HUGGING FACE

Hugging Face Response to the American Science Acceleration Project (ASAP) RFI

Submitted to: Offices of Senator Heinrich and Senator Rounds, United States Senate
June 2025

About Hugging Face

Hugging Face is a U.S.-based, community-driven company committed to democratizing responsible artificial intelligence (AI) and machine learning (ML). Our open-source platform is the most widely used resource for sharing and collaborating on ML models and datasets, currently hosting thousands of scientific models spanning protein folding to climate simulation.

Our mission aligns directly with ASAP's vision for democratized scientific innovation. Through our platform coordinating millions of users and hosting models that advance fields across the sciences, we have witnessed how open AI infrastructure accelerates research. Our work pioneering accessible AI tools, developing industry-standard libraries, and creating educational resources reflects our commitment to ensuring AI serves the broader scientific community rather than remaining concentrated in elite institutions.

Executive Summary

The American Science Acceleration Project presents a generational opportunity to ensure AI and open science serve as catalysts for inclusive, rapid, and responsible scientific discovery. As the world's largest hub for open ML— with over 1.7 million models serving 5 million users — we have seen firsthand how community-driven AI can accelerate research. Our BigScience initiative, involving over 1,000 researchers from 60 countries, demonstrated the potential of collaborative AI with the creation of the large, multilingual BLOOM model. Our status as the principal platform for AI system research has afforded us a unique perspective in evaluating the technology's scientific acceleration and our potential to augment this impact.

This response outlines recommendations aligned with the ASAP initiative, with specific question callouts wherever applicable.

1. Accelerating U.S. Scientific Innovation (Q1, Q8)

AI is rapidly transforming science through accelerated discovery and growing access to powerful tools. AlphaFold cut protein structure prediction from months to minutes, aiding over 2

million researchers globally. BenevolentAI identified a COVID-19 treatment candidate in just 48 hours by mining scientific literature, with clinical trials later showing a 20% proportional reduction in mortality. Microsoft and the Pacific Northwest National Laboratory used AI to discover new battery materials in under 80 hours, computationally screening 32 million candidates before synthesizing promising prototypes.

Crucially, many breakthroughs are being shared as full research artifacts — open-source code, model weights, and documentation — enabling direct experimentation and reuse. IBM-NASA's Prithvi climate models are freely available on Hugging Face for global disaster and climate research. Meta open-sourced its ESMFold protein structure predictor on GitHub with full training infrastructure. The OpenFold Consortium released end-to-end protein modeling tools, including training code for customization. Stanford's ChemBERTa models for molecular property prediction are accessible through DeepChem on Hugging Face. These artifacts empower institutions of all sizes to build on cutting-edge work rather than merely cite it.

To sustain this momentum, the U.S. must avoid concentrating AI capacity in a handful of elite institutions. Congress should fund open science and incentivize the release of reusable AI research artifacts. Federal Chief AI Officers across agencies can coordinate interagency efforts. Industry should contribute compute and technical expertise while maintaining openness. Civil society ensures accessibility, trustworthiness, grounding and representation, while academia continues to drive fundamental research and train AI-literate scientists. A collaborative, open infrastructure is key to unlocking AI's full potential in science.

Recommendations:

- Invest in open-source AI infrastructure to democratize access, building on successful projects like Hugging Face's Transformers library or Alphafold, which is now open source.
- Support model compression and efficient inference techniques to reduce barriers for under-resourced institutions[1].
- Incentivize cross-institutional collaboration through shared tools, standardized datasets, and accessible interfaces.
- Effectively coordinate inter-agency efforts across NSF, NIH, DOE, NIST, and other agencies via the Chief AI Officers (CAIOS).

2. Building Scientific Data Infrastructure (Q6)

Scientific data remains trapped in incompatible formats and ambiguous overlapping standards across disciplines and institutions. Furthermore, each federal agency maintains independent data standards, creating interoperability barriers that in turn limit AI adoption. The challenge

[1] We refer to efficiency in detail in our 2025 NSF National AI R&D Strategy Response.

extends beyond technical standards to cultural and institutional barriers where researchers lack incentives to properly document and share data.

Successful data ecosystems require domain-specific ontologies that capture field-specific knowledge while maintaining cross-domain compatibility, automated data quality assessment tools that identify and flag potential issues, version control systems designed for large scientific datasets, and standardized APIs enabling programmatic access across repositories.

Recommendations:

- Launch a National Scientific Data Standardization Initiative through NIST, with the aim to enable all federal science agencies and American institutions relying on federal grants to contribute to accelerating science with AI.
- Require adoption of FAIR principles (Findable, Accessible, Interoperable, Reusable) for all federally funded research data.
- Improve Data.gov with more open scientific datasets in critical research areas, standardized metadata and petabyte-scale storage.
- Develop privacy-preserving tools including differential privacy and homomorphic encryption for sensitive research data.
- Link data sharing requirements to grant funding, creating strong incentives for proper data curation.
- Implement unified data sharing and research standards beginning with pilot implementations.

3. Democratizing Compute Access and Infrastructure (Q2, Q7)

Recent advances in model efficiency have democratized access to sophisticated AI. Quantization techniques now enable extremely large models to run on consumer grade GPUs — wih one study showing 16x reduction in memory requirements. Knowledge distillation allows transfer of capabilities from large foundation models to compact versions optimized for specific scientific domains. However, significant disparities remain: most academic institutions access older-generation GPUs while industry leverages state-of-the-art H100 systems, cloud compute costs can be extremely cost prohibitive for sustained research projects, and 80% of researchers cite skill gaps as their primary adoption barrier. Compute resources extend beyond GPUs to include CPU time for density functional theory (DFT) simulations critical in materials science and chemistry.

Public compute infrastructure represents essential scaffolding for democratized scientific innovation. Like previous transformative infrastructure — from particle accelerators to the Internet — NAIRR must serve as public infrastructure enabling research impossible at

institutional scale. Out of the more than 150 proposals that the NAIRR pilot received, it was only able to award 35 projects in its first round of allocations, when the resource demand for scientific AI research encompasses thousands of potential applications. The marketing challenge remains critical — many researchers remain unaware of NAIRR's existence despite its valuable resources and partnerships.

Infrastructure funding should follow a tiered approach: core national resources (NAIRR, supercomputing centers) funded through direct federal appropriation, regional hubs supported through state-federal partnerships, institutional resources enhanced through competitive grants and equipment programs, and cloud resources subsidized through negotiated academic agreements.

Recommendations:

- Expand NAIRR into a persistent national research resource with dedicated multi-billion-dollar funding separate from project-based grants, supporting tens of thousands of projects annually.
- Adopt public-private partnership models demonstrated by successful national computing initiatives, including the U.S. DOE's Aurora exascale system (a collaboration between Argonne National Laboratory, Intel, and HPE), Norway's national supercomputing infrastructure through Sigma2, Spain's Barcelona Supercomputing Center MareNostrum facility, and France's Jean Zay Supercomputer.
- Negotiate academic-tier pricing with cloud providers through federal purchasing power for federal and state/local government funded institutions.
- Develop hybrid edge-cloud infrastructure combining local scientific computing with burst capacity for intensive training workloads.
- Create comprehensive outreach programs ensuring all eligible researchers can access NAIRR and other public compute resources.

4. Fostering Collaboration and Innovation (Q5, Q9)

The NSF AI Research Institutes demonstrate effective collaboration patterns, connecting over 500 institutions with $500+ million in funding. Multi-national, open collaboration, such as Cohere's Aya, can produce rigorous, scalable AI models while maintaining scientific standards. The project coordinated computational resources, data curation, and model development across time zones and disciplines. Effective collaboration requires shared evaluation frameworks across disciplines, common vocabularies bridging domain-specific terminology, career incentives rewarding interdisciplinary work in tenure and promotion decisions, and collaborative spaces—both physical and virtual—that facilitate serendipitous interactions.

Structured challenges have proven exceptionally effective at accelerating innovation while fostering broad participation. DARPA's AI Cyber Challenge allocated $29.5 million among 42

teams, resulting in 22 distinct vulnerability discoveries and 15 successful patches. The Subterranean Challenge yielded technologies currently implemented in agriculture, mining, and defense. The Vesuvius Challenge has awarded over $1.5 million in prizes to decode carbonized Herculaneum scrolls buried by Mount Vesuvius in 79 AD, with winning teams using AI to read ancient Greek text for the first time in 2,000 years. Platforms such as Kaggle engage millions globally in machine learning competitions, while DrivenData has provided over $4.8 million in prizes across climate, health, and conservation challenges, drawing 242,000 submissions from data scientists worldwide.

These successful models point toward similar grand challenge objectives that could guide American scientific acceleration: reducing drug discovery time by several orders of magnitude for targeted disease categories, accelerating materials discovery for clean energy while maintaining safety standards, enabling real-time climate modeling at high resolution globally, automating routine laboratory procedures while maintaining experimental rigor, and achieving near perfect accuracy in early disease detection from standard medical imaging.

Recommendations:

- Support long-term (4–5 year) collaborative grants with explicit requirements for interdisciplinary team composition, prioritizing domain scientists as leads.
- Build platforms for federated learning that enable secure research collaboration while respecting data sovereignty and privacy.
- Establish "matchmaking" services connecting domain scientists with AI researchers, expertise and tools.
- Create annual federal AI-for-science challenges with substantial prize pools addressing grand challenges.
- Require open-source outputs for all publicly funded competitions ensuring broad benefit.
- Formalize international compute sharing agreements with ally nations and key research partners.

5. Building AI Literacy and Trust in Scientific Practice (Q3, Q10)

Domain scientists often resist "black-box" AI systems that conflict with evidence-based scientific practices. Scientific AI must provide interpretable results, transparent methodologies, and rigorous validation processes aligned with scientific values of reproducibility and peer review. Successful human-AI collaboration requires iterative feedback loops where AI generates hypotheses that humans validate experimentally, mixed-initiative interfaces allowing scientists to guide AI systems, provenance tracking documenting AI contributions to each scientific claim, and fail-safe mechanisms ensuring human oversight of critical decisions.

Current AI education remains fragmented and inaccessible, creating bottlenecks where valuable scientific expertise cannot leverage AI capabilities. A comprehensive strategy must address K-12 education introducing computational thinking and AI concepts early, undergraduate programs requiring AI literacy across all STEM fields, graduate training combining deep domain knowledge with practical AI skills, professional development helping established researchers adopt AI tools, and industry partnerships providing real-world experience with cutting-edge AI applications.

Recommendations:

- Establish research into standardized explainability mechanisms for all AI systems used in federally funded scientific research.
- Develop comprehensive training programs teaching scientists to validate AI outputs using established methodologies.
- Establish peer-review guidelines for AI-assisted research, maintaining scientific rigor while acknowledging AI contributions.
- Create comprehensive documentation trails to record all AI involvement in scientific discoveries.
- Launch federal AI skills Initiatives and fellowship programs with substantial funding, providing comprehensive training from K-12 through professional development, and ensuring collaborations between domain-expert scholars and AI technologists.
- Develop no-code/low-code interfaces enabling domain scientists to apply AI without extensive programming knowledge.

6. Measuring Success and Enabling Deployment (Q4)

AI demonstrably accelerates scientific discovery in multiple critical fields. For instance, drug discovery timelines have been compressed from 5+ years to months. Materials research that once required years is now completed in weeks. However, speed alone inadequately captures AI's transformative potential.

The federal government needs centralized tracking systems aggregating AI usage across funded research, standardized reporting requirements for AI-assisted discoveries, real-time dashboards monitoring infrastructure utilization and research outputs, and predictive analytics identifying emerging trends and resource needs.

Accelerating deployment requires risk-based regulatory frameworks distinguishing low-risk research tools from high-stakes applications, sandboxes for AI experimentation in regulated domains like healthcare, streamlined ethics and compliance frameworks reducing IRB delays from months to weeks, modernized IP policies clarifying ownership of AI-assisted discoveries, and updated peer review processes accommodating AI contributions while maintaining quality standards.

HUGGING FACE

Recommendations:

- Develop comprehensive AI-science impact Metrics that capture the time from hypothesis to validated discovery, the quality and reproducibility of outputs, the breadth of interdisciplinary collaboration, equity of access across institutions, the efficiency and sustainability of developed technology, and downstream economic impact.
- Create domain-specific benchmarks, recognizing that AI impact varies across scientific fields, to help quickly gauge success in specific domains.
- Establish multidisciplinary evaluation teams with expertise spanning technical, social, and domain-specific fields to address the complexity of comprehensive AI impact assessment.
- Institute regular evaluation cycles aligned with national science goals and ASAP objectives, with annual reports to Congress.
- Track longitudinal outcomes measuring career trajectories of AI-trained scientists.
- Implement risk-based regulatory frameworks that accelerate deployment without sacrificing safety.

Conclusion: A Vision for American Scientific Leadership

American science faces a strategic choice about AI deployment. History shows that our scientific leadership emerged from distributed innovation, open collaboration, and inclusive participation rather than centralization. The American Science Acceleration Project offers the opportunity to extend these proven principles into the AI era, ensuring broad distribution of AI capabilities maximizes innovation potential and harnesses diverse perspectives across our research ecosystem.

By investing in open infrastructure, standardized datasets, collaborative frameworks, and inclusive education, ASAP can ensure that AI accelerates discovery while strengthening the democratic foundations of American science. Success will be measured not just in breakthrough discoveries but in the breadth of participation in creating those breakthroughs.

Submitted by:
Avijit Ghosh, Applied Policy Researcher, Hugging Face
Yacine Jernite, ML & Society Lead, Hugging Face
Irene Solaiman, Head of Global Policy, Hugging Face