WPI

Exploratory Cluster Analysis of Bumblebee and Plant Interaction Data

A Major Qualifying Project submitted to the faculty at

Worcester Polytechnic Institute

in partial fulfilment of the requirements for the

Degree of Bachelor of Science

Andrew Ressler
Bioinformatics and Computational Biology
Computer Science

Meadow Wicke
Bioinformatics and Computational Biology

Professor Carolina Ruiz, Project Advisor
Department of Computer Science,
Bioinformatics and Computational Biology Program, WPI

Professor Elizabeth F. Ryder, Project Co-advisor
Department of Biology and Biotechnology,
Bioinformatics and Computational Biology Program, WPI

This report represents the work of one or more WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on the web without editorial or peer review.

Table of Contents

Table of Figures

Table of Tables

Abstract

A decline in both pollinator and floral biodiversity has motivated conservation initiatives such as the Beecology Project to collect and analyze ecological data in hopes of sustaining native bumblebee and plant species populations. After incorporating native plant and *Bombus* subgenera classifications, we used clustering and visualization techniques to uncover patterns within our local bee-flower network. Our experiments showed significant variation in the clusters that occurred, with both flower shape and bee species appearing to play a role in many clusters.

Acknowledgments

We would like to acknowledge and thank everyone who contributed to our project's success.

- **Professor Carolina Ruiz** for giving us insight and technical expertise on the computational aspects of our project as well as providing relevant feedback to aid in our analysis of the bee-flower network.
- **Professor Elizabeth Ryder** for guiding us through the biological components of our project and providing relevant feedback to aid in our analysis of the bee-flower network.
- **Professor Rober Gegear** for providing valuable information regarding *Bombus* and plant species, as well as reviewing our biological classifications.
- **Sarun Paisarnsrisomsuk** for his help in reviewing and adding our floral classifications to the official Beecology database.
- **All the members of the Beecology and Bio-CS Bridge Teams** for their support and insights into the various aspects of the Beecology Project and Bio-CS Bridge.

Chapter 1: Introduction

Pollinators are a crucial component in many co-dependent biological networks and contribute greatly to global biodiversity. Approximately 80% of wild plant species (Potts, 2010) and 75% of cultivated plant species (Thomann, 2013) are dependent on insect pollination as means for sustainability and survival. Pollinators also contribute to higher trophic levels in ecosystems as many species are dependent on plant products for food (e.g., seeds and fruits), and nesting material. Unfortunately, in the past decade pollinator decline has become an increasingly prevalent issue, one of chief importance in environmental conservation due to the important roles pollinators play. Research has indicated that human-made drivers such as agrochemicals, climate change, and habitat destruction have directly contributed to this downward trend (Potts, 2010).

One of the insect genera most important for pollination in New England is *Bombus*, known colloquially as the bumblebee. Close to fifty bumblebee species are native to the North American continent, with twelve generally found in New England. When comparing current and historic data in the New England area, it is evident that the populations of many bumblebee species as well as the plants they are responsible for pollinating have declined both in numbers and geographic spread (Colla, 2008). Although it is obvious that the number of pollinators plays an important role, the diversity of the pollinator species is equally as important in maintaining balance in an ecosystem.

In response to pollinator decline (both in population and diversity), the Beecology Project was established to inform citizen scientists of this issue as well as encourage them to record their own observations of bee-flower interactions. This citizen-collected data, as well as historic data, is stored in the Beecology database for use in analysis of the bee-flower interaction network. Additionally, the Beecology Project works alongside the Bio-CS Bridge, an NSF-funded project which is developing an educational curriculum to engage students in using computational thinking and tools to analyze biological data.

In an effort to discover patterns in at-risk bumblebee pollinator species, we performed exploratory analysis of the bee-flower interaction network using the real world data collected in the Beecology database. The difference in networks between bees and native versus non-native plant species was identified as a major area of interest. This prompted us to add new ecological classifications to the database to further classify floral and *Bombus* attributes. With the addition of these new classifications, we used hierarchical clustering, visual, and manual analysis techniques to explore the data and investigate the interaction network. Our analysis indicated that flower shape plays a large role in bee-flower preference, and the distribution of *Bombus* species (and thus tongue lengths) throughout the clusters played far less important of a role. We also determined that bees do not seem to have a preference for either native or non-native plant species, but do show an affinity for plants of specific families.

Chapter 2: Background

2.1 Ecological Components

Codependent relationships between organisms are an essential part of sustaining life and are present throughout the global ecosystem. In providing recommendations, learning tools, and spreading awareness about the pollinator decline, it is essential that we understand the biology behind the bee-flower interaction.

2.1.1 Pollinators

Many plant species rely on animals like beetles, bees, and butterflies for reproduction. When an animal visits a flower in search of food (nectar and pollen), it can deposit pollen from a previously visited plant of the same species onto the female reproductive part, the stigma, of the flower. This pollination signals to the plant to begin reproducing, usually via a fruit or seed (About Pollinators, n.d.). While it is possible for plants to self-pollinate using natural forces such as water and wind to carry pollen from the stamen (the male part), to the stigma, the United States Department of Agriculture estimates that three-fourths of the global flowering plant population and approximately thirty-five percent of the world's food crops rely on pollinators as carriers to trigger reproduction (USDA, n.d.). Currently, the most well-documented and researched pollinator species are bees, which are responsible for pollinating sixty to seventy percent of the world's total flowering species (Brauman, 2018).

The role pollinators play in sustaining an ecosystem is also dependent on their physical, and species-specific characteristics. Both the functional diversity, which is the range and variety of functions organisms have that contribute to an ecosystem, and species diversity amongst pollinators is crucial in order to promote pollination and modulate how insects in a community deliver pollen. Pollinator attributes such as tongue length, flower preference, and sensory/cognitive abilities all contribute to the functional diversity of an ecosystem (Kremen, 2007). A 2013 study examined the effect of bee diversity on plant reproductive success. In the experiment, they established 55 self-contained mesocosms, each with a different combination and number of up to 5 bee and 16 plant species. An analysis of the seed production in each cage confirmed that both species richness and diversity resulted in a positive yield on seed production. Functional complementarity between the different species was apparent as the cages with more pollinator diversity had more visits from different species and had higher yields, as shown in Figure 1 (Frund, 2013).

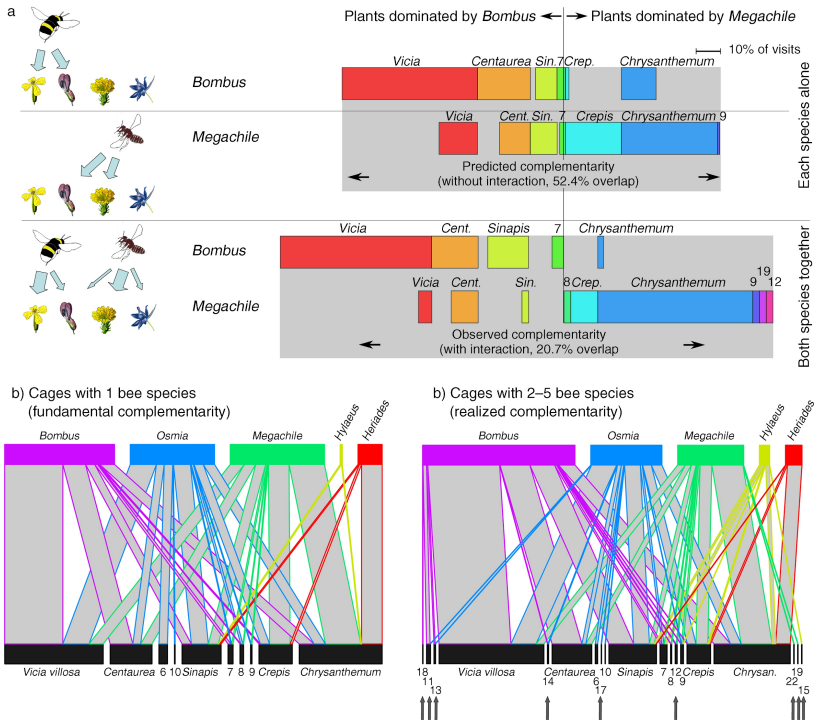

Figure 1: Functional Complementarity among the bees and plant species visited. When multiple bee species were present, they made more visits and had more diverse plant interactions (Frund, 2013).

Increased species diversity implies a more diverse set of functional diversity which can, in turn, cater to a wider variety of other life forms. Coevolution, how different species influence each other's evolution through natural selection (Rafferty, 2020), results in a biodiverse ecosystem that's very interdependent. Because so many species have coevolved to rely on each other for food, germination, and other functions, the reduction of biodiversity of one species can result in the reduction of biodiversity of another. For example, the loss of a long-tongued pollinator species in an otherwise self-sustaining ecosystem could lead to insufficient pollination of plants with deeper or less accessible stigmas. Over time, this could lead to the endangerment or extinction of said plant species, or be cause for the evolution of those plants to become more accessible for the present pollinators. With a more diverse pollinator population, an ecosystem can sustain a more biodiverse plant population as well.

2.1.2 Pollinator Decline

Despite the substantial role pollinators play in the global ecosystem, the population, geographic spread, and diversity among the different species has declined (Colla, 2008). While domesticated bees (i.e., honey bees, non-native) have been at the forefront of pollinator-flower interaction research and conservation efforts (Penn, 2019), other native species such as the variety of different bumblebees are also seeing a decrease in diversity, a

factor that can in turn also lead to a decrease in population. While honey bees contribute largely to the pollination of different crops and hold a larger economic importance as compared to their non-honey producing counterparts, they are not as important for sustaining native ecosystems. After analysis of pre- and post-1980s pollinator data, Biesmeijer and colleagues found that there had been a 30% decrease in the number of different bee species in both Britain and the Netherlands. This was consistent with their findings of decreased population size of the recorded bee species (Biesmeijer, 2006).

The decline in species richness and functional diversity has been attributed to several human-influenced factors that often work in tandem. Land use that causes habitat destruction and degradation is one of the main drivers for bee pollinator decline. As land becomes increasingly urbanized, the resources for survival and spatial availability for a colony decreases. Not only does urbanization destroy the natural resources for the pollinators, but also introduces new obstacles such as pesticides and potentially dangerous alien species, both animal and plant (Potts, 2010). Climate change is also a driving factor in the destruction of natural ecosystems and has harmful effects on wild bee populations. As pollinator species decline in both population and diversity, so too will the biodiversity of the plants that rely on pollinators. This decline can be substantially detrimental to both natural ecosystems and those in which humans are directly involved such as agriculture where pollinators contribute more than 24 billion dollars to the United States agricultural economy (National Archives, 2014).

2.1.3 *Bombus*

In New England alone, there have been 401 wild bee species reported across 6 different families; Andrenidae, Apidae, Colletidae, Halictidae, Megachilidae, and Melittidae. *Bombus*, the bumblebee, is 1 of 14 different genera of the Apidae family (Dibble, 2018). The *Bombus* genus includes 250 species found around the globe, 21 of which are present in the United States. *Bombus* species are unique in that they have the capability to thermoregulate and heat themselves to the minimum temperature of flight, 30℃, which allows them to be active from spring to fall. Unlike honey bees, bumblebees are able to forage under harsher conditions with their thermoregulatory ability which makes them an extremely important pollinator. While each subspecies of *Bombus* has unique characteristics and abilities, the variation in tongue length strongly influences the bee's flower preferences among all the species, as the ability to maximize nectar intake increases when the flower depth is the same length as the tongue (Colla, 2011). It is also important to note that although tongue length determines the types of flowers that a bee can feed from, bees with the same length tongue can still have unique floral preferences. Figure 2 depicts a guide chart of several Massachusetts natives according to abdomen color and tongue length.

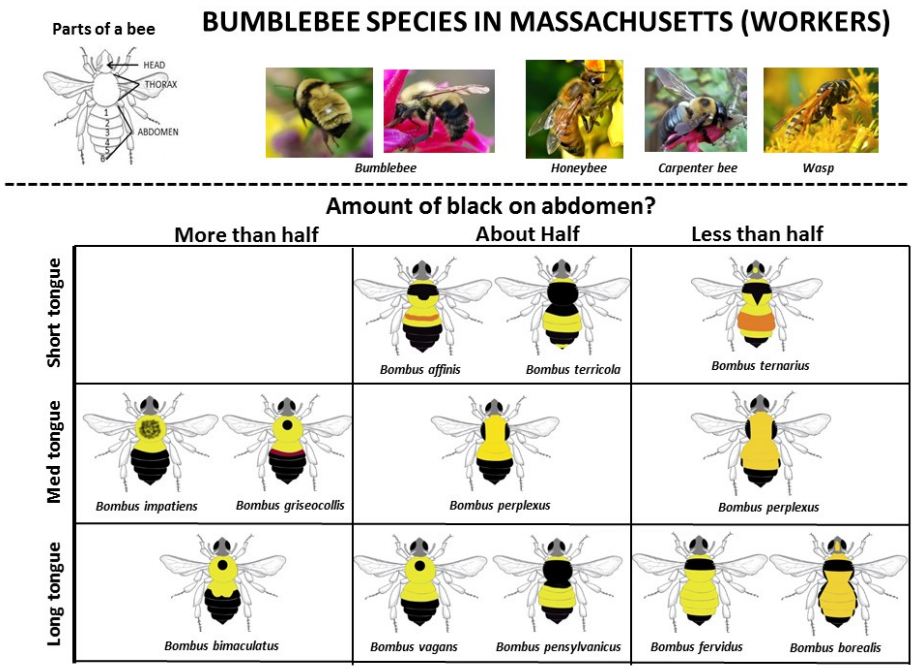

Figure 2: Bumblebee species in Massachusetts (https://beecology.wpi.edu).

In New England, the most common bumblebee is *Bombus impatiens (B. impatiens),* which can also be found all along the east coast. *B. impatiens* is a generalist both in terms of its food choices and habitation. It has a medium length tongue and a shaggy coat, ideal for holding onto pollen (Species Bombus, n.d.). While there is, comparatively, an abundance of *B. impatiens*, other New England natives including *B. affinis, B. terricola,* and *B. fervidus*, are in decline or at risk for extinction. In 2017 the U.S. Fish and Wildlife Service listed *B. affinis* as critically endangered, and later *B. terricola* and *B. fervidus* were recognized as endangered or declining (Dibble, 2018). The eleven *Bombus* species observed in our project belong to five different subgenera: bombus, pryobombus, cullumanobombus, subterraneobombus, and thoracobombus, which are shown in relation to each other on a *Bombus* phylogenetic tree in Figure 3.

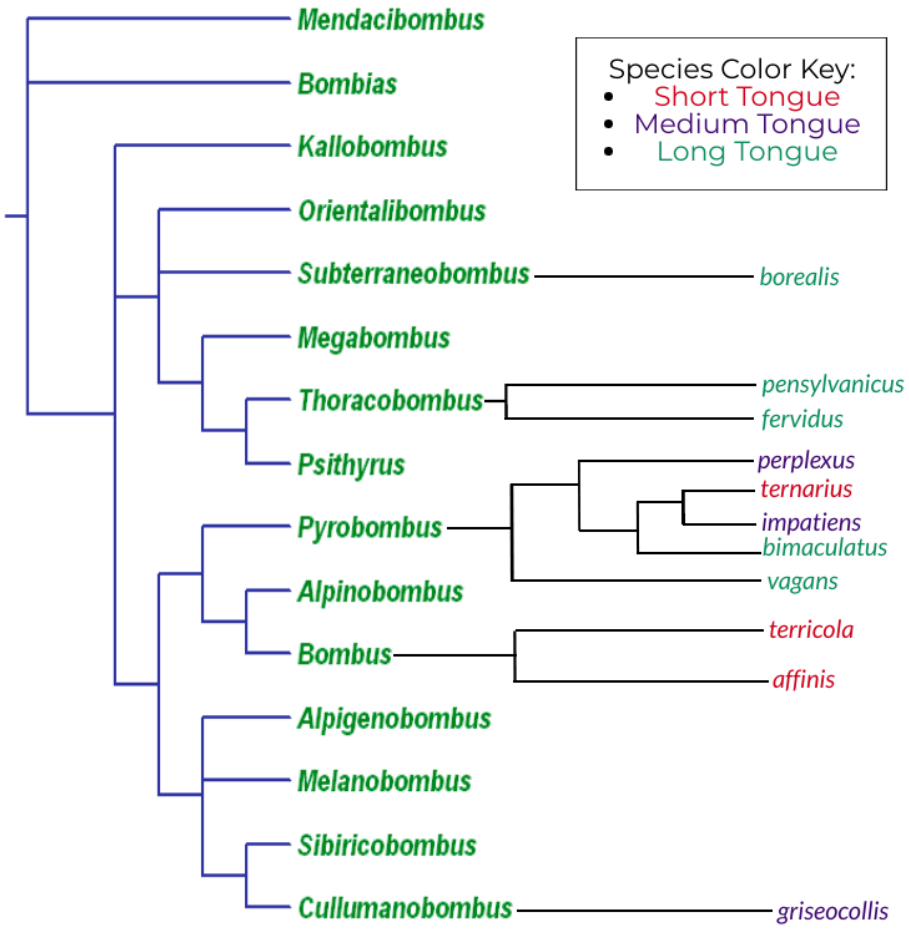

Figure 3: Phylogenetic tree of Bombus subgenera (simplified from 38 to 15 subgenera) and relevant Bombus species based on genetic similarity, modified from (Williams, 2000). Also included is a color key indicating tongue length.

2.1.4 New England Flower Biodiversity

New England is rich in plant biodiversity as the soil type, elevation, climate, and pollinator availability vary around the region. Plant diversity is essential in maintaining the health of an entire ecosystem as plants rarely occur in isolation and are often sensitive to changes in their environment. Species can be classified as either Native or Non-native in relation to where they are found. Native Species are historically indigenous to the specific area or ecosystem, whereas non-native species are introduced from another ecosystem (USDA, n.d.). While not all non-native plant species detrimentally impact native ecosystems, the introduction of a non-native plant species or the loss of a single native plant has the potential to disrupt other vegetation or animal life that it supports. A 2013 study by the USDA Forest Service found that two-thirds of all their monitored forest plots in the

Northeast contained at least 1 non-native species (USDA Forest Service. 2013). Today, 22% of New England's native plants are deemed rare or historic, as human intervention has driven many plant populations to cluster in the remnants of their former regional range. Among these declining plant species, a disproportionately high number requires insect pollination for reproduction (Farnsworth, 2015).

Bombus are some of the primary pollinators in the Northeast region. A variety of floral attributes such as flower shape and color contribute to the preferences of different *Bombus* species. Summarized in Figure 4 are some physical characteristics of several major plant families found in Massachusetts. Although a plant family is composed of multiple species each with their own unique physical traits, often flowers within the same family will share similar general characteristics.

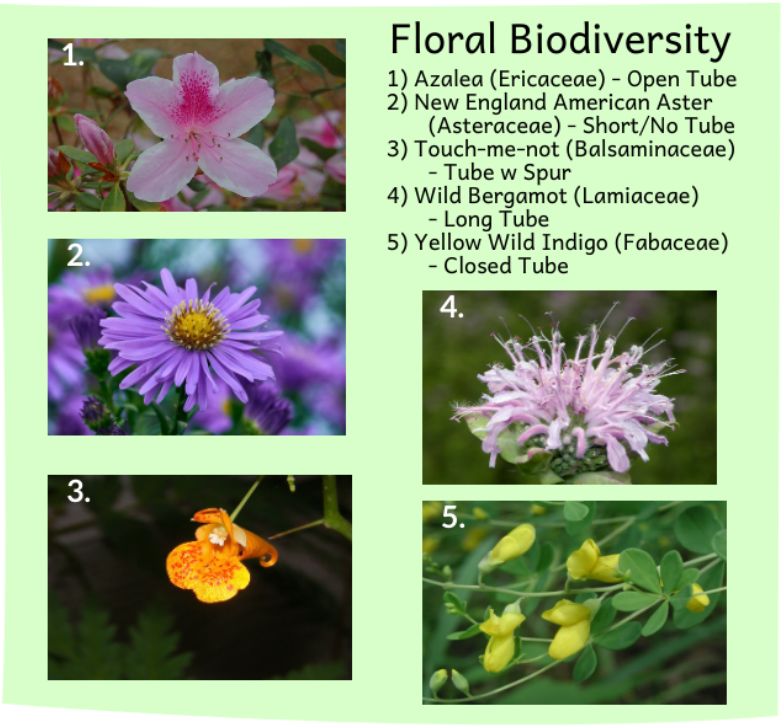

Figure 4: A variety of Massachusetts native plant species (plant family in perens) illustrating the differences in flower shape. Graphic created using images from ([vikisuzan, 2007], [McCranie, 2009],[pixabay],[Gould, 2005],[McGrady, 2018]).

While bumblebees are generalists in nature and many flowers fit the different needs of various tongue-lengthed bees (due to the variety in flower shapes), observations have shown that *Bombus* are especially drawn to flowers with a blue, purple, pink, or yellow color (Attracting Beneficial Bees, 2019). However, it is important to note that a bee's ability to see UV light (and thus colors/patterns not visible to the human eye) certainly plays a significant role in flower choice(Chittka, 1997). Additionally, *Bombus* tends to prefer

perennial plants as opposed to annuals, which other natives and honey bees prefer (Salman, 2019). Perennials regrow each season and typically produce more nectar than annuals, which live for one growing season (Dawson, 2011).

2.2 The Beecology Project

Established in 2016, the Beecology Project aims to educate a broad scope of audiences about pollinator decline across New England, as well as to encourage people across the region to collect and submit ecological data about native bee species to a central bumblebee data repository. The project consists of several components including but not limited to a mobile and web application for collecting and submitting pollinator data; analysis, visualization, and simulation tools; and an education curriculum for high school students (About the Beecology Project, n.d.). Although the project currently focuses on bumblebees, they hope to expand their database to include other struggling pollinators such as butterflies in the future. All of the software tools used for Beecology were developed as part of the Bio-CS Bridge, a related project sponsored by the National Science Foundation. The Bio-CS Bridge project is a team of university and high school-level biology and computer science faculty and students aiming to develop a curriculum to engage and encourage students to use computational tools to analyze biological data and use biology to motivate learning computer science concepts (About Bio-CS Bridge, n.d.).

The data gathered in the Beecology project is stored in a postgres database. The data comes primarily from observations made using the Beecology app, but also from historical data dating back to the late 19th century, primarily from the last few decades, for a total of close to 9,000 observations at the time of this writing. Each observation, or data point, consists of the following attributes: time, year, date, and month observed; latitude, longitude, and elevation of observation; the bee species, its behavior (nectar vs. pollen), its gender, its tongue length, and the months in which it starts and stops being active. About one-third of these data instances include data for the flower the bee was seen interacting with, and all of these came from submission to the app from 2017 on. For these data instances, the following flower characteristics were recorded as well: genus, species, common name, primary shape, and primary color.

2.3 Data Analysis

2.3.1 Exploratory Analysis

Analyzing the network of interactions between bees and flowers presents an opportunity to discover underlying patterns in bees' flower preference as well as to draw conclusions regarding native versus non-native plant pollination. Exploratory analysis differs from other forms of analysis in that it does not have a hypothesis to be proved or rejected; rather, insight is gained as the analysis proceeds and many of the most interesting questions which are answered through the analysis are not even known at the outset. This does not mean, however, that guiding questions or goals do not exist — it just means that the analysis is of an open-ended nature (Chatfield, 1986).

2.3.2 Clustering

Clustering, or cluster analysis, is the process of taking data instances and grouping them based on similarity in characteristics. An example of a basic clustering, with each cluster shaded differently, can be seen in Figure 5. Clustering can help reveal underlying patterns existing in a dataset that are not immediately apparent from basic metrics or other forms of analysis (Tan, 2019).

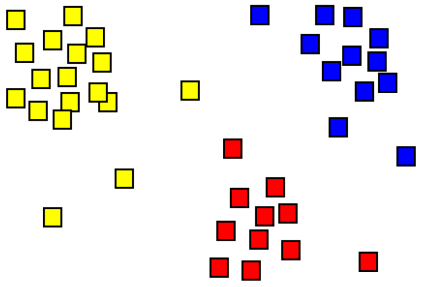

Figure 5: An example of clustered data instances (public domain).

A broad variety of algorithms exist for performing clustering. These algorithms can generally be grouped based on several aspects, one of which is hierarchical versus partitional. Partitional algorithms simply divide up the entire dataset into disjoint clusters. Hierarchical clustering algorithms yield a varying number of nested clusters by proceeding in a bottom-up way, where clusters are grouped into a bigger cluster, or in a top-down way, where clusters are broken down into two or more subclusters. The bottom-up method of hierarchical clustering is referred to as agglomerative clustering, which will be discussed more in Section 2.3.4 (Tan, 2019).

Another distinction between different types of clustering algorithms is exclusive, overlapping, or fuzzy. In exclusive clustering, every data point belongs to exclusively one cluster (hence the name). In overlapping clustering, a data point may belong to multiple clusters. Fuzzy clustering replaces the idea of belonging or not belonging to a given cluster with a weight that describes the extent to which a data point belongs to that cluster. Clustering can also be complete or partial, the former meaning every data point belongs to at least one cluster, with the latter meaning that not all data points are required to belong to one (Tan, 2019).

Clustering algorithms can also be classified in terms of what underlying metrics are being used to separate the data points into their clusters. There are many ways of doing so, only a few of which are discussed here. In centroid-based or prototype-based clustering, such as the k-means algorithm, data points are clustered based on their proximity or resemblance to a given centroid, which can be viewed as the "representative" value for that cluster. In conceptual clustering, data points are clustered based on specific shared attributes that are known ahead of time. In connectivity-based clustering, such as hierarchical clustering, the data points are clustered based on how close they are to one another. In the case of a well-separated clustering, each data point is closer to every other point in its cluster than it is to any other point in other clusters. Note that applying this type of clustering requires separation, or distance, to be defined (Tan, 2019).

2.3.3 Distance and Similarity Metrics

Given that data used in clustering experiments is generally high-dimensional, algorithms such as hierarchical clustering which make use of distances between data points must have a precise mathematical definition for distance. The most commonly used definition is Euclidean distance, which is the square root of the sum of the difference along each dimension squared, as given by the following formula, where x and y are two data points, n is the number of dimensions, i corresponds to a particular dimension, and x_i and y_i being the magnitudes of the data points in the ith dimension:

$$\mathrm{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance is an absolute sum of difference between the points in a Cartesian grid, which ignores diagonals. Cosine similarity is a way of measuring a distance by taking the cosine of the angle between the two data points, equal to the dot product of the normalized vectors composing the two data points. In clusterings that do not rely on a mathematical distance, such as conceptual clustering, similarity may be defined in a myriad

of ways, such as the number of attributes the data points have in common, with certain attributes perhaps weighted more highly than others in this determination (Tan, 2019).

2.3.4 Agglomerative Hierarchical Clustering

Agglomerative clustering is the bottom-up form of hierarchical clustering. At initialization, each data instance forms its own cluster, known as a singleton. The two closest clusters are then merged together, using the preferred distance metric (often Euclidean). Then, the next two closest are merged, and this continues until there is only one cluster, consisting of the entire dataset (Manning, 2008). The specific points during the process at which clusters group together is often displayed via a dendrogram. Analysis can be performed at any point along this dendrogram, by isolating the clusters that were in existence at that point in the algorithm's course. The key value of agglomerative clustering is that, for a set of *n* data points, it allows for analysis on *n* different clusterings, some of which may provide different insights from others. This is an advantage that many other clustering algorithms, such as k-means, lack (Manning, 2008).

2.3.5 Data Preprocessing and Encoding

Generally, a dataset has a number of attributes, which will be clustered on. It is good practice to ensure that all data points to be used in the clustering have defined values for each attribute, with unknowns or blanks excluded. Additionally, depending on the clustering experiment being run, some attributes may be excluded, either to focus analysis on a few attributes, or to see the effect on the clustering when a specific attribute is excluded. Some data points may be excluded as well, often because they are above or below a threshold of interest. This can help eliminate outliers and lead to a more rigorous clustering.

In the simplest cases of clustering, all data exists as numerical values in Cartesian coordinates, where distances can be objectively and mathematically calculated using a distance metric such as Euclidean or Manhattan. Clustering becomes more difficult when dealing with data that is non-numeric or needs scaling. The need for scaling arises when the range of the attributes in data differs. In a clustering where each attribute is intended to have equal weight, the data values under each attribute should be normalized so that each attribute has the same mean and standard deviation. Alternatively, in cases where the distribution is relatively similar across all the attributes, it may suffice to use min-max scaling. This approach takes the minimum value for the values under that attribute, sets it to 0, sets the maximum value to 1, and scales all the other values in between accordingly (Garcia, 2015).

Occasionally, one might wish to weigh one attribute more heavily than others in the clustering. This can be done by scaling that attribute to have a wider range. Alternatively, attributes which are intended to be deemphasized in the clustering can be scaled to smaller ranges, or if their impact is not desired at all, excluded from the clustering.

Any non-numeric data must be converted to a numeric form in order to be clustered. For data that is ordinal (can be ordered or made to fit on a spectrum), it can be converted to a range of properly-scaled numbers. For example, colors could be converted to a wavelength or a gradient pixel value, and then scaled. Data without a natural ordering is best hot-encoded: this means that every unique type of that attribute becomes its own dimension, with each data point having a value of 1 for the dimension that describes its own type and 0 for all the other related dimensions (Cerda, 2018). As an example of this, if a United States customer datasheet listed the state each customer was in, a customer from Montana would end up with a 1 in the dimension for Montana and a 0 in the other 49 dimensions, each of which correspond to one of the other 49 states. As can be seen from this example, hot-encoding often massively increases the dimensionality of the data.

2.3.6 Data Dimensionality Reduction

Working with high-dimensional data is difficult both from a comprehension and a visualization standpoint, so dimensionality reduction becomes necessary in order to perform meaningful analysis. Many methods exist for clustering higher-dimensional data and then reducing it to two or three dimensions, among them t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), multidimensional scaling (MDS), and spectral embedding (SE) for the purpose of visualization. All of these algorithms rely on complex linear algebra, making use of various tradeoffs to reduce so many dimensions, leading to various advantages and disadvantages. t-SNE operates by assigning every pair of data points a similarity score, and then places similar objects closely while avoiding placing dissimilar objects near each other, generally making use of Euclidean distance (Shah, 2019). UMAP functions similarly to t-SNE, but with some specific underlying assumptions which lead to it preserving more of the global structure (Dorrity, 2020). MDS records a distance matrix for the high-dimensional structure and attempts to preserve these distances as exactly as possible while transforming each data point into a two- or three-dimensional space. This leads to very accurate spacing between data points, but may fail to preserve global structure (Hout, 2012). SE makes use of the set of the eigenvalues of a calculated similarity matrix across all the data points to perform its reduction, and is also not ideal for preserving global structure (Bengio, 2016).

These dimensionality reduction techniques are often used in conjunction with other clustering methods. For example, performing agglomerative clustering on a

high-dimensional dataset and then using one of the dimensionality reduction techniques allows for visualization of that original clustering, assuming that color, shape or some other property is used to show which original cluster each data point belongs to. As each technique has advantages and disadvantages, it is best to apply several to get a better understanding of the clustering.

2.3.7 Evaluating Clusters

While clustering provides distinct groups of similar data, analysis of these clusters is necessary in order to evaluate the resulting clusters and to discover any patterns or draw meaningful conclusions about the data being interpreted. There are a variety of techniques to analyze the results of clustering analyses, including manual investigation of the clusters' contents, visualization, and quantitative analysis.

Manually examining the data points and summarizing the patterns in attributes between clusters is the most straightforward way to discover patterns in the data, but is also time-consuming and prone to human error. This can include calculating a variety of metrics on numerical data: mean, median, mode, variance, range, as well as examination of minima, maxima, outliers, and anything else that stands out about the data.

Visualization, generally requiring dimensionality reduction, is another powerful method for evaluating clusters. This can take the form of simple scatter plots, but dozens of visualization techniques can help lead to more insight, from violin plots to 3D displays to dendrograms. Two features are important for visualizing clusters: first, that it is readily apparent which cluster a data point belongs to. This means that if something is being visualized other than the layout of the clusters (such as a scatterplot comparing clusters with an attribute), color or some other property must be used to indicate which cluster each data point belongs to. Second, it is important that the attributes of a data point are either visible or easily accessible, as otherwise the amount of information that can be gleaned from the visualization is minimal.

Finally, there are quantitative methods for evaluating clustering. These fall into two categories: extrinsic and intrinsic. Extrinsic methods compare the clustering to a "ground truth" (perhaps using known information that was left out of that particular clustering, or to another clustering considered to be ideal), while intrinsic methods analyze the clustering itself, based on how close points within a cluster are to each other, and how far from points in other clusters they are. Especially in exploratory analysis, it is only through a combination of investigation, visualization, and quantitative methods that meaningful conclusions can be drawn from the cluster analysis.

Chapter 3: Methodology

3.1 Biological Classifications

Early in our project, we consulted with Dr. Robert Gegear, the founder of the Beecology Project, as well as other members of the Beecology team and found that additional biological information was necessary to perform in-depth analysis of the network. Classifying the floral attributes of native versus non-native, native ecoregion, and plant family, along with categorizing *Bombus* subgenera, provided valuable information to further the analysis of the bee-flower network and to add to the database.

3.1.1 Classifying Plant Native Status and Ecoregion

Prior to our project's contributions, the Beecology database contained information for each plant species entry such as the flower's shape, main color, and bloom period. It did not, however, have any information regarding the plant's native ecoregion (the general geographical region from which the species originates), nor did it identify whether or not a particular species was a native to where it was sighted. To aid in both our exploratory analysis and to provide useful biological information to the database, we classified each plant species as either native or non-native and provided the species' native egoregion.

Because we were mainly focused on the local bee-flower network interaction, the scope of our project was specific to the U.S. state of Massachusetts and the plants were classified as either native or non-native to MA specifically. It is important to note that plants do not follow state borders; thus, any plant species that were found to be native to bordering states (such as New Hampshire or Connecticut) were also classified as native to MA.

After running a simple SQL script on the entire Beecology dataset to remove duplicate floral entries, we generated a list of all the known plant species that had been entered in the database. For each of the 259 unique plant species, we manually cross-referenced verified botanical sources (primarily Go Botany: Native Plant Trust and Lady Bird Johnson Native Plant Database) to classify them as either native or non-native and to identify their native ecoregion. If the plant was not native to anywhere in the United States, the native ecoregion was labeled as the main country(ies) or continent(s) of origin. If the species was only native to the greater MA/ New England area, the native ecoregion was classified as Northern-Appliacian-US. In other cases where there were several native ecoregions, each was identified.

3.1.2 Plant Family

Identifying each plant species' scientific family was also an important contribution both to our project's analysis and the Beecology database. Although the database already had both plant species and plant genus as a flower attribute, due to the very high number of genuses represented (several hundred), we came to the consensus that adding a taxonomic attribute another level higher would enable us to more easily manage all the data and find patterns later on in our analysis. Using the same plant list as previously generated in our native/ non-native classification, we manually referenced a variety of botanical sources to classify each plant species' plant family.

3.2 Data Description

The database table of bee-flower interactions we worked with consisted of 8687 rows, each corresponding to an observation. In this context, an observation refers to an observation of a bumblebee. For data collected from 2017 on, using the Beecology app, the vast majority involved the bee photographed interacting with a flower, which was then submitted to the Beecology project (with accompanying data such as time and location, and information about the flower where it could be determined) and became a row in the Beecology database. The exact preprocessing we did before clustering is discussed later, but our clustering experiments made use of only nine of the twenty-four columns existent in the database. A description of each of the nine we worked with (month, elevation, bee species, behavior, gender, tongue length, flower shape, flower native classification, and flower family) is provided here. An example of the data can be seen in Figure 6.

Attributes

Observations

month	elevation	species	behavior	gender	tongue_lengt	flower_shape	flower_native	flower_family
9	59	fervidus	nectar	female	long	Long Tube	Non-Native	Fabaceae
9	129	impatiens	nectar	female	medium	Closed Tube	Native	Plantaginaceae
9	60	impatiens	nectar	female	medium	Closed Tube	Native	Campanulaceae
9	60	impatiens	nectar	female	medium	Closed Tube	Native	Campanulaceae
9	58	impatiens	nectar	female	medium	Long Tube	Non-Native	Asteraceae
9	58	impatiens	nectar	female	medium	Long Tube	Non-Native	Asteraceae
9	58	impatiens	nectar	female	medium	Long Tube	Non-Native	Asteraceae
9	58	impatiens	nectar	female	medium	Long Tube	Non-Native	Asteraceae
9	73	impatiens	nectar	female	medium	Closed Tube	Native	Campanulaceae
9	21	griseocollis	nectar	female	medium	Long Tube	Non-Native	Asteraceae
9	197	impatiens	nectar	female	medium	Closed Tube	Native	Plantaginaceae
9	129	vagans	nectar	female	medium	Closed Tube	Non-Native	Fabaceae
9	129	impatiens	nectar	female	medium	Long Tube	Non-Native	Lamiaceae
9	60	impatiens	nectar	female	medium	Closed Tube	Native	Campanulaceae

Figure 6: Sample rows from the dataset, displaying all attributes for 14 observations.

The month column consisted of integer values corresponding to the month of the Gregorian calendar in which the observation was made. Because bumblebees are not active in the winter, not all months were represented, with the values ranging from 4 (April) to 11 (November), and the vast majority being either 6 (June) or 7 (July).

The elevation column consisted of the elevation at which the observation was recorded, measured in meters above sea level. Decimal accuracy was provided to 5 places; however, in the vast majority of cases, only one nonzero value was provided past the decimal point, meaning the true accuracy is to the nearest meter or decimeter. The vast majority of observations were found at low elevations of less than 150 meters. (This was especially true after preprocessing was done due to data being constrained to the Massachusetts area, which will be discussed more later.)

The bee species column consisted of the bumblebee species involved in the observation. There were ten different species: *B. impatiens*, *B. borealis*, *B. ternarius*, *B. griseocollis*, *B. bimaculatus*, *B. vagans*, *B. perplexus*, *B. fervidus*, *B. terricola*, and *B. pensylvanicus*. (An eleventh species, *B. affinis*, was present in the overall data but did not occur in our dataset due to not having been found in the Massachusetts area recently.)

The behavior column consisted of either “nectar” or “pollen,” based on which of these the bee in question was pursuing when observed. Gender was also a binary attribute, labeled as either “male” or “female.”

Tongue length was classified into “short,” “medium,” and “long,” and was entirely determined by the bee species involved. Due to only some less-common species (such as *B. ternarius*) having short tongues, medium and long were far more common.

Flower shape was categorized into five possible values, defined entirely by the flower species: “Long Tube,” “Tube with Spur,” “Open Tube,” “Closed Tube,” and “Short/No Tube,” of which “Tube with Spur” was uncommon, and the other four all well-represented.

Flower native classification also had binary values (“native” or “non-native”), with similar numbers of each. Flower family referred to the scientific family name for that flower, when it could be determined. The information in both of these columns came from work we did, and is accordingly discussed further in the results.

3.3 Exploratory Analysis

In our search to uncover any naturally occurring patterns or themes in bee-flower preference or in the interaction network itself, we utilized hierarchical clustering to group the data instances according to their similarities. As our clustering analysis was exploratory, we went through multiple iterations of different preprocessing criteria to ensure that each attribute was weighted correctly in how we wanted to analyze the data. Due to the nature of exploratory analysis, there was no specific hypothesis we sought to address; rather, there were several questions which guided us in our work. For each iteration of the analysis, we investigated the compositions of each cluster, performed quantitative analyses and used other visualization techniques in order to draw conclusions and find naturally occurring patterns in the data.

3.3.1 Guiding Questions

Our analysis was guided by several questions of interest. The one we invested the most time into was, “Do *Bombus* prefer native or non-native flowers?” There were several associated questions:“Which species tend to prefer non-native flowers the most?” “ Which flower families were frequented the most in general by each bee species?” Another important guiding question was, “Do *Bombus* species and genera influence flower choice?” Perhaps the most important question was, “Do naturally occurring clusters exist in the data?”

3.3.2 Hierarchical Clustering

With the exception of a few early experiments using k-means, all the clustering experiments run during the project made use of hierarchical clustering, specifically the bottom-up implementation of it known as agglomerative clustering. The

AgglomerativeClustering class from the scikit-learn Python library was used to implement the clusterer and then fit to our data using built-in methods.

3.3.2.1 Data Preprocessing

The bee-flower interaction data was downloaded directly from the Beecology website and then imported into our Python projects as a CSV file, where it was converted into a dataframe using the Pandas library. The starting dataframe for all the clustering experiments had 8687 rows in it, each corresponding to an observation existing in the Beecology database. Once all our additional data about the plants (family, native status, and others mentioned above) was added, each row contained 24 columns, each for a different attribute, although this would be narrowed down to the nine columns of interest discussed in Section 3.2.

The first preprocessing action was to remove all data before the year 2017, as 2017 was the first year in which the Beecology app was deployed, allowing for submitting photos of bumblebees on flowers. Before this, the vast majority of data points had no plant information, making them irrelevant for our analysis. This reduced the dataset down to 3616 observations. Next, the data was whittled down to only observations from the approximate area of the state of Massachusetts, by excluding everything not within the latitude range of 41.2N and 42.9N and longitude range of 69.9W and 73.51W. The dataset then consisted of 3501 observations — as most of the app userbase is in the general Massachusetts region, this did not remove many data instances.

At this point, the dataframe was reduced to the columns of interest, which in the majority of clustering experiments consisted of month, elevation, bee species, behavior, gender, tongue length, flower shape, flower native classification, and flower family. Descriptions of data in each of these columns was provided in Section 3.2.

Next, any rows with unknown or missing values were removed, and the dataframe index reset. This left us with 2366 data points; this large reduction was primarily due to observations where the associated flower had not been identified. However, for the majority of the analysis, data points with *B. impatiens* as the bee species were also removed. This is because it is by far the most abundant species in the area, and dominates all the clustering experiments if left in. Consequently, the majority of analysis was done without it (although a few experiments were run with it included, or with exclusively *impatiens*, a dataset of 1346 observations). This left 1020 observations for the majority of the clustering experiments.

3.3.2.2 Encoding and Clustering

The dataframe now contained the exact data to be used in clustering. To work with scikit-learn's AgglomerativeCluster class to perform the hierarchical clustering, it had to be scaled and encoded properly. Month was scaled using a MinMaxScaler class from scikit-learn, setting the largest month value to 1, the smallest to 0, and everything else appropriately scaled in between. Tongue length was put on a scale from 0 to 1, with short corresponding to 0, medium to 0.5, and long to 1. Behavior, gender, and flower native classification, all being binary attributes, each had one of their values encoded to 0 and the other value to 1.

The remaining columns (elevation, species, flower shape, and flower family) had to be one-hot encoded. In the case of elevation, this involved a prior step of binning the values into quartiles, so that equal amounts ended up in each. For flower shape, 'long tube' and 'tube with spur' were encoded together, while the other three possible values ('open tube', 'short/no tube', 'closed tube') each received their own dimension. All this one-hot encoding increased the dimensionality of each observation (i.e., number of attributes) to just under 70. The dataframe was now fully ready to be passed into the scikit-learn clusterer, which could be accomplished with only a few lines of code.

3.3.3 Dendrograms

Because bottom-up hierarchical clustering starts with as many clusterings as there are data points and then joins them together, it lends itself naturally to visualization via a dendrogram. This was done partly through the use of a built-in scikit learn dendrogram function, with additional code written to generate a correct linkage matrix from our clustering and then plot the dendrogram using the Python library matplotlib.

3.3.4 Dimensionality Reduction

Due to high-dimensional data being difficult to visualize, dimensionality reduction techniques were used. The four that we used were Multidimensional Scaling (MDS), Spectral Embedding (SE), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). The first three were implemented using the appropriate classes for them from the scikit-learn library, while for UMAP we made use of the umap Python library.

3.3.5 Visualization Techniques

Other than the dendrogram, the first visualization technique used in every clustering experiment was to display the results of the various dimensionality reductions, after they had reduced the data points from the clustering to two dimensions. These points were then displayed as a scatterplot using matplotlib plotting functions. The points were color coded to indicate which cluster they had originally belonged to at a certain point in the hierarchical clustering — generally when there were four clusters exactly, although we also worked with three, five, and six clusters.

To visualize how the clusterings corresponded with the possible values for each attribute, one of the dimensionality reduction color-coded graphs (usually t-SNE) was plotted several times in a row, with the color-coding used to represent the different attributes (instead of cluster membership). For example, one subplot would have each color correspond to a bee species, one would have two colors for native vs. non-native, one would have each color correspond to a flower shape, and so on. This provided a quick and intuitive way to see how the attributes had been divided up across the clustering.

Additionally, more scatterplots were used to compare the breakdown between two attributes; for example, putting elevation on one axis and bee species on the other. For some of the clustering experiments, the centroid of each cluster was calculated, although these values were not used in the analysis.

3.3.6 Manual Analysis

After the clustering was performed, a column was added back on to the original (unencoded dataframe) with the numbers of the cluster each data point corresponded to. (Again, this was usually done with four clusters, but also for some experiments either three, five, or six.) This allowed the dataframe to then be separated into several different dataframes, each one corresponding to the observations in a particular cluster. From there, each cluster's data was outputted to a CSV file, where it could be manually analyzed.

Manual analysis of each clustering was performed to summarize the contents of each cluster and to aid discovery of any underlying patterns. A simple Excel script was run to count the number of different attribute instances of each data instance in the cluster. The data attributes summarized or counted are shown below in Table 1. The summary of each cluster was then recorded on a spreadsheet for later interpretation of the findings. Other metrics (such as average value for a numerical attribute) were also calculated using Excel functions.

General Observation Attributes	*Bombus* Attributes	Floral Attributes
Elevation: highest, lowest, and cluster average	Distribution of *Bombus* species	Number of native & non-native plants
Location: longitude / latitude	Distribution of tongue lengths	Distribution of flower families
Distribution of months	Distribution of gender	Distribution of flower shape

Table 1: Summary of attributes investigated in our analysis; including Native Classification, the distribution of different months, bee species and flower families, the number of each gender bee, the distribution of bee tongue length, the different flower shapes, as well as the highest, lowest, and average elevation for each cluster.

3.3.7 Clustering Experiments

A sequence of experiments was designed, with the initial ones being used to tune parameters and decide on the best approaches for data preprocessing, attribute encoding, and clustering.

Once the general clustering script was fully set up and debugged, there were many slight variations that were made for running different clustering experiments. Elevation was treated in different ways: scaling it from 0 to 1, from 0 to 2, thresholding out higher elevations, and then finally binning it into 4 quartiles and making each its own dimension. Most experiments were run with *B. impatiens* excluded, yet a few were run either with it included, or with just the *impatiens* data. Additional experiments were run with the bee species attribute excluded entirely (i.e., all non-*impatiens* observations present, but the bee species column removed from the dataframe).

Because of the nature of hierarchical clustering, for every clustering experiment there is a choice of how many clusters to use for the analysis. The standard number used was four clusters, yet for most of the experiments we also examined (through visualizations and manual analysis) the clusterings involving three, five, and six clusters.

3.3.8 Clustering by Attribute

In addition to clustering by observation, experiments clustering the data by attribute were designed. The approach can be thought of as essentially clustering on the transpose of the original dataframe depicted in Figure 6. Conceptually, this meant that each singleton

cluster began as one of the 70 or so dimensions, with attributes that are more similar to each other (based on the data) being combined into clusters sooner than others. Several clustering experiments were run with clustering by attribute, varying how the attribute encoding and (to a minor extent) which attributes were included The full dendrogram resulting from each clustering by attribute experiment was displayed so that the horizontal distances between attributes could be analyzed. All clustering by attribute experiments included all of the encoded data, both *impatiens* and non-*impatiens* observations.

The first clustering by attribute experiment encoded all attributes exactly as the clustering by observation experiments had. Later experiments made small changes to the encoding. The data that had been encoded binarily (gender, behavior, flower native classification status) was instead one-hot encoded. This allowed the attributes to be broken up further to aid in analysis; for example, instead of a singleton cluster simply labeled 'native_classification_status,' there were now two clusters, one labeled 'native' and the other 'non-native.' For one experiment, tongue length was changed from a scale into three discrete attributes (short, medium, and long). One experiment was run with month divided into the months themselves (June, July, etc.).

Chapter 4: Results

4.1 Biological Contributions

In this section, we summarize our biological classifications to help further our analysis and understanding of the bee-flower interaction network. Additionally, these classifications were reviewed and added as attributes to the official Beecology database.

4.1.1 Native Plant and Ecoregion Classifications

At the time of our analysis, the Beecology database contained 259 unique flower species entries for bee-flower interactions. We classified 113 native plant species, 136 non-native plant species, and found that 10 of the entries either did not provide enough information, or could be considered as either native or non-native. Summarized in Table 2 are all 58 plant families (later classified) with the varying proportions of native to non-native species. The overall native classification ratio of the plant species is very similar to the rate of non-native observations themselves, as 52% of the recorded plant species are non-native, and 51.4% of the observations in MA were with a non-native plant.

Plant Family	Sp.	Native	Plant Family	Sp.	Native	Plant Family	Sp.	Native
						Papaveraceae	1	0%
Acanthaceae	1	0%	Crassulaceae	2	0%	Phrymaceae	1	100%
Actaea	1	0%	Cucurbitaceae	3	33% (1)	Phytolaccaceae	1	100%
Amaryllidaceae	2	0%	Elaeagnaceae	1	0%	Plantaginaceae	9	44% (4)
Anacardiaceae	1	100%	Ericaceae	8	75% (6)	Polemoniaceae	1	0%
Apiaceae	2	50% (1)	Fabaceae	19	32% (6)	Polygonaceae	3	33% (1)
Apocynaceae	6	83% (5)	Grossulariaceae	1	0%	Pontederiaceae	1	100%
Aquifoliaceae	1	100%	Hostacaeae	1	0%	Primulaceae	1	100%
Asteraceae	69	57% (39)	Hydrangeaceae	4	0%	Ranunculaceae	10	50% (5)
Balsaminaceae	3	66% (2)	Hypericaceae	2	0%	Rhamnaceae	2	100%
Boraginaceae	4	0%	Iridaceae	2	50% (1)	Rosaceae	22	45% (10)
Brassicaceae	3	0%	Lamiaceae	29	24% (7)	Rubiaceae	2	100%

Buxaceae	1	0%	Liliaceae	1	100%	Salicaceae	1	100%
Campanulaceae	2	50% (1)	Lythraceae	2	50% (1)	Saururaceae	1	100%
Caprifoliaceae	2	50% (1)	Malvaceae	4	0%	Saxifragaceae	1	100%
Caryophyllaceae	1	0%	Myrsinaceae	1	0%	Scrophulariaceae	1	0%
Clethraceae	1	100%	Myrtaceae	1	0%	Solanaceae	4	0%
Commelinaceae	1	0%	Nymphaeaceae	1	100%	Styracaceae	2	0%
Convolvulaceae	2	50% (1)	Onagraceae	3	66% (2)	Verbenaceae	2	50% (1)
Cornaceae	1	100%	Orobanchaceae	1	100%	Violaceae	1	0%

Table 2: Plant families in the Beecology database. Plant, Plant families (green), Sp, Number of unique recorded species in each family (red), Native, Native composition based on the # of recorded species in that family (blue).

In identifying each species' native status, we also classified the native ecoregion for later use, as the Beecology project hopes to expand their reach and native classification is relative to the area of interest. Ecoregion of plant species that are not native to the United States were labeled as their continent of origin; in some cases, the native country was identified. Ecoregions of the United States were identified referencing the US Environmental Protection Agency's website, and are broken down into nine regions, as shown in Figure 7 below. Massachusetts falls within the Northern Appalachian ecoregion. The higher level of ecoregion specificity for the US, in comparison to non-US natives, enabled us to add multiple ecoregion classifications to plants that are native to many parts of the United States.

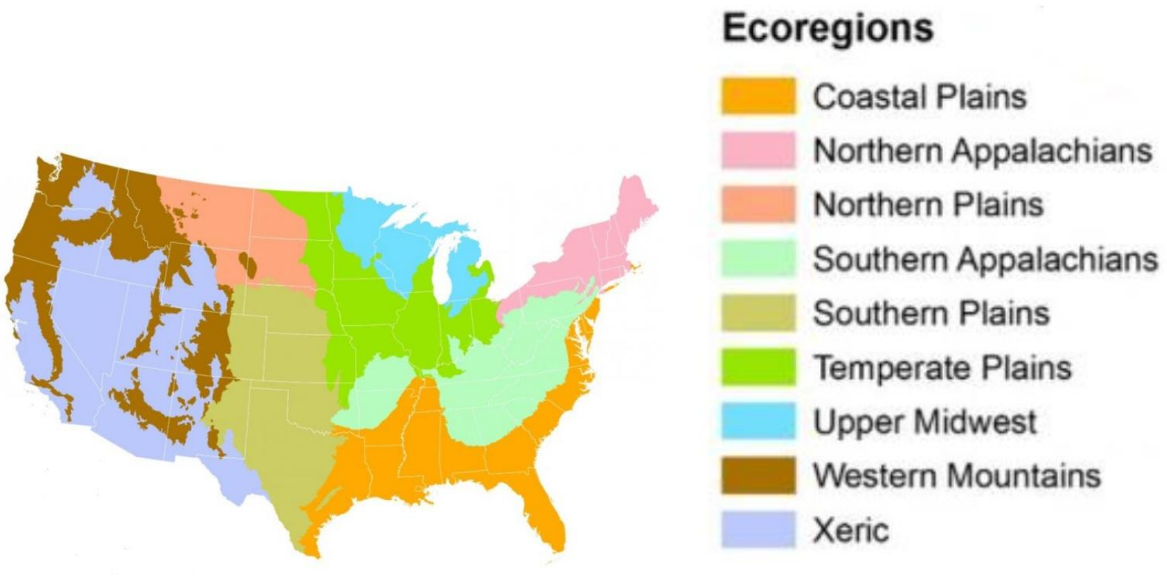

Figure 7: Map of the contiguous United States with a key depicting the nine assigned ecoregions (Environmental Protection Agency, 2020).

Our classifications showed the plants in the database have a variety of native ecoregions, ranging across five continents (Europe, Asia, Africa, North America, and South America), and inclusive of all nine U.S. ecoregions. 106 of the recorded species were completely non-native to the U.S., and 19 species were native to other regions within the US but not the northern appalachians. It is however, important to note that this flower data may not represent the entire range of floral observations with bees or completely encapsulate the bees' preference, as many observations are taken near urbanized areas where there might be a higher frequency of human-introduced, non-native plants. A fully detailed list of all 259 floral entries and their classifications is included in appendix A.

4.1.2 Plant Family Classifications

The 259 plant species in the database belong to 58 different plant families, the largest being Asteraceae which includes 69 recorded species. Interestingly, 28 (49%) of the species were the only recorded member of their family in the database. Table 2 summarizes the identified plant families and the number of observed species in each. Figure 8 is a corresponding phylogenetic tree illustrating the degree of similarity between the different plant families. Note that the number of observations corresponds to the number used in our dataset after data preprocessing, not the number in the overall Beecology database.

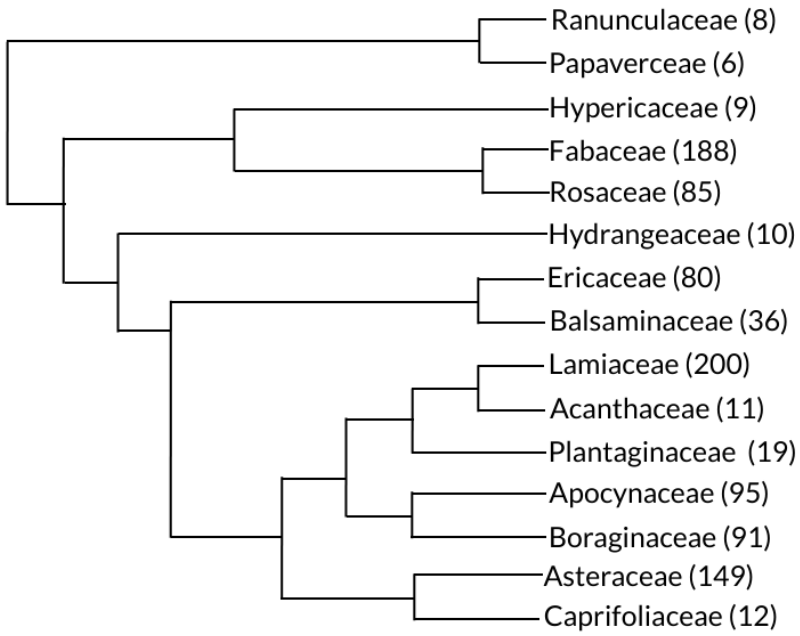

Figure 8: Phylogenetic tree relating the 15 identified plant families which have 5 or more recorded observations in our database. The number in parentheses indicates the total number of observations for that plant family in our analyzed dataset.

4.2 Analysis of the Bee-flower Network

4.2.1 Clustering By Observation

Through the use of hierarchical clustering, visualization techniques, and manual analysis, we identified several patterns in the interaction network and confirmed several predictions about *Bombus* behavior. Over the course of our project, we ran multiple clustering experiments, as summarized below in Table 3, in order to fine-tune our preprocessing criteria so that each attribute was properly weighted and scaled.

Clustering Experiment #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Year	≥2013				≥2017						
Months	4-10, encoded using the MinMaxScaler from 0 to 2										
Species*	A-i			X	A-i		i	A	X	i	A-i
Latitude / Longitude	Not limited		Limited to MA								
Gender	Binary Encoding										
Tongue Length	short: 0, medium: 0.5, long: 1										
Flower Shape	Inaccurately grouped				Grouped into 4 types						
Elevation	Encoded using the MinMaxScaler from 0 to 2										
Native Classification	N/A	Binary Encoding									
Flower Family	Hot Encoded					Hot Encoded, removing families with less than 5 observations					

*Table 3: Summary of our data preprocessing criteria used for each of our clustering experiments. *Species criteria indicating which Bombus species were included as a labeled feature in our clustering; A = All Bombus species, i = B. impatiens, A-i = All species excluding impatiens, X = Only A-i observations included but excluding the labeled species attribute itself from clustering.*

Out of the eleven clustering experiments we performed, we present here in-depth analysis of the three experiments that produced the most interesting results: Experiment #11: All Bombus species except *impatiens*, with *Bombus* species included as a labeled attribute in clustering; Experiment #9: All Bombus species except *impatiens*, but with the *Bombus* species attribute not included in the clustering (i.e., the species column was removed before performing clustering); and Experiment #10: clustering only *B. impatiens* observations. We analyzed *B. impatiens* separately due to the high volume of recorded observations with *B. impatiens* (1310) as compared to all other species (1020 total), to avoid skewing any underlying patterns that might be overshadowed by the sheer amount of *impatiens* data. In addition to the data summarized here, the complete analysis of each

cluster is included in the supplementary materials. The subsections below present the results and analysis of each of these three experiments.

4.2.1.1 All species except *B. impatiens*; *Bombus* Species Included as an Attribute (Clustering Experiment #11)

The first clustering experiment we discuss here included species as an attribute, with *B. impatiens* observations excluded but observations involving all other species included. Flower families present in fewer than 5 observations were excluded. For this clustering experiment, we analyzed four clusterings from the dendrogram, namely the clusterings consisting of three, four, five, and six clusters. Figure 9 shows the dendrogram highlighting the clustering that consists of five clusters. The results of the scatter plot visualizations for these five clusters are seen in Figures 10 and 11.

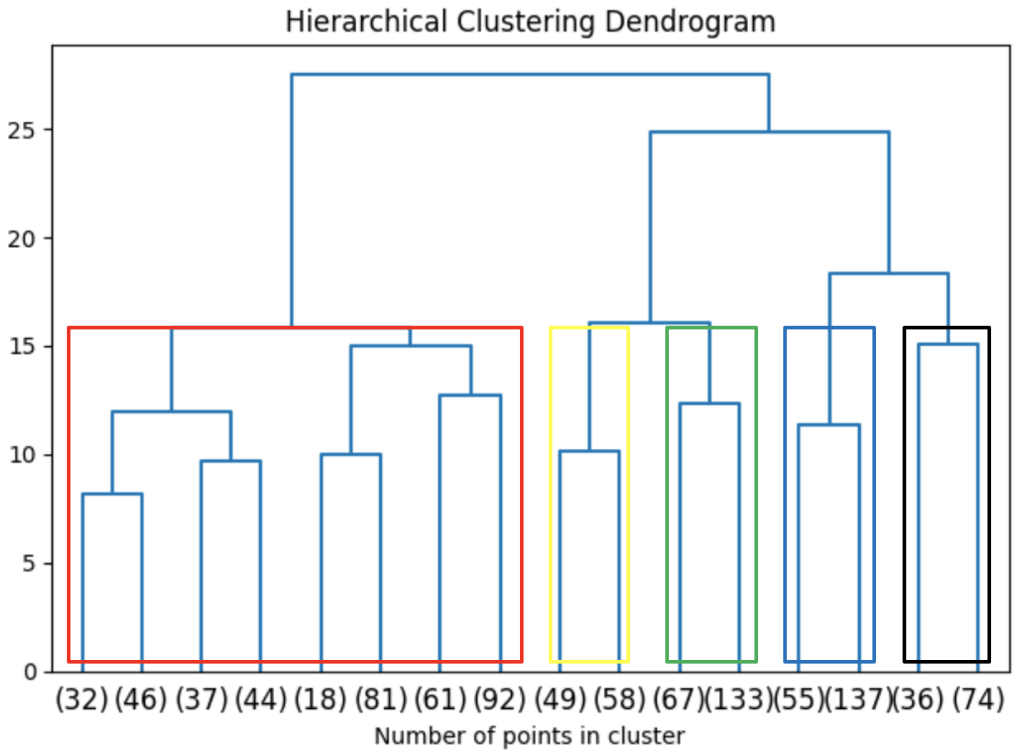

Figure 9: Dendrogram from Clustering Experiment #11. The clustering with five clusters is highlighted with each of the clusters in a different color. Height of a branch in the dendrogram is proportional to the distance between the clusters that the branch joins.

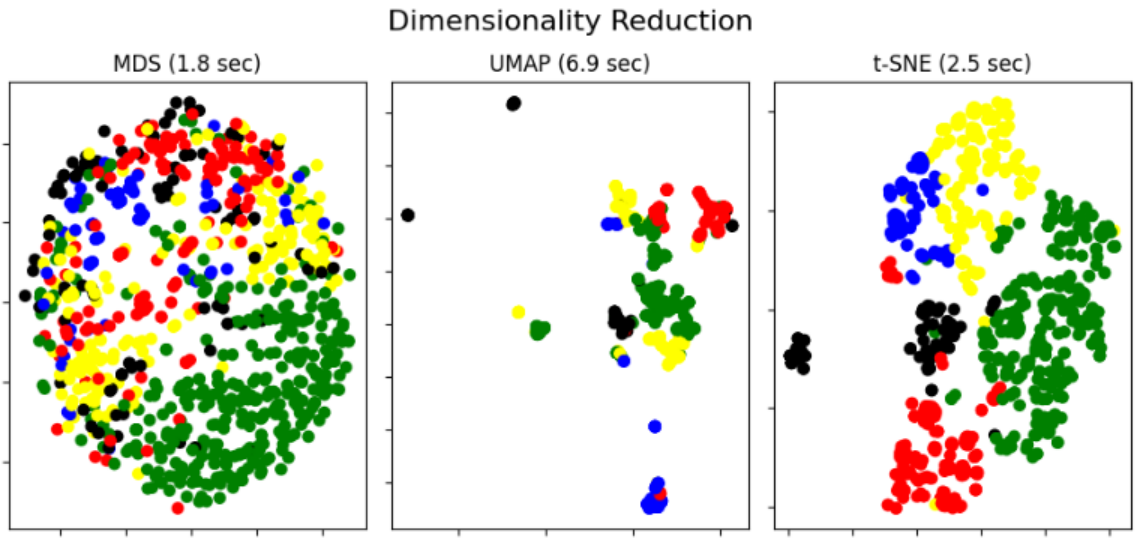

Figure 10: Visualizations of the resulting clustering consisting of five clusters in Clustering experiment #11, where each color corresponds to one of the hierarchical clusters.

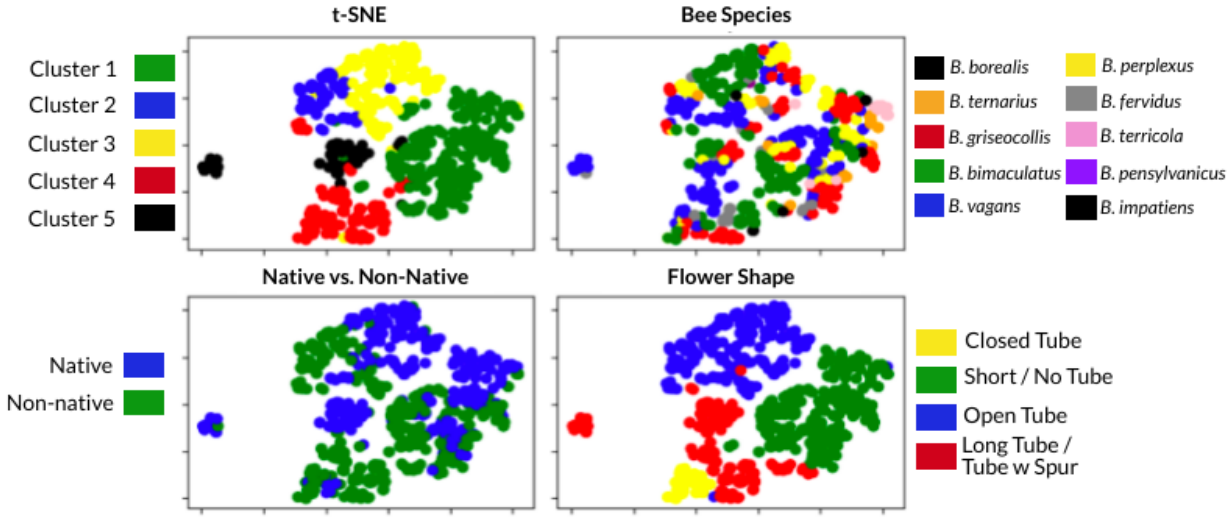

Figure 11: Visualizations of the resulting clustering with five clusters in Clustering experiment #11 , comparing t-SNE plots using four different coloring conventions: by cluster, by bee species, by native classification, and by flower shape, respectively.

Of the dimensionality reduction techniques, the 2D spacing generated by t-SNE separated out the five clusters from the hierarchical clustering the most clearly, meaning it was more in general agreement with the hierarchical clustering than any of the other dimensionality reduction techniques (Figure 10). Thus, for comparison purposes, we next color-coded the t-SNE scatter plot by various attributes, instead of coloring by cluster (Figure 11). Flower shape showed a strong correspondence to the clusters (bottom right of

Figure 11); we observed this correspondence in most of our subsequent findings. Neither bee species (top right of Figure 11) nor native status (bottom left of Figure 11) showed a strong relationship to the clusters.

Manual analysis of these clusters agreed that most clusterings are dictated primarily by flower shape, as the total number of each flower shape was generally concentrated within one or two clusters (Table 4). Interestingly, the distribution of bee species throughout the clusters was consistent with the expected flower shape; for example, long-tongued bees dominated clusters that predominantly contained long tube and tube with spur flower shapes. In addition, while short-tongued bees were rare overall in the database, the highest number of short-tongued bees were grouped in the same clusters that were primarily short/no tube shape. *Bombus* species distribution and flower shape for each cluster is summarized in Table 4. While we did measure other attributes and aspects of the clusters, such as the average elevation and distribution of observations throughout different months, we did not uncover any other notable patterns and thus focused on the most interesting results.

	1	**2**	**3**	**4**	**5**
Native %	46% (190)	52% (161)		37% (113)	
Flower Shape	---				
open tube	19	303		8	
short/no tube	373	1		14	
long tube	19	2		189	
tube w spur		1		36	
closed tube				55	
Species / Tongue Length Distributions	83% (40/48) of all recorded short tongue *Bombus* species	94% (289) long or medium tongue *Bombus* species		79% (240) long tongue *Bombus* species	
	1	**2**	**3**	**4**	**5**
Native %	46% (190)	11% (12)	75% (149)	9% (17)	87% (96)
Flower Shape					
open tube	19	105	198	8	
short/no tube	373		1	10	4
long tube	19	2		119	70

tube w spur				1		36
closed tube					55	
Species / Tongue Length Distributions	83% (40) of all recorded short tongue *Bombus* species		97% (104) long or medium tongue *Bombus* species	97.5% (195) long or medium tongue *Bombus* species	77% (148) long tongue *Bombus* species	84% (92) long tongue *Bombus* species
	1	2	3	4	5	6
Native %	28% (45)	58% (145)	11% (12)	75% (149)	9% (17)	87% (96)
Flower Shape	---					
open tube		19	105	198	8	
short/no tube	159	214		1	19	4
long tube		19	2		119	70
tube w spur				1		36
closed tube					55	
Species / Tongue Length Distributions	No short tongue *Bombus* species (0/159)	83% (40) of all recorded short tongue *Bombus* species	81% (87) long tongue *Bombus* species	96.5% (193) long or medium tongue *Bombus* species	76.5% (147) long tongue *Bombus* species	84% (92) long tongue *Bombus* species

Table 4: Summary of native percentage and the distribution of all flower families throughout nested clusterings consisting of three, five and six clusters in Clustering Experiment #11.

Additionally, many clusters in the five and six cluster analysis were heavily skewed native or non-native, despite the overall ratio in the data being closer to 50-50. This is particularly interesting in the six-cluster analysis, where a large number or the majority of a given flower shape were distributed amongst two clusters, one of which is heavily skewed native, the other skewed non-native, as summarized in Table 4. This evidence indicates that while flower shape still plays a large role in cluster formation, native status is also a determining factor.

4.2.1.2 All species except *B. impatiens*; *Bombus* Species Not Included as an Attribute in Clustering (Clustering Experiment #9)

The second clustering experiment discussed here included the same observations as previously clustered in 4.2.1.1, but excluding *Bombus* species attribute weight in the clusterings. This meant that the exact same observations were clustered on, but the bee species column was removed from the data before performing the clustering. We hypothesized that removing this attribute in the clusterings would uncover new patterns or validate the existence of other trends discovered from clustering with species. For this clustering experiment, we analyzed four clusterings from the dendrogram, namely the clusterings consisting of three, four, five, and six clusters. Figure 12 shows the dendrogram highlighting the clustering that consists of five clusters. The results of the scatter plot visualizations for these five clusters are seen in Figures 13 and 14.

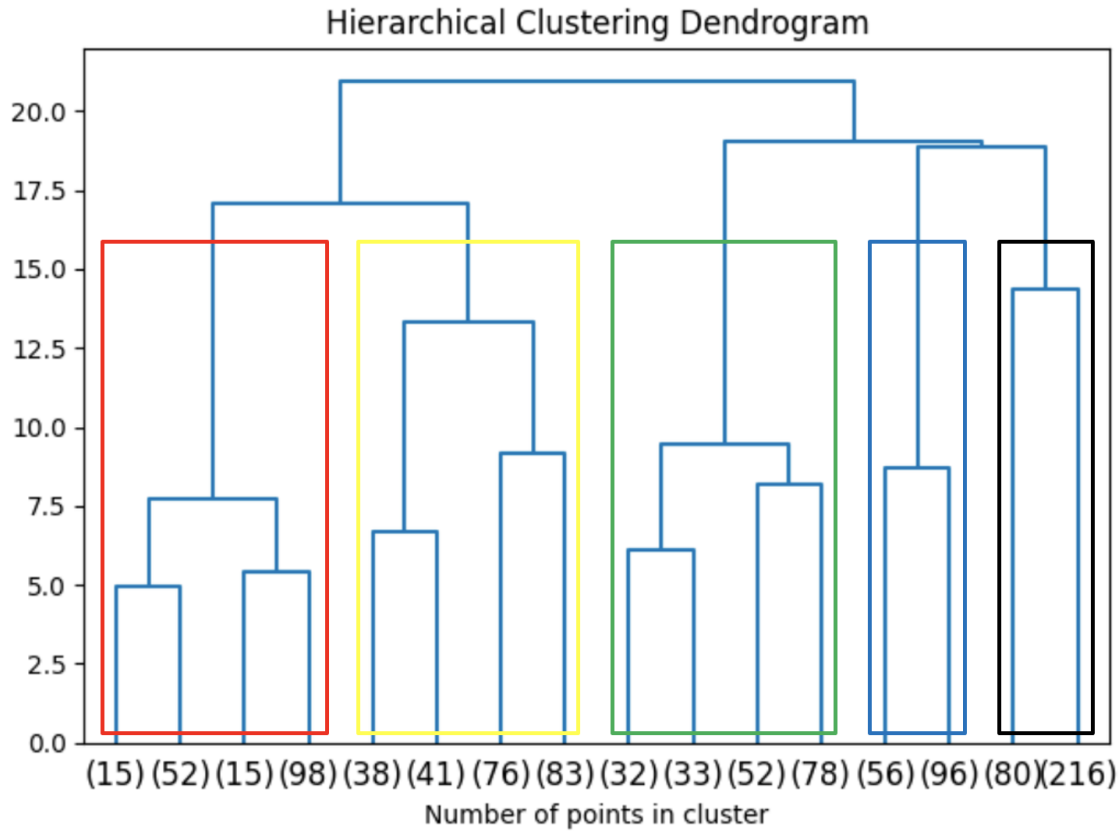

Figure 12: Dendrogram from Clustering Experiment #9. The clustering with five clusters is highlighted with each of the clusters in a different color. Height of a branch in the dendrogram is proportional to the distance between the clusters that the branch joins.

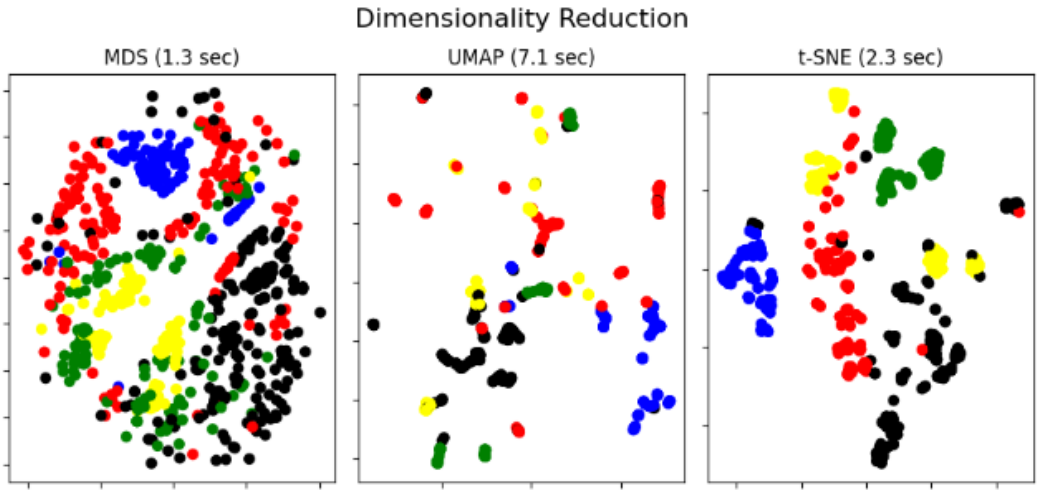

Figure 13: Visualizations of the resulting clustering consisting of five clusters in Clustering experiment #9, where each color corresponds to one of the hierarchical clusters.

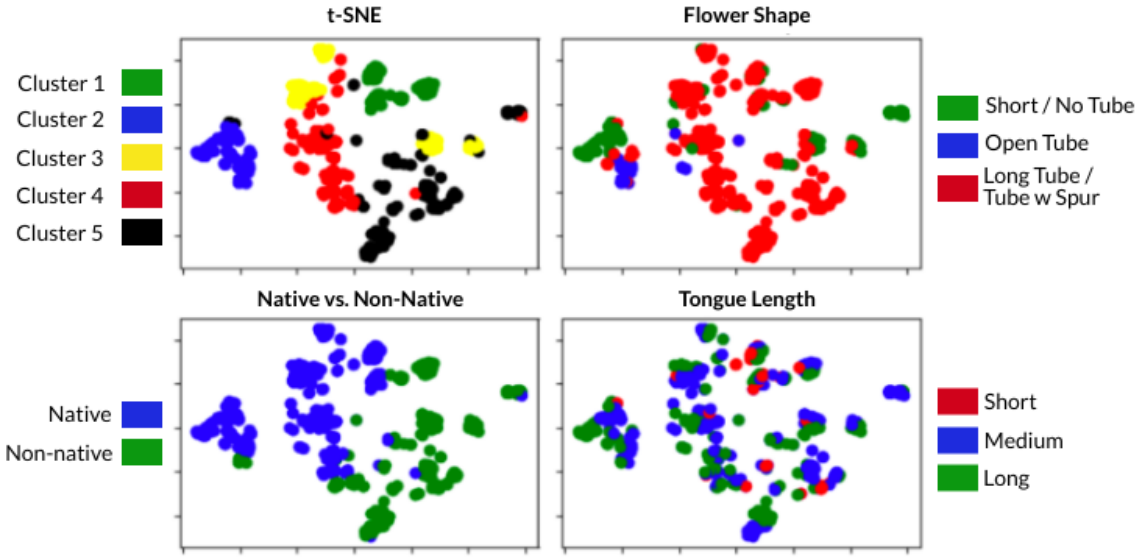

Figure 14: Visualizations of the resulting clustering with five clusters in Clustering experiment #9, comparing t-SNE plots using four different coloring conventions: by cluster, by flower shape, by native classification status, and by tongue length, respectively.

Like with other experiments, t-SNE was more in agreement with the five hierarchical clusters than any of the other dimensionality reduction techniques (Figure 12). Flower shape was somewhat predictive of cluster in these experiments, but interestingly, not to the same extent as when bee species had also been present. Meanwhile, native status had more of an effect than before. If bee species was added back to the dataframe and compared afterwards, it did not play a huge role in the clustering, as the manual analysis showed that species were relatively evenly distributed across each cluster. This is perhaps unsurprising,

given that species did not play a huge role in clustering even when it was included directly, as discussed in Section 4.2.1.1. In Clustering Experiment # 11, tongue length was somewhat correlated with cluster, but this association is not apparent here, likely because flower shape is also playing less of a role in clustering.

Unlike the experiment discussed in Section 4.2.1.1., removing bee species from the analysis yielded clusters that were grouped more by flower family, as opposed to flower shape (Table 5). In the clusterings, there are 1-4 clusters, (respective to 3, 4, 5, and 6 final clusters), that consist of only 1 plant family. In the clustering set of 6, the four lone plant families are Astraceae, Fabaceae, Lamiaceae, and Ericeae, which are some of the most frequently observed plant families in the dataset used for clustering (Figure 8). Interestingly, we see that while Astraceae and Lamiaceae have a relatively even ratio of native to non-native plants, the Fabaceae and Ericaceae plant families have a much more skewed distribution. Since almost the entire, if not the whole, plant family is isolated to the singular cluster, we can conclude that about 56% of the observations with Asteraceae or Lamiaceae plants are non-native species, 91% of Fabaceae observations are non-native, and that 99% of Ericaceae observations are native.

	1	**2**	**3**
Native %	14% (57)	56% (110)	75% (338)
Plant Families	Fabaceae (180)	Lamiaceae (195)	Asteraceae (152)
	Boraginaceae (86)		Apocynaceae (93)
	Rosaceae (79)		Ericaceae (80)
	Other (73)		Other (123)

	1	**2**	**3**	**4**
Native %	14% (57)	44% (85)	44% (67)	92% (271)
Plant Families	Fabaceae (180)	Lamiaceae (195)	Asteraceae (152)	Apocynaceae (93)
	Boraginaceae (86)			Ericaceae (80)
	Rosaceae (79)			Balsaminaceae (35)
	Other (73)			Other (88)

	1	**2**	**3**	**4**	**5**
Native %	9% (16)	17% (41)	44% (85)	44% (67)	92% (271)
Plant Families	Fabaceae (180)	Boraginaceae (86)	Lamiaceae (195)	Asteraceae (152)	Apocynaceae (93)

		Rosaceae (79)			Ericaceae (80)	
		Acanthaceae (12)			Balsaminaceae (35)	
		Other (61)			Other (88)	
	1	**2**	**3**	**4**	**5**	**6**
Native %	9% (16)	17% (41)	44% (85)	44% (67)	99% (79)	89% (192)
Plant Families	Fabaceae (180)	Boraginaceae (86)	Lamiaceae (195)	Asteraceae (152)	Ericaceae (80)	Apocynaceae (93)
		Rosaceae (79)				Balsaminaceae (35)
		Acanthaceae (12)				Plantaginaceae (19)
		Other (61)				Other (69)

Table 5: Summary of native percentage and the distribution of all flower families throughout nested clusterings consisting of three, four, five and six clusters in Clustering Experiment#9.

Additionally, we find that the native percentage of Fabaceae in cluster 1 (of Experiment #9 with five and six clusters) is consistent with our findings in cluster 5 of clustering #11.6, where the overall native percent is also 9% and fabaceae, while not the only plant family included, is the most prevalent (178 fabaceae / 192 total observations in cluster 5).

4.2.1.3 Clustering only *B. impatiens* observations (Clustering Experiment #10)

The third clustering experiment discussed here included solely *B. impatiens* observations. For this clustering experiment, we analyzed four clusterings from the dendrogram, namely the clusterings consisting of three, four, five, and six clusters. Figure 15 shows the dendrogram highlighting the clustering that consists of four clusters. The results of the scatter plot visualizations for these four clusters are seen in Figures 16 and 17.

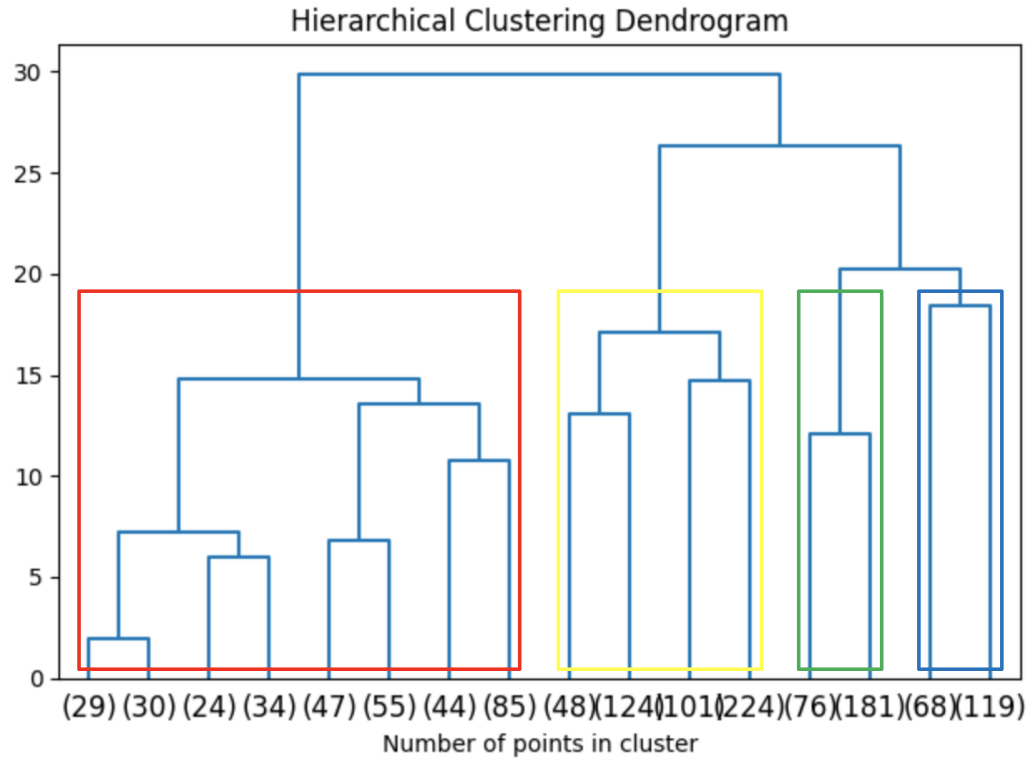

Figure 15: Dendrogram from Clustering Experiment #10. The clustering with four clusters is highlighted with each of the clusters in a different color. Height of a branch in the dendrogram is proportional to the distance between the clusters that the branch joins.

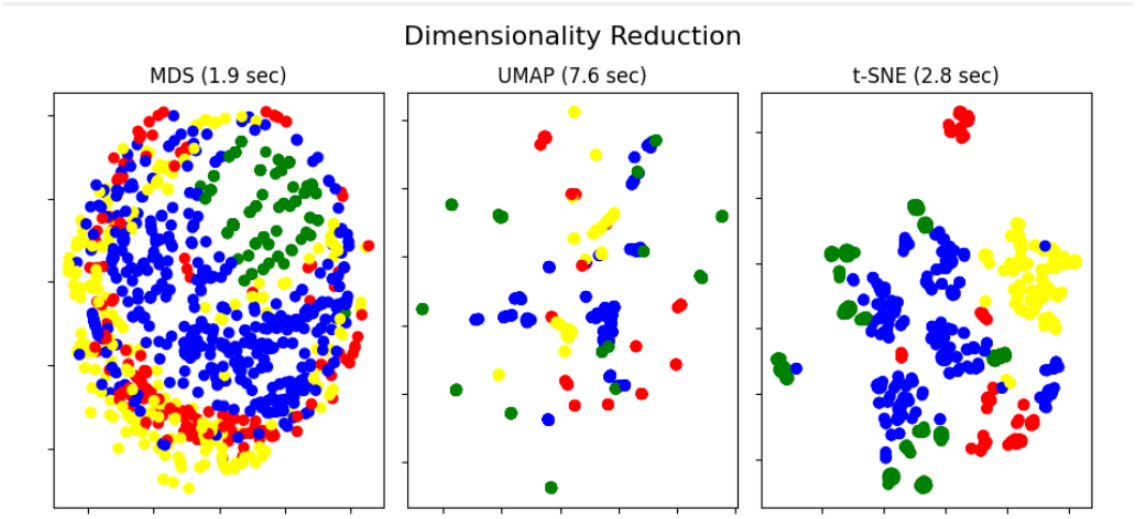

Figure 16: Visualizations of the resulting clustering consisting of four clusters in Clustering experiment #10, where each color corresponds to one of the hierarchical clusters

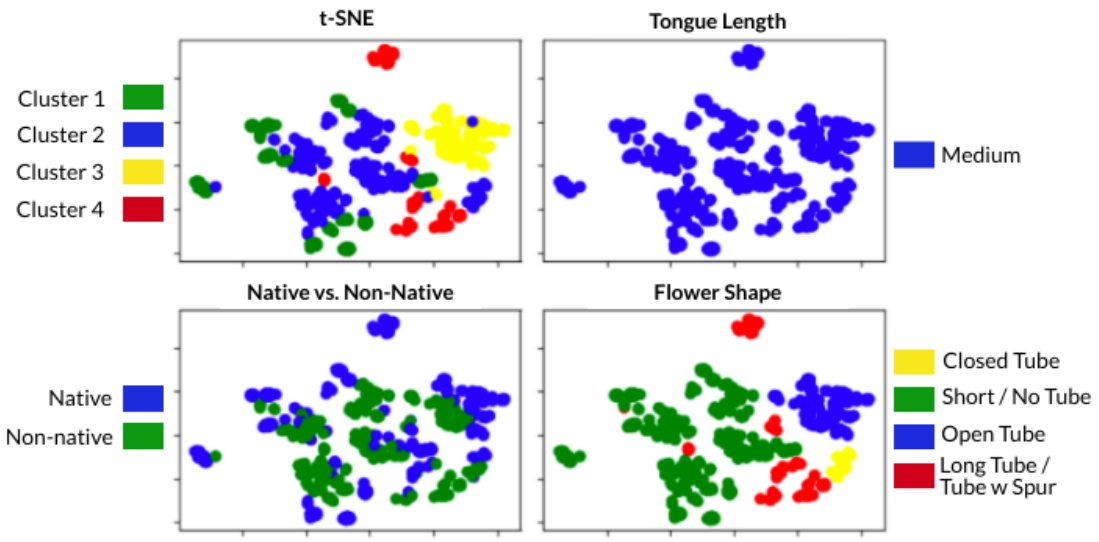

Figure 17: Visualizations of the resulting clustering with five clusters in Clustering experiment #10, comparing t-SNE plots using four different coloring conventions: by cluster, by tongue length, by native classification status, and by flower shape, respectively.

As seen in Figure 17, t-SNE spaced the data points in a way that did the best job of separating out the hierarchical clusters; however, not to the same extent as the clustering experiments discussed in 4.2.1.1 and 4.2.1.2 did. Figure 17, meanwhile, shows that the clustering was highly dependent on flower shape, as was often the case for clustering experiments, though again not to the same extent as discussed in 4.2.1.1. Native status was not as obvious a determining factor as flower shape, but did differ among the clusters. The breakdowns of these clusters are summarized in Table 6.

	1	**2**	**3**		
Native %	58% (263)		28% (109)	63% (295)	
Flower Shape	---				
open tube	247		3		
short/no tube			339	469	
long tube	134			1	
tube w spur	69				
closed tube			48		
	1	**2**	**3**	**4**	**5**
Native %	65%	50% (102)	28% (109)	83% (208)	40% (87)

	(161)					
Flower Shape						
open tube	247		3			
short/no tube			339		249	220
long tube		134			1	
tube w spur		69				
closed tube			48			
	1	**2**	**3**	**4**	**5**	**6**
Native %	65% (161)	90% (62)	30% (40)	28% (109)	83% (208)	40% (87)
Flower Shape	---					
open tube	247			3		
short/no tube				339	249	220
long tube			134		1	
tube w spur		69				
closed tube				48		

Table 6: Summary of native percentage and the distribution of all flower families throughout nested clusterings consisting of three, five and six clusters in Clustering Experiment #10.

As this clustering contains only *B. impatiens* observations, having 81% of the flower shapes being short/no tube or open tube is consistent with the medium tongue-length of the species. Additionally, the distribution of flower shape in this clustering is much less dispersed as several shapes are confined almost entirely to one cluster. This can allow us to make generalizations regarding the likelihood of native classification based on the shape. For example, cluster #2 of the 6-cluster analysis is 90% native and contains exclusively tube with spur, which is not found in any other clusters; thus, we can assume that most observations with recorded flower shape of tube with spur belong to native plant species. It is also worth noting that *impatiens* showed preferences for plant families that were also popularly visited amongst other *Bombus* species (such as Asteraceae, Fabaceae, and Rosaceae), and also favored non-native plants at similar rates as all the other bee species.

4.3 Clustering By Attribute

Clustering by attribute was an additional means of analyzing the collected data. This involved clustering on the transpose of the dataframe, essentially treating each column

(attribute) as a data point to be clustered rather than each row (observation). This resulted in each of the bottom-most (singleton) clusters in the hierarchical clustering consisting of one of the encoded attributes, whether that be a plant family, bee species, or one of the elevation ranges. This allowed for comparing how closely each attribute was clustered to each other attribute. For example, if two bee species visited highly similar types of flowers at the same locations and times of year, the species names should be closely clustered together.

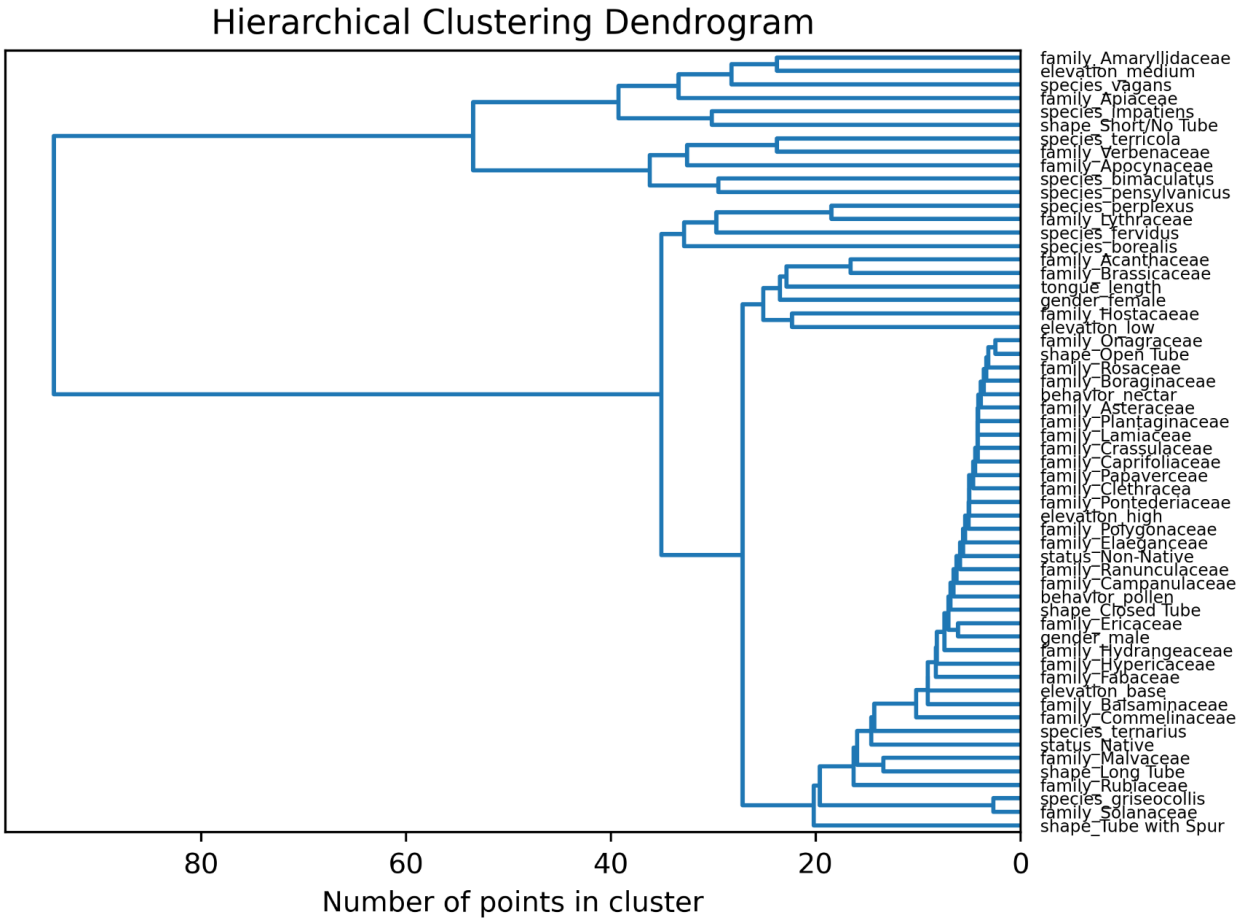

Figure 18: Dendrogram of Clustering by Attribute Experiment

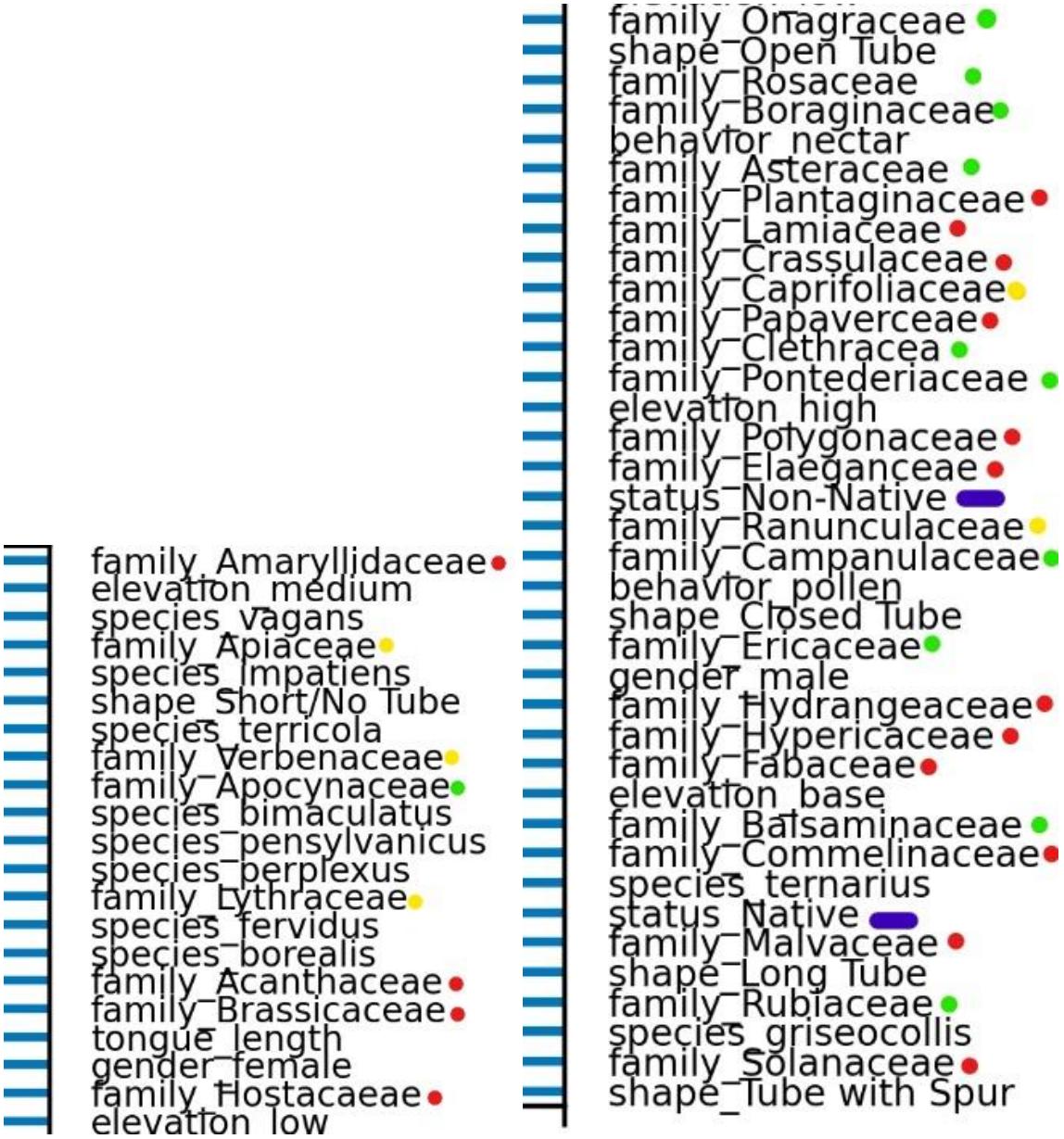

Figure 19: Enlargement of the clustered attributes, with dots next to flower families (red = majority of species in the family are non-native, green = majority native, yellow = approximately even split) and purple dashes next to the two native classification statuses. Note that the colorings are determined by percentage of native species in the family, not percentage of observations involving native species in that family.

The dendrogram for the final clustering by attribute experiment can be seen in Figure 18. A closeup of some of the sections can be seen in Figure 19, where flower families are marked based on how many of their species are native. The most significant result came from comparisons between the bee species and flower families in this dendrogram with

two phylogenetic trees, one for bees and one for flowers. (Figure 3 is contained in the background section while Figure 8 is earlier in the results section.) No significant correspondence was found with either, indicating that genetics may not play much of a role in *Bombus* flower selection.

Two closely related bees often display entirely different flower preferences, as exemplified by *B. ternarius* and *B. bimaculatus*, which are clustered far apart in the dendrogram despite being closely related. Additionally, bees with similar preferences may not be closely related, as seen in the case of *B. perplexus* and *B. fervidus*, which are clustered relatively closely to one another despite not being closely related at all.

Also of note is that the plant families clustered near either native classification status or non-native classification status did not necessarily consist of native or non-native flowers, respectively, as seen by Malvaceae (a flower family with majority non-native species) being clustered closely with the native status. The underlying reasons for this likely require further analysis. One potential explanation is that several Malvaceae species (such as linden trees) have been in the Northeast for centuries and may consequently be more similar to native species than to recent invasive species. Another potential explanation is that the Malvaceae species most visited by the bees happen to be the few native species within the Malvaceae family.

The fact that the bee species are generally interspersed between various flower families, rather than right next to other bee species, is of interest. The majority of flower families clustered near a bee species were often observed with that bee: for example, *B. ternarius* with the flower families Commelinaceae and Balsaminaceae, or *B. griseocollis* with the flower family Solanaceae. The fact that most bee species are not clustered near other species and instead by flowers they prefer provides more evidence for the idea that flower preferences are not determined by how closely related two bees are, and that these preferences can vary widely even between two closely related *Bombus* species.

Chapter 5: Discussion and Conclusion

The analysis of citizen-collected observation data provided valuable insights to help further the understanding of the local bee-flower interaction network. The two main components of our MQP were classifying biological attributes and analyzing the Beecology observations using hierarchical clustering, visualization, and manual analysis.

While many ecological attributes, such as the flower's main color and bloom period were already included in the database, our new classifications of floral native status, native ecoregion, and plant family not only helped inform our analysis of the hierarchical clustering, but also will serve as permanent fixtures in the Beecology database as new means of detailing floral observation entries. Additionally, these contributions to the Beecology database make future analysis of native versus non-native networks outside Massachusetts possible.

In comparing the analyses of the three clustering experiments, we can identify both unique characteristics and similarities in cluster formation and preference patterns. Flower shape played a pivotal role in the clustering, consistent with its known role in shaping bumblebees' flower preferences. Native vs. non-native classification status showed an interesting distribution across a number of clusters, yet overall, none of the bee species showed an obvious preference for native (or for non-native) flowers, perhaps indicating that their preferences depend more on what is available to them. Bee species played comparatively little role in the clustering experiments. Tongue length was somewhat associated with cluster in some experiments (e.g., clustering experiment #11), but this association was not observed consistently. The lack of a strong association could be explained by the paucity of short-tongue length bees in the data, in addition to the generalist nature of medium-tongued bees, which allows them to visit the vast majority of flower species.

Manual analysis also demonstrated that clustering by flower shape was one of the most apparent grouping patterns. Although the one clustering excluding the *Bombus* species did not show the same pattern, many other clustering experiments (not included in final analysis, detailed in supplementary materials) also showed this pattern in cluster grouping. Additionally, it was very common for the closed tube flower shape to be clustered with the long tube / tube with spur flower shape. The distribution of *Bombus* species throughout the clusters was remarkably consistent across all clustering experiments, with the species being well distributed among clusters. This meant that tongue length was also well-distributed, and not a very strong determining factor in the clustering. Flower shape, meanwhile, was consistently a strong determining factor for clustering, with individual shapes often being concentrated almost entirely in one cluster. Because of this distribution of flower shape, some flower families were more highly represented in some clusters than

others. Although the plant family distribution was not consistent across the three clustering experiments, species in the Asteraceae, Lamiaceae, and Fabaceae families were the three most visited overall.

Another particular area of interest was determining if there was a difference in native and non-native plant observations. Our analysis of the clusters showed that the percentage of native to non-native observations (51.5% to 48.5%) was quite close to the ratio of unique recorded native to non-native species in the database (54.6% to 45.4%). The ratio of native versus non-native observations was also remarkably consistent across species, indicating that bees may not be explicitly preferring either native or non-native species, but are simply visiting available plants in their area. This presents opportunities for future study, as it may be that some non-native species are acting as substitutions for native ones that have been crowded out. This might allow for repopulation of these native species through concerted efforts, without having to worry about destabilizing the bee species in their role as pollinators. Due to the biases in the dataset as discussed previously, we cannot draw any firm conclusions about exact bee preferences; instead, our conclusions consist solely of interesting correspondences. It is also of note that there were no large unexpected trends in the data (e.g., one species overwhelmingly preferring non-native plants).

Finally, the clustering by attribute experiments were most notable for what they indicated *not* to be the case. Specifically, they indicated that closely related bees do not necessarily have similar flower preferences, and that closely related flowers do not necessarily attract similar bees. This, when taken in conjunction with the clustering by observation experiments showing that species did not play a significant role in the clustering, strongly suggests that the flower preference of bumblebees is dependent on other factors, likely environmental, not underlying genetics.

References

About Bio-CS Bridge. (n.d.) Retrieved from https://beecology.wpi.edu/.

About Pollinators. (n.d.). Retrieved from https://www.pollinator.org/pollinators.

About the Beecology Project. (n.d.) Retrieved from https://beecology.wpi.edu/.

Attracting Beneficial Bees. 2019. Retrieved from: https://www.gardeners.com/how-to/attracting-beneficial-bees/5024.html#:~:text=Long%2Dtongued%20bumblebees%20are%20attracted,%2C%20monarda%2C%20columbine%20and%20snapdragons.

Bengio Y., Delalleau O., Le Roux N., Paiement JF., Vincent P., Ouimet M. (2006) Spectral Dimensionality Reduction. Studies in Fuzziness and Soft Computing, (207). Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-540-35488-8_28.

Biesmeijer, J. C., et al. (2006). Parallel Declines in Pollinators and Insect-Pollinated Plants in Britain and the Netherlands. Science, 313(5785), 351-354.

Brauman, K. A., Daily, G. C. (2008). Ecosystem Services. Encyclopedia of Ecology, Five-volume set (pp. 1148-1154). Elsevier Inc. DOI: https://doi.org/10.1016/B978-008045405-4.00621-2.

Brohée, S., Faust, K., Lima-Mendez, G., Vanderstocken, G., & Van Helden, J. (2008). Network Analysis Tools: From biological networks to clusters and pathways. Nature Protocols (3), (pp. 1616-1629). DOI: 10.1038/nprot.2008.100

Cerda, P., Varoquaux, G., & Kegl, B. (2018). Similarity encoding for learning with dirty categorical variables. Machine Learning. 107, 1477-1494.

Chatfield, C. (1986). Exploratory data analysis. European Journal of Operational Research, 23(1), 5-13. DOI: https://doi.org/10.1016/0377-2217(86)90209-2.

Chittka, L. (1997). Bee Color Vision is Optimal for Coding Flower Color, but Flower Colors are not Optimal for being Coded -- Why?. Israel Journal of Plant Sciences, 45(203), 115-127. DOI: https://doi.org/10.1080/07929978.1997.10676678.

Colla, S.R., & Packer, L. (2008) Evidence for decline in eastern North American bumblebees (Hymenoptera: Apidae), with special focus on Bombus affinis Cresson. Biodivers Conserv.17, 1379. DOI: https://doi.org/10.1007/s10531-008-9340-5.

Colla, S., Richardson, L., Williams, P. 2011. Bumble Bees of the Eastern United States. USDA Forest Service and the Pollinator Partnership. Retrieved from: https://www.fs.fed.us/wildflowers/pollinators/documents/BumbleBeeGuideEast2011.pdf.

Dawson, J. 2011. What's the Difference Between Annual and Perennial Flowers? Life Science. Retrieved from:

https://www.livescience.com/33266-whats-the-difference-between-annual-and-perennial-flowers.html#:~:text=Annual%20plants%20live%20for%20one,an%20annual%20into%20a%20perennial.

Dibble, A. C., et al. 2018. Bees and Their Habitats in Four New England States. Maine Agricultural and Forest Experiment Station. MISC report 448. ISSN: 1070-1516.

Dorrity, M.W., Saunders, L.M., Queitsch, C. et al. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat Commun 11, 1537 (2020). DOI: https://doi.org/10.1038/s41467-020-15351-4.

Environmental Protection Agency. (2020). Ecoregions used in the National Aquatic Resource Surveys. EPA. Retrieved from: https://www.epa.gov/national-aquatic-resource-surveys/ecoregions-used-national-aquatic-resource-surveys.

Farnsworth, E. 2015. State of New England's Native Plants. New England Wild Flower Society. Retrieved from: http://www.nativeplanttrust.org/documents/3/state-of-the-plants-brief_rev2.pdf.

Frund, J., Dormann, C. F., Holzschuh, A., Tscharntke, T. (2013). Bee diversity effects on pollination depend on functional complementarity and niche shifts. Ecology, 94(9).

Garcia, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. Springer Intelligent Systems Reference Library, 72. DOI: 10.1007/978-3-319-10247-4.

Go Botany: Native Plant Trust. Retrieved from: https://gobotany.nativeplanttrust.org/.

Gould, A. (2005). Wild Indigo, P7020021 [digital photo]. Flickr, retrieved from: https://www.flickr.com/photos/anitagould/25630246.

Hout, M., Papesh, M., & Goldinger, S. (2012). Multidimensional scaling. Wiley Interdiscip Rev Cogn Sci. 4(1): 93-103. DOI: 10.1002/wcs.1203.

Jafarkarimi, H., Tze Hiang Sim, A., & Saadatoost, R. (2012). A Naïve Recommendation Model for Large Databases - Volume 2 Number 3 (Jun. 2012). Retrieved from: http://www.ijiet.org/show-31-222-1.html.

Kremen, C., et al. (2007). Pollination and other ecosystem services produced by mobile organisms: a conceptual framework for the effects of land-use change. Review and Synthesis Ecology Letters. 10: 299-314. DOI:10.1111/j.1461-0248.2007.01018.x.

Lady Bird Johnson Wildflower Center Native Plant Database. Retrieved from: https://www.wildflower.org/plants/.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

McCraine, J. (2009). A Gerorge Taber azalea [digital photo]. Wikimedia Commons, retrieved from: https://commons.wikimedia.org/wiki/File:George_Taber_azalea.jpg.

McGrady, D. (2018). Monarda fistulosa (wild bee-balm, wild bergamot), Dover Plains, NY [digital photo]. Flickr, retrieved from https://www.flickr.com/photos/douglas_mcgrady/41770843790.

McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). "Birds of a Feather: Homophily in Social Networks". Annual Review of Sociology. 27: 415–444.

Mobley, M.W., and R.J. Gegear. 2018. Once size does not fit all: Sex and caste differences in the response of bumblebees to chronic neonicotinoid exposure. PLOS ONE:13(10): e0200041. DOI: https://doi.org/10.1371/journal.pone.0200041.

National Archives and Records Administration. (2014, June 20). Fact Sheet: The Economic Challenge Posed by Declining Pollinator Populations. National Archives and Records Administration. Retrieved from: https://obamawhitehouse.archives.gov/the-press-office/2014/06/20/fact-sheet-economic-challenge-posed-declining-pollinator-populations#:~:text=Pollinators%20contribute%20substantially%20to%20the,and%20vegetables%20in%20our%20diets.&text=Globally%2C%2087%20of%20the%20leading,35%25%20of%20global%20food%20production.

Otte, Evelien; Rousseau, Ronald (2002). "Social network analysis: a powerful strategy, also for the information sciences". Journal of Information Science. 28(6): 441–453.

Penn, J., Hu, W., & Penn, H.J. (2019). Support for Solitary Bee Conservation among the Public versus Beekeepers. American journal of agricultural economics, 101(5), 1386-1400.

Pixabay. Garden Flower Aster [digital photo]. Pixabay, retrieved from: https://pixabay.com/photos/search/american%20asters/.

Potts, S. G., et al. (2010). Global pollinator declines: trends, impacts and drivers. Trends in Ecology & Evolution, 25(6), 345-353.

Rafferty, J. P. & Thompson, J. N. (2020). Coevolution. Encyclopedia Britannica. Retrieved from: https://www.britannica.com/science/coevolution.

Salman, D. 2019. Providing Habitat for Bumblebees: Gardening with A Big Buzz. Retrieved from: https://www.highcountrygardens.com/plant-finder/bee-friendly-plants/providing-habitat-for-bumblebees.

Shah, R., Silwal, S. (2019). Using Dimensionality Reduction to Optimize t-SNE. ArXiv, abs/1912.01098.

Species Bombus impatiens - Common Eastern Bumble Bee. (n.d.). Retrieved from: https://bugguide.net/node/view/56797.

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to Data Mining. Pearson. ISBN-10: 0133128903. Retrieved from: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php.

Thomann, M., Imbert, E., Devaux, C., Cheptou, P. (2013). Flowering plants under global pollinator decline. Trends in Plant Science, 18(7), 353-359.

USDA. (nd). Insects and Pollinators. Natural Resources Conservation Service.

USDA. (nd). Native, Invasive, and Other Plant-Related Definitions. Natural Resources Conservation Service Connecticut.

USDA Forest Service. 2013. In the Northeast, forests with entirely native flora are not the norm. Retrieved from: https://www.sciencedaily.com/releases/2013/04/130430142106.htm.

Van der Maaten, L., Postma, E., Van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. J Mach Learn Res, 10.

Vikisuzan, (2007). Spotted Jewelweed, Touch-Me-Not ~ Impatiens capenis [digital photo]. Flickr, retrieved from: https://www.flickr.com/photos/7721261@N05/2145806144.

Williams, P. (2000). Bumblebees of the world. Natural History Museum, Retrieved from:. https://www.nhm.ac.uk/research-curation/research/projects/bombus/groups.html.

Appendix: Summary of Plant Classifications

Full table detailing our plant classifications including the flower species and flower common name (both previously included in the database), as well as the flower family, the flower's native ecoregion, and the duration of the flower (our newly added classifications).

Flower Species	Flower Common Name	Flower Family	Native Ecoregion(s)	Duration
farinacea	mealy sage	Acanthaceae	Mexico-US Southwest-US	perennial
simplex	chinese bugbane	Actaea	Asia	perennial
tuberosum	garlic chives	Amaryllidaceae	China	perennial
app.	daffodil	Amaryllidaceae	Europe	perennial
glabra	smooth sumac	Anacardiaceae	United-States	perennial
carota	queen anne's lace	Apiaceae	Europe	biennial
aurea	zizia	Apiaceae	United-States	perennial
herbacea	herbaceous periwinkle	Apocynaceae	Europe Asia	perennial
tuberosa	butterfly milkweed	Apocynaceae	Northeast-US Southeast-US Midwest-US Southwest-US	perennial
syriaca	common milkweed	Apocynaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
androsaemifolium	spreading dogbane	Apocynaceae	United-States	perennial
spp.	milkweed	Apocynaceae	United-States	perennial
incarnata	swamp milkweed	Apocynaceae	US-Excluding-West-Coast	perennial
opaca	american holly	Aquifoliaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
chinensis	chinese aster	Asteraceae	Asia	annual
dentata	leopard plant	Asteraceae	Asia	perennial
novi belgii	new york american aster	Asteraceae	East-Coast-US	perennial
cyanus	blue cornflower	Asteraceae	Europe	annual
intybus	chicory	Asteraceae	Europe	perennial
stoebe	spotted knapweed	Asteraceae	Europe	perennial
officinale	dandelion	Asteraceae	Europe	perennial
cyanus	cornflower	Asteraceae	Europe	annual
serriola	prickly lettuce	Asteraceae	Europe	annual
arvense	creeping thistle	Asteraceae	Europe Asia	perennial

	burdock	Asteraceae	Europe Asia	perennial
tataricus	tatarian aster	Asteraceae	Europe Asia	perennial
vulgare	common thistle	Asteraceae	Europe Asia	annual
sphaerocephalus	globe thistle	Asteraceae	Europe Asia	perennial
radicata.	cat's ear	Asteraceae	Europe Asia	perennial
arvensis	field sow thistle	Asteraceae	Europe Asia	perennial
superbum	shasta daisy	Asteraceae	Europe Asia	perennial
spp.	knapweed	Asteraceae	Europe Asia	perennial
rotundifolia	mexican sunflower	Asteraceae	Mexico Central-America	annual
	dahlia	Asteraceae	Mexico Central-America	perennial
	zinnia	Asteraceae	Mexico Central-America	annual
bipinnatus	garden cosmos	Asteraceae	Mexico Southwest-US	annual
spp.	goldenrod	Asteraceae	Northern-Appalachians-US	perennial
spp.	american aster	Asteraceae	Northern-Appalachians-US	perennial
novae angliae	northern blazing star	Asteraceae	Northern-Appalachians-US	perennial
laciniata	cutleaf coneflower	Asteraceae	Northern-Appalachians-US	perennial
pumilum	pasture thistle	Asteraceae	Northern-Appalachians-US	perennial
divaricata	white wood aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
patens	late purple american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
decapetalus	thin leaved sunflower	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
purpureum	purple joe pye weed	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial

horridulum	bull thistle	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	annual
juncea	early goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
laeve	smooth american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
caesia	axillary goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
racemosum	small white american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
rugosa	wrinkleleaved goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
dumosum	bushy american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
flexicaulis	zig zag goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
perfoliatum	boneset thoroughwort	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial

lateriflorum	calico american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
maculatum	spotted joe pye weed	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
macrophylla	large leaved aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	perennial
discolor	field thistle	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	biennial
hieraciifolius	american burnweed	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	annual
puniceum	purple stemmed american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	perennial
umbellata	tall white aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	perennial
canadensis	canada goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial

novae angliae	new england american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
lanceolatum	lance leaved aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
pilosum	awl american aster	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
nemoralis	gray goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
altissima	tall goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
bicolor	white goldenrod	Asteraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
patula	french marigold	Asteraceae	South-America	annual
spicata	blazing star	Asteraceae	Southeast-US	perennial
purpurea	eastern purple coneflower	Asteraceae	Southeast-US	perennial
lanceolata	lance leave tickseed	Asteraceae	Southeast-US	perennial
lettermannii	narrowleaf ironweed	Asteraceae	Southern-US	perennial
spp.	symphyotrichum	Asteraceae	United-States	perennial
annuus	daisy fleabane	Asteraceae	United-States	annual

spp.	sneezeweed	Asteraceae	United-States	annual
graminifolia	grass leaved goldenrod	Asteraceae	US-Excluding-Southwest	perennial
tripartita	three lobed beggar ticks	Asteraceae	US-Excluding-West-Coast	annual
hirta	black eyed susan	Asteraceae	US-Excluding-West-Coast	annual
annuus	common sunflower	Asteraceae	Western-US	annual
spp.	daisy	Asteraceae		perennial
spp.	thistle	Asteraceae		perennial
spp.	tagetes	Asteraceae		
glandulifera	ornamental jewelweed	Balsaminaceae	Asia	annual
pallida	pale jewelweed	Balsaminaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	annual
capensis	touch me not	Balsaminaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	annual
officinale	common comfrey	Boraginaceae	Europe	perennial
officinalis	borage	Boraginaceae	Europe	annual
saccharata	bethlehem lungwort	Boraginaceae	Europe	perennial
spp.	comfrey	Boraginaceae		perennial
	mustard family	Brassicaceae	Africa	
vesicaria	arugula	Brassicaceae	Europe	annual
odoratum	fragrant solomon's seal	Brassicaceae	Europe	perennial
	boxwood	Buxaceae	Europe Asia	perennial
rapunculoides	creeping bellflower	Campanulaceae	Europe	perennial
siphilitica	blue lobelia	Campanulaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
spp.	bellflower	Campanulaceae	United-States	perennial
	weigela	Caprifoliaceae	Asia	perennial

lonicera	northern bush honeysuckle	Caprifoliaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
flos cuculi	ragged robin lychnis	Caryophyllaceae	Europe	perennial
alnifolia	sweet pepperbush	Clethracea	Northern-Appalachians-US Coastal-Plains-US	perennial
virginiana	virginia spiderwort	Commelinaceae	Midatalantic-US Southeast-US	perennial
purpurea	common morning glory	Convolvulaceae	Mexico Central-America	annual
sepium	hedge false bindweed	Convolvulaceae	United-States	perennial
amomum	silky dogwood	Cornaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
arboroseum	garden stonecrop	Crassulaceae	Asia	perennial
spp.	sedum autumn joy	Crassulaceae		perennial
sativus	cucumber	Cucurbitaceae	Asia	annual
pepo	pumpkin	Cucurbitaceae	Mexico Central-America	annual
lobata	wild cucumber	Cucurbitaceae	United-States	perennial
multiflora	goumi	Elaeganceae	Asia	perennial
vulgaris	heather	Ericaceae	Europe	perennial
floribunda	mountain fetterbush	Ericaceae	Midatalantic-US Southeast-US	perennial
latifolia	mountain laurel	Ericaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
spp.	azalea	Ericaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
	azalea	Ericaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US	perennial
fuscatum	black highbush blueberry	Ericaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial

corymbosum	highbush blueberry	Ericaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial
spp	rhododendron	Ericaceae	United-States	perennial
vulgaris	bush bean	Fabaceae	Central-America South-America	annual
repens	white clover	Fabaceae	Europe	perennial
cracca	cow vetch	Fabaceae	Europe Asia	perennial
corniculatus	bird's foot trefoil	Fabaceae	Europe Asia	perennial
arborescens	siberian pea shrub	Fabaceae	Europe Asia	perennial
hybridum	alsike clover	Fabaceae	Europe Asia	perennial
officinalis	yellow sweet clover	Fabaceae	Europe Asia	annual
pratense	red clover	Fabaceae	Europe Asia Africa	perennial
varia	purple crown vetch	Fabaceae	Europe Asia Africa	perennial
australis	blue wild indigo	Fabaceae	Midatlantic-US	perennial
fasciculata	partridge pea	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	annual
perennis	sundile lupine	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial
marilandicum	maryland tick trefoil	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial
tinctoria	yellow wild indigo	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
hebecarpa	northern wild senna	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US	perennial

canadense	showy tick trefoil	Fabaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US	perennial
sativa	alfalfa	Fabaceae		perennial
	legume family	Fabaceae		annual
spp.	indigo	Fabaceae		perennial
uva crispa	european gooseberry	Grossulariaceae	Europe	perennial
ventricosa	blue plantain lily	Hostacaeae	China	perennial
paniculata	panicled hydrangea	Hydrangeaceae	Asia	perennial
spp.	hydrangea	Hydrangeaceae	Asia	perennial
macrophylla	mophead hydrangea	Hydrangeaceae	Asia	perennial
petiolaris	climbing hydrangea	Hydrangeaceae	Asia	perennial
perforatum	common st. john's wort	Hypericaceae	Europe	perennial
prolificum	shrubby st. john's wort	Hypericaceae	Midatalantic-US Southeast-US	perennial
spp.	iris	Iridaceae	Europe Asia	perennial
versicolor	northern blue flag iris	Iridaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
citriodorum	lemon basil	Lamiaceae	Africa Asia	annual
atriplicifolia	russian sage	Lamiaceae	Asia	perennial
hederacea	gill over the ground	Lamiaceae	Europe	perennial
angustifolia	english lavender	Lamiaceae	Europe	perennial
scorodonia	wood sage	Lamiaceae	Europe	perennial
reptans	carpet bugle	Lamiaceae	Europe	perennial
vulgare	oregano	Lamiaceae	Europe Asia	perennial
cataria	catnip	Lamiaceae	Europe Asia	perennial
racemosa	walkers low catmint	Lamiaceae	Europe Asia	perennial
byzantina	lamb's ear	Lamiaceae	Europe Asia	perennial
	thyme	Lamiaceae	Europe Asia	perennial
spicata	spearmint	Lamiaceae	Europe Asia	perennial
officinalis	common hedgenettle	Lamiaceae	Europe Asia	perennial
maculatum	spotted deadnettle	Lamiaceae	Europe Asia	perennial
bifida	split lipped hemp nettle	Lamiaceae	Europe Asia	annual
didyma	red beebalm	Lamiaceae	Midatlantic-US	perennial
foeniculum	anise hyssop	Lamiaceae	Midwest-US	perennial
canadensis	american wild mint	Lamiaceae	Northern-Appalachians-US	perennial

incanum	hoary mountain mint	Lamiaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
fistulosa	wild bergamot	Lamiaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial
virginianum	virginia mountain mint	Lamiaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US	perennial
virginiana	fall obedient plant	Lamiaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
vulgaris	self heal	Lamiaceae	United-States	perennial
pilosa	hairy hedgenettle	Lamiaceae	US-Excluding-Coastal-Plains	perennial
spp.	lavendula	Lamiaceae		perennial
spp.	lavender	Lamiaceae		perennial
sp.	snowy spires	Lamiaceae		perennial
	mint family	Lamiaceae		perennial
	giant hyssop	Lamiaceae		perennial
cernuum	nodding onion	Liliaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US	perennial
salicaria	purple loosestrife	Lythraceae	Europe Asia	perennial
verticillatus	swamp loosestrife	Lythraceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
syriacus	rose of sharon	Malvaceae	Asia	perennial
	marsh mallow	Malvaceae	Europe	perennial
trimestris	rose mallow	Malvaceae	Europe Asia	annual
spp.	linden tree	Malvaceae		perennial

vulgaris	yellow loosestrife	Myrsinaceae	Europe Asia	perennial
communis	myrtle	Myrtaceae	Asia	perennial
odorata	white water lily	Nymphaeaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
lindheimeri	guara	Onagraceae	Midatalantic-US Southeast-US	perennial
biennis	common evening primrose	Onagraceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	biennial
angustifolium	fireweed	Onagraceae	US-Excluding-Coastal-Plains	perennial
canadensis	wood betony	Orobanchaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
eximia	wild bleeding heart	Papaverceae	Southern-Appalachians-US	perennial
ringens	allegheny monkey flower	Phrymaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
americana	american pokeweed	Phytolaccaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
purpurea	purple foxglove	Plantaginaceae	Europe	perennial
vulgaris	toad flax	Plantaginaceae	Europe Asia	perennial
lanceolata	english plantain	Plantaginaceae	Europe Asia	perennial

spicata	spiked speedwell	Plantaginaceae	Europe Asia	perennial
hirsutus	hairy beardtongue	Plantaginaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
glabra	white turtlehead	Plantaginaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
digitalis	foxglove beardtongue	Plantaginaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
virginicum	culver's root	Plantaginaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
spp.	snapdragon	Plantaginaceae		annual
paniculata	garden phlox	Polemoniacaea	Midatlantic-US	perennial
japonica	japanese knotweed	Polygonaceae	Asia	perennial
maculosa	lady's thumb smartweed	Polygonaceae	Europe	annual
sagittata	arrow leaved tearthumb	Polygonaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
cordata	pickerelweed	Pontederiaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial

meadia	shooting star	Primulaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
rochebruneanum	meadow rue	Ranunculaceae	Asia	perennial
tomentosa	japanese anemone	Ranunculaceae	Asia	perennial
macrantha	hummingbird mint	Ranunculaceae	California	perennial
acris	tall buttercup	Ranunculaceae	Greenland	perennial
racemosa	black baneberry	Ranunculaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
pubescens	tall meadow rue	Ranunculaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
palustris	marsh marigold	Ranunculaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
	columbine	Ranunculaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
virginiana	virginia virgin's bower	Ranunculaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
spp.	monkshood	Ranunculaceae		perennial
alnifolia	alder leaved buckthorn	Rhamnaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Pacific-Northwest-US	perennial
americanus	new jersey tea	Rhamnaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Southern-Plains-US Northern-Plains-US Coastal-Plains-US	perennial
rugosa	beach rose	Rosaceae	Asia	perennial

japonica	japanese meadowsweet	Rosaceae	Asia	perennial
divaricatus	spreading cotoneaster	Rosaceae	Asia	perennial
oblonga	quince	Rosaceae	Asia	perennial
multiflora	multiflora rose	Rosaceae	Asia	perennial
recta	sulphur cinquefoil	Rosaceae	Europe	perennial
robertianum	red robin	Rosaceae	Europe	annual
rubiginosa	sweet briar rose	Rosaceae	Europe Asia	perennial
persica	peach	Rosaceae	Europe Asia	perennial
cerasus	sour cherry	Rosaceae	Europe Asia	perennial
odoratus	flowering raspberry	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US	perennial
carolina	carolina rose	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US Southern-Plains-US	perennial
alba	white meadowsweet	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
spiraea tomentosa	steeplebush	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
virginiana	virginia rose	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
allegheniensis	common blackberry	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
tomentosa	rosy meadowsweet	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
pensylvanica	pin cherry	Rosaceae	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US Northern-Plains-US	perennial
idaeus	raspberry	Rosaceae	US-Excluding-Coastal-Plains	perennial

virginiana	common strawberry	Rosaceae	US-Excluding-West-Coast	perennial
cinco de mayo	cinco de mayo rose	Rosaceae		perennial
spp.	rose	Rosaceae		perennial
occidentalis	common buttonbush	Rubiaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
repens	partridge berry	Rubiaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
petiolaris	meadow willow	Salicaceae	Northern-Appalachians-US Temperate-Plains-US	perennial
cernuus	lizard's tail	Saururaceae	Northern-Appalachians-US Southern-Appalachians-US Coastal-Plains-US Temperate-Plains-US	perennial
spp.	foamflower	Saxifragacaea	Northern-Appalachians-US Southern-Appalachians-US Temperate-Plains-US	perennial
thapsus	common mullein	Scrophulariaceae	Europe Asia Africa	biennial
dulcamara	climbing nightshade	Solanaceae	Europe Asia	perennial
philadelphica	tomatillo	Solanaceae	Mexico	annual
lycopersicon	garden tomato	Solanaceae	South-America Central-America	annual
carolinense	carolina nightshade	Solanaceae	Southeast-US	perennial
rostratum	horned nightshade	Solanaceae	Southeast-US	annual
grandifolius	bigleaf snowbell	Styracaeae	Coastal-Plains-US	perennial
spp.	styrax	Styracaeae		perennial
bonariensis	purpletop vervain	Verbenaceae	Europe	perennial
hastata	swamp verbena	Verbenaceae	United-States	biennial
tricolor	pansy	Violaceae	Europe Asia	perennial