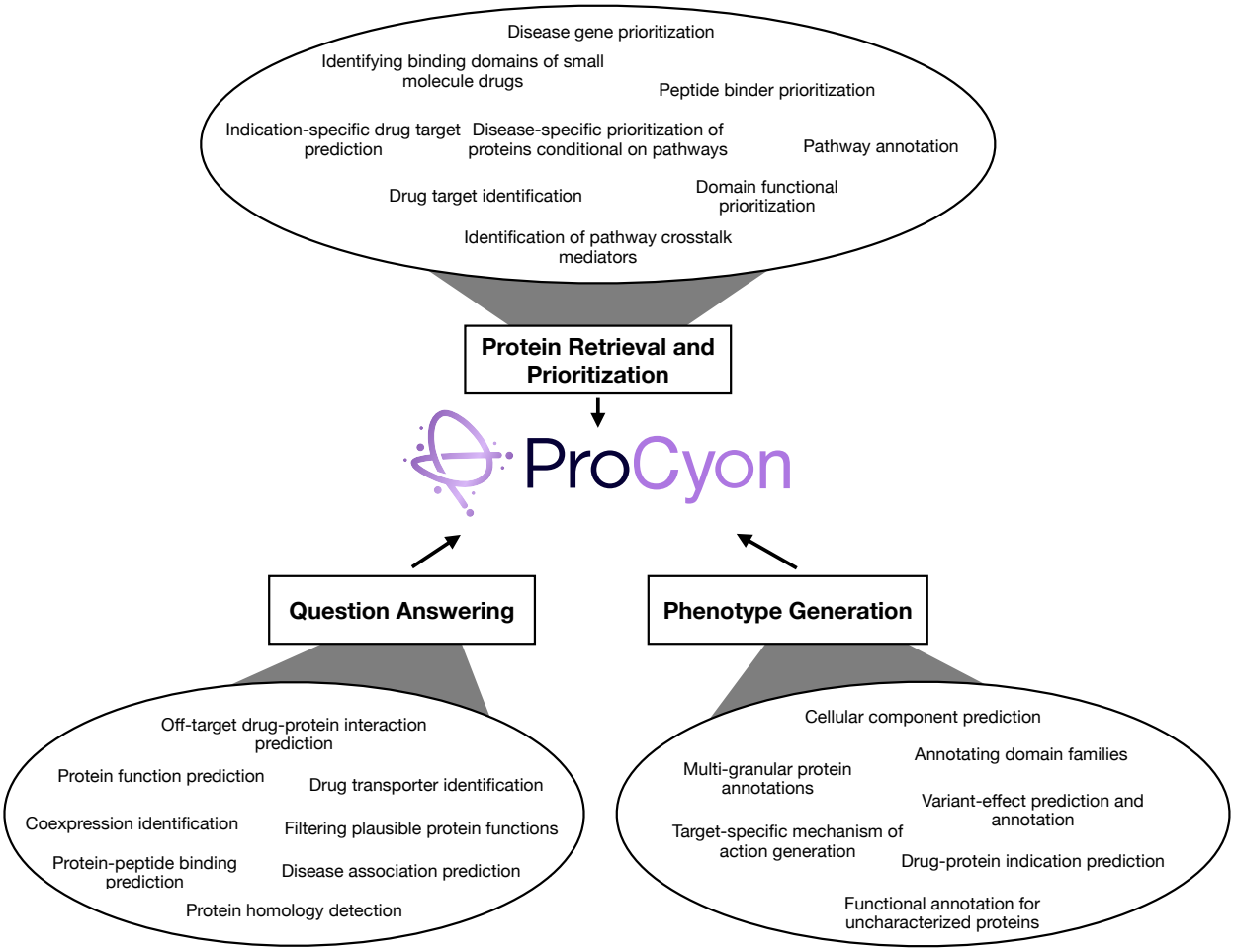
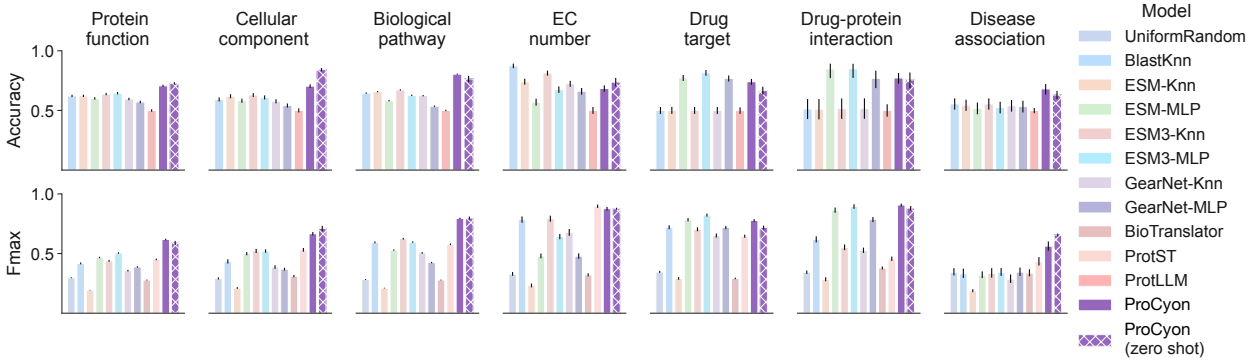

a
Poorly annotated proteins
36 proteins
RGPD8
MRRSKADVER...
DMBX1
MQHYGVNGYS...
ZNF608
MSVNISTAGK...
ProCyon
Generated Phenotypes
The directed movement of substances into, out of, or within a lysosome.
A cellular transport process in which transported substances are moved in membrane-bounded vesicles...
The chemical reactions and pathways resulting in the breakdown of a protein or peptide by hydrolysis of its peptide bonds, initiated by the covalent attachment of ubiquitin...
Transcriptional regulator that plays a key role in regulating gene expression through chromatin remodelling (By similarity).
This transcription factor is a component of the SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex that mediates the ubiquitination of target proteins.
Any process that stops, prevents or reduces the frequency, rate or extent of chromosome segregation (...).
A protein modification process in which one or more groups of a small protein, such as ubiquitin or a ubiquitin-like protein, are covalently attached to or removed from a target protein.
The aggregation, arrangement and bonding together of a set of components to form a non-membrane-bounded organelle.
Any process that activates or increases the frequency, rate or extent of cell cycle phase transition.
Clustering
Expert Annotation
Expert Score
Ubiquitin protease pathway 3
Lysosomal autophagy 3
Chromatin remodelling 1
Vesicular transport and trafficking 3
Cell cycle phase transition and checkpoint regulation 2
Non-membrane bound organelle assembly 3
Cell cycle metaphase-anaphase transition 2
Weak Moderate Strong
b
ProCyon Predicted Pathways
PD-Assoc.
Neuro
General
proteasome-mediated ubiquitin-dependent protein catabolic process
non-membrane-bounded organelle assembly
lysosomal transport
response to oxidative stress
vesicle-mediated transport
regulation of small GTPase mediated signal transduction
mitochondrial membrane organization
lipid metabolic process
mRNA processing
endoplasmic reticulum to Golgi vesicle-mediated transport
vascular endothelial growth factor signaling pathway
axonal transport
ATP transport
protein kinase activity
regulation of cell cycle phase transition
cell population proliferation
response to metal ion
regulation of cell cycle
phosphate-containing compound metabolic process
nucleotide metabolic process
cell surface receptor protein tyrosine kinase signaling pathway
transforming growth factor beta receptor signaling pathway
regulation of DNA-templated transcription
chromatin remodeling
positive regulation of DNA-templated transcription
transcription by RNA polymerase II
G protein-coupled receptor signaling pathway
regulation of DNA-templated transcription initiation
extracellular matrix assembly
epigenetic regulation of gene expression
spermatogenesis
10^{-2}
10^{-1}
Group Enrichment Score
PD Assoc. vs. Neuro Control
PD Assoc. vs. General Control
Expert Score
3 = Strong Association
2 = Moderate Association
1= Weak Association
0 = Not Associated
−0.1 0.0 0.1 0.2 0.3
PD Assoc. - Neuro Control
PD Assoc. - General Control

Figure 6

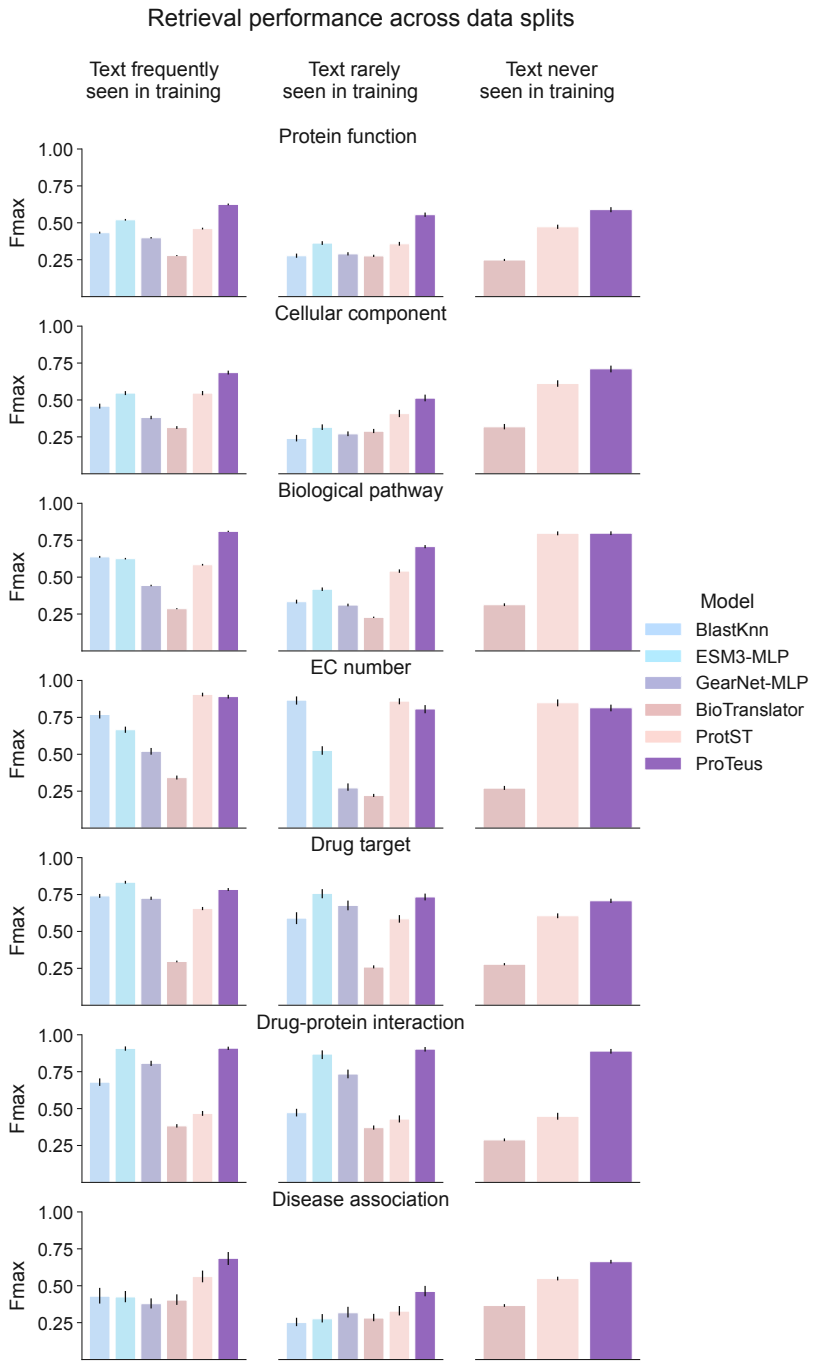
Figure 6: ProCyon on predicting functions of poorly-characterized proteins that play a role in Parkinson's disease. a) We designed a study of proteins with known association to PD but with poor functional annotation in literature. We generated captions with ProCyon, clustered them via manual and automatic methods, manually annotated the pathways identified by each cluster with the help of two independent experts, and provided the pathways to a physician-scientist who specializes in PD and related disorders, who scored their PD association on a scale of 0-3 (with 0 being no association and 3 strong association). **b)** An association study for the pathways selected by ProCyon. The leftmost heatmap presents an enrichment score (described in Methods Sec. 4.13) for each pathway compared to a large set of established PD-related proteins ("PD Assoc."), of "Neuro controls" proteins expressed in the nervous system but not in PD or other neurodegenerative conditions, and "General controls" proteins not expressed in neurological tissues. We then show differences between "PD Assoc.", "Neuro" and "General" by highlighting in green statistically-significant differences in enrichment scores across each pathway (gray is non-significant differences). Finally, we group pathways according to the PD association score provided by the expert, as shown in the right side of the panel.



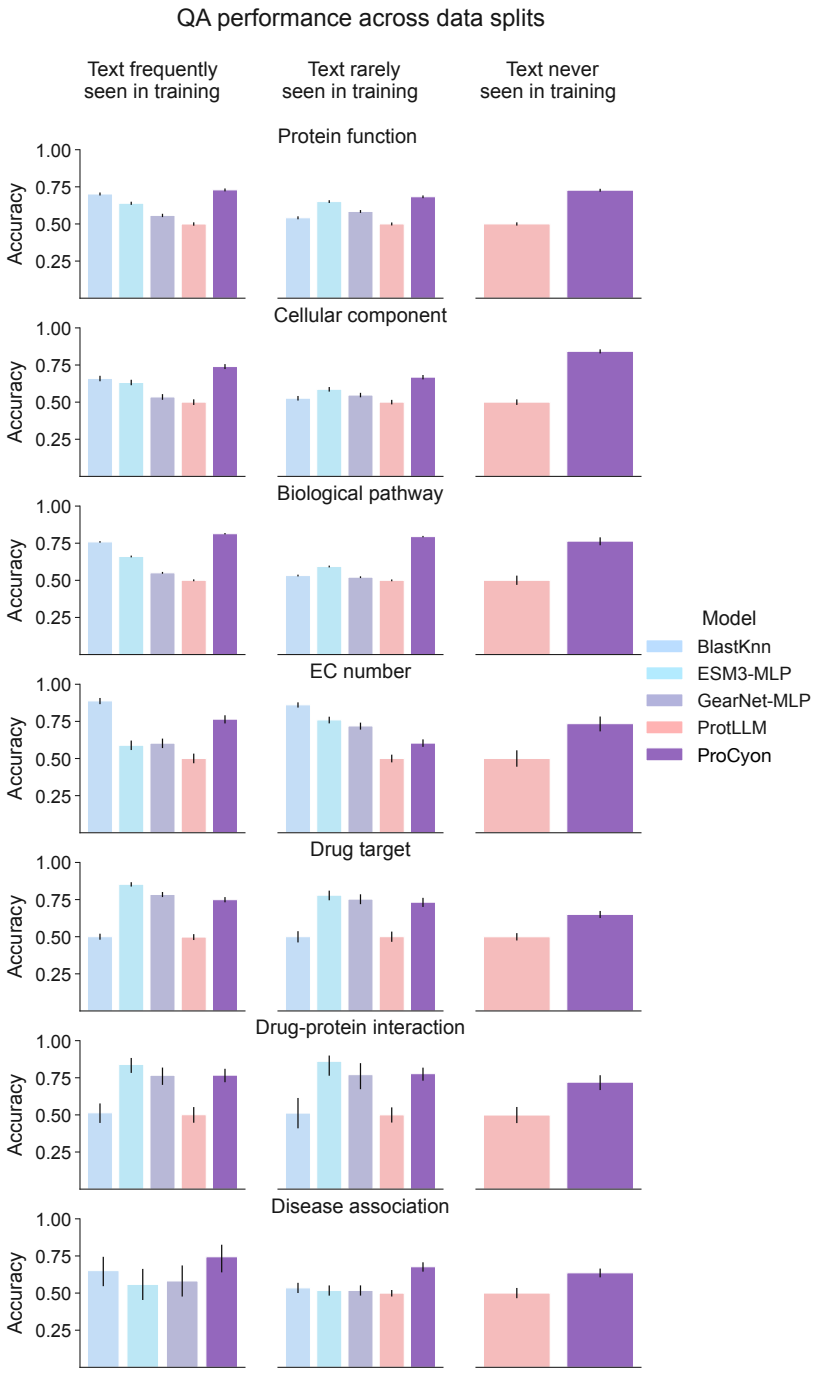
Extended Data Figure 1: PROCYON can accomplish an array of tasks that can be mapped into its core components mentioned in Figure 1. Here, we list some possible mappings for different tasks that PROCYON can perform; this is not an exhaustive list as PROCYON's free-text processing components allow for a wide array of tasks.



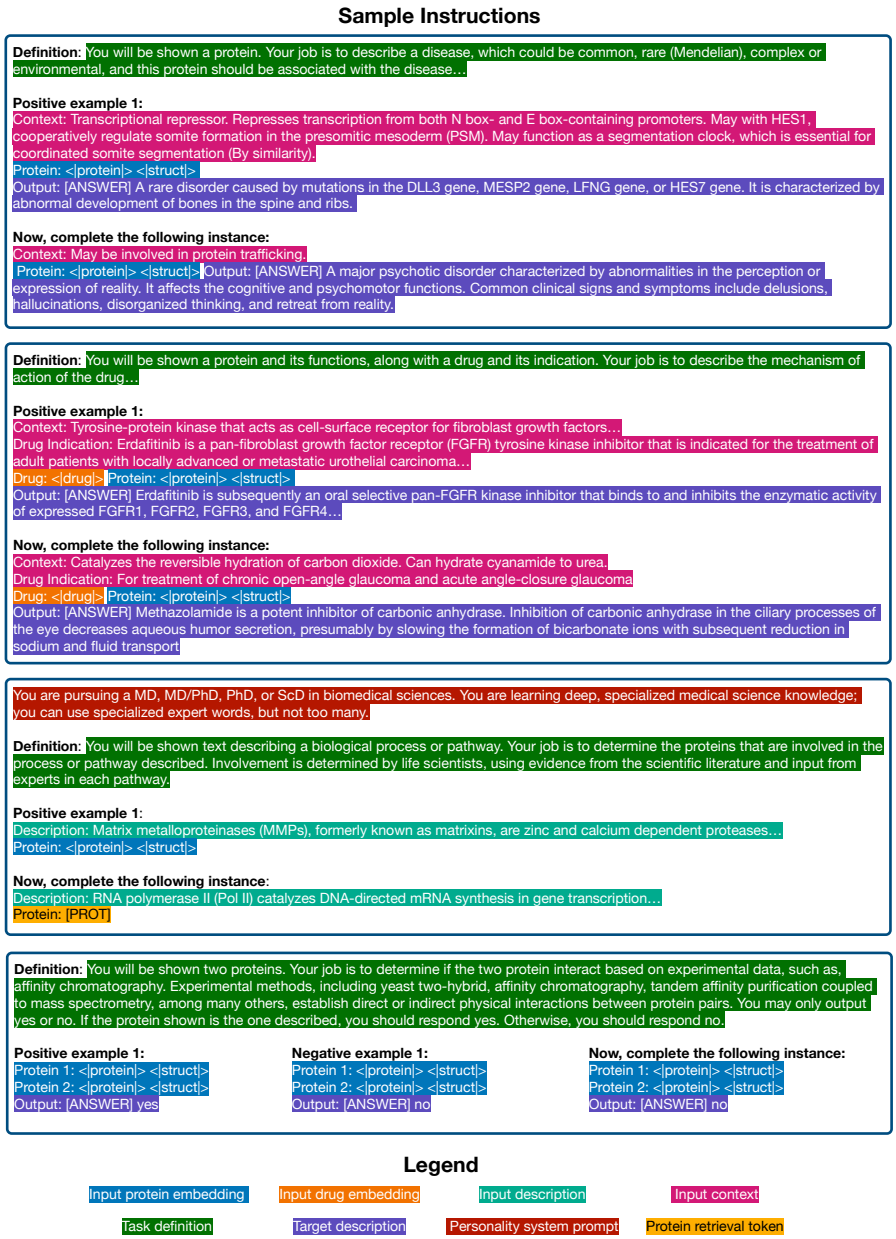
Extended Data Figure 2: Full suite of benchmarking results for PROCYON compared against various baselines and other models. The tasks in this benchmark capture common steps in biomedical research workflows, such as identifying a protein's function or relevant biological pathway, selecting drugs that target a given protein, or understanding a protein's disease associations. By defining tasks across knowledge domains, we ensure that our benchmark can probe each model's capacity to reason over a broad spectrum of biomedical knowledge.



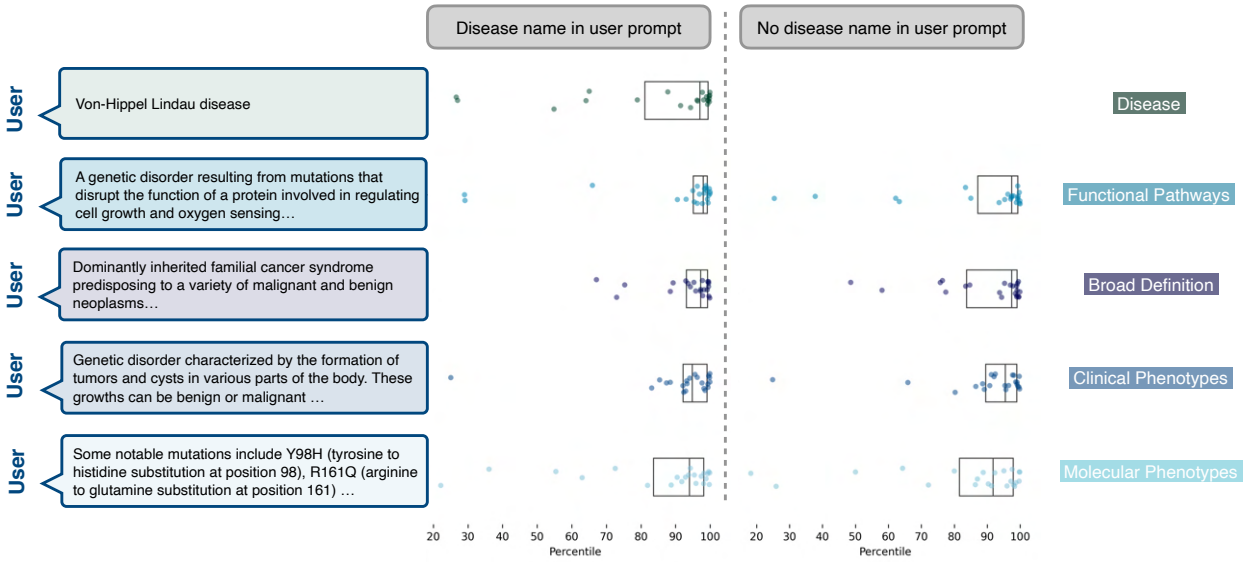
Extended Data Figure 3: Retrieval performance of PROCYON and other models across different dataset splits of PROCYON-INSTRUCT. As described in our methods, dataset splits are constructed rigorously to minimize information leakage across splits from similar phenotypes. Additionally, splits are constructed that allow us to test PROCYON'S performance on texts that are never seen in training. We note that PROCYON'S generalizability is shown by consistent performance across splits with increasingly novel texts compared to the training set.



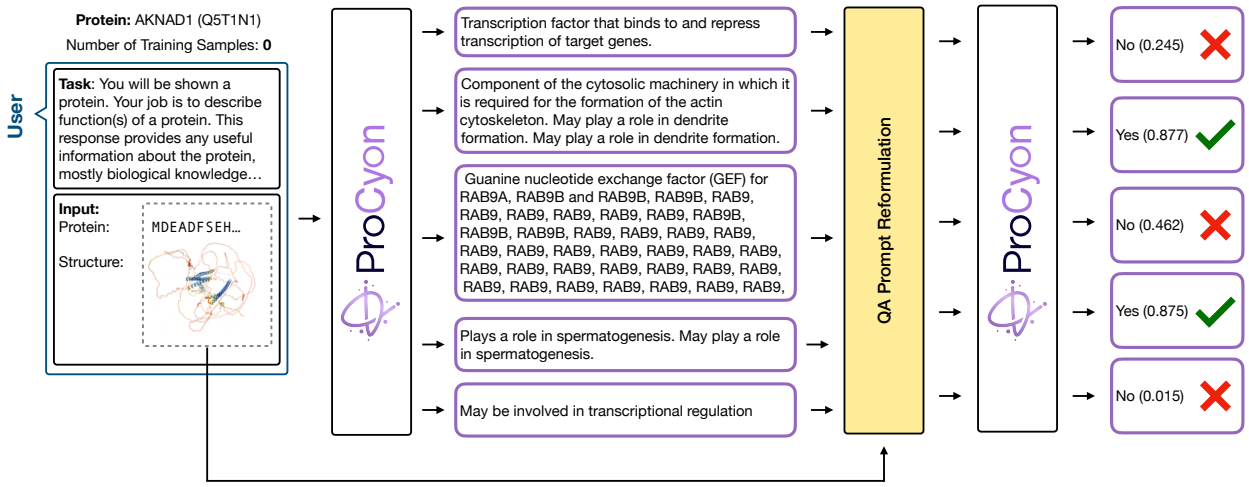
Extended Data Figure 4: question answering performance of PROCYON and other models across different dataset splits of PROCYON-INSTRUCT. Analogous to Extended Data Figure 3, except for the question answering task.



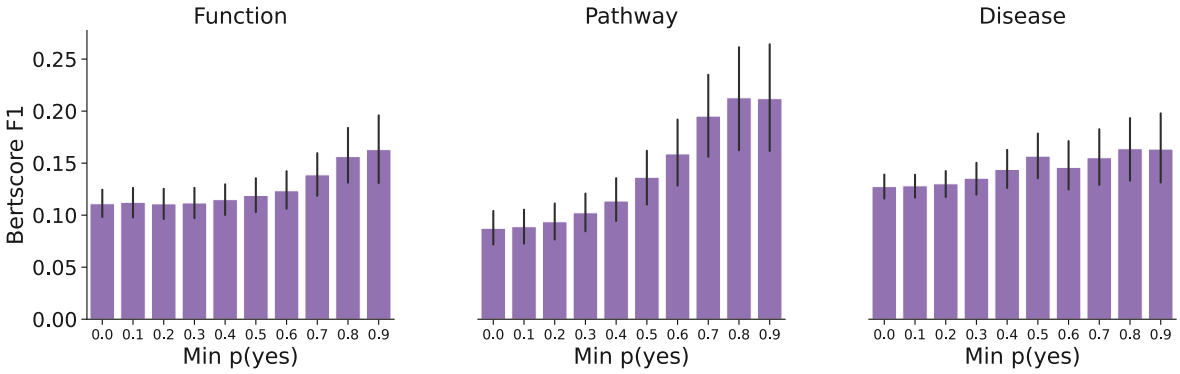
Extended Data Figure 5: Example instructions for four different tasks in PROCYON-INSTRUCT, including disease-gene association, target-specific mechanism of action (MoA) generation, pathway annotation, and experimental protein interaction prediction. These examples show how the various knowledge domains can be represented as a unified instruction template which interleaves multimodal inputs (i.e., proteins, domains, etc.) into the text within each instruction.



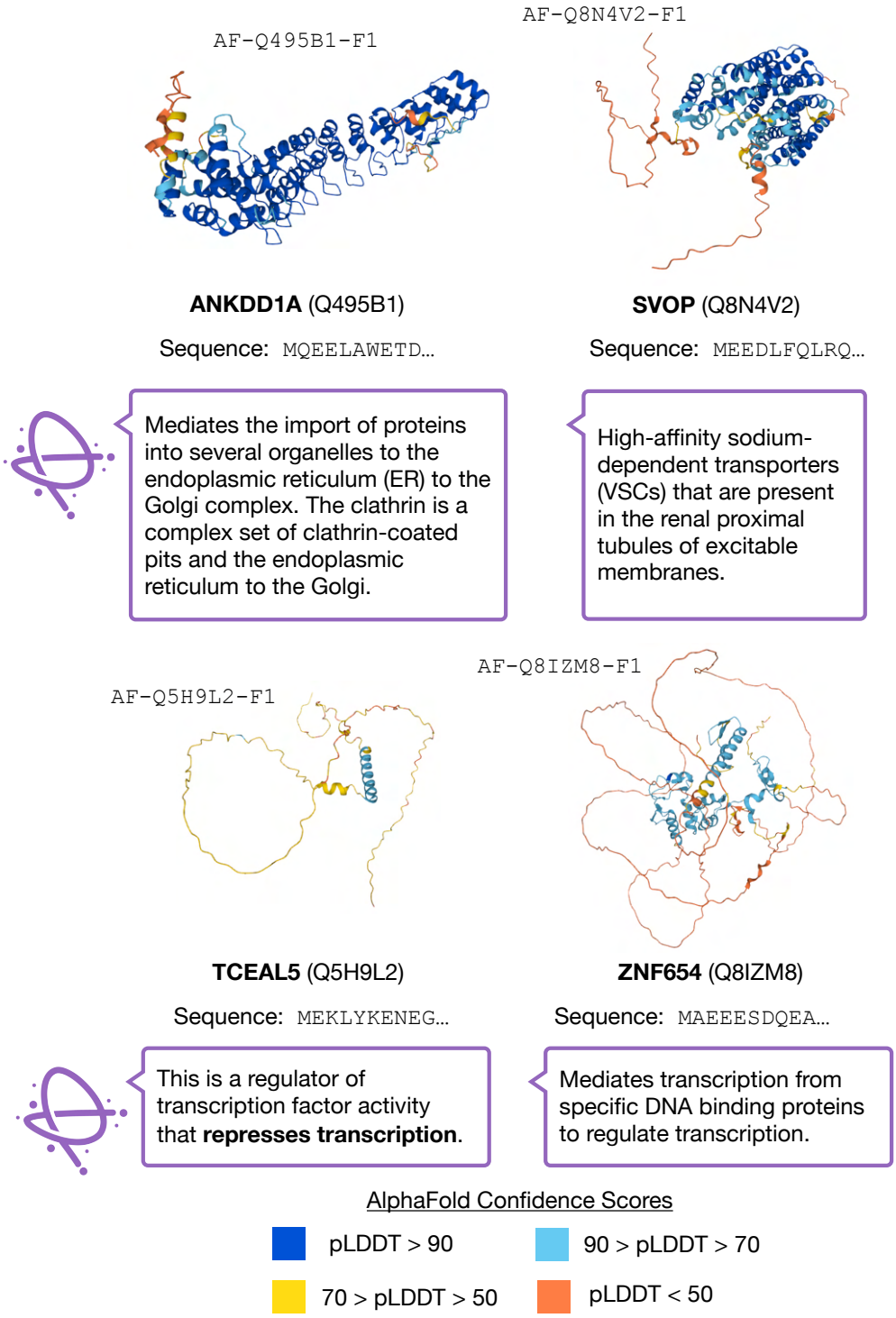
Extended Data Figure 6: PROCYON can successfully retrieve the causal gene of 22 monogenic diseases using as input prompt descriptions providing different information, and with or without the disease name. The disease descriptions correspond to the disease name, clinical phenotype, functional pathways, molecular phenotype and a broad definition. The causal gene was originally identified experimentally through cloning and linkage analysis.



Extended Data Figure 7: QA filtering for the example shown in Figure 4. PROCYON originally produces N texts (normally 10, but 5 are shown), and then we reformulate the output description as a QA query and pass it back to the model, getting output yes/no scores (as described in Section 2.3). Note that this procedure removes incoherent text (middle example with "RAB9" repeats) and vague general text ("May be involved in transcriptional regulation") when necessary. We also believe it removes biologically-infeasible texts since it leverages a direct training mechanism (QA) of PROCYON.



Extended Data Figure 8: QA filtering improves caption quality Shown are BertScore F1 scores with higher QA filter thresholds ("Min p(yes)"). As the threshold increases, the quality of captions improve. This holds across multiple knowledge domains, and QA filtering seems to be more effective for function and pathway domains.



Extended Data Figure 9: PROCYON captions for poorly-characterized proteins determined by [86] to be associated to Parkinson's Disease via α-synuclein toxicity screening. Each protein has no phenotypic relations in PROCYON-INSTRUCT, representing natural zero-shot examples for PROCYON. We show PROCYON descriptions for biological processes the proteins may be involved in. Shown is also the confidence scores for AlphaFold-predicted structures of each protein; this highlights the poor experimental characterization of each protein considered.

Add annotation that identifies that the structure is AlphaFold generated
- *Mention uncertainty of the structure*
- *Take generated text and put it in a box, say it's ProTeus*

Online Methods

1 Datasets

We introduce our dataset PROCYON-INSTRUCT, which contains diverse datasets curated from 12 sources that span biological *knowledge domains*. This dataset is available for download at https://huggingface.co/datasets/mims-harvard/ProCyon-Instruct.

1.1 Sourcing PROCYON-INSTRUCT

First, to specify the types of datasets in PROCYON-INSTRUCT, we define five different *knowledge domains* into which each dataset can be binned:

Function. The function knowledge domain annotates proteins and domains with pathway-level biological information, including molecular functions and biological processes in which each protein is implicated. To curate such annotations, we source data from Gene Ontology [95], Reactome [96], Enzyme Commissions [97], and UniProt [26].

Protein domains. This knowledge domain contains data on domain families, which are subunits of proteins that are shared across many proteins and communicate distinct, shared functions. We derive functional information about domains from Gene Ontology [95] and Pfam [98, 99].

Disease. The disease knowledge domain annotates proteins with diseases they are associated with, where associations are defined based on established evidence linking genetic variations or mutations to the diseases. To curate such annotations, we source data from DisGeNet [100] and Online Mendelian Inheritance in Man (OMIM) [101].

Therapeutics. The therapeutics knowledge domain annotates proteins with their associations with drugs, including their involvement in the drug mechanism of action, uptake, and metabolism. To curate such annotations, we source data from DrugBank [102] and IUPHAR/BPS Guide to Pharmacology [103].

Interaction. The interaction knowledge domain encompasses interaction between multiple proteins, domains, or peptides. We curate the STRING dataset [40] which includes protein-protein relationships of coexpression, homology, and experimental evidence. Finally, we include a Protein-Peptide binding dataset from PDBBind [66] which encompasses physical interactions between biological macromolecules more broadly.

Detailed information on the downloading and processing of each separate dataset is included

in Supplementary Note 4. There are a number of resources we do not include for each knowledge domain. Some exclusions are due to the restricted access to the full repository of data. For some others, exclusions are due to the need of external validation. For example, we utilize Open Targets database for external validation, which is thus not included.

1.2 Evaluation Splitting for PROCYON-INSTRUCT

All datasets in PROCYON-INSTRUCT, with the exception of protein-protein interaction and protein-peptide binding datasets, are composed of samples of associations (phenotype, protein or domain instance). For most datasets, an association means an annotation of a textual description of a phenotype to a protein or domain instance based on some pre-specified rule. For the DrugBank dataset specifically (protein-drug associations), associations are further stratified into "target", "carrier", "enzyme", and "transporter", representing different roles of proteins at different stages of a drug within human body. To robustly evaluate the capability of models to annotate proteins with phenotypes beyond controlled vocabularies, we split each dataset into one training and multiple test sets, evaluating different levels of generalizability. More specifically, unique phenotypes from a specific dataset are splitted into three different phenotype sets: many-shot, few-shot, and zero-shot, with a ratio of approximately[1] 0.7:0.15:0.15. We then split about 20% of all associations containing phenotypes in the many-shot phenotype set as the many-shot test set, with the rest being in the training set. For each phenotype in the few-shot phenotype set, k different associations containing the phenotype are splitted into the training set (k is set to be either 2 or 5 depending on the average number of associations for phenotypes in each dataset), with the rest into the few-shot test set. All associations containing any zero-shot phenotypes are kept in the zero-shot test set. In summary, the many-shot test set contains annotations with phenotypes that are seen much during training, the few-shot test set contains annotations with phenotypes that are only seen at most k times during training, and the zero-shot test set contains annotations with phenotypes that are not seen during training. Three validation sets are also split in a similar manner for model development purpose. For the few-shot and zero-shot phenotype sets, we split phenotypes into 1:2 (validation:test). Training sets from all datasets are merged as the final training set.

We define rigorous splits for each dataset to ensure minimal leakage across the train and test sets. Details on each specific split are included in Supplementary Note 1.

[1] As will be specified in later subsections, it is oftentimes not possible to split the phenotypes with an exact ratio as desired due to the stringency brought by additional requirements in our splitting strategy for those datasets. The exact ratios of splits for each dataset can be found in Supplementary Table S1.

1.3 Generating dataset rephrasings

Language from curated biomedical databases often consist of boilerplate terminology that is standardized across terms in the database. For example, the Gene Ontology text for ”lipoxin A4 biosynthetic process” reads ”The chemical reactions and pathways resulting in the formation of lipoxin A4. Lipoxin A4 is a C20 hydroxy fatty acid having (5S)-, (6R)- and (15S)-hydroxy groups as well as (7E)- (9E)-, (11Z)- and (13E)-double bonds.” while the text for ”lipoxin B4 metabolic process” reads ”The chemical reactions and pathways involving lipoxin B4. Lipoxin B4 is a C20 hydroxy fatty acid having (5S)-, (14R)- and (15S)-hydroxy groups as well as (6E)- (8Z)-, (10E)- and (12E)-double bonds.” Both sentences have roughly the same grammatical syntax, a structure that is represented across much of the Gene Ontology database. These boilerplate descriptions might be comprehensive and detailed enough for experts in the field, but this can make such terminology inconcise and impractical for a broader scientific audience. In addition, this lack of diversity in text is detrimental to language model training, where diversity of the dataset is crucial for learning [104–107]. Therefore, we implement a data augmentation strategy whereby we prompt GPT models [25] to rephrase the entities from each database to represent the diversity of user inputs that the model might encounter in deployment. To further increase diversity, we include prompting that encourages GPT to output descriptions in the style of researchers with different levels of expertise.

To generate such rephrasings, we prompt `GPT-3.5-turbo-0125` to generate responses in a JSON format that includes several definitions of expertise levels. We define three expertise levels: 1) junior scientists, which are individuals at a Bachelor’s degree level who have basic medical science knowledge but no specialized biological knowledge, 2) mid-level scientists, which are individuals pursuing a graduate or professional degree in the biomedical field and thus are developing deep knowledge of biomedicine, and 3) senior scientists, who have specialized knowledge of biomedical concepts and have more than five years of postdoctoral research experience. In addition, we define two levels of description transformations 1) rephrasing: an “alternate phrasing” of the phenotype and 2) summarisation: a shortened and more concise version of the phenotype. GPT is prompted to generate a “rephrasing” and “summarisation” for each expertise level. The full description of the expertise levels are included in Supplementary Note 6. We include the full prompt to GPT in Supplementary Note 7. Finally, we include a table with all fields that were rephrased in Supplementary Table 5. We do not publicly release these rephrased entities; please contact the authors of this work to request access to these augmented descriptions.

2 PROCYON Model

Overview. We introduce PROCYON, a novel multimodal foundation model that is built on top of transformer deep learning architectures. PROCYON is designed to process multiple biomedical modalities in the input—including protein sequence, protein structure, ligand structure, and natural language—while generating text and prioritizing proteins in the output space. The core of PROCYON is built on decoder-only transformer-based language models, specifically the LLaMA family of models [39, 108]. PROCYON is inspired by recent developments in multimodal vision-language models (VLMs), including LLaVA [109], BLIP [44, 110], FROMAGe [111], and Flamingo [112], which center around LLMs for processing multimodal information. PROCYON can interleave multimodal inputs into textual prompts, enabling the specification of in-context examples and diverse types of input queries, such as prompting the model for protein-protein interactions. In addition, PROCYON is trained to retrieve and prioritize proteins and arbitary amino acid sequences based on user input queries. In this section, we will describe the exact specifications for our model architecture, which is in supplement to the model implementation given in our codebase https://github.com/mims-harvard/ProCyon. We fully release all of PROCYON's model weights and codebase as well as extensive demos and steps for using PROCYON such that the research community can build on and explore our model's capabilities.

2.1 Capabilities and modes of PROCYON

PROCYON is a foundation model that has a variety of multimodal capabilities. We start this section by describing how a text-only LLM can be made to take multimodal inputs via multimodal token composition, and then we proceed to describe how we can leverage different "modes" of the model to accomplish diverse tasks at inference time. In terms of input and output structures, PROCYON has two main modes: retrieval and text generation. For text generation, we define two main sub-modes, including question answering (QA) and phenotype generation. Details for inference of PROCYON are included in Methods Sec. 2.3, and training details are described in full in Methods Sec. 3.

Multimodal token composition. PROCYON is built on top of a pre-trained generative large language model (LLM) which is a decoder-only transformer architecture [113]. The input to the LLM module is a set of token embeddings $\{\mathbf{t}_1, ..., \mathbf{t}_n | \mathbf{t}_i \in \mathbb{R}^{d_{\text{input}}}\}$. The LLM is pretrained on a large text-only corpus, i.e., each $\mathbf{t}_i$ represents a piece of a word as defined by the Byte-Pair En-

coding (BPE) tokenizer [114]. We adopt the base tokenizer used by LLaMA-2 and LLaMA-3 for PROCYON-SPLIT and PROCYON-FULL, respectively.

An interleaved multimodal input consists of a combination of text, protein sequence, protein structure, and drug ligand structure inputs. For example for a sequence `sequence_1` with corresponding structure `structure_1`, an interleaved input with text might be "Describe the role of `sequence_1` `structure_1` in obesity". To denote these multimodal inputs in the model, we introduce new input tokens to the tokenizer. `<|protein|>` denotes a protein sequence input, `<|struct|>` denotes a protein structure input, and `<|drug|>` denotes a drug ligand input. During tokenization, we keep track of the indices of each multimodal input token and use this later for inserting multimodal embeddings.

Each input modality gets a specific encoder that has been especially designed with inductive biases to match the given modality. To encode the protein sequences, we utilize a large protein language model which has been pretrained on over 60 million proteins [16]. For protein structures, we use an equivariant geometric deep learning model built to process 3-dimensional molecular structures [21]. For ligand drug structures, we use another geometric deep learning model that achieves state-of-the-art performance for molecular classification tasks [115]. More details on exact specifications of the encoders for each modality are included in Methods Sec. 2.2. During a forward pass, we deconstruct a user input into the respective modalities provided to the model, and we pass each input entity into the encoders for their respective modality; this then gives us embeddings for each multimodal input entity.

The next step is to connect the output spaces of the protein and drug encoders to that of the LLM. We use three-layer multi-layer perceptrons (MLPs) to accomplish this task which we refer to as "Connectors". Specifically, we use the "Token Connectors" to project the embeddings of proteins and drugs into the input space of the LLM. Here, we describe only the token connectors (other connectors are defined in subsequent sections). We define three token connectors: 1) $\mathbf{C}_p^t$ which takes as input the output embedding from the protein language model, 2) $\mathbf{C}_s^t$ which takes as input the protein structure embedding, and 3) $\mathbf{C}_d^t$ which takes the drug/ligand structure as input. The input to each connector matches the dimensionality of the output embedding from each multimodal encoder, and the output from the final layer of the MLP matches the dimensionality of the input token embeddings for the LLM. Once each unified embedding is received, we take these embeddings and pass them to the LLM in the index of the multimodal indicator tokens described previously. Thus, each multimodal entity becomes a kind of "soft token" input to the model, which allows it

to flexibly encode a wide variety of modalities that can be interleaved as complex instructions.

Retrieval mode. The retrieval mode is defined as a query in which the input is a phenotype and the output is a list of prioritized amino acid sequences, which can be either proteins, domains, or peptides, depending on user preferences. Practically, output is returned to the user as an ordered list of proteins p_i with corresponding retrieval scores s_i, i.e., $\{(p_1, s_1), ..., (p_n, s_n)\}$ where $s_i < s_{i-1}$. For retrieval demonstrations, we return the lists in descending order, where the first protein returned is the most relevant according to the model. Inputs can be any combination of phenotype given as either text or other amino acid sequences. Some example tasks that can be formulated as retrieval tasks are disease-gene prioritization, drug target identification, and homologous protein retrieval. See Extended Data Figure 1 for more examples.

To enable the pre-trained LLM to accomplish this task, we introduce a special token `[PROT]` that denotes a retrieval task. This is similar to the approach taken by FROMAGe [111], which grounds language model outputs in libraries of images. We then access the hidden representation of the last layer of the LLM that corresponds to the next-token prediction of the `[PROT]` token, $\mathbf{z}_p^l \in \mathbb{R}^{d_{\text{LM}}}$, where d_{LM} is the hidden dimensionality of the last layer in the language model. We then define a two MLP connectors to aid in the retrieval procedure. First, $\mathbf{C}_p^q$ is known as the Query Connector, and it projects the output embedding from the language model $\mathbf{z}_p^l$ (the query) to a shared retrieval latent space space $\mathbf{z}_r^l \in \mathcal{Z}^r$. Second, we define a Retrieval Connector $\mathbf{C}_p^r$, which projects a set of candidate protein sequences $\mathcal{P}$ into the retrieval embedding space $\mathcal{Z}^r$. This shared latent space is then optimized via contrastive learning (Methods Sec. 3.4) and leveraged for multimodal retrieval (inference procedure described in Methods Sec. 2.3).

Text generation mode. After establishing PROCYON's multimodal token composition capabilities as being able to process any combination of interleaved multimodal information, its abilities in text generation very clearly follow. We group both QA and phenotype generation queries under this section as both of these capabilities are fundamentally text generation; in QA, the output is a simple "yes" or "no" answer, which is supervised as such (see Extended Data Figure 7), and in phenotype generation, the output is free-form text in the form of a short description or a paragraph. After constructing multimodal inputs via the methods previously described, text generation can be done using the mechanisms of the LLM, so no additional architectural changes are needed to enable these capabilities.

2.2 Architecture Specifications

The architecture of PROCYON can be broken down into several modules: the LLM, the encoders (sequence, protein structure, and ligand structure), and the connectors.

LLM. The base of PROCYON is a text-only, pretrained LLM. For PROCYON-SPLIT, we use LLaMA-2 (7B) [108], and LLaMA-3 [39] is used for PROCYON-FULL. This decision was made because during our development of PROCYON, LLaMA-3 models (which show far superior language modeling abilities than LLaMA-2) were publicly released as we had already began training PROCYON-SPLIT. Both LLaMA-2 and LLaMA-3 are decoder-only transformer models trained with causal language modeling objectives (see Equation 4) and can be inferenced for text generation via autoregressive generative sampling. Please reference the original publications of the LLaMA family of models [39, 108] for exact specifications of the LLM's architecture.

Protein sequence encoder. To encode protein sequence, we use the ESM-2 (3B) model [16], which is a bidirectional, encoder-only transformer model. We freeze ESM-2 and use a mean pooling on all tokens in order to obtain one embedding vector per protein sequence. For proteins longer than the context for ESM-2 (1024 amino acids), we use a split-and-merge strategy. Thus, we split the protein into 1024 amino acid chunks, pass each chunk through the encoder in separate batches, and then pool across all output token representations for the mean pooling. This strategy ensures minimal information loss as opposed to truncation of the protein, keeping output embeddings for each amino acid and merging them in the final pooling operation.

Protein structure encoder. We utilized the GearNet architecture to extract embeddings from protein structures. More specifically, we chose the pre-trained variant with the multiview contrast objective, as this demonstrated superior performance across a range of tasks, based on the original results [21]. The pre-trained GearNet, sourced from the original Zenodo repository (https://zenodo.org/records/7723075), was used with its weights kept frozen throughout the PROCYON training and finetuning. We configured the GearNet model with an input dimensionality of 21, reflecting the number of input features for each node in the protein structure graph. The chosen model architecture consists of 6 hidden layers, each with 512 hidden units. Moreover, we set the edge input dimensionality to 59, and the number of relation types to 7, which corresponds to the number of relations of the relational graph network backbone model of GearNet.

Ligand structure encoder. We utilize the pre-trained Transformer-M model to extract embeddings

from small molecule ligand structures [115], which makes use of 2D and 3D structural information of molecules. More specifically, we choose the default 12-layer ("L12.pt") configuration and weights of Transformer-M, which has been pre-trained on predicting HOMO-LUMO energy gap and 3D Position Denoising with 3.37 million by the authors [115]. The pre-trained Transformer-M was used with its weights kept frozen throughout the PROCYON training and finetuning. The 2D and 3D inputs are originally converted from SMILES and 3D coordinates, respectively, obtained through DrugBank. For molecules without 3D information, we only feed 2D inputs to the model.

Connectors. Connector modules bridge the gap between inputs of non-text modalities—e.g., protein sequence, protein structure, drug ligand structure—and the text tokenization space of the LLM. Each connector is a multilayer perceptron (MLP) with varying parameter sizes. The parameter size changes are a major change in architecture from PROCYON-SPLIT to PROCYON-FULL, and their exact parameter numbers are given in Supplementary Table 6 for PROCYON-FULL and Supplementary Table 7. The role of these connector modules are defined in other sections. A table of all connector modules, their architecture and their respective tasks are shown in Supplementary Tables 8, 9 and 10.

Embedding caches. Because our encoder bases are frozen during training, we utilize embedding caches to greatly speed up training and inference. Embedding caches, which have been used in other multimodal models such as [23, 24], work by precomputing embeddings for each protein or drug ligand instance in the training dataset and stored in a matrix $\mathbf{E} \in \mathbb{R}^{N,d}$, where N is the number of samples and d is the dimensionality of the embedding. Thus, during forward passes, one can use an index i to represent the protein sequence, protein structure, or drug structure instance and then index the embedding matrix $\mathbf{z}_i = \mathbf{E}[i, :]$. This avoids unnecessary forward passes through the frozen encoder.

2.3 PROCYON Inference

In this section, we describe how inference is performed on PROCYON, detailing the core capabilities. This section does not describe training objectives and optimization, which is discussed throughout Methods Sec. 3.

Retrieval mechanism. Retrieval inference is performed by searching over a large database of embeddings in order to find proteins most related to the input query. Given is a user input query x_i and a library of amino acid sequences $\mathcal{S}$ (domains, proteins, peptides, etc.). We first compute the

embedding of x_i by passing it to PROCYON and then take the output embedding from the [PROT] token. This is then passed to the query connector $\mathbf{C}_p^q$, where a query embedding $\mathbf{z}_i$ is obtained. Then, we pass all of the amino acid sequence library $\mathcal{S}$ to the protein language model of PROCYON; this output is directly fed to the retrieval connector $\mathbf{C}_p^r$, producing an output embedding matrix $\mathbf{Z}_\mathcal{S}$. We then compute the similarity between $\mathbf{z}_i$ and the embedding matrix $\mathbf{Z}_\mathcal{S}$ (using cosine similarity) to obtain similarity scores $\mathbf{y}$. We then rank the similarity scores in descending order to obtain our ranking of similarities, and then we match the indices of this sorting with the original amino acid sequences in $\mathcal{S}$. Ultimately, we return to the user $\mathcal{S}^*$ and $\mathbf{y}^*$, which are sorted amino acid sequences and similarity scores, respectively (as is shown in Figure 1d). Users may then subset to a certain top-k number of important proteins, but we leave this as a post-processing step for completeness. Such an inference is common in CLIP-style multiimodal models [116], which seek to match some query to multimodal entities via a shared latent space.

This mechanism gives the user not only the ranked list of amino acids but the similarity scores as given by the model. In this way, PROCYON performs *grounded* retrieval, meaning that its outputs are not dependent on free-text generation, which is known to produce hallucinations [117], but rather specifically linked to the input amino acid library. We note that users may provide their own library of amino acid sequences or optionally use our default sequences, which are all human proteins in UniProt.

Question answering. At inference time for QA tasks, we use a greedy algorithm which selects the top token prediction after the "[ANSWER]" token provided to the model. For some tasks such as the QA filter, we use the raw token score predicted at the "yes" and "no" tokens. However, we found empirically that a well-trained (converged) PROCYON model tended to predict almost all the softmax probability in the "yes" and "no" tokens for QA, i.e., no other tokens were included.

Phenotype generation. For phenotype generation (i.e., generating free-form textual responses to input instructions), we use diverse beam search [118], a modification of the beam search algorithm (a common algorithm for generating text from language models). Diverse beam search encourages separate beams to generate text distinct from the other beams, thereby resulting in semantically distinct generations; we verified this result qualitatively. In all of our experiments, we use a beam size of 10, beam group size of 2, and a diversity penalty of 0.8.

Phenotype generation with QA filtering. Since we generate multiple phenotype descriptions via beam search, we want to also provide some sense of confidence by the model for the gener-

ated description. We found through empirical testing that simply ranking by log-likelihood of the sequence (as typically given by beam search algorithms) did not effectively correlate with "correctness" of the answer. For instance, the model would default to more-likely seen texts (such as more often-represented GO terms) as higher log-likelihood. Therefore, we employ a strategy called "QA filtering" where we recycle the outputs of the model and reformulate them as separate QA inputs along with the original query protein or domain. This strategy is described in Extended Data Figure 7. Then, we take the "yes" token index in the output of the PROCYON model as the certainty of the relation between the generated text and the protein/domain. This "certainty" score can be thresholded at 0.5 (>0.5 being "yes", <0.5 being "no"), or one can take the top-k descriptions as ranked by the "yes" score. This requires no additional training or weights but rather an additional N forward passes to the model, where N is the number of generated descriptions to be filtered. This capability arises from the fact that PROCYON generates and can process in its input free-form text, even text that it generates.

QA filtering has many benefits and downfalls. The first benefit is the empirical evidence that it benefits caption quality with respect to the BERTScore metric, as seen in Extended Data Figure 8. Another benefit is that the nature of QA training gives PROCYON greater confidence on scoring more commonly-seen phenotypes, thereby providing a filter for commonly-seen descriptions to which phenotype generation might default. Let t_i be a very commonly-observed textual description in one of our databases, e.g., a disease that has many associated proteins. This means that t_i is seen more often by PROCYON during training, thereby giving it the chance to learn more negative examples and become more certain in predictions about this phenotype. Therefore, PROCYON is more confident in more commonly-seen examples, so if PROCYON generates a more commonly-observed text such as t_i, we are more confident in QA filtering to identify if this phenotype does or does not match the protein. A major downfall is that it is upper-bounded by the performance of PROCYON in QA for a given knowledge domain. If PROCYON does not converge or is not accurate for a given task or dataset description, then QA filtering is less effective. Another downfall is that seemingly-incoherent outputs may present out-of-distribution samples to PROCYON in the QA reformulation; however, note that this behavior was not observed empirically during testing. In fact, QA filtering was often able to catch incoherent outputs such as the middle example generated description in Extended Data Figure 7, which repeats the term "RAB9".

Overall, QA filtering provides a self-contained "check" on the outputs of PROCYON by leveraging its free-text understanding capabilities in a systematic manner to improve predictions.

2.4 Implementation of PROCYON model

PROCYON is implemented in PyTorch `2.2.0+cu121` and wrapped with the Huggingface library (`transformers==4.31.0`). We follow a similar API to that of Huggingface LLMs in order to maximize user compatibility and familiarity. More details are found in our codebase, including demonstrations on how to query the model.

3 Training PROCYON

In this section, we describe the training procedure for PROCYON and the design decisions made during the large-scale training of this model.

3.1 Instruction tuning and interleaved multimodal inputs

To train PROCYON, we use instruction tuning, which expresses inputs in the form of an instruction, thereby increasing the model's ability to understand new, unseen tasks at inference time. Instruction tuning has been studied primarily in text-only LLMs as a way to increase performance on natural language tasks after large-scale pretraining, and it is an important alignment step in tuning large language models to more effectively follow human instructions and act as "chatbots" [42–46]. In this section, we will describe how to formulate instructions using interleaved multimodal inputs in order to train PROCYON.

Instruction templates. Because our data are not initially provided as instructions, we define templates for each task and dataset. We follow the SuperNaturalInstructions [119] formulation of instruction templates, which defines three critical components for instruction templates: 1) task definitions, 2) in-context examples, and 3) input instance.

Task definitions. We define task definitions as short paragraphs of text that describe the task to be performed by the model. We manually curate task definitions for each dataset across the three core capabilities of the model: QA, retrieval, and phenotype generation, as previously described. Thus, for our 20 pretraining datasets (including each relation within data sources) and three core capabilities, we have 60 task definitions. The task definition is always at the beginning of the instruction (with the exception of rephrased instructions), and it is preceded with a "`Definition:`" prefix to indicate to the language model that the following paragraph is a task definition. We release all of our task definitions along with our training data and codebase.

In-context examples. Following [119], we also provide in-context examples of each task to be

performed by the model. In-context examples are completed, correct examples of the task to be performed. We use other (protein, description) pairs from the dataset being trained on to formulate the in-context examples. However, we fix the samples used for the in-context examples for each dataset, i.e., the in-context example given for one dataset are fixed for all example in this dataset, following the approach given by [119]. For retrieval and phenotype generation capabilities, we use one (1) in-context example, but for QA, we provide one positive example (i.e., where the answer is "yes") and one negative example (where the answer is "no"). Positive examples (including the examples for retrieval and phenotype generation) are followed by a preamble "`Positive Example:`" and negative examples are followed by a preamble "`Negative Example:`". We use these examples during inference (from the training databases) except for zero-shot tasks, where no in-context examples are provided.

Note that the ability to represent multimodal in-context examples follows directly from multimodal interleaved inputs, which allows the representation of PROCYON to be as flexible as for text-only LLMs.

Input instance. The input instance within the instruction is given after the task definition and in-context examples at the end of the instruction. This is the actual input query to the model, which we use for supervision during training. For text generation capabilities (QA and phenotype generation), we introduce a new token "`[ANSWER]`" which denotes to the model when it should output a response to the prompt. We train the language modeling objective (described in Methods Sec. 3.4) on the tokens given only after the `[ANSWER]` token. Full training objectives are given in Methods Sec. 3.4.

Rephrasing instruction template components. To increase diversity of the instructions seen during training, we use GPT-4 to rephrase the instructions given to the model. This rephrasing is in addition to the rephrasing of database descriptions that we perform for training, which is described in Methods Sec. 1.3. Here, we rephrase the task definitions that we originally write for each task. We prompt GPT-4 (version `GPT-4-1106-preview`) to generate rephrased task definitions matching the expertise levels described in Methods Sec. 1.3 and Supplementary Note 8. Our full prompt is given in Supplementary Note 9. At train time for PROCYON-FULL, we sample first the expertise level and then the rephrasing level. Expertise level is tied across both task definitions and entity descriptions, but rephrasings are sampled independently for each of these components. This means that the task definition and the input description given for an example will be from the

same expertise level, but these may be given as a summary and a rephrasing, a rephrasing and a simplification, etc. We then use a "personality system prompt", as shown in Figure 4e, which tells PROCYON to process inputs and generate outputs that match the style of the requested expertise level. These prompts are given in full within Supplementary Note 8.

From protein-text pairs to each PROCYON capability. PROCYON is trained for the three capabilities discussed in Methods Sec. 2.1. For enabling these capabilities, we design instructions that match each capability. Each sample from our databases (excluding the interaction knowledge domain which is discussed in the subsequent section) is a protein, text pair (p, t), where p is associated to the phenotype described by the textual description t. Let $\mathcal{F}$ denote the PROCYON model and I denote some human-provided instruction (encoded via the instruction tuning formatting). Then, each task can be broken down as a mapping between p and t conditional on instructions. If we define each mode of PROCYON as $\mathcal{F}^{\text{R}}$ as the retrieval mode, $\mathcal{F}^{\text{QA}}$ as the QA mode, and $\mathcal{F}^{\text{C}}$ as the phenotype generation mode, then each function is defined as such:

$$\mathcal{F}^{\text{R}}(t|I) \mapsto p; \mathcal{F}^{\text{QA}}(p, t|I) \mapsto \{\text{“yes”}, \text{“no”}\}; \mathcal{F}^{\text{C}}(p|I) \mapsto t \tag{1}$$

Thus, we construct instructions to reflect these tasks. For retrieval instructions, we give the model the instruction and text, prompting for the protein that matches the query text and instruction combination. For QA instructions, we give the model the instruction, protein, and text and prompt the model to predict if the text describes a property of the protein ("yes") or not ("no"). For phenotype generation instructions, we give the model the instruction and protein, prompting the model to generate the text t that matches the protein. In this way, one pair of protein-text (p, t) is transformed into three samples phrased in different ways to the model via instruction tuning.

Instructions for molecular interactions. Molecular interactions, including protein-protein interactions for the STRING dataset and protein-peptide interactions for the PDB dataset, are encoded as instructions. As a reminder, PROCYON can handle interleaved multimodal inputs (see Methods Sec. 2.1), so encoding protein-protein interactions is very straightforward. Denote a protein-protein (or peptide-protein for generality) sample as (p_i, p_j, r), where p_i and p_j are two proteins and r is the relation describing their relationship, such as homology, coexpression, or experimental interaction. Without loss of generality, this can also refer to a general physical interaction such as binding that is referred to in the PDB dataset. We encode the relation as part of the instruction, which we can denote as $I(r)$ for the relation-augmented instruction. This is as simple as men-

tioning to the model that the relation being predicted is, for example, coexpression. Then, we can construct retrieval and QA examples as:

$$\mathcal{F}^{\text{R}}(p_i|I(r)) \mapsto p_j; \mathcal{F}^{\text{R}}((p_i, p_j)|I(r)) \mapsto \{\text{“yes”}, \text{“no”}\} \tag{2}$$

Thus, we train PROCYON models on retrieval and QA examples from datasets in the interaction knowledge domain. Phenotype generation is omitted because no text is naturally associated with this interaction data. All instruction templates, including task definitions, used in our training process is available in our codebase which can be found at https://github.com/mims-harvard/ProCyon.

Example instructions are shown in Extended Data Figure 5, with each component highlighted that is discussed in this section.

3.2 Context augmentation

We introduce several strategies to improve model training through methods known as context augmentation (CA). Because input to PROCYON can be expressed flexibly as interleaved multimodal-text data, the model can encode arbitary context to perform a task. Therefore, we design strategies that leverage domain-specific intuition to increase PROCYON's performance. We define three kinds of CA techniques, Ontology CA, Disease CA, and Drug CA.

Ontology CA utilizes the ontologies of Gene Ontology and Reactome databases to retrieve upstream ancestors in the ontology to insert as context to the input prompt; this narrows the search space for PROCYON given the high number of protein-phenotype connections. Disease CA retrieves a functional description of the protein during training to further connect the molecular phenotypes to high-level clinical descriptions of diseases. Drug CA uses the indication and mechanism of action to boost prediction of other given properties of the molecule that are relevant to therapeutics tasks. More details are given in Supplementary Note 10, 11 and 12 and in Supplementary Table 11. A summary of all datasets that use each CA technique is given in Supplementary Table 12.

3.3 Data processing details for training PROCYON

In this section, we summarize a collection of design choices not outlined in previous sections that were helpful for training PROCYON.

Epoch scaling per task. We implement an epoch scaling technique to combat overfitting on textual data during training of PROCYON. Since we train PROCYON to simultaneously perform retrieval, QA, and phenotype generation tasks, we must carefully balance the training resources used for both of these main tasks. To balance such differences during training, we scale the frequency of training for the main tasks (QA, retrieval, and phenotype generation). Define a standard whole-task training batch at step i as $\mathbf{B}_i = \{B^i_{\text{QA}}, B^j_{\text{R}}, B^k_{\text{C}}\}$ where $B_{\text{QA}}, B_{\text{R}}, B_{\text{C}}$ are batches of QA at QA step i, retrieval at retrieval step j, and phenotype generation (captioning) at captioning step k, respectively. An epoch scaling factor is an integer n_τ associated to one of the given tasks, thus $n_{\text{QA}}, n_{\text{R}}, n_{\text{C}}$ is the epoch scaling factor for QA, retrieval, and captioning, respectively. n_τ controls the frequency at which a task is seen during a given training step; for example, a set of $n_{\text{QA}} = 2, n_{\text{R}} = 1, n_{\text{C}} = 1$ would yield a training step as $\{B^i_{\text{QA}}, B^{i+1}_{\text{QA}}, B^j_{\text{R}}, B^k_{\text{C}}\}$. Over training, this amounts to each task being seen at different numbers of epochs. Epoch scaling parameters for PROCYON are reported in Supplementary Table 13.

Two-stage training. For training PROCYON-SPLIT, we utilize a two-stage training process whereby the connector layers are pretrained using limited data while keeping the language model frozen. This was originally inspired by [109], which follows a similar architecture but for vision-language models. In this work, they pretrain the connector layers while keeping the language model frozen, thereby providing a better initialization of connector weights before tuning language model weights in the second stage. Thus, we trained a model PROCYON-STAGE1 on the UniProt dataset without using instruction tuning. This means that the model is prompted with a simple retrieval query or caption query without the added text of an instruction. This allows the newly-introduced tokens to be tuned as well as the connector layers. We report PROCYON-STAGE1 hyperparameters alongside hyperparameters for other versions of PROCYON in Supplementary Tables 14 and 13. More details on the dataset used to train PROCYON-STAGE1 is included in Supplementary Note 13.

For training PROCYON-FULL, we decided to not follow this protocol since work that was published during the duration of training PROCYON-SPLIT suggested that the two-stage training procedure is not necessary [120].

Integrating rephrasings during training. During training of PROCYON-FULL, we integrate rephrasings for task descriptions (Methods Sec. 3.1) and phenotypes as a sampling procedure. For each sample drawn in a minibatch, we decide with 0.5 probability if that entity will be rephrased, for the entities that we consider to be rephrased (see Supplementary Table 5). This probability was

determined via internal testing on a limited model that is not reported here. After deciding that an entity will be rephrased, we then select the expertise level and granularity level independently and uniformly. Please see Methods Sec. 3.1 for more information on how the rephrased instructions are constructed.

3.4 PROCYON Training Objectives

We formalize PROCYON training into three primary task types (and their shorthand for this section): protein retrieval and prioritization ("retrieval"), question answering ("QA"), and phenotype generation ("captioning"). In this section, we describe the training procedures and objectives for each of these tasks.

Retrieval training. We train the model for retrieval based on contrastive learning, which optimizes a joint embedding space for outputs from the protein language model and from the LLM. Specifically, we adopt the InfoNCE [121] framework for multimodal representational alignment, which has been a popular choice in vision-language models [116] as well as some protein-language models [23]. This is formalized as an output text embedding $\mathbf{z}_i^t$ and output protein embedding $\mathbf{z}_i^p$. Thus, our objective is:

$$\mathcal{L}_{\text{ret}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_i^t, \mathbf{z}_i^p)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^t, \mathbf{z}_j^p)/\tau) \cdot \mathbf{m}_{i,j}} + \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_i^p, \mathbf{z}_i^t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^p, \mathbf{z}_j^t)/\tau) \cdot \mathbf{m}_{i,j}} \tag{3}$$

Where θ are the weights of PROCYON, N is the batch size, $\mathbf{z}_i^t$ is the retrieval embedding from the LLM, $\mathbf{z}_i^p$ is the retrieval embedding from the PLM, τ is a learnable temperature parameter, and $\mathbf{m}_{i,j}$ is a mask between samples i, j to account for mismatches in negatives across the batch (i.e., batch conflicts due to in-batch negative sampling). Note that the denominator in either term is guaranteed to be non-zero since by definition, $\mathbf{m}_{i,i} = 1$.

Sampling negatives for Retrieval Optimization. We perform in-batch negative sampling, as is standard practice in contrastive learning frameworks due to efficiency and simplicity [122, 123]. In the distributed training framework, we pool negative samples from all proteins across the global batch for each forward pass; such an operation is accomplished through an `all-gather` call which is implemented in PyTorch distributed [124]. This effectively allows us to increase our batch size to leverage all protein embeddings across the global training of the model. However,

we found that during development, the model encountered many false-negatives when using naive in-batch negative sampling, leading to signficant conflict in retrieval training. Therefore, we also implement a negative sampling mask, whereby when a protein is falsely sampled as an in-batch negative sample, we set the distance to 0 in the distance matrix for the conflicting pair. Further details can be found in the PROCYON codebase.

Text generation training. For text generative tasks, we use causal langauge modeling as an objective. As previously discussed, we align QA and phenotype generation tasks (as discussed in Figure 1) under "text generation" for the purposes of training. Both QA and captioning are trained as language modeling tasks, using the causal language modeling objective. Let a training sample X be broken down into a sequence of tokens $\{x_1, x_2, \ldots, x_N\}$. For our purposes, we mask out the part of the sequence that consists of the instruction, so let the instruction be tokens $\{x_1, x_2, \ldots, x_M\} \subset X$ where $M < N$. Thus, our causal generative loss is defined as such:

$$\mathcal{L}_{\text{gen}}(\theta) = -\sum_{i=M}^{N} \log P(x_i \mid x_1, x_2, \ldots, x_{i-1}; \theta) \tag{4}$$

Both QA and captioning tasks are trained on this objective. Please consult Methods Sec. 3.1 for more information on the content of text trained on within the instructions.

Loss rescaling by dataset. Each dataset in our training set contains different numbers of samples, meaning that there is high imbalance between tasks during training. For example, function prediction via Gene Ontology makes up about 29.6% of the dataset, while drug target prediction via DrugBank only accounts for 2.1% of the samples. Further, our datasets are highly non-injective, meaning that one text may be overrepresented for many proteins. For example, following the previous example, Gene Ontology datasets have 52.2 proteins per textual description, e.g., a given text is seen 52.2 times on average per epoch during training; in contrast, drug target prediction via DrugBank has 4.5 proteins per textual description. This leads to the possibility of the model overfitting on descriptions from Gene Ontology and other databases with high levels of this non-injectivity. To combat this task imbalance and the non-injectivity of the datasets, we implement a manual loss rescaling method whereby the loss from rarely-seen databases are upweighted for open-ended phenotype generation. The choice of rescaling weights was made based on a combination of task imbalance as well as observed difficulty on tasks learned by training PROCYON-SPLIT.

Practically, loss rescaling is implemented as a constant multiplier $\lambda_{\mathcal{D}_i}$ for each dataset $\mathcal{D}_i$.

Formally, for some loss function $\mathcal{L}$ for a batch on dataset $\mathcal{D}_i$, the loss-rescaled version $\mathcal{L}^s$ is defined as:

$$\mathcal{L}^s(\mathbf{x}|\mathcal{D}_i) = \lambda_{\mathcal{D}_i}\mathcal{L}(\mathbf{x}) \tag{5}$$

We implement loss rescaling on PROCYON-FULL but not PROCYON-SPLIT, as this phenomenon was observed while training PROCYON-SPLIT. Values used for the loss rescaling are shown in Supplementary Table 15.

Total loss function. In our final loss function, we perform a sum across the QA, retrieval, and captioning batch losses. Assume we receive a batch $B_{\text{QA}}, B_{\text{R}}, B_{\text{C}}$; note that with epoch scaling (discussed in Methods Sec. 3.3), not every batch may be constructed this way, but consider this as a batch WLOG. Consider each batch to come from datasets $\mathcal{D}_{B_{\text{QA}}}, \mathcal{D}_{B_{\text{R}}}, \mathcal{D}_{B_{\text{C}}}$, which are the datasets for the QA, retrieval, and captioning batches, respectively. Let $\mathcal{L}(B)$ denote the loss function computed over a batch B, i.e., over all samples in the batch. Thus, our total loss function $\mathcal{L}_{\text{PROCYON}}$ is defined as such:

$$\mathcal{L}_{\text{PROCYON}}(B_{\text{QA}}, B_{\text{R}}, B_{\text{C}}) = \lambda_{\text{QA}}\mathcal{L}^s_{\text{gen}}(B_{\text{QA}}|\mathcal{D}_{B_{\text{QA}}}) + \lambda_{\text{R}}\mathcal{L}^s_{\text{ret}}(B_{\text{R}}|\mathcal{D}_{B_{\text{R}}}) + \lambda_{\text{C}}\mathcal{L}^s_{\text{gen}}(B_{\text{C}}|\mathcal{D}_{B_{\text{C}}}) \tag{6}$$

where $\lambda_{\text{QA}}, \lambda_{\text{R}}, \lambda_{\text{C}}$ correspond to scalar weights for the QA, retrieval, and captioning tasks, respectively. We then optimize our model for this loss function with the AdamW optimizer.

Targeted modules during training. During training, we optimize the weights of the LLM and the weights of the connectors. The weights of the protein language model (PLM), protein structure encoder, and ligand structure encoder are frozen and not tuned. These decisions follow results from popular vision-language models such as LLaVA [109] which, in a similar architectural setup, train the LLM and connector (analagous to our connector modules) with full-finetuning while freezing the vision encoder (to which our multimodal encoders are analagous).

3.5 Ablations for PROCYON

To better understand the crucial components of PROCYON, we conduct a number of ablations. Our ablations are **inference-time ablations**, whereby we seek to understand which parts of the inference mechanisms are important for PROCYON'S performance. This is to better understand

user recommendations for querying PROCYON. Please see limitations below for a discussion on this point. All ablations are made on PROCYON-SPLIT in order to evaluate fairly on held-out datasets.

Ablation 1: Structure token. We ablate the use of the structural token for QA prediction tasks. QA is the sole focus for this experiment because structure is not explicitly used in the retrieval task. PROCYON shows little-to-no change in performance when removing the structural token from the input, demonstrating the robustness of our model when protein structures might be poorly-characterized or not experimentally-determined (Supplementary Table 16).

Ablation 2: ICL examples. In this experiment, we consider the case where in-context learning (ICL) examples are excluded from the input. ICL examples are intrinsically a part of the instruction tuning framework used to train PROCYON (Methods Sec. 3.1). We find that taking out ICL examples has approximately no effect for retrieval (Supplementary Table 17) while for QA, there seems to be a negligible negative impact on performance when excluding ICL examples compared to their inclusion (Supplementary Table 18). We hypothesize this is due to the ICL examples providing a template, which is important for narrowing the options of the LLM when answering QA-style questions, i.e., limiting answers to "yes" or "no". Thus, we recommend using an ICL example for QA tasks with PROCYON, but for retrieval, no ICL examples are needed.

Ablation 3: Context augmentation. In this ablation, we take out the context augmentation (Methods Sec. 3.2) at inference time to understand if the model uses the CA mechanism. We ablate three CA mechanisms: ontology CA , disease CA , and drug CA. For ontology CA, we see no significant drop in performance when removing CA from the input prompt; in fact, performance slightly improves when removing CA, most likely due to the stochasticity of choosing upstream neighbors at inference time (Supplementary Table 19). For disease CA, we see a slight drop in performance in DisGeNet and more mixed results in OMIM (Supplementary Table 20). Because DisGeNet is far larger than OMIM, we consider these results to be more reliable, thus we can conclude the disease CA provides a marginal benefit to modeling performance. For drug CA, we clearly see that it is crucial to model performance, with performance dropping to almost random (0.5 accuracy) when omitting it (Supplementary Table 21). This is most likely due to the challenging nature of drug prediction tasks, which require extensive knowledge of biological systems as well as molecular components of both drugs and proteins. Thus, the use of drug CA demonstrates how the free-text capabilities of PROCYON can be used to produce more compelling results, a property that is not

achievable by many baseline models compared in this study.

Limitations. Because of the high cost of training PROCYON, we do not conduct training ablations, i.e., we do not train separate instances of PROCYON to understand training choices. We acknowledge that this is a limitation of our work; however, we note that the monetary and energy cost is prohibitive given the long training times for our model (see Supplementary Tables 23). We plan for future work to explore the design components of the PROCYON architecture and training procedure. We hope that as the community explores similar architectures to that of PROCYON, we can derive a better understanding of the design decisions for training such a model.

3.6 PROCYON Training Implementation

Training duration. Training PROCYON was a significant effort that included multi-GPU and many days of training on a large distributed cluster. Training duration can be measured in either a data-centric manner or a compute-centric manner. By a data-centric method, we can measure train duration in numbers of epochs, where an epoch is one iteration over every sample for all training datasets considered in that training. We also break down our reporting of training epochs by task, e.g., QA, retrieval, and caption. Exact training epochs are found in Supplementary Table 22.

Using a compute-centric method to calculate training time, we measure training time in terms of wall clock duration of training iterations. GPU time is then calculated by multiplying the wall clock duration of training with the number of GPUs used during training. Note that this does not correspond to raw GPU usage, as some CPU time may be considered during the training iterations for non-GPU operations such as instruction construction. The total GPU training time for PROCYON-FULL was 287.28 days, PROCYON-SPLIT was 173.91 days, and PROCYON-BIND was 2.39 days. Full train time statistics are found in Supplementary Table 23.

Training hardware. We train all versions of PROCYON on Nvidia H100 GPUs provided by the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. We used 48 GPUs to train PROCYON-SPLIT and 32 GPUs to train PROCYON-FULL; this was due to shifting resource availability, explaining why PROCYON-FULL resulted in longer training time overall.

Training infrastructure. To train PROCYON, we use state-of-the-art methods in distributed neural network training. For our data and model-parallelism, we use Sharded Data Parallel (ZeRO Stage 2) as implemented in the Deepspeed library [125] version `0.12.4`. ZeRO Stage 2 conveniently

keeps the data parallel world size equal to the number of GPUs being used, so each GPU was used for parallel data iteration, as implemented by Deepspeed. More details on our full training stack are available in our codebase (https://github.com/mims-harvard/ProCyon).

4 Analysis and application of PROCYON

We now describe the methods used for various analyses, benchmarking, and probing of PROCYON conducted throughout this work.

4.1 Benchmarking and Evaluation

All benchnarking results are shown in Figure 2gf and in Extended Data Figures 2, 3, 4. When benchmarking question answering and retrieval performance, we wanted to ensure fair comparison of PROCYON to our various baselines. For this purpose, we implemented a standardized evaluation framework and integrated all baselines within this framework. The desired datasets and splits are specified via a YAML config file, and similarly the models to be benchmarked are specified in a separate YAML config file alongside any relevant model-specific arguments. Each model is wrapped in a standardized API, allowing us to use the same data loading and evaluation logic across all models to ensure they're evaluated fairly and on the same set of examples.

All benchmarking of PROCYON was performed using the PROCYON-SPLIT model, and all benchmarking results use evaluation splits only, averaged across non-zero shot splits (i.e. many-shot and few shot) unless explicitly specified as zero-shot results.

Question answering evaluation. For each dataset and split, question-answer pairs are constructed in batches. Each batch consists of eight positive (protein, text) relationships, which are then supplemented with eight negative (protein, text) relationships. Negative relationships are constructed as follows: for each positive (protein, text) pair, a single protein is sampled from the set of all proteins in our overall dataset, with sampling probability according to dissimilarity to the selected protein (i.e. biasing towards less similar proteins). The sampled protein and the original text then form a negative example. In this manner, each batch contains a balance of positive and negative examples for QA evaluation.

Each batch of (protein, text) pairs is then passed to a given model following a standardized API, and the model returns a single binary prediction for each received example. Accuracy is calculated as the mean number of correct predictions across the entire dataset split. 95% confidence intervals for accuracy are calculated via bootstrapping using scipy's (v1.13.0) `bootstrap` function:

`bootstrap(n_resamples=9999, method='BCa', confidence_level=0.95).`

Retrieval evaluation. For each dataset and split, we first identify the set of (protein, text) relationships to evaluate. For a given text, we begin by treating all proteins it's related to in this split as the positive retrieval targets, and all other proteins as negative targets. To avoid conflict with training data, we remove all (protein, text) relations in the training set from consideration. For example, this avoids treating a (protein, text) relation in the train set as a negative relation at test time. This is necessary because our splits are constructed at the level of (protein, text) relations, and thus a given text can appear in multiple data splits. In this manner, we construct a cleaned set of positive and negative retrieval targets per text.

The set of texts and the full set of proteins in our overall datasets are then passed to a given model following a standardized API, and the model returns a score for each (protein, text) pair, where a higher score corresponds to higher retrieval ranking.

We evaluated each model's retrieval performance using $F_{\max}$, the maximum F_1 score achieved across all possible thresholds for separating positives and negatives. To evaluate a model's performance across the entire dataset split, we first calculate $F_{\max}$ for each text separately, using only its positive protein relations and considering all other proteins as negatives. To account for severe label imbalance (most texts relate to very few of our $\approx 18,000$ proteins), we randomly sample multiple negatives per positive sample. In particular, for each text, we sample 10 negative proteins for each positive, or if this number would exceed the total number of proteins, we use the full set of proteins as negatives. We perform this sampling 5 times and return the average of the samples as the $F_{\max}$ score for this text. The final $F_{\max}$ score for the entire dataset split is then the average of the $F_{\max}$ scores of each text. Similarly to QA, 95% confidence intervals are calculated via bootstrapping at the text level (i.e. resampling per-text $F_{\max}$ scores) using scipy's (v1.13.0) `bootstrap` function: `bootstrap(n_resamples=9999, method='BCa', confidence_level=0.95).`

4.2 Baseline model methods

We compare PROCYON against a number of baseline models to evaluate it's capabilities with respect to different types of modeling approaches. This includes five protein-only models, including sequence-only models BLAST [49] (kNN approach), GearNet [21] (both kNN and MLP approach), and ESM-2 [16] (both kNN and MLP approach). We also compare to multimodal models, including ProtST [23], BioTranslator [22], ProtLLM [24], and ESM-3 [18] (both kNN and MLP approach). Detailed information on the implementation of these baseline models is included

in Supplementary Notes 14 and 15.

4.3 PROCYON retrieval performance on varying evidence sources

These results are shown in Figure 3b. To validate the abilities of PROCYON to retrieve proteins across different diseases using as ground truth (disease, gene) associations derived from OpenTargets (OT) [91] and collected through different evidence sources (e.g., genetic associations, somatic mutations, animal models, affected pathway, literature and known drug). For each disease, we derive the direct disease-gene associations for each data source, resulting in 3372 diseases being included in the analysis. Exact details on data curation are included in Supplementary Note 16.

We query PROCYON using the following task description:
"You will be shown text describing a disease, which could be common, rare (Mendelian), complex or environmental. Your job is to determine the proteins that are associated with the disease described. The associations are provided by various expert curated resources, representing relationships where there is established evidence linking genetic variations or mutations to the occurrence, risk, or progression of diseases. The associations indicate whether a protein, through its variation or mutation, plays a role in the manifestation of a disease, either by directly causing it, serving as a marker for it, being implicated in its mechanism, or being targeted for therapeutic intervention."

We use the disease description available in OMIM as input description, and retrieve the ranked list of proteins. Finally, for each disease and source, we derive the ranked percentile for each protein in the ground truth list and calculate the median, resulting in a median percentile score per disease and source.

4.4 Disease gene prioritization with increasing evidence

This section addresses the experiment discussed in Figure 3c. The list of diseases and input prompts are derived from the Diagnostic Manual of Mental Disorders (DSM) fifth edition. The DSM provides the standard classification of mental disorders that is commonly used by mental health specialists in the United States. It provides a structured text that is characterized by multiple sections that are consistent across the majority of the diseases. For this analysis, we used only the sections "Diagnostic Features" and "Associated Features supporting Diagnosis" [52].

To select the diseases to query the model for, we first identify the subset of diseases available both in the DSM-5 and in OpenTargets (OT). We complete the mapping from DSM-5 to OT manually by using the OT online dashboard to check for diseases https://platform.opentargets.org/,

resulting in 63 diseases being included in the analysis. Then, when available, we manually extract the "Diagnostics" and "Associated Features supporting Diagnosis" descriptions for each disease, resulting in 51 diseases.

To assess the performance of PROCYON, we extract the overall ground truth direct gene-disease associations from OT (data were downloaded on the 10th June 2024) for the 51 diseases, regardless of the database and source used to derive them. We apply a threshold of 0.8 on the gene-disease associations confidence scores, in order to keep only high confidence associations. We conduct the analysis only on diseases with at least one gene-disease association that passed the threshold, resulting in 30 diseases.

Finally, we query PROCYON with a task definition that reads: 'You will be shown text describing a disease, which could be common, rare (Mendelian), complex or environmental. Your job is to determine the proteins that are associated with the disease described. The associations are provided by various expert curated resources, representing relationships where there is established evidence linking genetic variations or mutations to the occurrence, risk, or progression of diseases. The associations indicate whether a protein, through its variation or mutation, plays a role in the manifestation of a disease, either by directly causing it, serving as a marker for it, being implicated in its mechanism, or being targeted for therapeutic intervention.'. We query the model five times for each disease, each time using a different input description. We first query the model just with the disease name, the "Diagnostics Features" and the "Associated Features supporting diagnosis". Then, to increase the amount of evidence input into the model, we query PROCYON using "Disease + Diagnostics Features", and "Disease + Diagnostics Features + Associated Features supporting diagnosis". Finally, for each disease, we derive the ranked percentile for each protein in the ground truth list and calculate the median, resulting in a median percentile score per disease.

4.5 Embedding space generation

This section addresses the visualization in Figure 3d. To compare the density of proteins with the same annotation across different databases (Drugbank, EC, GO, Reactome, GTOP) before and after multimodal fine-tuning, we visualize the protein retrieval embedding space of the initialized space (based on that of ESM-2) and the tuned space of PROCYON.

First, we identify the corresponding annotation for each protein across the databases and rank the annotations per database based on how many proteins they are associated with. For the visualization, we use either the annotation with the highest or second highest number of proteins.

We select the second in the ranking in cases where the first one is overlapping across multiple databases, making it hard to visualize. Specifically, we use the annotation "Fostamatinib" for Drugbank (first in the ranking, N = 289), "Metabolism" for Reactome (second in the ranking, N = 2046), "CD molecules" for GTOP (second in the ranking, N = 46), "Extracellular space" for GO (second in the ranking, N = 461), and the annotation "RING-type E3 ubiquitin transferase" for EC (second in the ranking, N = 218).

To generate the visualization, we scale the embeddings to unit variance and zero mean. Then, we compute the principal component analysis with 30 components, the nearest neighbors distance matrix and a neighborhood graph of observations and we embed the neighborhood graph using UMAP [126]. Afterwards, we filter the embeddings to include only those of the proteins with the annotations of interest and calculate the Gaussian kernel density estimation of the proteins of interest in the protein retrieval embedded space. Finally, we overlay the estimated protein density to the UMAP visualization of the protein retrieval embedding space.

4.6 Compositional and complex protein retrieval

Retrieving pleiotropic proteins. This addresses the result in Figure 3e. For this experiment, we consider pleiotropic proteins as proteins annotated to more than one GO biological process terms that are not an ancestor of one another and do not share any offspring GO terms ("non-overlapping"). With this, we consider GO biological process terms above level 2, and curate mutually non-overlapping pairs, triplets, and quadruplets of terms that have between 3 to 20 overlapping protein annotations. Due to the vast combinatorial space, we restrict our experiment to randomly sampled 1000 pairs, 500 triplets, and 200 quadruplets (decreasing number because of permutation of queries required as specified below). For each pair, triplet, or quadruplet of signaling pathways, we query PROCYON in different ways and obtain the quantiles of the associated pleiotropic proteins.

In the "composite query" approach, we query PROCYON with a task definition defining the retrieval target, along with the descriptions of pathways for each pair, triplet, or quadruplet of pathways. Specifically, the task definition reads: "You will be shown text descriptions of two[three/four] biological processes. Your job is to determine if a protein is involved in both processes described. Involvement is determined by whether or not there is evidence from an experiment directly supporting the annotation of the protein." For each pair, triplet or quadruplet of pathways, we query the model with all permutations of pathway description combinations (2

for pairs, 6 for triplets and 24 for quadruplets). The quantiles of proteins from each query are aggregated with geometric mean.

In the "aggregated individual ranks" approach, we query PROCYON with a task definition defining the retrieval target, along with the description of one of the two, three, or four pathways taken from its mapped GO biological process term. Specifically, we adopt the same task definition as used for the biological process of GO. For each pair, triplet or quadruplet of signaling pathways, we query the model twice, three times, or four times, each time with a different pathway. The quantiles of proteins from each query are aggregated with geometric mean.

In the "random composite query" approach, we query PROCYON with a task definition defining the retrieval target, along with the descriptions of two, three, or four randomly sampled GO biological process terms from our database. The task definition is identical to the "composite query" case. For each pair of pathways, we query the model with all permutations of pathway description combinations. The quantiles of proteins from each query are aggregated with geometric mean.

In the "random pathway" approach, we query PROCYON with a task definition defining the retrieval target, along with the description of a randomly sampled GO biological process term from our database. The task definition is identical to the "aggregated individual ranks" case.

Retrieving signaling pathway crosstalk proteins.. This section describes the analysis with results in Figure 3f. We extract signaling pathway crosstalk data from [53], which is a "carefully curated database that documents the scientific literature that supports crosstalk between pairs of pathways". We ensure the data extracted are human. The proteins and pathways are mapped to proteins and GO terms in our database, respectively. In total, we obtain 84 pairs of 16 unique signaling pathways, each annotated with on average 2.8 crosstalk proteins.

For each pair of signaling pathways, we query PROCYON in different ways and obtain the quantiles of the annotated crosstalk proteins. The "composite query" approach aims at directly prioritizing proteins with desired composite role with composite queries, where PROCYON is queried with a task definition defining the retrieval target and pathway crosstalk, along with the descriptions of the two signaling pathways taken from their respectively mapped GO biological process terms concatenated with a newline character. Specifically, the task definition reads: "You will be shown descriptions of two signaling pathways. Your job is to retrieve a protein that is involved in the crosstalk between the two pathways. Crosstalk refers to the case where one or more components

of one signaling pathway affects the other." Since we do not consider directionality of crosstalk in this experiment, for each pair of signaling pathways, we query the model twice, each time with a different pathway following the other. The quantiles of proteins from each query are aggregated with geometric mean.

The "aggregated individual ranks" approach aims at prioritizing proteins with desired composite role through individual queries and combining two rank lists, where PROCYON is queried with a task definition defining the retrieval target and pathway crosstalk, along with the description of one of the two signaling pathways taken from its mapped GO biological process term. Specifically, the task definition reads: "You will be shown description of one signaling pathway. Your job is to retrieve a protein that is involved in the crosstalk between this pathway and another. Crosstalk refers to the case where one or more components of one signaling pathway affects the other." For each pair of signaling pathways, we query the model twice, each time with a different pathway. The quantiles of proteins from each query are aggregated with geometric mean.

The "random composite query" approach serves as a negative control, where PROCYON is queried with a task definition defining the retrieval target and pathway crosstalk, along with the descriptions of two randomly sampled GO biological process term from our database. The task definition is identical to the "composite query" case. For each pair of pathways, we query the model twice, each time with a different pathway following the other. The quantiles of proteins from each query are aggregated with geometric mean.

The "random pathway" approach serves as a negative control, where PROCYON is queried with a task definition defining the retrieval target and pathway crosstalk, along with the description of one randomly sampled GO biological process term from our database. The task definition is identical to the "aggregated individual ranks" case.

Retrieving STING for neuronal inflammatory stress response. This point addresses the result in Figure 3g. Using the queries in Supplementary Note 3, we query the model with increasing information about the STING pathway to observe it's ability to prioritize scientific evidence in retrieving proteins. Each level of text is permuted 100 times by randomly dropping 5% of words. All queries are subsequently fed into PROCYON to obtain the quantiles of STING along with a confidence interval on the bootstrap procedure.

For the control group of proteins, we obtain genes expressed (median TPM >= 1) in all sequenced adult CNS tissues from the GTEx Portal in May 2024 ("GTEx_Analysis_2017-06-

05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz"), and select those that are initially ranked in the top 10% (above 90% quantile) of proteins, when PROCYON is queried with the first level of evidence.

4.7 AKNAD1 Phenotype Generation

This section addresses the generation of phenotypes for AKNAD1 in Figure 4a. To generate these outputs, we use a simple phenotype generation query for PROCYON that copies the structure of the prompt used for UniProt during training. In the input prompt, we insert the protein AKNAD1 (UniProt: Q5T1N1). We use PROCYON-FULL for this prediction. This protein appears in none of our training data, i.e., there are no relations to any phenotypes for this model in PROCYON-INSTRUCT. Further, AKNAD1 has a poor functional annotation more broadly, as its UniProt annotation score is 3 / 5. This protein was chosen due to the fact that it is not in our training data, but external evidence of its function exists in the Human Protein Atlas (HPA). Thus, this is a zero-shot prediction for PROCYON on a protein that has been poorly annotated in literature.

We use diverse beam search to generate the phenotype, and we generate 10 generated phenotypes for this protein. After generating these descriptions, we perform QA filtering (see REF SECTION) to derive the top-3 captions for this protein. This gives us the three descriptions seen above "Generated Phenotype".

Next, we validate the generated phenotypes for this protein with immunofluorescent and expression evidence through the Human Protein Atlas [8] (Human Protein Atlas proteinatlas.org). For the top description, we find immunofluorescence evidence supporting that this protein is localized to the cytoskeleton. Specifically, immunofluorescent staining of AKNAD1 (antibody HPA030270, https://www.proteinatlas.org/ENSG00000130590-AKNAD1/antibody) in cell line HAP1, which is a "near-haploid human cell line derived from KBM7, a human myeloid leukemia cell line developed from a 39-year-old male patient in the blastic phase of chronic myeloid leukemia" has confirmed its localization to cytosol, microtubules, and cytokinetic bridge.

For the second description, we find two lines of gene expression evidence supporting that the protein is overexpressed in spermatids, leading to a likely conclusion of the protein being involved in spermatogenesis. In a meta-analysis of single-cell RNA sequencing literature and databases conducted by HPA [8], a group enrichment (nTPM at least four times any other cell type) of the expression of AKNAD1 is observed in both early spermatids (90.9 nTPM) and late spermatids (29.6 nTPM). In addition, inferred from RNA-seq of unfractionated tissue samples

[127], very high enrichment of AKNAD1 expression in late and early spermatids were observed (mean correlation=0.696, 0.695, respectively, with 3 reference transcripts selected to represent the two cell types, specifically CEP55, KPNA5, PBK (late spermatids); PRM1, PRM2, TNP1, respectively, data extracted through HPA [8] (Human Protein Atlas proteinatlas.org).

4.8 Comparisons to Text-only LLMs

This section addresses the comparison to Text-only LLMs in Figure 4b, 4c and 4d. To understand how adding rich and generalizable protein representations enables PROCYON to reason over proteins, we compared our model against text-only LLMs. For this evaluation, we focused on the question answering and phenotype generation tasks, and compared to state-of-the-art LLMs at a range of model sizes.

For a fair comparison, we prompt all LLMs with the same prompts used for PROCYON, but with differing methods of how the protein is encoded in text. We probe the ability of text-only LLMs to understand proteins across a range of representations by considering three possible text encodings of proteins: amino acid sequence (-AA), UniProt ID (-UP), and HGNC gene ID (-HG). In each case, we directly replace PROCYON'S protein sequence and structure tokens with the corresponding encoding. We acknowledge that the amino acid representation is likely to perform very poorly due to full protein amino acid sequences being unlikely to appear in the training corpuses of text-only LLMs; however, we also note that this is the only representation that can accommodate completely novel or unseen proteins, as the UniProt and HGNC encodings are limited to proteins which have received such IDs.

For our evaluations, we consider LLMs across three different size classes:

- Parameter count similar to PROCYON (1X size)
  - Meta-Llama-3-8B - weights obtained from https://huggingface.co/meta-llama/Meta-Llama-3-8B
  - Meta-Llama-3-8B-Instruct - weights obtained from https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
  - Both Llama models were queried using the `transformers` Python package, with text generation parameters: `max_new_tokens=300, top_p=0.9, temperature=0.7`
- Parameter count larger than PROCYON (10X size)
  - Mistral Large - queried via Microsoft Azure AI Services Chat Completions API. API

version `2024-05-01-preview`, model `mistral-large`, temperature of 0.5, and max tokens of 350.

- Parameter count significantly larger than PROCYON (>50X size)
  - GPT-4o - queried via Microsoft Azure AI Services Chat Completions API. API version `2024-05-01-preview`, model `gpt-4o`, temperature of 0.5, and max tokens of 350.

For all LLM comparisons, we use the PROCYON-FULL model to ensure that our model has exposure to as much data as possible for a better comparison to the internet-scale training corpuses of the LLMs we compare against.

QA evaluation. Due to the scale of our datasets, we have an extremely large number of QA prompts to evaluate over, even when only considering positive QA examples, i.e., question-answer pairs generated from the relations in our dataset, where the ground truth label would be positive. Since evaluating all such QA prompts would be prohibitively expensive, we subsample our datasets to maintain a reasonable cost. For each dataset and split, we randomly sample 128 question-answer pairs, where the sampled set consists of 64 positive examples and 64 negative examples. Since we have 3-4 splits per dataset, this corresponds to roughly 400-500 prompts per dataset, which strikes a reasonable balance between sample size and cost.

Given the selected prompts, we generate QA results for the selected LLMs by using the text-only prompts with proteins encoded using one of the methods described above. Our prompt explicitly states "You may only output yes or no", which PROCYON and GPT-4o mostly abide by (modulo capitalization), but other models require more post-processing. In particular, we found that Mistral Large tended to output rationales before outputting an answer in the format "Output: <answer>", whereas the Llama-3-8B models often didn't output an answer at all. To extract outputs from LLM answers, we first check if the answer begins with "[Yy]es" or "[Nn]o" and use that as the answer if so, otherwise we use the regex `''Output: (\w+)"` to attempt to recover a yes or no answer. We find that this approach identifies a yes or no answer for all PROCYON responses, all GPT-4o responses, and all but 2 Mistral Large responses, with Llama-3-8B models frequently outputting nonsensical text.

Given the extracted yes or no answers from each LLM, we are then able to proceed with QA evaluation as performed for our baseline models, including bootstrapping for 95% confidence

interval calculation. For Llama-3-8B models in particular, we treat outputs without a yes or no answer as incorrect, resulting in binary accuracies below 50%.

Phenotype generation evaluation. Similar to QA, we have to downsample our datasets for evaluating text-only LLM's phenotype generation performance. To ensure we sample proteins with a range of representation in scientific literature (and by proxy, representation in an internet-scale LLM training corpus), we use UniProt annotation scores as a proxy for how well-studied a given protein is. UniProt annotation scores range from 1 (least studied) to 5 (most studied). We select three of our datasets for evaluation: UniProt, Reactome, and DisGeNet, which each represent distinct knowledge domains. Within each dataset, we select all proteins with annotation scores of 1 or 2, and randomly sample up to 50 proteins each per annotation score of 3, 4, and 5, resulting in a total of 600 proteins across all three datasets.

For phenotype generation from PROCYON, we use Diverse Beam Search [118] with 5 beam groups each of size 2, resulting in a total of 5 generated phenotypes per prompt. We then select amongst these captions using our QA filtering process, in which each caption is passed back to the model alongside the corresponding protein as a question answering prompt. The resulting probability of the "yes" token is then used as a confidence metric for the caption, and for each protein, we retain the single generated phenotype with the highest confidence.

Text generation for text-only LLMs is performed using the same prompt as is used for PROCYON, except with the protein token replaced with the corresponding textual representation of the protein. Text generation parameters are as listed above.

Given a set of generated phenotypes for a single protein, we generate reference-based performance metrics compared to the set of ground-truth annotation texts from the corresponding database. As a single protein may be associated with multiple phenotypes (e.g., a target of multiple drugs, or association with multiple diseases), a protein can have multiple reference phenotypes to compare against. To capture the case in which the generated text matches one association, but none of the others, we calculate metrics against all reference texts, and take the max of these values as the single metric by which to evaluate the generated text. For the performance metric, we use BERTScore F1 [128], an extension of n-gram matching metrics that also accounts for semantic similarity between words rather than requiring exact matches (e.g. the words "support" and "maintain" can be a high-scoring match). We calculate BERTScore with the parameters `rescale_with_baseline=True, lang="en-sci"`, and the hashcode `allenai/`

`scibert_scivocab_uncased_L8_no-idf_version=0.3.12(hug_trans=4.31.0)-rescaled`. Given a single BERTScore F1 metric for each protein in the evaluated dataset split, we then calculate 95% confidence intervals via bootstrapping as previously described.

LLM-based holistic evaluation. To perform LLM-based ranking, we provide the judge LLM with the original phenotype generation prompt, the generated text from PROCYON, the generated text from another model, and the set of reference texts for the given protein. We prompt the model to select which of the two models' outputs is more correct relative to the provided reference texts, or to select a tie if both models' outputs seem equally correct (or incorrect).

To combat the issue of positionality bias, in which the judge is biased towards either the first or second model, we provide each comparison twice, once with PROCYON first and once with PROCYON second. We then only use results in which the judge's decision is consistent across both permutations. In the case where the judge selects a tie in one permutation, and a winner in the other, we use the single winner as the result of the comparison.

To avoid self-bias from any of the text-only LLMs we compare against, we use a different model for judging, Anthropic's Claude-3.5-Sonnet.

The full prompt used to generate the LLM-based evaluations is provided in Supplementary Note 17.

4.9 Domain prioritization for drug targets

In Figure 5a,b, we show how PROCYON can be used to find the binding domain for drugs given the protein targeted by the drug. We extract experimentally validated domain-drug interaction data ("DrugDomain_v1.0_data_PDBs_ECODv291.txt") from [61]. Specifically, we consider only domains families and proteins in our database. As some proteins have multiple sections of the same domain family, we map the domains to specific domains on the protein by their positions on the protein, as provided by `pdb_start` and `pdb_end` in the data. To account for annotation differences across protein databases, we accept a mapping as long as `pdb_start` is before the start as annotated in our domain database, while `pdb_end` is after the end. Interacting residues are further compared to obtain the shift, when possible. This processing workflow leaves us with 521 samples (pairs of drugs and domains). The final samples are (drug, protein) pairs with a mapping from the protein to a domain on which the drug binds; thus, each sample contains a set of domains $(D_1, \ldots, D_i^*, \ldots, D_N)$, where all D_i are domains and D_i^* is the binding domain for the drug.

After these processing steps, we then filtered by a few criteria. First, we filtered for all

samples where there were at least 2 Pfam domains that did not bind to the drug for that given sample. Thus, all samples considered contain proteins have at least three domains, including the binding domain. Second, for simplicity, we filter out all instances where multiple domains bind the drug. After both filtering steps, we're left with 35 samples total.

We construct input instructions in the same manner as for drug-target retrieval during training, using identical instruction inputs. However, when we compute the retrieval scores, we score them over the set of domains for the given protein. Please see Methods Sec. 2.3 for more information. For computing results, we calculate the absolute rank (an integer, where 1 is the best) of the binding domain in the prioritized list. The ranking is determined by the similarity score output by the model. We calculate these ranks across all 35 samples in the dataset and plot them in a bar plot in Figure 5b. To aid in understanding of the analysis, we also show a few more examples of lists with the correct binding domains, including one where PROCYON misidentifies the binding domain as rank 2 in the list.

We further compare our results to a simple random reference. Importantly, we filter for multi-domain proteins in this experiment, so a random ranking, i.e., a model that does not represent understanding of drug-domain binding properties, is unlikely to produce a similar ranking to that of PROCYON, given the number of domains for each protein. Therefore, the random reference takes the expected rank of the correct binding domain (the middle of the list, i.e., the list size divided by two), and subtracts the actual rank produced by PROCYON from it. Thus, we see that Rank 1 proteins are on average 2.5 ranks above the expected random rank. Full results for this experiment are included in Supplementary Table 24.

4.10 Protein-peptide binding prediction

In this analysis, shown in Figure 5c, we examine the transferability of PROCYON to a novel task and therapeutic modality: protein-peptide binding prediction. Further, we test PROCYON on synthetic peptides, demonstrating that the model generalizes to large distribution shifts at test time. The results for this experiment are shown in Figure 5c.

First, we curated a database of protein-peptide physical interactions from PDBBind. Specifically, we followed the training set evaluation from CAMP framework [66]. This dataset contains 1565 tuples of proteins-peptides, containing the conformations of the complexes, as well as the corresponding binding affinity scores. First, we filtered the original dataset, so that it only consists of pairs of proteins-peptides, since a subset of the samples consists of more than two structure

components. This resulted to a total of 1202 protein-peptide samples. Moreover, we included the union of inhibition constant K_i, dissociation constant K_d, and half-maximal inhibitory concentration $IC50$ values for each protein-peptide complex. It is noteworthy that a subset (680) of peptides in the protein-peptide dataset have their own accession Uniprot ID. In these subset of samples, the interactions were treated as PPI relations. After filtering, this dataset consists of 1097 examples of protein-peptide complexes. The use of this database is notable as PROCYON is not trained explicitly on physical binding data during training. PROCYON is trained on protein interactions from STRING database; however, these interactions and similarities are limited to homology, coexpression, and experimental interaction data, so physical binding data is needed to tune the model for this specific task.

Second, we finetune PROCYON-FULL on this data. We design interaction instruction templates for both QA and retrieval in the same manner as for STRING datasets during training (see Methods Sec. 3.1 for more information). Thus, PROCYON is tuned on only QA and retrieval with the PDB binding data to produce PROCYON-BIND, which predicts physical protein-peptide binding interactions. PROCYON-BIND is trained for 6.54 epochs for both QA and retrieval tasks on the PDB binding dataset. More details about this model are included alongside hyperparameter reports for other versions of PROCYON (Supplementary Information).

Next, we curate a dataset of interactions between synthetic peptides and angiotensin converting enzyme 2 (ACE2). Our dataset, previously generated by Zhang et al. [67], consists of 58 detected binders for ACE2 and 5072 non-binders taken from a large randomized library of synthetic peptides generated with similar methods gathered by Brown et al. [71]. All peptides are 13 amino acids long due to the library's synthetic construction.

As an initial test of this dataset, we perform testing of the peptides on the original ESM-2 3B model [16]. To test ESM-2's capability to identify these peptides, we test use the proximity of the embedding of each peptide to the embedding of ACE2 as a binding prediction score. ESM-2 achieves a very poor 0.3923 AUROC. This matches results seen from Brown et al. [71], which also found the identification of these ACE2 binders to be challenging for unsupervised machine learning models.

Next, inference is performed to rank peptides for their likelihood of binding to ACE2. Our prompts for both QA and retrieval for this task are identical to that used during training on the PDB Binding data, including the same in-context examples and the same task definition (see Methods Sec. 3.1). We perform both QA prediction and retrieval prediction for the peptides included in the

ACE2-binding set. Importantly, the binding nature of each peptide (i.e., if it is a binder or not) is not given to the model during inference but rather is predicted by the model. For QA scores, we take the difference between the "yes" and "no" token outputs as the singular score of the model; this normalizes the score to a range between [-1,1] where a lower score is considered less likely by the model to be an ACE2 binder and higher is considered more likely to be an ACE2 binder.

For retrieval, our query takes in ACE2 as the input and the retrieval embedding for the prompt is compared to all embeddings of the peptides, thereby calculating the retrieval similarity score comparing the input to all candidate peptides. Each peptide is then ranked by its similarity score, where a higher similarity score corresponds to a prediction of being more likely to be a binder and a lower similarity score is less likely to be a binder. As this procedure follows our typical retrieval inference procedure, we refer the reader to Methods Sec. 2.3 for more details.

We then combine the QA and retrieval results to arrive at the final binding score. The rationale for this comes from the geometry of the latent space whereby there is no observable linear separation between binders and non-binders, thereby hindering the maximum capacity of retrieval in this case. Such lack of separability is most likely due to the challenging nature of this task and that this set of synthetic peptides is indeed out-of-distribution compared to the PDB-binding dataset used to finetune PROCYON-BIND. QA is not limited by this lack of linear separability, but we retain the retrieval scores since retrieval performs very well during training of PROCYON-BIND and thus should recover some signal of the binders. To combine QA and retrieval scores, we first normalize each score to $[0, 1]$ based on min-max normalization; then we take the mean of each score, i.e., $(s_{\text{QA}} + s_{\text{ret}})/2$. This produces the final combined "Predicted Binding Score" from PROCYON-BIND. We then plot the distributions of scores for binders and non-binders beside each other and conduct a Mann-Whitney U test for statistical significance. The final results of the separation gave 0.6480 AUROC for separation of binders and non-binders for the ACE2 dataset while the Mann-Whitney U test showed that PROCYON-BIND achieved a statistically-significant separation of binders and non-binders ($p = 1.04e - 4$).

Limitations. In this analysis, we do not claim PROCYON-BIND is a state-of-the-art protein-peptide prediction model. Many models exist for prediction of protein-peptide binding, including AlphaFold3 [15]. Rather, this analysis demonstrates the transferability of PROCYON to new tasks, including those for which it was not trained and on an entirely new modality than its training set, i.e., peptides.

4.11 Variant prediction with PROCYON

In Figure 5de, we study how PROCYON can generalize to new variants of proteins and potentially explain the biological meaning of variants.

Descriptions of variant effects with PROCYON. In this analysis we examine the ability of PROCYON to [129] describe the changes made by mutations in a single protein and [130] correlate the changes in generated captions with biology. To do so, we focus on mutations for Alzheimer's Disease (AD). Specifically we take a list of current benign and pathogenic AD mutations, mutate the wildtype protein with each mutation, and generate phenotypes for the resulting sequence. We only take into account single nucleotide polymorphisms for mutations and apply one mutation at a time. We hypothesize that there will be a meaningful difference in the captions generated for benign variants and pathogenic variants.

We retrieved all variants in Clinvar [131] that were associated with Alzheimer's and filtered variants to those with at least one star (1191 variants). We then took variants with the following germline classification: "Benign", "Likely benign", "Pathogenic", "Likely pathogenic" (535 variants, 380 benign, 155 pathogenic). We mutated each wildtype protein and passed it into PROCYON to generate 100 captions for GO-process and also passed the generated captions through a QA filter generating a probability for how likely PROCYON believes the generated caption accurately reflects the input protein sequence. For each protein, we divided benign and pathogenic sequences and used a BLEU score threshold of 0.7 to filter out captions that appeared in both benign and pathogenic sequences. We then manually review captions to find meaningful differences between benign and pathogenic sequence captions.

The genes considered are MAPT, PRNP, PSEN1, APP, PSEN2, CALHM1, ABCD1, LLRRK2, TREM2, VCP, GBE1, APOE, and CSF1R. For PSEN2, we observed that benign variants had the following caption, "Any process that activates or augments the rate or extent of calcium ions (Ca2+) into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore." while pathogenic variants did not have this caption. For APP (Amyloid Precursor Protein), pathogenic variants produce captions which discuss heat shock proteins [130,132], Reg-family proteins [133], and nucleotide exchange factors [134]. Benign variants of APP produce captions discussing phospholipids and transport proteins, which are never mentioned in pathogenic variant captions.

Correlating variant descriptions with protein structure. In this analysis, we explore how PRO-

CYON captions correlate with protein structure by introducing point mutations into a known structure and observing caption changes. Focusing on PSEN1, a key protein in Alzheimer's disease (AD), we apply single nucleotide polymorphisms (one at a time) associated with AD to the PSEN1 wildtype amino acid sequence. We then compare the resulting captions to those of the wildtype protein, identifying unique changes. For ease of analysis we identify functional keywords described by the captions instead of the raw words and use these for analysis.

To generate captions we asked PROCYON to create a caption for the sequence using GO-Process. We then passed the generated caption with the input sequence into a QA filter where we asked PROCYON to give a probability that the caption was appropriate for the input sequence. If the probability of a "Yes" was greater than 0.7, we accepted the caption. If not, the caption was discarded. We repeated the process until 20 accepted captions were generated for each input sequence. We then used a BLEU score to compare captions to the wild-type captions, using a filter of 0.3 to filter out similar captions.

We use an atomic structure and AD-relevant variants of PSEN1 from a study by Bai et al. [75]. PSEN1 is organized in multiple trans-membrane (TM) sections numbered from one to nine. We applied and generated captions for the following mutations in PSEN1: I143V (TM2) , L226F (TM5), M233L (TM5), F237I (TM5), L248R (TM6), I213L (TM4), I202F (TM4), V261F (TM6), and F177L (TM3). The mutations clustered into four groups by caption: Group 1 contains F177L (TM3), Group 2 contains V261F (TM6), Group 3 contains V261F (TM6), and Group 4 contains I143V (TM2), L226F (TM5), M233L (TM5), F237I (TM5), and L248R (TM6). Captions in group 1 described regulation of cell interactions, captions in group 2 described ion transport, regulate cell interaction, and proteolytic activity, captions in group 3 described ion transport, proteolytic activity, and sodium channel transport, and captions in group 4 describe ion transport, proteolytic activity, and cell to cell communication.

Limitations. We do not claim that PROCYON is a state-of-the-art variant effect prediction model. Rather, this analysis shows the potential of a framework such as PROCYON, which bridges explicit multimodal protein representations with natural language generation, to yield deep insight into the effects of variants. In order to improve the variant effect prediction capabilities of PROCYON, one would need to train on variant data such as that from ClinVar [131] or other databases such as those used by popular models such as AlphaMissense [135].

4.12 Retrieving with compositional queries across knowledge domains

In this analysis shown in Figure 5f, we examine the abilities of PROCYON to perform composite queries across knowledge domains, performing drug target retrieval that is conditional on the indication of the drug, i.e., the disease that the drug is treating.

First, we searched for drugs which roughly matched the following categorization: indicated for at least two diseases for which those diseases had distinct protein targets. Thus if Drug X is indicated for Disease A and Disease B, Drug X meets our criteria if there exists two targets Protein 1 and Protein 2 where Protein 1 is associated with Disease A but not Disease B and Protein 2 is associated with Disease B but not Disease A. From this, we found Bupropion, a norepinephrine and dopamine reuptake inhibitor that is used to treat major depressive disorder (MDD) and smoking cessation, among others (see Drugbank entry [65]). The drug targets both the norepinephrine receptor (NET, UniProt: P23975), which is indicated for treatment of MDD, and the cholinergic receptor nicotinic alpha 3 subunit (AChR, CHRNA3, UniProt: P32297), which is indicated distinctly in treating nicotine addiction (i.e., would be targeted for smoking cessation). We also included the analysis of a third target, dopamine receptor DAT (UniProt: Q01959), which is indicated in both smoking cessation and MDD (i.e., the impact is not necessarily distinct).

Second, we gathered the appropriate textual information for constructing the query. We gathered descriptions of Nicotine addiction and MDD from StatPearls, focusing on molecular information about the disease, i.e., not clinical diagnosis information. On August 08, 2024, we retrieved the StatPearls description for Depression from the "Pathophysiology" section on the webpage entry [136]. We used the depression entry [84] instead of MDD because the MDD description [136] does not provide explicitly a "Pathophysiology" section. In addition, since MDD corresponds to a subtype of depression [52] with a relatively strong biological component, the "Pathophysiology" section of depression was almost exclusively focused on MDD. On 08/08/2024, we retrieved the StatPearls description for Nicotine addiction [83]. The exact descriptions used in the forward pass to the model are given in Supplementary Note 18.

4.13 Analyzing poorly-characterized proteins for Parkinson's Disease

This section encompasses the analysis described in Figure 6, where we study the pathways predicted by PROCYON for poorly characterized proteins that are hypothesized to have association in Parkinson's disease. Analysis of these experiments is described in Results.

Gathering the PD-uncharacterized list. We seek to study the ability of PROCYON to uncover phenotypes of proteins that are poorly functionally characterized in literature. To get this list, we gather all proteins in OpenTargets that are hypothesized to be associated to Parkinson's disease (PD) [91] (Accessed on the 12th August 2024). We filter the list at those above 0.1 association score threshold. We then cross-reference this list with the proteins in our database, and we gather the proteins in the list from OpenTargets that do not appear in any of our training databases. This creates the PD-uncharacterized list, a list of 36 proteins with a link to Parkinson's disease that are considered **zero-shot examples** for PROCYON since it was not trained on any phenotypes for these given proteins. This full list is given in Supplementary Table 25.

Generation of descriptions with PROCYON. We follow the standard pipeline to generate descriptions of the proteins in PD-uncharacterized. We use prompts in the style of GO biological process to generate descriptions of the pathways in which each protein might be involved. We use QA filtering with a threshold of 0.6 to get the final list of descriptions for each protein.

Automatic clustering. The automatic clustering is performed by grouping together similar descriptions that have a high n-gram overlap similarity. First, we take ROUGE-L scores for every pair of generated descriptions $\{(d_i, d_j)|d_i, d_j \in \mathcal{D}\}$, resulting in a $(|\mathcal{D}|, |\mathcal{D}|)$ distance matrix. This distance matrix is then converting into a binary adjacency matrix by taking all distances greater than 0.95 (which are considered to be "copies" by ROUGE-L) as edges between descriptions. A connected components algorithm is then performed on the resulting graph to find all connected components, i.e., groups of descriptions that are similar to each other. This results in 9 clusters, with an minimum membership of 3 and a maximum membership of 7. These clusters are then used to collapse down the number of descriptions for each of the proteins in the PD uncharacterized set.

Manual clustering and curation of pathways. The protein descriptions that are not assigned to a cluster through the automatic approach are manually evaluated by two experts. The descriptions are annotated independently by two experts and assigned a cluster label (i.e., a pathway) according to currently available literature (see Supplementary Note 20). The two independent annotations are examined for consensus, and a unique label is then derived. The same approach is followed to assign pathway labels to the automatically derived cluster, where the group descriptions rather than individual protein descriptions are used. Finally, the proteins with the same manually assigned pathway annotation are grouped together.

Enrichment testing. To test quantitatively whether the pathways generated by PROCYON are

implicated in PD, we test pathway enrichment among a gene list of well-known PD-associated genes ("PD-Associated") to two control groups: one of neurologically-expressed genes ("Neuro control") and one of genes expressed outside of neurological tissue ("General control") . Exact details on these lists is included in Supplementary Note 19.

We then use these lists to compute the enrichment of each pathway derived from PD-uncharacterized within the lists. First, we calculate a proportion expressing the coverage of each GO term by the proteins in the list. The proportion is calculated as the number of proteins in the given protein list $\mathcal{P}$ that are associated to the cluster's assigned GO term $\mathcal{G}_i$ over the size of that GO term:

$$\text{Proportion}(\mathcal{G}_i, \mathcal{P}) = \frac{\sum_{p_j \in \mathcal{P}} \mathbf{1}_{\mathcal{G}_i}(p_j)}{|\mathcal{G}_i|}; \mathbf{1}_{\mathcal{G}_i}(p_j) = \begin{cases} 1 & \text{if } p_j \in \mathcal{G}_i \\ 0 & \text{else.} \end{cases} \tag{7}$$

By dividing by the size of the GO term as opposed to the size of the control list, this formula considers the "specificity" of the pathway, i.e., how many proteins within the pathway are found in these given lists. We can then directly compare the number of proteins in a given list to another list by subtraction, i.e., $\text{Proportion}(\mathcal{G}_i, \mathcal{P}_1) - \text{Proportion}(\mathcal{G}_i, \mathcal{P}_2)$. We calculate this difference for PD-associated vs. the neuro and general lists, respectively. We then show these differences of proportions as bars in Figure 6b.

To balance the statistical comparison we perform downsampling of both the neuro and general control lists during computation of the proportions. This is built off the intuition that a large randomly-gathered list of genes would have a high likelihood of including the associated proteins in the GO term since our proportion calculation (Equation 7) does not account for size of the list $\mathcal{P}$. Thus, we randomly sample each control list to the size of the PD-associated list (962) 30 separate times, then we take the median proportion from those 30 separate trials.

To compare statistical significance of the proportions, we perform a one-sided proportion z-test where the alternative hypothesis is that the proportion in the PD-associated list is higher than in the control list. We show the results of this significance test in Figure 6b as well as in Supplementary Tables 26 and 27. In Figure 6b, we highlight the bar in the barplot as green if the difference in proportions is indeed significant.

Describing poorly-annotated proteins identified through alpha-synuclein screens. Next, we examine a list of proteins that were discovered to be involved with the molecular pathways leading to the toxic effect of α-synuclein. We first gathered data from a study that ran α-synuclein

toxicity screens on yeast and then transferred the discovered genes based on homologous proteins in humans. Specifically we gathered the list of proteins from Supplementary table S9 in Khurana et al. [86]. Then, we filtered for proteins that were not present in our training data, i.e., proteins on which our model had seen no linkage to phenotypes. This left us with only four proteins, ANKDD1A (UniProt: Q495B1), SVOP (UniProt: Q8N4V2), TCEAL5 (UniProt: Q5H9L2), and ZNF654 (UniProt: Q8IZM8). We then generated descriptions in both the style of the UniProt database (i.e., general functions), and in the style of GO biological process. We selected a few descriptions to show in Extended Data Figure 9. We show the protein structures as well as generated descriptions below the structure and sequence of each protein. Protein structures are gathered from AlphaFoldDB [137], and we include the pLDDT scores (i.e., confidence of the structure at a per-residue granularity level) to illustrate the poor annotation of the proteins from a structure perspective.

References

1. Pawłowski, K. Uncharacterized/hypothetical proteins in biomedical 'omics' experiments: is novelty being swept under the carpet? *Briefings in Functional Genomics and Proteomics* **7**, 283–290 (2008).

2. Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods* **19**, 774–779 (2022).

3. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. The structural coverage of the human proteome before and after alphafold. *PLoS computational biology* **18**, e1009818 (2022).

4. Kustatscher, G. *et al.* An open invitation to the understudied proteins initiative. *Nature Biotechnology* **40**, 815–817 (2022).

5. Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).

6. Schaffer, L. V. & Ideker, T. Mapping the multiscale structure of biological systems. *Cell systems* **12**, 622–635 (2021).

7. Zheng, F. *et al.* Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).

8. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

9. Li, M. M. *et al.* Contextual ai models for single-cell protein biology. *Nature Methods* **21**, 1546–1557 (2024).

10. Kwon, J. J., Pan, J., Gonzalez, G., Hahn, W. C. & Zitnik, M. On knowing a gene: A distributional hypothesis of gene function. *Cell Systems* (2024).

11. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nature methods* **10**, 221–227 (2013).

12. Zhou, N. *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology* **20**, 1–23 (2019).

13. Omenn, G. S. *et al.* The 2022 report on the human proteome from the hupo human proteome project. *Journal of Proteome Research* **22**, 1024–1042 (2022).

14. Omenn, G. S. *et al.* The 2023 report on the proteome from the hupo human proteome project. *Journal of Proteome Research* **23**, 532–549 (2024).

15. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 1–3 (2024).

16. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

17. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

18. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model 2024.07.01.600583 (2024).

19. Somnath, V. R., Bunne, C. & Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems* **34**, 25244–25255 (2021).

20. McDermott, M. B., Yap, B., Szolovits, P. & Zitnik, M. Structure-inducing pre-training. *Nature Machine Intelligence* **5**, 612–621 (2023).

21. Zhang, Z. *et al.* Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations* (2023).

22. Xu, H., Woicik, A., Poon, H., Altman, R. B. & Wang, S. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications* **14**, 738 (2023).

23. Xu, M., Yuan, X., Miret, S. & Tang, J. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, 38749–38767 (PMLR, 2023).

24. Zhuo, L. *et al.* ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8950–8963 (Association for Computational Linguistics, Bangkok, Thailand, 2024).

25. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

26. Consortium, T. U. *et al.* Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).

27. Seal, R. L. *et al.* Genenames. org: the hgnc resources in 2023. *Nucleic Acids Research* **51**, D1003–D1009 (2023).

28. Liang, P. P., Wu, C., Morency, L.-P. & Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, 6565–6576 (PMLR, 2021).

29. Algaba, A. *et al.* Large language models reflect human citation patterns with a heightened citation bias. *arXiv:2405.15739* (2024).

30. Gallifant, J. *et al.* Language models are surprisingly fragile to drug names in biomedical benchmarks. *arXiv preprint arXiv:2406.12066* (2024).

31. Kratz, A. *et al.* A multi-scale map of protein assemblies in the dna damage response. *Cell Systems* **14**, 447–463 (2023).

32. Kong, J. *et al.* Prediction of immunotherapy response using mutations to cancer protein assemblies. *Science Advances* **10**, eado9746 (2024).

33. Lytle, N. K. *et al.* A multiscale map of the stem cell state in pancreatic adenocarcinoma. *Cell* **177**, 572–586 (2019).

34. Porter, L. L. *et al.* Many dissimilar nusg protein domains switch between α-helix and β-sheet folds. *Nature Communications* **13**, 3802 (2022).

35. Chakravarty, D. *et al.* Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications* **15**, 7296 (2024).

36. Pillai, A. *et al.* De novo design of allosterically switchable protein assemblies. *Nature* **632**, 911–920 (2024).

37. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* 1–13 (2024).

38. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024).

39. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

40. Szklarczyk, D. *et al.* The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638–D646 (2023).

41. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

42. Wei, J. *et al.* Finetuned language models are zero-shot learners (2022). ArXiv:2109.01652 [cs].

43. Chung, H. W. *et al.* Scaling instruction-finetuned language models (2022). ArXiv:2210.11416 [cs].

44. Dai, W. *et al.* Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (Curran Associates Inc., Red Hook, NY, USA, 2024).

45. Honovich, O., Scialom, T., Levy, O. & Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor (2022). ArXiv:2212.09689 [cs].

46. Longpre, S. *et al.* The flan collection: Designing data and methods for effective instruction tuning (2023). ArXiv:2301.13688 [cs].

47. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

48. Ittisoponpisan, S., Alhuzimi, E., Sternberg, M. J. & David, A. Landscape of pleiotropic proteins causing human disease: structural and system biology insights. *Human Mutation* **38**, 289–296 (2017).

49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

50. Ko, Y. S., Parkinson, J. & Wang, W. Benchmarking text-integrated protein language model embeddings and embedding fusion on diverse downstream tasks. *bioRxiv* (2024). https://www.biorxiv.org/content/early/2024/08/26/2024.08.24.609531.full.pdf.

51. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations* (2023).

52. American Psychiatric Association, D., American Psychiatric Association, D. *et al. Diagnostic and statistical manual of mental disorders: DSM-5*, vol. 5 (American psychiatric association Washington, DC, 2013).

53. Sam, S. A., Teel, J., Tegge, A. N., Bharadwaj, A. & Murali, T. Xtalkdb: a database of signaling pathway crosstalk. *Nucleic Acids Research* **45**, D432–D439 (2017).

54. Woo, M. S. *et al.* Sting orchestrates the neuronal inflammatory stress response in multiple sclerosis. *Cell* **187**, 4043–4060.e30 (2024).

55. Goodacre, N. F., Gerloff, D. L. & Uetz, P. Protein domains of unknown function are essential in bacteria. *MBio* **5**, 10–1128 (2014).

56. Bateman, A. *et al.* The pfam protein families database. *Nucleic acids research* **32**, D138–D141 (2004).

57. Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36**, 46595–46623 (2023).

58. Anthropic. Claude 3.5 sonnet (2024).

59. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

60. Tseng, Y.-M. *et al.* Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171* (2024).

61. Medvedev, K. E., Schaeffer, R. D. & Grishin, N. V. Drugdomain: The evolutionary context of drugs and small molecules bound to domains. *Protein Science* **33**, e5116 (2024).

62. Ren, L., Cao, X., Geng, P., Bai, F. & Bai, G. Study of the inhibition of two human maltase-glucoamylases catalytic domains by different α-glucosidase inhibitors. *Carbohydrate research* **346**, 2688–2692 (2011).

63. Roy, D., Kumar, V., Acharya, K. K. & Thirumurugan, K. Probing the binding of syzygium-derived α-glucosidase inhibitors with n-and c-terminal human maltase glucoamylase by docking and molecular dynamics simulation. *Applied biochemistry and biotechnology* **172**, 102–114 (2014).

64. Shaer, D. A., Musaimi, O. A., Albericio, F. & de la Torre, B. G. 2023 fda tides (peptides and oligonucleotides) harvest. *Pharmaceuticals (Basel)* **17**, 243 (2024). Published: 2024-02-13.

65. Knox, C. *et al.* Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research* **52**, D1265–D1275 (2024).

66. Lei, Y. *et al.* A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications* **12**, 5465 (2021).

67. Zhang, G. *et al.* Rapid de novo discovery of peptidomimetic affinity reagents for human angiotensin converting enzyme 2. *Communications Chemistry* **5**, 8 (2022).

68. Wang, Q. *et al.* Structural and functional basis of sars-cov-2 entry by using human ace2. *Cell* **181**, 894–904 (2020).

69. Lan, J. *et al.* Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature* **581**, 215–220 (2020).

70. Jia, H., Neptune, E. & Cui, H. Targeting ace2 for covid-19 therapy: opportunities and challenges. *American journal of respiratory cell and molecular biology* **64**, 416–425 (2021).

71. Brown, J. S. *et al.* Unsupervised machine learning leads to an abiotic picomolar peptide ligand (2023).

72. Cai, Y., An, S. S. A. & Kim, S. Mutations in presenilin 2 and its implications in alzheimer's disease and other dementia-associated disorders. *Clinical interventions in aging* 1163–1172 (2015).

73. Pizzo, P. *et al.* Presenilin-2 and calcium handling: molecules, organelles, cells and brain networks. *Cells* **9**, 2166 (2020).

74. Ge, M. *et al.* Role of calcium homeostasis in alzheimer's disease. *Neuropsychiatric Disease and Treatment* **18**, 487 (2022).

75. Bai, X.-c. *et al.* An atomic structure of human γ-secretase. *Nature* **525**, 212–217 (2015).

76. Hurley, E. M., Mozolewski, P., Dobrowolski, R. & Hsieh, J. Familial alzheimer's disease-associated psen1 mutations affect neurodevelopment through increased notch signaling. *Stem cell reports* **18**, 1516–1533 (2023).

77. Zhou, B. *et al.* Notch signaling pathway: architecture, disease, and therapeutics. *Signal transduction and targeted therapy* **7**, 95 (2022).

78. Wu, B., Yamaguchi, H., Lai, F. A. & Shen, J. Presenilins regulate calcium homeostasis and presynaptic function via ryanodine receptors in hippocampal neurons. *Proceedings of the National Academy of Sciences* **110**, 15091–15096 (2013).

79. U.S. Food and Drug Administration. Label for WELLBUTRIN SR (NDA No. 020358, Supplement No. 068) (2024). Accessed: August 14, 2024.

80. U.S. Food and Drug Administration. Label for ZYBAN (NDA No. 020711, Supplement No. 052) (2021). Accessed: August 14, 2024.

81. Shalabi, A. R., Walther, D., Baumann, M. H. & Glennon, R. A. Deconstructed analogues of bupropion reveal structural requirements for transporter inhibition versus substrate-induced neurotransmitter release. *ACS Chemical Neuroscience* **8**, 1397–1403 (2017).

82. Clark, A. *et al.* Bupropion mediated effects on depression, attention deficit hyperactivity disorder, and smoking cessation. *Health Psychology Research* **11** (2023).

83. Aslam, S. P., Leslie, S. W. & Morris, J. Nicotine addiction and smoking: Health effects and interventions. In: StatPearls [Internet]. StatPearls Publishing (2024). Available from: https://www.ncbi.nlm.nih.gov/books/NBK537066/.

84. Arif, H. & Suma P., C. Depression. In: StatPearls [Internet]. StatPearls Publishing (2024). Available from: https://www.ncbi.nlm.nih.gov/books/NBK430847.

85. Stefanis, L. -Synuclein in Parkinson's Disease. *Cold Spring Harbor Perspectives in Medicine* **2**, a009399 (2012). Publisher: Cold Spring Harbor Laboratory Press.

86. Khurana, V. *et al.* Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. *Cell systems* **4**, 157–170 (2017).

87. Cooper, A. A. *et al.* -Synuclein Blocks ER-Golgi Traffic and Rab1 Rescues Neuron Loss in Parkinson's Models. *Science* **313**, 324–328 (2006). Publisher: American Association for the Advancement of Science.

88. Sidhu, A., Wersinger, C. & Vernier, P. -Synuclein regulation of the dopaminergic transporter: a possible role in the pathogenesis of Parkinson's disease. *FEBS Letters* **565**, 1–5 (2004).

89. Dagra, A. *et al.* -Synuclein-induced dysregulation of neuronal activity contributes to murine dopamine neuron vulnerability. *npj Parkinson's Disease* **7**, 1–22 (2021). Publisher: Nature Publishing Group.

90. Tiwari, P. C. & Pal, R. The potential role of neuroinflammation and transcription factors in Parkinson disease. *Dialogues in Clinical Neuroscience* **19**, 71–80 (2017).

91. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353–D1359 (2023).

92. Sweeney, P. *et al.* Protein misfolding in neurodegenerative diseases: implications and strategies. *Translational neurodegeneration* **6**, 1–13 (2017).

93. Zhang, W., Xiao, D., Mao, Q. & Xia, H. Role of neuroinflammation in neurodegeneration development. *Signal transduction and targeted therapy* **8**, 267 (2023).

94. Tan, J. M., Wong, E. S. & Lim, K.-L. Protein misfolding and aggregation in parkinson's disease. *Antioxidants & redox signaling* **11**, 2119–2134 (2009).

95. Consortium, T. G. O. *et al.* The gene ontology knowledgebase in 2023. *GENETICS* **224**, iyad031 (2023).

96. Milacic, M. *et al.* The reactome pathway knowledgebase 2024. *Nucleic Acids Research* **52**, D672–D678 (2024).

97. McDonald, A. G., Boyce, S. & Tipton, K. F. Explorenz: the primary source of the iubmb enzyme list. *Nucleic Acids Research* **37**, D593–D597 (2009).

98. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

99. Paysan-Lafosse, T. *et al.* Interpro in 2022. *Nucleic acids research* **51**, D418–D427 (2023).

100. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* gkz1021 (2019).

101. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. Omim.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research* **47**, D1038–D1043 (2019).

102. Wishart, D. S. *et al.* Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

103. Harding, S. D. *et al.* The iuphar/bps guide to pharmacology in 2024. *Nucleic Acids Research* **52**, D1438–D1449 (2024).

104. Zhou, C. *et al.* Lima: Less is more for alignment (2023). ArXiv:2305.11206 [cs].

105. Chan, S. C. Y. *et al.* Data distributional properties drive emergent in-context learning in transformers (2022). ArXiv:2205.05055 [cs].

106. Xie, S. M. *et al.* Doremi: Optimizing data mixtures speeds up language model pretraining (2023). ArXiv:2305.10429 [cs].

107. Shin, S. *et al.* On the effect of pretraining corpora on in-context learning by a large-scale language model (2022). ArXiv:2204.13509 [cs].

108. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

109. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *NeurIPS* (2023).

110. Xue, L. *et al.* xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872* (2024).

111. Koh, J. Y., Salakhutdinov, R. & Fried, D. Grounding language models to images for multimodal inputs and outputs (2023).

112. Alayrac, J.-B. *et al.* Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **35**, 23716–23736 (2022).

113. Vaswani, A. *et al.* Attention is all you need. In *Vaswani2017AttentionIA* (2017).

114. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In Erk, K. & Smith, N. A. (eds.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016).

115. Luo, S. *et al.* One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023* (OpenReview.net, 2023).

116. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).

117. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).

118. Vijayakumar, A. *et al.* Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

119. Wang, Y. *et al.* Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Goldberg, Y., Kozareva, Z. & Zhang, Y. (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5085–5109 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).

120. Karamcheti, S. *et al.* Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)* (2024).

121. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

122. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607 (PMLR, 2020).

123. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).

124. Li, S. *et al.* Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704* (2020).

125. Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16 (IEEE, 2020).

126. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

127. Dusart, P. *et al.* A tissue centric atlas of cell type transcriptome enrichment signatures. *bioRxiv* 2023–01 (2023).

128. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations* (2020).

129. O'brien, R. J. & Wong, P. C. Amyloid precursor protein processing and alzheimer's disease. *Annual review of neuroscience* **34**, 185–204 (2011).

130. Vendredy, L., Adriaenssens, E. & Timmerman, V. Small heat shock proteins in neurodegenerative diseases. *Cell Stress and Chaperones* **25**, 679–699 (2020).

131. Landrum, M. J. *et al.* Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980–D985 (2014).

132. Campanella, C. *et al.* Heat shock proteins in alzheimer's disease: role and targeting. *International journal of molecular sciences* **19**, 2603 (2018).

133. Chen, Z., Downing, S. & Tzanakakis, E. S. Four decades after the discovery of regenerating islet-derived (reg) proteins: current understanding and challenges. *Frontiers in cell and developmental biology* **7**, 235 (2019).

134. Meshref, M. *et al.* The role of rin3 gene in alzheimer's disease pathogenesis: a comprehensive review. *Molecular Neurobiology* **61**, 3528–3544 (2024).

135. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* **381**, eadg7492 (2023).

136. Bains, N. & Abdijadid, S. Major depressive disorder. In: StatPearls [Internet]. StatPearls Publishing (2024). Available from: https://www.ncbi.nlm.nih.gov/books/NBK559078/.

137. Varadi, M. *et al.* Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439–D444 (2022).