THEBIOCOLLECTION: Unified Pre-Training Scale LLM Corpus for Biology

Hyunjin Seo[1,*], Hyeon Hwang[1,*], Gyubok Lee[2], Jay Shin[1], Jimin Park[3], Taesoo Kim[4], Sanghoon Lee[5], Hongjoon Ahn[1,‡], Sungjun Han[1,‡], Sangwon Jung[1,‡]

[1] **Trillion Labs** [2] **KAIST** [3] **SK Biopharmaceuticals Co., Ltd.** [4] **Lunit Inc.** [5] **AIGEN Sciences Inc.**

[*] `First Author`

[‡] `Project Lead`

arXiv:2607.08803v1 [q-bio.QM] 9 Jul 2026

Abstract

The push toward large language models for biology (BioLM) has created a need for training corpora that can endow models with a genuine understanding of biology. However, existing biological resources, such as molecular databases, protein repositories, genomic annotations, single-cell atlases, and pathway databases, are scattered across heterogeneous formats and remain unorganized into a cohesive corpus for language model training. We present THEBIOCOLLECTION, a 52.6B-token pre-training-scale corpus that converts these disparate resources into a unified, training-ready form spanning small molecules, proteins, genomic sequences, cells, and pathways. Beyond consolidating existing data, THEBIOCOLLECTION enriches each record with tool-computed biological properties and introduces new instruction tasks for capabilities that current corpora barely cover. We pair the corpus with THEBIOCOLLECTION-EVAL, a matched suite probing recognition, generation, and prediction across molecular, protein, genomic, cellular, and cross-domain settings. Holding the base Gravity-16B-A3B architecture fixed, training on THEBIOCOLLECTION more than doubles its overall score on THEBIOCOLLECTION-EVAL with gains in every domain, while leaving general linguistic ability nearly intact.

🤗 **Training corpus:** THEBIOCOLLECTION

🤗 **Evaluation dataset:** THEBIOCOLLECTION-EVAL

🤗 **Model:** Gravity-bio-16B-A3B

1. Introduction

Large language models (LLMs) for biology (BioLM) (Jang et al., 2026; Li et al., 2026; Wang et al., 2025a,b; Xia et al., 2025) aim to extend LLMs beyond general text understanding toward the representation and reasoning of biological systems and generation of biological entities. Unlike task-specific predictors trained for a single modality or endpoint in biology (Cui et al., 2024; Lin et al., 2023; Zhou et al., 2024), BioLM seek to operate over diverse biological entities and processes, including small molecules, proteins, genomic sequences, cell states, and pathways. Understanding this broader spectrum is important as many biological questions involve not only recognizing isolated entities, but also understanding their properties, functions, interactions, and effects across different biological contexts (Mazein et al., 2024; Mohamed et al., 2021; Simon et al., 2024).

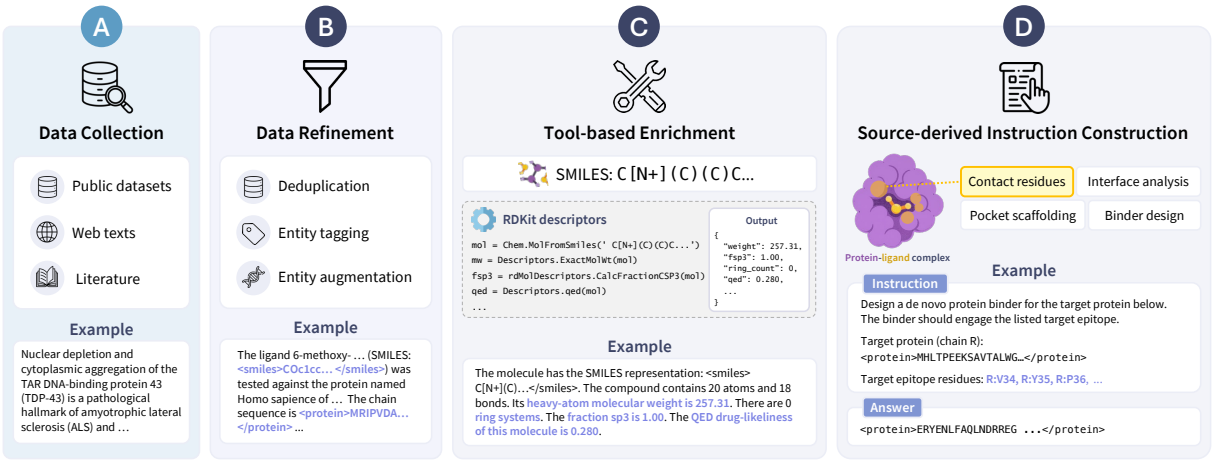

Figure 1 | The construction pipeline of THEBIOCOLLECTION. We first integrate scattered biological resources and refine them into structured texts. Subsequently, we enrich the biological information in free-text stream with computational tools. Finally, we construct new instruction datasets for domains that are underrepresented in the corpora.

Realizing this vision requires a large-scale training corpus for biology that exposes the model to the breadth and structure of biology. The field has already produced many useful resources, including molecular databases, protein repositories, genomic annotations, single-cell atlases, pathway databases, and computational biology tools. However, these resources are rarely organized as a large-scale, LLM training-friendly data corpus for BioLM; instead, they are scattered across heterogeneous formats such as tables, sequence files, graph edges, structured annotations, or isolated instruction schemas (Cai et al., 2026; Richard et al., 2024; Seo et al., 2026; Shen et al., 2024; The Tabula Sapiens Consortium, 2022; Yu et al., 2024). Nevertheless, no prior effort has consolidated these resources into a large-scale pre-training corpus that teaches BioLMs broad spectrum of biological knowledge spanning molecules, proteins, genomes, cells, and their cross-domain relationships.

To address this gap, we introduce THEBIOCOLLECTION, a 52.6B-token pre-training scale corpus for biology that turns heterogeneous biological resources into LLM training friendly data. THEBIOCOLLECTION is built through a construction pipeline that collects resources across biological domains, refines them through deduplication, entity tagging, and entity augmentation, enriches them with tool-computed biological properties, and renders them as instruction-form data with programmatically verifiable answers. This pipeline makes the resulting corpus both broad in biological coverage and directly usable for training BioLMs. To systematically assess biological capability after training, we also release THEBIOCOLLECTION-EVAL, a separate evaluation suite matched to the breadth of THEBIOCOLLECTION. The suite spans recognition, generation, and prediction across molecular, protein, genomic, cellular, and cross-domain settings.

To isolate the effect of the corpus, we hold the base architecture of Gravity-16B-A3B (Trillion Labs, 2026) fixed and train it on THEBIOCOLLECTION, comparing the result against the base model. Our evaluations using THEBIOCOLLECTION-EVAL reveal that the model trained with our corpus improves task performance on biology without specialized architectures or additional training tricks. Specifically, training on THEBIOCOLLECTION more than doubles the base model's overall score with consistent gains in every biological domain. Furthermore, in our ablation study, the model trained with our corpus outperforms the model further trained only

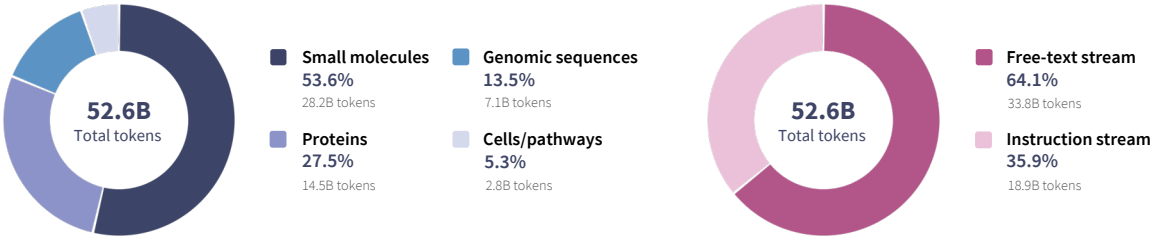

Figure 2 | Token statistics of THEBIOCOLLECTION. *Left*: the 52.6B-token corpus split across small molecules, proteins, genomic sequences, and cells and pathways. *Right*: the same corpus split into free-text and instruction streams. Token counts exclude scientific literature used during training.

with web-text on biological capability while preserving general linguistic ability with minimal degradation. These results suggest that biological capability and general linguistic ability are not mutually exclusive and can be learned jointly without interference.

To summarize, our contributions are as follows:

- **THEBIOCOLLECTION corpus.** We introduce THEBIOCOLLECTION, a 52.6B-token biological corpus that turns scattered public datasets into LLM-ready training data, spanning small molecules, proteins, genomic sequences, cells, and pathways, with tool-computed biological properties exposed directly in texts. We further construct new instruction datasets for tasks that existing corpora cover poorly, including protein binding and genomic feature localization.
- **THEBIOCOLLECTION-EVAL.** We release THEBIOCOLLECTION-EVAL, a matched evaluation suite that covers recognition, generation, and prediction across molecular, protein, genomic, cellular, and cross-domain settings, covering a broad spectrum of biological applications.
- **Controlled evidence of corpus effect.** Holding the base LLM architecture fixed, we show that training on THEBIOCOLLECTION improves broad biological task performance without additional training tricks, while maintaining general linguistic performance through the inclusion of only a limited amount of web text in the training corpus.

2. Dataset Construction

THEBIOCOLLECTION is a 52.6B-token corpus that places small molecules, proteins, genomic sequences, cells, pathways, and cross-domain relationships under a shared language interface. Its composition is summarized in Figure 2. We build the corpus by collecting public, commercially usable resources across all five domains (Section 2.1) and transforming them into LLM training friendly data through three complementary steps: refining database records into self-contained descriptions (Section 2.2), enriching them with tool-computed biological features (Section 2.3), and constructing instruction datasets from both existing resources and newly generated datasets that no existing instruction-tuning corpus covers (Section 2.4). To measure whether these data translates into biological understanding, we pair the corpus with THEBIOCOLLECTION-EVAL, a matched evaluation suite spanning the same domains together with cross-domain reasoning (Section 2.5).

Table 1 | External resources used to construct free-text stream in THEBIOCOLLECTION, grouped by biological domain.

Domain	Source datasets
Small molecules	PubChem (Kim et al., 2025); DrugBank (Knox et al., 2024); BindingDB (Liu et al., 2025); ChEMBL (Zdrazil et al., 2024); USPTO-50K (Schneider et al., 2016).
Proteins	UniProt Knowledgebase (The UniProt Consortium, 2025); AlphaFoldDB (Bertoni et al., 2026); Protein Data Bank (PDB) (Berman et al., 2000).
Genomic sequences	GENCODE (Frankish et al., 2021); ENCODE (Moore et al., 2020); RNAcentral (The RNAcentral Consortium, 2021); Rfam (Kalvari et al., 2021).
Cells/Pathways	L1000 (Subramanian et al., 2017); Human Protein Atlas (HPA) single-cell type data (Uhlén et al., 2015); GeneRIF (Mitchell et al., 2003); NCBI Gene (Maglott et al., 2005); Gene Ontology / GOA human (The Gene Ontology Consortium, 2023); Cell Ontology (Diehl et al., 2016); CELLxGENE papers (CZI Cell Science Program et al., 2025); JUMP Cell Painting (Chandrasekaran et al., 2024); HuBMAP (Börner et al., 2025).
Cross-domain KGs	Hetionet (Himmelstein et al., 2017); DRKG (Drug Repurposing Knowledge Graph Team, 2020); DGIdb (Freshour et al., 2021); STRING (Szklarczyk et al., 2023); OmniPath (Türei et al., 2016); Reactome (Milacic et al., 2024).
Broad scientific literature	PubMed; bioRxiv; medRxiv.

2.1. Data collection

We build THEBIOCOLLECTION only from public resources that permit commercial use, and we keep each entity's source identifiers throughout so every record stays traceable to its origin. The collected resources span five domains targeted by the corpus, including small molecules, proteins, genomic sequences, cells/pathways, and cross-domain relationships. As summarized in Table 1, these sources provide complementary biological signals, ranging from molecular characteristics to protein sequences and functions, genomic and transcriptomic annotations, cellular identities and perturbation responses, and knowledge-graph-linked relationships across biological entities. Finally, to retain general scientific and language ability that structured data alone does not provide, we include broad scientific literature from PubMed[1], bioRxiv[2], and medRxiv[3]. These scientific sources are included for training but are excluded from the token counts reported for THEBIOCOLLECTION. The remaining subsections describe how we transform these resources into LLM training friendly data and enrich them with additional biological signals beyond public resources.

2.2. Dataset refinement

Curated biological databases contain rich biological knowledge, but often in formats that language models cannot consume directly (Kim et al., 2025; Liu et al., 2025; Schneider et al., 2016; Subramanian et al., 2017; Zdrazil et al., 2024). We therefore transform each database record into a self-contained natural-language description that consolidates entity's key attributes into a single textual record. Depending on the modality, these descriptions may include physicochemical properties, perturbation profiles, structural information, or functional annotations. We then render each value with a plain language statement of what it indicates. Symbolic identifiers

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.biorxiv.org/
[3] https://www.medrxiv.org/

including SMILES strings and protein, DNA, or RNA sequences are placed inline and wrapped in dedicated tagging tokens (*e.g.*, `<smiles>...</smiles>`) that keep them distinct from the surrounding text, following Wang et al. (2025b); Xia et al. (2025). The resulting records keep an entity's sequence, properties, and interpretation within a single context, turning structured database entries into language the model can learn from.

Beyond single-entity descriptions, we also link entities across domains by gathering evidence about the same entity from several databases into one record. A knowledge graph serves as the skeleton and the source databases as the payload: we use cross-domain knowledge graphs (Hetionet (Himmelstein et al., 2017), DRKG (Drug Repurposing Knowledge Graph Team, 2020), STRING (Szklarczyk et al., 2023), OmniPath (Türei et al., 2016)) to decide which entities belong together through typed relations such as compound–target binding, gene–pathway participation, protein–protein interaction, and transcription-factor–target regulation, then fill in each linked entity with the records already held in the underlying databases. As a result, a single record reads as a mechanistic chain rather than an isolated fact: a protein anchor, for instance, carries its function and pathway membership, the compounds that bind it with their measured affinities, and the transcriptional response those perturbations induce in a given cell line, so that molecule, protein, pathway, and cell-domain evidence appear together in one context window.

2.3. Tool-computed feature narratives

We verbalize biological signals that live in databases and computational tools rather than in text. We run standard computational-biology tools over source records and render their outputs as natural-language narratives with explicit provenance, exposing structured, tool-computable signals that public free text rarely contains. We apply the same recipe across modalities; the full feature sets and extraction pipelines are given in Appendix B.

Small molecules. We enrich the refined PubChem molecular narratives produced from Section 2.2 with RDKit-computed descriptors (Landrum et al., 2026). After parsing and canonicalization, we derive deterministic features such as ring-system composition, scaffold and complexity measures, stereochemistry, functional-group and structural-alert matches, and fingerprint summaries, and incorporate them into the corresponding narratives.

Proteins. We construct protein structure/function narratives by linking UniProt sequences to AlphaFold database and Protein Data Bank (PDB) structures. These records combine deterministic structure-derived features such as physicochemical properties, secondary structure, fold topology, and solvent exposure using Biopython (Cock et al., 2009) and DSSP (Kabsch and Sander, 1983), with UniProt Knowledgebase-derived annotations such as domains, InterPro/Pfam families, transmembrane regions, GO terms, and subcellular localization (The UniProt Consortium, 2025).

Genomic sequences. We represent regulatory DNA intervals and RNA transcripts as structured biological entities. For DNA, we use Python-based processing pipelines built on indexed FASTA access (pyfaidx (Shirley et al., 2015)), Biopython parsing, and deterministic GTF/BED interval handling to associate genomic sequences from GENCODE (Frankish et al., 2021) and ENCODE (Moore et al., 2020) with coordinate and assembly metadata, regulatory evidence, and sequence-derived features such as GC content, CpG statistics, entropy, homopolymer length, and genomic feature overlap. For RNA, we build transcript-level narratives from GENCODE and

Table 2 | Imported instruction datasets curated for THEBIOCOLLECTION.

Domain	Source datasets	Task coverage
Small molecules	Mol-Instructions (Fang et al., 2023); SMolInstruct (Yu et al., 2024); MolLangBench (Cai et al., 2026); ChEBI-20-MM (Liu et al., 2024); ChemData700K (Zhang et al., 2024); Language-Plus-Molecules (LPM-24) (Edwards et al., 2024); TxGemma (Wang et al., 2025a)	Text-guided molecule design and generation, molecule captioning, SMILES/name/text conversion, atom- and functional-group recognition, molecular property and toxicity prediction, forward synthesis, retrosynthesis, and reaction prediction.
Proteins	ProteinLMBench (Shen et al., 2024); BioReason-Pro (Fallahpour et al., 2026); TxGemma (Wang et al., 2025a)	Protein sequence-to-function prediction, protein function description, and drug–target or affinity-style prediction.
Cells/pathways	Cell2Sentence (Levine et al., 2024); ENCODE-C2S (Moore et al., 2020); Tabula Sapiens (The Tabula Sapiens Consortium, 2022); PerturBench Norman and Replogle K562 (Norman et al., 2019; Replogle et al., 2022; Wu et al., 2026); TxGemma (Wang et al., 2025a)	Cell-type classification, tissue-aware cell-state interpretation, perturbation-response and differential-expression signature prediction, drug synergy prediction, and omics response modeling.

RNA resources including RNAcentral (The RNAcentral Consortium, 2021) and Rfam (Kalvari et al., 2021). These narratives incorporate features derived from Biopython-based FASTA and GenBank parsing (Sayers et al., 2025), codon and ORF analyses, and sequence-motif profiling.

Cellular profiles. We convert cellular profiles into language that captures four signals: expression, perturbation response, spatial context, and morphology. Single-cell records describe cell states from marker genes and tissue context (Diehl et al., 2016; Uhlén et al., 2015). Perturbation records link CRISPRa/CRISPRi interventions to the differential-expression signatures they induce (Norman et al., 2019; Replogle et al., 2022; Wu et al., 2026). Spatial records from HuBMAP (Börner et al., 2025) place each measured spot in context, reporting its tissue, coordinates, local neighborhood, and spot-local transcript features. We build them from AnnData/H5AD inputs (Virshup et al., 2024) and recover each neighborhood with SciPy KD-tree radius queries (Virtanen et al., 2020). JUMP Cell Painting records (Chandrasekaran et al., 2024) capture how each perturbation shifts cell morphology. We derive them from CellProfiler profiles (Carpenter et al., 2006), normalize features against controls with pandas/NumPy (Harris et al., 2020; McKinney, 2010), and retrieve similar morphologies with scikit-learn nearest-neighbor search (Pedregosa et al., 2011).

2.4. Instruction-tuning datasets

The instruction corpus draws on two complementary sources. The first is imported instruction datasets: the public datasets refined via Section 2.2, which already expose biological entities through natural-language instructions. The second is source-derived instruction tasks that we construct ourselves. Across the full instruction corpus, approximately 30% of examples include few-shot support and the rest are zero-shot, following the FLAN collection (Longpre et al., 2023).

For the imported datasets, our guiding principle is careful curation. We select commercially usable datasets that broaden the task coverage of THEBIOCOLLECTION and add them to the corpus with associated tagging tokens. Table 2 summarizes these datasets and the tasks they cover. The source-derived tasks, by contrast, are generated directly from non-text biological re-

sources, and we design them so that answers stay structured, operational, and programmatically checkable. Crucially, these tasks cover capabilities that existing public biological instruction datasets rarely address, broadening the task coverage of THEBIOCOLLECTION. We detail two such task families below, and Appendix C gives full construction details.

Protein binding. We derive binding tasks from protein–ligand, protein–protein, and protein–peptide complexes by computing target–binder interfaces from heavy-atom contacts. The resulting instructions ask the model to recover masked binding-site residues, summarize contact patterns, scaffold binding pockets, or generate binder sequences conditioned on target epitope residues. After structural-quality filtering, deduplication, and evaluation-overlap removal, this yields 12M binding instruction records before context-length filtering.

DNA/RNA feature localization. We cast regulatory DNA and transcript RNA annotations as exact span-recovery tasks. Given a sequence window, the model returns the feature label, coordinates, and subsequence, making every answer programmatically checkable. DNA tasks cover regulatory elements, open-chromatin peaks, splice-related sites, and transcription-factor motifs; RNA tasks cover coding and untranslated regions, exon junctions, family hits, miRNAs, and tRNA anticodons. We discard ambiguous or low-quality windows and any record that fails deterministic span self-checks.

2.5. Evaluation suite: THEBIOCOLLECTION-EVAL

THEBIOCOLLECTION-EVAL measures whether training on THEBIOCOLLECTION improves biological reasoning capability across domains of organization. We build it by drawing the subtasks from many scattered existing benchmarks and combining them with tasks we construct ourselves, then standardizing everything into one cross-domain suite. The construction of evaluation datasets and the task organization are illustrated in Figure 3. This consolidates evaluations that were previously fragmented across separate resources and formats, and it mirrors the breadth of the corpus, spanning small molecules, proteins, genomics, and cells together with cross-domain reasoning. We sample 100 and 50 records per subtask in single-domain and cross-domain, respectively, yielding 1,650 examples across 18 tasks, and for any task whose format also appears in training, we separate train and test entities to prevent overlap. The statistics of the evaluation dataset are shown in Table 11. We show an example question for each subtask in Table 12, and provide per-task evaluation metrics in Appendix E. We group the tasks by biological domain as follows:

- **Small molecules:** The small-molecule tasks span recognition, generation, and reaction prediction. They cover molecular recognition and reconstruction from structural descriptions (Cai et al., 2026), description-guided molecule design (Fang et al., 2023), and forward synthesis (Yu et al., 2024).
- **Proteins:** The protein tasks evaluate function prediction and design. Function prediction asks the model to predict the natural-language names of a protein's InterPro IDs from its sequence. Design tasks consist of functional description-conditioned generation (Fang et al., 2023) and binder generation against small-molecule and protein targets.
- **Genomic sequences:** The genomics tasks evaluate DNA and RNA feature annotation as structured span recovery. Specifically, the DNA tasks cover candidate cis-regulatory element (cCRE), open-chromatin, and splice-site localization, whereas the RNA tasks require identifying Rfam families and tRNA anticodons. Each task asks the model to identify the

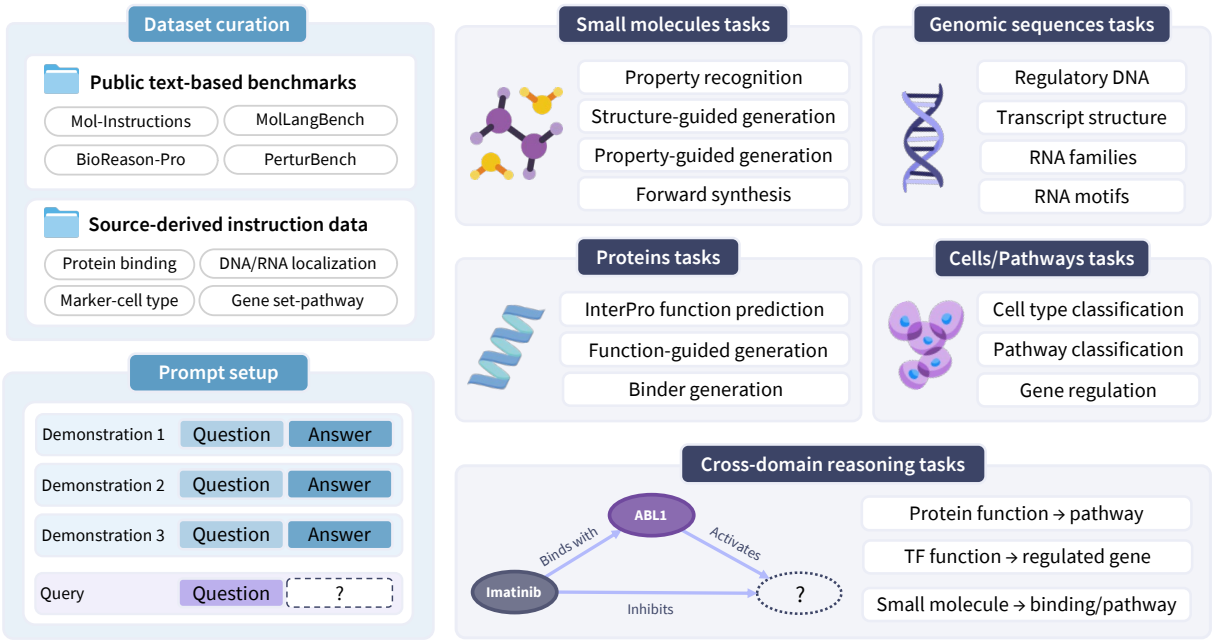

Figure 3 | Illustration of THEBIOCOLLECTION-EVAL. *Left*: Evaluation dataset construction. We combine established public benchmarks with the instruction datasets that we newly construct from primary biological sources, and convert single-domain tasks into a 3-shot format; cross-domain tasks are evaluated zero-shot. *Right*: Evaluation task organization. Within each domain, we adopt tasks that are commonly studied in the corresponding domain-specific literature. To probe whether models transfer knowledge across domains, we further introduce cross-domain reasoning tasks constructed from biological knowledge graphs.

biological feature type and return its label, exact coordinates, and corresponding subsequence, mirroring the interface used during training.

- **Cells and pathways:** The cell and pathway tasks evaluate marker-based cell-type identity (Ianevski et al., 2022), Hallmark pathway recognition (Chen et al., 2013; Liberzon et al., 2015), and K562 CRISPRi perturbation-response direction (Replogle et al., 2022). Each is posed as multiple-choice or as up/down classification over candidate genes.
- **Cross-domain reasoning:** The cross-domain tasks test whether the model can chain evidence across domains rather than within a single one. Each is a two-hop multiple-choice question linking an entity's function or identity to a knowledge-graph relation: a protein's function to a pathway it participates in, a transcription factor's function to a gene it regulates, and a small molecule to the target and pathway it acts through (Drug Repurposing Knowledge Graph Team, 2020; Milacic et al., 2024; Türei et al., 2016).

Filtering and decontamination. We clean the corpus in three steps:

(1) **Validity checks.** Every record passes source-specific checks: malformed molecules and protein sequences are dropped, binding records must satisfy contact and interface constraints, and a DNA/RNA span is kept only when its target feature is unambiguously recoverable.

(2) **Deduplication.** Reading the corpus one record at a time, we assign each a stable key (normalized text hash, canonical molecule identifier, sequence hash, genomic coordinate, document ID, or instruction–answer hash) and drop any record whose key has already appeared, along with residual leakage and artifact rows.

Table 3 | Performance across diverse biological domains: **Base** vs. **Ours**. **Base** is the base model checkpoint before training; **Ours** adds the THEBIOCOLLECTION corpus on top of general scientific and web text. Scores are averaged first across metrics within a task, then across subtasks within a domain; the overall score averages all subtasks. Full subtask-by-metric results are provided in Appendix D. Bold indicates the best model in each row.

Domain	Task	Base	Ours	Δ
Small molecules	Molecule reconstruction/design	0.200	**0.522**	+0.322
	Forward synthesis	0.213	**0.619**	+0.406
	Molecular property recognition	0.280	**0.390**	+0.110
	Domain average	0.223	**0.513**	+0.290
Proteins	Text-conditioned functional protein design	0.243	**0.522**	+0.279
	Binder design	0.234	**0.645**	+0.411
	Protein function prediction	0.000	**0.055**	+0.055
	Domain average	0.159	**0.407**	+0.248
Genomic sequences	DNA regulatory/splice span localization	0.134	**0.516**	+0.382
	RNA family/anticodon span localization	0.238	**0.396**	+0.158
	Domain average	0.175	**0.468**	+0.293
Cells/pathways	Cell type recognition	0.470	**0.580**	+0.110
	Hallmark program recognition	0.520	**0.750**	+0.230
	Perturbation response prediction	0.015	**0.498**	+0.483
	Domain average	0.335	**0.609**	+0.274
Overall		0.223	**0.499**	+0.276

(3) **Decontamination.** We hold THEBIOCOLLECTION-EVAL out of training with task-specific keys and exact string matching, e.g., genomics by sequence hash, accession, and coordinate, and protein binder-generation targets by excluding any exact, subsequence, or shared 15-mer overlap.

3. Experiments

3.1. Training details

We train the base Gravity-16B-A3B (Trillion Labs, 2026) on THEBIOCOLLECTION. We choose Gravity-16B-A3B for two reasons: it exposes a pre-anneal checkpoint, and its pretraining is known to contain no biological corpus, so any biological capability we observe comes from our data rather than the base. Beginning before annealing matters as most high-impact learning happens during annealing, so starting before it lets us measure the annealing effect of THEBIOCOLLECTION cleanly without the effect of another corpus entangling with ours. We mix our corpus with general scientific and web text from OLMoCR science PDFs, DCLM-CC, EAI-DCLM, PubMed, bioRxiv, and medRxiv as replay, to counteract forgetting and preserve broad language ability. We use a sequence length of 8,192 and a global batch of 8.4M tokens ($1{,}024 \times 8{,}192$), an annealed learning rate of 4×10^{-4}, and weight decay of 0.01, trained on 16 NVIDIA B200 GPUs.

To attribute the measured gains to the corpus rather than to the training process itself, we compare against the base model under the same architecture. **Base** is the base Gravity-16B-A3B checkpoint without any training, and **Ours** is the model trained on THEBIOCOLLECTION, starting from that same checkpoint.

3.2. Evaluation protocol

For all evaluation tasks except cross-domain reasoning, we adopt a fixed three-shot prompting setting, prepending three task-specific demonstration examples to each query; the cross-domain tasks are evaluated zero-shot. We then aggregate by averaging within a subtask, across subtasks within a domain, and finally across all subtasks for the overall score. To keep these aggregated scores directionally consistent, we average only higher-is-better metrics and exclude the one lower-is-better metric, SMILES Levenshtein distance, which we still report per subtask in Appendix D.

3.3. Main results

We report the performance of the model trained with our corpus in Table 3. As shown, training on THEBIOCOLLECTION improves performance at every biological domain. The model raises the overall score from 0.223 to 0.449, more than doubling the baseline, with consistent domain gains across genomic sequence (0.175 → 0.468), small molecule (0.223 → 0.513), protein (0.159 → 0.407), cell/pathway (0.335 → 0.609).

The gains concentrate on tasks whose supervision is structured and tool-derived rather than literature-style text. This is especially pronounced in DNA regulatory/splice-site localization (0.134 → 0.516) and protein binder design (0.234 → 0.645), where the supervision is difficult to retrieve from public free-text records and is instead supplied by matched instruction tasks. Beyond these, the corpus also improves recognition tasks such as molecular property recognition (0.280 → 0.390) and cell-type and hallmark-program recognition (0.470 → 0.580, 0.520 → 0.750), for which the tool-driven narratives (Section 2.2) supply the signal implicitly: RDKit descriptors render each molecule's computed structural properties into text, and marker-gene narratives describe each cell by its expression state. Although these narratives are written as descriptions rather than task-formatted supervision, they teach the property and identity cues that recognition requires.

3.4. Ablation study

Contribution of the sole biological corpus. To isolate the contribution of the THEBIOCOLLECTION corpus from the effect of annealing on general text alone, we introduce **Text-annealing only**, the model obtained by training the base checkpoint on just the general scientific and web text described above, with no THEBIOCOLLECTION data. Since text-annealing only and Ours share the same base checkpoint, the same annealing process, and the same general text, their only difference is the presence of the THEBIOCOLLECTION corpus, so any gap between them isolates the contribution of our biological data on downstream tasks (Table 5).

Table 4 | Scientific free text used in annealing.

Source	Docs	Tokens
PubMed	6.43M	48.8B
bioRxiv	0.47M	4.6B
medRxiv	0.11M	0.58B
Total	**7.01M**	**54.0B**

General text alone already lifts overall performance from 0.223 to 0.385. Much of this comes from scientific sources: the annealing mixture includes about 54B tokens of PubMed, bioRxiv, and medRxiv text, comparable in scale to THEBIOCOLLECTION itself (Table 4), which already describes molecules, proteins, and pathways in prose. Annealing on it, therefore, transfers real biological signal, leaving the text-only model far above the base and even competitive where free text states facts explicitly, such as cross-domain reasoning (0.493). Training on THEBIOCOLLECTION raises overall performance further to 0.499, with the clearest gains where free

Table 5 | Performance comparison: **Text-annealing only (TA only)** vs. **Ours**. **TA only** is trained on general scientific and web text alone, with no THEBIOCOLLECTION data; **Ours** adds the THEBIOCOLLECTION corpus on top of that same general text. Since the two share the same base checkpoint, annealing process, and general text, the gap between them isolates the contribution of the THEBIOCOLLECTION corpus. For biological domains, scores are averaged first across metrics within a task, then across subtasks within a domain; the overall score averages all subtasks. General-language benchmarks are reported on their native scales and excluded from the biological overall score. Bold indicates the best model in each row.

Domain	Task	TA only	Ours	Δ
Small molecules	Molecule reconstruction/design	0.432	**0.522**	+0.090
	Forward synthesis	0.477	**0.619**	+0.142
	Molecular property recognition	0.300	**0.390**	+0.090
	Domain average	0.410	**0.513**	+0.103
Proteins	Text-conditioned functional protein design	**0.586**	0.522	-0.064
	Binder design	0.297	**0.645**	+0.348
	Protein function prediction	0.000	**0.055**	+0.055
	Domain average	0.294	**0.407**	+0.113
Genomic sequences	DNA regulatory/splice span localization	0.208	**0.516**	+0.308
	RNA family/anticodon span localization	0.253	**0.396**	+0.143
	Domain average	0.226	**0.468**	+0.242
Cells/pathways	Cell type recognition	0.550	**0.580**	+0.030
	Hallmark program recognition	0.700	**0.750**	+0.050
	Perturbation response prediction	**0.624**	0.498	-0.126
	Domain average	**0.625**	0.609	-0.016
Biological overall		0.385	**0.499**	+0.114
General language	MMLU (5-shot)	**61.8**	59.7	-2.1
	ARC-c (10-shot)	**58.0**	57.0	-1.0
	HellaSwag (10-shot)	**77.6**	76.5	-1.1
	Winogrande (0-shot)	**70.6**	70.5	-0.1
	PIQA (10-shot)	**81.7**	81.1	-0.6
	Average	**69.9**	69.0	-0.9

text falls short, on structured and tool-derived tasks: genomic localization improves by 0.242 and binder design by 0.348, with forward synthesis and molecule generation rising comparably. Cross-domain reasoning also improves, reaching 0.507 over the text-only model's 0.493. On text-conditioned functional protein design, the THEBIOCOLLECTION model scores slightly below the text-only model (0.522 vs 0.586), but the text-only model produces degenerate sequences (nondegeneracy 0.160 for binder generation, 0.540 for functional design), whereas THEBIOCOLLECTION generates diverse ones (1.000 and 0.930). Adding THEBIOCOLLECTION also lowers perturbation-response prediction, which we attribute to the thin cell-domain coverage of the current corpus; enriching this is left to future work.

Preservation of general-language ability. The gains above come from shifting the training mixture heavily toward biological data, which raises the concern that general-language ability may degrade in return. We check this by comparing both models on five standard language benchmarks (Bisk et al., 2020; Clark et al., 2018; Hendrycks et al., 2021; Sakaguchi et al., 2021; Zellers et al., 2019) in Table 5. Despite the biological corpus dominating the mixture, the model trained with THEBIOCOLLECTION stays within 0.9 points of the text-annealing only model

Table 6 | Cross-domain two-hop reasoning: **Base** vs. **Ours**. **Base** is the base model checkpoint before training; **Ours** adds the THEBIOCOLLECTION corpus on top of general scientific and web text. Each subtask is a four-way multiple-choice question linking an entity to a knowledge-graph relation. Bold indicates the best model in each row.

Subtask	Base	Ours	Δ
Protein function → pathway	0.420	**0.680**	+0.260
TF function → regulated target gene	0.260	**0.480**	+0.220
Small molecule → binding target, pathway	0.260	**0.360**	+0.100
Average	0.313	**0.507**	+0.194

on average (69.0 vs 69.9), and the largest single-benchmark drop is only 2.1 points, on MMLU. The broad biological capability that THEBIOCOLLECTION adds therefore comes with little forgetting, leaving general-language ability nearly intact.

3.5. Cross-domain understanding

The suite in Section 3.3 scores each biological domain in isolation. To test whether ours yields genuinely cross-domain reasoning, we evaluate three two-hop relational tasks, each posed as four-way multiple choice. Every question begins from one entity's function or identity and requires a second hop into a knowledge-graph relationship: a protein's function to a pathway it participates in (SwissProt → Reactome/Hetionet), a transcription factor's function to a gene it directly regulates (SwissProt → OmniPath), and a small molecule to the (target protein, pathway) pair it acts through (DRKG → Reactome/Hetionet). Entity names and gene symbols are withheld, so the first hop cannot be solved by surface lookup.

We report the zero-shot accuracy in Table 6. Across all three relation types, the model trained on our corpus consistently outperforms the base model, raising the average accuracy from 0.313 to 0.507, with the largest improvements on protein function-to-pathway. This suggests that THEBIOCOLLECTION strengthens cross-domain knowledge integration, not just single-domain task performance.

4. Related Work

4.1. Biological foundation language models

A growing body of work places biological entities under a shared language interface, so one model can read and generate across molecules, proteins, and nucleic acids (Chaves et al., 2024; Pei et al., 2023, 2024; Wang et al., 2025a). The most ambitious treat many entity types as one "language": NatureLM (Xia et al., 2025) models small molecules, proteins, DNA, RNA, and materials as sequences in a single model and SciReasoner (Wang et al., 2025b) aligns language with heterogeneous scientific representations and adds reasoning-oriented post-training. LOGOS (Li et al., 2026) extends this idea by representing small molecules, proteins, and material entities and their 3D spatial interactions through a shared text-based scientific grammar. These establish the language-interface paradigm THEBIOCOLLECTION builds on, but each centers on a model rather than a corpus, and stays at the molecular sequence levels.

4.2. Modality-specific foundation models

Outside the language-model paradigm, modality-specific models pair a pretrained model with a large domain corpus: ESM2 (Lin et al., 2023) and ProGen2 (Nijkamp et al., 2023) for protein sequences, the Nucleotide Transformer (Dalla-Torre et al., 2025) and DNABERT-2 (Zhou et al., 2024) for genomes, Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2024) for single cells, and Evo (Brixi et al., 2026; Nguyen et al., 2024) for cross-modal genomic modeling. Each is powerful within its modality but does not span levels or expose entities through language.

4.3. Biological corpora and benchmarks

A parallel line releases the data and evaluations such models need. Instruction corpora such as Mol-Instructions (Fang et al., 2023) cover molecule, protein, and biomolecular-text tasks, while benchmark suites standardize evaluation within a domain: TAPE and PEER (Rao et al., 2019; Xu et al., 2022) and ProteinGym (Notin et al., 2023) for proteins, GUE (Zhou et al., 2024) and BEND (Marin et al., 2023) for genomics, and the Therapeutics Data Commons (Huang et al., 2021) for drug discovery. These remain largely single-domain and tied to one or a few modalities. THEBIOCOLLECTION instead pairs a corpus with a matched evaluation suite across molecular, protein, genomic, cellular, and cross-domain settings, and goes beyond aggregation by enriching entries with tool-computed properties verbalized into language and by constructing new instruction data for practical tasks that existing resources underserve.

5. Conclusion

We introduced THEBIOCOLLECTION, an open pre-training scale corpus for biology on training BioLM, spanning small molecules, proteins, genomic sequences, cells, and pathways. We convert diverse biological resources into a shared language interface and polish them into structured textual records. We further enrich these records with biological features computed by computational tools, and build instruction tasks for capabilities that conventional corpora underrepresent, such as protein binding, DNA/RNA feature localization, and cross-domain reasoning. In addition, we present THEBIOCOLLECTION-EVAL to assess broad biological capabilities across domain-specific and cross-domain reasoning tasks. Holding the Gravity-16B-A3B architecture fixed, training a base model on THEBIOCOLLECTION substantially improves overall biological task performance across all domains. These results highlight that carefully constructed data can be the main driver of practical biological capability in BioLM.

Acknowledgments

This work was partially supported by the Ministry of Science and ICT (MSIT), Republic of Korea, through the National IT Industry Promotion Agency (NIPA), as part of the Domain-Specific Foundation Model Project (Grant No. PJT-26-100004). We thank Mingu Kang, Cedric Caruzzo, Wonseok Lee, Steve Immanuel, Donggeun Yoo, Gihyeon Lee, Jin Woo Oh, Aisha Urooj, Laurent Dillard and Aaron Valero from Lunit, Jeongwook Lee and Sunggi An from Seoul National University, Saebom Leem and Hyunkyu Jung from KAIST for their helpful support and contributions throughout the project.

References

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. Nucleic Acids Research, 28(1):235–242, 2000.

D. Bertoni, M. Tsenkov, P. Magana, S. Nair, I. Pidruchna, M. Querino Lima Afonso, A. Midlik, U. Paramval, D. Lawal, A. Tanweer, M. Last, R. Patel, A. Laydon, D. Lasecki, N. Dietrich, H. Tomlinson, A. Žídek, T. Green, O. Kovalevskiy, A. Lau, S. Kandathil, N. Bordin, I. Sillitoe, M. Mirdita, D. Jones, C. Orengo, M. Steinegger, J. R. Fleming, and S. Velankar. Alphafold protein structure database 2025: a redesigned interface and updated structural coverage. Nucleic Acids Research, 54(D1):D358–D362, 2026.

Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 7432–7439, 2020.

K. Börner, P. D. Blood, J. C. Silverstein, et al. Human biomolecular atlas program (hubmap): 3d human reference atlas construction and usage. Nature Methods, 22(4):845–860, 2025.

G. Brixi, M. G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, D. Chang, A. Fanton, G. A. Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. Nature, 652(8112):1349–1361, 2026.

A. Bushuiev, R. Bushuiev, P. Kouba, A. Filkin, M. Gabrielova, M. Gabriel, J. Sedlar, T. Pluskal, J. Damborsky, S. Mazurenko, and J. Sivic. Learning to design protein-protein interactions with enhanced generalization. In International Conference on Learning Representations, 2024.

F. Cai, J. Bai, T. Tang, G. He, J. Luo, T. Zhu, S. Pilla, G. Li, L. Liu, and F. Luo. Mollangbench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation. In International Conference on Learning Representations, 2026.

S. Candido, T. Hayes, A. Derry, R. Rao, Z. Lin, R. Verkuil, B. Z. Wu, J. S. Lee, E. S. Bruguera, J. A. Keval, et al. Language modeling materializes a world model of protein biology. bioRxiv, pages 2026–06, 2026.

A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology, 7: R100, 2006. doi: 10.1186/gb-2006-7-10-r100.

P. P. Chan and T. M. Lowe. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Research, 44(D1):D184–D189, 2016.

S. N. Chandrasekaran, B. A. Cimini, A. Goodale, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nature Methods, 21:1114–1121, 2024.

J. M. Z. Chaves, E. Wang, T. Tu, E. D. Vaishnav, B. Lee, S. S. Mahdavi, C. Semturs, D. Fleet, V. Natarajan, and S. Azizi. Tx-llm: A large language model for therapeutics. arXiv preprint arXiv:2406.06316, 2024.

E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. BMC Bioinformatics, 14:128, 2013.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11):1422–1423, 2009.

H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature methods, 21(8): 1470–1480, 2024.

CZI Cell Science Program, S. Abdulla, B. Aevermann, et al. CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Research, 53(D1):D886–D900, 2025.

H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. De Almeida, H. Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22 (2):287–297, 2025.

A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. Van Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake, and C. J. Mungall. The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. Journal of Biomedical Semantics, 7:44, 2016.

Drug Repurposing Knowledge Graph Team. Drug repurposing knowledge graph (drkg). https://github.com/gnn4dr/DRKG, 2020. Accessed: 2026-06-12.

J. Durairaj, Y. Adeshina, Z. Cao, X. Zhang, V. Oleinikovas, T. Duignan, Z. McClure, X. Robin, G. Studer, D. Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. BioRxiv, 2024.

C. Edwards, Q. Wang, L. Zhao, and H. Ji. L+M-24: Building a dataset for Language+Molecules @ ACL 2024. In Proceedings of the 1st Workshop on Language + Molecules, pages 1–9, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.langmol-1.1.

A. Fallahpour, A. Seyed-Ahmadi, P. Idehpour, O. Ibrahim, P. Gupta, J. Naimer, K. Zhu, A. Shah, S. Ma, A. Adduri, et al. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. bioRxiv, pages 2026–03, 2026.

Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. arXiv preprint arXiv:2306.08018, 2023.

A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, et al. Gencode 2021. Nucleic Acids Research, 49(D1):D916–D923, 2021.

S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. Song, M. Griffith, O. L. Griffith, and A. H. Wagner. Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Research, 49(D1):D1144–D1151, 2021.

L. Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research, 29(8):1363–1375, 2019.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, et al. Array programming with NumPy. Nature, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In International Conference on Learning Representations, 2021.

D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife, 6:e26726, 2017.

K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv preprint arXiv:2102.09548, 2021.

A. Ianevski, A. K. Giri, and T. Aittokallio. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nature Communications, 13(1):1246, 2022.

Y. Jang, L. Zhu, J. Fawkes, A. K. Denton, D. Beaini, and E. Noutahi. Towards autonomous mechanistic reasoning in virtual cells. arXiv preprint arXiv:2604.11661, 2026.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12):2577–2637, 1983.

I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microrna families. Nucleic Acids Research, 49(D1): D192–D200, 2021.

S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. Pubchem 2025 update. Nucleic Acids Research, 53 (D1):D1516–D1525, 2025.

C. Knox, M. Wilson, C. M. Klinger, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. Nucleic Acids Research, 52(D1):D1265–D1275, 2024.

D. Kovtun, M. Akdel, A. Goncearenco, G. Zhou, G. Holt, D. Baugher, D. Lin, Y. Adeshina, T. Castiglione, X. Wang, et al. Pinder: The protein interaction dataset and evaluation resource. bioRxiv, 2024.

A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones. miRBase: from microrna sequences to function. Nucleic Acids Research, 47(D1):D155–D162, 2019.

G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, P. Gedeck, et al. rdkit/rdkit: 2025_09_5 (q3 2025) release, 2026. URL https://doi.org/10.5281/zenodo.18428170. Version Release_2025_09_5.

D. Levine, S. A. Rizvi, S. Lévy, N. Pallikkavaliyaveetil, D. Zhang, X. Chen, S. Ghadermarzi, R. Wu, Z. Zheng, I. Vrkic, A. Zhong, D. Raskin, I. Han, A. H. de Oliveira Fonseca, J. O. Caro, A. Karbasi, R. M. Dhodapkar, and D. van Dijk. Cell2sentence: Teaching large language models the language of biology. In International Conference on Machine Learning, 2024.

M. Li, Y. Liu, J. Ye, B. Su, J.-R. Wen, and Z. Wang. Speaking the language of science: Toward a general-purpose generative foundation model for the natural sciences. arXiv preprint arXiv:2606.16905, 2026.

A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database hallmark gene set collection. Cell Systems, 1(6):417–425, 2015.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

P. Liu, J. Tao, and Z. Ren. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. arXiv preprint arXiv:2402.04119, 2024.

T. Liu, L. Hwang, S. K. Burley, C. I. Nitsche, C. Southan, W. P. Walters, and M. K. Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. Nucleic Acids Research, 53(D1):D1633–D1644, 2025.

S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In International Conference on Machine Learning, 2023.

D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. Nucleic acids research, 33(suppl_1):D54–D58, 2005.

F. I. Marin, F. Teufel, M. Horlacher, D. Madsen, D. Pultz, O. Winther, and W. Boomsma. Bend: Benchmarking dna language models on biologically meaningful tasks. arXiv preprint arXiv:2311.12570, 2023.

P. Martins, D. Mariano, F. C. Carvalho, et al. Propedia v2.3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. Frontiers in Bioinformatics, 3, 2023.

I. Mazein, A. Rougny, A. Mazein, R. Henkel, L. Gütebier, L. Michaelis, M. Ostaszewski, R. Schneider, V. Satagopam, L. J. Jensen, et al. Graph databases in systems biology: a systematic review. Briefings in Bioinformatics, 25(6):bbae561, 2024.

W. McKinney. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May, et al. The reactome pathway knowledgebase 2024. Nucleic Acids Research, 52(D1): D672–D678, 2024.

J. A. Mitchell, A. R. Aronson, J. G. Mork, L. C. Folk, S. M. Humphrey, and J. M. Ward. Gene indexing: Characterization and analysis of nlm's generifs. In AMIA Annual Symposium Proceedings, pages 460–464, 2003.

S. K. Mohamed, A. Nounu, and V. Nováček. Biological applications of knowledge graph embedding models. Briefings in bioinformatics, 22(2):1679–1693, 2021.

J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. A. Davis, A. Dobin, R. Kaul, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. Nature, 583(7818):699–710, 2020.

A. Morehead, C. Chen, A. Sedova, and J. Cheng. DIPS-Plus: The enhanced database of interacting protein structures for interface prediction. Scientific Data, 10(1):509, 2023.

E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. Science, 386(6723):eado9336, 2024.

E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani. Progen2: exploring the boundaries of protein language models. Cell systems, 14(11):968–978, 2023.

T. M. Norman, M. A. Horlbeck, J. M. Replogle, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science, 365(6455):786–793, 2019.

P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In Advances in Neural Information Processing Systems, 2023.

D. Ovek Baydar, I. Rauluseviciute, D. R. Aronsen, et al. JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models. Nucleic Acids Research, 54 (D1):D184–D193, 2026.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

Q. Pei, W. Zhang, J. Zhu, K. Wu, K. Gao, L. Wu, Y. Xia, and R. Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In Conference on Empirical Methods in Natural Language Processing, 2023.

Q. Pei, L. Wu, K. Gao, X. Liang, Y. Fang, J. Zhu, S. Xie, T. Qin, and R. Yan. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. In Findings of the Association for Computational Linguistics, 2024.

R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song. Evaluating protein transfer learning with tape. In Advances in Neural Information Processing Systems, 2019.

J. M. Replogle, R. A. Saunders, A. N. Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. Cell, 185(14):2559–2575.e28, 2022.

G. Richard, B. P. De Almeida, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, M. Lopez, A. Laterre, M. Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. bioRxiv, pages 2024–04, 2024.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 64(9):99–106, 2021.

E. W. Sayers, M. Cavanaugh, L. Frisse, K. D. Pruitt, V. A. Schneider, B. A. Underwood, L. Yankie, and I. Karsch-Mizrachi. Genbank 2025 update. Nucleic Acids Research, 53(D1):D56–D61, 2025. doi: 10.1093/nar/gkae1114.

N. Schneider, N. Stiefl, and G. A. Landrum. What's what: The (nearly) definitive guide to reaction role assignment. Journal of Chemical Information and Modeling, 56(12):2336–2346, 2016.

H. Seo, H. Ahn, J. Park, S. Han, G. Lee, S. Yang, J. S. Brown, L. Chen, G. E. Nesr, F. Eweje, et al. Vibeproteinbench: An evaluation benchmark for language-interfaced vibe protein design. arXiv preprint arXiv:2605.10978, 2026.

Y. Shen, Z. Chen, M. Mamalakis, L. He, H. Xia, T. Li, Y. Su, J. He, and Y. G. Wang. A fine-tuning dataset and benchmark for large language models for protein understanding. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2024.

M. D. Shirley, Z. Ma, B. S. Pedersen, and S. J. Wheelan. Efficient "pythonic" access to fasta files using pyfaidx. PeerJ PrePrints, 3:e970v1, 2015. doi: 10.7287/peerj.preprints.970v1.

E. Simon, K. Swanson, and J. Zou. Language models for biological research: a primer. Nature Methods, 21(8):1422–1429, 2024.

A. Subramanian, R. Narayan, S. M. Corsello, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171(6):1437–1452.e17, 2017.

D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51(D1):D638–D646, 2023.

The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. Genetics, 224(1): iyad031, 2023.

The RNAcentral Consortium. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Research, 49(D1):D212–D220, 2021.

The Tabula Sapiens Consortium. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science, 376(6594):eabl4896, 2022.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. Nucleic Acids Research, 53(D1):D609–D617, 2025.

C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, et al. Transfer learning enables predictions in network biology. Nature, 618(7965):616–624, 2023.

Trillion Labs. Gravity-16b-a3b-base, 2026. URL https://huggingface.co/trillionlabs/Gravity-16B-A3B-Base.

D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods, 13(12):966–967, 2016.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al. Proteomics. tissue-based map of the human proteome. Science, 347(6220):1260419, 2015.

I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. A. Wolf. anndata: Access and store annotated data matrices. Journal of Open Source Software, 9(101):4371, 2024. doi: 10.21105/joss.04371.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

E. Wang, S. Schmidgall, P. F. Jaeger, F. Zhang, R. Pilgrim, Y. Matias, J. Barral, D. Fleet, and S. Azizi. Txgemma: Efficient and agentic llms for therapeutics. arXiv preprint arXiv:2504.06196, 2025a.

Y. Wang, C. Tang, H. Deng, J. Xiao, J. Liu, J. Wu, J. Yao, P. Li, E. Su, L. Wang, et al. Scireasoner: Laying the scientific reasoning ground across disciplines. arXiv preprint arXiv:2509.21320, 2025b.

Y. Wu, E. Wershof, S. Schmon, M. Nassar, B. Osiński, R. Eksi, Z. Yan, R. Stark, K. Zhang, and T. Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. In Advances in Neural Information Processing Systems, 2026.

Y. Xia, P. Jin, S. Xie, L. He, C. Cao, R. Luo, G. Liu, Y. Wang, Z. Liu, Y.-J. Chen, et al. Nature language model: deciphering the language of nature for scientific discovery. arXiv preprint arXiv:2502.07527, 2025.

M. Xu, Z. Zhang, J. Lu, Z. Zhu, Y. Zhang, M. Chang, R. Liu, and J. Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. In Advances in Neural Information Processing Systems, 2022.

B. Yu, F. N. Baker, Z. Chen, X. Ning, and H. Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. arXiv preprint arXiv:2402.09391, 2024.

B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. Mendez Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kiziloren, A. Gaulton, A. P. Bento, and A. R. Leach. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, 52(D1): D1180–D1192, 2024.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? In Proceedings of the 57th annual meeting of the association for computational linguistics, pages 4791–4800, 2019.

D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, Y. Li, and W. Ouyang. Chemllm: A chemical large language model, 2024.

Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In International Conference on Learning Representations, 2024.

N. Zhu, Y. Ming, C. Zhang, C. Sen, C. Li, J. Guo, and H. Duan. PPIKB: A comprehensive knowledge base and analysis platform for protein-peptide interactions based on literature and patents. bioRxiv, 2025.

Appendix

A. Dataset Collection Details

We collect existing biological resources that already encode useful molecular, protein, genomic, cells, and cross-domain knowledge, and normalize them into a shared corpus schema.

Small molecules. We draw on large public chemical and bioactivity resources. PubChem (Kim et al., 2025) provides broad chemical identity and property coverage, DrugBank (Knox et al., 2024) contributes drug-level annotations such as development stage and mechanism of action, and BindingDB (Liu et al., 2025) and ChEMBL (Zdrazil et al., 2024) supply protein–ligand bioactivity and binding measurements. USPTO-50K (Schneider et al., 2016) provides reaction records for forward-synthesis supervision.

Proteins. We collect protein sequences, structures, and annotations from established repositories. UniProt/SwissProt (The UniProt Consortium, 2025) provides curated sequences and functional annotations, while the AlphaFold database (Bertoni et al., 2026) and the Protein Data Bank (PDB) (Berman et al., 2000) provide predicted and experimental structures.

Genomic sequences. We integrate established annotation resources. ENCODE (Moore et al., 2020) provides regulatory annotations and assay-derived evidence over functional genomic elements, GENCODE (Frankish et al., 2021) supplies reference gene and transcript annotations, and RNAcentral and Rfam (Kalvari et al., 2021; The RNAcentral Consortium, 2021) provide non-coding RNA sequences and RNA family annotations.

Cells and pathways. We include resources that describe cellular states, expression, perturbation responses, and pathway-level biology. L1000 (Subramanian et al., 2017) provides large-scale transcriptional perturbation profiles, and the Human Protein Atlas (Uhlén et al., 2015) contributes single-cell-type expression data. GeneRIF (Mitchell et al., 2003) and NCBI Gene (Maglott et al., 2005) supply gene-level functional summaries, while Gene Ontology/GOA (The Gene Ontology Consortium, 2023), Reactome (Milacic et al., 2024), and the Cell Ontology (Diehl et al., 2016) provide functional, pathway, and cell-type annotations. We also draw on CELLxGENE (CZI Cell Science Program et al., 2025) for aggregated single-cell data.

Cross-domain knowledge graphs. To link entities across domains, we integrate biomedical knowledge graphs that encode typed relations among compounds, proteins, genes, pathways, and cell contexts: Hetionet (Himmelstein et al., 2017), DRKG (Drug Repurposing Knowledge Graph Team, 2020), DGIdb (Freshour et al., 2021), STRING (Szklarczyk et al., 2023), OmniPath (Türei et al., 2016), and Reactome (Milacic et al., 2024).

B. Tool-generated Dataset Details

Small molecules. For small molecules, we convert PubChem records into computationally enriched molecule narratives using RDKit (Landrum et al., 2026). We first parse each PubChem SMILES string with RDKit, canonicalize the molecular graph when parsing succeeds, and retain source-grounded identifiers and descriptions such as SMILES, IUPAC name, InChI, molecular

formula, atom and bond counts, molecular weight, exact mass, charge state, hydrogen-bond donor/acceptor counts, rotatable bonds, and reported logP values. We then add deterministic RDKit-derived structural features that are rarely available in ordinary free-text descriptions, including ring counts and ring subtypes, aromatic/aliphatic/saturated ring composition, carbocyclic and heterocyclic ring counts, heteroatom count, amide bonds, spiro and bridgehead atoms, stereocenter counts, fraction of sp3 carbons, Murcko scaffold, Bertz complexity, Labute ASA, Hall–Kier alpha, kappa shape indices, valence and radical electron counts, functional-group matches, PAINS/Brenk alerts, and Morgan/MACCS fingerprint summaries. These computed properties are rendered into natural-language descriptions with explicit provenance, turning machine-computable molecular graph information into language-facing pretraining examples while preserving the original PubChem molecular identity.

Proteins. We generate structure, binding, and function narratives from UniRef50/UniProt-linked sequences and structures, using AlphaFold database (AFDB) (Bertoni et al., 2026) and Protein Data Bank (PDB) (Berman et al., 2000) as structure sources. We extract secondary structure and backbone hydrogen-bond annotations using DSSP (Kabsch and Sander, 1983), FreeSASA for solvent accessibility, and rule-based geometric analyses for Ramachandran regions, contact maps, long-range contacts, salt bridges, disulfide bonds, π-stacking, cation-π interactions, radius of gyration, and relative contact order. The resulting structure narratives describe each protein's sequence, residue length, structure source, physicochemical properties, structural confidence, secondary-structure composition and residue spans, fold topology, compactness, interaction network, solvent exposure, residue composition, motifs, domains, GO annotations, active sites, and low-confidence or potentially disordered regions. We further construct binding narratives from protein–protein, protein–ligand, and protein–peptide complex resources, rendering interface residues and partner context as language-facing examples. Together, these sources expose proteins as physical, functional, and interaction-domain entities, rather than only amino-acid strings or free-text mentions.

Genomic sequences. We generate narrative records from regulatory DNA intervals and RNA transcript annotations, rather than treating nucleotide sequences as plain strings. For DNA, we use GENCODE (Frankish et al., 2021) genome annotations and ENCODE (Moore et al., 2020) cCRE/peak resources to construct sequence-grounded regulatory narratives over genomic intervals. Each record combines the raw sequence with coordinate metadata, organism and assembly information, regulatory class or assay evidence when available, and deterministic sequence features such as length, GC/AT content, CpG count and observed-to-expected ratio, entropy, homopolymer runs, palindromic k-mers, ambiguous-base content, and simple composition flags. For RNA, we build transcript-domain narratives from GENCODE transcript sequences and related RNA resources (Kalvari et al., 2021; The RNAcentral Consortium, 2021), adding computable features such as base composition, dinucleotide counts, GC/AU skew, AUG and stop-codon counts, ORF statistics, polyadenylation signal motifs, U-rich windows, and deterministic stem-loop candidates.

Cells. We convert cellular profile resources into language-facing records that capture expression, spatial context, perturbation response, and morphology. Single-cell records describe cell states from ranked or highly expressed marker genes and associated tissue context, while perturbation records connect CRISPRa/CRISPRi perturbations to pseudobulk differential-expression signatures. For spatial biology, we generate HuBMAP-derived spatial transcriptomics narratives (Börner et al., 2025) that attach each spot to its tissue, assay, coordinate, detected-gene

count, total transcript count, local neighborhood density, nearest-neighbor distance, and locally abundant transcripts. For morphology, we generate JUMP Cell Painting narratives (Chandrasekaran et al., 2024) from matched-control-normalized CellProfiler profiles, summarizing perturbation-domain phenotype strength, replicate consistency, morphology-family shifts across nuclei, cytoplasm, and whole cells, and the top contributing image-derived measurements.

Cross-domain linking. Cross-domain KG-linked records are constructed as chain-style training examples that co-locate evidence across biological domains. We use a knowledge-graph structure to decide which entities should appear together, then hydrate those entities with payloads from the original databases. Gene-anchored chains combine UniProt/SwissProt protein function and sequence, BindingDB compound–target measurements, Reactome pathway context, STRING interaction context, and LINCS L1000 perturbation signatures (Liu et al., 2025; Milacic et al., 2024; Subramanian et al., 2017; Szklarczyk et al., 2023; The UniProt Consortium, 2025). Drug-anchored chains connect DrugBank/PubChem molecular identity to L1000 transcriptional responses through shared compound identifiers (Kim et al., 2025; Knox et al., 2024; Subramanian et al., 2017). Cell-anchored chains connect CellxGene expression profiles to SwissProt annotations of marker genes (CZI Cell Science Program et al., 2025; The UniProt Consortium, 2025). Additional chains use Hetionet, DRKG, STRING, and OmniPath/DoRothEA to link compounds, proteins, pathways, transcription factors, and cell-domain expression evidence (Drug Repurposing Knowledge Graph Team, 2020; Garcia-Alonso et al., 2019; Himmelstein et al., 2017; Szklarczyk et al., 2023). These records are cross-domain because the supervision is the structured co-occurrence of molecular, protein, pathway, perturbation, and cell-context evidence, rather than a single-database fact or an unrelated mixture of domains.

C. In-house Instruction Dataset Details

Protein binding. We construct a protein-binding instruction dataset from structural complex records, not from free-text protein descriptions. For protein–ligand binding, we use PLINDER experimental and predicted-linked complexes (Durairaj et al., 2024); for protein–protein binding, we use PINDER holo complexes (Kovtun et al., 2024), PPIRef50K (Bushuiev et al., 2024), DIPS-Plus (Morehead et al., 2023), PDB-derived domain-domain complexes (Berman et al., 2000), and high-confidence AlphaFold-derived complexes (Bertoni et al., 2026); and for protein–peptide binding, we use PDB direct complexes, Propedia (Martins et al., 2023), and PPIKB-derived peptide–protein records (Zhu et al., 2025). For each complex, we parse the target and binder chains, compute heavy-atom contacts using a 5.0 Å cutoff, derive target-side and binder-side interface residues, and retain contact metadata such as contact count and contact type. These structured interfaces are then rendered into answer-supervised tasks: protein–ligand examples ask for binding-site residue listing, masked binding-site recovery, ligand-conditioned protein scaffolding, observed-interface summarization, or reference binder sequence generation; protein–protein examples ask for a binder sequence conditioned on a target protein and epitope residues; and protein–peptide examples ask for target-interface residues, peptide-interface residues, masked peptide-interface recovery, peptide binder generation, or protein–peptide interface summarization. We apply source-specific filters on sequence length, ligand size, contact count, interface size, and experimental resolution where available, deduplicate by unit/interface/prompt hashes, and exclude protein sequences overlapping binder-generation evaluation examples. The append step selects 12.0M binding instruction records before the final context-length filtering stage.

Table 7 | Full subtask performance for the small molecule domain. Bold indicates the best model in each scored row.

Subtask	Metric	Base	Ours
Description-guided molecular design	Validity ↑	0.200	**0.890**
	MACCS Tanimoto ↑	0.237	**0.454**
	Morgan Tanimoto ↑	0.097	**0.149**
	SMILES Levenshtein ↓	191.1	**99.7**
Molecular reconstruction	Validity ↑	0.060	**0.930**
	MACCS Tanimoto ↑	0.425	**0.508**
	Morgan Tanimoto ↑	0.180	**0.202**
	SMILES Levenshtein ↓	103.0	**50.7**
Molecular recognition	Both fields exact ↑	0.220	**0.290**
	result_1 exact ↑	0.300	**0.500**
	result_2 exact ↑	0.320	**0.380**
Forward synthesis	Canonical exact ↑	0.000	**0.140**
	Validity ↑	0.340	**1.000**
	MACCS Tanimoto ↑	0.334	**0.756**
	Morgan Tanimoto ↑	0.177	**0.578**
	SMILES Levenshtein ↓	64.9	**16.2**

DNA/RNA feature localization. DNA and RNA feature-localization instructions are constructed as exact span-recovery tasks over nucleotide sequences. For DNA, we use ENCODE SCREEN cCRE intervals and ENCODE open-chromatin peaks (Moore et al., 2020), GENCODE reference genome and transcript annotations for transcription start sites and splice donor/acceptor sites (Frankish et al., 2021), and JASPAR position-frequency matrices for transcription-factor motif hits (Ovek Baydar et al., 2026). Each example extracts a sequence window containing exactly one target feature, surrounds it with flanking sequence, and asks the model to return valid JSON with the feature label, 1-indexed inclusive start and end coordinates, and the exact subsequence copied from the input. For RNA, we build the same instruction interface from GENCODE CDS/UTR annotations (Frankish et al., 2021), RNAcentral exon-junction records (The RNAcentral Consortium, 2021), Rfam family hits (Kalvari et al., 2021), miRBase mature-miRNA segments (Kozomara et al., 2019), and GtRNAdb tRNA anticodons (Chan and Lowe, 2016). We reject ambiguous windows, windows with excessive unknown bases, features that do not occur exactly once when uniqueness is required, and records whose target subsequence fails a deterministic self-check. This makes regulatory and transcript annotations trainable as precise language-mediated extraction tasks, rather than as passive sequence narratives.

D. Full Experimental Results

In this section, we report subtask-wise performance across small molecules (Table 7), proteins (Table 8), genomic sequences (Table 9), and cell/pathway (Table 10).

D.1. Small molecules

For small molecules, the base model struggles to generate valid molecular outputs: generation validity never exceeds 0.20, and forward synthesis yields no exact-match products. In contrast, training on THEBIOCOLLECTION improves both validity and structural accuracy. The model trained on our corpus increases chemical validity to 0.89–1.00 and consistently achieves higher

Table 8 | Full subtask performance for the protein domain. Bold indicates the best model in each scored row.

Subtask	Metric	Base	Ours
InterPro function prediction	InterPro ID precision ↑	0.000	**0.075**
	InterPro ID recall ↑	0.000	**0.035**
Text-conditioned functional protein design	Functional consistency ↑	0.321	**0.427**
	Nondegeneracy ↑	0.050	**0.930**
	pLDDT ↑	0.439	**0.472**
	pTM ↑	0.160	**0.258**
Binder design	ipTM ↑	0.131	**0.289**
	Complex pLDDT ↑	0.452	**0.646**
	Nondegeneracy ↑	0.120	**1.000**

fingerprint similarity to the reference molecules. The most pronounced improvement is observed in forward synthesis, where correct products emerge despite zero exact matches from the vanilla model.

D.2. Proteins

In the protein domain, the vanilla model collapses to degenerate, repetitive sequences, with nondegeneracy near 0.05. We observe a paradoxical relationship between fold confidence (pLDDT) and nondegeneracy, suggesting that structural confidence alone can be misleading. We attribute this behavior to the tendency of structure prediction models to assign high confidence to low-complexity repetitive sequences when they match memorized local sequence patterns. For example, repeated motifs such as `QVQVQ...` may readily induce plausible local structural arrangements through learned subsequence priors despite lacking meaningful biochemical diversity. In contrast, our trained model raises nondegeneracy to 0.93 for monomer design and 1.00 for binder generation.

Additionally, we find that training on THEBIOCOLLECTION provides the smallest gains for protein function prediction. We read this as a coverage gap, where function-prediction instructions form a relatively small and narrow slice of the corpus next to the much larger generation and structure tasks; the model sees few examples to improve here. Scaling up and diversifying that supervision is the natural future direction.

D.3. Genomic sequences

Genomic sequence tasks are particularly challenging because they require recovering structured annotations rather than predicting a single label. The base model rarely recovers complete structured answers, with Full JSON exact-match scores remaining near zero across most tasks. In contrast, our trained model substantially improves genomic feature localization, achieving perfect performance on splice-site localization and large gains on cCRE, open-chromatin, and RNA-element localization. These improvements are particularly pronounced for structured span recovery; for example, Full JSON exact match on splice-site localization increases from 0.13 to 1.00, indicating that the model learns to identify and extract biologically meaningful sequence features.

Table 9 | Full subtask performance for the genomic sequence domain. Bold indicates the best model in each scored row.

Subtask	Metric	Base	Ours
cCRE class and genomic span	Full JSON exact ↑	0.000	**0.010**
	Coordinate exact ↑	0.000	**0.420**
	Categorical exact ↑	0.030	**0.290**
	Subsequence exact ↑	0.000	**0.170**
Open chromatin peak and summit	Full JSON exact ↑	0.000	**0.080**
	Coordinate exact ↑	0.000	**0.220**
	Categorical exact ↑	0.130	**0.850**
	Subsequence exact ↑	0.000	**0.150**
Splice-site type, strand, and motif span	Full JSON exact ↑	0.130	**1.000**
	Coordinate exact ↑	0.820	**1.000**
	Categorical exact ↑	0.290	**1.000**
	Subsequence exact ↑	0.210	**1.000**
Rfam family and RNA element span	Full JSON exact ↑	0.000	**0.370**
	Coordinate exact ↑	0.790	**0.840**
	Categorical exact ↑	0.800	**0.840**
	Subsequence exact ↑	0.000	**0.380**
tRNA amino acid and anticodon span	Full JSON exact ↑	0.010	**0.100**
	Coordinate exact ↑	0.270	**0.400**
	Categorical exact ↑	0.010	**0.120**
	Subsequence exact ↑	0.020	**0.120**

Table 10 | Full subtask performance for the cell/pathway domain. Bold indicates the best model in each scored row.

Subtask	Metric	Base	Ours
Cell type from marker genes	Exact match ↑	0.470	**0.580**
Hallmark pathway from gene set	Exact match ↑	0.520	**0.750**
Gene regulation after K562 CRISPRi knockdown	Direction accuracy ↑	0.000	**0.468**
	Macro-F1 ↑	0.045	**0.349**
	Candidate coverage ↑	0.000	**0.675**

D.4. Cell/Pathway

Cell and pathway tasks improve consistently across all subtasks. Cell-type recognition increases from 0.47 to 0.58, while hallmark-program recognition improves from 0.52 to 0.75. The largest gains are observed in perturbation-response prediction, where the vanilla model scores zero on both direction accuracy and candidate coverage, whereas our model reaches 0.468 direction accuracy and 0.675 candidate coverage. These results suggest that the corpus improves coverage across diverse cell-related tasks, whereas the vanilla model struggles on perturbation response (gene regulation) prediction.

E. Evaluation Details

We evaluate THEBIOCOLLECTION with task-specific metrics that match the structure of each biological output. For all tasks, model generations are first parsed into the expected answer

Table 11 | Evaluation dataset statistics of THEBIOCOLLECTION-EVAL.

Domain	Evaluation subtask	Queries
Small molecules	Description-guided molecule design	100
	Molecular reconstruction	100
	Molecular recognition	100
	Forward synthesis prediction	100
	Subtotal	**400**
Proteins	Text-conditioned protein design	100
	Binder generation	100
	Protein function prediction	100
	Subtotal	**300**
Genomic sequences	cCRE localization	100
	Open-chromatin localization	100
	Splice-site localization	100
	Rfam hit localization	100
	tRNA anticodon localization	100
	Subtotal	**500**
Cells/pathways	Tabula Sapiens cell-type classification	100
	Hallmark pathway classification	100
	Replogle K562 CRISPRi perturbation-response prediction	100
	Subtotal	**300**
Cross-domain reasoning	Protein function → pathway	50
	TF function → regulated target gene	50
	Small molecule → binding target, pathway	50
	Subtotal	**150**
All domains	**Total queries**	**1,650**

format and normalized with the same deterministic post-processing used for scoring. For the main summary table, scores are averaged first across the reported metrics within a subtask and then across subtasks within each biological domain. The statistics of our evaluation dataset is in Table 11, and the representative subtask-wise evaluation queries are shown in Table 12. We detail the metrics we adopted for each domain as follows.

Small molecules

- **Validity:** the fraction of model outputs that can be parsed as chemically valid SMILES by RDKit.
- **Exact match (exact):** compares the predicted and reference answers. For a canonical exact match, the predicted and reference molecules are matched after canonical SMILES normalization.
- **MACCS and Morgan Tanimoto similarity:** fingerprint overlap between the predicted and reference molecules.
- **SMILES Levenshtein distance:** edit distance between canonical predicted and reference SMILES strings.

Proteins

- **InterPro precision:** the fraction of predicted functional annotations that are correct.
- **InterPro recall:** the fraction of gold annotations recovered by the model.
- **Functional consistency:** the cosine similarity between embeddings of the generated protein

and the reference protein from ESM-C (Candido et al., 2026).

- **Nondegeneracy:** whether generated protein sequences avoid low-complexity or repetitive artifacts. A sequence can be syntactically valid while still being biologically implausible; nondegeneracy separates valid amino-acid output from collapsed or repetitive generations.
- **Fold quality:** pLDDT, pTM, and ipTM are fold-confidence metrics computed from ESMFold2-fast (Candido et al., 2026). These metrics assess the foldability of the given sequence in either a single-chain or a complex context.

Genomic sequences

- **Full JSON exact match (Full JSON exact):** whether the complete predicted JSON object matches the gold answer after canonical serialization.
- **Coordinate exact match (Coordinate exact):** compares only coordinate fields such as start and end positions.
- **Categorical exact match (Categorical exact):** compares non-coordinate annotation fields, such as feature type, class, strand, family, or motif label, depending on the subtask.
- **Subsequence exact match (Subsequence exact):** compares the sequence field copied from the input window.

Cell/Pathway

- **Direction accuracy:** the fraction of candidate genes whose predicted direction matches the gold differential-expression direction after perturbation.
- **Macro-F1:** computes F1 separately for up-regulated and down-regulated gene sets and averages the two classes.
- **Candidate coverage:** the fraction of candidate genes for which the model emits a valid direction prediction.

Cross-domain understanding

- **Accuracy:** each task is a four-way multiple-choice question; we select the option with the highest model likelihood and count it correct when it matches the gold choice.

F. Example Corpus Records

We provide representative records from THEBIOCOLLECTION for each major corpus-construction stage described in Section 2. These examples illustrate how structured biological resources are transformed into language-model-readable training records, from dataset refinement and feature enrichment to instruction-based supervision.

Dataset Refinement (Section 2.2)

The examples below illustrate how biological databases are converted into natural-language records while preserving the underlying biological information.

Example 1. Protein–ligand binding record

```
The ligand 2-{[4-(but-2-yn-1-ylamino)benzene]sulfonyl}ethane-1-thiol
(SMILES: <smiles>CC#CCNc1ccc(cc1)S(=O)(=O)CCS</smiles>,
InChI: <inchi>InChI=1S/C12H15NO2S2/c1-2-3-8-13-11-4-6-12(7-5-11)17(14,15)10-9-16/h4-7,13,
16H,8-10H2,1H3</inchi>)
```

```
was tested against the protein Disintegrin and metalloproteinase domain-containing protein
17 [215-477,S266A,N452Q].

The target protein name is Homo sapiens of ADA17_HUMAN and chain sequence is
<protein>RADPDPMKNTCKLLVVADHRFYRYMGRGEESTTTNYLIELIDRVDDIYRNTAWDNAGFKGYGIQIEQIRILKSPQEVKPGE
KHYNMAKSYPNEEKDAWDVKMLLEQFSFDIAEEASKVCLAHLFTYQDFDMGTLGLAYVGSPRANSHGGVCPKAYYSPVGKKNIYLNSGLT
STKNYGKTILTKEADLVTTHELGHNFGAEHDPDGLAECAPNEDQGGKYVMYPIAVSGDHENNKMFSQCSKQSIYKTIESKAQECFQERSN
KV</protein>.
PDB IDs of target chain are 2A8H,1ZXC,3G42,2I47,2OI0,3EWJ,3EDZ,3E8R,3B92,3CKI,1BKC

The experiment was performed at pH 7.5000 and 22.00 C.
The ligand exhibited strong binding strength, with a reported Ki (nM) of 2.
```

Example 2. Cellular perturbation record

```
In HA1E (kidney) cells, palmitoylethanolamide (<smiles>CCCCCCCCCCCCCCCC(=O)NCCO</smiles>) at
0.04 uM for 24 h up-regulates SUV39H1, CLPX, SRC and down-regulates GJA1, HMGCS1, DNM1L (|z|
>= 2.0), based on LINCS L1000 signature REP.A023_HA1E_24H:B24.
```

Tool-Computed Feature Narratives (Section 2.3)

These examples demonstrate records enriched with tool-computed biological features that are not explicitly present in the original source databases.

Example 3. Molecular feature narrative

```
The molecule has the SMILES representation: <smiles>C1C=CC2=CSC=C2O1</smiles>
and its IUPAC systematic name is 2H-thieno[3,4-b]pyran. It also has the corresponding InChI:
<inchi>InChI=1S/C7H6OS/c1-2-6-4-9-5-7(6)8-3-1/h1-2,4-5H,3H2</inchi>.

The compound contains 15 atoms, including 9 heavy atoms, and 16 bonds. The molecule is
Achiral, and it has 0 charged atoms, and a total formal charge of 0. Its molecular formula
is C7H6OS, giving an exact mass of 138.01393598 g/mol and a molecular weight of 138.19
g/mol.

The molecule features 2 hydrogen bond acceptors, 0 hydrogen bond donor, 0 rotatable bond,
and an XlogP3-AA value of 1.9, indicating moderate polarity, balanced aqueous solubility and
membrane permeability.

There are 2 rings in the molecular graph, including 1 aromatic, 1 aliphatic, and 0 saturated
rings. Ring subtype counts include 0 aromatic carbocyclic, 1 aromatic heterocyclic, 0
aliphatic carbocyclic, 1 aliphatic heterocyclic, 0 saturated carbocyclic, and 0 saturated
heterocyclic rings. The graph contains 2 heteroatoms, 0 amide bonds, 0 spiro atoms, and 0
bridgehead atoms. Its fraction sp3 is 0.14. No potential stereocenters are detected in the
parsed molecular graph. The Murcko scaffold is C1=Cc2cscc2OC1. Graph complexity descriptors
include Bertz complexity 236.8, Labute ASA 57.6 A^2, Hall-Kier alpha -0.8, and Kier kappa
values 5.1, 1.9, and 0.8. The parsed graph has 46 valence electrons and 0 radical electrons.

Additional physicochemical descriptors: heavy-atom molecular weight 132.14, topological
polar surface area (TPSA) 9.2 A^2, molar refractivity 38.8, QED drug-likeness 0.533, 0
Lipinski rule-of-five violations, NHOH count 0, N+O count 1.

Structural alerts: 0 PAINS, 0 Brenk. Substructure map (0-based atom indices follow the given
SMILES): Functional groups -> none. Aromatic rings -> 1 thiophene at [3, 4, 5, 6, 7].
Ring-junction atoms (shared by two rings): [3, 7]. Per-atom detail: Atom 5 is S (aromatic);
1-hop neighbors: 4 (aromatic), 6 (aromatic); 2-hop neighbors [3, 7]; 3-hop neighbors [2, 8].
Atom 0 is C; 1-hop neighbors: 1 (single), 8 (single); 2-hop neighbors [2, 7]; 3-hop
neighbors [3, 6].
```

Example 4. Spatial transcriptomics narrative

```
HuBMAP tissue-context spatial record:

sample=HBM332.KQWB.454,
assay=Slide-seq [Salmon],
organism=Homo sapiens,
region=Kidney (Right),
spot=CAGATCCACAACGC,
coordinate=x=1977.5, y=4192.2,
annotation=not reported.

Observed RNA capture:
total_counts=222.1,
detected_genes=236.

Retained spot-local transcripts:
MALAT1, MT-RNR2, CYTB, ND2, UMOD, ALDOB, ATP6, GUK1.

Spatial context:
12 neighboring spots within 50 um,
nearest_neighbor=17.4 um.
```

Instruction-Tuning Datasets (Section 2.4)

The examples below illustrate instruction–answer pairs designed to capture biological tasks that are difficult to obtain from conventional text-based biological datasets.

Example 5. Masked protein binding-site infilling

```
Binding-site residues in the protein sequence are masked for ligand-conditioned infilling.
Predict the original amino acid identity at each masked binding-site position. Positions are
1-indexed within the protein chain.

Ligand (SMILES): <smiles>NCC1=CC([N+](=O)[O-])C=CC1=O</smiles>

Protein (chain A), binding-site residues masked:
<protein>MNTV<mask><mask><mask>SAPIEVTIGIDQYSFNVKENQPFHGIKDIPIGHVHVIHFQHADNSSMRYGYWFD
CRMGNF<mask>IQYDPKDGLYKMME<mask>RDGA<mask<mask>EN<mask>VHNFKERQMMVSYPKIDEDDTWYNLTEFVQ
MDKIRKIVRKDENQFSYVDSSMTTVQENELSDPAHSLNYTVINFKSREAIRPGHEMEDFLDKSYYLNTVMLQGIFKNSSNYFGEL
QFAFLNAMFFGNYGSSLQWHAMIELICSSATVPKHMLDKLDEILYYQIKTLPEQYSDILLNERVWNICLYSSFQKNSLHNTEKIM
ENKYPELL</protein>
Masked positions: A9, A10, A11, A72, A87, A92, A93, A96

List the recovered amino acids by position as chain:residue amino acid and 1-indexed
position tokens (e.g., A:V15).

Answer: A:P9, A:F10, A:T11, A:Y72, A:E87, A:K92, A:F93, A:I96
```

Example 6. Genomic motif localization

```
You are given a human GRCh38 genomic DNA sequence, taken from the genome forward (+) strand,
containing one high-confidence JASPAR transcription factor (TF) binding motif plus flanking
DNA. The target motif is a high-confidence JASPAR pattern embedded in flanking genomic DNA.

Sequence:
<dna>AATTCATTACAGGCTTCAGAGGCTGAGCTTTTTGTGTTTCATTGGTTGAACACTTGATTATGGATTACAACTTATTTTGAACCACA
AGCCATCAAAATTTAGATGACTTGTTTCCAAAACTGGAGTACTTTTTTTGTAGAGCATTTGGATTGACTCATGAAGTATGTTTTAAAGTTA
AAAGCTGCCTAATGTCAGCAATTTCCTATCATTTCATTATTTCAAGTGTCAAATATTTCTCAAATCCAACTTGACAGGCAGCAAAGAGAGA
ACCGCACACTCTACTGATGAAATGTTTTTAATCCAATTTA</dna>

Task: Locate the TF motif hit.

- "strand" is "+" if the motif matches the sequence as written,
  or "-" if it matches the reverse complement.
- "subsequence" is always the forward-strand characters from
  start to end, regardless of "strand".
```

```
Coordinate rules:
- Positions are 1-indexed.
- All reported interval endpoints are inclusive.
- The output subsequence field must exactly match the characters
  of the input sequence over the interval specified for that task.
- The sequence contains exactly ONE target element.
  Return exactly one object.

Output rules:
- Output valid JSON only. No markdown fences, no comments,
  no explanation.
- Use exactly the keys shown. Do not add or rename keys.

Return JSON only:

{"motif_hits":[{"motif_name":"<string>","start":<int>,"end":<int>,
"strand":"<+ or ->","subsequence":"<string>"}]}

Answer:
{"motif_hits":[{"motif_name":"FOS","start":151,"end":158,
"strand":"+","subsequence":"TGACTCAT"}]}
```

Table 12 | Example evaluation questions by subtask. For convenience, each example is the final held-out question only, with few-shot demonstrations removed.

Domain	Evaluation subtask	Example question
Small molecule	Description-guided molecule design	Synthesize a molecule that matches the given characteristics. The molecule appears as colorless crystals. Insoluble in water.
Small molecule	Molecular reconstruction	Reconstruct the molecule described below and answer with one SMILES string only. Begin with a six-membered benzene ring. At position 1, attach an amide connected to a piperidine ring whose nitrogen bears a sulfonyl-fused 1,2,5-oxadiazole/benzene system; at position 4, attach an isopropyl group.
Small molecule	Molecular recognition	Analyze the molecule and answer as `result_1: <value> ; result_2: <value>`. SMILES: `<smiles>C(=CCC=CCC)CC...</smiles>`. Task: ester. Report the number of esters and ester atom indices.
Small molecule	Forward synthesis prediction	Predict the product of the reaction from the given reactants and reagents: `<smiles>C1CCOC1.CCN(CC)...</smiles>`.
Protein	Text-conditioned protein design	Create a protein sequence with the necessary features to perform the desired function. The designed protein must contain at least one GTP cyclohydrolase II domain, include an accessible Mg(2+) binding site, and contain a structurally stable nucleophile for GTP cyclohydrolase activity.
Protein	Binder generation	Design a de novo protein binder for a target protein. Use only the 20 standard amino acids and produce a single chain of exactly 204 residues. The binder should engage the listed target epitope residues on chain R; output only the binder sequence.
Protein	Protein function prediction	Given protein information for human SWI/SNF-related chromatin regulator SMARCA4, with InterPro annotations hidden, predict likely InterPro domain, family, repeat, site, or homologous-superfamily IDs. Output only semicolon-separated InterPro IDs. Sequence: `<protein>XRRDTALETALN...</protein>`.
Genomics	cCRE localization	Given a human GRCh38 DNA sequence containing exactly one candidate cis-regulatory element, locate the cCRE and report its class and exact span. Return JSON with one region containing the label, start, and end. Sequence: `<dna>TCAAAAGAAAAA...</dna>`.
Genomics	Open-chromatin localization	Given a human GRCh38 DNA sequence measured in K562, locate the single open-chromatin peak and its summit position. Return JSON with assay, biosample, start, end, and summit position. Sequence: `<dna>AGTGTGCTCACA...</dna>`.
Genomics	Splice-site localization	Given a human GRCh38 DNA sequence containing one annotated exon-intron boundary, locate the splice site and report site type, strand, boundary position, and 2-bp motif span. Sequence: `<dna>CCTGGGTCAGAG...</dna>`.
Genomics	Rfam hit localization	Given an RNA sequence containing one conserved non-coding RNA family element, identify the Rfam family and report the element span. Sequence: `<rna>GCCUGGCUGGCU...</rna>`.
Genomics	tRNA anticodon localization	Given a tRNA sequence, locate the annotated anticodon triplet and report the carried amino acid. Sequence: `<rna>GCCGCAAUAGCU...</rna>`.
Cell/pathway	Cell type classification	Given marker genes `CD24, PAX5, DNTT, PTPRC, ...` from the immune system, choose the single best supported cell type from candidates such as effector CD4+ T cells, memory CD4+ T cells, Pro-B cells, naive T cells, and classical monocytes. Return only the option letter.
Cell/pathway	Hallmark pathway classification	Given gene set `WNT1, CSNK1E, JAG1, DVL2, ...`, choose which biological program or pathway is most strongly represented from candidate Hallmark programs. Return only the option letter.
Cell/pathway	Perturbation-response prediction	Cell line: K562. Perturbation type: CRISPRi knockdown. Perturbed gene: `HINFP`. Given candidate genes `TNNI3, AIF1, MT-ND1, MT-CO3, ...`, predict which are up- or down-regulated and return JSON with `up` and `down` lists.
Cross-domain	Protein function → pathway	In which of the following pathways does this human protein participate? Protein description: "Receptor for ghrelin, coupled to G-alpha-11 proteins".
Cross-domain	TF function → regulated gene	A human transcription factor is described below and directly stimulates the transcription of one of the following genes (OmniPath). Which one? Function: "Transcriptional activator which binds specifically to the MEF2 element. . . found in numerous muscle-specific, growth factor- and stress-induced genes".
Cross-domain	Small molecule → target, pathway	The small molecule below binds a human protein target (DRKG). Which option correctly pairs its target with a pathway that target participates in? SMILES: `<smiles>Cc1cccc...</smiles>`.