INDIANA UNIVERSITY
BLOOMINGTON

HANDOVER DOCUMENT

Project Title:
Understanding the Dynamics of the Emerging Community of Structural Biologists in Africa (BioStruct-Africa)

Authors:

1. Mohammad Fahad Shahul Hameed
2. Pravin Raj Senguttuvan
3. Athish GR
4. Sushil Amalan John Moses
5. Vimal Aditya Raj

Project Summary

This project delivers an interactive visual analytics system to analyze applicant and participant data from BioStruct-Africa's capacity-building workshops conducted in Mali (2022), Cameroon (2024), and Kenya (2025). The system was designed to support strategic program evaluation by enabling exploration of geographic reach, disciplinary composition, gender distribution, career stages, and workshop-specific demand patterns.

Two primary visualization components were implemented:

1. **A multi-layer interactive Sankey diagram for categorical flow analysis**
2. **A geospatial flow map using Kepler.gl to visualize applicant origins and mobility patterns across Africa**

The project strictly adheres to GDPR constraints, avoids public sharing, and uses only privacy-compliant, aggregated representations.

Data Overview and Governance

The core dataset (cleansed_biostruct_data_v4.csv) was provided by the project sponsors after initial discussions, with the following guarantees:

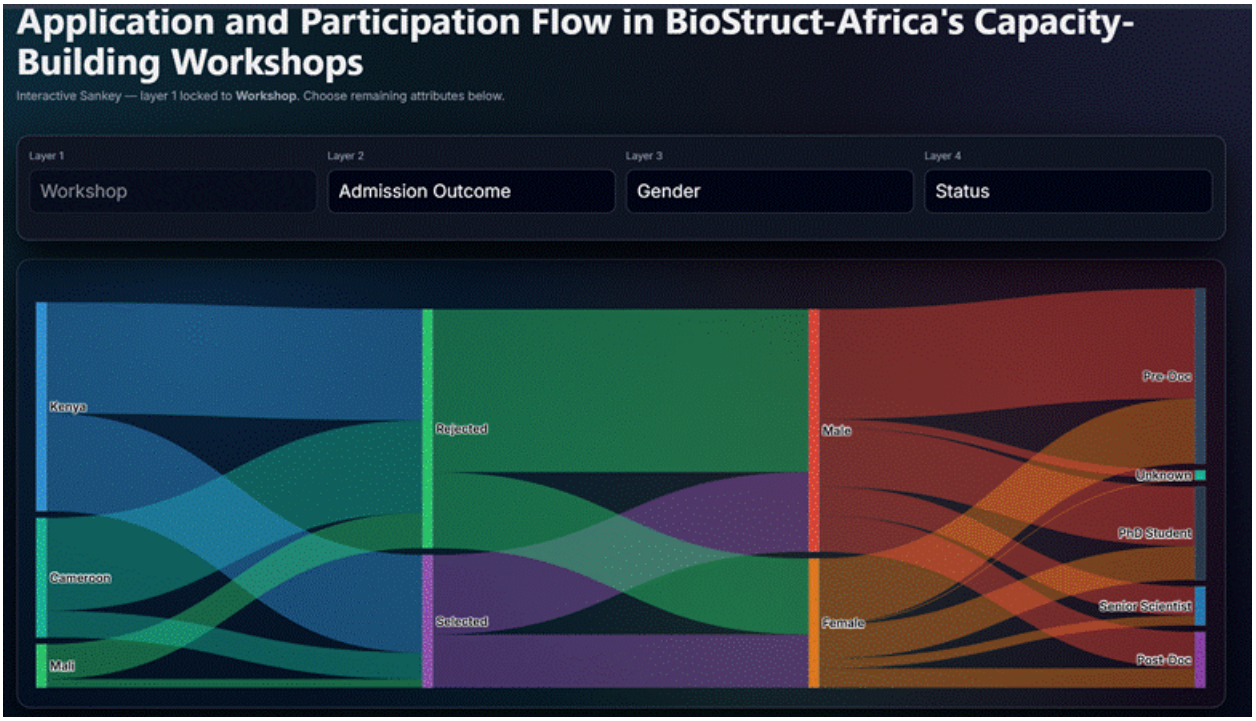
- Column structure was fixed and stable
- Dataset size ~200 applicants (170 confirmed, ~20 pending at time of analysis)
- Primary institutional affiliation removed for GDPR compliance
- Creative freedom granted for visualization design

Important:
All data standardization, cleaning, harmonization, and analytical restructuring were performed entirely by the project team, including:

- **Capitalization normalization (e.g., *PhD / Phd / phd → PhD Student*)**
- **Country name correction (e.g., *"Egyp" → Egypt*)**
- **Scientific discipline clustering into meaningful reporting categories**
- **Removal of redundant and low-signal attributes**

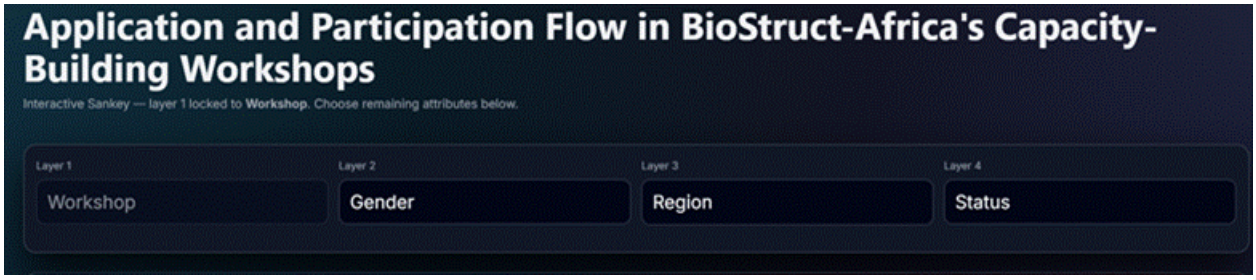
Visualization Architecture

A. Sankey Diagram (Flow Analysis)



The Sankey diagram represents applicant flows across four configurable analytical layers. Each layer can be dynamically reassigned using dropdown controls, allowing users to reorder how attributes are compared.

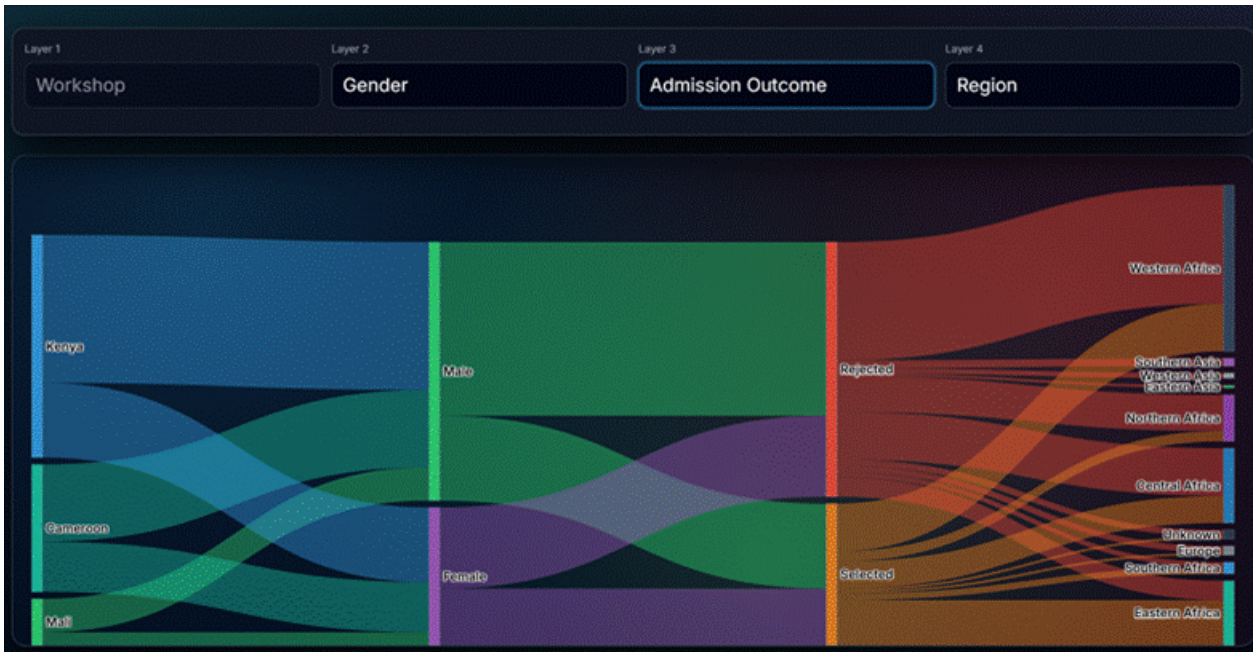
Layer Structure



- Layer 1: Workshop (Kenya, Cameroon, Mali)
- Layer 2: Admission Outcome / Gender / Status / Field of Research
- Layer 3: Gender / Status / Field of Research
- Layer 4: Status or Field of Research (depending on configuration)

This structure enables analysts to examine how applicant composition changes depending on which attribute is foregrounded.

Color Encoding



Each workshop is assigned to a consistent color family that persists across all layers. This allows users to visually trace cohorts originating from a specific workshop through multiple categorical splits without ambiguity. Color consistency was chosen to reduce cognitive load and improve path traceability.

Flow Thickness

Ribbon thickness represents absolute applicant counts. Percentages displayed in tooltips (Male:65%) are derived as:

$$\textbf{Percentage} = \frac{\textbf{Count for attribute}}{\textbf{Total applicants in current selection}} \times 100$$

For example, “Male (65%)” reflects
(Number of male applicants ÷ total applicants) × 100.

Sankey Interactions and Advanced Features

Hover Tooltips (Scrollable)

Hovering over a node or flow activates a persistent tooltip displaying:

- Absolute counts
- Percentage contribution
- Attribute breakdown across all visible layers

If the tooltip content exceeds the visible area, it supports mouse-based scrolling.

Click-to-Highlight Path Tracing

Clicking any node or flow highlights the entire connected path from source to destination across all layers, while dimming unrelated flows. This enables precise inspection of specific applicant cohorts (e.g., *Kenya → Selected → Female → PhD Student*).

Clicking on empty space resets the view.

Clicking on any empty area of the Sankey canvas resets the visualization to its default state. This action clears all active flow selections, removes any highlighted paths, and restores the full set of connections across all categories. It allows users to quickly return to the global overview after exploring a specific pathway or subset of the data, without needing to reload the visualization or manually undo interactions.

Common Questions About the Sankey

Q: I clicked something and now I can't see anything else! Help! A: Click on the empty gray space around the diagram. Everything will come back.

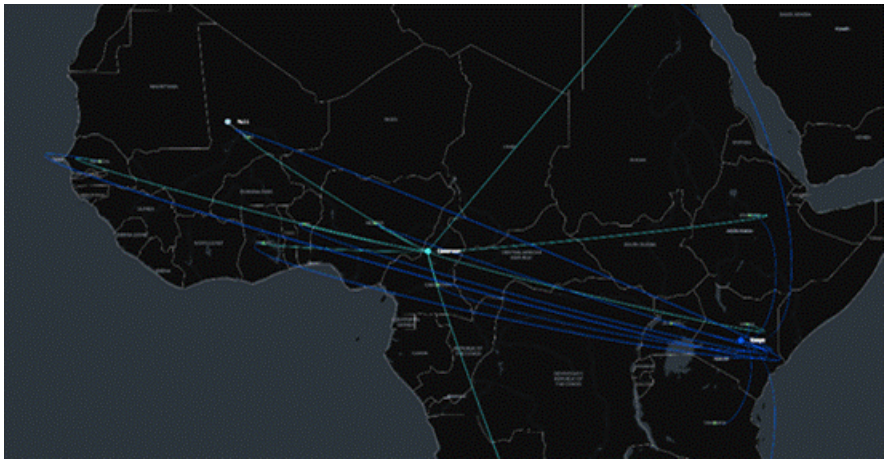
Q: The white tooltip box disappeared before I finished reading it! A: Move your mouse INTO the white box. It will stay visible while you're inside it. You can even scroll it if it's long!

Why are some ribbons so thin I can barely see them? A: That means very few people have that specific combination. For example, "Male + Senior Scientist + Chemistry + Southern Africa" might only be 1-2 people.

Q: Can I see just females? A: Not directly with a click, but YES with the dropdowns! Set Layer 2 to "Gender" and watch the split happen early in the flow.

Q: I want to print this or save it. How? A: Right-click on the diagram and choose "Save image as" or "Print." The colors will print nicely!

B. Kepler.gl Geospatial Visualization

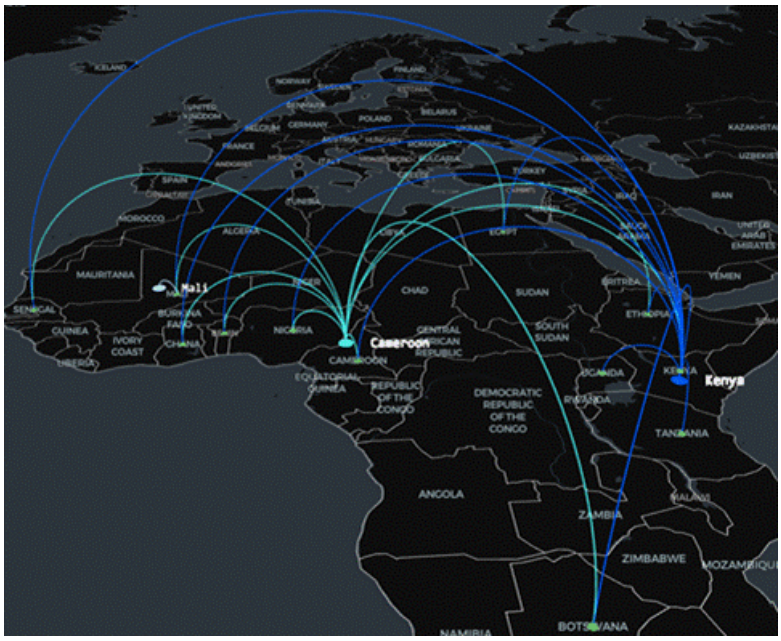


The Kepler map visualizes country-to-workshop flows using arc layers, where:

- Each arc represents applicants from one country to one workshop
- Arc thickness encodes applicant volume
- Arc color encodes destination workshop
- Workshop locations are shown as fixed destination points

Country coordinates were derived from a standardized country-centroid reference file to ensure consistent spatial placement. GEO JSON CODED FILE

Interaction Capabilities



Users can:

- Hover over arcs to view origin, destination, and applicant counts
- Hover over countries to view total applicants and workshop distribution
- Toggle workshop layers on/off to reduce clutter
- Zoom and pan for regional or continental analysis

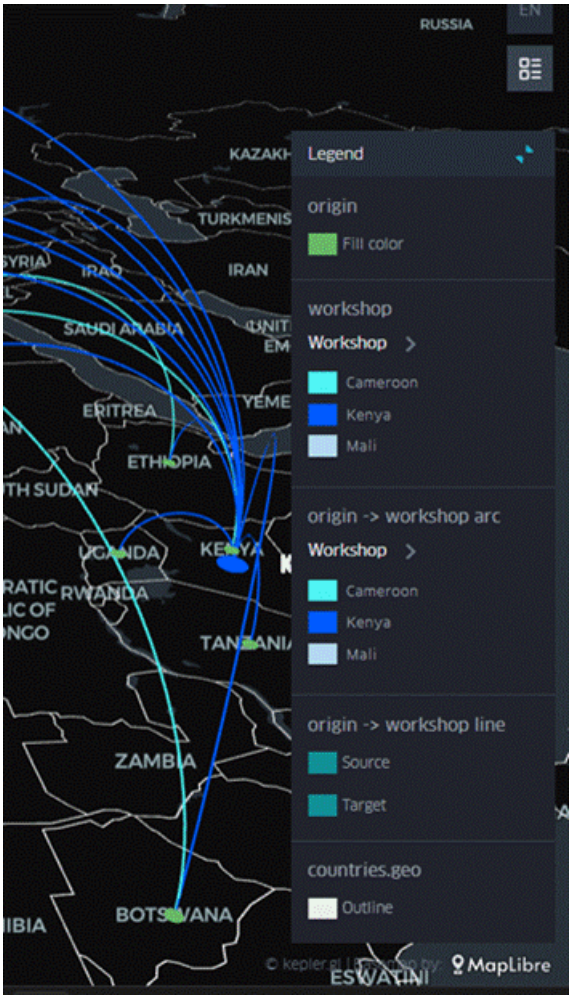
This design supports both macro-level pattern detection and country-specific inspection.

Initial Map View Overview

When the visualization is first loaded, the map displays the African continent using a dark, neutral basemap to emphasize foreground data layers. Three prominently highlighted point markers indicate the locations of the BioStruct-Africa workshops: Mali in West Africa, Cameroon in Central Africa, and Kenya in East Africa. These workshop locations act as destination nodes for the visualization.

From multiple countries across Africa and beyond, curved flow lines (arcs) originate and converge toward these workshop points. Each arc represents aggregated applicant movement from a country of affiliation to a specific workshop location. The curved geometry is intentionally used to reduce visual overlap and to clearly convey directional flow, similar to flight-path representations commonly used in geospatial flow maps.

The colors:

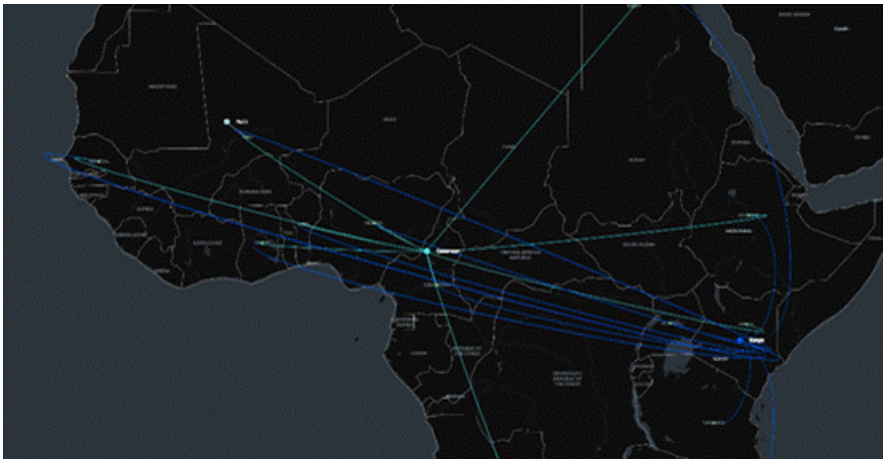


The Kepler.gl visualization uses a consistent and destination-centric color encoding scheme to clearly communicate geographic participation patterns. Countries of applicant origin are represented as green point markers, positioned at the geographic centroids of each country. These origin points indicate where applicants are affiliated and serve as the starting locations for all flow connections. Workshop destinations are encoded using distinct colors to enable immediate visual differentiation: Cameroon is shown in teal, Kenya in blue, and Mali in light blue. This color scheme is applied consistently across workshop points, connecting arcs, and the legend to maintain visual clarity and coherence.

Curved origin-to-workshop arcs inherit the color of the destination workshop rather than the origin country. This design choice intentionally emphasizes where applicants are traveling to, allowing viewers to compare the geographic reach and draw of each workshop at a glance. Directionality is conveyed through the arc geometry and endpoint placement, with internal source–target definitions handled by Kepler's arc and line layers, avoiding the need for arrows that could increase visual clutter. Country boundaries are displayed using a neutral light-gray outline from the countries.geo layer, providing

necessary geographic context while ensuring that the primary analytical elements—workshop destinations and participation flows—remain visually dominant.

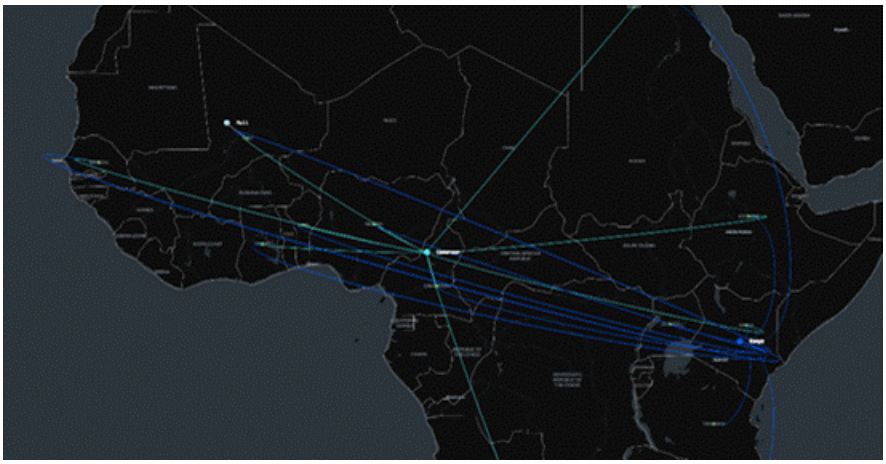
The lines (arcs):

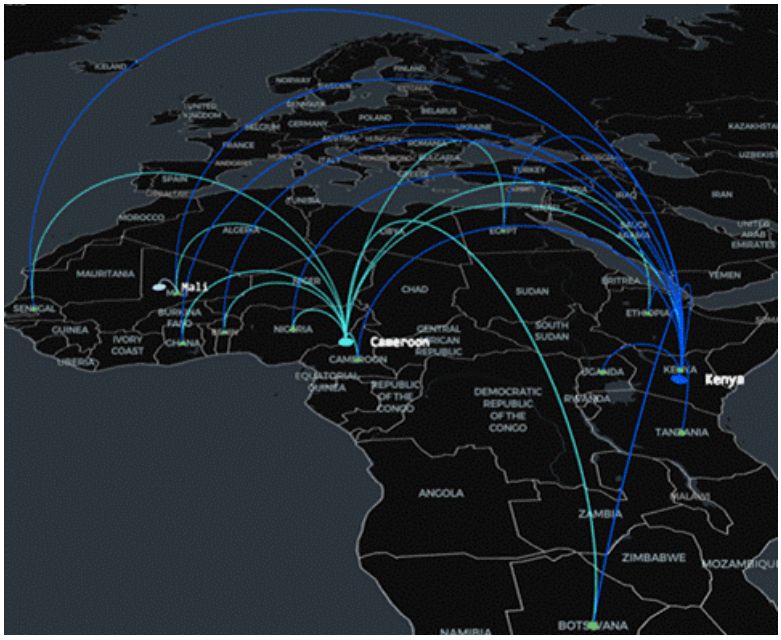


The curved lines in the Kepler.gl map represent aggregated participation flows from countries of applicant origin to the BioStruct-Africa workshop locations. Each line corresponds to a country–workshop pairing and encodes the volume of applicants traveling from that country to a specific workshop. Line thickness is proportional to participant count: thicker lines indicate higher numbers of applicants, while thinner lines represent lower participation. This quantitative encoding allows viewers to quickly distinguish major contributor countries from those with smaller applicant pools without relying on numeric labels.

The lines are rendered as smooth geodesic arcs rather than straight segments to reflect long-distance movement across the globe and to reduce visual overlap in dense regions. All lines are colored by destination workshop, ensuring consistency with the workshop point markers and legend. This destination-colored encoding shifts analytical focus toward workshop reach and draw, making it easier to compare how widely each workshop attracts participants across Africa and beyond. Together, line curvature, thickness, and color convey both the scale and direction of participation flows while maintaining readability in a geographically dense visualization.

MOVING THE MAP (Click and Drag)





You can interact with the Kepler.gl map by clicking and holding the left mouse button anywhere on the canvas and dragging the cursor. While dragging, the entire map translates smoothly in the direction of movement, allowing you to reposition the view. When you release the mouse button, the map remains fixed at the new location. During this interaction, all visual elements, including country boundaries, origin points, and flow lines—move together as a single coordinated layer, preserving spatial relationships.

This panning interaction is primarily used to explore different geographic regions in detail. It allows users to center specific countries or workshop locations of interest, examine peripheral regions such as Northern Africa or island nations, and reduce visual clutter by repositioning dense clusters. Panning is especially useful when combined with zooming and hover-based tooltips, enabling focused inspection of participation flows without losing broader geographic context.

ZOOMING IN and OUT (Scroll Wheel)

To zoom on the Kepler.gl map, place your mouse cursor over the country or region you want to examine and use the mouse scroll wheel. Scrolling downward (rolling the wheel away from you) zooms in, bringing the map closer and revealing finer spatial detail, while scrolling upward (rolling the wheel toward you) zooms out to show a wider geographic area. Zooming is centered on the cursor location, allowing precise control over the focal region.

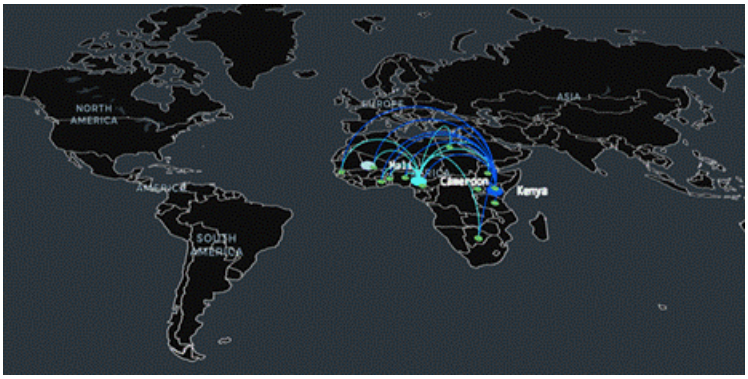
This interaction helps users shift between macro- and micro-level views of participation patterns. Zooming out is useful for understanding continental-scale connectivity and cross-regional flows, whereas zooming in enables closer inspection of individual countries, workshop hubs, and dense clusters of origin–destination lines.

ZOOM IN:

When you zoom in on the Kepler.gl map, the visualization reveals finer spatial detail. Country boundaries become more distinct, labels such as country and major city names appear more clearly, and individual origin–destination lines separating countries and workshops are easier to distinguish. At this level, overlapping flows that may appear merged at a broader scale become visually separable, allowing users to inspect specific participation paths in greater detail.

This zoomed-in view is particularly useful for analyzing dense regions such as West and East Africa, where many applicants flows overlap. By increasing the level of detail, users can identify subtle patterns—such as secondary contributor countries or weaker participation links—that are not immediately visible at lower zoom levels.

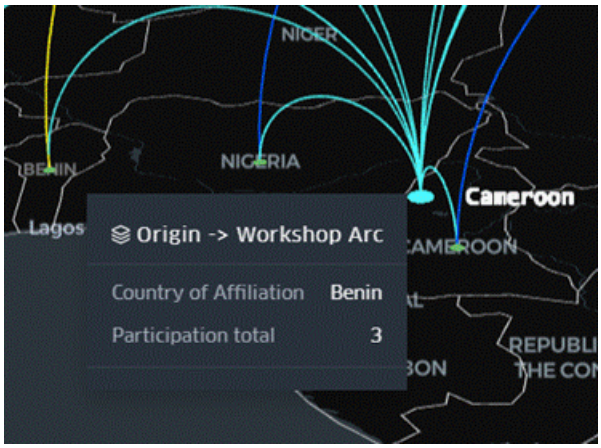
ZOOM OUT:



When you zoom out on the Kepler.gl map, the visualization shifts from local detail to a continental overview. The map scale decreases, allowing a much larger geographic area to be viewed at once, often encompassing all of Africa and surrounding regions. Individual origin–destination lines begin to visually aggregate, forming broader flow patterns rather than distinct paths. This aggregation helps emphasize overall movement trends and dominant workshop hubs rather than fine-grained country-level connections.

Zooming out is particularly useful for understanding the "big picture" of BioStruct-Africa participation. It allows users to quickly assess which regions contribute most strongly, how participation spans across Africa and beyond, and how the three workshops (Mali, Cameroon, and Kenya) function as regional anchors. In contrast, zooming in is best suited for precise inspections such as identifying the exact origin of a specific flow or examining participation from a single country in detail.

HOVERING Over Countries (Just Moving Mouse, No Clicking)



When hovering over a country on the Kepler.gl map, users can access contextual participation details through an interactive tooltip. By simply resting the mouse cursor over a country or its associated

origin–destination arc—without clicking—a tooltip appears near the cursor after a brief delay. This tooltip dynamically displays the country of affiliation, the total number of applicants originating from that country, and, where available, a breakdown of how those applicants are distributed across the three workshops (Kenya, Cameroon, and Mali). The information shown is computed directly from the aggregated dataset and reflects the same counts used to generate the visual flows.

This interaction is particularly useful for rapid exploration and comparison. Users can quickly answer questions such as how many applicants originated from a specific country, or whether a given country contributed more participants to one workshop than another, without manually interpreting line thickness or counting arcs. Once the cursor moves away, the tooltip disappears automatically, ensuring that the map remains uncluttered and responsive while supporting on-demand, detail-oriented analysis.

HOVERING Over Lines/Arcs

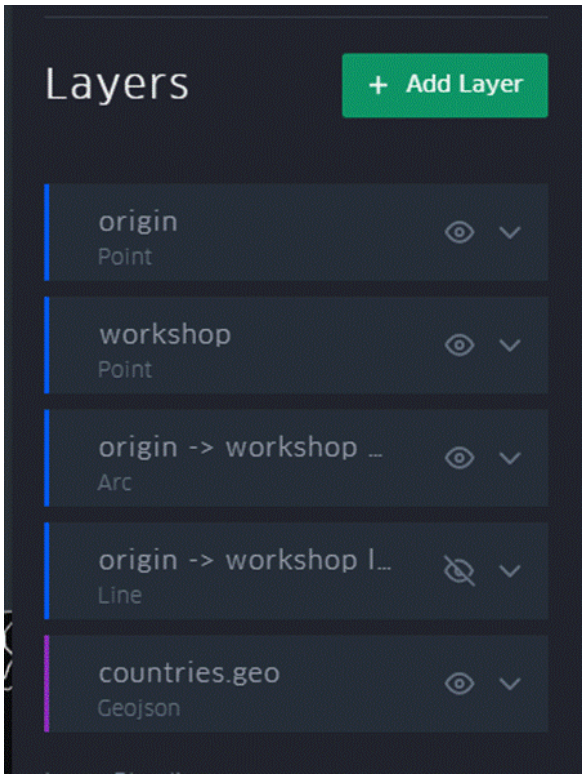
When the user hovers over a curved origin–destination line on the Kepler.gl map, the interaction highlights that specific flow to support precise interpretation. The hovered arc temporarily increases in thickness and is visually emphasized (rendered in a brighter highlight color), allowing it to stand out clearly from overlapping flows. At the same time, a contextual tooltip appears adjacent to the cursor. This

tooltip reports the origin country, the destination workshop location, and the exact number of applicants represented by that flow (for example, “Nigeria → Kenya: 42 applicants”).

This interaction is designed to disambiguate dense visual regions where multiple arcs overlap and where line thickness alone may be insufficient for exact quantification. By isolating a single flow and exposing its numeric value on demand, users can confidently validate specific origin–destination relationships, confirm participant volumes between a given country and workshop, and cross-check high-level patterns observed in the visualization with precise counts from the underlying data.

CLICKING the Legend/Layer Controls (Top Right Corner)

Look at the top-right corner of the map. You'll see small controls and layer names.



The map includes an interactive layer and legend control panel located in the top-right corner of the interface. This panel lists all active visualization layers, including the origin country points, workshop destination points (Mali, Cameroon, and Kenya), origin-to-workshop arc layers, optional straight-line layers, and the base countries GeoJSON layer. Each layer is accompanied by a visibility toggle (eye icon), allowing users to selectively enable or disable individual components of the visualization.

Toggling a workshop layer off immediately removes all corresponding flow arcs and destination markers from the map. For example, disabling the Mali workshop layer hides all flows and points associated with Mali, allowing users to focus exclusively on patterns related to Cameroon and Kenya. Re-enabling the layer restores the arcs and markers with their original color encodings. This functionality is particularly useful for reducing visual clutter, isolating workshop-specific participation patterns, and performing comparative analysis across workshops without interference from overlapping flows.

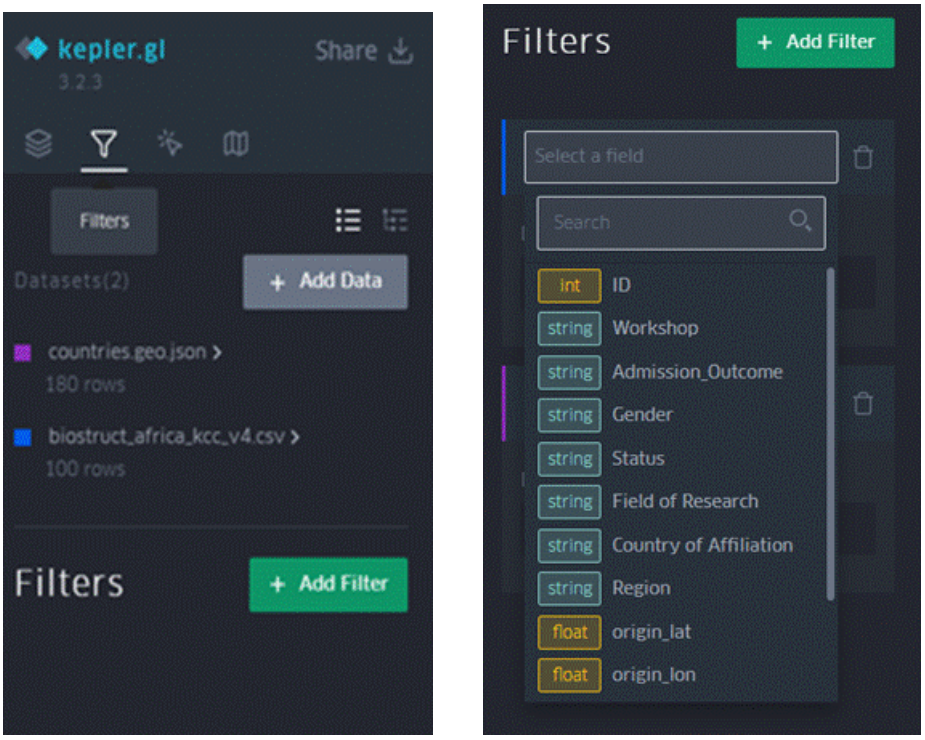
The legend also serves as a visual key for interpreting color encodings. Distinct colors are consistently applied to represent each workshop across point, arc, and line layers, ensuring interpretability even when multiple workshops are displayed simultaneously. The ability to dynamically control layer visibility allows users to tailor the visualization to specific analytical tasks, such as examining a single workshop's geographic reach or comparing cross-workshop participation **trends.**

CLICKING the "3D" Button (If Available)

Some Kepler.gl configurations include a 3D map toggle that allows users to switch from a traditional 2D map view to a three-dimensional, globe-like perspective. When the 3D mode is activated, the map surface tilts and flow arcs are rendered with greater vertical elevation, making long-distance connections and overlapping flows more visually distinguishable. Users can interact with the 3D view by clicking and dragging to rotate the map, enabling exploration of spatial relationships from multiple angles. This mode is particularly useful for presentations and exploratory analysis, as it enhances depth perception and

reduces visual occlusion in areas with dense flow patterns. To return to the standard planar view, users can disable the 3D mode by toggling back to the 2D option, restoring the flat map layout.

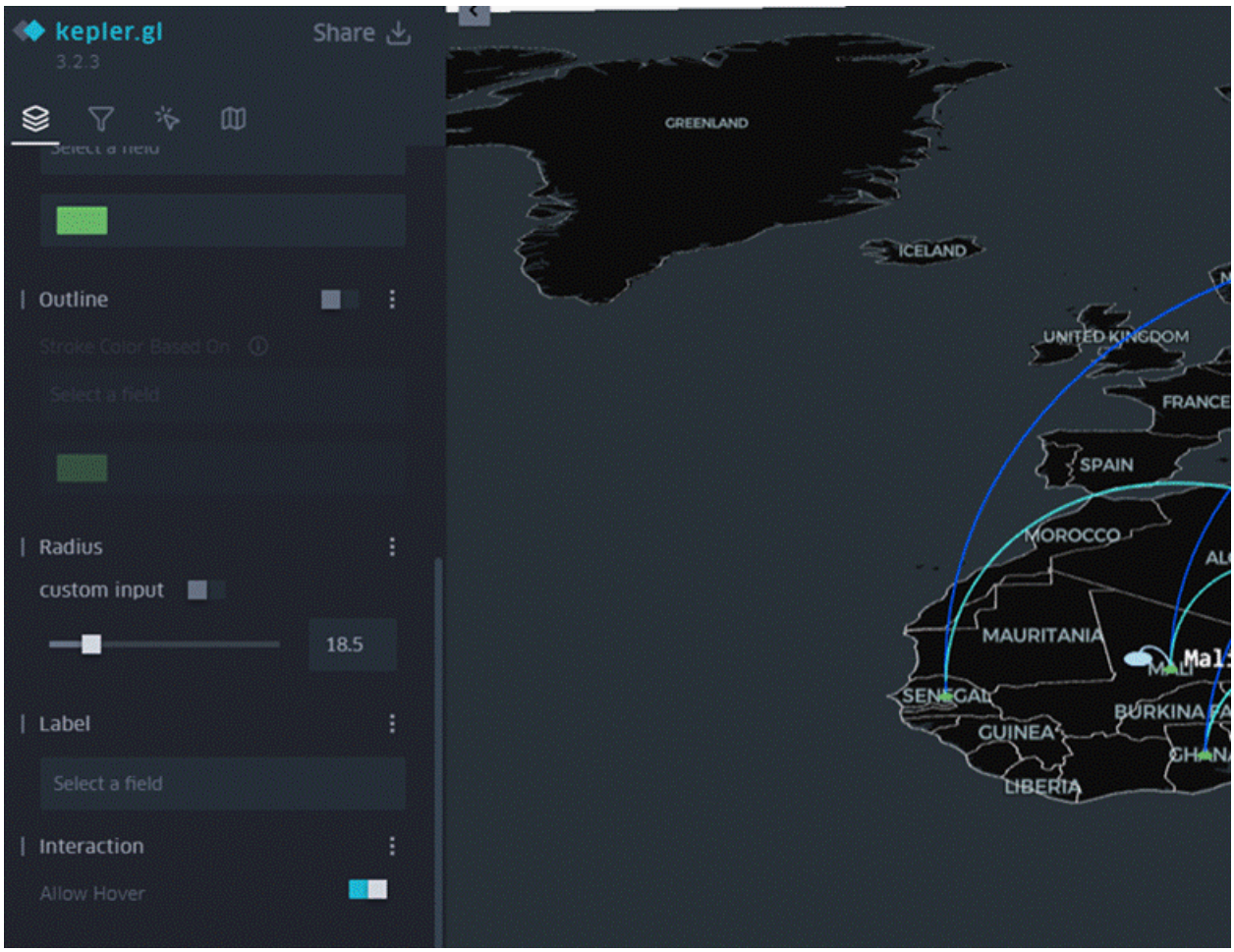
THE FILTER PANEL



Some Kepler maps show a panel on the left side with sliders and options. The Filter Panel in Kepler.gl enables users to dynamically subset the dataset based on specific attributes without modifying the underlying data. Filters operate across all active layers, allowing users to isolate patterns and reduce visual complexity during exploration. Supported filter fields include categorical variables (e.g., *Workshop*, *Gender*, *Admission Outcome*, *Field of Research*, *Country of Affiliation*, *Region*) as well as numeric variables (e.g., *ID*, latitude/longitude fields).

When a filter is applied, Kepler.gl immediately updates all visible layers—points, arcs, and lines—so that only records matching the selected criteria remain visible. For example, filtering by *Workshop = Kenya* hides all flows and points related to Cameroon and Mali, enabling focused analysis of a single workshop. Similarly, filtering by *Gender* or *Admission Outcome* allows stakeholders to examine demographic or selection-specific participation patterns without redrawing the visualization.

The filter interface supports multi-filter combinations, making it possible to perform compound queries such as viewing only *Female PhD Students from West Africa attending the Kenya workshop*. Filters can be enabled, adjusted, or removed in real time, providing an efficient exploratory workflow while preserving GDPR compliance, since no raw records are exposed or altered. This interaction mechanism is essential for targeted analysis, presentation clarity, and stakeholder-driven exploration of the BioStruct-Africa participation data.



Kepler.gl provides interactive styling controls that allow users to fine-tune the visual representation of flow arcs for clarity and interpretability. The **Arc Thickness** slide controls the stroke width of all origin-to-workshop connections, enabling users to emphasize higher-volume flows or reduce visual dominance when the map becomes crowded. Increasing thickness makes high-participation routes more prominent, while decreasing it helps declutter dense regions.

The **Arc Height** slider adjusts the vertical curvature of the arcs. Higher arc values create more pronounced curves, which improve separation between overlapping flows and enhances depth perception, especially when multiple routes intersect. Lower arc values flatten the curves, producing a more compact, map-aligned view that can be preferable for precise geographic interpretation.

Additional **color pickers** allow customization of arc and point colors by workshop, supporting visual distinction between Mali, Cameroon, and Kenya workshops. **Show/Hide toggles** enable users to temporarily disable specific layers or workshops, further reducing clutter and supporting focused analysis. Together, these controls provide flexible visual tuning without altering the underlying data, allowing the map to adapt to both exploratory analysis and presentation needs.

Patterns You Can See

1. The "Starburst" Pattern Around Kenya:

- Lines coming from ALL directions
- North, West, South, even from Asia and Europe
- Means: Kenya workshop is internationally attractive

2. The "Regional Cluster" Around Mali:

- Most lines come from nearby countries (Senegal, Burkina Faso, Guinea)
- Few long-distance lines
- Means: Mali attracts mostly local/regional applicants

3. The "Bridge" Pattern for Cameroon:

- Lines from both West AND East Africa
- It's in the middle, so it draws from both sides
- Means: Cameroon is a meeting point

4. Lines from Nigeria Everywhere:

- Nigeria has thick lines to ALL three workshops
- Means: Nigeria sends the most applicants overall
- They're a major contributor to the program

5. Empty Spaces (Southern Africa):

- Fewer lines from South Africa, Namibia, Botswana
- Means: These regions are underrepresented
- Opportunity for future outreach

Design Rationale

- Dark basemap was selected to improve contrast for colored arcs and reduce visual noise.
- White typography was implemented across Sankey labels and metrics to improve readability, as explicitly requested by the client.
- Granular attributes (e.g., individual countries) were intentionally removed from the Sankey to reduce clutter and improve interpretability.
- Scientific discipline text was standardized and clustered to reflect *actual research domains*, not noisy self-reported strings.

Common Questions About the Kepler Map

Q: I zoomed in too far and now I'm lost! How do I get back? A: Scroll your mouse wheel UP (toward you) to zoom out. Keep going until you see the whole continent again.

Q: Why do some lines overlap and I can't see them all? A:

1. Turn off some workshops using the layer controls (top right)
2. OR zoom in to separate them
3. OR increase "Arc height" slide to make them curve higher

Q: Can I see just female applicants or just PhD students? A: No - this map only shows geography (where people are from). Use the Sankey diagram to filter by gender or career stage.

Q: Why are there lines going to Europe or Asia? I thought this was Africa workshops! A: A few applicants came from outside Africa (maybe African scientists working abroad). The workshops are open to anyone interested.

Q: How do I save or share this map? A: Take a screenshot (Windows: Win+Shift+S, Mac: Cmd+Shift+4) or click Export button if available.

Q: The map is moving by itself / won't stop moving! A: You might be touching your laptop trackpad accidentally. Lift your hands for a second. Click once to stop momentum scrolling.

Q: I want to measure distance between countries. Can I? A: Not directly, but you can estimate: Long lines = far apart, Short lines = close together. Each line's curve height also gives a hint about distance.

Project Timeline and Iterations

Initial Requirements (10/21/2025)

- Dataset delivery pending
- GDPR compliance mandatory
- Visualization freedom granted
- IU infrastructure approved

Intermediate Review (11/25/2025)

Client requested:

1. Additional Kepler map focused only on Africa and workshop flows
2. Removal of country-level granularity from Sankey
3. Scientific discipline normalization and correction
4. Further data cleaning
5. Attribute renaming for clarity

All requested changes were implemented.

Final Client Feedback (12/08/2025)

Client confirmed:

- No further meetings required
- Visualizations are intuitive
- Minor fixes requested (capitalization, title, white text, tooltip persistence, path highlighting)

All items were delivered successfully.

Deliverables (Followed GDPR guidelines)

- **cleansed_biostruct_data_v4.csv – Final cleaned dataset**
- **Interactive Sankey visualization (HTML)**
- **Kepler.gl Africa-focused workshop flow map**
- **Jupyter notebooks for EDA and data standardization**

- **Project report documentation**
- **This handover document**

All files are stored in a private Google Drive folder and must not be shared publicly.

Outstanding Tasks

1. Kepler Tooltip Enhancement:
   Add workshop-level breakdown information when hovering over countries.
2. Final Sankey Typography Check:
   Ensure all labels and metrics are rendered in white as per client request.

Final Notes

This system was designed for analysts, program coordinators, and decision-makers with domain familiarity. All design and analytical choices prioritize interpretability, reproducibility, and scalability. Future extensions (e.g., longitudinal analysis, institution-level mapping) can be integrated without architectural changes.

==========