ARPA-H Customer Experience Hub
PART OF THE ARPANET-H HEALTH INNOVATION NETWORK

Network Survey for APECx Program
Safe & Secure Biotechnology Platforms

Overview:

This Network Survey supports ARPA-H's program for Antigens Predicted for Broad Viral Efficacy through Computational Experimentation (APECx). The APECx program is developing a comprehensive toolkit to enhance the discovery of effective vaccines against a wide range of viruses. The questions in this survey will help ARPA-H ensure end-user adoption by understanding the needs of researchers that will use the toolkits and large datasets potentially generated by the APECx program for medical countermeasure design. The survey seeks to gather input from experts on how the APECx toolkit and viral protein data can be safely and securely used while balancing access for the scientific community for research purposes. The results of this survey may be used to shape future opportunities supporting the transition of the APECx program.

APECx is developing an AI toolkit that aims to overcome the limitations of traditional vaccine development methods by enabling the modeling of broadly effective antigens that target a wide range of viral pathogens. To do this, APECx is combining technological advances along a variety of fronts, including:

- Addressing the shortage of structural and functional data for viral proteins by developing novel high-throughput (HTP) characterization methods and creating new datasets
- Harnessing relevant datasets that are already publicly available, such as protein structures, mass spectrometry data, sequencing information (eg. DNA or RNA sequences), peptide binding assays, protein microarrays, next-generation sequencing (NGS), and immunological data
- Creating techniques to harmonizing multi-omics, structural, and functional datasets
- Applying artificial intelligence and machine learning techniques to create new predictive modeling tools for designing broadly effective antigens

The Problem:

As the APECx program gathers extensive biological and immunological data, including viral DNA sequences, to predict antigens, it highlights the critical role of AI tools in vaccine development. These tools have tremendous potential for research purposes, but open access must be balanced with safety and security concerns stemming from their potential misuse. Additionally, the datasets used to create these toolkits have significant value to the research community. But, without careful planning, it is likely that these datasets could be shared in a way that impedes researchers' ability to use and access them. ARPA-H also seeks to find the optimal solution for making APECx's data and toolkits available, whether by housing it within suitable widely-used platforms, building on existing infrastructure/repositories, or (if necessary) creating a bespoke solution to house both the data and toolkits.

Survey Goals:

- Understand how to **balance open access** for the scientific community with security controls to prevent misuse of the APECx toolkit or datasets
- Understand how researchers would need to **interface** with the APECx tools and datasets

- Determine **key features** for the development of user-friendly interfaces and comprehensive documentation to facilitate maximum adoption of the APECx toolkit and datasets within the scientific community
- Understand the landscape of **existing platforms** that are critical for data generation and manipulation of biological systems related to vaccine and therapeutic development and the understanding of mechanisms of pathogenesis
- Find users/**early adopters**/applications. Learn how users would leverage the future APECx platform and their specific use-cases
- Inform a framework/**architecture** for sharing APECx's data and modeling tools

Instructions:

The Network Survey can be found at the Customer Experience Hub website.
All organizations with relevant experience and expertise are invited to respond, including those that are not members of the Customer Experience Hub network.
The survey is divided into three sections and is estimated to take 30 minutes to complete:

1. Data Management and Interoperability
2. Platform Usability, Features, and Characteristics
3. Platform Security

Respondents may skip sections or questions that are outside their area of expertise.
Response Limit:

- There's a 5,000-character limit for each question. Please be detailed in your responses.
- Please answer each question to the best of your ability. If you're unsure, feel free to skip the question.
- Responses shall be written in English.

Submission:

- You may save your responses as a draft and edit as needed.
- Once the survey is complete, make sure to click Submit.
- Feedback should be provided by Friday, May 17th, 11:59PM CST.

Who Should Respond:

The ARPA-H APECx program is seeking input from the scientific community, specifically (1) **potential users** of the APECx toolkit and datasets, and (2) **experts in data** management, interoperability, storage, and security. Respondents may come from academic, government, non-profit, startup, or commercial sectors, and may answer as individuals or on behalf of their organization. Relevant areas of experience/expertise include:

- Protein engineering
- Functional and structural protein characterization, multi-omics data
- Data management, applications, interoperability, for large datasets
- Developing and architecting platforms to host data
- Data security and management
- Data harmonization, metadata management, and ontology development
- Vaccine and Ag development
- AI and Machine Learning

Note: It is not a requirement to be expert in all (or even the majority) of these areas.

This survey is being administered by the Customer Experience Hub, Advanced Technology International (ATI), part of the Advanced Research Projects Agency for Health (ARPA-H) nationwide health innovation network, ARPANET-H. Contact information provided to complete the survey is confidential and will only be used for survey follow-up. Individual responses to survey questions are confidential and will only be accessible to select personnel in ARPA-H, and ARPA-H's Customer Experience Hub, ATI. Please do not include medical imaging data, patient data, or personally identifiable information of others in your response. Survey findings and key insights will only be presented in an aggregated format with no identifying information. Findings will be made available to all respondents through a public findings document. This survey does not constitute an endorsement of any initiative from ARPA-H or ATI.

For more information about ARPANET-H, the agency's nationwide health innovation network that connects people, innovators, and institutions to accelerate better health outcomes for everyone, visit the ARPANET-H webpage.

Organizations interested in becoming a part of the Customer Experience Hub consortium – or Spoke network – can learn more at the How to Become a Spoke webpage on the Customer Experience Hub website.

Network Survey:

Background & Demographic Information

Organizational Information

- Organization Legal Name*
- Address of Principal Offices*
- City
- State
- Country*
- Zip Code
- Organization Principal Phone Number
- Organization Principal Website*
- Are you already a Customer Experience Hub Spoke? Note that Spoke membership is NOT required to respond to this Network Survey.*
    - Yes or No (check box)

Contact Information

- Contact Name*
- Title*
- Contact Phone Number*
- Contact Email*
- Confirm Contact Email*

How did you hear about this Network Survey?

- CX Hub email
- Forwarded email from someone in my network
- ARPA-H Vitals newsletter
- ARPA-H website

- ARPA-H press release
- CX Hub website
- Social media
- Other (please specify)

Demographics

1. **Institution Type [check box]**
   - Institutions of Higher Education (IHE)
   - For-Profit Organization
   - Non-Profit Organization (excl IHE)
   - Government

2. **Please select your primary and secondary sector/industry [check box]**
   a. Research Organization
   b. Healthcare System/Organization/Provider
   c. Incubator/Accelerator
   d. Manufacturer
   e. Venture Capital/Investment Firm
   f. Entrepreneurial Support/Innovator Training
   g. Network/Association
   h. Consultancy
   i. Community-based Organization (CBO)
   j. Patient advocacy groups
   k. Insurance company/Payer
   l. Clinical or Contract Research Organization (CRO)
   m. Pharmaceutical
   n. Medical device
   o. Biotechnology
   p. Digital Health
   q. Pharmacy/Retail

3. **Please select all designations that apply to your organization [check box]**
   a. Small Business
   b. Startup (incl definition)
   c. Professional associations
   d. Tribally owned organization
   e. Federally Funded Research and Development Centers (FFRDCs)
   f. Women-Owned
   g. LGBTQ-Owned
   h. Minority-Owned
   i. B Corp
   j. Other

4. **(If small business), choose all that apply to your organization [check box]**
   a. Socially or Economically Disadvantaged (8a certification)
   b. Veteran-Owned Small Business (VOSB)
   c. Historically Underutilized Business (HUBZone location)
   d. Service-Disabled Veteran-Owned Small Business (SDVOSB)
   e. Small Disadvantaged Business (SDB)
   f. Economically Disadvantaged Women-Owned Small Businesses (EDWOSB)
   g. Not applicable

Capability Areas:
As a reminder, the Survey is **segmented into three sets of questions**. The questions you see will depend on your selection below of one or more of the categories. Please note that questions become progressively more technical in each section. **All questions are optional**; only answer what is applicable to your expertise. (5,000 character limit per question)

1. **Data Management and Interoperability**
2. **Platform Usability Features and Characteristics**
3. **Platform Security**

Section 1: Data Management and Interoperability

1. What specific research applications do you have for the APECx toolkit or dataset?
2. What potential trade-offs exist when using public clouds for storing biomedical data, particularly concerning data ingress and egress costs?
3. How should the datasets be managed and made available at the program's end to enable the public to further develop and enhance new tools?
4. What tools do you as researchers use to interrogate pathogens and the systems they interact with?
5. What examples/models for sharing data/modeling tools exist that could be used to share the APECx dataset and toolkit? Are any of these models preferred? How do they protect and manage underlying IP?
6. What types of annotations do you recommend for these types of data:
    a. Sequences and variants
    b. Microarray data
    c. Array elements (printed spots or features)
    d. Others
7. How would providing specific ontologies for experimental data benefit AI/ML model development? Alternatively, what is the simplest way for experimentalists to provide descriptions for their data, such as free-text descriptions for sample annotations?
8. What are the primary challenges in acquiring data for training AI to predict consensus antigenicity or single antigen prediction? What data sources/databases do you use? Are there specific challenges associated with these data sources/databases?
9. What are the optimal metrics and methods for standardizing the diverse data types collect by different teams participating in APECx, including genomic, proteomic, structural, and functional characterization data? Additionally, what is the best data reporting format among FASTA, GFF, VCF, and BED?
10. What are the minimal sets of ontologies needed to store omics data, structural data, and experimental data?
    a. Furthermore, what problems could arise if certain requirements are not met or if the reporting format is not consistent?
    b. Various microarray platforms and experimental designs used by APECx will produce data in different formats and units, normalized differently. How can we enable data integration and limit standardization errors in data being generated from multiple researchers in parallel?

11. When allowing researchers to access data including viral structures and assay reporting through API/web services, what specifications, parameters, and types of output should be expected? For example, single multi-FASTA file vs XML file for each sequence ID vs JSON format.

Section 2: Platform Usability Features and Characteristics

A) Usability Requirements and Features:
   1. What are advantages and disadvantages of allowing users to contribute data to the dataset? Why? What controls are needed?
   2. What are crucial features of a database that enable easy retrieval and download of dataset(s) (i.e., keyword search and phase suggestions)?
   3. How can one-click downloads for accessing immunological and biochemical data be compared to other download methods (ex: selection-based), in terms of ease and efficiency?
   4. How can access be provided through various channels (i.e. website, FTP, API) to accommodate users with different preferences or technical requirements needs?
   5. Please provide examples on the best ways data repositories can improve their findability and interoperability and integrate multimodal information (i.e. sequences, structures, etcs.). If possible, please provide examples.

B) Visualization/Bioinformatic Tools:
   6. What visualizations / bioinformatic tools are most useful for Ag design and prediction (ex: phylogenetic analysis)?
   7. How important is the integration of visualization/ bioinformatic tools (i.e. sequence similarity network) for sequence data and structural information? Please expand on the tools that could be used. Expand on the significance of including such tools within a database platform, particularly for users without coding backgrounds who may rely on pre-built visualization features?

C) Metadata Standards:
   8. What common metadata associated with biological sequences such as DNA, RNA, or protein sequences are lacking in databases, and how does this impact usability? (For example, sequence version, sequence features, and alignment information)
   9. What common metadata associated with biochemical assays are lacking in databases, and how does this impact usability? (For example, assay controls, instrumentation, and validation parameters)
   10. What are the primary objectives of community-wide metadata standards, and how are they expected to enhance interoperability in virology research in the near future?

D) Virus-specific Database:
   11. What specific criteria are used to evaluate the FAIR properties of virus databases, and how are these criteria applied?
   12. How do experimental biases related to enrichment methods affect the quality of data in virus databases, and what measures can be taken to mitigate these biases?

Section 3: Platform Security

1. What would be the best architecture of the database that offers security and access permission?
2. What layers of access control, user authentication, and data authentication are optimal to enhance sharing and collaboration but prevent misuse of the APECx toolkit?
3. What measures (if any) can be put in place to monitor and detect attempts to manipulate or misuse the model's outputs?
4. What security measures should be in place to prevent unauthorized access to AI tools and datasets used for antigen prediction, reducing the risk of intellectual property theft or unauthorized replication of proprietary antigen designs?
5. What strategies can be employed to defend against adversarial attacks targeting biased models, particularly for models/tools developed for the APECx program utilizing biochemical and immunological data?
6. How susceptible are AI models for predicting antigen design to algorithmic biases and adversarial manipulations that could result in the creation of harmful or ineffective antigens?
7. What steps can be taken to enhance the robustness and security of AI algorithms used in antigen prediction to prevent malicious tampering or exploitation?
8. What considerations should organizations bear in mind when developing cloud interoperability strategies to ensure seamless communication between disparate cloud environments hosting biomedical data?
9. What risks are associated with data siloing when heavily relying on a specific public cloud provider for storing biomedical data? How are current public databases such as the PDB managed, and what funding mechanisms enable this?
10. What strategies, technologies, or approaches can be employed to verify data integrity and prevent unauthorized alterations?

Thank you for taking the time to complete this Survey!

1. Please let us know if you have any additional comments, questions, or concerns?
2. I agree to receiving communication about the Safe & Secure Biotechnology Platforms
    a. [checkbox] Yes, I would like to receive communications about the Safe & Secure Biotechnology Platforms

End of Survey – Submit

Appendix A:

Understanding the pros and cons of available databases (examples shown below in Table 2) and requirements for developing a database or system of interoperable databases that increase effectiveness and utilization by end users.

Table 2. Example databases according to category.

Knowledge databases:	Genomic sequence databases:	Other -omics databases:	Virus-specific databases:
International Committee on Taxonomy of Viruses (ICTV) ViralZone VIPERdb Virus-Host DB	1. Bacterial and Viral Bioinformatics Resource Center (BV-BRC) 2. NCBI Virus 3. NCBI Viral Genomes 4. Reference Viral Database (RVDB) 5. Virus Orthologous Groups Database (VOGDB) 6. Virxicon 7. ZOVER	1. Integrated Microbial Genomes/Virus (IMG/VR) 2. Multi-omics Portal of Virus Infection (MVIP) 3. Viral Host Range DB	1. Global Initiative on Sharing All Influenza Data (GISAID) 2. COVID-19 Data Portal 3. Coronavirus Antiviral & Resistance Database (COVDB) 4. LANL HIV Database 5. EuResist 6. HIV Drug Resistance DB 7. Hepatitis B Virus database (HBVdb) 8. The Papillomavirus Episteme (PaVE) 9. Virus Variation Resource (NCBI VVR) 10. Proviral Sequence Database (PSD)