1. Statement of Work: The Geometry of Organic Chemistry: Student Research

1.1. **Objectives and Significance.** In 1998 a group of biochemists led by R. Lahana began work on a new drug intended for suppressing the immune system. They started by considering the 64 billion possible ways of putting together 7 in a row of 35 different amino acids. Rather than educated guessing, they used a virtual library of the structures that could be formed from those 35 building blocks. This means that the computer was programmed to list the possibilities in order based on shared structural characteristics in what is known as a combinatorial library. "Combinatorial" describes objects that are constructed from basic building blocks according to specified rules. The searchers were able to screen out unlikely candidates and narrow the list to "only" 279,936 critically bioactive choices. Further computer searches winnowed the field down to 26 finalists. The winning chemical, that is, the one predicted to be most active, indeed proved in tests to be about 100 times more potent than the current lead compound.

Modern rational drug designers (and chemists in general) rely more and more upon mathematical methods for organizing the options present in the creation of new molecules. This proposal is for research within a developing area of pure mathematics called *geometric combinatorics*, which combines combinatorial questions with geometrical constructions. This research will have significant direct applications to mathematical chemistry.

For example: we plan to study the polyhexes, which consist of groups of a certain number of hexagons which share at least one side with another in the group. Three hexagons can be arranged in three ways as in Figure 1.

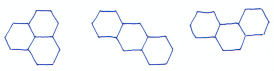


Figure 1. 3-cell polyhexes.

But these arrangements look very familiar to an organic chemist, since they are the pictures of *polycyclic benzenoid hydrocarbons*. This name refers to the way that carbon often occurs in a molecule as a hexagonal ring of six atoms. One of these rings alone is the molecule benzene, C_6H_6. There has been much recent research into the enumeration of hydrocarbons. The state of knowledge here is that it is still unknown how to calculate the number of possible hydrocarbons of a given size. Many partial results have been discovered. Enumeration of hydrocarbons is closely related to purely mathematical constructions like the polyhexes. Especially so when we restrict our attention to special polyhexes, such as the tree-like ones with a chosen "root" edge. Figure 2 shows the five of these with 2 or fewer hexagons, including the one with zero hexagons.

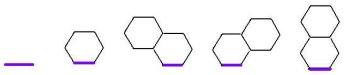


Figure 2. $\leq$ 2-cell polyhexes.

Since the method of study is geometry, the researcher will not only be counting and listing the number of ways a construction might turn out, but also geometrical ways in which the possible outcomes are related–the big picture. The mathematical principles we plan to focus on involve special shapes: 2-dimensional polygons, 3-dimensional polyhedra, and 4-dimensional polytopes. Recall that dimension can be defined as the number of coordinates needed to describe a point. In fact, polytopes can be dimension n for any whole number n. Figure 3 shows some familiar polytopes. The 4-d hypercube is least familiar, and

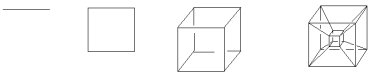


Figure 3. Segment, square, cube, hypercube.

we must note that the picture is only an approximation. Notice the pattern however; the cube is made of 2 squares (front and back) with corners connected, and the hypercube is made of 2 cubes (outer and inner) again with corresponding corners connected. The pattern continues for higher dimensions, but is more and more difficult to draw. All of these shapes, however, have corners (technically *vertices*) and sides (*facets*) just as in a cut diamond. Our overall goal is to discover how to arrange a family of combinatorial structures at the vertices in such a way that the facets take on meaning.

The applications include finding useful geometric arrangements of the list of hydrocarbons. If collections of molecules could be arranged around facets of a polytope then there might be revealed interesting insights into the properties of those molecules and their relationships. Imagine that a new molecule is a

good anticancer agent or AIDS fighter, but also has toxic side effects. Knowing the mathematical structure of a related family of molecules could allow researchers to make good choices when searching for near neighbors of their molecule which keep its lifesaving properties and avoid the toxicity.

This knowledge should also accelerate the computer processes of building and searching libraries of molecules. The number of facets of the polytopes we are planning to study grows much more slowly than the number of vertices. If the facets had meaning in terms of the chemical properties of the molecules, then the search process could be sped by screening for entire groups of molecules that share a facet. Even better, perhaps only certain facets need be represented in the library. This would help in the building stage, which can be the most time consuming.

Another advantage is that the polytope can be thought of as a solid in space, made of all its interior points rather than just the vertices. This can lead to the solution of problems by use of continuous optimization techniques. By this we mean that a property such as the conductivity of the molecule we are building might be represented by a continuous function on the polytope. Then we could find a point (not necessarily a vertex) somewhere in the solid polytope where that function value is at a maximum. Finally we could find the nearest vertex to that point and predict its associated molecular structure to realize maximum possible conductivity.

Chemistry is just one example of the potential power of this new area of mathematical research. Any sort of problem that is characterized by a large variety of possibly complex constructions based on simple building blocks can be studied in this way. Other examples might be the placement of robotic network nodes in an urban area, or the construction of the genome of a protozoa, or the positioning of communication satellites or towers, or the inputs and outcomes of an economic model.

Our primary objective is to thoroughly research and publish results about the mathematical theory involved in these applications, in preparation for a federal grant effort. Two main avenues of pure mathematics will be further developed if this proposal is funded. They are A) research into the properties of an extended family of polytopes which are combinatorially created; and B) research into ways that combinatorially generated objects can be organized by those polytopes. Then this latter information will be applied to the questions of mathematical chemistry. Some of this work has already been done by the principal investigator (S. Forcey), presented at various conferences and submitted for publication. Collaborative efforts with researchers at other institutions are also underway, and grant proposals may be joint efforts.

Our secondary objective is to engage undergraduate math majors and masters degree candidates in this area of research. It is elementary enough to allow students to quickly be able to participate in the actual research, designing and conducting experiments, looking for patterns in the collected data, and helping to formulate and prove conjectures. The principal investigator has directed three masters theses and three senior research projects, one in this precise area and the others in closely related topics. Two of those theses became part of collaboratively published papers. The future funding requested of the NSF will include tuition and stipend money for students.

1.2. **Plan and Methods A: Shapes and sizes.** When we speak of a combinatorial sequence we mean a list of whole numbers each of which counts the ways of building something given more and more building blocks. For instance the sequence 1,1,3,10,36,137,... gives the numbers of tree-like polyhexes with n hexagons. Another sequence which begins 1,1,2,5,14,42,... gives the numbers of binary trees with n leaves. For example here are the binary trees with 1 or 2 leaves:  . The trees with 3 or 4 leaves can be seen in Figure 4. This sequence of numbers of trees is known as the Catalan numbers, and there is a very large list of other interpretations of (or things you can count with) that same sequence. The fact about the Catalan numbers we want to focus on is the way in which the trees can be organized as the vertices of a polytope. For example in Figure 4 we show the trees with 4 leaves all arranged around the pentagon. The edges

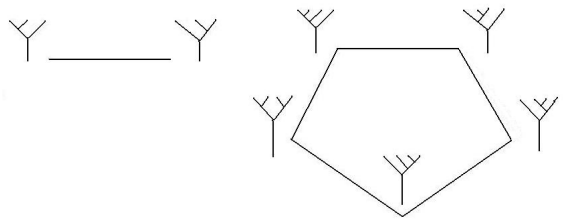


FIGURE 4.

of this pentagon can represent simple branch moves from one type of tree to another. The move is also seen between the two trees with 3 leaves. When we have five leaves, the trees arrange themselves to make the first shape in Figure 5. It is called the 5th associahedron , or $\mathcal{K}(5)$. The second shape in Figure 5

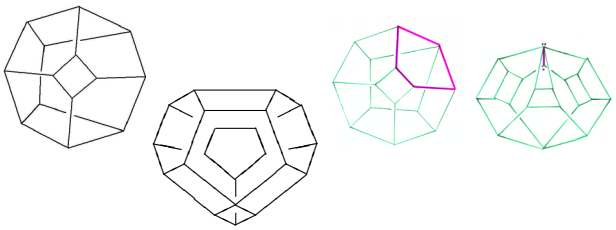
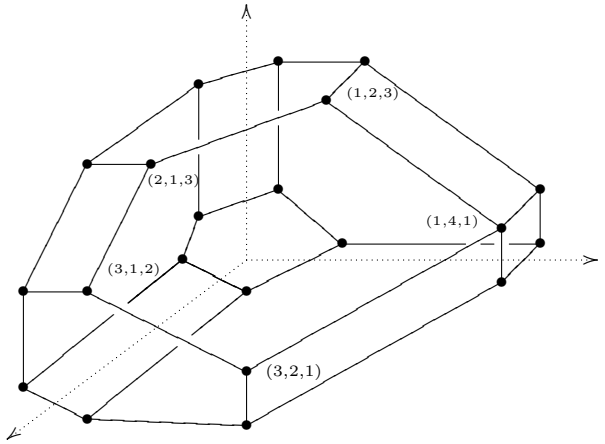


FIGURE 5. $\mathcal{K}(5)$, $\mathcal{J}(4)$, $\mathcal{CK}(4)$ and $\mathcal{N}(3)$

is called the 4th multiplihedron, or $\mathcal{J}(4)$. Its vertices correspond to trees with paint applied to their branches in various ways. The number of vertices was discovered by the principal investigator to be derived by a simple transformation of the Catalan numbers. The other two shapes in Figure 5 are new inventions of the principal investigator that correspond to further sorts of combinatorics. The third shape is called the 4th composihedron, or $\mathcal{CK}(4)$ and the last shape is called the 3rd naturahedron, or $\mathcal{N}(3)$. First we discuss further polytope research and then we will return to the combinatorial interpretations.

In 2004 it was discovered by J.L. Loday that a simple algorithm existed for finding the actual points in space that are the vertices of the associahedron. This algorithm was generalized to the multiplihedra in a recently submitted paper of the principal investigator. Here is the actual geometric version of the multiplihedron with some of the points labeled:



Just turn the page 90 degrees clockwise to see the picture of $\mathcal{J}(4)$ that is in Figure 4 of this paper.

S. Devadoss, a collaborator of the principal investigator, has also developed a generalization of the associahedra to be based upon any graph. Here a graph is just any simple collection of points and connecting lines. In collaboration, the graph multiplihedra for any graph has been conceptualized. The principal investigator has recently developed a way to find the vertices in space that give the graph multiplihedra. Figure 6 shows two of the new graph multiplihedra that the principal investigator invented.

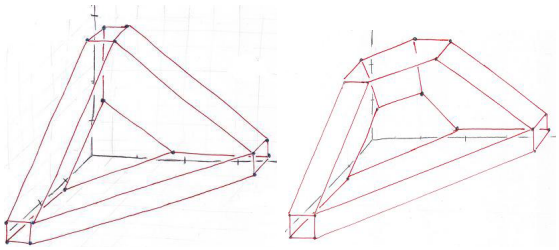


FIGURE 6. Graph multiplihedra.

The principal investigator has also developed related methods for finding the vertices in space for the composihedra and graph composihedra. These will also be described in forthcoming publications.

Among the open questions to be researched or assigned to students are: What are all the geometrical properties of the various polytopes–centers, volumes, symmetries, edge lengths and facet areas? Also of interest are the combinatorial properties–number of vertices, numbers of faces, numbers of triangulations, and space tiling properties. When the numbers of vertices of a particular sequence of polytopes are known, then there is the opportunity to find other (molecular) interpretations of those numbers which the polytopes also help to organize.

1.3. **Plan and Methods B: Sorting and counting.** It turns out that the n^{th} composihedron has the same number of vertices as the number of all the rooted tree-like polyhexes with up to n cells. One of the first questions we would like to answer is how the polyhexes might be arranged according to the vertices of $\mathcal{CK}(n)$, and what the edges and facet groupings of the diagrams and their corresponding hydrocarbons might mean. The same sort of question will also be asked for many combinatorial problems involving the other shapes in Figures 5 and 6.

In Figure 7 we show the rooted polyhexes from Figure 2 arranged around a pentagon.

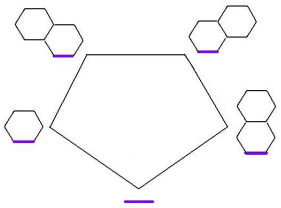


FIGURE 7. Polyhexes around a pentagon: 1^{st} guess.

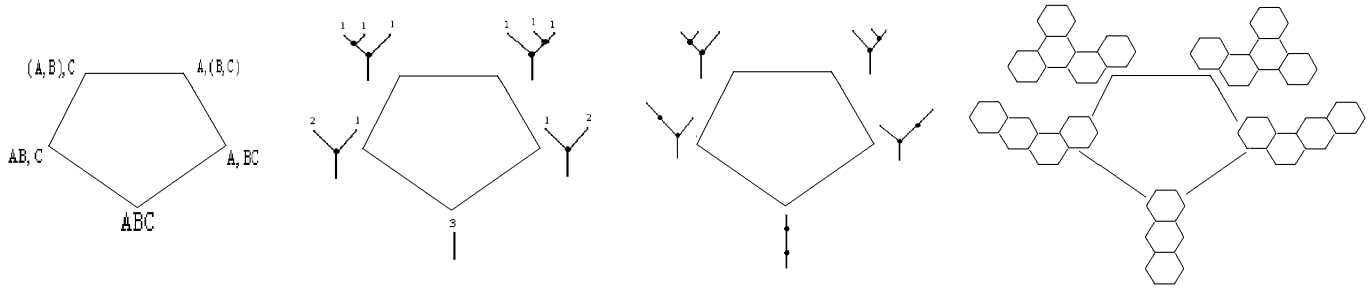


Figure 8. More pentagons.

This is only one possible arrangement. The question is how to choose the "right" arrangement so that it extends meaningfully to an arrangement around the 3-dimensional composihedron of all 15 of the tree-like polyhexes with 3 or fewer hexagons.

In fact, we want to find a recipe for putting the polyhexes with n or fewer hexagons at the vertices of the n-dimensional composihedron. The tools for attacking the problem of finding a meaningful recipe include a list of known one-to-one complete correspondences (bijections) between the polyhexes and other combinatorial objects. Four of these other types of objects are shown in Figure 8 with their correct arrangements around pentagons: strings of words made with a given alphabet, trees with a whole number assigned to each leaf, trees with extra long branches, and branching polyhexes. Others with unknown arrangements (in addition to the rooted tree-like polyhexes) include paths from point to point on a grid and the symmetric polyhexes with $2n + 1$ hexagons. By linking together the various bijections we hope to find useful new ones. Another tool is to understand the polytope edges as moves made between objects.

The number of different polyhexes grows quickly as the number of n hexagons increases. The sequence of total numbers of tree-like rooted polyhexes starts out 1,2,5,15,51,188... and then eventually grows as quickly as 5^n. The sequence of composihedra however has numbers off acets which start out 0,2,5,10,19,36... and that grow only as quickly as 2^n. This is important to the possible applications of molecular library searching, since it means that a facet-based search has the potential to proceed much more quickly.