1. PROGRAM OVERVIEW

The Intelligent Generator of Research (IGoR) program aims to eliminate longstanding inefficiencies in research and accelerate the development of effective therapies for complex diseases by creating an AI-enabled interoperable research ecosystem. IGoR would develop: 1) mechanistic disease models that encode causal biological relationships across scales, 2) an AI orchestration layer that identifies knowledge gaps and designs optimal experiments, 3) a layered protocol architecture that enables any qualified laboratory to execute the same experiment reproducibly, and 4) a distributed marketplace of validated laboratories that execute standardized protocols and return gold-standard data. Together, these components form a cycle of hypothesis generation, experimentation, and model refinement that enables researchers to create validated knowledge at least 10x more rapidly than conventional approaches. Ultimately, IGoR will empower researchers at every level to pursue bold, unconventional research directions that are currently too slow, too complex, or too resource-intensive.

2. BACKGROUND

Slow, at times irreproducible biomedical research delays effective treatments for disease and reduces confidence in science. Based on some estimates, more than 70% of researchers have failed to reproduce another scientist's experiments[i], and up to 89% of preclinical work cannot be fully reproduced[ii,iii]. At the same time, diseases involving multifactorial causes, poorly understood mechanisms, and interactions across molecular, cellular, and tissue scales remain inadequately treated. The research ecosystem cannot integrate knowledge quickly or reliably enough to identify effective interventions for complex diseases. The successful direction of GLP-1 receptor agonists towards new indications illustrates the cost: decades of serendipitous discovery were required to finally connect disparate mechanistic dots that, with better tools, might have been quickly identified.

Research today often follows a centuries-old pattern: individual labs generate hypotheses constrained by local expertise and instrumentation, run experiments described in insufficient detail, and publish selected results that are difficult to reproduce. Consequently, mechanistic knowledge fragments across subfields, each with its own preferred approach. Representations of isolated pathways abound, while integrative, multiscale models that encode causal, time-dependent biological relationships are rare. Recent efforts to address these problems using AI have demonstrated promise but remain incomplete. Frontier AI models can uncover hidden biological interactions, but experimental confirmation through traditional approaches often takes many months if not readily automated.

Greater synthesis could be achieved if diverse confirming experiments could be readily requested; however, there is currently no widely adopted standard for specifying experiments precisely enough to transfer them reliably across laboratories. Cloud laboratories have made significant advances but have not fully resolved this bottleneck because their proprietary environments rely heavily on low-level languages, and protocol transfer remains labor-intensive. Conversely, AI-driven autonomous laboratories have achieved speed, but only for a narrow set of methods and limited goals with straightforward

objective functions. An ecosystem is needed that enables a researcher to request a broad array of experiments outside of their own laboratory with minimal effort.

Recent demonstrations of large language models (LLMs) connected to automated laboratories (e.g., for protein engineering or reaction optimization) have shown that AI can accelerate narrowly scoped experimental loops. However, these systems typically operate on single-objective, single-modality problems with well-defined fitness functions. They optimize within a known space rather than navigating the open-ended, multi-scale uncertainty that characterizes complex disease research. No current platform combines the high experimental variety, mechanistic reasoning, and cross-laboratory interoperability needed for complex disease research with the speed of execution required to substantially accelerate discovery.

IGoR is not simply an LLM connected to a laboratory. It is a distributed research ecosystem in which mechanistic models define what to learn, an orchestration layer determines how to learn it across diverse experimental modalities, and a validated marketplace ensures that the resulting data are trustworthy and interoperable. IGoR aims to create an ecosystem that will comprehensively support the researcher with AI-driven knowledge integration, rapid generation of standardized experimental procedures, and quick access to a distributed network of laboratories. It will power a continuous cycle of well-grounded hypothesis generation, automated experiment execution, and model refinement. IGoR will empower researchers at every level — regardless of local instrumentation or expertise — to efficiently and effectively pursue increasingly complex, resource-intensive research.

3. PROGRAM DESCRIPTION

IGoR seeks to accelerate therapeutic development by creating an advanced platform that amplifies human scientific judgment. Every component of IGoR should make researchers more capable, not less necessary. The system succeeds when a scientist can pursue research directions that were previously too complex, too slow, or too resource-intensive — not when the scientist is removed from the loop.

ARPA-H is soliciting proposals to build a system that integrates four major components:

1. Modular, mechanistic, multiscale models of complex diseases — digital twins that encode causal, time-dependent biological relationships and serve as the shared memory of the research ecosystem
2. An AI orchestration layer that interrogates these models and the scientific literature to identify knowledge gaps, generates testable hypotheses, designs optimal experiments, and explains its reasoning to human researchers
3. A layered protocol architecture that separates experimental intent from instrument-level implementation, enabling any qualified laboratory to execute the same experiment reproducibly
4. A distributed network of validated laboratories that execute standardized protocols and return high-quality data and metadata for automatic ingestion into the disease models

These components comprise a continuous research cycle: First, a human researcher uses a rich, AI-enabled interface to interrogate mechanistic models and laboratory capabilities, generating proposed experiments. Next, a layered protocol architecture translates experiments into interoperable protocols and offers them to a marketplace of laboratories. These laboratories then rapidly execute these experiments and produce *model-ready* data. Finally, the mechanistic models are updated with the new results. The cycle time should be limited primarily by the physical time required to run experiments. Human-driven hypothesis generation, protocol design, and laboratory selection will occur substantially faster than they do today.

A central design principle is high cohesion and low coupling: each component has a clearly defined responsibility, and the interfaces — what information flows between them, in what format, and at what cadence — are as important as their internal capabilities. Proposers must describe their initial approach to defining and maintaining these interfaces, recognizing that precise boundaries will evolve during the program.

NATIONAL HEALTH IMPACT

Complex diseases — such as neurodegenerative conditions, chronic autoimmune disorders, complex pain syndromes, and emerging infectious diseases — impose enormous burdens on patients and health systems. Current research approaches are too slow and fragmented to effectively address these areas. Knowledge generated in one subfield takes years to reach researchers in adjacent fields, and promising mechanistic insights often cannot be experimentally validated because the required capabilities do not exist at an investigator's institution. No single university, company, or funding agency can build the shared research infrastructure needed to change this.

IGoR will address these gaps by creating an infrastructure that makes the full breadth of biomedical knowledge and experimental capability accessible to any qualified researcher. It will ensure that new findings are immediately integrated into shared mechanistic models and will dramatically reduce the time from hypothesis to validated insight. If successful, IGoR will enable earlier identification of therapeutic targets, faster characterization of disease mechanisms, and more efficient allocation of research resources — particularly for diseases where current treatments have remained largely unchanged for decades and where growing threats to Americans' health receive relatively low funding.

Teams will propose their own complex area of human disease involving large knowledge gaps, multifactorial causes, or poorly understood mechanisms of action for existing treatments. The program is not seeking to solve well-bounded problems, such as creating a better binder to a known drug target. Rather, IGoR aims to discover new, non-obvious targets; leverage models that reason about how interventions behave in broader biological contexts; and account for dynamic aspects of biology, such as circadian rhythms, developmental timing, or disease progression. Proposers must justify their choice of disease focus, demonstrating that the disease area is:

- Tractable through mechanistic or quantitative modeling
- Characterized by multifactorial causes, time-dependent relationships, or multi-scale interactions as described above
- Amenable to meaningful progress within the program timeframe

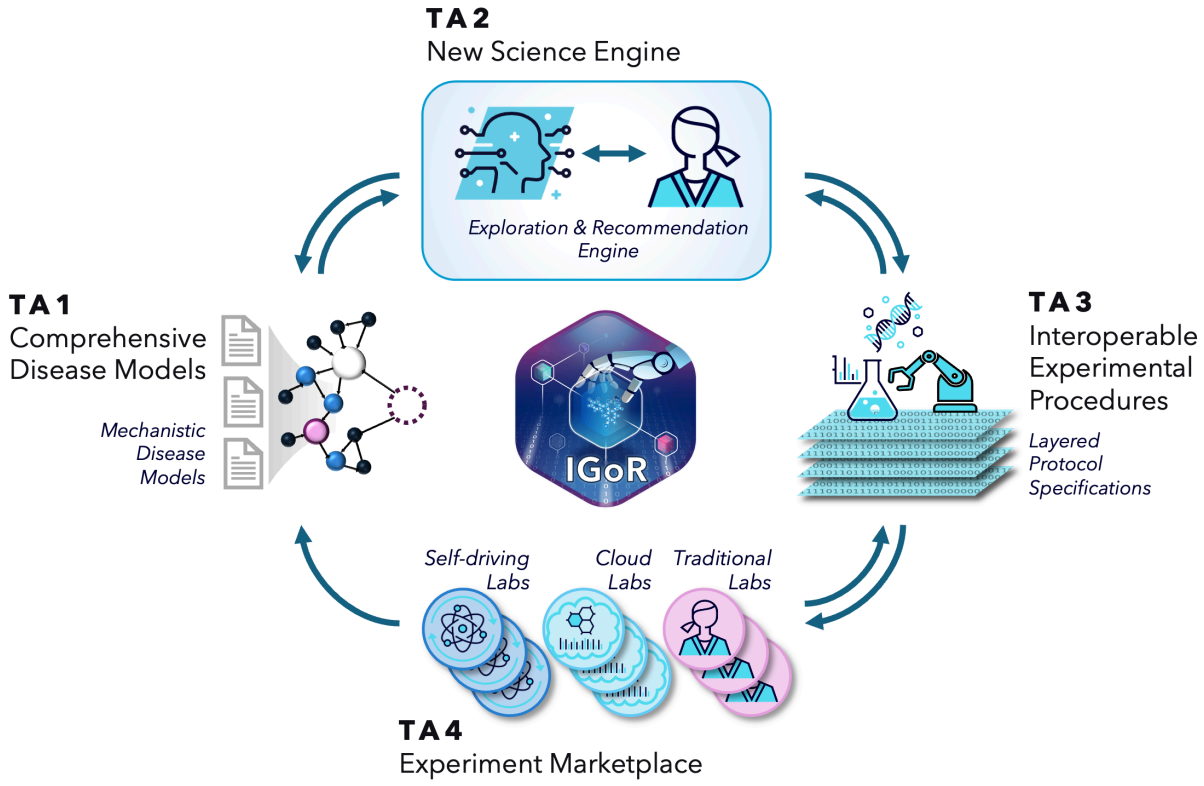
By Phase III, teams must extend their approach to a related disease area.

Experiments in cell/tissue culture and lower-level invertebrate animals (e.g., insects, nematodes) are in scope. Limited use of vertebrate and higher-level invertebrate animal experiments is allowed for testing strong predictions of a working system; however, they must be directed *solely* by the human researcher. Human clinical trials are out of scope.

4. IGoR TECHNICAL AREAS

4.1 Technical Area Definitions

The four components described above will be implemented through four corresponding Technical Areas (TAs). The interfaces between TAs are as important as the internal capabilities of each TA. The boundaries described below reflect ARPA-H's current view of the most natural decomposition, but they are not rigid. Proposers may place specific capabilities in a different TA than described here, provided they justify the choice and demonstrate that the overall closed-loop system remains coherent.



4.1.1 TA1: Comprehensive Disease Models

Understanding a complex disease requires assembling knowledge from dozens of subfields into a coherent, testable picture. Today, this integration happens informally: each laboratory maintains its own mental model shaped by its own data and disciplinary lens, and critical cross-field connections go unnoticed. TA1 will replace these fragmented, implicit models

with shared, computable representations that any researcher or AI system can query, update, and build upon.

TA1 will develop digital twins that encode causal biological relationships across time and length scales. These models will integrate literature, experimental data, and expert knowledge from many subfields into a single representation. The models will empower researchers to interpret new data and formulate hypotheses in the context of everything that is known about a disease, rather than through the lens of a narrow subfield.

TA1 models are the shared memory of the IGoR system. They are updated when new experimental data are returned from TA4, ensuring that learning is grounded. TA1 models must be:

- Modular: Composed of interoperable sub-models. Existing standards should be leveraged and extended where needed.
- Mechanistic: The theory or mechanism of action is known for each sub-model (e.g., protein–protein interactions, signaling cascades, metabolic pathways).
- Multiscale: Capable of predicting tissue- or organ-scale effects from changes at the molecular level.
- Verifiable: For each sub-model, the mechanism of action can be experimentally validated. The model must be structured so that specific predictions can be tested.

Objective 1: Disease Model Architecture. Teams must establish the representational framework for their disease models, including the languages, schemas, and ontologies used to encode biological knowledge. The architecture must support automated ingestion of new experimental data from TA4. Proposers should describe how their architecture will accommodate the integration of diverse data types, including omics data, imaging data, functional assay results, and literature-derived knowledge, and how it will handle uncertainty, conflicting evidence, and data from heterogeneous sources. Proposers should also describe how the architecture will leverage existing biological databases, ontologies, and model repositories.

Objective 2: Knowledge Gap Identification. The proposed system must not only represent what is known, but also identify what is *not* known — specifically, the knowledge gaps whose resolution would most advance understanding of the disease. This capability sits at the interface between TA1 and TA2: TA1 models encode the current state of knowledge and can identify structural gaps (e.g., missing edges in a causal graph, poorly constrained parameters, conflicting sub-models), while TA2 uses these gaps to generate hypotheses and propose experiments. Proposers should describe where they place this capability and justify their choice. ARPA-H recognizes that the optimal boundary between TA1 and TA2 for knowledge gap identification is an open question and expects that it will be refined during the program, particularly through the interface definition process described in Section 5.5.

Out of scope for TA1:

- Development of wholly new foundation models, such as training a new LLM from scratch. However, fine-tuning existing models or developing new specialized models for disease representation and gap identification is in scope.

- Models that are purely correlational without mechanistic grounding.
- Models of non-human diseases, except insofar as they inform human disease biology.

4.1.2 TA2: New Science Engine

TA2 will develop an orchestration layer, a "New Science Engine," that plans, verifies, and explains scientific workflows. This is the reasoning core of the IGoR system: it leverages frontier models, synthesizes literature and mechanistic models from TA1 to find knowledge gaps, surfaces conceptually adjacent knowledge, and proposes optimal experiments. The proposed experiments would be unconstrained by any single researcher's prior knowledge or local resources.

A key design requirement for TA2 is flexibility and evolvability. IGoR will not prescribe specific AI models. The program anticipates that AI model development will progress rapidly during and beyond the program's lifetime, and performers must design TA2 so that future users can incorporate the latest models without re-engineering the system.

Secondly, TA2 architectures that support multiple parallel and competing streams of deliberation are of interest. Architectures with designed roles for dissent, verification, and reconciliation, are of particular interest. Proposers should describe how their orchestration approach leverages insights from team science and organizational design to structure AI reasoning, rather than relying on single-thread chain of thought.

TA2 is the principal interface for the researcher and must explain its reasoning through narratives and visualizations so that scientists can effectively direct the research. TA2 will enable a researcher to see and compare experimental options, including modalities they have never used and that are not available locally, with an understanding of how each option would change the TA1 disease model. This is a critical capability: IGoR aims to free researchers from the constraint of designing experiments based only on what they know and what equipment their institution owns.

Objective 1: Hypothesis Generation and Experiment Design. TA2 must identify knowledge gaps (in coordination with TA1, as described above), generate testable hypotheses to address those gaps, and design experiments that would most efficiently resolve them. This includes defining the specific knowledge that must be collected, such as cell type, time frame, stimuli, response, and phenotype, in sufficient detail to enable generation of an experimental procedure.

TA2 is not a wrapper around a frontier LLM. Systems that merely prompt an LLM to suggest experiments and route them to an automated laboratory do not meet the requirements of this program. TA2 must demonstrate mechanistic grounding (via TA1), multi-modal experimental design (via TA3), and distributed execution across heterogeneous laboratories (via TA4) — capabilities that require architectural innovation beyond what current LLM-to-lab integrations provide.

Objective 2: Explainability and Human Interface. The researcher and AI must operate as a collaborative unit in which the AI surfaces diverse perspectives, competing hypotheses, and structured disagreements, while the researcher exercises judgment over research direction.

The design goal is augmentation, not automation: the system must make researchers more productive and more creative, expanding the range of hypotheses they can consider and the experiments they can access, while preserving and strengthening — not atrophying — their scientific agency. The system must be designed so that human expertise and AI capabilities complement each other, and the human role is not merely supervisory or reduced to approving AI recommendations. TA2 must provide:

- Clear narratives explaining why a particular knowledge gap was identified and why a proposed experiment is expected to resolve it.
- The ability for researchers to incorporate their own insights and domain expertise.
- Sufficient detail (cell type, time frame, stimuli, response, phenotype, etc.) to enable TA3 protocol generation.

Approaches that anchor LLM outputs to mechanistic models (TA1) to reduce hallucinations and improve factual grounding are expected.

Out of scope for TA2:

- Development of new frontier or foundational AI models (e.g., pre-training a new LLM). Fine-tuning, prompt engineering, and development of specialized smaller models are in scope.
- Systems that make irreversible commitments of resources (e.g., ordering experiments, allocating laboratory time) without researcher approval. Autonomous AI deliberation, including multi-agent debate and self-organized reasoning, is encouraged provided that consequential decisions require human authorization.
- System designs in which the primary effect is to replace human researchers rather than to amplify their capabilities.

4.1.3 TA3: Interoperable Experimental Procedures

Once a hypothesis has been selected, the next step is to generate clear procedures for executing experiments and to identify the optimal laboratory to which it will be assigned. This addresses a current infrastructure gap: today, transferring an experiment from one laboratory to another requires extensive negotiation, protocol adaptation, and troubleshooting. TA3 will make this transfer as straightforward as sending a data file.

To this end, TA3 will define intuitive yet comprehensive protocol representations so that any qualified laboratory — whether a contract research organization (CRO), cloud lab, or conventional academic lab — can execute the same experiment in a highly reproducible way. TA3 will enable a researcher to describe an experiment in declarative terms and have it executed at any qualified laboratory, with the protocol stack handling the translation to site-specific instruments, reagents, and conditions. Approaches that enable a high diversity and novelty of experiments, even while maintaining standardization required for reproducibility and data aggregation, are of particular interest.

Objective 1: Layered Protocol Architecture. TA3 performers should describe a protocol stack supporting, at a minimum, the following abstractions:

- Intent: Specifies what needs to be done in declarative terms — the scientific question, the variables to be tested, the expected outputs, and the quality requirements. Focuses on *what*, not *how*.
- Protocols: Define standard processes, such as dispensing, heating, mixing, measuring, using shared formats that include error handling and fault tolerance across diverse laboratory environments (e.g., built-in checks, retries, and feedback loops). Relies on calibrated parameters and standard reactions without requiring low-level instrument specifics.
- Calibration: Standardizes parameters and sensitivity across devices, enabling interoperability. Each measurement records its context, source, and uncertainty using open formats. Calibration layer conformance will be established by creating, in collaboration with the Independent Verification and Validation (IV&V) partner, a set of required calibration artifacts.
- Hardware: Contains machine-specific settings and instructions. Isolates hardware specifics so that differences in equipment are managed through adaptive control and common capability interfaces.

A critical goal of the layered architecture is to eliminate procedural variation with weak scientific justification that reflects local preference rather than variables of interest. These variations create systematic batch effects that severely limit the ability to aggregate datasets for model training and cross-laboratory comparison. The TA3 protocol stack must distinguish between parameters that are scientifically meaningful and parameters that are procedurally arbitrary (and therefore should be locked to a consensus default). Changes to locked parameters must be justified through a Request-for-Comments (RFC) process and supported by evidence that the change improves scientific outcomes. This discipline is essential to ensure that the data will be interoperable at the level required for TA1 model training.

The protocol stack must encode rich metadata including cell lines, reagents, calibration status, quality control parameters, power analyses, and uncertainty estimates.

Proposers must explain how their protocol architecture can accommodate at least three distinct experimental modalities (e.g., imaging, biochemical assays, omics, mechanical measurements, functional assays).

Objective 2: Standards Development and Interoperability. TA3 will be developed in collaboration with the IV&V partner, using community-driven standards processes to converge on protocol representations. Proposers should plan to engage on:

- Bake-offs: Interactive sessions where performers challenge each other with increasingly complex interoperability tasks that will test the interoperability of common experiments and showcase extensions of the protocol standard.
- Request-for-Comments (RFC) process: A structured, collaborative method will be used to propose, review, and adopt or change standards in protocol specification.

- Open standards: The resulting data and metadata layer will be as open as possible — open schemas, ontologies, and reference implementations — while allowing extensible namespaces for proprietary data and processes.

Proposers must describe how they will engage with the broader scientific community, equipment manufacturers, and standards bodies to ensure that the protocol standards developed under IGoR have adoption potential beyond the program.

Out of scope for TA3:

- Development of closed or proprietary protocol systems that do not support open-access data sharing or interoperability.
- Protocol development for vertebrate animals or clinical trials. TA3 protocols will focus on cell/tissue culture and lower invertebrates.

4.1.4 TA4: Experiment Marketplace

TA4 will make ordering a validated experiment as routine as querying a database. TA4 will create a network of validated laboratories, such as cloud labs, core facilities, CROs, and independent research labs, that can execute standardized TA3 protocols and return high-quality data and metadata. The marketplace will enable researchers to request experiments from any qualified laboratory in the network, receive results in a standardized format, and have those results automatically ingested into TA1 disease models.

A critical design consideration for TA4 is the balance between standardization and flexibility. The marketplace must ensure that experiments are reproducible and that data is trustworthy, but it must not lock in the *types* of experiments that can be performed. The long-term goal is not to offer a fixed menu of assays, but to create an infrastructure in which innovative experimental concepts can be validated, onboarded, and made available to the research community. The marketplace should grow over time, incorporating new instruments, assay types, and laboratory partners.

Objective 1: Validated and Reproducible Experimentation. TA4 performers will establish experimental workflows that produce reliable, well-characterized data suitable for updating TA1 models. This includes:

- Defining and validating cell lines, reference materials, and standard samples that ensure reproducible execution of key experimental procedures.
- Developing quality control (QC) parameters throughout protocols and workflows that inform experimental success and variability, including sources and characterization of cell lines, handling and culture conditions, sample QC, and reagent and buffer QC.
- Demonstrating reproducibility across laboratories by executing the same experiment at multiple TA4 sites and achieving predefined concordance thresholds.
- Monitoring the number of exceptions (situations requiring human intervention) raised per experiment as a key metric and implementing automated exception handling wherever possible.

In Phase I, experiments may involve the growth, manipulation, and characterization of cultured cells. In Phase II, teams must expand to multicellular systems (e.g., organoids, microbiomes, or invertebrate animals) while increasing the number of cell lines and experimental modalities. Performers are strongly encouraged to develop flexible protocols and workflows that can be readily adapted to additional areas of study.

Labs will progressively increase the fraction of experimental workflows that can be executed autonomously, with the number of exceptions per experiment serving as a critical metric. Proposals should describe planned experimental workflows and how they align with program goals.

Objective 2: Marketplace Operations. TA4 will develop the operational infrastructure for a distributed research marketplace, including:

- Simple interfaces for researchers to request experiments and receive results.
- Continuously updated information on laboratory capabilities, experimental precision, speed of execution, cost, and availability.
- Two-way communication of laboratory capabilities, calibration data, experimental procedures, results, and exceptions.
- Mechanisms for onboarding new laboratories and new experimental capabilities into the marketplace, including validation against IV&V standards.

Each team must include at least two TA4 laboratories. In Phase I, reproducibility and concordance metrics will be assessed within each team's own laboratories. In Phase II, teams will begin cross-team experimentation: laboratories from each team must demonstrate the ability to execute experiments requested by another team's TA2 engine, using a common TA3 protocol standard. By Phase III, the marketplace will operate as a unified network in which any qualified researcher can request experiments from any participating laboratory across all teams. Proposers must describe how experimental data and metadata will be returned to TA1 in formats that support automated model updating.

Out of scope for TA4:

- Experiments in vertebrate and higher-level invertebrate animals (see Section 2 for limited exceptions conducted outside the TA4 marketplace).
- Human clinical trials.
- Experiments that do not return data and metadata in formats compatible with TA1 model ingestion.

5. Program Structure

5.1 Proposal Scope

To be responsive to this solicitation, proposals must address all four TAs (TA1, TA2, TA3, & TA4). Proposals not addressing all four TAs will not be reviewed. An organization may serve as a sub-performer under multiple proposals. If ARPA-H selects two or more proposals in which the same sub-performer is proposed to perform work under the same Technical Area, ARPA-H reserves the sole right to determine whether the proposed efforts

are substantially distinct or whether a material overlap exists. In the event ARPA-H determines that such an overlap exists, ARPA-H may, at its discretion, (i) fund the overlapping work only once, or (ii) require the sub-performer to elect the project in which it will participate.

Each proposal team is expected to include performers with the expertise and capabilities to address all four TAs. This may involve a single organization with breadth across all areas or, more likely, a team of organizations, each contributing expertise in one or more TAs. A single team member may cover more than one TA. Proposers must describe how their team is organized, how the TA responsibilities are distributed, and how the team will coordinate across TAs to ensure closed-loop operation. Team formation is the sole responsibility of the proposing teams. Proposers must submit a single proposal led by one Principal Investigator under a single prime performer that addresses program Phase I, Phase II, and Phase III.

5.2 Phases

IGoR is a 5-year, 3-phase program structured as follows:

Phase I — Concept and Component Development (18 months). Phase I focuses on establishing the foundational capabilities for each TA and demonstrating initial integration within each team.

- Establish the core architecture for each TA: disease models, AI orchestration, protocol stack, and laboratory workflows.
- Demonstrate an initial closed-loop cycle within each team, from hypothesis through experiment execution to model update.
- Participate in program-wide standards activities, including workshops and IV&V packaging conventions.

Key activities are detailed in Table 1.

Phase II — Integration and Interoperability (18 months). Phase II focuses on demonstrating robust operation within each team and beginning cross-team interoperability.

- Show that closed-loop operation produces measurable improvements in TA1 model predictive performance.
- Begin cross-team experimentation: laboratories from one team execute experiments designed and specified by another team.
- Expand experimental complexity to multicellular systems and validate reproducibility across ≥3 laboratories.

Key activities are detailed in Table 1.

Phase III — Scaling, Generalization, and Transition (24 months). Phase III focuses on demonstrating the system at scale, generalizing to new disease areas, and preparing for transition.

- Extend the system to a second proposed disease area and demonstrate that external researchers can use IGoR end to end.
- Operate a unified experiment marketplace across all teams, validated through a collaborative process.
- Deliver all computational artifacts to the public repository and establish at least one sustained partnership or adoption path.

Key activities are detailed in Table 1.

5.3 Program Milestones and Metrics

The following tables present the program milestones and metrics. These represent minimum requirements and serve to bound the scope of the effort while affording maximal flexibility, creativity, and innovation. Proposers should propose additional quantitative metrics appropriate to their specific approach for each phase. Achievement of all metrics, as agreed to by ARPA-H, is the basis for continuation to subsequent phases.

Table 1: Program Milestones

	Phase I	Phase II	Phase III
System-Level	Domain-Driven Design workshop completed; shared interface specification published. Initial walking skeleton demonstration within each team (TA2 proposes experiment → TA3 generates protocol → TA4 executes → data returned to TA1)	Cross-team interface specification updated based on Phase I learnings; all teams confirm interoperability test plan. Closed-loop operation with measurable improvement in TA1 predictive performance for at least one disease area	Transition plan delivered; at least one sustained partnership or adoption path established. Cross-team interoperability demonstrated end-to-end; external researchers (not on performer teams) use IGoR to design and execute experiments
TA1: Comprehensive Disease Models	Disease model architecture documented (languages, schemas, ontologies); existing databases and model repositories surveyed and integration plan delivered. At least one mechanistic disease model demonstrated with algorithmic detection of ≥3 specific, quantitative knowledge gaps; model accepts and processes at least one TA4 experimental data return	Model documentation reviewed and accepted by Computational IV&V partner; model cards delivered for all sub-models. Automatic model updates from TA4 experimental data demonstrated; model generates ≥3 novel predictions not present in training literature, of which ≥1 is experimentally confirmed via TA4	All models deposited in public repository in certified executable form; documentation meets open-access standard. Model extended to second disease area with algorithmic gap detection; model generates novel, experimentally validated hypotheses in both disease areas

	Phase I	Phase II	Phase III
TA2: New Science Engine	AI orchestration architecture documented; interface with TA1 model and TA3 protocol generation demonstrated. Given a TA1 model and literature, system identifies plausible knowledge gaps and proposes candidate experiments; ≥50% of proposed experiments judged high-value by expert panel	Operation demonstrated with ≥2 model backends (including ≥1 open-weight); researcher usability study completed with ≥10 domain scientists. ≥75% of proposed experiments judged high-value by expert panel; system generates experiment designs that lead to measurable TA1 model improvement	≥85% of proposed experiments judged high-value; system demonstrated on second disease area; efficiency of experimental design approach quantified relative to conventional baseline
TA3: Interoperable Experimental Procedures	Initial protocol schema defined; calibration artifacts defined (in collaboration with IV&V partner where possible). Same protocol executed at two (2) TA4 laboratories (within the same team) with comparable outcomes across ≥1 experiment	RFC process operational; ≥2 RFCs issued and responded to by all teams; protocol schema updated based on Phase I findings. Protocols communicate with ≥3 laboratories (including ≥1 from a different team), each executing ≥3 core experiments with predefined reproducibility thresholds met	Open data and metadata layer delivered (schemas, reference implementations, benchmark datasets); engagement with ≥1 external standards body or equipment manufacturer documented. Protocols communicate with ≥5 laboratories across teams; cross-team protocol interoperability demonstrated at connect-a-thon
TA4: Experiment Marketplace	Summary table of experimental workflows delivered; ≥1 cell line & instrument validated on IV&V test artifacts at each of the team's two laboratories; two-way communication of capabilities established. Intra-team reproducibility demonstrated: same experiment executed at both team laboratories with ≥85% concordance	≥2 additional instruments validated at each lab; marketplace interface demonstrated for experiment request and result return; exception handling baseline established. Cross-team experiment execution demonstrated: ≥1 laboratory from a different team executes an experiment with ≥85% concordance; exceptions per experiment reduced ≥50% relative to Phase I	Marketplace onboarding documentation published; all laboratory capability manifests current and publicly queryable. Unified marketplace operational across all teams; experiments ordered and executed across teams; ≥90% concordance on standardized experiments; connect-a-thon completed
Standards	Technical and governance framework for IV&V collaboration established; TA3 protocol schema	Bake-offs and RFC process operational; cross-team standards convergence documented. At least one	Sustained partnership or adoption path established (e.g., NIH center, cloud lab provider, academic

	Phase I	Phase II	Phase III
	sufficiently specified to support cross-team laboratory onboarding. Domain-Driven Design workshop completed with shared interface specification published	cross-team interoperability test completed successfully using common standards	consortium). Open data + metadata layer delivered and adopted by all teams; connect-a-thon demonstrates full interoperability
Computational IV&V	Standards workshop completed; documentation and packaging conventions published; IV&V environment operational. 100% of TA1 models and software delivered in containerized form; IV&V partner confirms build-and-execute on independent environment for 100% of artifacts	All artifacts pass IV&V documentation review; model cards complete for all sub-models. Reproducibility verified on ≥2 hardware configurations for all TA1 models; cross-team model comparison initiated	Public repository operational with long-term hosting and versioning. All artifacts deposited in public repository in certified executable form; cross-team comparisons completed and published; open-access certification issued

Table 2: Program Metrics

	Phase I	Phase II	Phase III
Marquee Metrics			
Disease / Therapeutic Area	1	1	2nd related
Experimental Cycle Time	Baseline established	≥4x improvement	≥10x improvement
Laboratories generating data	≥2, with ≥80% concordance across ≥1 experiment	≥2, with ≥90% concordance across ≥3 experiments	≥3, with ≥90% concordance across ≥3 experiments (including cross-team labs)
TA1: Comprehensive Disease Models			
Knowledge gaps identified	≥3 known gaps algorithmically detected	≥3 novel gaps identified (not in training literature)	≥3 novel gaps for second disease area
Sub-models integrated	≥3 mechanistic sub-models	≥10 sub-models spanning ≥2 biological scales	≥15 sub-models spanning ≥3 scales

	Phase I	Phase II	Phase III
TA4 experimental data returns ingested	≥1	≥10	≥25
Model update latency (time from data receipt to model update)	Baseline established	≤24 hours	≤4 hours
Prediction accuracy (on held-out TA4 data)	Prediction methodology defined; baseline established using available data	Statistically significant improvement over literature-only model	≥2x improvement over Phase I baseline
Explainability	Explain 1 specific set of known pathways	Explain ≥75% of pathways	Explain ≥90% of pathways
TA2: New Science Engine			
Hypothesis generation quality (expert panel)	≥50% judged high-value	≥75% judged high-value	≥85% judged high-value
Experimental procedures automatically generated	≥3	≥5	≥8
Model backend diversity	≥1	≥2 (including ≥1 open-weight)	≥3
Researcher usability (domain scientists)	N/A	≥10 researchers evaluated; ≥70% rate system as useful	≥20 researchers (including external); ≥80% rate as useful
TA3: Interoperable Experimental Procedures			
Labs supporting the layered protocol stack	≥2	≥3 (including ≥1 cross-team)	≥5 (pooled across teams)
Experimental modalities supported	≥2	≥3	≥4 (including cross-team)
RFCs issued and resolved	N/A	≥2	≥4 (cumulative)
TA4: Experiment Marketplace			

	Phase I	Phase II	Phase III
Cell lines and instrumentation validated on IV&V artifacts	≥1	≥3	≥8
Inter-laboratory reliability (on standard procedures)	80%	90% (including ≥1 laboratory from another team)	90% (across all marketplace laboratories)
Exception handling (fraction handled autonomously)	0% (manual handling)	≥30%	≥70%
Cross-team laboratory integration	Marketplace architecture supports external lab onboarding (demonstrated)	≥1 cross-team lab executes experiment with ≥85% concordance	All team laboratories operate as full marketplace participants
Fraction of procedures executed automatically	Baseline established	≥30%	≥50%
IV&V			
TA1 model reproducibility	100% of models build and execute on IV&V environment	100% reproduce on ≥2 hardware configs	100% deposited in public repository with open-access certification
TA3 & TA4 protocol reproducibility	100% of components build and execute on IV&V environment	100% reproduce on IV&V environment	100% deposited in public repository with open-access certification
Documentation completeness	Model cards and dependency manifests delivered for all artifacts	All artifacts pass IV&V documentation review	All artifacts meet open-access documentation standard

The 10x target improvement in Experimental Cycle Time refers to the time from initial hypothesis to validated experimental insight — the full cycle of identifying a knowledge gap, designing an experiment, executing it, and updating the disease model. The baseline for this metric will be established in Phase I by measuring cycle times for the team's initial experiments using conventional approaches. Subsequent phases will measure improvement against this baseline.

5.4 Independent Verification and Validation (IV&V)

ARPA-H will contract with one or more qualified organizations to provide independent verification and validation across the IGoR program. IV&V responsibilities span two areas, each covering different TAs:

IV&V Area	Covers	Purpose
Mechanistic Models	TA1	Verify that disease models produce consistent, scientifically sound predictions
Experimental Infrastructure	TA3, TA4	Validate protocol standards, calibration artifacts, and inter-laboratory reproducibility

These areas may be performed by the same or different organizations. Performers do not need to budget for IV&V partners, but proposals and budgets should account for the effort required to collaborate with them, including containerizing software, maintaining documentation, transferring materials, and attending biannual reviews.

Mechanistic Model IV&V (TA1)

Beyond software reproducibility, TA1 disease models require scientific verification: confirming that model predictions are internally consistent, that sub-models compose correctly, and that results are comparable across performer teams. IV&V will include:

- *Cross-team model comparison.* In Phase II and Phase III, conduct head-to-head evaluations of TA1 models from different performer teams using common datasets and standardized evaluation protocols, reporting comparative performance to the Program Manager and to performer teams.
- *Model composition verification.* Verify that sub-models can be combined, separated, and updated without introducing inconsistencies — confirming that the modularity required by TA1 is genuine, not nominal.
- *Prediction consistency.* Verify that model predictions are stable under perturbation of inputs within expected ranges, and that uncertainty estimates are calibrated against experimental outcomes as TA4 data accumulate.

Phased expectations:

- *Phase I:* Model cards delivered for all sub-models; IV&V partner reviews model architecture documentation and confirms that the framework supports modular composition and automated data ingestion.
- *Phase II:* Cross-team model comparisons initiated using common datasets. Model generates ≥3 novel predictions not present in training literature, of which ≥1 is experimentally confirmed via TA4; IV&V partner independently verifies the confirmation.
- *Phase III:* Cross-team model comparisons completed and results published. IV&V partner certifies that models in the public repository produce documented, reproducible predictions in both disease areas.

Experimental Infrastructure IV&V (TA3, TA4)

For experiments to be trustworthy and reproducible across laboratories, protocol standards and laboratory calibration must be independently verified. IV&V will include:

- *Protocol standards.* Coordinate the establishment of standards for experimental processes, cell lines, and calibration artifacts. Participate in bake-offs and the RFC process for protocol standards development.
- *Calibration artifacts.* Define, in collaboration with TA3 performers, a set of required calibration artifacts that laboratories must validate against before executing marketplace experiments.
- *Inter-laboratory reproducibility.* Provide independent assessment of reproducibility when the same experiment is executed at multiple TA4 sites, verifying that concordance thresholds are met.
- *Cross-team laboratory validation.* In Phase II and Phase III, independently assess whether laboratories from one team can execute experiments specified by another team's TA3 protocol stack and produce concordant results.

Phased expectations:

- *Phase I:* Calibration artifacts defined; at least one instrument and one cell line validated at each of the team's two laboratories. IV&V partner confirms that intra-team reproducibility meets ≥85% concordance threshold.
- *Phase II:* IV&V partner independently verifies cross-team experiment execution and assesses whether concordance thresholds are met across ≥3 laboratories.
- *Phase III:* IV&V partner certifies marketplace readiness: all participating laboratories meet calibration and reproducibility standards. Independent assessment of connect-a-thon results.

Expert Panel

The validity and impact of discovered drug targets or mechanisms of action will be evaluated by a panel of independent experts assembled by ARPA-H. While ARPA-H will determine panel composition, performers may recommend experts to include or exclude. Non-disclosure agreements will be established prior to panelists receiving proprietary program information.

5.5 Project Management, Integration, and Collaboration

Project Management. Proposals should clearly describe plans for overall project management and for integration across technical areas. Each team is responsible for closed-loop operation of its own system across all four TAs. Proposers must describe how their team will coordinate across TAs, including explicit plans for interaction among collaborators and sub-performers. The person or people directly leading integration should be identified as well as the hours they will dedicate to integration each year. Proposals should identify the individual responsible for day-to-day project management, cross-team integration, delivery tracking, dependency management, and reporting. This individual should have project management expertise, experience managing large multi-team efforts, and a suitable level of effort to manage the project. Additionally, the proposal should name the individual responsible for interface design, cross-team interoperability, and architectural consistency. This individual should have demonstrated experience building systems that interoperate with external partners (e.g., open-source infrastructure, protocol implementations, API-first platforms). This "software architect" role is distinct from the PI

and Project Manager. Proposers should describe how this role will exercise architectural authority over implementation decisions, especially those accelerated by AI coding tools.

Cross-team collaboration. All performer teams are expected to interact and work collaboratively with other teams in developing methods, technologies, and tools, using open, timely, and effective communication, information exchange, and reporting. Performers across all teams will attend common meetings and technical exchanges. In particular:

- Phase I: Teams will participate in a Domain-Driven Design workshop, convened by ARPA-H at program kickoff, to collaboratively define consensus TA boundaries, interfaces, and information flows. Each performer team must be substantially represented at this workshop. The workshop will produce a shared interface specification that all teams will implement, enabling cross-team interoperability in later phases. This workshop is a program milestone.
- Phases I–II: Teams will participate in working groups to develop common protocol standards (TA3), including active participation in bake-offs and the RFC process. Full cooperation in the establishment of the research protocol suite is expected, including thoughtful and routine participation in the generation of and response to RFCs.
- Phases II–III: Teams will demonstrate cross-team interoperability, with laboratories from one team executing experiments designed by another team's TA2 and specified using a common TA3 protocol standard.
- Phase III: Teams will participate in a connect-a-thon or equivalent interoperability demonstration, in which TAs from different teams are combined and evaluated for end-to-end performance.

Associate Performer Agreements (APAs) are agreements between performer teams on a program that establish the terms for sharing information, data, and materials. APAs specify what types of information will be freely shared, what protections apply, and what obligations each party assumes. They are a standard mechanism in ARPA-H programs to enable cross-team collaboration while protecting legitimate proprietary interests.

Initially, teams will operate as self-contained units. By Phase III, cross-team collaboration will be required, and APA language will be included in awards to facilitate the open exchange of information. Each performer will work with other performers to develop an APA that specifies the types of information that will be freely shared across teams.

5.6 Commercial Transition

Throughout the program, performers will work closely with government, private sector, and research partners to refine the commercialization approach for IGoR components. To ensure long-term sustainability and broad adoption, performers will develop technologies in a manner that creates incentives for ongoing investment, development, and adoption. Examples may include:

- Establishing public-private partnerships for sustained operation of the experiment marketplace.

- Pursuing business or academic funding models that support the use of AI-powered research tools.
- Enabling new funding and collaboration models where philanthropies, agencies, and patient groups can invest directly in defined disease-understanding milestones.
- Demonstrating proof of concept for integrating IGoR technologies into realistic research workflows for one or more diseases of interest.

A key ingredient in IGoR's long-term success will be the open data and metadata layer (TA3) and the demonstrated interoperability of the experiment marketplace (TA4), which together create the foundation for a self-sustaining research ecosystem.

Works Cited

[i] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature. https://doi.org/10.1038/533452a

[ii] Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. PLOS Biology. https://doi.org/10.1371/journal.pbio.1002165

[iii] Errington, T. M., et al. (2021). Investigating the replicability of preclinical cancer biology. eLife. https://doi.org/10.7554/eLife.71601