RESEARCH ARTICLE SUMMARY

ARTIFICIAL INTELLIGENCE

Autonomous biomedical research with an artificial intelligence agent

Kexin Huang*†, Serena Zhang†, Hanchen Wang†, Yuanhao Qu†, Yingzhou Lu†, Ryan Li† *et al.*

Full article and list of author affiliations: https://doi.org/10.1126/science.adz4351

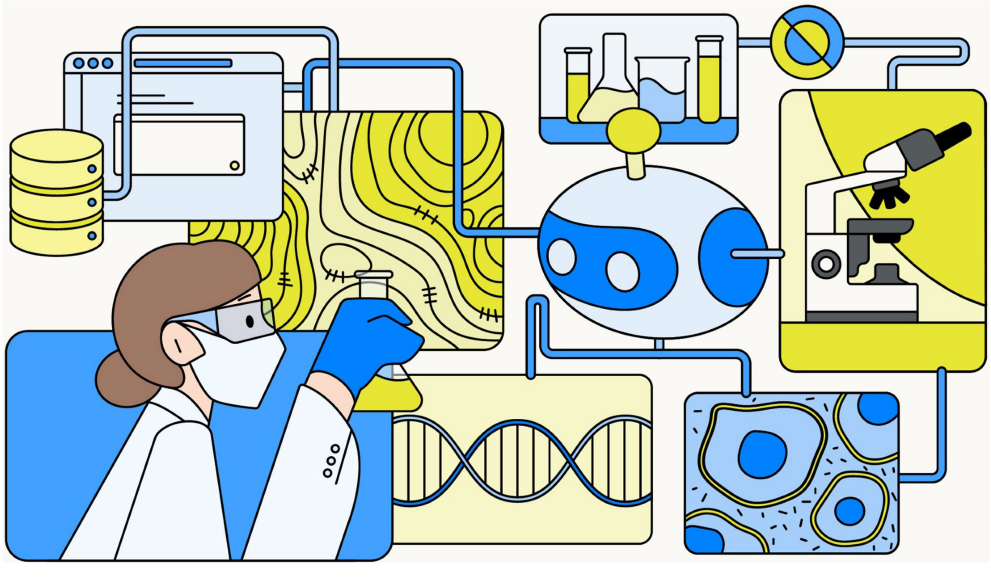
INTRODUCTION: Modern biology and medicine research generate data far faster than can be analyzed. A single study can require the use of dozens of specialized software programs, combing through many years of literature results, designing detailed experimental protocols, and evaluating statistics. As a result, valuable datasets sit unexamined, and connections across existing knowledge go unmade.

RATIONALE: We designed Biomni as a single artificial intelligence (AI) system meant to handle many kinds of research. An AI agent—software that plans and acts on its own—needs to be able to access specialized tools and datasets. The first step to designing Biomni was assembling these digital resources into one shared workspace of 150 analysis tools and dozens of software packages and databases spanning 25 areas of biology. Given this base of established methods, we designed the Biomni system to process and respond to questions in plain English using large language models. The agentic framework enables complex workflow planning, writing, executing code to analyze data, error checks, and adjustments.

RESULTS: Across 443 questions spanning 10 kinds of biomedical tasks, Biomni achieves an average accuracy of 57%, considerably higher than that achieved by other agentic systems on the same benchmarks. On three expert-level tasks, it matched specialists in accuracy while only requiring a fraction of the time. We also assessed Biomni in real-world experiments across very different disciplines. We used Biomni to develop an analysis pipeline to process previously published smartwatch readings and recovered known early warning signs of COVID-19 infection. Biomni designed a gene-editing cloning protocol that bench scientists ran exactly as written, with sequencing confirming success. We redesigned a protein for greater heat stability, with Biomni creating a computational optimization procedure that proposed three rational mutations consistent with state-of-the-art protein thermostability engineering principles. We also prompted Biomni to write ready-to-run code to drive a laboratory robot through a multistep drug dose–response experiment. Each of these experiments drew on computational and wet-lab methods curated during optimization of Biomni but without requiring extensive specialized knowledge and coding ability of the human deciding the research agenda.

CONCLUSION: As a general-purpose agent, Biomni can design workflows for experiments in fields ranging from genetics, immunology, pharmacology, to clinical medicine, without being retooled or retrained on additional, domain-specific protocols. We propose that such agentic AI systems will be able to aid in accelerating cumbersome analysis tasks, thus enabling scientists to focus on framing questions, judging results, and pursuing the creative leaps that machines do not make. Biomni is openly available, and its developers emphasize responsible use as these tools become more powerful. □

*Corresponding authors: Kexin Huang (kexinh@cs.stanford.edu); Jure Leskovec (jure@cs.stanford.edu) †These authors contributed equally to this work. **Cite this article as** K. Huang *et al.*, *Science* **393**, eadz4351 (2026). DOI: 10.1126/science.adz4351



Biomni is an AI agent that enables biomedical research. A scientist asks a question in plain English; Biomni picks the right tools, plans the steps, and writes and runs its own code to deliver an answer. [Illustration: Jasmine Zhang]

ARTIFICIAL INTELLIGENCE

Autonomous biomedical research with an artificial intelligence agent

Kexin Huang[1,2]*†, Serena Zhang[1,2]†, Hanchen Wang[1,3]†, Yuanhao Qu[2,4,5,6]†, Yingzhou Lu[6]†, Ryan Li[1]†, Yusuf Roohani[1,7], Lin Qiu[8], Shiyi Cao[9], Gavin Li[1], Junze Zhang[4,6], Di Yin[4,6], Rick Wierenga[10], Deniz Kavi[11], Sherry Liu[11], Tianwei She[2], Shruti Marwaha[12], Jennefer N. Carter[12], Xin Zhou[6], Matthew T. Wheeler[12], Jonathan A. Bernstein[13], Mengdi Wang[14], Peng He[15], Jingtian Zhou[7], Michael P. Snyder[6], Le Cong[4,6], Aviv Regev[3], Jure Leskovec[1]*

Biomedical research is increasingly constrained by repetitive, fragmented workflows that slow discovery. We introduce Biomni, a general-purpose biomedical artificial intelligence agent that autonomously executes diverse research tasks. To map the biomedical action space, Biomni's action-discovery agent mines tools, databases, and protocols from thousands of publications across 25 domains, building a unified agentic environment. Its general-purpose architecture integrates large language model reasoning with retrieval-augmented planning and code-based execution, dynamically composing workflows without predefined templates. Systematic benchmarking shows strong generalization across heterogeneous tasks—causal gene prioritization, drug repurposing, rare-disease diagnosis, microbiome analysis, and molecular cloning—without task-specific tuning. Real-world case studies demonstrate Biomni interpreting multimodal datasets, optimizing protein stability, orchestrating wet-lab instruments, and generating experimentally testable protocols. Biomni envisions artificial intelligence augmenting human scientists and accelerating discovery.

Biomedical research is a key pillar of modern science and medicine, driving discoveries in disease mechanisms, diagnostics, and therapeutics (*1*–*4*). Yet, with the growth in large-scale experiments, data, tools, and literature, progress is increasingly slowed by fragmented, complex workflows that make it difficult to efficiently integrate specialized tools, exhaustive literature reviews, intricate experimental design, and rigorous statistical modeling (*5*, *6*). A vast volume of valuable biomedical data is underused (*7*), many sophisticated analyses are not conducted, and many connections for past knowledge and literature are not made, not for lack of value but because the demand for expert researchers far exceeds the supply. This mismatch between data abundance and limited human bandwidth highlights an urgent need for approaches that can effectively scale expertise, streamline workflows, and unlock the full potential of biomedical research.

Artificial intelligence (AI) agents have substantially reshaped fields such as software engineering (*8*), law (*9*), materials science (*10*), and health care (*11*) by automating repetitive tasks, enhancing productivity, and enabling capabilities that were previously difficult to achieve. The development of AI agents also offers opportunities to reshape aspects of biomedical research (*12*). We envision such agents tackling diverse biomedical research tasks across subfields, thus augmenting human biologists' specialized expertise. Capable of efficiently managing thousands of concurrent tasks, AI agents could enhance human productivity and accelerate the pace of biomedical discovery.

Previous and current biomedical AI agents are largely specialist systems built for narrow tasks (*13*–*20*), which limits their ability to generalize across biomedical domains. Virtual Lab orchestrates multiagent large language model (LLM) workflows for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) nanobody design (*14*), and BioDiscoveryAgent runs closed-loop gene-perturbation selection within predefined screens (*15*). CellAgent autonomously configures and executes single-cell RNA sequencing (scRNA-seq) analysis pipelines (*17*), and scBaseCount uses agents to curate and continually expand a single-cell data repository (*21*)—both powerful but confined to single-cell omics. More recent efforts extend agents to additional modalities and tasks (*22*): SpatialAgent applies multiagent reasoning to spatial transcriptomics for spatially aware inference and hypothesis exploration (*23*), and parallel advances in self-evolving, reinforcement learning (RL)–driven, and automated hypothesis-generation architectures show promise (*24*), yet each remains restricted to a task-specific action space (*25*).

Enabling an AI agent to handle a broad range of biomedical tasks introduces substantial technical challenges—most notably, the need to tightly couple advanced reasoning (*26*) with the ability to execute highly specialized biomedical actions (*27*). Although LLM-based reasoning has seen major advances (*28*), such LLMs need access to an environment that explicitly defines the biomedical action space, which is inherently diverse, domain specific, and complex. Moreover, a truly capable system requires an agentic architecture that can natively interact with this biomedical environment to select and execute diverse tasks without relying on predefined workflows.

In this work, we present Biomni, a general-purpose biomedical AI agent purpose-built to assist and partially automate biomedical research across a wide range of subfields. Biomni formulates hypotheses, performs complex bioinformatics analyses, and designs rigorous experimental protocols. To enable this capability, we first constructed a unified and comprehensive biomedical action space by systematically analyzing 2500 biomedical research papers spanning 25 distinct subfields, curated from literature repositories. From this foundation, we developed an LLM-powered action-discovery agent capable of reading papers and extracting key tasks, tools, and databases essential to driving biomedical discoveries. These elements are then selected and implemented into Biomni-E1, the foundational environment that defines the biomedical action space for agentic interaction. Biomni-E1 includes 150 specialized biomedical tools, 105 software packages, and 59 databases. We then designed Biomni-A1, a general-purpose agent architecture capable of flexibly executing a broad spectrum of biomedical tasks by using tools and datasets provided by Biomni-E1. Given a user query, the agent first uses a retrieval system to identify the most relevant tools, databases, and software needed. It then applies LLM-based reasoning and domain expertise to generate a detailed, step-by-step plan. Each step is expressed through executable code, enabling precise and flexible compositions of biomedical actions—an essential feature given the domain's reliance on highly specialized tools and data resources. This integrated system allows Biomni to generate solutions for challenging, large-scale biomedical problems with efficiency but also to generalize to tasks across previously unseen areas of biomedical research.

Rigorous benchmarking demonstrates Biomni's strong performance across established biomedical quality and assurance benchmarks and

[1]Department of Computer Science, Stanford University School of Engineering, Stanford, CA, USA. [2]Phylo, Inc., South San Francisco, CA, USA. [3]Research and Early Development, Genentech, South San Francisco, CA, USA. [4]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [5]Cancer Biology Program, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [7]Arc Institute, Palo Alto, CA, USA. [8]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [9]Department of EECS, University of California, Berkeley, CA, USA. [10]Retro Bioscience, Redwood City, CA, USA. [11]Tamarind Bio, South San Francisco, CA, USA. [12]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [13]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [14]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA. [15]Department of Pathology, University of California San Francisco, San Francisco, CA, USA.
***Corresponding author. Email: kexinh@cs.stanford.edu (K.H.); jure@cs.stanford.edu (J.L.)** †These authors contributed equally to this work.

robust generalization performance in challenging, realistic scenarios never encountered during development. We highlight Biomni's practical capabilities through five case studies: (i) analyzing wearable sensor data; (ii) performing comprehensive bioinformatics analyses on massive raw datasets, such as scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data; (iii) designing laboratory protocols to assist wet-lab researchers; (iv) optimizing a protein sequence for better thermostability; and (v) orchestrating robotics wet-lab instruments. With Biomni, we introduce a scalable, general-purpose biomedical AI agent, pointing toward a future in which AI agents work alongside human researchers to accelerate biomedical discovery from basic research to translation.

Overview of Biomni

Curating a unified biomedical action space is challenging owing to its inherent complexity and vastness. We systematically address this problem by using an AI-driven approach (Fig. 1A). We leveraged the 25 subject categories defined by bioRxiv, selecting the 100 most recent publications per category. An action-discovery LLM agent processed each paper sequentially, extracting essential tasks, tools, databases, and software necessary to replicate or generate the described research. This comprehensive set of resources constitutes the essential actions required to perform a large set of biological research tasks.

We then curated Biomni-E1, an environment for a biomedical AI agent to perform a wide range of actions (Fig. 1B). Identified tools were rigorously verified by human experts, along with corresponding test cases. These tools (tables S1 to S18) were specifically chosen for their nontrivial nature, encompassing complex code, domain-specific know-how, or specialized AI models. Recognizing the inherent flexibility required by biological software, which cannot always be simplified into static functions, we constructed an execution environment pre-installed with 105 widely used biological software packages (tables S24 to S31), supporting Python, R, and CLI. For database integration, we categorized resources into two distinct groups. The first group consists of massive relational databases accessible through web application programming interfaces (APIs) (e.g., PDB, OpenTarget, and ClinVar) (tables S19 and S20). Rather than creating numerous individual retrieval tools, we implemented a unified function per database. Each function accepts natural language queries and internally uses an LLM to parse database schemas and generate executable queries dynamically. Databases without web interfaces were downloaded into a data lake and preprocessed locally into structured flat files for seamless integration with the agent, for a total of 59 databases in Biomni-E1 (tables S21 to S23).

To build a general-purpose agent capable of tackling diverse biomedical tasks, we require a specialized agentic architecture—one that avoids hardcoding workflows for each individual task. This led to the development of Biomni-A1, which incorporates several core innovations critical for operating across the biomedical research landscape. First, we introduce an LLM-based resource selection mechanism designed to navigate the complexity and specialization of the biomedical environment, dynamically retrieving a tailored subset of resources based on the user's goal (see fig. S1 for ablation experiment on the retrieval step). Second, recognizing that biomedical tasks often require rich procedural logic, Biomni-A1 uses code as a universal action interface—allowing it to compose and execute complex workflows involving loops, parallelization, and conditional logic. This approach also enables the agent to interleave calls to

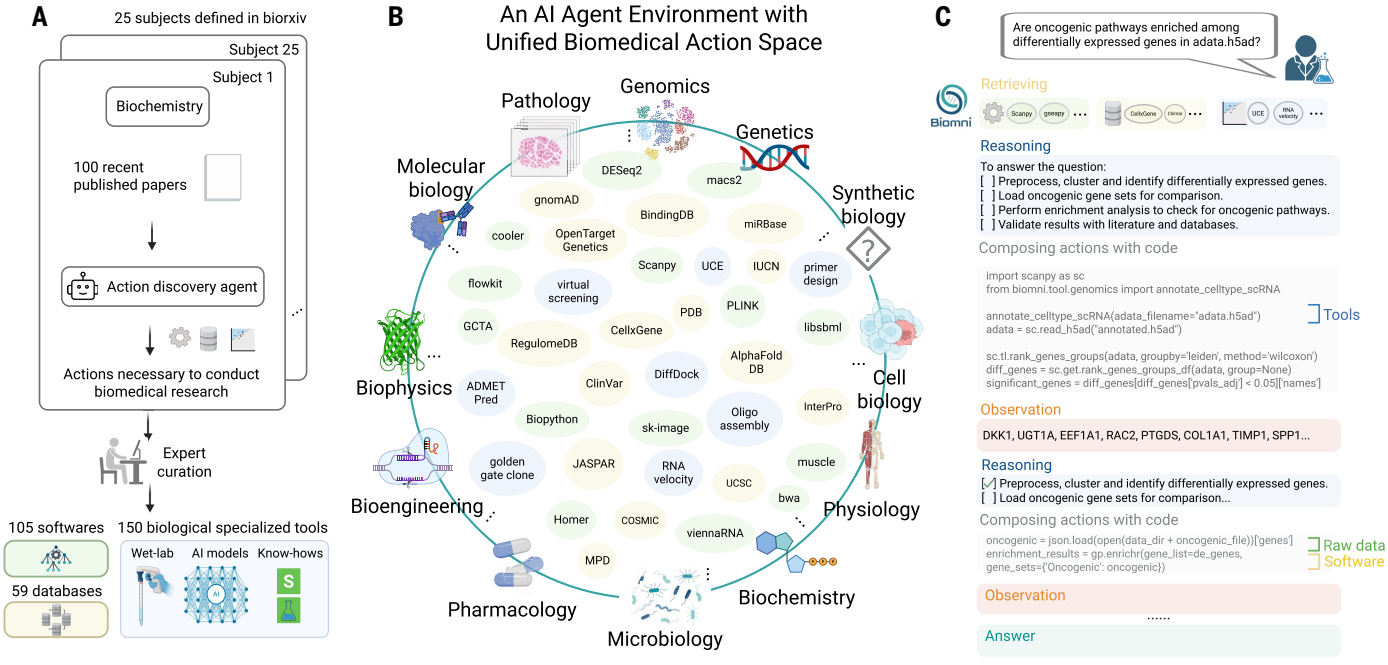


Fig. 1. Overview of the unified biomedical action space and agent environment in Biomni. (**A**) Workflow for systematically curating the unified biomedical action space. Actions necessary to conduct biomedical research were extracted from 2500 recent bioRxiv publications across 25 biomedical subfields using an AI-driven discovery agent. Extracted actions were rigorously validated and curated by human experts, resulting in the integration of 105 biomedical software packages, 150 specialized biological tools (including wet-lab protocols, AI-driven predictive models, and domain-specific know-how), and 59 comprehensive biomedical databases. (**B**) Illustration of the unified biomedical action space, spanning diverse biomedical subfields such as genetics, genomics, synthetic biology, cell biology, physiology, microbiology, pharmacology, bioengineering, biophysics, molecular biology, and pathology. Representative tools and databases integrated into Biomni's environment are shown, highlighting its general-purpose capabilities. (**C**) Example workflow demonstrating Biomni's reasoning and action composition process to autonomously answer a complex biological question. Biomni retrieves relevant tools based on the user's query, formulates a structured reasoning plan, and composes executable code to perform comprehensive bioinformatics analyses, iteratively refining its reasoning on the basis of observations until converging on a final, precise answer.

software, tools, databases, and raw data operations that do not conform to predefined function signatures, supporting flexible and dynamic integration of heterogeneous resources. Third, the agent adopts an adaptive planning strategy: It formulates an initial plan grounded in biomedical knowledge and iteratively refines it throughout execution, enabling responsive, context-aware behavior. Together, these innovations enable Biomni-A1 to generalize to previously unseen tasks and domains, dynamically composing actions and interfacing with software, data, and tools (Fig. 1C).

Biomni excels on general biomedical research benchmarks

To establish baseline competence across a broad range of biomedical research problems, we first evaluate Biomni on standardized benchmarks that span common biological tasks (Fig. 2). We first evaluate performance on Biomni-Eval1, which consists of 443 queries spanning 10 representative biomedical research tasks (Fig. 2A): CRISPR delivery, causal gene detection (gene-centric, ontology-based, and pathway-based), variant prioritization, database query, DNA sequence query, patient gene detection, rare disease diagnosis, and perturbation screen

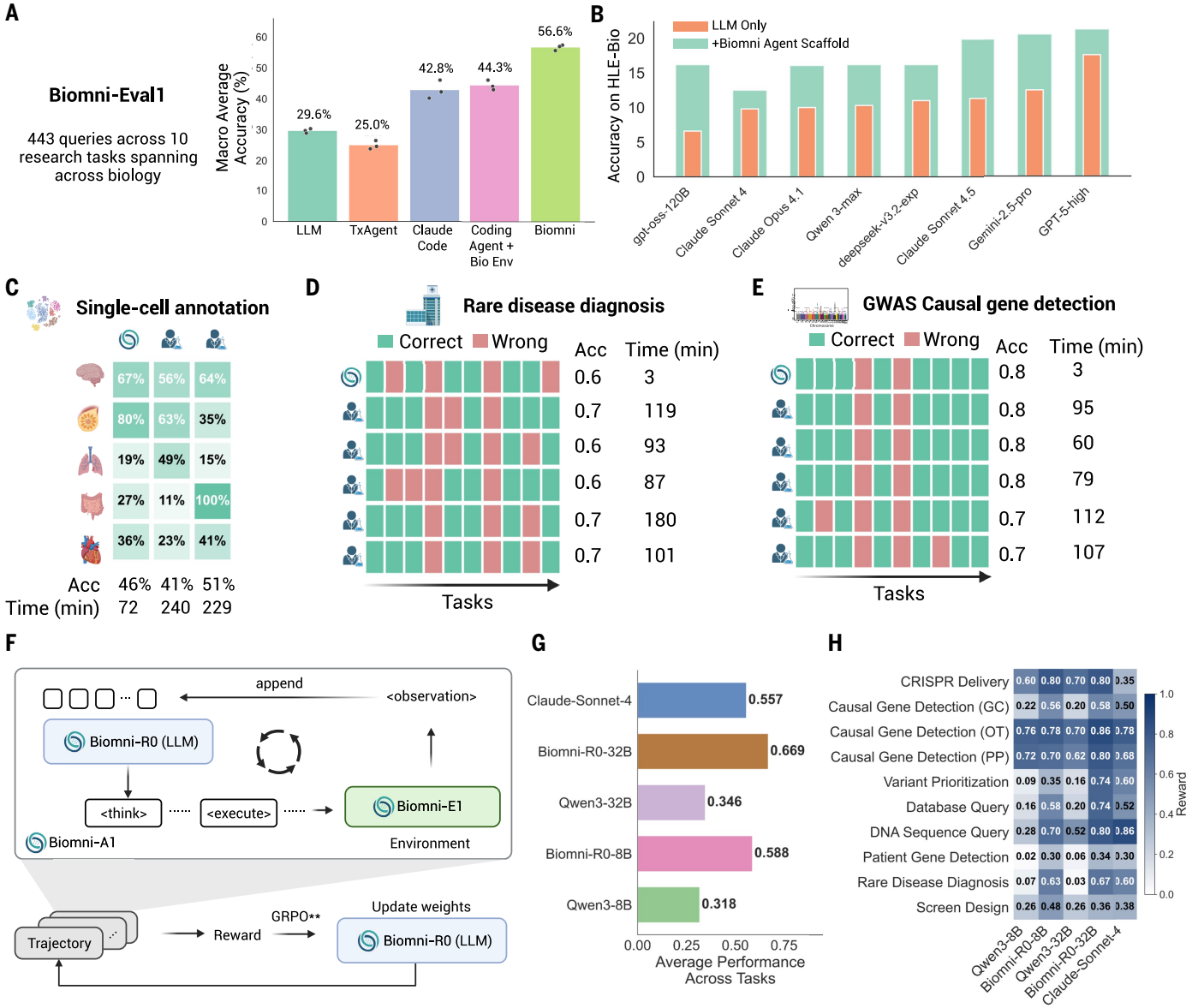


Fig. 2. Biomni performance across general benchmarks, expert-level tasks, and RL. (**A**) Performance on Biomni-Eval1, comprising 443 queries across 10 biomedical research tasks, including CRISPR delivery, causal gene detection, variant prioritization, database and DNA sequence queries, patient gene detection, rare disease diagnosis, and screen design. The data points are three independent runs, and we report the average accuracy across 443 queries. (**B**) Accuracy on Humanity's Last Exam (HLE-Bio) across multiple frontier LLM backbones, comparing LLM-only baselines with the Biomni-A1 agent scaffold and Biomni-E1 agent environment. (**C**) scRNA-seq annotation accuracy and execution time compared with those of human experts, showing expert-comparable accuracy with reduced analysis time. (**D**) Rare disease diagnosis accuracy and execution time, with Biomni matching expert accuracy while substantially reducing time to completion. (**E**) GWAS causal gene detection accuracy and execution time, demonstrating expert-level performance with markedly faster execution. (**F**) Overview of the Biomni agent architecture and RL training procedure. (**G**) Average task performance before and after RL, showing substantial gains for Biomni-R0 models at 8B and 32B scales. Biomni-R0 was fine-tuned from the Qwen3 open-weight model (*37*) and was first distilled from the frontier model Claude Sonnet 4 (i.e., teacher model) and then performed RL to further specialize to biomedical tasks (see materials and methods). It outperformed the teacher frontier model after tuning. (**H**) Task-level performance comparison across specialized biomedical tasks, illustrating consistent improvements from RL across task categories.

design. Across these tasks, Biomni achieves an average accuracy of 57%, outperforming a base LLM (Claude Sonnet 4.5) (30%), a therapeutics-specific agent called TxAgent (25%), a general coding agent called Claude Code (43%), and a bioinformatics software–based agent using ReAct equipped with the Biomni-E1 (44%).

We further evaluate general biomedical reasoning using Humanity's Last Exam (HLE-Bio) (Fig. 2B) to test robustness across unfamiliar domains without task-specific tuning. Across multiple frontier LLM backbones, the Biomni agent scaffold yields consistent improvements over LLM-only baselines, with absolute accuracy gains of 6 to 12 percentage points, which indicates that the observed improvements are not driven by a particular underlying model but by the Biomni-A1 agent scaffold and Biomni-E1 environment. Additional trajectory and error analyses are available in figs. S6 to S17.

Biomni performs biomedical research tasks at expert level while shortening the time

Having established strong performance on benchmarks, we next assessed whether these gains translate to realistic end-to-end biomedical research workflows by comparing with human experts. We evaluated Biomni on three representative tasks—scRNA-seq annotation, rare disease diagnosis, and genome-wide association study (GWAS) causal gene detection—with direct comparison with human expert performance (Fig. 2, C to E; see materials and methods and Data, code, and materials availability statement for prompts and responses). Experts were recruited on the basis of having at least 5 years of experience in the relevant task domains and included postdoctoral researchers and faculty members from academic institutions. The prompt was developed by the authors. In single-cell annotation (Fig. 2C), Biomni achieves 45.8% accuracy, in the range of the strong expert baseline (40.5 and 50.9%), and reduces the average analysis time from 235 min on average by experts to 72 min by agents per dataset. For rare disease diagnosis (Fig. 2D), Biomni attains 60% accuracy across cases, matching five expert performances (60 to 70%), and completes analyses in roughly 3 min compared with 116 min on average for human experts. In GWAS causal gene detection (Fig. 2E), Biomni reaches 80% accuracy across queries, again comparable to expert performance, and reduces execution time from 91 min on average by experts to 3 min by agents per locus. Across tasks, Biomni consistently matches expert accuracy while requiring substantially less time.

RL enables scalable improvement on specialized biomedical tasks

Despite strong generalization, performance on several specialized tasks remains below expert level, and qualitative analysis shows that models do not consistently exploit the full capabilities of the Biomni-E1 environment (figs. S6 to S9), motivating a RL approach to scalably improve task-specific performance. Figure 2, F to H, introduces Biomni-R0, a RL-trained open-source model that optimizes end-to-end task success through direct interaction with the Biomni-E1 environment. Using expert-annotated rewards, RL training substantially improves average task performance: Biomni-R0-8B increases from 0.32 to 0.59, and Biomni-R0-32B increases from 0.35 to 0.67 (Fig. 2G). Biomni-R0-8B exceeds the performance of larger closed-source models, such as Claude Sonnet 4 (0.56), and scaling to 32B yields an additional 10% absolute gain. Task-level results in Fig. 2H show consistent improvements across CRISPR screen design, causal gene detection, variant prioritization, database querying, and rare disease diagnosis, which indicates that RL provides a practical mechanism for systematically hill-climbing performance on specialized biomedical agent tasks.

Biomni performs population-scale autonomic and chronobiological analysis from raw wearable data

To evaluate Biomni's performance in real-world biomedical workflows, we invited scientists to apply it directly to their own research questions. Because biomedical research spans across a wide range of tasks, we select five use cases for illustrative purpose, with an additional four use cases available in figs. S2 to S5.

We first applied Biomni to two large, previously published wearable datasets with gold-standard human analyses. In one benchmark, we used the COVID-19 physiological response cohort from Alavi *et al.* (*29*), consisting of raw minute-resolution Fitbit heart rate and step data from 1027 participants, comprising more than 1.4 billion heart rate measurements and 37 million step records, with a mean monitoring duration of 170 days per individual. The dataset is highly heterogeneous, containing long-term longitudinal recordings across individuals with diverse baseline physiology and activity patterns (Fig. 3A).

Starting exclusively from raw heart rate and step data, Biomni autonomously generated and executed a complete end-to-end analysis pipeline (see Fig. 3B; materials and methods, section G; and the Data, code, and materials availability statement). The agent extracted six validated COVID-19–associated physiological biomarkers, including elevated resting heart rate, reduced circadian amplitude, decreased heart rate variability, elevated nocturnal heart rate, reduced daily steps, and blunted chronotropic response. It further reconstructed population-level 24-hour circadian structure, integrated biomarkers into a multidimensional risk score (0 to 6), and generated publication-quality visualizations (Fig. 3C). The resulting biomarkers, effect sizes, and correlations closely match those reported in the original study and subsequent literature (Fig. 3D), including a negative correlation between resting heart rate and circadian amplitude ($r = -0.34$) and a positive correlation between daily steps and heart rate variability ($r = +0.54$). These results indicate that Biomni can independently reproduce expert-level autonomic and chronobiological analyses directly from raw wearable data at the population scale.

Biomni automates complex multiomics analysis to decipher transcriptional regulation of skeletal lineages

To test whether Biomni could generalize to complex omics workflows, we used it to analyze a recently published multiomics dataset of the developing human skeleton (*30*). This dataset comprises 336,162 single-nucleus RNA-seq (snRNA-seq) and single-nucleus ATAC-seq (snATAC-seq) paired with spatial transcriptomics data collected from human embryos between 5 and 11 weeks postconception (Fig. 3E). Although the original study emphasized developmental trajectories and disease mechanisms, we were interested in exploring gene regulatory mechanisms across emerging skeletal cell types—a technically demanding task typically requiring extensive bioinformatics support.

We asked Biomni to investigate transcriptional regulation across skeletal lineages using a detailed instruction (see materials and methods, section H). Biomni autonomously planned and executed a 10-stage analysis pipeline: (i) loading and exploring all datasets; (ii) preparing RNA-seq data for analysis; (iii) configuring pySCENIC to retrieve motifs; (iv) running GRNBoost2 to infer gene regulatory networks (GRNs); (v) pruning networks using cisTarget; (vi) calculating regulon activity with AUCell; (vii) extracting accessibility data from ATAC-seq; (viii) filtering predicted targets using ATAC-seq accessibility; (ix) analyzing activity patterns across cell types, developmental stages, and anatomical regions; and (x) summarizing findings and preparing a report. This pipeline's output included transcription factor–target gene links and filter regulons based on motif enrichment and chromatin accessibility correlations (Fig. 3F).

The full run, completed in around 5 hours, handled real-time execution issues (e.g., variable name mismatches) by subsampling and debugging locally. Throughout the run, Biomni maintained all intermediate outputs—code, figures, and logs—organized in a reproducible notebook structure for validation and inspection (see Data, code, and materials availability statement). The agent summarized all the analysis and generated a report describing the analysis and key findings (see materials and methods, section H).

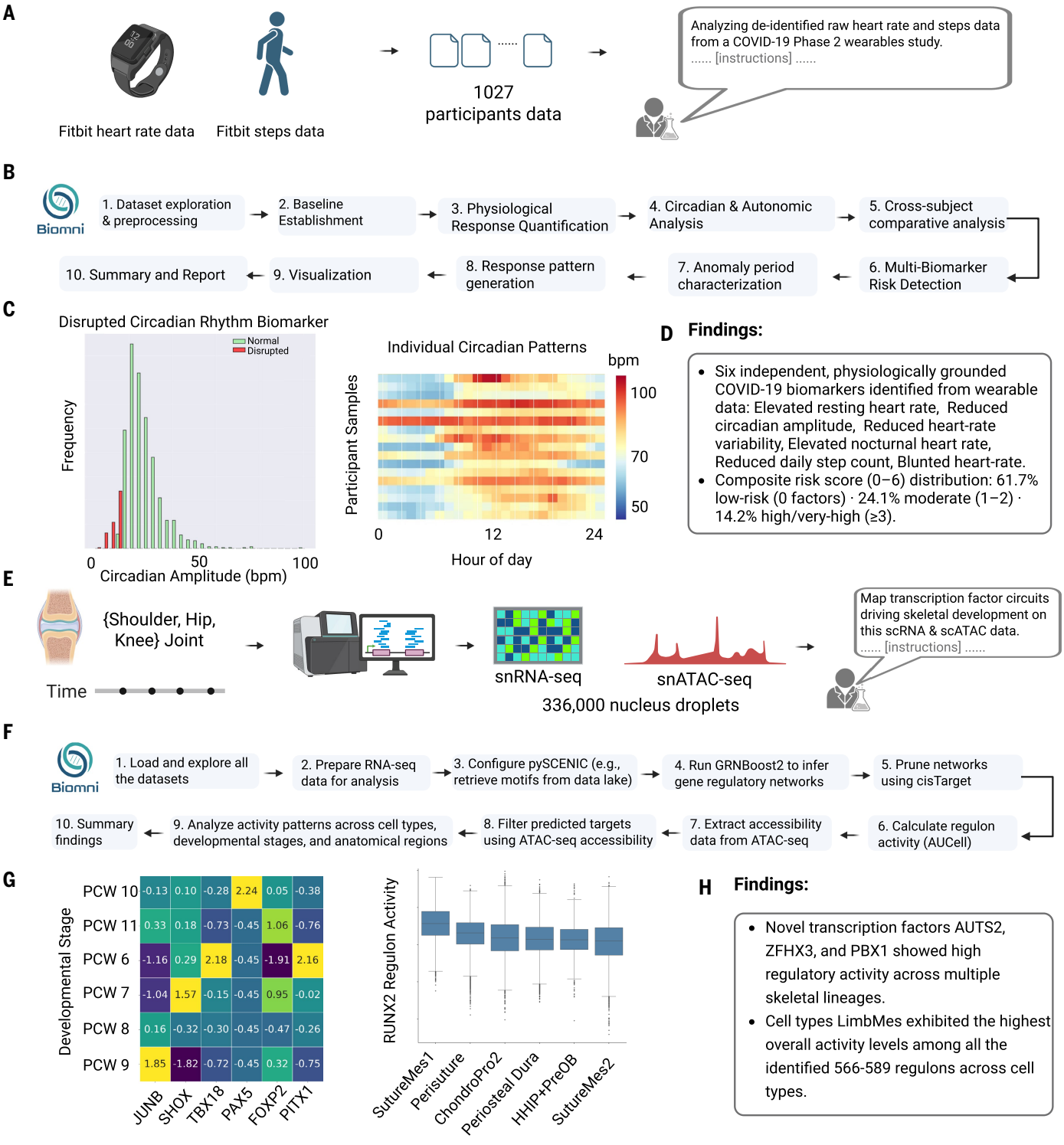


Fig. 3. Biomni autonomously executes complex multimodal biomedical analyses to generate hypotheses. (**A** to **D**) Overview of the COVID-19 phase 2 wearables study and analysis workflow. (A) High-frequency heart rate and step count data were collected using Fitbit devices from 1027 participants during the COVID-19 pandemic. Data were processed to establish individual physiological baselines and to detect illness-related deviations. (B) Ten-step analytical pipeline applied to the full cohort using Biomni software. (C) Key figures generated by Biomni, including disrupted circadian rhythm biomarkerd (left) and individual circadian patterns (right). (D) Biomni identified that key physiological correlations confirm interconnected mechanisms. (**E** to **H**) Biomni autonomously analyzed single-cell multiomics data from ~336,000 nucleus droplets, combining snRNA-seq and snATAC-seq across human embryonic joint development (shoulder, hip, and knee). (F) A detailed workflow diagram showing Biomni's 10-step analysis pipeline for GRNs with multiomics. (G) Two key figures generated from Biomni. (Left) Heatmap of regulator activity by developmental stage, with color intensity indicating activity levels. (Right) Boxplot of RUNX2 regulon activity by cell type, showing variation in expression across different cell populations. (H) Biomni identified additional transcription factors and cell type hypotheses for further experimental validation.

In its final GRN analysis (Fig. 3H), Biomni recapitulated known regulatory relationships between key osteogenic transcription factors, such as RUNX2 and HHIP, confirming how they are regulated by a shared set of antiosteogenic transcription factors, including TWIST1, LMX1B, and ALX4 (*30*). These findings align with a previous study's findings about the balanced regulation needed for proper bone formation and suture patency (*30*). Furthermore, Biomni nominated several putative regulators that have not been prominently highlighted in prior human skeletal GRN studies, including AUTS2, ZFHX3, and PBX1, which showed high inferred regulatory activity across multiple skeletal cell types. This signal is supported by complementary evidence from motif enrichment and transcription factor footprinting, concordant AUCell regulon activity, ATAC coaccessibility of predicted targets, and cross-region and cross-stage reproducibility, with explicit decision criteria, negative controls, and per–transcription factor evidence (Fig. 3, F to H, and Discussion). PBX1 is a well-established skeletal regulator (*31*), whereas ZFHX3 and AUTS2 have only limited or indirect skeletal reports [in mice (*32*) or zebrafish (*33*)]; their broad activity identified by Biomni suggests underappreciated roles across diverse skeletal lineages. Biomni reported that these regulators were particularly active in osteoblasts, preosteoblasts, and various chondrocyte populations, which suggests that they play important but previously unrecognized roles in the transcriptional control of skeletal cell fate determination during human skeletal development. Finally, Fig. 3, G and H, reveals how Biomni's visualizations effectively captured both temporal dynamics of regulator activity and cell type–specific variations in key regulons, such as RUNX2. This example demonstrates how Biomni enables researchers to autonomously perform complex multiomics analysis and rapidly generate testable hypotheses without specialized programming expertise.

Biomni designs wet-lab–validated experimental protocol for cloning

To evaluate Biomni's ability to support real-world experimental design, we focused on a core task in molecular biology: cloning. This process is central to countless workflows in research and biotechnology and requires complex reasoning, from designing high-fidelity primers to choosing the right assembly method and validating constructs. Whereas general-purpose LLMs have struggled to perform such tasks owing to limited domain knowledge and tool access (*34*), Biomni integrates LLM reasoning with dynamic tool execution, enabling expert-level performance in molecular biology tasks.

To rigorously evaluate this task, we first collaborated with an expert group of gene-editing researchers to design an open-ended cloning benchmark and expert user study (Fig. 4A). Our benchmark consisted of 10 realistic, representative cloning tasks covering Golden Gate, Gibson, Gateway, and restriction cloning—each with options including single-fragment versus pooled assembly. The benchmark also included essential validation steps, such as designing Sanger sequencing primers and analyzing restriction digests. We posed these tasks to four entities: an LLM, Biomni, a human trainee (Stanford biology graduate with previous experience in cloning, S.Z.), and a senior human expert (Stanford genetics postdoctoral researcher with >5 years of cloning experience, D.Y.). Each was asked to generate a complete, end-to-end protocol along with the final cloned plasmid map. A blinded expert reviewer assessed the outputs (see tables S32 and S33 for rubrics). Biomni produced protocols and designs that matched the human expert in accuracy and completeness as determined by manual examination of the outputs by an independent expert. Biomni often provided comparable levels of detail and anticipated the same edge cases. By contrast, the human trainee's submissions were frequently incomplete or suboptimal, reflecting the experience gap typical in early-stage researchers. Notably, Biomni completed all tasks autonomously in a fraction of the time taken by the expert.

To further validate Biomni in a real-world setting, we assigned it a practical cloning task—cloning a guide RNA targeting the human B2M gene into the lentiCRISPR v2 Blast construct (Fig. 4B). Biomni successfully executed the task through a comprehensive workflow (Fig. 4C). First, it analyzed the plasmid structure using annotation and pattern search tools to identify key features necessary for cloning. It then designed three Cas9 single guide RNAs (sgRNAs) targeting B2M using specialized knockout sgRNA design tools. For the cloning process, Biomni generated forward and reverse oligos with BsmBI overhangs to enable directional insertion of the sgRNA sequence. It produced detailed protocols (Fig. 4D) for oligo annealing, double-stranded DNA formation, and Golden Gate cloning into the target vector. Biomni also provided complete bacterial transformation instructions, including heat-shock steps and antibiotic selection. For quality control, it designed a U6 promoter sequencing primer to verify sgRNA insertion and simulated the Golden Gate assembly to produce the final plasmid map (see materials and methods, section I).

We followed Biomni's protocol exactly to perform the wet-lab experiment (Fig. 4E). Colonies appeared on the plate the next day, and two were cultured, miniprepped, and sequenced using the Biomni-designed primers—both showing perfect alignment. This case illustrates how scientists can rely on Biomni to autonomously design complex molecular biology experiments with accuracy comparable to that of human experts but in a fraction of the time.

Biomni orchestrates AI models for iterative molecular design

AI models are increasingly embedded in computational protein-engineering workflows, but deploying and combining them often requires substantial technical expertise in software configuration and graphics processing unit (GPU) infrastructure. Biomni addresses this barrier by integrating protein design models—including AlphaFold-2 (*2*) and ThermoMPNN (*35*)—as native tools that scientists can invoke and combine through natural language instructions. To illustrate this capability, we asked Biomni to identify candidate mutations predicted to improve the thermostability of a provided protein sequence. Biomni constructed and executed a multistep computational workflow (Fig. 5A) using AlphaFold-2 to predict the initial three-dimensional (3D) structure, ThermoMPNN to assess predicted thermostability, and iterative evaluation to propose and assess candidate variants. Over three optimization cycles, the workflow identified three candidate mutations with cumulative predicted thermostability improvement from ThermoMPNN (Fig. 5B and materials and methods, section O). The resulting mutations represent computational hypotheses that remain to be experimentally validated and are subject to limitations of the underlying prediction models. In this case study, the thermostability task illustrates how Biomni makes specialized AI models accessible within an integrated scientific workflow. By handling model deployment, software dependencies, and computational infrastructure, Biomni democratizes these tasks by enabling scientists to invoke and coordinate these models through natural language instructions without requiring specialized technical expertise.

Biomni bridges computational analysis and wet-lab execution through automated protocol generation

A persistent bottleneck in biological research is the disconnect between computational design and physical experimentation: Insights from in silico analysis must be painstakingly translated into laboratory protocols, often requiring days of manual coordination. To address this gap, we integrated Biomni with PyLabRobot (*36*), the open-source Python framework for laboratory automation, enabling researchers to generate and execute production-ready liquid handling protocols directly from natural language. We validated this capability on Hamilton STAR systems in production environments. Given a simple prompt describing a liquid transfer, Biomni generates complete, executable code with proper deck configuration, error handling, and resource cleanup (Fig. 5C). For complex workflows, we tested Biomni on a cell viability assay measuring dose response of eight compounds across 12-point serial dilutions in triplicate. From a hardware specification,

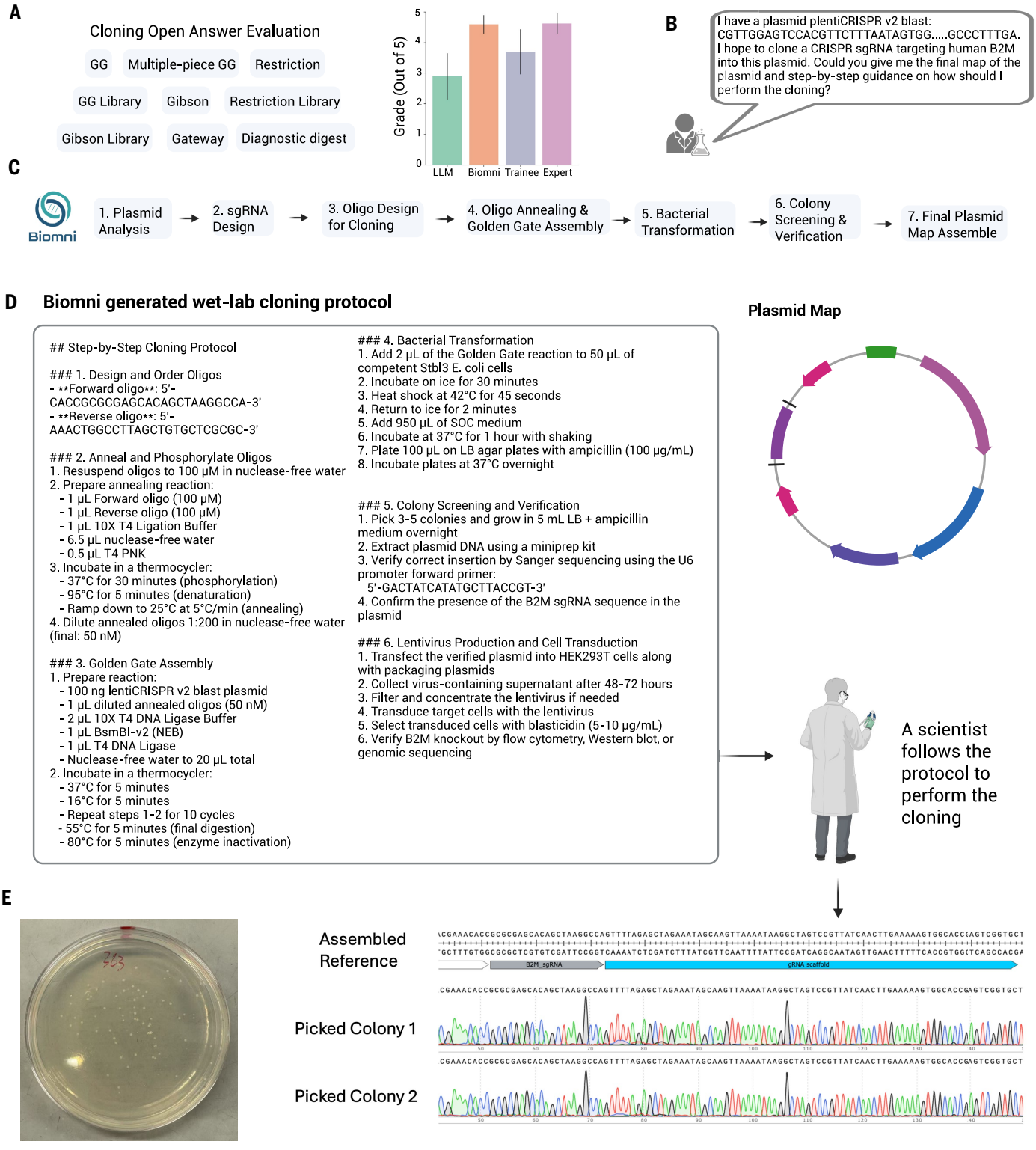


Fig. 4. Biomni designs wet-lab experimental protocol. (**A**) Open-ended cloning benchmark on 10 real cloning scenarios. We compared against base LLM, trainee-level human, and expert-level human scientists. We found that Biomni has similar accuracy to that of the expert-level scientist and higher accuracy than that of the trainee-level human, and it uses much less time. (**B**) Example of a user request to Biomni for cloning an sgRNA targeting the human B2M gene into the lentiCRISPR v2 Blast plasmid. (**C**) Biomni's automated stepwise workflow, including plasmid analysis, sgRNA design, oligo synthesis, Golden Gate assembly, bacterial transformation, colony screening, and final plasmid mapping. (**D**) Biomni-generated detailed cloning protocol with step-by-step instructions and comprehensive plasmid map, enabling laboratory scientists to execute the experiment autonomously. (**E**) Validation of Biomni's cloning protocol through successful colony growth on selection plates, followed by Sanger sequencing confirming perfect alignment of sgRNA insertion in picked colonies, demonstrating Biomni's robust capability for precise and reliable experimental design.

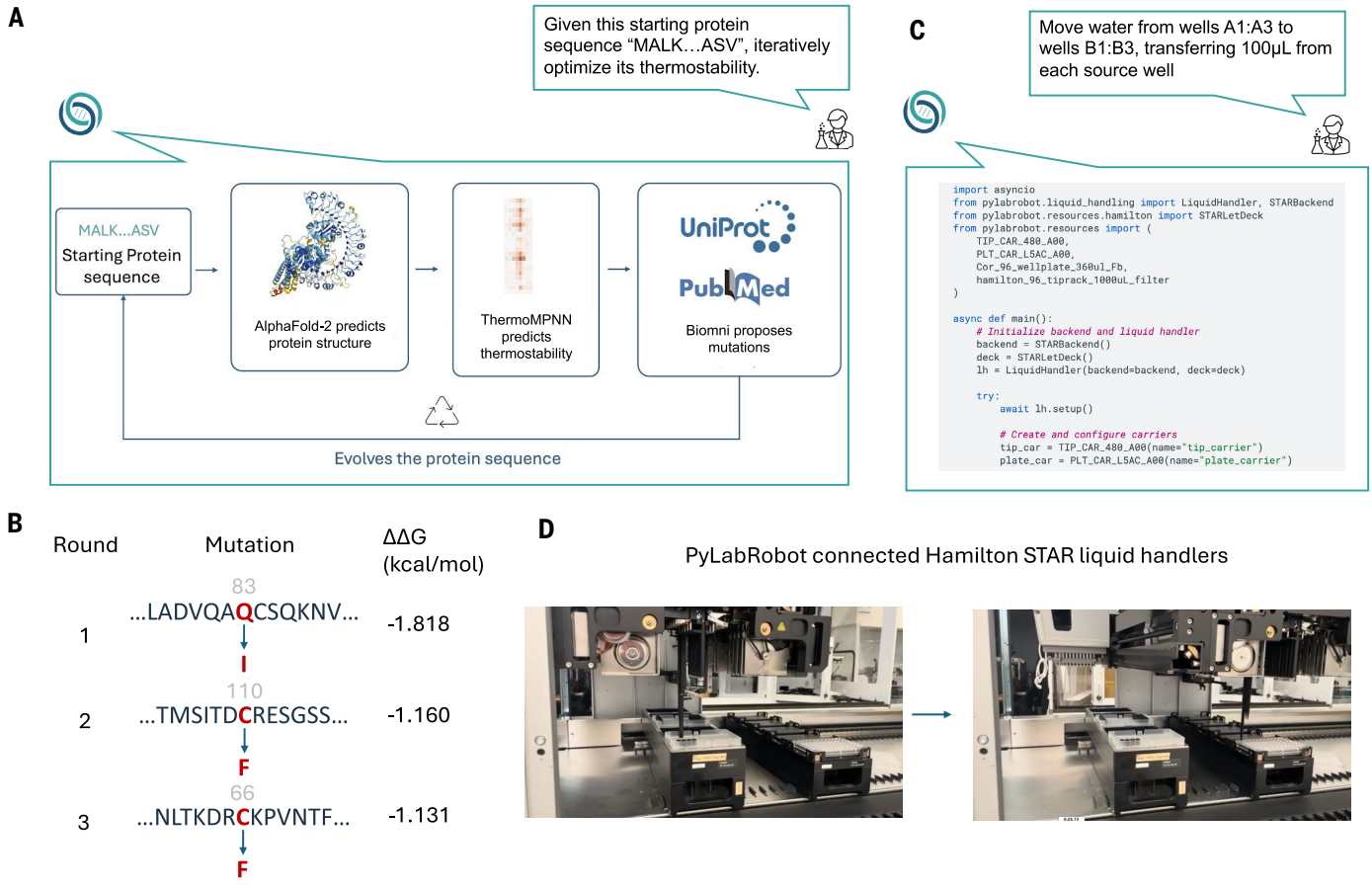


Fig. 5. AI-guided protein optimization and automated experimental execution. (**A**) Computational protein optimization pipeline. Biomni integrates structure prediction (AlphaFold-2) and thermostability estimation (ThermoMPNN) to iteratively optimize a starting sequence ("MALK...ASV") for improved predicted thermostability. (**B**) Example optimization trajectory. Over three iterative rounds, Biomni proposed three mutations, yielding cumulative predicted ΔΔG improvement of −4.108 kcal/mol. These are computational predictions for illustrative purposes and are yet to be experimentally validated. (**C**) AI-driven automation of experimental protocols. Biomni converts natural language descriptions of laboratory tasks (e.g., "Move water from wells A1-A3 to B1-B3, transferring 100 µl each") into fully executable PyLabRobot code. (**D**) The code was converted and validated on Hamilton STAR platforms for end-to-end digital-to-physical experiment execution.

Biomni produced a comprehensive protocol with intelligent tip selection, mathematically correct serial dilutions, and volume validation (see materials and methods, section N). This integration demonstrates how Biomni can serve as an interface between experimental intent and executable automation code, an important step toward closing the loop between dry-lab analysis and wet-lab execution.

Discussion

Biomni is a powerful platform for biomedical research, demonstrating robust generalization across diverse subfields and laying the groundwork for AI agents as integral collaborators in scientific discovery. Its zero-shot performance across complex tasks—including those in genetics, genomics, microbiology, immunology, pharmacology, and clinical medicine—underscores its potential to boost research productivity and accelerate discovery.

By automating complex, labor-intensive workflows, which normally require both expert knowledge and coding skills, Biomni enables researchers to redirect their efforts toward creative hypothesis generation, experimental innovation, and cross-disciplinary collaboration. In the context of target and drug discovery for biopharma, Biomni could potentially be applied to design workflows to prioritize targets, design perturbation screens, or repurpose drugs. In clinical settings, Biomni can support gene prioritization and rare disease diagnosis. For consumer health, Biomni supports the integration of wearable data and multiomics analyses for health monitoring and intervention.

Nonetheless, several limitations remain. Although Biomni's unified environment spans a wide range of biomedical tools and databases, the evaluated tasks represent only a subset of the field, and key domains remain unexplored. Additionally, in the action-discovery agent, our decision to prioritize the most recent literature makes the agent appear timely but risks overlooking foundational concepts and techniques that have faded from current discourse, despite their enduring relevance. Future versions should consider a larger range of publications when defining the environment. In addition, complex multistep analyses, such as the scRNA-scATAC multiomics workflow, often benefit from more structured prompts that explicitly define intermediate analytical steps. Although high-level prompts are sufficient for many tasks, workflows requiring substantial domain knowledge and implicit analytical conventions achieve greater robustness and reproducibility when key steps are specified, suggesting the limitation on generalizability of the current AI agent. Moreover, although Biomni approaches human-level performance in tasks such as database querying, sequence analysis, and molecular cloning, it still struggles in areas requiring nuanced clinical judgment, experimental reasoning, or deep biological thinking and synthesis. No system yet captures the full scope of human biomedical expertise. As reflected in our benchmarks, Biomni has not

achieved expert-level performance across all task categories. Performance remains uneven across the diverse landscape of biomedical tasks, and Biomni does not yet perform strongly across the board. Because the space of possible biomedical applications is vast and manuscript space is limited, we present five representative use cases spanning diverse biological domains to illustrate the range of Biomni's capabilities rather than exhaustively covering all potential use cases. We expect continued improvements as foundation models evolve and the agentic environment expands and as Biomni is used by human experts and trainees to facilitate or augment their work.

These limitations open promising directions for future development. Training biomedical reasoning agents with RL could enable continuous self-improvement in planning and execution. Integrating multimodal data—text, images, and structured inputs—may further deepen reasoning capabilities. Equipping Biomni to autonomously discover and incorporate new tools and databases, as well as to incorporate more historical methods (which may have high utility but can be easily forgotten by human users), would ensure adaptability and long-term relevance.

Finally, we recognize that increasingly capable AI systems in the life sciences raise important biosecurity considerations. Although Biomni is designed to accelerate beneficial biomedical research, agentic systems capable of literature synthesis, protocol generation, and automated analysis could also lower barriers to the misuse of biological knowledge. Responsible development therefore requires careful evaluation of both risks and benefits. In this work, we adopt several principles to mitigate potential concerns: focusing the system on widely used academic datasets and tools, emphasizing transparency through open-source release, and aligning our development practices with emerging community discussions and policy frameworks on AI-biosecurity risks. We believe that openness, rigorous evaluation, and engagement with the broader biosecurity and policy community are essential to ensure that scientific AI agents are developed in ways that maximize societal benefit while minimizing potential harms.

Looking ahead, Biomni and its successors could become foundational infrastructure in an AI-powered biomedical ecosystem, working seamlessly with human experts to unlock insights into health and disease. This hybrid partnership may reshape biomedical research—automating hypothesis generation, scaling discovery pipelines, and enabling medical innovation to advance faster than before.

Materials and methods are available in the supplementary materials.

REFERENCES AND NOTES

1. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013). doi: 10.1126/science.1231143; pmid: 23287718
2. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
3. C. H. van Dyck *et al.*, Lecanemab in early Alzheimer's disease. *N. Engl. J. Med.* **388**, 9–21 (2023). doi: 10.1056/NEJMoa2212948; pmid: 36449413
4. C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, G. Kroemer, Hallmarks of aging: An expanding universe. *Cell* **186**, 243–278 (2023). doi: 10.1016/j.cell.2022.11.001; pmid: 36599349
5. R. Botvinik-Nezer *et al.*, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020). doi: 10.1038/s41586-020-2314-9; pmid: 32483374
6. I. Thiele, B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010). doi: 10.1038/nprot.2009.203; pmid: 20057383
7. E. Gibney, R. Van Noorden, Scientists losing data at a rapid rate. *Nature* 10.1038/nature.2013.14416 (2013). doi: 10.1038/nature.2013.14416
8. T. Ridnik, D. Kredo, I. Friedman, Code Generation with AlphaCodium: From Prompt Engineering to Flow Engineering. arXiv:2401.08500 [cs.LG] (2024).
9. J. Cui *et al.*, Chatlaw: A Multi-Agent Legal Assistant based on a Role-Aligned Mixture-of-Experts Architecture. arXiv:2306.16092 [cs.CL] (2023).
10. G. Tom *et al.*, Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024). doi: 10.1021/acs.chemrev.4c00055; pmid: 39137296
11. C. Peng *et al.*, A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **6**, 210 (2023). doi: 10.1038/s41746-023-00958-w; pmid: 37973919
12. H. Wang *et al.*, Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023). doi: 10.1038/s41586-023-06221-2; pmid: 37532811
13. Y. Qu *et al.*, CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments. bioRxiv 2024.04.25.591003 [Preprint] (2024). doi: 10.1101/2024.04.25.591003
14. K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. bioRxiv 2024.11.11.623004 [Preprint] (2024). doi: 10.1101/2024.11.11.623004
15. Y. Roohani *et al.*, BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments. arXiv:2405.17631 [cs.AI] (2025).
16. E. Wang *et al.*, TxGemma: Efficient and Agentic LLMs for Therapeutics. arXiv:2504.06196 [cs.AI] (2025).
17. Y. Xiao *et al.*, CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis. arXiv:2407.09811 [cs.AI] (2024).
18. J. Gottweis *et al.*, Accelerating scientific discovery with Co-Scientist. *Nature* 10.1038/s41586-026-10644-y (2026). doi: 10.1038/s41586-026-10644-y
19. M. Hu *et al.*, Evaluation of large language models for discovery of gene set function. *Nat. Methods* **22**, 82–91 (2025). doi: 10.1038/s41592-024-02525-x; pmid: 39609565
20. D. Ferber *et al.*, Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat. Cancer* **6**, 1337–1349 (2025). doi: 10.1038/s43018-025-00991-6; pmid: 40481323
21. N. D. Youngblut *et al.*, scBaseCamp: An AI agent-curated, uniformly processed, and continually expanding single cell data repository. bioRxiv 2025.02.27.640494 [Preprint] (2025). doi: 10.1101/2025.02.27.640494
22. J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022). doi: 10.1038/s41591-022-01981-2; pmid: 36109635
23. H. Wang *et al.*, SpatialAgent: An autonomous AI agent for spatial biology. bioRxiv 2025.04.03.646459 [Preprint] (2025). doi: 10.1101/2025.04.03.646459
24. A. Madaan *et al.*, "Self-Refine: Iterative Refinement with Self-Feedback" in *Advances in Neural Information Processing Systems*, vol. 36, A. Oh *et al.*, Eds. (Curran Associates, Inc., 2023), pp. 46534–46594.
25. S. S. Sahoo *et al.*, Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *J. Am. Med. Inform. Assoc.* **31**, 2114–2124 (2024). doi: 10.1093/jamia/ocae074; pmid: 38657567
26. J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" in *Advances in Neural Information Processing Systems*, vol. 35, S. Koyejo *et al.*, Eds. (Curran Associates, Inc., 2022), pp. 24824–24837.
27. S. Yao *et al.*, "React: Synergizing reasoning and acting in language models" in *11th International Conference on Learning Representations (ICLR)* (ICLR, 2023).
28. DeepSeek-AI *et al.*, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] (2025).
29. A. Alavi *et al.*, Real-time alerting system for COVID-19 and other stress events using wearable data. *Nat. Med.* **28**, 175–184 (2022). doi: 10.1038/s41591-021-01593-2; pmid: 34845389
30. K. To *et al.*, A multi-omic atlas of human embryonic skeletal development. *Nature* **635**, 657–667 (2024). doi: 10.1038/s41586-024-08189-z; pmid: 39567793
31. J. A. Gordon *et al.*, Pbx1 represses osteoblastogenesis by blocking Hoxa10-mediated recruitment of chromatin remodeling factors. *Mol. Cell. Biol.* **30**, 3531–3541 (2010). doi: 10.1128/MCB.00889-09; pmid: 20439491
32. G. A. Gomez *et al.*, Evaluation of potential roles of *zinc finger homeobox 3* (*Zfhx3*) expressed in chondrocytes and osteoblasts on skeletal growth in mice. *Calcif. Tissue Int.* **115**, 445–454 (2024). doi: 10.1007/s00223-024-01265-6; pmid: 39085428
33. Z. Geng, Y. T. Tai, Q. Wang, Z. Gao, AUTS2 disruption causes neuronal differentiation defects in human cerebral organoids through hyperactivation of the WNT/β-catenin pathway. *Sci. Rep.* **14**, 19522 (2024). doi: 10.1038/s41598-024-69912-4; pmid: 39174599
34. J. M. Laurent *et al.*, LAB-Bench: Measuring Capabilities of Language Models for Biology Research. arXiv:2407.10362 [cs.AI] (2024).
35. H. Dieckhaus, M. Brocidiacono, N. Z. Randolph, B. Kuhlman, Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2314853121 (2024). doi: 10.1073/pnas.2314853121; pmid: 38285937
36. R. P. Wierenga, S. Golas, W. Ho, C. Coley, K. M. Esvelt, PyLabRobot: An open-source, hardware-agnostic interface for liquid-handling robots and accessories. *Device* **1**, 100111 (2023). doi: 10.1016/j.device.2023.100111; pmid: 37502883
37. A. Yang *et al.*, Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] (2025).
38. K. Huang, Biomni Manuscript Data [Data set], Zenodo (2026). doi: 10.5281/zenodo.21865829
39. K. Huang, Biomni - Publication Version, Zenodo (2026). doi: 10.5281/zenodo.20100497

ACKNOWLEDGMENTS.

We thank E. Alsentzer, A. Lee, members of J. Leskovec's laboratory, and members of E. Ashley's laboratory for providing helpful feedback. **Funding:** K.H. and J.L. acknowledge the support of the National Science Foundation under nos. CCF-1918940 (Expeditions), DMS-2327709 (IHBEM), and IIS-2403318 (III); the National Institutes of Health under no. 1U24NS146314-01; the Gates Foundation; the Stanford Data Applications Initiative; the Wu Tsai Neurosciences Institute; the Stanford Institute for Human-Centered AI; the Allen Institute; Genentech; SAP; and SCBX. K.H. acknowledges the support of a Stanford Bio-X fellowship. Research reported in

this publication was supported in part by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (NIH) under award no. U01NS134358. L.C. acknowledges support from the NIH under grants R01GM141627, R01AG091819, and R35HG011316; the Donald and Delia Baxter Foundation; and the Weintz Family Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Figures were created with BioRender.com. **Author contributions:** K.H., Y.R., and J.L. conceived the study. K.H. and J.L. supervised the project. K.H. designed and developed the framework. K.H., S.Z., H.W., Y.Q., and Y.L. implemented tools and databases. K.H. designed and implemented the generalist agent architecture. K.H. and R.L. designed the action-discovery agent. S.Z. performed benchmarks on Q&A tasks. K.H., H.W., and Y.L. collected and implemented the benchmarks on realistic tasks. X.Z. provided advice on the microbiome benchmark. H.W., J.Zho., P.H., and K.H. performed the multiomics integration case study. Y.L. and K.H. performed the wearable data analysis case study. Y.Q., J.Zha., D.Y., S.Z., Y.L., and K.H. performed the wet-lab case study. K.H., S.M., J.N.C., M.T.W., and J.A.B. performed the rare disease diagnosis case study. R.L. performed the qualitative trace analysis. R.L., L.Q., and G.L. provided support for software. K.H., S.Z., H.W., Y.Q., A.R., and Y.L. wrote the draft paper. All authors discussed the results and contributed to the final manuscript. **Competing interests:** K.H., S.Z., Y.Q., L.C., and J.L. have equity in Phylo, Inc. A.R. is an officer of Genentech and Roche and a member of the corporate executive committee of Roche; the Genentech board of directors; and the boards of directors of the Broad Institute, the Allen Institute, and the Human Cell Atlas Inc. H.W. is an employee of Genentech. L.C. is an adviser or has equity interest in Acrobat Genomics, AutoBio, Arbor Biotechnologies, RococoBio, and Rootpath Genomics. M.P.S. received support from the Anu and BV Jagadeesh Foundations. All other authors declare no competing interests. **Data, code, and materials availability:** All data used in Biomni are publicly available and can be automatically downloaded using the open-sourced package at https://github.com/snap-stanford/biomni. The raw data of all plots in the main experiments are available at Zenodo (*38*). Biomni is open-sourced at https://github.com/snap-stanford/biomni. The code repository version used to generate the results in this manuscript is available at Zenodo (*39*). A user-friendly interface is available at https://biomni.phylo.bio. The Biomni-R0 model weights are available at https://huggingface.co/biomni/Biomni-R0-32B-Preview. All materials used in this study are described in the materials and methods, and requests for materials can be made to the corresponding authors. AI tools were used for manuscript editing (ChatGPT) and to assist with code generation (Claude Code). All AI-assisted manuscript text and code were reviewed and verified by the authors.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adz4351
Materials and Methods; Figs. S1 to S17; Tables S1 to S33; References (*40*–*196*); MDAR Reproducibility Checklist

Submitted 2 June 2025; resubmitted 10 March 2026; accepted 10 June 2026; published online 9 July 2026

10.1126/science.adz4351

Autonomous biomedical research with an artificial intelligence agent

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, Junze Zhang, Di Yin, Rick Wierenga, Deniz Kavi, Sherry Liu, Tianwei She, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec

Science **393** (6813), eadz4351. DOI: 10.1126/science.adz4351

Editor's summary

Biology research requires complex planning and adaptation depending on the specific project, but many of the specific tasks are tedious and rote. Some elements of research design, particular coding and protocol development, may be amenable to automation. Huang *et al*. developed a general-purpose artificial intelligence model to support basic biomedical research through information retrieval and generation of custom protocols, data analysis methods, and code. An agentic framework supports layered planning and decision making, and integration with specialized tools and databases permits this system to handle open-ended tasks and use established methods for customized tasks. The authors evaluated their model on a range of queries and demonstrated integration with laboratory automation. — Michael A. Funk