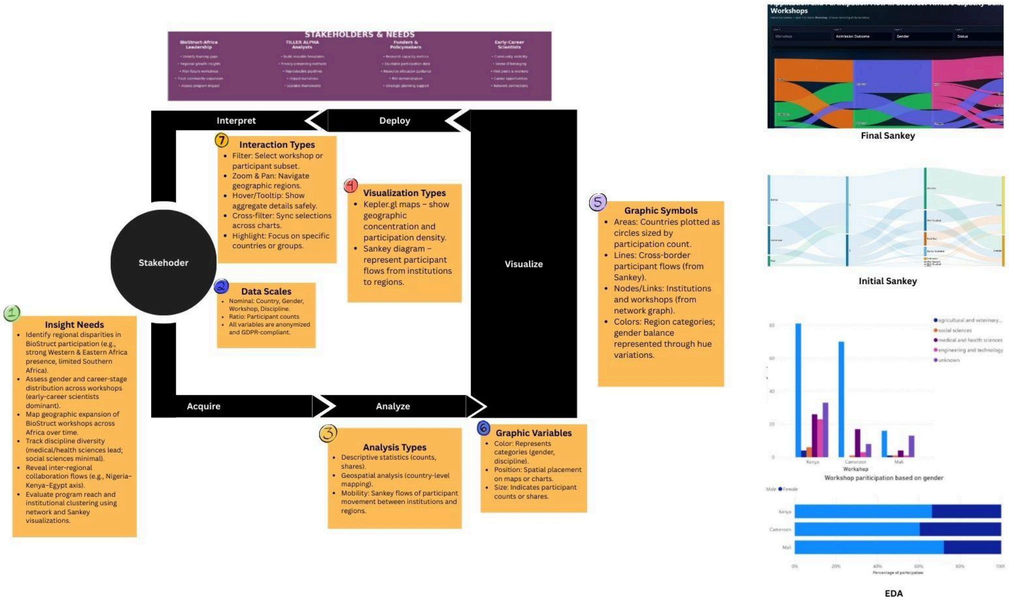
Understanding the Dynamics of the Emerging Community of Structural Biologists in Africa (BioStruct)

Team Members: Mohammed Fahad, Sushil A, Pravin Raj, Vimal A, Athish G Affiliation: Indiana University – ENGR-E483/E583 Information Visualization Sponsors: Katharina Cramer (TILLER ALPHA), Emmanuel Nji (BioStruct-Africa)



STAKEHOLDERS & NEEDS

BioStruct-Africa Leadership
- Identify training gaps
- Regional growth insights
- Plan future workshops
- Track community expansion
- Assess program impact

TILLER ALPHA Analysts
- Build reusable templates
- Privacy-preserving methods
- Reproducible pipelines
- Impact narratives
- Scalable frameworks

Funders & Policymakers
- Research capacity metrics
- Equitable participation data
- Resource allocation guidance
- ROI demonstration
- Strategic planning support

Early-Career Scientists
- Community visibility
- Sense of belonging
- Find peers & mentors
- Career opportunities
- Network connections

Insight Needs
- Identify regional disparities in BioStruct participation (e.g., strong Western & Eastern Africa presence, limited Southern Africa).
- Assess gender and career-stage distribution across workshops (early-career scientists dominant).
- Map geographic expansion of BioStruct workshops across Africa over time.
- Track discipline diversity (medical/health sciences lead; social sciences minimal).
- Reveal inter-regional collaboration flows (e.g., Nigeria–Kenya–Egypt axis).
- Evaluate program reach and institutional clustering using network and Sankey visualizations.

Data Scales
- Nominal: Country, Gender, Workshop, Discipline.
- Ratio: Participant counts
- All variables are anonymized and GDPR-compliant.

Analysis Types
- Descriptive statistics (counts, shares).
- Geospatial analysis (country-level mapping).
- Mobility: Sankey flows of participant movement between institutions and regions.

Visualization Types
- Kepler.gl maps – show geographic concentration and participation density.
- Sankey diagram – represent participant flows from institutions to regions.

Graphic Symbols
- Areas: Countries plotted as circles sized by participation count.
- Lines: Cross-border participant flows (from Sankey).
- Nodes/Links: Institutions and workshops (from network graph).
- Colors: Region categories; gender balance represented through hue variations.

Graphic Variables
- Color: Represents categories (gender, discipline).
- Position: Spatial placement on maps or charts.
- Size: Indicates participant counts or shares.

Interaction Types
- Filter: Select workshop or participant subset.
- Zoom & Pan: Navigate geographic regions.
- Hover/Tooltip: Show aggregate details safely.
- Cross-filter: Sync selections across charts.
- Highlight: Focus on specific countries or groups.

Fig. 1. Visual Abstract – Data Visualization Logic (DVL) Framework for BioStruct-Africa. This figure summarizes the visual analytics workflow developed for the BioStruct-Africa project, illustrating how stakeholder needs are translated into data-driven insights and visual representations.

1. INTRODUCTION AND PRIOR WORK

Structural biology enables understanding of disease mechanisms and the design of new treatments. Yet, African researchers contribute to less than 8% of global scientific output (Nji et al., 2025). BioStruct-Africa—a non-profit registered in Sweden and Ghana—addresses this gap by training African scientists in advanced computational tools such as AlphaFold, cryo-electron microscopy, and X-ray crystallography.TILLER ALPHA, a European analytics company, collaborates with BioStruct-Africa to analyze and visualize data from training workshops. The project aims to create interactive, human-centered visualizations that reveal how Africa's community of structural biologists is growing and connecting.

Existing visualizations of scientific collaboration rely heavily on publication data (Dosso et al., 2022; Pan et al., 2021), which excludes early-career researchers. Our project introduces new ways of representing growth, participation, and inclusivity—emphasizing empathy and community representation alongside scientific rigor.

1.1 Stakeholder Groups

The BioStruct-Africa leadership is interested in understanding regional growth and identifying training gaps across the network. TILLER ALPHA analysts are focused on building reusable, privacy-preserving visualization templates that can be applied across similar datasets. Funding agencies and policymakers need clear insights on research capacity and equitable participation to inform their decisions and resource allocation. Early-career scientists are seeking belonging and visibility within the growing African research network, looking for ways to connect and establish their presence in the community.

1.2 Stakeholder Needs

Stakeholders—including BioStruct-Africa program directors, training coordinators, funders, and partner institutions—require analytical tools that allow them to clearly understand the evolving landscape of workshop participation across Africa. One major need is the ability to identify regional disparities in engagement, particularly the strong presence of applicants from Western and Eastern

Africa compared to the more limited representation from Southern African countries. Stakeholders also emphasized the importance of assessing gender balance and career-stage composition across workshops, as preliminary observations suggest that early-career scientists, especially Pre-Doctoral and PhD students, constitute the majority of applicants. Understanding whether this demographic profile is consistent across years and regions is essential for guiding targeted outreach and capacity-building strategies.

Another key requirement is the capacity to track geographic expansion of BioStruct-Africa's training programs over time. As the organization continues to broaden its reach, stakeholders want tools that reveal how participation patterns shift with each new workshop location and how effectively the program penetrates different regions. This extends to examining disciplinary diversity, where Natural Sciences—particularly biomedical and molecular fields—currently dominate, while disciplines such as social sciences and computational subfields appear only marginally. Stakeholders need visual analytics that can surface such imbalances and help guide future interdisciplinary recruitment.

Furthermore, stakeholders are interested in understanding inter-regional collaboration flows, such as recurring participation links between countries like Nigeria, Kenya, and Egypt. These flows offer insight into existing research ecosystems, regional scientific mobility, and potential areas for strengthening multi-country collaborations. They also seek tools that evaluate overall program reach and institutional clustering, using methods such as Sankey flows and network visualizations to illuminate how applicants move across workshops, how institutions are represented, and where capacity-building efforts are most effective. Together, these needs highlight the importance of a comprehensive, intuitive visualization system that supports strategic planning, diversity monitoring, and long-term evaluation of BioStruct-Africa's impact.

2. DATA ACQUISITION

2.1 Data Sources

The primary dataset is an Excel workbook provided by TILLER ALPHA and BioStruct-Africa, containing two sheets labeled raw_data and metadata. This extract encompasses 308 anonymized applications submitted to three BioStruct-Africa workshops held in Kenya with 173 applications, Cameroon with 99 applications, and Mali with 36 applications, with each application labeled according to its selection outcome.

The dataset contains no direct identifiers, and all processing procedures strictly comply with GDPR and institutional ethics guidelines. Results are reported exclusively in aggregated form with minimum cell-size thresholds set between 3 and 5 observations, supplemented by categorical roll-ups such as "Other countries/regions" to prevent potential re-identification of individuals. The core variables captured include a pseudonymous ID for each applicant, workshop location, selection outcome, gender, career stage, scientific discipline recorded as free text, primary institutional affiliation, and country of primary affiliation. An auxiliary reference file,
United Nations Statistics Division (UNSD) data which is maintained by the UN and is one of the most widely used global standards for country names, geographic regions, and statistical groupings.

The official Research Organization Registry (ROR) dataset is an openly maintained global registry of research institutions., integrated to enhance institutional disambiguation. Each affiliation was cross-checked against ROR identifiers, improving global interoperability and ensuring one-to-one mapping between textual names and persistent institutional IDs.

The OECD Schema to Web of Science Category Mapping (2022) is a standardized crosswalk created by the Organisation for Economic Co-operation and Development (OECD) to align scientific fields across multiple classification systems. OECD data Provides the discipline harmonization framework linking workshop "Scientific Discipline" free-text entries to standardized OECD major fields via Web of Science categories. This enabled classification consistency and comparability across disciplines

These reference datasets were integrated into the data pipeline to enrich country and institutional attributes before analysis and visualization.

2.2 Data Description, Quality and Coverage

The primary dataset contains one row per applicant across the BioStruct-Africa workshops in Mali (2022), Cameroon (2024), and Kenya (2025). The raw dataset originally contained inconsistent formatting, duplicated category labels, misspellings (e.g., "Eygpt," "Pre-," "Phd student"), and multiple representations of the same disciplines. Our team performed a complete data cleaning and standardization process to resolve these issues, ensuring uniformity across all variables and enabling reliable visualization.

The dataset includes several key columns. "Workshop Location" identifies the specific event an applicant applied to, forming the foundational node structure for the Sankey diagram. "Country of Affiliation" records the applicant's reported country and serves as the geographic source in the Kepler flow map. "Gender" captures the applicant's demographic identity. "Admission Outcome" differentiates between *selected* and *rejected* applicants; although rejection appears in the data as a binary status, it primarily reflects limited workshop capacity rather than applicant quality. The "Career Status" field categorizes applicants into stages such as *Pre-Doc*, *PhD Student*, *Post-Doc*, *Faculty*, and *Research Scientist*. During cleaning, our team corrected capitalization differences, merged redundant categories, and standardized terminology to ensure accurate cohort comparisons.

"Field of Research" contains the harmonized scientific discipline derived by clustering applicant-reported fields using mapping references such as the OECD to Web of

Science schema. The raw dataset included dozens of highly granular or inconsistently named research areas; our team consolidated these into a coherent, reproducible set of discipline clusters that align with international scientific taxonomies. The data also contains unique applicant identifiers, which we used to validate row uniqueness, remove duplicates, and guarantee one-to-one correspondence across analytical steps.

For geospatial mapping, our team integrated country-level latitude and longitude centroids, generating *origin_lat* and *origin_lng* fields used in Kepler.gl to visualize origin-to-workshop flows. Workshop coordinates (*dest_lat*, *dest_lng*) were similarly standardized. As part of our quality improvements, we validated all country names against an external reference (UNSD country classifications) to ensure proper joins during map construction.

Despite initial inconsistencies, the cleaned dataset now exhibits high overall coverage and reliability across all key attributes. Privacy-sensitive fields, such as institutional affiliation, were intentionally excluded to maintain GDPR compliance. Some applicants reported multiple affiliations or disciplinary areas; however, to preserve interpretability, the dataset retains only each applicant's primary country and primary field designation. The standardized structure produced through our cleaning efforts ensures strong analytical validity and supports the construction of clear, interpretable visualizations.

2.3. Data Cleaning

We did advanced text-normalization and multi-stage fuzzy-matching pipeline designed to clean and standardize institutional and regional data. The approach begins with comprehensive normalization that includes lowercasing, accent removal using unidecode, and the elimination of special symbols. Importantly, the pipeline deliberately avoids aggressive suffix removal of terms like "university" or "college" to preserve meaningful distinctions between institutions.

A key innovation is the introduction of a local alias dictionary that correctly links known variations of institutional names, such as mapping "OAU" to "Obafemi Awolowo University". The fuzzy matching operates through six sequential stages, starting with alias resolution and progressing through exact matching, token sort ratio for word-order insensitive comparisons, token set ratio for substring awareness, partial ratio as a fallback, and finally a global fallback search across all institutions when no closer match is found.

Matching record receives three critical fields: u_match_name for the standardized institution name, match_method indicating which stage produced the match, and match_confidence providing a quality grade from A to based on similarity thresholds, where grade A represents, matches with 95% or higher confidence, grade C represents85% or higher, and so forth. This refinement successfully addressed inconsistencies related to spacing and punctuation while maintaining the distinct identity of each institution. However, the process also revealed edge cases requiring further attention, such as distinguishing between Cairo University and American University in Cairo, which will be addressed through manual alias corrections in subsequent iterations. The final processing version integrates the fuzzy-matched institutions with verified regional mappings from the UNSD methodology dataset, ensuring global consistency and removing dependency on ISO2 codes.

2.3.1 Region Enrichment

Using the UNSD Dataset, each country was systematically mapped to a specific subregion within Africa to replace vague geographic classifications. Generic "Africa" entries were disaggregated into five distinct subregions: Northern Africa, Western Africa, Middle Africa, Eastern Africa, and Southern Africa. This granular mapping provides more precise geographic context for institutional analysis. Additionally, a secondary validation script was implemented to address ISO2 code inconsistencies, ensuring that countries such as Ivory Coast, Botswana, and The Netherlands were correctly remapped to their appropriate subregions after initial misclassifications were identified.

2.3.2 Country and Institution Mapping

To enhance data quality, a comprehensive correction process was implemented across multiple dimensions. Common typographical errors were systematically identified and corrected, with frequent misspellings such as "Egyp", "Eygpt", and "Egpyt" all standardized to "Egypt". Country name harmonization addressed linguistic and formatting variants, converting entries like "Côte d'Ivoire" to the standardized form "Ivory Coast" for consistency. Additionally, institutional name typos were automatically corrected when they demonstrated 98% or higher similarity with reference universities in the master dataset, ensuring that minor spelling variations did not fragment what should be recognized as the same institution.

2.3.3 Major Scientific Discipline Calculation

The calculation of the major scientific discipline was performed through a structured, reproducible mapping pipeline implemented inside the project's Jupyter notebook (data_cleansing_and_eda_v3.ipynb). The raw dataset contained free-text discipline responses such as “biochem,” “structural biology,” “drug discovery,” “public health,” and similar variations, many of which were inconsistent or ambiguous. To convert these heterogeneous entries into standardized analytical categories, the notebook first normalized all text—lowercasing, removing punctuation, and trimming noise tokens—to enable reliable matching. Each cleaned value was then compared against an externally validated classification system derived from the OECD Fields of Science Schema, which was cross-referenced with the Web of Science Category Mapping (2022) to ensure contemporary scientific alignment.

Using this combined schema, the notebook applied a rule-based system paired with keyword detection to map each applicant's free-text discipline into one of the consolidated reporting categories: Natural Sciences, Medical & Health Sciences, Engineering & Technology, Agricultural & Veterinary Sciences, Social Sciences, or Unknown (when insufficient information was available). This mapping is fully deterministic and reproducible, ensuring that identical inputs always resolve to the same high-level discipline. By aggregating these mapped categories, the analysis enabled cross-workshop comparisons of disciplinary composition and revealed meaningful trends—such as the dominance of Natural Sciences across all workshops and the relative concentration of Agricultural and Social Sciences within the Mali cohort. These standardized discipline groups were also critical for building clean, interpretable Sankey flows and workshop-distribution comparisons.

3. DATA ANALYSIS

The analytical workflow was implemented in a fully reproducible Jupyter notebook supported by a Power BI file for additional exploratory checks and stakeholder-ready summaries. The notebook handles the full data standardization pipeline: normalizing country names, harmonizing career-stage labels, and mapping free-text discipline responses into the standardized major scientific fields described earlier. Missing values are consistently retained as "Unknown" rather than dropped, preserving analytical transparency and ensuring that filters and Sankey layers reflect the real completeness of the submitted data.

For in depth review check our repo:prasengu/E583-Biostruct: This repository is created for client project discussion and visualization creation for biostruct Africa

The analysis spans both *applications* and *selected participants*, allowing comparison between the full applicant pool and the final workshop cohorts. Descriptive summaries quantify distributions across workshop, gender, country, discipline, and career stage. While the dataset does not yet contain precise application dates, workshop ordering provides a coarse time dimension; once temporal attributes are added, the same codebase will automatically generate year-over-year and cumulative trend series. Geographic analysis operates at the country level to characterize regional origins and cross-border reach, supporting the geospatial arc visualizations used later in Kepler.gl. A lightweight network specification complements this analysis: country nodes are connected through weighted edges representing co-attendance or recurrent flows across workshops, highlighting clusters of regional scientific mobility.

The dataset was further enriched with UNSD regional classification, ensuring global statistical consistency and enabling region-level aggregation for Sankey and mobility-flow visuals. All intermediate transformations are logged to ensure GDPR compliance and analytical reproducibility.

The results of the analysis reveal several notable patterns. First, participation is highly concentrated in Western and Eastern Africa, with Nigeria, Cameroon, Kenya, Ghana, Sudan, Egypt, and Uganda serving as prominent contributors. The gender distribution demonstrates a clear imbalance, with male applicants comprising approximately 65% of the pool and female applicants only 35%, a pattern consistently visible across workshops and reproduced in Sankey flows where male streams are visibly thicker

Kenya attracts the largest number of women overall, though none of the workshops come close to achieving gender parity.

Disciplinary analysis shows that Natural Sciences dominate across all three workshops, consistent with the structural-biology focus of the initiative. Kenya exhibits the strongest concentration in these fields, whereas Cameroon displays a more balanced profile with substantial representation in Medical & Health Sciences. Mali, although smaller in total applicants, includes proportionally more individuals from Agricultural, Veterinary, and Social Science fields, suggesting workshop-specific appeal that differs from the other two hosts.

Career-stage analysis indicates that Pre-Docs and PhD students form the majority of applicants, but senior scientists and postdoctoral researchers also participate in meaningful numbers, illustrating BioStruct-Africa's appeal across academic stages. When combined, these results gender distribution, disciplinary profiles, regional engagement, and workshop-specific patterns provide a rich, multi-perspective understanding of the structural biology community emerging across Africa.

4. VISUALIZATIONS

4.1 Geo Maps Visualization (kepler.gl)

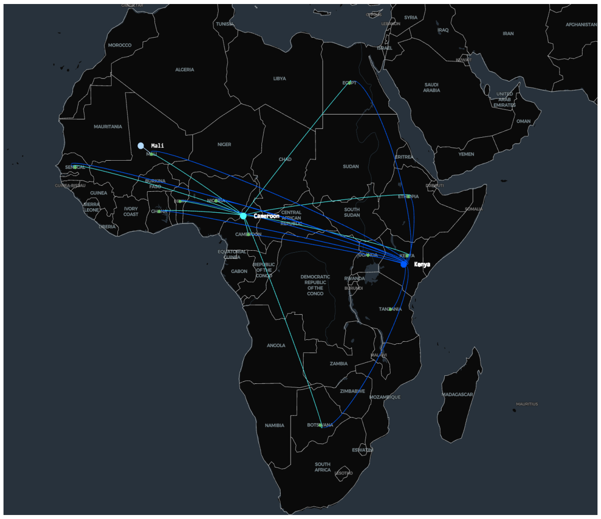
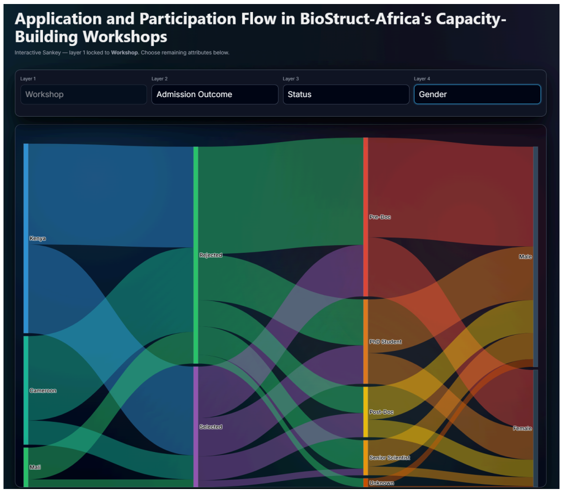


Fig. 3. Kepler diagram representing country workshops conducted by BioStruct Africa

To extend beyond static network views, a Kepler.gl visualization was introduced to capture the geographic spread and distribution of BioStruct workshops and participants across the African continent. The Kepler.gl layer is now generated from cleaned data using centroid coordinates from an integrated regional lookup aligned with UNSD regions.

This ensures proper spatial grouping for all African countries and compatibility with global datasets. The data preparation phase involved associating each country with its centroid coordinates, including latitude and longitude values, while aggregated metrics such as the number of participants and workshops per country were computed from the enriched dataset. The visualization incorporates multiple interactive layers to convey different dimensions of the data: a point layer represents each African country with circles whose size corresponds to participant count and color denotes the specific region, a heatmap layer provides an intensity overlay to highlight areas of regional concentration, and an optional time filter enables temporal analysis to show participation growth trends from 2017 through 2025 when year data is available. This geographic visualization reveals important patterns, demonstrating that Western and Eastern Africa exhibit the highest workshop participation density, with Lagos, Nairobi, and Cairo emerging as recurring centers for structural biology training across the continent.

4.2 Sankey Visualizations



Application and Participation Flow in BioStruct-Africa's Capacity-Building Workshops

Interactive Sankey — layer 1 locked to **Workshop**. Choose remaining attributes below.

Layer 1: Workshop

Layer 2: Admission Outcome

Layer 3: Status

Layer 4: Gender

Fig. 4. Sankey Diagram representing workshops conducted by BioStruct Africa

A Sankey diagram was developed to visualize the flow of participants between their home institutions and workshop regions, providing an intuitive depiction of cross-regional collaboration and mobility patterns within the BioStruct network. The diagram was constructed using data from fuzzy matched file with nodes representing both institutions and regions, while links illustrate the number of participants

transitioning from a given institution to a specific workshop region. The Sankey data were derived from the enriched file and regional mappings validated via UNSD methodology, ensuring accurate regional labeling and consistent cross-region flow representation. The visualization was implemented using Plotly and exported as an interactive HTML file for broader accessibility. Several planned variants are under consideration to provide different analytical perspectives, including flows from country to workshop, institution to region, and discipline to workshop. Early insights from this visualization reveal that institutions in Western Africa send the highest number of participants to cross-regional workshops, with particularly strong flows directed toward Eastern African hubs such as Kenya and Tanzania, highlighting the importance of these locations as regional training centers.

5. USAGE AND CRITIQUE OF AI TOOL

AI tools (e.g., ChatGPT for design ideation, Explain Data, ArcGIS AI layers) can aid clustering and interpretation, but outputs are used cautiously due to risks of bias, misrepresentation of underrepresented regions, privacy breach.

6. INTERPRETATION OF RESULTS

The combined analytic and visualization pipeline reveals a series of clear and meaningful insights into the composition, diversity, and dynamics of BioStruct-Africa's workshop applicants. Taken together, the Sankey and geospatial analyses provide a coherent picture of who applies, who participates, and how the emerging structural biology community is distributed across regions and scientific fields.

6.1 GEOGRAPHIC PARTICIPATION

Across all three workshops, applicants originate from a wide geographic footprint spanning Western, Eastern, Northern, Central, and Southern Africa, with a smaller number from Europe, the Middle East, and Asia. The geospatial arc visualization highlights Nigeria, Cameroon, Ghana, Kenya, Sudan, Egypt, and Uganda as the most prominent contributors. These countries consistently show strong outbound arcs toward all workshops, with Cameroon and Kenya in particular drawing the broadest regional engagement. The Mali workshop shows more modest geographic reach, yet still demonstrates attractivity from several West African countries. Collectively, these patterns reveal BioStruct-Africa's success in reaching a scientifically diverse pan-African audience.

6.2 GENDER DISTRIBUTION

Gender analysis indicates a persistent imbalance across the applicant pool, with male applicants comprising approximately two-thirds of all submissions, and females representing the remaining one-third. This distribution appears consistently across workshops and is clearly visible in the Sankey diagram, where male flows dominate across both selected and rejected categories. Kenya attracts the largest number of female applicants in absolute terms, but none of the workshops reach gender parity. This finding underscores the need for continued efforts to promote gender diversity in structural biology training and research participation.

6.3 ADMISSION OUTCOMES

Admission outcomes reveal strong workshop-specific differences. While the majority of applicants are categorized as "Rejected," the client clarified that these rejections are largely capacity-driven rather than performance-based. The

Kenya workshop, which received the highest number of applications, shows the steepest drop from application to acceptance. Cameroon displays a more balanced acceptance profile, whereas Mali due to its smaller applicant pool exhibits more proportional selection rates. These distinctions suggest that workshop size, logistical constraints, and regional interest jointly drive selection patterns.

6.4 CAREER STAGE COMPOSITION

Career stage analysis shows that BioStruct-Africa attracts an academically diverse population. Pre-Doctoral researchers and PhD students form the largest share of applicants, closely aligning with the organization's training mission. Postdoctoral researchers and senior scientists also appear in meaningful numbers, with Cameroon and Kenya drawing the highest concentrations of senior-stage applicants. This distribution highlights the workshops' dual appeal: they serve both as foundational training experiences for early-career researchers and as collaborative networking opportunities for established scientists.

6.5 SCIENTIFIC DISCIPLINE TRENDS

The structured discipline-mapping process reveals that Natural Sciences overwhelmingly dominate the applicant pool, particularly in molecular biology, biochemistry, structural biology, and related biomedical sciences. This trend appears consistently across all workshops and aligns with BioStruct-Africa's core focus areas. Cameroon exhibits a more heterogeneous disciplinary mix compared to Kenya, which shows a stronger concentration in the Natural Sciences. Mali shows proportionally higher representation in Agricultural, Veterinary, and Social Sciences, suggesting that workshop themes and regional expertise may influence disciplinary diversity.

6.6 INTEGRATED INTERPRETATION

Together, these results illustrate a rapidly growing and geographically diverse community of early-career and senior researchers engaging with BioStruct-Africa's training ecosystem. The gender imbalance highlights an important equity challenge, while the disciplinary and regional results demonstrate strong scientific relevance and broad continental demand. The Sankey and Kepler visualizations jointly reveal that workshop selection patterns, research backgrounds, and cross-border mobility all contribute to a rich and evolving landscape of scientific capacity-building in Africa. These insights provide a powerful foundation for future planning, outreach strategies, and targeted program expansion.

7. EVALUATION

7.1 Problems Identified During Evaluation

During the initial evaluation phase, several data and visualization challenges were identified through peer and client feedback. Many records contained free-text affiliations with variations and abbreviations such as "Univ Lagos", "Lagos State Univ.", and "LASU", which caused redundancy and incorrect aggregation in institutional visualizations.
The initial dataset included entries labeled simply as "Africa," making regional comparisons impossible and limiting the ability to assess cross-regional participation trends. Free-text scientific disciplines were inconsistent or too granular to be compared across workshops, making it difficult to quantify interdisciplinary participation or training diversity. Early network visualizations revealed institutional and country linkages but lacked spatial context, prompting stakeholders to request an interactive map to visualize participation growth and workshop locations geographically. Additionally, expanding the dataset to include new sources such as ROR, OECD, and UNSD introduced schema mismatches and required complex harmonization efforts.

7.2 Redesign and Resolution

The redesign addressed each issue systematically through both data enrichment and visualization refinement. A multi-stage fuzzy-matching pipeline was implemented and combined with the Research Organization Registry (ROR) dataset to ensure consistent and globally recognized institutional identifiers, significantly reducing duplication errors. The UNSD Methodology was adopted to replace generic "Africa" tags with specific subregions such as Western, Eastern, and Southern Africa, improving the accuracy of regional roll-ups and geospatial comparability. Free-text disciplines were mapped to standardized OECD major fields using the "OECD Schema to Web of Science Category Mapping (2022)" dataset, enabling consistent analysis of disciplinary representation across workshops. New visualizations were introduced, including Kepler.gl for interactive, region-based geospatial mapping and Sankey diagrams to show participant flow between institutions and regions, which enhanced storytelling and interpretability for stakeholders. All enrichment steps were consolidated into a unified file, improving reproducibility and scalability for future datasets.

8. CHALLENGES AND LIMITATIONS

Despite the significant improvements achieved through extensive data cleaning, harmonization, and visualization refinement, several challenges and limitations remain. First, the dataset contained institution names with limited contextual information, making high-confidence matching difficult and often requiring manual review or alias correction. In addition, the metadata lacked temporal detail; because applicant data were not tagged by year beyond their workshop association, it was not possible to conduct longitudinal or trend-based analyses across cohorts.

9. FUTURE WORK

This project lays the groundwork for continued data integration and analytics development, with several promising future directions identified. Building an automated enrichment pipeline that regularly syncs with ROR and OECD taxonomies would streamline data processing and ensure continuous updates. Integrating year-based metadata would enable visualization of the evolution of BioStruct workshops over time in Kepler.gl, providing valuable longitudinal insights. Combining participant-level metadata with workshop co-attendance data could infer collaboration strength and network centrality, offering enhanced collaboration mapping capabilities. Developing an interactive web dashboard that combines network, Sankey, and geospatial views would provide stakeholders and donors with comprehensive analytical tools in a single interface. Finally, the standardized design established in this project allows for future integration of workshops outside Africa, enabling global benchmarking and comparative analysis across continents.

10. ACKNOWLEDGEMENTS

We thank our sponsors Katharina Cramer (TILLER ALPHA) and Emmanuel Nji (BioStruct-Africa) for their time and feedback. We also acknowledge Indiana University's ENGR-E483/E583 faculty and TAs for their guidance.

11. REFERENCES

Code Repository : prasengu/E583-Biostruct: This repository is created for client project discussion and visualization creation for biostruct Africa

1. BioStruct-Africa & TILLER ALPHA GmbH. (2025). *cleansed_biostruct_data_v4.csv* [Dataset]. Provided to the project team for workshop applicant analysis.

2. United Nations Statistics Division (UNSD). (2023). *Standard Country or Area Codes for Statistical Use (M49).*
https://unstats.un.org/unsd/methodology/m49/

3. Research Organization Registry (ROR). (2025). *ROR Data Dump, Version 1.73 (2025-10-28).*
Available at: https://ror.org/data/

4. Organisation for Economic Co-operation and Development (OECD). (2022). *OECD Fields of Science and Technology (FOS) Classification Schema and Web of Science Category Mapping.*
Retrieved from: https://www.oecd.org/sti/inno/38235147.pdf

5. Natural Earth. (2024). *Countries GeoJSON – Admin 0 – Countries.*
https://www.naturalearthdata.com/downloads/110m-cultural-vectors/

6. Kepler.gl. (2024). *Open-source geospatial analysis tool.*
Available at: https://kepler.gl/

7. D3.js. (2024). *Data-Driven Documents – Sankey Layout Module.*
Available at: https://github.com/d3/d3-sankey