Development of an Exploratory Visualization Tool for Comparison of Biotic and Abiotic Amino Acids

Kristina Bridgwater
Computer Science
UMBC
Baltimore, MD, USA
kbridgw1@umbc.edu

Paul Heiss
Computer Science
UMBC
Baltimore, MD, USA
pheiss1@umbc.edu

Nathan Vanbourgondien
Computer Science
UMBC
Baltimore, MD, USA
natha1@umbc.edu

***Abstract*—The focus of this work is to produce a prototype visualization tool which will allow researchers to compare multiple sets of user-defined amino acids in the Euclidean space. The focus is not currently on data analysis, but rather on the framework that will eventually support data analysis by distributed teams.**

In general, data sets visualized by the tool will be defined by quantitative measures of amino acid properties like size, charge and hydrophobicity. The tool's initial visualization capabilities will include a search-able and sort-able data table as well as visualizations to show correlation between data sets and specific information per amino acid chain.

The goal is that this tool will evolve into one used by research communities distributed between North America, Europe, and Japan, and expedite amino acid research.

***Index Terms*—visualization, amino acids, D3, react, node, chemistry space**

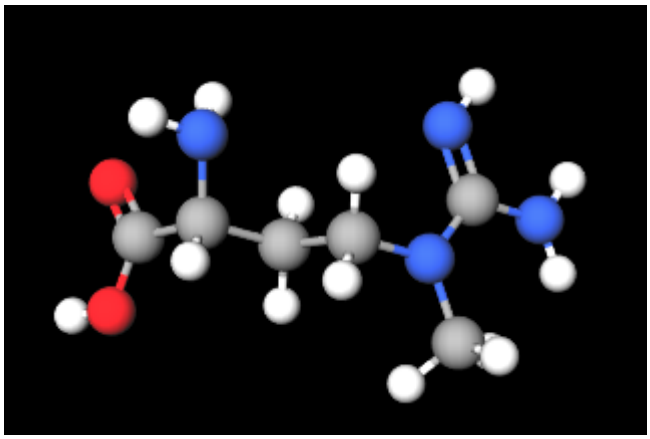

Fig. 1. 3D rendering of an amino acid as defined by a SMILES formula. The visualized image can be rotated and zoomed.

I. Introduction

The focus of this work is to produce a visualization tool which will allow researchers to compare multiple sets of user-defined amino acids. The data sets will be defined by quantitative measures of amino acid properties such as size, charge and hydrophobicity.

The sections below describe how the visualizations will present the available data to the end user and what technology tools will enable the team to visualize the data. Additionally, after the presenting the visualizations, known issues with the toolkit as well as longer term goals for the project are discussed.

II. Background and Related Works

Our client for this effort is Dr. Stephen Freeland, an Evolutionary Biologist at UMBC, who was suggested to the team by Dr. Engel, a Computer Science faculty member and Associate Vice President for Research Development, also at UMBC. Team member Kristina Bridgwater interviewed Dr. Freeland [3] in her search to conduct a project relating to bioinformatics. Dr. Freeland's hope for the project is that the prototype visualization can evolve into a tool utilized by multiple research communities distributed between North America, Europe, and Japan. Freeland's wish is that this work will extend his initial collaborative approach to design a system for visualizing quantitative factors to determine the similarities between amino acids[1].

The team conducted preliminary email exchanges with Dr. Freeland [3] to expand our domain knowledge regarding the project and to assist us in creating this proposal's text. The interviews formed the basis of the project and the visualization tool that was developed.

According to Dr. Freeland, there are at least three distinct areas of investigation that are interested in the chemical structures of biological material beyond the "natural alphabet": synthetic biology, looking to artificially increase life's alphabet; meteoritics, analyzing the organic chemistry of planetary material to answer questions about cosmic origins; and biochemistry, studying how our current understanding of life has changed over time to become

Table SPLOM

Amino Acid Data Table

ID or SMILES

	ID	SMILES	mib_vol	LogP_Jchem	pKa_uncap	Type
>	2	C(=O)(CN)O	67.727	-3.409459573	2.307456882	Computational
>	2	C(=O)(CN)O	67.727	-3.21	2.31	Abiotic
>	2	C(=O)(CN)O	67.727	-3.409459573	2.307456882	Coded
>	11	C[C@@H](N)C(=O)O	84.313	-2.84078791	2.474897471	Computational
>	3	CC(N)C(=O)O	84.313	-2.96	2.47	Abiotic
>	11	C[C@@H](N)C(=O)O	84.313	-2.84078791	2.474897471	Coded
>	16519	C(=O)(CNC)O	85.401	-3.186260954	2.056675239	Computational
>	16519	C(=O)(CNC)O	85.401	-2.78	2.06	Abiotic

8 rows 1-8 of 1977

Fig. 2. Vis1 - Sortable Table. *Navigation between visualization appears at the top left: selecting the link **Table** navigates to this visualization; and selecting the link **SPLOM** shows the scatter plot matrix shown in Figure 3. By clicking on the column label, data shown in each column is sortable, either alphabetically or numerically, as appropriate. Data can also be filtered for specific values by entering a string in the text box below the column label. A global search for an ID or SMILES formula is also available in the top right corner.*

what we know today. Dr. Freeland further stated that these three fields range dramatically in size and scope but overlap in their interest of comparing amino acids, both individually and in sets. These fields would be interested in seeing both structural visualization and quantifiable property ("descriptor") comparison, combined in a Euclidean space that can be subject to further mathematical and statistical analysis. This is also known more generally as the concept of "chemistry space".

In the course of his previous works to visualize amino acids [4], Freeland explored relationships between amino acids before and after prebiotic earth by exploring the use of chemical space. Their initial work resulted in a testing framework which established three tests to assist in quantifying the results. The problem of course, as described by Dobson[2], is the sheer vastness of chemical space, which encompasses all possible organic molecules, and how so little of it has been explored. Dobson states that this is what drives the need for a data visualization tool: by understanding this space better, significant biological findings can be discovered. This is the motivation for this project.

III. Data Sets

The data was provided by the sponsor in multiple CSV files. For the initial release of this project, data was processed from three files: the abiotic_cleaned.csv file, the coded_cleaned.csv file, and the computational_cleaned.csv file. The structure of the files were rows of data that each had columns for an identifier, a text string in the format of a Simplified Molecular Input Line Entry System (SMILES) formula for an amino acid, then one or more quantitative columns. For this initial tool release and the paper, the columns included were mib_vol,LogP_Jchem,and pKa_uncap.

The combination of the data sets resulted in more than two thousand rows of data. This data was parsed, combined and stored as a JSON (JavaScript Object Notation) file for later use by the tools described in the Technology section.

One of the columns in the data set contained a SMILES formula. The SMILES language is a typographical notation for specifying chemical structures [8]. This format is the defacto standard for describing molecular structures in a concise, textual format that can be used by the computer. SMILES formulas utilize the atomic symbol from the periodic table to represent each atom in a complex chemical structure. According to conventions of the specification, upper case letters refer to non-aromatic atoms and lower case letters refer to aromatic atoms. When an atomic symbol has more than one letter, the second letter is lower case.

The sequence of specific atomic elements form the amino acid chain. This SMILES formula and will be used to drive a 3D model of the amino acid. Each amino acid chain will have unique quantitative attributes, and the SMILES formula may be used as an index to pull other third part data into the visualization tool.

The data is available in the form of a multi-tab CSV file. There are multiple data sets:, one for the twenty "natural" amino acids; one for those found in abiotic chemistry; and one for those used successfully within synthetic biology. For each amino acid, primary fields would be an arbitrary unique identifier, a 3D molecular structure represented in SMILES format[5], and a variable number of real number fields, each representing a quantifiable property (such as molecular volume in angstroms, or hydrophobicity measured as LogP). This is limited data, however. The amino acid index is comprised of more than twenty values representing both phyciso- and biochemical properties of amino acids, and specifies how

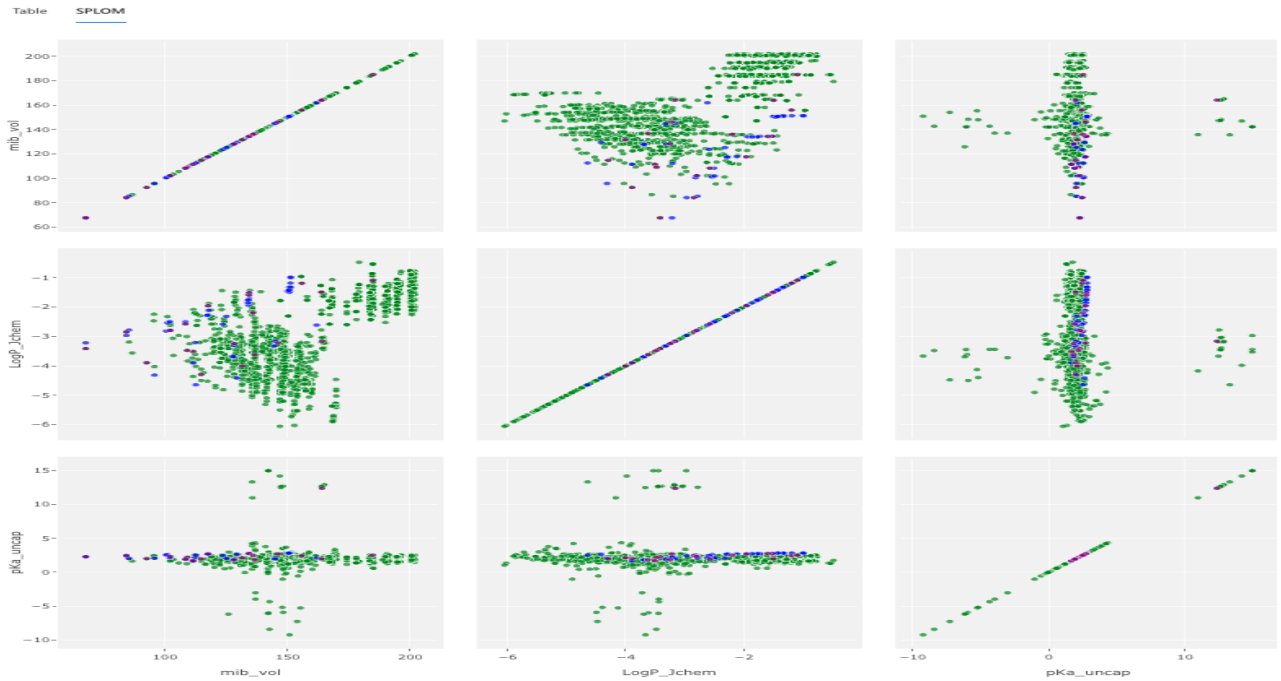

Fig. 3. Vis2 - Scatter Plot Matrix. *Navigation between visualization appears at the top left: selecting the link **SPLOM** navigates to this visualization; and selecting the link **Table** shows the sortable table shown in Figure 2. The SPLOM depicts the pair-wise data correlation of quantitative attributes. There is a subplot for each pair of data attributes. Panning and zooming is available for the sub-graphs. Mousing-over a point on a graph shows all its data, while click a data point loads that amino acid's 3D rendering.*

both parts of a descriptive mutation matrix are created and how the correlation coefficients are determined[6]. But all twenty are not needed to build the initial visualization tool.

Using the provided data sets, the team created an initial visualization landing page showing a data table displaying the available data. Described in more detail int eh following sections, the table is searchable and sortable. Additional visualizations to show correlation between data sets, as well as specific information per amino acid change, are also part of the tool's design.

IV. Visualization Tool Design

The team decided that a browser-based tool to present the data would be ideal. With the tool deployed onto an appropriate web server, users other than the project's sponsor will be able to utilize the tool to visualize their data. Thus the first design decision was made.

Subsequent design decisions included how to navigate the visualizations, how to potentially load other data files, and how to show alternative data that is not included in the supplied data sets. After considerable discussions, the team decided on a design to includes the following features:

- **Vis1** A sortable table shown in Figure 2
- **Vis2** A scatter plot matrix shown in Figure 3
- **Vis3** A 3D rendering shown in Figure 1

Vis 3 is presented in a dynamic section where additional data, such as amino acid's properties from the file, can also be displayed. This container, including the 3D rendering, will be populated when a user clicks on an amino acid in one of the primary visualizations.

The design is flexible to enable subsequent researchers to add and/or eliminate features as needed. Furthermore, the code is designed in distinct modules (e.g. vis1, vis2, etc.) to allow for independent development of distinct features by various team members. This allows for easier merging and refactoring of code from a GIT repository to allow the tool to be deployed on a public facing web server. The designed was presented and approved by Dr. Freeland before commencing on the development work.

V. Technology Stack

Tableau is a commercially available data visualization tool that others had used to visualize amino acid data [7]. But during the alpha release, the team determined that using Tableau would not afford sufficient design flexibility and may preclude users from using the developed visualization tool if they did not own a Tableau license. Additionally, since the team decided that a web-based interface was appropriate, migrating to web-based tools made sense. The primary tools used and their respective uses are:

- **React** is a web-based framework for developing user interfaces;

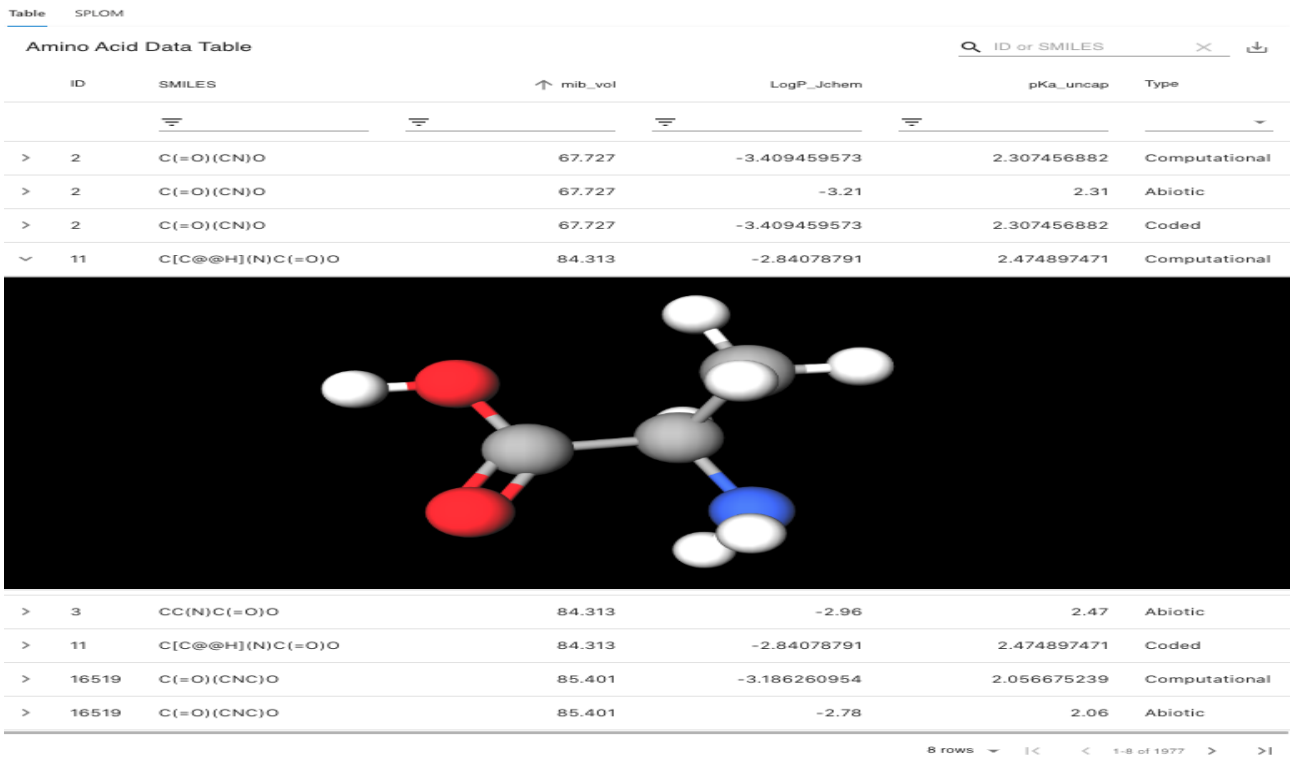

Fig. 4. Vis1a - Sortable Table Expanded with a 3D Rendering. *Navigation between visualization appears at the top left: selecting the link* ***Table*** *navigates to this visualization, pre-expansion; and selecting the link* ***SPLOM*** *shows the scatter plot matrix shown in Figure 3. By clicking on the > to the left of the ID, the 3D rendering of the selecting amino acid's SMILE formula is shown beneath the selected data row. The row can be collapsed by clicking agin on the symbol the the left of the ID.*

- **Node.js** is a server based framework for rendering web apps and accessing data;
- **D3, Plotty and material-table** are visualization libraries;
- **MolView API** is a 3D amino acid rendering;
- **Vercel.com** is public facing web server.

All of the tools utilize the JavaScript language for development. This made integration easier, although the learning curve on all available options for each tool is significant. Fortunately, with the help of Google searches and tools like StackOverflow, most questions can be quickly researched and solved.

There were some challenges with the technology stack, especially with the use of the React framework and its hierarchical design. In the document object model (DOM) for a Hyper Text Markup Language (HTML) page, React encourages passing values from top down through the DOM elements to generate a dynamic user interface (UI). If something is changed by the user or the page in a higher level context, that change can then be reflected by changing the properties of children components in the DOM and triggering a lifecycle update. This effectively re-renders components to reflect the changes on the visual page.

For this visualization tool, the components are laid out such that elements that need to communicate with each other are isolated by several levels of context. This proves to be difficult design implementation as React is designed to handle changes from top down. While there are several methods to simulate passing values from child to parent, React does not natively support the functionality.

In order to enable communication between separated components, a function can be declared and passed down from the root parent component to its children. This function's context must be bound to the original DOM component which can be achieved with the function this.function.bind(this)). If a function bound in such a way alters values in the original component, the function can be passed down through the component hierarchy and eventually used to simulate the passing of information upstream. This same problem can be solved through global context objects, but they are to be used sparingly and only for information used by a wide variety of components.

Nonetheless, these challenges were overcome and the tool was successfully implemented based on the initial designs.

VI. Visualizations

Once Dr. Freeland approved the team's design approach for the tool, the initial visualizations for the project were developed. The visualizations included a sortable table with filtering capabilities; a scatter plot matrix; a three

dimensional (3D) model of the amino acid, and a chart of specif attributes for a selected amino acid . Those visualizations are fused through a navigation menu and are described and depicted in the following sections. The primary tool for the visualization development is the D3 JavasScript library.

A. Visualization Tool Description

The visualization's tool's user interface (UI) design is simple and intuitive. The landing page shows the sort-able table (Vis1) loaded with the projects data. The navigation to other visualizations is at the top left of the landing page and are marked by the words *Table* and *SPLOM*. When *Table* is selected, the user is shown the landing page with Vis1, the sortable table. When *SPLOM* is selected, the user is shown Vis2, the scatter plot matrix. These navigation choices will always be available at the top left corner of the visualization tool, and the features of both Vis1 and Vis2 are described in subsequent sections.

The simple navigation design allows for quick, descriptive navigation among available visualizations, is easily extendable for new features, and provides a UI known to most users.

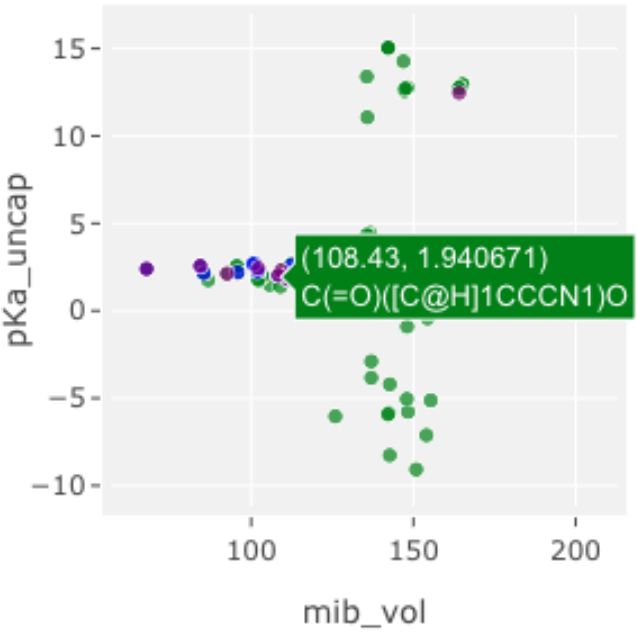

Fig. 5. Descriptive information for a single amino acid data point. This visualization is available on the SPLOM chart during mouse-over events. The data shown are any quantitative attributes loaded in the data sets as well as the amino acid's SMILES formula.

B. Vis1: Sortable Table

React has a component called the **material-table**. This table was used for the project and the implementation is shown in Figure 2.

By clicking on the column label, data shown in each column is sortable, either alphabetically or numerically, as appropriate. Data can also be filtered for specific values by entering a string in the text box below the column label. A global search is for an ID or SMILES formula is also available in the top right corner.

This table, while powerful, has some limitations, the primary one of which is the data in all columns are textual. This means that columns which should be treated as numerics can not be easily rendered once converted, and this limits the team from adding functionality such as an inequality search, range searches, etc. on numeric data. Additionally, the filtering looked for exact matches on strings, so to correct for this, the floor function was implemented during the filtering of numeric columns.

C. Vis2: Scatter Plot Matrix

A scatter plot matrix (SPLOM) is a matrix of scatter plots used to visualize the relationship between a pair of variables. For the initial tool release we are plotting three variables, thus our scatter plot matrix is a 3x3 grid containing nine scatterplots. The off diagonal plots show the correlation between two variables, while the diagonal plots are simply the auto-correlations of the variable and show the range of the data.

An important feature of the scatter plot matrix is that users are provided descriptive information of the plotted data points when they mouse-over a specific data point, such as depicted in Figure 5. This is a blown-up image from the full scatter plot matrix rendered in Figure 3, where the abiotic_cleaned.csv data points are blue while the computational_cleaned.csv data points are green.

Users having the ability to visualize the correlation of the loaded data sets, as well details for from values of individual data points, is a powerful feature.

Additionally, zoom and panning features have been enabled on the SPLOM to allow the user to visualize all data sets simultaneously, to move to a specific graph (panning), or to enlarge an area of the graph (zoom) at an increased magnification. Engaging the users with the data allows them to increase their understanding of the data as well as to provide better visual channel separation and encoding.

If a user selects a data point in the SPLOM, that data point is highlighted in all SPLOM subplots, and a 3D rendering of the amino acid change is depicted accompanied by the quantitative attributes for the selected data point.

D. Vis3: 3D Model

The SMILES formula is precise and concise, but does not provide the user a visual representation of how the atoms comprising the amino acid are bonded.

To enable this feature, an embedded image from the website *https://molview.org/* was utilized via an application programming interface (API) call. Molview, as described on the company's website, is an intuitive, Open-Source web-application to be mainly used as web-based data visualization platform. The team experienced no issues using the MolView API, and the feature added visual appeal to our final tool.

When the 3D model is rendered, the data from 6 is also shown to the user. This data is dynamic based on the data loaded and shows all attributes of the selected and rendered amino acid.

ID: 4787

SMILES: C1(=C(C=NC1)[C@@H](N)C(=O)O)N

mib_vol: 135.678

LogP_Jchem: -4.88059519

pKa_uncap: 4.254454517

Fig. 6. Data attributes shown after user selected an amino acid on the SPLOM. This data is simultaneously shown along with the 3D rendering of the SMILES formula.

VII. Known Issues

There were a few technical issues while developing the visualization tool. Some involved the HyperText Markup Language (HTML) Document Object Model (DOM) which wanted to be controlled both both the React and D3 frameworks. This was resolved in the tool's final implementation.

There were minor nuances of each framework which increased the time to debug some issues. For example, the projects code is in a directory labeled *src*, and our data was in a directory labeled *sampledData* under the same parent node. The data could not be loaded from *sampleData* with the appropriate Node.js function call because Node requires the data to be in the *src* directory. We assume this is a security feature, or may be a side effect of not doing the development on a web server. Regardless, it is these type of simple issues that consume time when developing this initial version of the tool.

Other functional issues that were experienced were limitations with the React data table that was implemented. The filtering, while adequate for an initial release, does not enable advanced features such as inequalities and range searches.

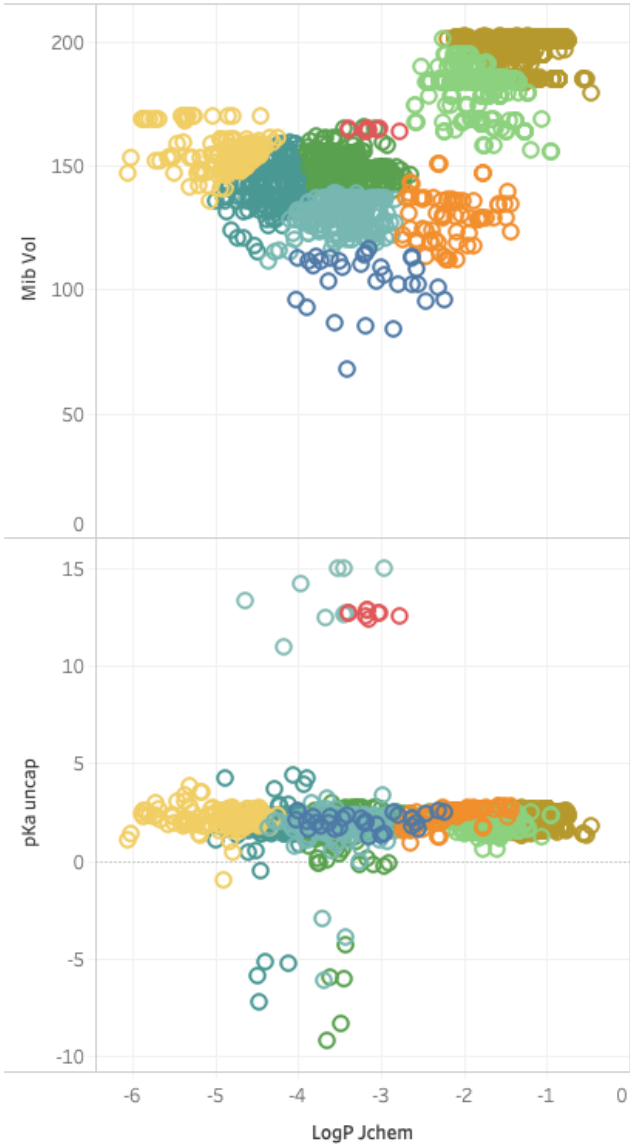

Fig. 7. Sample data analysis visualization showing k-means clustering where there are ten groups (e.g. k=10) of similar amino acid chains across three quantitative attributes..

VIII. Longer Term Goals

In order for the tool to be utilized by distributed research teams, a few new features and capabilities must be added. This includes dynamic data loading from users files as well as data visibility. Both of these file management features present challenges.

Data loading on a web server can be handled through the File Transfer Protocol (FTP), or even secure FTP (SFTP), but some consideration needs to be given to naming conventions, file format validation, and column naming conventions, especially when combining multiple data sets that may have different column names representing the same data. Data visibility is also important to consider if user data is public, private, or shareable.

Additional features that may work well in the in the tool include more third party data as well a other visualizations, such as one that will cluster the data into groups as shown in Figure 7. The clustering was a prototype implementation that was considered by the project team by was not desired by the sponsor at this time.

REFERENCES

[1] B. Bulka, M. desJardins, and S. Freeland, "An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices," *BMC Bioinformatics*, vol. 7, no. 329, 2006. [Online]. Available: https://doi.org/10.1186/1471-2105-7-329

[2] C. M. Dobson, "Chemical space and biology," *Nature*, vol. 432, December 16, 2004.

[3] S. Freeland, "Investigatory interviews," in *Zoom Meetings, Fall Semester*, September 2020.

[4] M. Ilardo1 and S. Freeland, "Testing for adaptive signatures of amino acid alphabet evolution using chemistry space," *Journal of Systems Chemistry*, vol. 1, May 2014. [Online]. Available: http://www.jsystchem.com/content/5/1/1

[5] C. A. James. (2007-2016) Open smiles specification. [Online]. Available: http://opensmiles.org/opensmiles.html

[6] S. Kawashima and M. Kanehisa, "Aaindex: Amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, p. 374, January 2000. [Online]. Available: https://doi.org/10.1093/nar/28.1.374

[7] J. Roszik and S. Woodman, "Hotspotter: efficient visualization of driver mutations," *BMC Genomics*, vol. 15, December 2014. [Online]. Available: https://doi.org/10.1186/1471-2164-15-1044

[8] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, February 1988.