AN APPLICATION OF ITERATED LINE GRAPHS TO BIOMOLECULAR CONFORMATION

DANIEL B. DIX

ABSTRACT. Graph theory has long been applied to molecular structure in regard to the covalent bonds between atoms. Here we extend the graph $\mathcal{G}$ whose vertices are atoms and whose edges are covalent bonds to allow a description of the conformation (or shape) of the molecule in three dimensional space. We define GZ-trees to be a certain class of tree subgraphs Γ of a graph $\mathrm{AL}^2(\mathcal{G})$, which we call the amalgamated twice iterated line graph of $\mathcal{G}$, and show that each such rooted GZ-tree (Γ, r) defines a well-behaved system of molecular internal coordinates, generalizing those known to chemists as Z-matrices. We prove that these coordinates are the most general type which give a diffeomorphism of an explicitly determined and very large open subset of molecular configuration space onto the Cartesian product of the overall position and orientation manifold and the internal coordinate space. We give examples of labelled rooted GZ-trees, describing three dimensional (3D) molecular structures, for three types of molecules important in biochemistry: amino acids, nucleotides, and glucose. Finally, some graph theoretical problems natural from the standpoint of molecular conformation are discussed.

CONTENTS

Date: October 24, 2000.

Key words and phrases. line graph, molecular conformation, conformational analysis, Z-matrix, Z-tree, internal coordinates, iterated line graph, embeddings of graphs in three dimensional space, amino acid, nucleotide, glucose, furanose ring, ring closure equations, pyranose ring.

1. Introduction

Graph theory has been applied to chemistry almost since it was invented; see [6] for the history and a survey of different applications. The abstract graph as a mathematical notion is well suited to the study of the topology of molecules, and many chemical properties can be studied on this basis [43]. However, for many purposes one needs to have good ways of describing the three dimensional shapes of molecules [7]. This is especially true in biochemistry and molecular biology, where the exact shape of a molecule essentially determines how this molecule behaves as it interacts with other (usually very large) molecules; see [32] and [29].

The simplest approach to molecular shape is to specify the Cartesian coordinates of each atom in the molecule with respect to some chosen coordinate system. A complete specification of this sort could be called a molecular *configuration*, because it would be a point in the configuration manifold of the molecule, thought of as a classical mechanical system [4], [1]. Usually, however, one is not interested in the absolute position in space nor the absolute orientation of the molecule. Orbits of molecular configurations under the action of the Lie group G_a of all proper Euclidean motions (i.e. translations and proper rotations) are called molecular *conformations*. A coordinate system on the space of all molecular conformations is called a system of *internal coordinates*. Mathematical study of the dynamics of molecular conformations was initiated in [20], and has been subsumed in the general reduction theory for classical mechanical systems possessing a Lie group of symmetries; see [26] and references therein. There are many systems of internal coordinates which could be applied to molecules [34], but we are only interested in those which reflect the network of covalent bonds connecting the atoms. This network is described by the *molecular graph* $\mathcal{G}$ of the molecule, where vertices are atoms and edges are covalent bonds between them. In this paper we study a general class of internal coordinate systems which are intimately related to the graph $\mathcal{G}$. These systems are well known to chemists, since they describe the conformation of the molecule in terms of *bond lengths*, *bond angles*, and *torsion angles*. Such systems provide the simplest way to understand and adjust the internal geometry of a molecule. They are natural because chemical forces very effectively constrain the values of bond lengths and bond angles, and the aspects of real molecular flexibility are for the most part expressed in terms of the torsion angles.

Chemists typically specify a molecular conformation using a data structure called a *Z-matrix* [22], [34], [17] and [11]. (Unfortunately the term Z-matrix has many meanings throughout the scientific literature. For example, our Z-matrix seems to be totally unrelated to the Hosoya Z matrix of [37].) See section 4.1 for information about the historical origin of the Z-matrix. Behind the data structure is an internal coordinate system *of Z-matrix type*. We show how these types of internal coordinate systems can be represented by certain rooted tree subgraphs of a new type of graph related to the twice iterated line graph of $\mathcal{G}$. In fact this abstract viewpoint leads us to introduce a new class of internal coordinate systems more general than those of Z-matrix type. These systems all have the same good behavior as Z-matrix systems; in fact we characterize precisely *all* the systems which have this sort of good behavior. We say an internal coordinate system is *well-behaved* if we can describe the corresponding coordinate chart (in the differential atlas of the conformation manifold) explicitly. Thus the special conformations where the coordinate system breaks down are all known in advance—there are no surprises to worry about.

However we do not intend to convey the impression that well-behaved internal coordinate systems are the best ones for every purpose. Nothing that we discuss will overcome one of the weaknesses of Z-matrix type internal coordinates, namely that ring structures involve constraints on the coordinates which are nontrivial to deal with (see especially section 5.2). While this weakness has serious consequences for computational geometry optimization (i.e. finding the molecular conformation with the lowest chemical potential energy) [38], [5], [39], it does not cause problems for us since our purpose is different. Our focus will be on the description and study of molecular geometries. We will not deal with the dynamics of molecules as described in internal coordinates (but see [33]); nor will we consider geometry optimization. We are interested in molecular flexibility, but from a geometric rather than an energetic perspective. Ring structures give rise to interesting mathematical issues from our point of view. We will not however be able to address all these issues here.

We have been motivated to pursue this study by the desire to bring the geometrical structure of the molecules of life under scrutiny by mathematicians. It is difficult for mathematicians to muster the perseverance needed to study mathematical objects which are unwieldy and *ad hoc* in appearance. (Utility, for the chemist, compensates for any deficiency in purely mathematical beauty.) The nicest mathematical treatment of molecular geometry of which we are aware is [12]. Their approach, which is based almost entirely on the distances between atoms, is motivated by measurement techniques which yield these distances directly. Our approach is instead to introduce a *theory* of internal coordinates for biomolecules, which is very natural and strongly connected to graph theory. What emerges is a mathematical object which we call a *3D molecule*, which can be manipulated mathematically much like a child builds with Legos. In this work we define and give examples of these objects, but in future work we intend to show how these manipulations can be performed. The examples we chose are not necessarily the simplest, but biochemically the most interesting. This is because of our belief that out of the study of the molecules of life the most interesting new mathematics will arise.

The key idea which connects the molecular graph $\mathcal{G}$ to its embedding in three dimensional space is that of the *line graph*. Some chemists have suspected that some such connection must exist (see [16]). But the connection between what we call the *amalgamated twice iterated line graph* $AL^2(\mathcal{G})$ of $\mathcal{G}$ and internal coordinate systems is so natural that it is surprising that it has not yet been noticed. $AL^2(\mathcal{G})$ is the correct context within which to study internal coordinates because almost every important biomolecular configuration extends in a geometrically natural way to an assignment of a Cartesian coordinate system to each of the vertices of $AL^2(\mathcal{G})$ (see section 3.1). Edges in $AL^2(\mathcal{G})$ can then be labelled by coordinate transformations, from which the internal coordinates of the molecule are naturally derived. In previous treatments of this subject a coordinate system was assigned to each atom (see e.g. [18]), but this involved making an arbitrary choice among several equally valid alternatives, and no general theory has emerged from that approach (except perhaps for the concept of *discrete Frenet frames* for chain molecules [40]). Those treatments are not difficult to reinterpret in our theoretical framework. Other internal coordinate systems not derived from this theoretical foundation can be imagined, and no doubt have been used in particular cases [45], but systems derived from subgraphs of $AL^2(\mathcal{G})$ are easier to analyze rigorously in general. Some

organic molecules have portions which are collinear, and this causes trouble with our scheme. But this local collinearity is hardly ever present in the molecules of life. Such problems are sufficiently rare in biomolecules that we can work around them.

We need to acknowledge several sources of inspiration and to mention some related work. The comprehensive discussions in [34] and [12] are very helpful. The papers [18], [14], [41], [21], [31], [40], [15] give a sample of some seminal and recent activity. Other references will be mentioned in later sections.

The plan of this paper is as follows. First we will discuss some preliminary topics which allow our treatment to be self-contained. Here we fix certain notation and make basic definitions. In particular, we define the graph $AL^2(\mathcal{G})$ in section 2.2. In section 2.3 we show how graph theoretical hypotheses can translate into coordinatization theorems in a simplified situation. This section is the foundation of all our subsequent work. In section 3.1 the basic idea relating the graph $AL^2(\mathcal{G})$ and molecular conformation comes into the light. In section 3.2 the internal coordinates are given precise definition. Z-matrix internal coordinate systems are associated to certain subgraphs of $AL^2(\mathcal{G})$ called *Z-trees*. These are defined in section 4.1, along with their canonical generalization, the GZ-trees. In this same section our main theorem is stated. This theorem gives necessary and sufficient conditions on a tree subgraph of $AL^2(\mathcal{G})$ that it define a well-behaved internal coordinate system. These conditions are of a purely graph theoretical character. The chemists and molecular physicists (e.g. experts in molecular vibration) have long been proceeding under the assumption that a theorem like this is true (see [10]), but a careful statement and proof apparently has not been given in the literature. The ideas behind the proof also seem to be new; see sections 4.2 and 4.3. In section 5 we give three examples of 3D molecules, which are Z-trees whose edges have been labelled with the appropriate internal coordinates. Our examples are the monomers of biopolymers, namely amino acids (proteins), nucleotides (DNA and RNA), and glucose representing the monosaccharides (polysaccharides). Our intention, besides giving intrinsically interesting examples, is to set the stage for later work where the mathematical scheme of polymerization (the linking together of similar molecules into long chains, which then typically fold into complicated three dimensional shapes) will be discussed. While studying these examples many fundamental mathematical problems are encountered, especially in view of the new formalism presented in this paper. Finally in the last section (section 6) we formulate a few graph theoretical problems which are suggested by this work. This section is for the graph theorist who is interested in questions suggested by applications. We hope that this new connection between graph theory and molecular conformation will lead to progress in both fields.

Acknowledgements. The author offers his thanks to P. G. Wolynes for extending him hospitality and for many useful conversations during his sabbatical at the University of Illinois. Thanks also to Zan Schulten, Todd Martinez, Tom Hughes, and Vassiliy Lubchenko for many stimulating conversations. Thanks also to László Székely and David Sumner for references on graph theory, Cathy Murphy for references on nucleic acids, and Ralph Howard for help with LaTeX and figures.

2. Preliminaries

2.1. **Space, Poses, Affine Symmetries.** In order to explain our coordinate systems in the clearest possible manner we will distinguish between space and its coordinate representations. Suppose W is a real 4-dimensional vector space, and $V \subset W$ is a 3-dimensional subspace. We assume V is equipped with an inner product (i.e. the usual dot product) and an orientation, i.e. a choice of a distinguished equivalence class of bases $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ of V, where two such bases are equivalent if the matrix transforming one into the other has positive determinant. Bases which are elements of the distinguished equivalence classs are said to be *positively oriented*. The dot product of two vectors $\mathbf{U}, \mathbf{V} \in V$ will be denoted by $\mathbf{U} \cdot \mathbf{V}$; the associated norm is defined by $\|\mathbf{U}\| = (\mathbf{U} \cdot \mathbf{U})^{1/2}$. The orientation in V allows us to define the vector (or cross) product $\mathbf{U} \times \mathbf{V}$ in the usual way in terms of any positively oriented orthonormal basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ of V. Let X denote a fixed distinguished nonzero coset in the one dimensional quotient vector space W/V; this set will be our model for the 3-dimensional space within which molecules live. There are several reasons why this model of space is better than $\mathbb{R}^3$ or even V. First of all, like real space and unlike V, X does not have a distinguished origin. Secondly, X and V, in contrast to $\mathbb{R}^3$, have no preferred directions. However, X and V are closely related in that any two points $\mathbf{x}, \mathbf{y}$ of X determine a unique $\mathbf{v} \in V$ such that $\mathbf{y} - \mathbf{x} = \mathbf{v}$. The difference $\mathbf{y} - \mathbf{x}$ is well-defined in the vector space W. (We could define everything without reference to the space W but not much would be thereby gained, and the exposition would be longer.)

If $\mathbf{e}_0 \in X$ is given then a bijection $V \to X \colon \mathbf{v} \mapsto \mathbf{e}_0 + \mathbf{v}$ is determined. Any positively oriented orthonormal basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ of V determines a unique orientation preserving isometric (dot product preserving) linear isomorphism $\mathbb{R}^3 \to V$. Hence the basis $E = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ of W determines (by composition of these two bijections), and is determined by, a choice of Cartesian coordinate system $\mathbb{R}^3 \to X$. $\mathbf{e}_0$ is the origin of this system. We will call (following [31]) the basis E a *pose*, a word from the robotics literature meaning "position and orientation". As we have seen, poses can be identified with Cartesian coordinate systems. Let $\mathcal{S}$ denote the set of all Cartesian coordinate systems on X, or equivalently all poses for X.

Affine changes of Cartesian coordinate systems (or poses) can then be represented as 4×4 real matrices:

$$(\mathbf{e}_0', \mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3') = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)T, \qquad T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & a_{11} & a_{12} & a_{13} \\ b & a_{21} & a_{22} & a_{23} \\ c & a_{31} & a_{32} & a_{33} \end{pmatrix},$$

which means the same thing as

$$\begin{aligned} \mathbf{e}_0' &= \mathbf{e}_0 + \mathbf{e}_1 a + \mathbf{e}_2 b + \mathbf{e}_3 c, \\ \mathbf{e}_1' &= \mathbf{e}_1 a_{11} + \mathbf{e}_2 a_{21} + \mathbf{e}_3 a_{31}, \\ \mathbf{e}_2' &= \mathbf{e}_1 a_{12} + \mathbf{e}_2 a_{22} + \mathbf{e}_3 a_{32}, \\ \mathbf{e}_3' &= \mathbf{e}_1 a_{13} + \mathbf{e}_2 a_{23} + \mathbf{e}_3 a_{33}. \end{aligned}$$

The 3×3 real matrix $A = (a_{ij})$ is a rotation matrix, and the triple $(a, b, c)^T$ determines a shift of the origin. Let G_p denote the Lie group of all such 4×4 matrices T. Let $\mathbf{1}$ denote the 4×4 identity matrix, and let its four column vectors

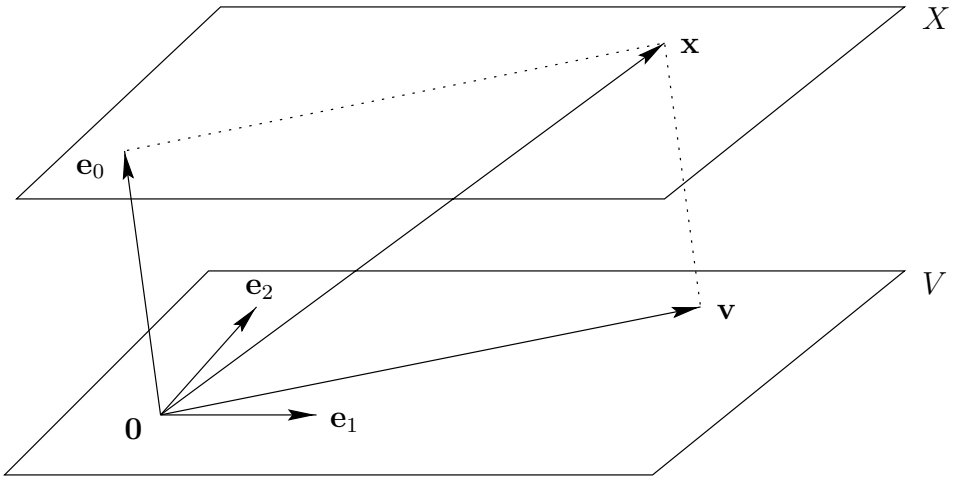

FIGURE 1. $X \subset W$ is a coset in W/V. Since both $\mathbf{e}_0 \in X$ and $\mathbf{x} \in X$ we have $X = \mathbf{e}_0 + V = \mathbf{x} + V$.

be $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$. Any point $\mathbf{x} \in X$ can be written as $\mathbf{x} = \mathbf{e}_0 + \mathbf{e}_1 x_1 + \mathbf{e}_2 x_2 + \mathbf{e}_3 x_3$ for some real numbers (x_1, x_2, x_3) with respect to the pose E. The coordinates (x'_1, x'_2, x'_3) of the same point $\mathbf{x}$ with respect to the new pose $E' = (\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ can be related to the previous coordinates by the relation

$$\begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & a_{11} & a_{12} & a_{13} \\ b & a_{21} & a_{22} & a_{23} \\ c & a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 \\ x'_1 \\ x'_2 \\ x'_3 \end{pmatrix},$$

which follows from the fact that $\mathbf{x} = (\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)(1, x'_1, x'_2, x'_3)^T$ and then using the formula $E' = ET$. The subscript p in G_p stands for "passive", since the point in space is not moving; its coordinate representation is however changing. Thus the same matrix T describes two different aspects of the transformation: the change of pose, and the change of coordinates. The group G_p acts on the set $\mathcal{S}$ on the right, and multiplies column vectors $(1, x_1, x_2, x_3)^T$ on the left. If we identify a 4×4 matrix T (of the above form) with its associated linear map $\mathbb{R}^4 \to \mathbb{R}^4$ (which on 4-tuples of the above form yields a bijection $\mathbb{R}^3 \to \mathbb{R}^3$), and if the elements of $\mathcal{S}$ are interpreted as bijections $\mathbb{R}^3 \to X$, then the right action of G_p on $\mathcal{S}$ is by composition of mappings.

Now define G_a to be the set of all linear maps $M \colon W \to W$, such that $M(X) \subset X$, $M(V) \subset V$, and such that $M|_V$ is an orientation preserving isometry. The subscript a stands for "active", since now points in X are being moved. $M \in G_a$ is called a *proper affine symmetry* of X because it is uniquely determined by its action on X. To see this suppose $M_1, M_2 \in G_a$ agree on X. Then for all $\mathbf{v} \in V$ we have $M_1(\mathbf{v}) = M_1(\mathbf{y} - \mathbf{x}) = M_1(\mathbf{y}) - M_1(\mathbf{x}) = M_2(\mathbf{y}) - M_2(\mathbf{x}) = M_2(\mathbf{y} - \mathbf{x}) = M_2(\mathbf{v})$, where $\mathbf{x}, \mathbf{y} \in X$ satisfy $\mathbf{y} - \mathbf{x} = \mathbf{v}$. Thus M_1 and M_2 agree also on V. Since an element of X and a basis of V together form a basis of W, we see that M_1 and M_2 agree on W. Thus elements of G_a can be thought of as mappings $X \to X$. It is clear that the composition of any two mappings in G_a is again in G_a. Each mapping $M \in G_a$ is a linear isomorphism $W \to W$ (since a basis is mapped into

a basis), and its inverse is again in G_a. Hence G_a is a group. It is in fact a Lie group isomorphic to G_p, but the isomorphism $G_a \to G_p$ depends on a choice of a pose $E \in \mathcal{S}$. To see this recall that such a pose induces a bijection $E\colon \mathbb{R}^3 \to X$, which is the Cartesian coordinate system associated to E. If $M \in G_a$, we can think of it as a bijective map from X to X. We can form the composition $M \circ E$ of these two maps to obtain another bijection $E'\colon \mathbb{R}^3 \to X$, which corresponds to another pose $E' \in \mathcal{S}$. The pose E' can be calculated another way, namely we can apply the map M to each of the vectors of E: i.e. if $E = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ then $E' = (M(\mathbf{e}_0), M(\mathbf{e}_1), M(\mathbf{e}_2), M(\mathbf{e}_3))$.

$$\begin{array}{ccc} X & \xrightarrow{M} & X \\ {\scriptstyle E}\uparrow & & \uparrow{\scriptstyle E} \\ \mathbb{R}^3 & \xrightarrow{T} & \mathbb{R}^3 \end{array}$$

There is a unique 4×4 matrix $T \in G_p$ such that $E' = ET$. The mapping $M \mapsto T$ is the desired group isomorphism $G_a \to G_p$. All of these group isomorphisms induce the same manifold structure on G_a, making it a well defined Lie group. Clearly G_a acts on W on the left, and hence on X, V, and $\mathcal{S}$ on the left. Thus the set $\mathcal{S}$ is equipped with a right action of G_p and a left action of G_a. Both of these actions are transitive and fixed point free, and $(ME)T = M(ET)$. Thus as a manifold, $\mathcal{S}$ is diffeomorphic to the Lie group G_p, although not via a canonical diffeomorphism.

2.2. **Amalgamated Iterated Line Graphs.** In this section we give the definition of an apparently new type of graph associated to an arbitrary graph $\mathcal{G}$. We review the basic definitions for the nonexpert. A nice reference for graph theory is [44]. The chemical interpretation of this new graph will be explained in section 3.

A (simple and finite) *graph* $\mathcal{G}$ consists of a finite set $\mathrm{vert}\mathcal{G}$, whose elements are called *vertices*, together with a set $\mathrm{edge}\mathcal{G}$, whose elements are two element subsets of $\mathrm{vert}\mathcal{G}$ which are called *edges*. Suppose $A_1, A_2, A_3 \in \mathrm{vert}\mathcal{G}$ are distinct vertices and $b = \{A_1, A_2\}, b' = \{A_1, A_3\}$ are distinct edges of $\mathcal{G}$. Since $b \cap b' = \{A_1\}$, we say the edges b and b' share the vertex A_1. The *line graph* $L(\mathcal{G}) = L^1(\mathcal{G})$ associated to the graph $\mathcal{G}$ is defined as follows. $\mathrm{vert}L^1(\mathcal{G})$ is the set of all edges of $\mathcal{G}$. $\mathrm{edge}L^1(\mathcal{G})$ contains $a = \{b, b'\}$ whenever b and b' are distinct edges of $\mathcal{G}$ which share a vertex of $\mathcal{G}$, i.e. $b \cap b' \neq \emptyset$. Define the mapping $\alpha_1\colon \mathrm{edge}L^1(\mathcal{G}) \to \mathrm{vert}\mathcal{G}$ as follows: if $a = \{b, b'\} \in \mathrm{edge}L^1(\mathcal{G})$ then $\alpha_1(a)$ is defined to be the unique element of $\mathrm{vert}\mathcal{G}$ such that $b \cap b' = \{\alpha_1(a)\}$.

Since $L^1(\mathcal{G})$ is itself a graph, we can discuss its line graph $L^2(\mathcal{G}) = L(L^1(\mathcal{G}))$, called the *twice iterated line graph* of $\mathcal{G}$. Thus $\mathrm{vert}L^2(\mathcal{G}) = \mathrm{edge}L^1(\mathcal{G})$ and elements of $\mathrm{edge}L^2(\mathcal{G})$ are pairs $w = \{a_1, a_2\}$, where a_1, a_2 are distinct edges of $L^1(\mathcal{G})$ such that $a_1 \cap a_2 \neq \emptyset$. As before we define the mapping $\alpha_2\colon \mathrm{edge}L^2(\mathcal{G}) \to \mathrm{vert}L^1(\mathcal{G})$ such that $\{\alpha_2(\{a_1, a_2\})\} = a_1 \cap a_2$.

These abstract constructions can be a bit confusing, so it is helpful to interpret them in the case where the graph $\mathcal{G}$ is associated to a molecule. Thus $\mathrm{vert}\mathcal{G}$ is the set of *atoms* (or atomic cores or nuclei) in the molecule; we assume different atoms of the same type, e.g. Hydrogen, are given distinct names, e.g. H_1 and H_2, so that they are distinguishable. If A_1, A_2 are distinct atoms of the molecule, then $b = \{A_1, A_2\} \in \mathrm{edge}\mathcal{G}$ if in the given molecule atom A_1 is covalently bonded to atom A_2. Thus edges of $\mathcal{G}$ are called *bonds*. An edge $a = \{b, b'\}$ in $L^1(\mathcal{G})$ (also a vertex in $L^2(\mathcal{G})$) consists of a pair of bonds sharing a common atom. It

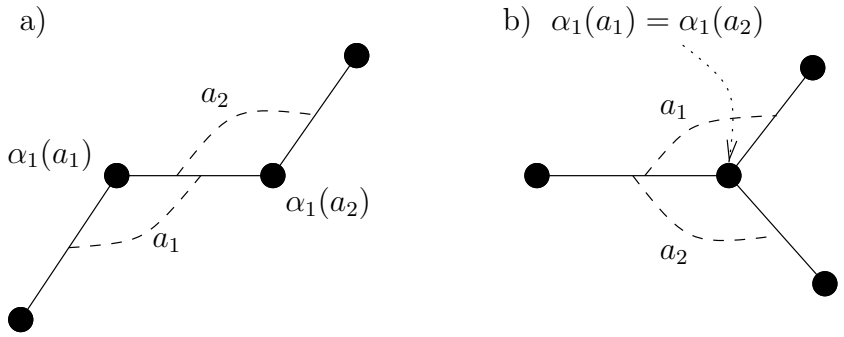

FIGURE 2. (a) $w = \{a_1, a_2\}$ is a dihedral. (b) $w = \{a_1, a_2\}$ is an improper. In both cases the common bond shared by a_1 and a_2 is $\alpha_2(w)$.

is natural to call a an *angle*, since when the molecule is embedded in space the two bonds form an angle whose vertex is the atom $\alpha_1(a)$. An edge $w = \{a_1, a_2\}$ in $L^2(\mathcal{G})$ is called a *wedge*, since generically when the molecule is embedded in space the plane of the angle a_1 intersects the plane of the angle a_2 along a line containing the common bond $\alpha_2(w)$. Wedges classify into two disjoint categories. If $w = \{a_1, a_2\}$ and $\alpha_2(w) = \{\alpha_1(a_1), \alpha_1(a_2)\}$, then the wedge w is called a *dihedral*. If $\alpha_1(a_1) = \alpha_1(a_2)$ then the wedge w is called an *improper*. Clearly any wedge must be either a dihedral or an improper, but not both. Thus the graphs $L^1(\mathcal{G})$ and $L^2(\mathcal{G})$ have natural chemical interpretations.

Now we will give the abstract definition of another graph built out of $\mathcal{G}$, $L^1(\mathcal{G})$, and $L^2(\mathcal{G})$, which will be extremely useful when discussing molecular conformation (see section 3). We call it the *amalgamated twice iterated line graph* of $\mathcal{G}$, and symbolize it as $AL^2(\mathcal{G})$. $\mathrm{vert}AL^2(\mathcal{G})$ consists of all triples (a, b, A), where $a \in \mathrm{vert}L^2(\mathcal{G}) = \mathrm{edge}L^1(\mathcal{G})$, $b \in \mathrm{vert}L^1(\mathcal{G}) = \mathrm{edge}\mathcal{G}$, $A \in \mathrm{vert}\mathcal{G}$, and $A \in b \in a$. An unordered pair of distinct vertices in $AL^2(\mathcal{G})$ will comprise an edge of $AL^2(\mathcal{G})$ if the two triples differ in only a single component, and if the two element set consisting of the components of the triples which do not agree is an edge in $L^2(\mathcal{G})$, $L^1(\mathcal{G})$, or $\mathcal{G}$ as appropriate. Thus there are three disjoint categories of edges in $AL^2(\mathcal{G})$. The unordered pair $\{(a_1, b_1, A_1), (a_2, b_2, A_2)\} \subset \mathrm{vert}AL^2(\mathcal{G})$ is in $\mathrm{edge}AL^2(\mathcal{G})$ if one of the following conditions holds.

0) $a_1 = a_2, b_1 = b_2, A_1 \neq A_2$, and $\{A_1, A_2\} \in \mathrm{edge}\mathcal{G}$.
1) $a_1 = a_2, b_1 \neq b_2, A_1 = A_2$, and $\{b_1, b_2\} \in \mathrm{edge}L^1\mathcal{G}$.
2) $a_1 \neq a_2, b_1 = b_2, A_1 = A_2$, and $\{a_1, a_2\} \in \mathrm{edge}L^2\mathcal{G}$.

We denote the subsets of $\mathrm{edge}AL^2(\mathcal{G})$ consisting of those unordered pairs satisfying conditions (0), (1), or (2) by $\mathrm{edge}_0AL^2(\mathcal{G})$, $\mathrm{edge}_1AL^2(\mathcal{G})$, or $\mathrm{edge}_2AL^2(\mathcal{G})$, respectively. Such edges are said to be *of type* 0, *type* 1, or *type* 2 respectively. The same notation will apply to any subgraph of $AL^2(\mathcal{G})$.

This graph is closely related to certain general graph theoretical constructions. Recall that a graph $\mathcal{H}$ is a *subgraph* of $\mathcal{G}$ if $\mathrm{vert}\mathcal{H} \subset \mathrm{vert}\mathcal{G}$ and $\mathrm{edge}\mathcal{H} \subset \mathrm{edge}\mathcal{G}$. A subset S of the set of vertices of $\mathcal{G}$ can be used to construct a subgraph $\mathcal{H} = \mathcal{G}[S]$, called the *induced subgraph* of $\mathcal{G}$ determined by S: we set $\mathrm{vert}\mathcal{H} = S$ and $\mathrm{edge}\mathcal{H} = \{e \in \mathrm{edge}\mathcal{G} \mid e \subset S\}$. There is a general construction of the Cartesian product of two or more graphs (see page 175 of [44]). If $\mathcal{G}_1$ and $\mathcal{G}_2$ are graphs then their

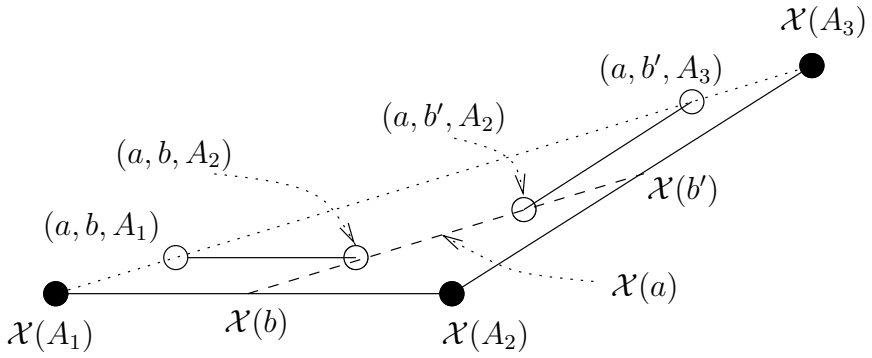

FIGURE 3. An embedding of some vertices and edges of $AL^2(\mathcal{G})$ associated with the angle $a = \{b, b'\}$, where $b = \{A_1, A_2\}$ and $b' = \{A_2, A_3\}$. Vertices of $AL^2(\mathcal{G})$ are shown as open circles. Edges of type 0 of $AL^2(\mathcal{G})$ are represented by solid lines connecting two vertices. An edge $\{(a, b, A_2), (a, b', A_2)\}$ of type 1 is represented by a dashed line connecting those two vertices. Edges of type 2 are not shown, but would connect two vertices in different planes determined by different angles a_1 and a_2 (see Figure 4).

Cartesian product $\mathcal{G}_1 \Box \mathcal{G}_2$ is defined as follows: $\mathrm{vert}(\mathcal{G}_1 \Box \mathcal{G}_2) = \mathrm{vert}\mathcal{G}_1 \times \mathrm{vert}\mathcal{G}_2$, and $\{(v_1, v_2), (v_1', v_2')\} \subset \mathrm{vert}(\mathcal{G}_1 \Box \mathcal{G}_2)$ is in $\mathrm{edge}(\mathcal{G}_1 \Box \mathcal{G}_2)$ if either $v_1 = v_1'$, $v_2 \neq v_2'$, and $\{v_2, v_2'\} \in \mathrm{edge}\mathcal{G}_2$ or $v_1 \neq v_1'$, $v_2 = v_2'$, and $\{v_1, v_1'\} \in \mathrm{edge}\mathcal{G}_1$. If we define $S \subset \mathrm{vert}(\mathcal{G} \Box L^1(\mathcal{G}) \Box L^2(\mathcal{G}))$ to consist of those triples (a, b, A) where $A \in b \in a$, then it is clear that $AL^2(\mathcal{G})$ is the subgraph of $\mathcal{G} \Box L^1(\mathcal{G}) \Box L^2(\mathcal{G})$ induced by S. The subset S defines the ternary relation of *incidence*.

Given an embedding of the graph $\mathcal{G}$ of a molecule it is possible and conceptually very useful to be able to give a corresponding embedding of the graph $AL^2(\mathcal{G})$. If A is an atom then let $\mathcal{X}(A) \in X$ denote its position in three dimensional space. Suppose (a, b, A) is a vertex in $AL^2(\mathcal{G})$. Suppose $a = \{b, b'\}$, $b = \{A_1, A_2\}$, $b' = \{A_2, A_3\}$. Let us assume the generic situation that the three atoms involved occupy distinct points of space and are not collinear. Thus the bonds b and b' represent noncollinear line segments sharing the point $\mathcal{X}(A_2)$. We will associate (a, b, A) to a certain point on the space triangle spanned by the points $\mathcal{X}(A_1)$, $\mathcal{X}(A_2)$ and $\mathcal{X}(A_3)$. The exact method we use to do this is not so important; but the method we suggest is easy to draw in diagrams. When thought of as an edge in $\mathcal{G}$ we draw b as a line segment connecting $\mathcal{X}(A_1)$ and $\mathcal{X}(A_2)$, but when thinking of it as a vertex in $L^1(\mathcal{G})$ we represent it by the midpoint $\mathcal{X}(b) = [\mathcal{X}(A_1) + \mathcal{X}(A_2)]/2$ of this line segment. Thus the edge a in $L^1(\mathcal{G})$ is represented by a line segment joining the midpoint $\mathcal{X}(b)$ of bond b to the midpoint $\mathcal{X}(b')$ of bond b'. However, when viewing a as a vertex in $L^2(\mathcal{G})$ we should think of it as the midpoint $\mathcal{X}(a) = [\mathcal{X}(b) + \mathcal{X}(b')]/2$ of this line segment. There are two possibilities: either $A = A_1$ or $A = A_2$. If $A = A_2$ we associate the triple (a, b, A) with the midpoint $[\mathcal{X}(a) + \mathcal{X}(b)]/2$. If $A = A_1$ then we associate the triple (a, b, A) with the point of intersection of the segment connecting $\mathcal{X}(A_1)$ and $\mathcal{X}(A_3)$ and the line through $[\mathcal{X}(a) + \mathcal{X}(b)]/2$ parallel to the segment connecting $\mathcal{X}(A_1)$ and $\mathcal{X}(A_2)$. (See Figure 3.) Edges in $AL^2(\mathcal{G})$ are represented by line segments connecting points in space representing

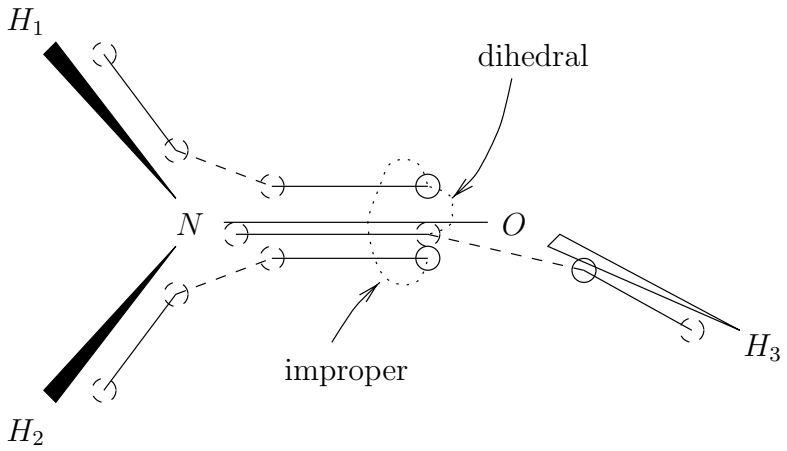

FIGURE 4. A depiction of edges of type 2 in $AL^2(\mathcal{G})$ when $\mathcal{G}$ is the graph of a molecule of hydroxylamine, i.e. NH_2OH. Edges of type 2 are drawn as dotted curves, connecting appropriate vertices. The N—H_1 and N—H_2 bonds are shaded as is common in chemical diagrams to indicate that the Hydrogen atoms are closer to the viewer than the Nitrogen atom. Similarly, the H_3 atom is further from the viewer than the Oxygen atom.

vertices of $AL^2(\mathcal{G})$. (See Figure 4 for edges of type 2.) If the embedding of $\mathcal{G}$ is bad enough then this construction might not result in an embedding of $AL^2(\mathcal{G})$, but it appears to work fine for the subgraphs of $AL^2(\mathcal{G})$ that we will consider when we restrict attention to the realistic embeddings of the biomolecular graphs $\mathcal{G}$. We stress the point that these embeddings are merely for the purpose of easily visualizing subgraphs of $AL^2(\mathcal{G})$, and are not directly connected to the detailed specification of a particular embedding of the molecular graph $\mathcal{G}$. This latter task can be accomplished independently of how we visualize $AL^2(\mathcal{G})$.

2.3. **Coordinatizing an Orbit Space.** In this section we exhibit a natural left G_a-equivariant diffeomorphism between $\mathcal{S}^{\text{vert}\Gamma}$ and $\mathcal{S} \times G_p^{\text{edge}\Gamma}$, where Γ is any rooted tree graph (defined below), and where we are using the notation introduced in section 2.1. (If Y, Z are sets then Z^Y denotes the set of all mappings $f\colon Y \to Z$.) Since G_a acts transitively and without any fixed point on the left of $\mathcal{S}$ and trivially (every point is a fixed point) on $G_p^{\text{edge}\Gamma}$, we obtain a diffeomorphism of the orbit space $G_a\backslash\mathcal{S}^{\text{vert}\Gamma}$ with the space $G_p^{\text{edge}\Gamma}$. We use a restriction of this diffeomorphism in our main theorem. But this section introduces some of the main ideas in a simpler context, and hence has a mainly pedagogical purpose. The reader might understand the manifold $\mathcal{S}^{\text{vert}\Gamma}$ to represent the configuration of a molecule, where the vertices of Γ are associated with individual atoms; however instead of merely specifying the position of each atom, we are specifying a family of poses associated with that atom. The group G_a of proper affine symmetries of space acts on this set of configurations, and we are mostly interested in the set of orbits $G_a\backslash\mathcal{S}^{\text{vert}\Gamma}$ under this left action. An orbit $G_a\mathcal{E}$, where $\mathcal{E} \in \mathcal{S}^{\text{vert}\Gamma}$ could be thought of as a molecular conformation. See section 3.1 for precise definitions of configuration and conformation. The factor $\mathcal{S}$ represents the overall position and orientation of

the molecule, and the manifold $G_p^{\text{edge}\Gamma}$ (roughly) represents the space of internal coordinates.

Suppose Γ is a graph (see section 2.2 for definitions) and $v, v' \in \text{vert}\Gamma$ are distinct. A *path* in Γ connecting v to v' is an ordered list $(v_0, v_1, \ldots, v_n)$ of distinct vertices of Γ such that $v_0 = v$, $v_n = v'$, and $\{v_{i-1}, v_i\} \in \text{edge}\Gamma$ for all $i = 1, 2, \ldots, n$. This path is of *length* $n \geq 1$. The graph Γ is said to be *connected* if for any two distinct vertices v, v' of Γ there exists a path in Γ connecting v to v'. An ordered list $(v_0, v_1, \ldots, v_n)$ of vertices ($n \geq 3$) of Γ such that $v_0 = v_n$, $\{v_{i-1}, v_i\} \in \text{edge}\Gamma$ for all $i = 1, 2, \ldots, n$, and $v_i = v_j$ for some $0 \leq i < j \leq n$ implies that $(i, j) = (0, n)$ is called a *cycle*. A connected graph Γ is said to be a *tree* if it possesses no cycle. A graph Γ is said to be *rooted* if a distinguished vertex r of Γ has been chosen; r is called the root vertex, or simply the root, and the pair (Γ, r) is called a rooted graph.

A (simple finite) *digraph* Δ is a finite set $\text{vert}\Delta$ together with a set $\text{oedge}\Delta \subset \text{vert}\Delta \times \text{vert}\Delta$. Each $(v_p, v_c) \in \text{oedge}\Gamma$ is an *oriented edge*. We will call v_p the *parent* and v_c the *child* of the oriented edge (v_p, v_c). An *oriented graph* is a digraph Δ such that whenever $(v_p, v_c) \in \text{oedge}\Delta$ we have $v_p \neq v_c$ and $(v_c, v_p) \notin \text{oedge}\Delta$. If Δ is an oriented graph then define $\text{vert}\Gamma = \text{vert}\Delta$ and $\text{edge}\Gamma = \{\{v_p, v_c\} \mid (v_p, v_c) \in \text{oedge}\Delta\}$. Taken together, $\text{vert}\Gamma$ and $\text{edge}\Gamma$ define the *underlying graph* Γ associated to the oriented graph Δ. The mapping $\text{oedge}\Delta \to \text{edge}\Gamma \colon (v_p, v_c) \mapsto \{v_p, v_c\}$ is a bijection.

A basic property of trees is that any two distinct vertices are connected by a unique path in Γ (see Theorem 2.1.3, page 52 of [44]). If (Γ, r) is a rooted tree, then we define $\text{vert}\Delta = \text{vert}\Gamma$ and we define the set $\text{oedge}\Delta$ to consist of all ordered pairs (v_{n-1}, v), where $v \in \text{vert}\Gamma \setminus \{r\}$ and where $(v_0, v_1, \ldots, v_n)$ is the unique path in Γ from the root vertex $r = v_0$ to the vertex $v = v_n$. In this manner every rooted tree defines an oriented graph, whose underlying graph is the original tree. Because of the bijection $\text{oedge}\Delta \to \text{edge}\Gamma$, we will treat rooted trees as oriented graphs without explicitly mentioning ordered edges or the digraph Δ.

Now we are ready to define a mapping $\phi \colon \mathcal{S}^{\text{vert}\Delta} \to \mathcal{S} \times G_p^{\text{oedge}\Delta}$, where (Δ, r) is a rooted oriented graph. Suppose $\mathcal{E} \colon \text{vert}\Delta \to \mathcal{S}$ is given. Then $\phi(\mathcal{E}) = (\mathcal{E}(r), \mathcal{A}_\mathcal{E})$, where $\mathcal{A}_\mathcal{E} \colon \text{oedge}\Delta \to G_p$ is defined on the oriented edge $e = (v_p, v_c) \in \text{oedge}\Delta$ by the rule that $\mathcal{A}_\mathcal{E}(e)$ is the unique element of G_p such that $\mathcal{E}(v_c) = \mathcal{E}(v_p)\mathcal{A}_\mathcal{E}(e)$. $\mathcal{A}_\mathcal{E}(e) \in G_p$ is uniquely determined because the right action of G_p on $\mathcal{S}$ is transitive and fixed point free. In order to see that the map ϕ is smooth it suffices to show that $\mathcal{A}_\mathcal{E}(e)$ depends smoothly on $\mathcal{E}$. Let $E \in \mathcal{S}$ be fixed. Then a diffeomorphism $G_p \to \mathcal{S} \colon T \mapsto ET$ is determined. Let $T_\mathcal{E} \colon \text{vert}\Delta \to G_p$ be defined by the rule $\mathcal{E}(v) = ET_\mathcal{E}(v)$. For each fixed v we regard $T_\mathcal{E}(v)$ as a "coordinate" expression of $\mathcal{E}(v)$. So we must show that $\mathcal{A}_\mathcal{E}(e)$ depends smoothly on $T_\mathcal{E}(v_p)$ and $T_\mathcal{E}(v_c)$. Since $ET_\mathcal{E}(v_c) = ET_\mathcal{E}(v_p)\mathcal{A}_\mathcal{E}(e)$, we have by the fact that the right action of G_p on $\mathcal{S}$ is fixed point free that $T_\mathcal{E}(v_c) = T_\mathcal{E}(v_p)\mathcal{A}_\mathcal{E}(e)$, or $\mathcal{A}_\mathcal{E}(e) = T_\mathcal{E}(v_p)^{-1}T_\mathcal{E}(v_c)$, which is an unquestionably smooth dependence.

Theorem. *Suppose (Δ, r) is a rooted oriented graph and Γ is the underlying graph of Δ. If Γ is a tree we do not necessarily assume that Δ is the oriented graph associated to (Γ, r). Suppose $\phi \colon \mathcal{S}^{vert\Delta} \to \mathcal{S} \times G_p^{oedge\Delta}$ is the smooth map defined above. Then ϕ is a diffeomorphism if and only if Γ is a tree.*

Proof. First suppose ϕ is a diffeomorphism. In particular it is injective. If Γ is not connected, then there is a connected component subgraph Γ' of Γ (i.e. a maximal

connected subgraph) which does not contain the root r as a vertex (see page 18 of [44]). If $\mathcal{E}_1 \in \mathcal{S}^{\mathrm{vert}\Gamma} = \mathcal{S}^{\mathrm{vert}\Delta}$ and $M \in G_a \setminus \{\mathbf{1}\}$ are given then define

$$\mathcal{E}_2(v) = \begin{cases} M\mathcal{E}_1(v) & \text{if } v \in \mathrm{vert}\Gamma', \\ \mathcal{E}_1(v) & \text{if } v \in \mathrm{vert}\Gamma \setminus \mathrm{vert}\Gamma'. \end{cases}$$

Since G_a acts fixed point free on $\mathcal{S}$ we see that $\mathcal{E}_1 \neq \mathcal{E}_2$. However we claim that $\phi(\mathcal{E}_1) = \phi(\mathcal{E}_2)$, which contradicts the injectivity of ϕ, showing that Γ must be connected. $\phi(\mathcal{E}_1) = (\mathcal{E}_1(r), \mathcal{A}_{\mathcal{E}_1})$, where $\mathcal{E}_1(v_c) = \mathcal{E}_1(v_p)\mathcal{A}_{\mathcal{E}_1}((v_p, v_c))$ for every oriented edge $(v_p, v_c) \in \mathrm{oedge}\Delta$. Since $r \notin \mathrm{vert}\Gamma'$ we have $\mathcal{E}_1(r) = \mathcal{E}_2(r)$. If $\{v_p, v_c\} \in \mathrm{edge}\Gamma'$ where $(v_p, v_c) \in \mathrm{oedge}\Delta$ then we may apply M to both sides of the equation $\mathcal{E}_1(v_c) = \mathcal{E}_1(v_p)\mathcal{A}_{\mathcal{E}_1}((v_p, v_c))$ to obtain $\mathcal{E}_2(v_c) = \mathcal{E}_2(v_p)\mathcal{A}_{\mathcal{E}_1}((v_p, v_c))$. If $(v_p, v_c) \in \mathrm{edge}\Gamma \setminus \mathrm{edge}\Gamma'$, we also have the equality $\mathcal{E}_2(v_c) = \mathcal{E}_2(v_p)\mathcal{A}_{\mathcal{E}_1}((v_p, v_c))$. Thus for all oriented edges $(v_p, v_c) \in \mathrm{oedge}\Delta$ we have $\mathcal{A}_{\mathcal{E}_1}((v_p, v_c)) = \mathcal{A}_{\mathcal{E}_2}((v_p, v_c))$. Thus $\phi(\mathcal{E}_2) = (\mathcal{E}_2(r), \mathcal{A}_{\mathcal{E}_2}) = (\mathcal{E}_1(r), \mathcal{A}_{\mathcal{E}_1}) = \phi(\mathcal{E}_1)$ as claimed.

If ϕ is a diffeomorphism then it must also be a surjective mapping. Suppose the graph Γ contains a cycle $(v_0, v_1, \ldots, v_n)$, $n \geq 3$. Because Δ is an oriented graph we can define the following.

$$\omega_i = \begin{cases} 1 & \text{if } (v_{i-1}, v_i) \in \mathrm{oedge}\Delta, \\ -1 & \text{if } (v_i, v_{i-1}) \in \mathrm{oedge}\Delta, \end{cases}$$

for $i = 1, \ldots, n$. Then for all $\mathcal{E} \in \mathcal{S}^{\mathrm{vert}\Gamma}$ we have the identity

$$\mathcal{A}_{\mathcal{E}}((v_0, v_1))^{\omega_1} \cdot \mathcal{A}_{\mathcal{E}}((v_1, v_2))^{\omega_2} \cdot \cdots \cdot \mathcal{A}_{\mathcal{E}}((v_{n-1}, v_n))^{\omega_n} = \mathbf{1}.$$

Since there are plenty of mappings $\mathcal{A} \in G_p^{\mathrm{oedge}\Delta}$ which do not satisfy this identity, we see that the mapping ϕ is not surjective. This contradction shows that no cycle can exist in Γ. Hence when ϕ is a diffeomorphism, the graph Γ must be a tree.

Now suppose Γ is a tree. Let Δ' be the oriented graph associated to the rooted tree (Γ, r). Then the mapping $\Omega \colon G_p^{\mathrm{oedge}\Delta'} \to G_p^{\mathrm{oedge}\Delta} \colon \mathcal{A}' \mapsto \mathcal{A}$ is a diffeomorphism, where for each $(v_1, v_2) \in \mathrm{oedge}\Delta$ define

$$\mathcal{A}((v_1, v_2)) = \begin{cases} \mathcal{A}'((v_1, v_2)) & \text{if } (v_1, v_2) \in \mathrm{oedge}\Delta', \\ \mathcal{A}'((v_2, v_1))^{-1} & \text{if } (v_2, v_1) \in \mathrm{oedge}\Delta'. \end{cases}$$

Let $\phi' \colon \mathcal{S}^{\mathrm{vert}\Delta'} \to \mathcal{S} \times G_p^{\mathrm{oedge}\Delta'}$ be the smooth mapping associated to Δ' as defined above. It suffices to prove that ϕ' is a diffeomorphism since then $\phi = (1_{\mathcal{S}} \times \Omega) \circ \phi'$ will also be a diffeomorphism. Thus without loss of generality we assume $\Delta = \Delta'$, and drop the primes. Furthermore, since $\mathrm{vert}\Delta = \mathrm{vert}\Gamma$ and the map $\mathrm{oedge}\Delta \to \mathrm{edge}\Gamma \colon (v_p, v_c) \mapsto \{v_p, v_c\}$ is a bijection, we may regard ϕ as defining a mapping $\mathcal{S}^{\mathrm{vert}\Gamma} \to \mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$, which we will also denote by ϕ.

We wish to define a smooth map $\psi \colon \mathcal{S} \times G_p^{\mathrm{edge}\Gamma} \to \mathcal{S}^{\mathrm{vert}\Gamma}$ which will be the inverse of ϕ. Suppose $(E, \mathcal{A}) \in \mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$ is given. Suppose $v \in \mathrm{vert}\Gamma$; then define $\psi(E, \mathcal{A})(v) = E$ when $v = r$, and when $v \neq r$

$$\psi(E, \mathcal{A})(v) = E\mathcal{A}(\{v_0, v_1\}) \cdot A(\{v_1, v_2\}) \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\}),$$

where $(v_0, v_1, \ldots, v_n)$ is the unique path in Γ with $v_0 = r$ and $v_n = v$. This mapping is clearly well-defined and smoothly dependent on the arguments $(E, \mathcal{A}) \in \mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$.

It remains to prove that both $\phi \circ \psi$ and $\psi \circ \phi$ are identity maps on their respective domains. Suppose $(E, \mathcal{A}) \in \mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$, and define $\mathcal{E} = \psi(E, \mathcal{A})$. Let $\phi(\mathcal{E}) =$

$(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})$. Clearly $\mathcal{E}(r) = E$. So we must show that $\mathcal{A}_{\mathcal{E}} = \mathcal{A}$. Suppose $e = \{v_p, v_c\} \in \text{edge}\Gamma$, where $(v_p, v_c) \in \text{oedge}\Delta$. Let $(v_0, v_1, \ldots, v_n)$ be the unique path in Γ with $v_0 = r$ and $v_n = v_c$. Noting the definition of the parent we see that $v_p = v_{n-1}$. Thus $(v_0, v_1, \ldots, v_{n-1})$ is the unique path in Γ with $v_0 = r$ and $v_{n-1} = v_p$. By the definition of $\mathcal{E} = \psi(E, \mathcal{A})$ we have that

$$\mathcal{E}(v_p) = E\mathcal{A}(\{v_0, v_1\}) \cdot A(\{v_1, v_2\}) \cdot \cdots \cdot \mathcal{A}(\{v_{n-2}, v_{n-1}\}),$$
$$\mathcal{E}(v_c) = E\mathcal{A}(\{v_0, v_1\}) \cdot A(\{v_1, v_2\}) \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\}).$$

Therefore $\mathcal{E}(v_c) = \mathcal{E}(v_p)\mathcal{A}(\{v_{n-1}, v_n\}) = \mathcal{E}(v_p)\mathcal{A}(\{v_p, v_c\}) = \mathcal{E}(v_p)\mathcal{A}_{\mathcal{E}}(e)$. Thus $\mathcal{A}_{\mathcal{E}}(e) = \mathcal{A}(e)$ as we desired to prove. Thus $\phi \circ \psi$ is the identity on $\mathcal{S} \times G_p^{\text{edge}\Gamma}$.

To show that $\psi \circ \phi$ is the identity on $S^{\text{vert}\Gamma}$, let $\mathcal{E} \in S^{\text{vert}\Gamma}$ be given, and let $\phi(\mathcal{E}) = (\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})$. We must show for every $v \in \text{vert}\Gamma$ that $\mathcal{E}(v) = \psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v)$. First of all, if $v = r$ this is clear. We consider the root to be connected to itself via a path of length zero. Suppose the equation $\mathcal{E}(v) = \psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v)$ is true for all $v \in \text{vert}\Gamma$ which are connected to the root by a path of length less than $n \geq 1$. Suppose $v \in \text{vert}\Gamma$ is connected to the root by a path $(v_0, v_1, \ldots, v_n)$ of length n. Thus $v_0 = r$ and $v_n = v$. The parent of v is v_{n-1}, which is connected to the root by a path of length less than n, so by the induction hypothesis we have $\mathcal{E}(v_{n-1}) = \psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v_{n-1})$. By the definition of ψ we have

$$\begin{aligned}\psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v_n) &= \mathcal{E}(r)\mathcal{A}_{\mathcal{E}}(\{v_0, v_1\}) \cdot \cdots \cdot A_{\mathcal{E}}(\{v_{n-2}, v_{n-1}\}) \cdot \mathcal{A}_{\mathcal{E}}(\{v_{n-1}, v_n\}), \\ &= \psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v_{n-1})\mathcal{A}_{\mathcal{E}}(\{v_{n-1}, v_n\}), \\ &= \mathcal{E}(v_{n-1})\mathcal{A}_{\mathcal{E}}(\{v_{n-1}, v_n\}) = \mathcal{E}(v_n).\end{aligned}$$

The last equality follows from the definition of $\mathcal{A}_{\mathcal{E}}$. Thus by induction we have established the equality $\mathcal{E}(v) = \psi(\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})(v)$ for all $v \in \text{vert}\Gamma$. This finishes the proof that $\psi \circ \phi$ is the identity on $S^{\text{vert}\Gamma}$, and so both ϕ and ψ are diffeomorphisms. □

Throughout the remainder of this paper we adopt the notation used in the above proof, namely whenever (Γ, r) is a rooted tree it is assumed to be equipped with its associated orientation, and we have a diffeomorphism $\phi\colon \mathcal{S}^{\text{vert}\Gamma} \to \mathcal{S} \times G_p^{\text{edge}\Gamma}$.

To say that ϕ is equivariant under the left action of G_a means simply that for all $M \in G_a$ and all $\mathcal{E} \in \mathcal{S}^{\text{vert}\Gamma}$ we have that $\phi(M\mathcal{E}) = ((M\mathcal{E})(r), \mathcal{A}_{M\mathcal{E}}) = (M\mathcal{E}(r), \mathcal{A}_{\mathcal{E}})$. G_a has a left action on $\mathcal{S}^{\text{vert}\Gamma}$ by the rule: $(M\mathcal{E})(v) = M\mathcal{E}(v)$ in terms of the left action of G_a on $\mathcal{S}$. Thus we certainly have the equation $(M\mathcal{E})(r) = M\mathcal{E}(r)$. So it remains to show that $\mathcal{A}_{M\mathcal{E}} = \mathcal{A}_{\mathcal{E}}$. Let $e = \{v_p, v_c\} \in \text{edge}\Gamma$. Then $(M\mathcal{E})(v_c) = (M\mathcal{E})(v_p)\mathcal{A}_{M\mathcal{E}}(e)$. Thus $M\mathcal{E}(v_c) = M\mathcal{E}(v_p)\mathcal{A}_{M\mathcal{E}}(e)$. Multiplying on the left by M^{-1} we obtain $\mathcal{E}(v_c) = \mathcal{E}(v_p)\mathcal{A}_{M\mathcal{E}}(e)$. By the definition of $\mathcal{A}_{\mathcal{E}}$ we therefore have $\mathcal{E}(v_p)\mathcal{A}_{M\mathcal{E}}(e) = \mathcal{E}(v_p)\mathcal{A}_{\mathcal{E}}(e)$. This implies that $\mathcal{A}_{M\mathcal{E}}(e) = \mathcal{A}_{\mathcal{E}}(e)$, as we desired to show. It follows that ψ is also left equivariant.

The map ϕ induces a map $\tilde{\phi}\colon G_a \backslash \mathcal{S}^{\text{vert}\Gamma} \to G_p^{\text{edge}\Gamma}$ which is a bijection. Since the left action of G_a on both $\mathcal{S}^{\text{vert}\Gamma}$ and $\mathcal{S} \times G_p^{\text{edge}\Gamma}$ is fixed point free and proper, the orbit spaces $G_a \backslash \mathcal{S}^{\text{vert}\Gamma}$ and $G_p^{\text{edge}\Gamma}$ are manifolds, and the induced bijection $\tilde{\phi}$ is a diffeomorphism (see Proposition 4.1.23 on page 266 of [1]). This diffeomorphism captures the intuitive fact that the aspects of molecular configuration which are invariant under spatial translations and proper rotations, i.e. molecular conformation, can be expressed in terms of the relative positions and orientations between the parts of the molecule.

3. From Configuration to Coordinates

3.1. **Conforming Pose Assignments.** Suppose $\mathcal{G}$ is the graph for a molecule. We will use the notation of section 2.2. A *molecular configuration* is an assignment of a point in space to each atom of $\mathcal{G}$, i.e. an element $\mathcal{X} \in X^{\mathrm{vert}\mathcal{G}}$. Here X is our model of three dimensional space as in section 2.1. $X^{\mathrm{vert}\mathcal{G}}$ is a real manifold of dimension $3N$, where N is the number of atoms (vertices) of $\mathcal{G}$. Not every point of this manifold represents a realistic molecular configuration when the covalent bonding network is considered. For example the forces between the nuclei in the molecule are such that the energy of a configuration tends to infinity as any two nuclei approach one another in spatial position. If two atoms are covalently bonded then the distance between the positions of their nuclei is relatively fixed, meaning that there is a high energetic penalty for distances which deviate significantly from the ideal distance. The measure of the angle between two bonds which share a common atom are also relatively fixed. These aspects have to do with the energy of a particular molecular configuration, and so are outside the focus of this work, despite their importance. However, we wish to describe molecular configurations in terms which will make the restrictions imposed by energy considerations easy to deal with.

Our approach to this problem is to associate with the molecular configuration $\mathcal{X}$ a family of poses which are especially well conformed to the configuration $\mathcal{X}$. Then we can use the formalism developed in section 2.3 to define internal coordinates. We will not try to decide on a single pose for each atom $A \in \mathrm{vert}\mathcal{G}$; there are several natural such poses, all of which are equally valid. Instead, we will assign a pose to each vertex (a, b, A) of $AL^2(\mathcal{G})$, provided the geometry of the angle is nondegenerate so that a unique such pose is determined. Suppose $a = \{b, b'\}$, $b = \{A_1, A_2\}$, $b' = \{A_2, A_3\}$, and either $A = A_1$ or $A = A_2$. Define the vectors $\mathbf{U} = \mathcal{X}(A_1) - \mathcal{X}(A_2)$, and $\mathbf{V} = \mathcal{X}(A_3) - \mathcal{X}(A_2)$ in V. The nondegeneracy conditions we require are the following:

(1) $\|\mathbf{U}\| > 0$, $\|\mathbf{V}\| > 0$, and
(2) $|\mathbf{U} \cdot \mathbf{V}| < \|\mathbf{U}\| \, \|\mathbf{V}\|$.

These two conditions depend only on the angle a of the vertex (a, b, A). Under these nondegeneracy conditions we define the pose $\mathcal{E}(a, b, A) = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ by the rule:

$$\begin{aligned}
\mathbf{e}_0 &= \mathcal{X}(A), \\
\mathbf{e}_3 &= \begin{cases} -\mathbf{U}/\|\mathbf{U}\| & \text{if } A = A_1, \\ \mathbf{U}/\|\mathbf{U}\| & \text{if } A = A_2, \end{cases} \\
\mathbf{e}_1 &= \frac{\mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})}{\|\mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})\|}, \\
\mathbf{e}_2 &= \mathbf{e}_3 \times \mathbf{e}_1.
\end{aligned}$$

The first nondegeneracy condition insures that $\mathbf{e}_3$ is well-defined, and the second insures that $\mathbf{e}_1$ is well-defined since

$$\|\mathbf{U}\|^2 \|\mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})\|^2 = \|\mathbf{U}\|^2 [\mathbf{V} \cdot \mathbf{V} - (\mathbf{e}_3 \cdot \mathbf{V})^2] = \|\mathbf{U}\|^2 \|\mathbf{V}\|^2 - (\mathbf{U} \cdot \mathbf{V})^2 > 0.$$

Thus the geometric significance of the amalgamated twice iterated line graph should now be clearer: there are assignments of naturally conformed poses to vertices when a nondegenerate molecular configuration is given.

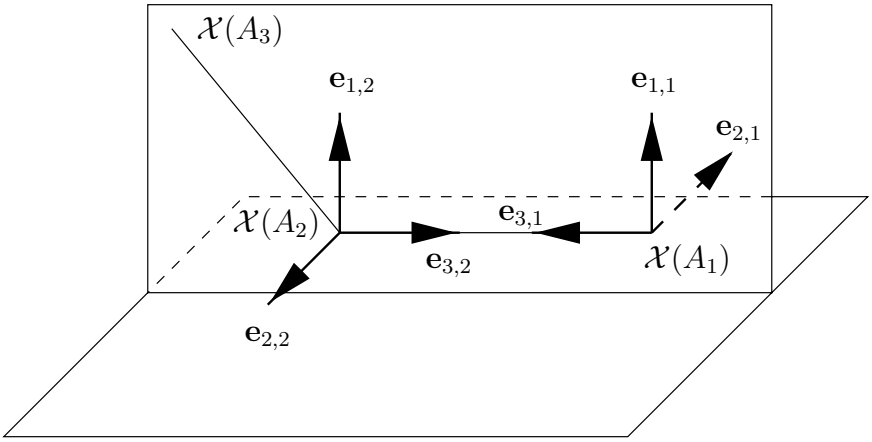

FIGURE 5. Poses at A_1 and A_2 conformed to the angle $a = \{b, b'\}$ and to the bond $b = \{A_1, A_2\}$. The bond b' is $\{A_2, A_3\}$. The pose $(\mathbf{e}_{0,i}, \mathbf{e}_{1,i}, \mathbf{e}_{2,i}, \mathbf{e}_{3,i})$ at $\mathcal{X}(A_i)$ is associated to the triple (a, b, A_i), $i = 1, 2$. The origin $\mathbf{e}_{0,i}$ of each pose is the position $\mathcal{X}(A_i)$ of A_i.

To assign a pose to every vertex of $AL^2(\mathcal{G})$ would require rather stringent nondegeneracy requirements on the molecular configuration. For example it would not be possible if the angle between two bonds sharing an atom was π radians; such bond angles do occur occasionally in biologically interesting molecules. For example, the O—C—O bond angle of the Carbon Dioxide molecule, the N—$H \cdots O$ (Hydrogen-) bond angle in the α-helices of proteins (see page 170 of [32]), or the (Histidine 93)N—Fe—O bond angle in the Heme group of hemoglobin (see page 218 of [32]). These situations are rare enough that we can work around them; thus we only assign poses to the vertices of a subgraph Γ of $AL^2(\mathcal{G})$, where the bond angles failing to meet the nondegeneracy conditions are not included in Γ.

We use the following notation. If $Y_1, \ldots, Y_n$ are sets then $\pi_j \colon Y_1 \times \cdots \times Y_n \to Y_j$, $j = 1, \ldots, n$, denote the projection mappings onto the jth factor. If $Z \subset Y_1 \times \cdots \times Y_3$ then $\pi_j(Z) \subset Y_j$ is the set of all projections of elements of Z via the mapping π_j.

Definition. If Γ is any subgraph of $AL^2(\mathcal{G})$, then we define $X_\Gamma \subset X^{\mathrm{vert}\mathcal{G}}$ to consist of those molecular configurations which satisfy the nondegeneracy conditions (1) and (2) for each $a \in \pi_1 \mathrm{vert}\Gamma$.

For each angle $a \in \pi_1 \mathrm{vert}\Gamma$ the set of molecular configurations not satisfying nondegeneracy condition (1) is the union of two closed submanifolds of codimension 3. Away from the union of these two, the set of molecular configurations not satisfying nondegeneracy condition (2) is a submanifold of codimension 2, whose closure is the union of all three submanifolds. This exceptional set has empty interior. Hence X_Γ is an open dense subset of the manifold $X^{\mathrm{vert}\mathcal{G}}$.

For most biologically important molecules it seems to be possible to choose the subgraph Γ so that the all the low energy configurations that occur under normal biological conditions lie in X_Γ. For such well-chosen Γ the subset X_Γ should be considered to be very large, since it contains the only configurations of interest. If one is however interested in reactions taking place under highly energetic conditions, such as in the damage done to biological systems by elevated temperatures or

exposure to high frequency radiation, then it may be necessary to consider molecular configurations which lie outside of X_Γ. Such configurations might lie in $X_{\Gamma'}$ for some other subgraph Γ' of $AL^2(\mathcal{G})$. To each judiciously chosen subgraph Γ we will associate a system of internal coordinates whose domain of applicability will be X_Γ. Even exceptional configurations will hopefully be covered by considering multiple systems. However in the remainder of this work we will consider only a single system (see however section 6).

The pose assignment $\mathcal{E}$ we have described depends on the molecular configuration $\mathcal{X}$; hence we write $\mathcal{E} = \gamma(\mathcal{X})$. This gives us a map $\gamma\colon X_\Gamma \to \mathcal{S}^{\text{vert}\Gamma}$. Let $\mathcal{S}_\Gamma$ denote the range of this mapping. The elements $\mathcal{E} \in \mathcal{S}_\Gamma$ are called *conformed pose assignments*.

The group G_a acts on $X^{\text{vert}\mathcal{G}}$ on the left by the rule: for all $A \in \text{vert}\mathcal{G}$ and all $\mathcal{X} \in X^{\text{vert}\mathcal{G}}$ and all $M \in G_a$ we have $(M\mathcal{X})(A) = M\mathcal{X}(A)$, in terms of the left action of G_a on X. Since the nondegeneracy conditions involve only concepts (dot products of vectors in V obtained by subtracting two points of X) which are invariant under this group action, we see that the set X_Γ is left invariant. An orbit $G_a\mathcal{X}$, where $\mathcal{X} \in X^{\text{vert}\mathcal{G}}$ is called a *molecular conformation*. The left action of G_a on $\mathcal{S}^{\text{vert}\Gamma}$ was defined in section 2.3. As discussed in section 2.1 each $M \in G_a$ is actually a linear isomorphism of the 4-dimensional vector space W, hence inspection of the formulae defining γ shows that the mapping γ is left equivariant: $\gamma(M\mathcal{X}) = M\gamma(\mathcal{X})$, for all $\mathcal{X} \in X_\Gamma$ and all $M \in G_a$. So $\mathcal{S}_\Gamma$ is left invariant under the action of G_a. Thus conformed poses rotate and translate with the molecule.

Theorem. *If Γ is a subgraph of $AL^2(\mathcal{G})$ and $\pi_3(\text{vert}\Gamma) = \text{vert}\mathcal{G}$ then γ is a smooth embedding, and $\mathcal{S}_\Gamma$ is an embedded submanifold of $\mathcal{S}^{\text{vert}\Gamma}$.*

Proof. Suppose $\mathcal{X}_1, \mathcal{X}_2 \in X_\Gamma$ such that $\gamma(\mathcal{X}_1) = \gamma(\mathcal{X}_2)$. For each $(a, b, A) \in \text{vert}\Gamma$ we have that $\pi_1[\gamma(\mathcal{X}_i)(a, b, A)] = \mathcal{X}_i(A)$, $i = 1, 2$. Thus $\mathcal{X}_1(A) = \mathcal{X}_2(A)$ for all $A \in \pi_3(\text{vert}\Gamma)$. Thus the map γ will be injective if we have the condition $\pi_3(\text{vert}\Gamma) = \text{vert}\mathcal{G}$ on the graph Γ.

By introducing a single fixed pose $E \in \mathcal{S}$, we can coordinatize $X^{\text{vert}\mathcal{G}}$ as $(\mathbb{R}^3)^{\text{vert}\mathcal{G}}$, X_Γ as $X_{\Gamma,E} \subset (\mathbb{R}^3)^{\text{vert}\mathcal{G}}$, $\mathcal{S}_\Gamma$ as $\mathcal{S}_{\Gamma,E}$, and $\mathcal{S}^{\text{vert}\Gamma}$ as $G_p^{\text{vert}\Gamma}$. As a manifold we have $G_p = \mathbb{R}^3 \times \text{SO}(3)$. The coordinate expressions defining γ are then seen to define a smooth mapping γ_E. We have the following commutative diagram:

$$\begin{array}{ccc} X_{\Gamma,E} & \xrightarrow{\gamma_E} & (\mathbb{R}^3 \times \text{SO}(3))^{\text{vert}\Gamma} \\ {\scriptstyle\text{inclusion}}\downarrow & & \downarrow{\scriptstyle \mathcal{E} \mapsto \pi_1 \circ \mathcal{E}} \\ (\mathbb{R}^3)^{\text{vert}\mathcal{G}} & \xrightarrow{\mathcal{X} \mapsto \mathcal{X} \circ \pi_3} & (\mathbb{R}^3)^{\text{vert}\Gamma} \end{array}$$

In the above $\pi_1\colon \mathbb{R}^3 \times \text{SO}(3) \to \mathbb{R}^3$ is the projection mapping. We claim that γ_E is an open map onto $\mathcal{S}_{\Gamma,E}$ (equipped with the subspace topology). Suppose U_A is an open subset of $\mathbb{R}^3$ for each $A \in \text{vert}\mathcal{G}$ such that $\prod_{A \in \text{vert}\mathcal{G}} U_A \subset X_{\Gamma,E}$. To see that $\gamma_E(\prod_{A \in \text{vert}\mathcal{G}} U_A)$ is open in the subspace topology of $\mathcal{S}_{\Gamma,E}$ we seek an open subset $\mathcal{U}$ of $(\mathbb{R}^3 \times \text{SO}(3))^{\text{vert}\Gamma}$ such that $\gamma_E(\prod_{A \in \text{vert}\mathcal{G}} U_A) = \mathcal{S}_{\Gamma,E} \cap \mathcal{U}$. For this purpose we define the open set $\mathcal{V} = \prod_{v \in \text{vert}\Gamma} U_{\pi_3(v)} \subset (\mathbb{R}^3)^{\text{vert}\Gamma}$. Since the right-most vertical mapping $\mathcal{E} \mapsto \pi_1 \circ \mathcal{E}$ in the above commutative diagram is continuous, the inverse image $\mathcal{U}$ of $\mathcal{V}$ under this mapping is open in $(\mathbb{R}^3 \times \text{SO}(3))^{\text{vert}\Gamma}$. If $\mathcal{X} \in \prod_{A \in \text{vert}\mathcal{G}} U_A$ and $\mathcal{E} = \gamma_E(\mathcal{X})$ then the commutivity of the above diagram means that $\pi_1 \circ \mathcal{E} = \mathcal{X} \circ \pi_3 \in \mathcal{V}$ and thus $\mathcal{E} \in \mathcal{U}$. Therefore $\gamma_E(\prod_{A \in \text{vert}\mathcal{G}} U_A) \subset \mathcal{S}_{\Gamma,E} \cap \mathcal{U}$. To show the

reverse inclusion suppose $\mathcal{E} \in \mathcal{S}_{\Gamma,E} \cap \mathcal{U}$, so that $\mathcal{E} = \gamma_E(\mathcal{X})$ for some $\mathcal{X} \in X_{\Gamma,E}$. Therefore for all $v = (a, b, A) \in \mathrm{vert}\Gamma$ we have $\mathcal{X}(A) = (\mathcal{X} \circ \pi_3)(v) = (\pi_1 \circ \mathcal{E})(v) \in U_{\pi_3(v)} = U_A$. Since $\pi_3(\mathrm{vert}\Gamma) = \mathrm{vert}\mathcal{G}$ we have that $\mathcal{X} \in \prod_{A \in \mathrm{vert}\mathcal{G}} U_A$. Thus $\mathcal{E} \in \gamma_E(\prod_{A \in \mathrm{vert}\mathcal{G}} U_A)$, and the reverse inclusion is true. Thus γ_E is an open map onto $\mathcal{S}_{\Gamma,E}$.

From this commutative diagram and the condition $\pi_3(\mathrm{vert}\Gamma) = \mathrm{vert}\mathcal{G}$ it is also clear that for every $\mathcal{X} \in X_{\Gamma,E}$ the linear map $D\gamma_E(\mathcal{X})\colon (\mathbb{R}^3)^{\mathrm{vert}\mathcal{G}} \to (\mathbb{R}^3 \times \mathbb{R}^9)^{\mathrm{vert}\Gamma}$ is injective, so that γ_E is immersive at each point of its domain. Thus γ is an injective immersion which is an open map onto $\mathcal{S}_\Gamma$ equipped with its subspace topology. Therefore γ is an embedding, and $\mathcal{S}_\Gamma$ is an embedded submanifold of $\mathcal{S}^{\mathrm{vert}\Gamma}$ (see 1.6Fa on page 51 of [1]). □

Under the assumptions of this theorem the codomain restricted map $\gamma\colon X_\Gamma \to \mathcal{S}_\Gamma$ is a diffeomorphism, hence it is permissible to regard elements of $\mathcal{S}_\Gamma$ as molecular configurations. This justifies the intuition we gave in section 2.3. In that case the map γ determines a bijection $G_a \backslash X_\Gamma \cong G_a \backslash \mathcal{S}_\Gamma$. Thus orbits $G_a\mathcal{E}$, for $\mathcal{E} \in \mathcal{S}_\Gamma$, can also be considered as being molecular conformations.

3.2. **Defining Internal Coordinates.** Suppose Γ is a subgraph of $AL^2(\mathcal{G})$, and let a root $r \in \mathrm{vert}\Gamma$ be chosen and fixed. Suppose Δ is an oriented graph whose underlying graph is Γ. The left G_a-equivariant map $\gamma\colon X_\Gamma \to S^{\mathrm{vert}\Gamma}$ of the previous section is well-defined, with range $\mathcal{S}_\Gamma$. Also the left G_a-equivariant map $\phi\colon \mathcal{S}^{\mathrm{vert}\Delta} \to \mathcal{S} \times G_p^{\mathrm{oedge}\Delta}$ of section 2.3 exists, and therefore we can map $\mathcal{S}_\Gamma = \gamma(X_\Gamma) \subset \mathcal{S}^{\mathrm{vert}\Gamma}$ into the subset $\phi(\mathcal{S}_\Gamma)$ of $\mathcal{S} \times G_p^{\mathrm{oedge}\Delta}$. Define $G_\Gamma = \pi_2\phi(\mathcal{S}_\Gamma) \subset G_p^{\mathrm{oedge}\Delta}$. Clearly $\phi(\mathcal{S}_\Gamma) \subset \mathcal{S} \times G_\Gamma$. But if $(E, \mathcal{A}) \in \mathcal{S} \times G_\Gamma$ let $\mathcal{E} \in \mathcal{S}_\Gamma$ and $E' \in \mathcal{S}$ such that $\phi(\mathcal{E}) = (E', \mathcal{A})$. Let $M \in G_a$ such that $ME' = E$. Then $(E, \mathcal{A}) = (ME', \mathcal{A}) = M(E', \mathcal{A}) = M\phi(\mathcal{E}) = \phi(M\mathcal{E})$, since ϕ is left G_a-equivariant. Therefore $(E, \mathcal{A}) \in \phi(\mathcal{S}_\Gamma)$. Hence $\phi(\mathcal{S}_\Gamma) = \mathcal{S} \times G_\Gamma$. Our task in this section will be to describe the subset G_Γ in more detail and to define internal coordinates.

Let $\mathcal{E} \in \mathcal{S}_\Gamma$ be a conformed pose assignment. Using the orientation coming from Δ we identify $\mathrm{oedge}\Delta$ with $\mathrm{edge}\Gamma$. The set $\mathrm{edge}\Gamma$ can be decomposed as a union of disjoint subsets $\mathrm{edge}\Gamma = \mathrm{edge}_0\Gamma \cup \mathrm{edge}_1\Gamma \cup \mathrm{edge}_2\Gamma$, since Γ is a subgraph of $AL^2(\mathcal{G})$. Suppose $\{(a, b, A_1), (a, b, A_2)\} \in \mathrm{edge}_0\Gamma$, where (a, b, A_1) is the parent vertex and (a, b, A_2) is the child vertex. Let $(\mathbf{e}_{0,i}, \mathbf{e}_{1,i}, \mathbf{e}_{2,i}, \mathbf{e}_{3,i}) = \mathcal{E}(a, b, A_i)$, for $i = 1, 2$. Then $b = \{A_1, A_2\}$, and $\mathbf{e}_{0,2} = \mathbf{e}_{0,1} + \mathbf{e}_{3,1}\|\mathbf{e}_{0,2} - \mathbf{e}_{0,1}\|$, $\mathbf{e}_{3,2} = -\mathbf{e}_{3,1}$, $\mathbf{e}_{1,1} = \mathbf{e}_{1,2}$, and $\mathbf{e}_{2,2} = \mathbf{e}_{3,2} \times \mathbf{e}_{1,2} = -\mathbf{e}_{3,1} \times \mathbf{e}_{1,1} = -\mathbf{e}_{2,1}$. (See Figure 5.) This implies that

$$(\mathbf{e}_{0,2}, \mathbf{e}_{1,2}, \mathbf{e}_{2,2}, \mathbf{e}_{3,2}) = (\mathbf{e}_{0,1}, \mathbf{e}_{1,1}, \mathbf{e}_{2,1}, \mathbf{e}_{3,1}) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ l & 0 & 0 & -1 \end{pmatrix}$$

where $l = \|\mathbf{e}_{0,2} - \mathbf{e}_{0,1}\|$, which is the length of the bond b, and which as a consequence of the nondegeneracy conditions is positive. Let $T_0(l)$ denote the above 4×4 real matrix, and let G_0 denote the set of all matrices of the form $T_0(l)$, $l > 0$. These matrices are equal to their own inverses, i.e. $T_0(l)^{-1} = T_0(l)$.

Suppose $\{(a, b_1, A), (a, b_2, A)\} \in \mathrm{edge}_1\Gamma$, where (a, b_1, A) is the parent vertex and (a, b_2, A) is the child. Let $(\mathbf{e}_{0,i}, \mathbf{e}_{1,i}, \mathbf{e}_{2,i}, \mathbf{e}_{3,i}) = \mathcal{E}(a, b_i, A)$, for $i = 1, 2$. Clearly $a = \{b_1, b_2\}$, and $\{A\} = b_1 \cap b_2$. Let $c = \mathbf{e}_{3,1} \cdot \mathbf{e}_{3,2}$. By nondegeneracy, we have that $c \in (-1, 1)$. Also set $s = \sqrt{1 - c^2}$. $(c, s) = (\cos\theta, \sin\theta)$ are the cosine and sine of

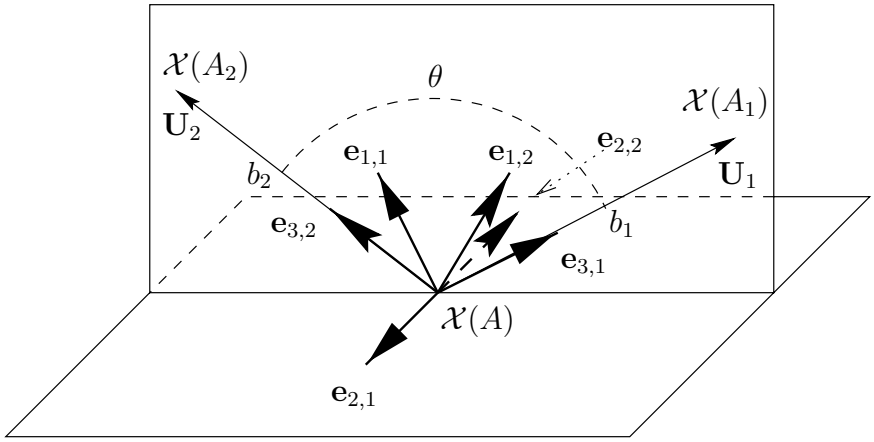

FIGURE 6. Poses associated with the triples (a, b_1, A) and (a, b_2, A), which are connected by an edge of type 1. $a = \{b_1, b_2\}$, and $b_i = \{A, A_i\}$, and $\mathbf{U}_i = \mathcal{X}(A_i) - \mathcal{X}(A)$, for $i = 1, 2$.

the angle θ between bonds b_1 and b_2. Clearly $\mathbf{e}_{0,1} = \mathbf{e}_{0,2}$, and $(\mathbf{e}_{1,2}, \mathbf{e}_{2,2}, \mathbf{e}_{3,2})$ can be obtained from $(\mathbf{e}_{1,1}, \mathbf{e}_{2,1}, \mathbf{e}_{3,1})$ by first rotating about the y-axis by an angle of θ, and then rotating about the z-axis by an angle of π. (See Figure 6.) Therefore

$$(\mathbf{e}_{0,2}, \mathbf{e}_{1,2}, \mathbf{e}_{2,2}, \mathbf{e}_{3,2}) = (\mathbf{e}_{0,1}, \mathbf{e}_{1,1}, \mathbf{e}_{2,1}, \mathbf{e}_{3,1}) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -c & 0 & s \\ 0 & 0 & -1 & 0 \\ 0 & s & 0 & c \end{pmatrix}.$$

Let $T_1(c)$ denote the above 4×4 real matrix, and let G_1 denote the set of all matrices in the form $T_1(c)$, $c \in (-1, 1)$. These matrices are also equal to their own inverses, i.e. $T_1(c)^{-1} = T_1(c)$.

Finally suppose $\{(a_1, b, A), (a_2, b, A)\} \in \mathrm{edge}_2\Gamma$, where (a_1, b, A) is the parent and (a_2, b, A) is the child. Let $(\mathbf{e}_{0,i}, \mathbf{e}_{1,i}, \mathbf{e}_{2,i}, \mathbf{e}_{3,i}) = \mathcal{E}(a_i, b, A)$, for $i = 1, 2$. Clearly $a_1 \cap a_2 = \{b\}$, $\mathbf{e}_{0,1} = \mathbf{e}_{0,2}$, and $\mathbf{e}_{3,1} = \mathbf{e}_{3,2}$. Thus $(\mathbf{e}_{1,2}, \mathbf{e}_{2,2}, \mathbf{e}_{3,2})$ can be obtained from $(\mathbf{e}_{1,1}, \mathbf{e}_{2,1}, \mathbf{e}_{3,1})$ by rotating about the z-axis through some angle φ. (See Figure 7.) Let $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$. Then we have

$$(\mathbf{e}_{0,2}, \mathbf{e}_{1,2}, \mathbf{e}_{2,2}, \mathbf{e}_{3,2}) = (\mathbf{e}_{0,1}, \mathbf{e}_{1,1}, \mathbf{e}_{2,1}, \mathbf{e}_{3,1}) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & x & -y & 0 \\ 0 & y & x & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

for some $z = x + iy = e^{i\varphi} \in S^1$. Let $T_2(z)$ denote the above 4×4 real matrix, and let G_2 denote the group of all matrices of the form $T_2(z)$, $z \in S^1$. $T_2(z)^{-1} = T_2(\overline{z})$, so the orientation of the edge in $\mathrm{edge}_2\Gamma$ does make a difference.

G_0, G_1, and G_2 are one dimensional submanifolds of G_p, diffeomorphic to $(0, \infty)$, $(-1, 1)$, and S^1 respectively. We may regard $T_0 \colon (0, \infty) \to G_0$, $T_1 \colon (-1, 1) \to G_1$, and $T_2 \colon S^1 \to G_2$ as the diffeomorphisms. We have seen that if $(\mathcal{E}(r), \mathcal{A}_\mathcal{E}) = \phi(\mathcal{E})$, and $e \in \mathrm{edge}_i\Gamma$, then $\mathcal{A}_\mathcal{E}(e) \in G_i$, for $i = 0, 1, 2$. This property follows directly from the fact that $\mathcal{E} \in \mathcal{S}_\Gamma$. Thus we have that $G_\Gamma \subset G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$.

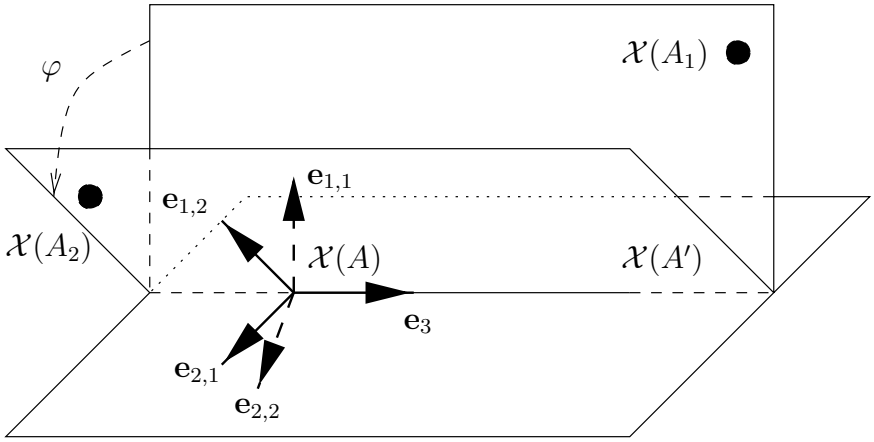

FIGURE 7. Poses associated with triples (a_1, b, A) and (a_2, b, A) which are connected by an edge of type 2. $b = \{A, A'\}$, and $b_i \setminus b = \{A_i\}$ and $a_i = \{b, b_i\}$ for $i = 1, 2$. The positions $\mathcal{X}(A_i)$ drawn suggest that this wedge is a dihedral, but we have not drawn the bonds b_i so that the improper case is also covered.

Using the diffeomorphisms T_0, T_1, T_2, we obtain a diffeomorphism

$$\kappa\colon (0,\infty)^{\mathrm{edge}_0\Gamma} \times (-1,1)^{\mathrm{edge}_1\Gamma} \times (S^1)^{\mathrm{edge}_2\Gamma} \to G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}.$$

Elements of $(0,\infty)^{\mathrm{edge}_0\Gamma} \times (-1,1)^{\mathrm{edge}_1\Gamma} \times (S^1)^{\mathrm{edge}_2\Gamma}$ constitute a list of *internal coordinates* associated to Γ.

It will be useful to have explicit expressions for the internal coordinates in terms of the molecular configuration $\mathcal{X} \in X_\Gamma$. If $e = \{(a, b, A_1), (a, b, A_2)\} \in \mathrm{edge}_0\Gamma$, then $\pi_2[(\phi \circ \gamma)(\mathcal{X})](e) = T_0(\|\mathcal{X}(A_1) - \mathcal{X}(A_2)\|)$. Thus the explicit expression for the *bond length coordinate* is:

$$l(e) = \|\mathcal{X}(A_1) - \mathcal{X}(A_2)\|.$$

If $e = \{(a, b_1, A), (a, b_2, A)\} \in \mathrm{edge}_1\Gamma$, then let $b_i = \{A, A_i\}$, $i = 1, 2$. According to our pose assignment construction, $\mathbf{U}_i = \mathcal{X}(A_i) - \mathcal{X}(A)$, and $\mathbf{e}_{3,i} = \mathbf{U}_i/\|\mathbf{U}_i\|$, $i = 1, 2$. We defined $c = \mathbf{e}_{3,1} \cdot \mathbf{e}_{3,2}$ (see Figure 6). Thus we have that $\pi_2[(\phi \circ \gamma)(\mathcal{X})](e) = T_1(c)$. Thus the explicit expression for the *bond angle cosine coordinate* is

$$c(e) = \frac{\mathcal{X}(A_1) - \mathcal{X}(A)}{\|\mathcal{X}(A_1) - \mathcal{X}(A)\|} \cdot \frac{\mathcal{X}(A_2) - \mathcal{X}(A)}{\|\mathcal{X}(A_2) - \mathcal{X}(A)\|}.$$

Finally, if $e = \{(a_1, b, A), (a_2, b, A)\} \in \mathrm{edge}_2\Gamma$, where (a_1, b, A) is the parent and (a_2, b, A) is the child, then let $\tilde{e} = ((a_1, b, A), (a_2, b, A)) \in \mathrm{oedge}\Delta$ denote the ordered edge. Let $a_i = \{b, b_i\}$, $b_i \setminus b = \{A_i\}$, $i = 1, 2$, and $b = \{A, A'\}$. According to our pose assignment construction, $\mathbf{U} = \mathcal{X}(A') - \mathcal{X}(A)$, and $\mathbf{e}_3 = \mathbf{e}_{3,1} = \mathbf{e}_{3,2} = \mathbf{U}/\|\mathbf{U}\|$. Also $\mathbf{V}_i = \mathcal{X}(A_i) - \mathcal{X}(\alpha_1(a_i))$, $i = 1, 2$. (Recall that α_1 maps an angle to the common atom of its two bonds.) $\alpha_1(a_i) \in \{A, A'\} = b$, $i = 1, 2$, so we can always write $\mathbf{V}_i = \mathcal{X}(A_i) - \mathcal{X}(A) + \mathbf{e}_3\beta_i$ for some $\beta_i \in \mathbb{R}$. Define $\mathbf{V}'_i = \mathcal{X}(A_i) - \mathcal{X}(A)$, $i = 1, 2$. (See Figure 8.) Then $\mathbf{V}_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}_i) = \mathbf{V}'_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}'_i)$, $i = 1, 2$. Thus we have that

$$\mathbf{e}_{1,i} = \frac{\mathbf{V}'_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}'_i)}{\|\mathbf{V}'_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}'_i)\|}, \quad i = 1, 2.$$

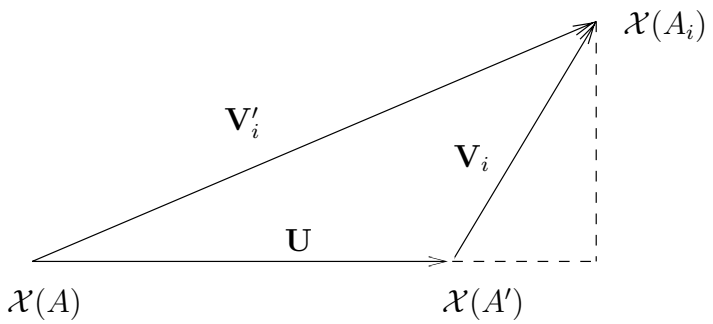

FIGURE 8. An illustration of the fact that $\mathbf{V}_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}_i) = \mathbf{V}'_i - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V}'_i)$ in the case where $\alpha(a_i) = A'$. Here $\mathbf{e}_3 = \mathbf{U}/\|\mathbf{U}\|$. When $\alpha(a_i) = A$ then the above equality is obvious since in that case $\mathbf{V}_i = \mathbf{V}'_i$.

$\mathbf{e}_{1,2}$ is obtained from $\mathbf{e}_{1,1}$ by rotating about $\mathbf{e}_3$ through an angle of φ, where $z = e^{i\varphi} = \cos(\varphi) + i\sin(\varphi) = x + iy$ (see Figure 7). Thus $x = \mathbf{e}_{1,1} \cdot \mathbf{e}_{1,2}$, and $y = \mathbf{e}_{1,1} \times \mathbf{e}_{1,2} \cdot \mathbf{e}_3$. Thus we have that $\pi_2[(\phi \circ \gamma)(\mathcal{X})](e) = T_2(z)$. Thus the explicit expression for the *wedge angle coordinate* is

$$z(\tilde{e}) = \mathbf{e}_{1,1} \cdot \mathbf{e}_{1,2} + i\mathbf{e}_{1,1} \times \mathbf{e}_{1,2} \cdot \mathbf{e}_3,$$

where

$$\mathbf{e}_3 = \frac{\mathcal{X}(A') - \mathcal{X}(A)}{\|\mathcal{X}(A') - \mathcal{X}(A)\|},$$
$$\mathbf{e}_{1,i} = \frac{\mathcal{X}(A_i) - \mathcal{X}(A) - \mathbf{e}_3[\mathbf{e}_3 \cdot (\mathcal{X}(A_i) - \mathcal{X}(A))]}{\|\mathcal{X}(A_i) - \mathcal{X}(A) - \mathbf{e}_3[\mathbf{e}_3 \cdot (\mathcal{X}(A_i) - \mathcal{X}(A))]\|}, \quad i = 1, 2.$$

Wedge angle coordinates are frequently defined from an ordered quadruple of atoms (see page 27-29 of [34] or page 103 of [22]), and the sign convention along the axis of rotation might differ slightly from our convention; thus care should be exercised. We employ the same convention for dihedrals as well as impropers. The wedge angle coordinate φ associated to a dihedral edge is called a *torsion angle*. We will use the term *dihedral angle* to be synonymous with torsion angle, even though some authors make a distinction between them (see [35]). The sign of an improper wedge angle conveys important information about chirality, which cannot be recovered from knowledge of only the distances between atoms (see [12]). Our scheme includes this information because Γ (actually Δ) is an oriented graph.

These are the usual expressions for chemical internal coordinates. However, here they emerge naturally from our general scheme. We use the complex variable z rather than the angle φ in order to avoid choosing a branch, i.e. $\varphi \in (-\pi, \pi]$ or $\varphi \in [0, 2\pi)$, etc.. This is useful if we want to examine the global structure of the manifold of internal coordinates, as we have done. Another advantage of c and z over the corresponding angles is that there is no need to involve transcendental functions, such as inverse cosine or inverse sine. However, for purely descriptive purposes we will use the angles as coordinates.

4. The Main Theorem

4.1. **Z-trees, GZ-trees, Statement.** Chemists today typically specify the geometry of a molecule using a Z-matrix, which is a matrix-like structure suitable for data entry into computer programs. Although various types of internal coordinate systems had long been used for the study of the vibrations of small molecules [45], a general systematic and well-behaved method of setting up such systems seems to have originated with J. A. Pople. In work by his graduate student M. A. Gordon [19] a computer program for automatically generating molecular geometries given only the name of the chemical compound was written which used Pople's idea of Z-matrix style internal coordinates. These internal coordinates were seen merely as a method of generating Cartesian coordinates of the atomic positions from "chemical" information arising from generalized rules of molecular structure. Gordon's program employed many matrices and the 26th such matrix contained the internal coordinates of the molecule, hence the name "Z-matrix". Several years later Z-matrix style internal coordinates were encorporated into the data entry system of the Gaussian 70 program [22] for quantum chemistry calculations. The good features of Z-matrices were widely recognized and they were adopted by computational chemists, although a couple of syntactical variants of the Gaussian Z-matrix also came into use (such as the MOPAC style Z-matrix, [11]). The basic idea of Z-matrix internal coordinates is to specify the position of every atom after the third using its spherical coordinates relative to a well-conformed Cartesian coordinate system based at one of the atoms already specified. The author was not able to locate in the literature a proof that Z-matrix internal coordinates are well-behaved on all molecular configurations not involving collinear angles, even though this appears to be a well-known fact to chemists [10]. Perhaps the difficulty of formulating this statement precisely has discouraged such efforts, although an elementary proof based on spherical coordinates could probably be given without using our formalism.

Rather than defining Z-matrices we intend to define the corresponding internal coordinate systems in the context of our theory. Suppose $\mathcal{G}$ is a molecular graph with $N \geq 3$ atoms. Suppose Γ is a subgraph of $AL^2(\mathcal{G})$. We say Γ is a *Z-tree* if there is an increasing sequence $(\Gamma_1, \ldots, \Gamma_{N-2})$ of subgraphs of Γ satisfying the following conditions:

(1) Γ_1 is a linear graph (called the *trunk*) of three edges and four vertices based on a single angle $a_1 \in \mathrm{vert}L^2(\mathcal{G})$. So there exists $a_1 = \{b_1, b_2\} \in \mathrm{vert}L^2(\mathcal{G})$, where $b_1 = \{A_1, A_2\}$ and $b_2 = \{A_2, A_3\}$, such that the vertices of Γ_1 are

$$(a_1, b_1, A_1)\text{———}(a_1, b_1, A_2)\text{----}(a_1, b_2, A_2)\text{———}(a_1, b_2, A_3),$$

and each pair of consecutive vertices in the above is an edge of Γ_1.

(2) For each $2 \leq j \leq N-2$, Γ_j is obtained by attaching a linear chain (called a *branch*) of three edges and three vertices arising from a single new atom to a single vertex of Γ_{j-1}. So there exists $(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j) \in \mathrm{vert}\Gamma_{j-1}$ (called the *vertex of attachment*), and $A_{j+2} \in \mathrm{vert}\mathcal{G} \setminus \pi_3(\mathrm{vert}\Gamma_{j-1})$ (called the *new atom*) such that $b_{j+1} = \{\tilde{A}_j, A_{j+2}\} \in \mathrm{edge}\mathcal{G}$, such that if $a_j = \{\tilde{b}_j, b_{j+1}\}$, then $\mathrm{edge}\Gamma_j \setminus \mathrm{edge}\Gamma_{j-1}$ contains the three edges formed from pairs of consecutive vertices from the following:

$$(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j)\cdot\cdots\cdot(a_j, \tilde{b}_j, \tilde{A}_j)\text{----}(a_j, b_{j+1}, \tilde{A}_j)\text{———}(a_j, b_{j+1}, A_{j+2}).$$

Also, $\mathrm{vert}\Gamma_j \setminus \mathrm{vert}\Gamma_{j-1}$ contains the last three vertices in the above.

(3) $\Gamma_{N-2} = \Gamma$.

Via the map $\phi \circ \gamma$, a rooted Z-tree (Γ, r) defines a system of internal coordinates on X_Γ. This system could be called an *internal coordinate system of Z-matrix type*. When we say that this system of internal coordinates is *well-behaved* we mean that $\phi \circ \gamma$ defines a left G_a-equivariant diffeomorphism between X_Γ and $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$. Thus in a well-behaved internal coordinate system we can characterize exactly and easily which molecular configurations are well-represented, namely all those in X_Γ, and which internal coordinate values represent them, namely those in $G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$, or equivalently those in $(0,\infty)^{\mathrm{edge}_0\Gamma} \times (-1,1)^{\mathrm{edge}_1\Gamma} \times (S^1)^{\mathrm{edge}_2\Gamma}$.

We will prove that internal coordinate systems of Z-matrix type are well-behaved, as a corollary of our main result. Thus our purpose is to generalize that result about Z-trees to the widest possible class of subgraphs of $AL^2(\mathcal{G})$. This is not an idle generalization since subgraphs other than Z-trees arise naturally in examples (see section 6) and especially in the study of polymerization, i.e. the linking of two separate molecules to form a larger molecule. We call a tree subgraph Γ of $AL^2(\mathcal{G})$ a *generalized Z-tree*, or a *GZ-tree*, if the rooted tree (Γ, r) (for some choice of a root vertex) gives rise via the map $\phi \circ \gamma$ to a homeomorphism between X_Γ and $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$. Our main result will give graph theoretical conditions on Γ which are necessary and sufficient for it to be a GZ-tree.

Main Theorem. *Suppse (Γ, r) is a rooted tree subgraph of $AL^2(\mathcal{G})$, where $\mathcal{G}$ is a graph with $N \geq 3$ vertices. ($AL^2(\mathcal{G})$ is the amalgamated twice iterated line graph of $\mathcal{G}$; see section 2.2.) Let X, $\mathcal{S}$, G_p, G_a be three dimensional space, the set of all Cartesian coordinates systems on X, the group of all transformations of Cartesian coordinate systems, and the group of all proper affine symmetries of X, respectively (see section 2.1). Let $X^{\mathrm{vert}\mathcal{G}}$ be the set of all molecular configurations of the molecule whose graph is $\mathcal{G}$, and $X_\Gamma \subset X^{\mathrm{vert}\mathcal{G}}$ the dense open subset of non-degenerate configurations with respect to Γ (see section 3.1). Let $\gamma\colon X_\Gamma \to \mathcal{S}^{\mathrm{vert}\Gamma}$ be the left G_a-equivariant map which extends each nondegenerate molecular configuration $\mathcal{X} \in X_\Gamma$ into a conformed pose assignment $\mathcal{E} = \gamma(\mathcal{X})$ (see section 3.1). Let $\phi\colon \mathcal{S}^{\mathrm{vert}\Gamma} \to \mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$ be the left G_a-equivariant diffeomorphism constructed in section 2.3. Let G_0, G_1, G_2 be the one dimensional real submanifolds of G_p defined in section 3.2, diffeomorphic to $(0,\infty), (-1,1), S^1$ respectively. Let $\mathrm{edge}\Gamma = \mathrm{edge}_0\Gamma \cup \mathrm{edge}_1\Gamma \cup \mathrm{edge}_2\Gamma$ be the decomposition discussed in section 2.2.*

Then X_Γ is homeomorphic to

$$\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$$

via the map $\phi \circ \gamma$ if and only if Γ satisfies the following conditions.

1. $\pi_3(\mathrm{vert}\Gamma) = \mathrm{vert}\mathcal{G}$, *i.e. for every atom $A \in \mathrm{vert}\mathcal{G}$ there exists a vertex $(a, b, A) \in \mathrm{vert}\Gamma$ based at that atom.*
2. *For every $A \in \mathrm{vert}\mathcal{G}$ the subgraph Γ_A of Γ induced by $\mathrm{vert}\Gamma_A = \{(a, b, A') \in \mathrm{vert}\Gamma \mid A' = A\}$ is connected.*
3. *For every $(a, b, A) \in \mathrm{vert}\Gamma$, where $A \neq \alpha_1(a)$, there exists $\{(a, b, \alpha_1(a)), (a, b, A)\} \in \mathrm{edge}_0\Gamma$.*
4. *For every $a \in \pi_1\mathrm{vert}\Gamma$, where $a = \{b_1, b_2\}$, there exists $\{(a, b_1, \alpha_1(a)), (a, b_2, \alpha_1(a))\} \in \mathrm{edge}_1\Gamma$.*

(5) *For every bond* $b \in \pi_2(\text{vert}\Gamma)$*, where* $b = \{A_1, A_2\}$ *there exists* $a \in \pi_1(\text{vert}\Gamma)$ *such that* $\{(a, b, A_1), (a, b, A_2)\} \in \text{edge}_0\Gamma$.
(6) *For every vertex* $(a, b, A) \in \text{vert}\Gamma$ *the subgraph* $\Gamma_{b,A}$ *of* Γ *induced by* $\text{vert}\Gamma_{b,A} = \{(a', b', A') \in \text{vert}\Gamma \mid b' = b, A' = A\}$ *is connected.*

Furthermore when these conditions hold $\phi \circ \gamma$ *defines a left* G_a*-equivariant diffeomorphism onto its image.*

The next two sections will explain the proof of this theorem. However, in the remainder of this section we will prove the following.

Theorem. *Suppose* Γ *is a Z-tree subgraph of* $AL^2(\mathcal{G})$*. Then* Γ *is a tree and it satisfies conditions (1)-(6) stated in the above theorem. Consequently, every internal coordinate system of Z-matrix type (i.e. associated to a Z-tree* Γ*) is well-behaved.*

Proof. Suppose $(\Gamma_1, \ldots, \Gamma_{N-2})$ is an increasing sequence of subgraphs of Γ with the three properties which make Γ into a Z-tree. As in the definition of a Z-tree let $(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j)$, $j = 2, \ldots, N-2$, be the vertices of attachment, and A_j, $j = 1, \ldots, N$, be the atoms, b_j, $j = 1, \ldots, N-1$ the bonds, and a_j, $j = 1, \ldots, N-2$ the angles. By Theorem 2.1.3, page 52 of [44], in order to show that Γ is a tree, it is sufficient to show that it is connected with the number of edges being one fewer than the number of vertices. Clearly these assertions hold for Γ_1. Suppose they hold for some Γ_j, where $1 \leq j < N-2$. Γ_{j+1} is obtained by attaching three new vertices and three new edges to Γ_j in such a manner that the result is still connected. By induction the assertions hold for $\Gamma_{N-2} = \Gamma$. Therefore Γ is a tree.

Clearly a Z-tree contains vertices based at all N atoms $A_1, \ldots, A_N$ of $\mathcal{G}$. Thus condition (1) is true.

Let $A \in \text{vert}\mathcal{G}$, and let $\text{vert}\Gamma_A = \{(a, b, A) \in \text{vert}\Gamma \mid b \in \text{vert}L^1(\mathcal{G}), a \in \text{vert}L^2(\mathcal{G})\}$. To verify condition (2) we must show that the subgraph Γ_A of Γ induced by $\text{vert}\Gamma_A$ is connected. Let $1 \leq J \leq N-2$ be as small as possible such that $\text{vert}\Gamma_A \cap \text{vert}\Gamma_J \neq \emptyset$. If $J = 1$ then $\text{vert}\Gamma_A \cap \text{vert}\Gamma_1$ consists either of a single vertex (and no edge) or a pair of vertices and a single edge connecting them. Either way the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_1$ is connected. If $J > 1$ then $\text{vert}\Gamma_A \cap \text{vert}\Gamma_J$ consists of a single vertex (and no edge). Hence the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_J$ is connected. Now suppose for $J < j \leq N-2$ the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_{j-1}$ is connected. Assume $\tilde{A}_j = A$, since otherwise $\text{vert}\Gamma_A \cap \text{vert}\Gamma_{j-1} = \text{vert}\Gamma_A \cap \text{vert}\Gamma_j$, and the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_j$ is automatically connected. Since $\Gamma_j \setminus \Gamma_{j-1}$ consists of three vertices and three edges, and the third vertex is associated to a new atom not in $\pi_3(\text{vert}\Gamma_{j-1})$, we see that the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_j$ is obtained from the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_{j-1}$ by adding two vertices and two edges, all associated with the atom A. Thus the subgraph of Γ_A induced by $\text{vert}\Gamma_A \cap \text{vert}\Gamma_j$ is connected. By induction therefore we have shown that Γ_A is connected. Hence condition (2) is true.

To verify conditions (3) and (4) note that $\pi_1(\text{vert}\Gamma) = \{a_1, \ldots, a_{N-2}\}$. Corresponding to a_1 we see that Γ contains the required two edges of type 0 and the required edge of type 1. Also for $2 \leq j \leq N-2$ the graph Γ contains the edges $\{(a_j, b_{j+1}, \tilde{A}_j), (a_j, b_{j+1}, A_{j+2})\}$ of type 0, and $\{(a_j, \tilde{b}_j, \tilde{A}_j), (a_j, b_{j+1}, \tilde{A}_j)\}$ of type 1. If $\tilde{b}_j = \{\tilde{A}_j, A'\}$, then the vertex $(a_j, \tilde{b}_j, A')$ is not in Γ, since Γ contains only three vertices involving the angle a_j. Thus conditions (3) and (4) are both true.

To verify condition (5) note that $\pi_2(\text{vert}\Gamma) = \{b_1, \ldots, b_{N-1}\}$. Corresponding to b_1 and b_2 the graph Γ contains the edges $\{(a_1, b_1, A_1), (a_1, b_1, A_2)\}$ and $\{(a_1, b_2, A_2), (a_1, b_2, A_3)\}$ of type 0. For $2 \leq j \leq N-2$ and corresponding to b_{j+1} the graph Γ contains the edge $\{(a_j, b_{j+1}, \tilde{A}_j), (a_j, b_{j+1}, A_{j+2})\}$ of type 0. Hence condition (5) is true.

Finally suppose $(a, b, A) \in \text{vert}\Gamma$ and let $\Gamma_{b,A}$ be the subgraph of Γ induced by the subset $\text{vert}\Gamma_{b,A} = \{(a', b', A') \in \text{vert}\Gamma \mid b' = b, A' = A\}$. To verify condition (6) we must show that $\Gamma_{b,A}$ is connected. Let $1 \leq J \leq N-2$ be as small as possible such that $\text{vert}\Gamma_{b,A} \cap \text{vert}\Gamma_J \neq \emptyset$. If $J = 1$ then this intersection consists of a single vertex. If $J > 1$ then this intersection also consists of a single vertex, either $(a_J, b_{J+1}, \tilde{A}_J)$ or (a_J, b_{J+1}, A_{J+2}). Either way the subgraph of $\Gamma_{b,A}$ induced by this intersection is connected. If $J = N-2$ then we are done, so suppose $J < N-2$. Suppose for $J < j \leq N-2$ the subgraph of $\Gamma_{b,A}$ induced by the intersection $\text{vert}\Gamma_{b,A} \cap \text{vert}\Gamma_{j-1}$ is connected. If $\text{vert}\Gamma_{b,A} \cap (\text{vert}\Gamma_j \setminus \text{vert}\Gamma_{j-1}) = \emptyset$ then the subgraph of $\Gamma_{b,A}$ induced by the intersection $\text{vert}\Gamma_{b,A} \cap \text{vert}\Gamma_j$ is connected. If $\text{vert}\Gamma_{b,A} \cap (\text{vert}\Gamma_j \setminus \text{vert}\Gamma_{j-1}) \neq \emptyset$ then note that $b \in \{b_1, \ldots, b_{J+1}\}$ and $A \in \{A_1, \ldots, A_{J+2}\}$. Therefore we know that $\text{vert}\Gamma_{b,A} \cap (\text{vert}\Gamma_j \setminus \text{vert}\Gamma_{j-1})$ can contain neither $(a_j, b_{j+1}, \tilde{A}_j)$ nor (a_j, b_{j+1}, A_{j+2}). Thus it must contain $(a_j, \tilde{b}_j, \tilde{A}_j)$, and we must have $b = \tilde{b}_j$ and $A = \tilde{A}_j$. Therefore $(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j) \in \text{vert}\Gamma_{b,A} \cap \text{vert}\Gamma_{j-1}$. Thus the subgraph of $\Gamma_{b,A}$ induced by the intersection $\text{vert}\Gamma_{b,A} \cap \text{vert}\Gamma_j$ is connected, since the new vertex $(a_j, \tilde{b}_j, \tilde{A}_j)$ will be connected to $(\tilde{a}_j, \tilde{b}_j, \tilde{A}_j)$ by the edge of type 2 connecting these two vertices in Γ. Hence by induction we have that $\Gamma_{b,A}$ is connected, and hence condition (6) is true. □

4.2. **Proof of Necessity.**

Proof. Here we begin the proof of our main theorem, which states necessary and sufficient conditions on a tree subgraph Γ of $AL^2(\mathcal{G})$ that the map $\phi \circ \gamma$ determine a homeomorphism between X_Γ and $\mathcal{S} \times G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$. In this section we will show that the conditions (1)-(6) on Γ in the statement of the theorem are necessary.

To prove the necessity of (1) suppose $\phi \circ \gamma$ induces a homeomorphism between X_Γ and $\mathcal{S} \times G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$ and yet there exists $A \in \text{vert}\mathcal{G} \setminus \pi_3(\text{vert}\Gamma)$. Let $S_A = \{(a, b, A') \in \text{vert}\Gamma \mid \text{there exists } b' \in a \text{ such that } A \in b'\}$. Let $\mathcal{X}_0 \in X_\Gamma$ be given. Let $\mathcal{E}_0 = \gamma(\mathcal{X}_0)$. For each $(a, b, A') \in S_A$ let $\text{dir}(a, b, A')$ be the set of all $\mathbf{v} \in V$ such that $\|\mathbf{v}\| = 1$, and $\mathcal{X}_0(A'') - \mathcal{X}_0(A') = \mathbf{v}x$ for some $x \in \mathbb{R} \setminus \{0\}$, where A, A', A'' are the three distinct atoms which are elements of bonds in a. Now choose $\mathbf{v} \in V$ such that $\|\mathbf{v}\| = 1$ and $\mathbf{v} \notin \text{dir}(a, b, A')$ for all $(a, b, A') \in S_A$. Define for each integer $n \geq 1$ and for each $A' \in \text{vert}\mathcal{G}$:

$$\mathcal{X}_n(A') = \begin{cases} \mathcal{X}_0(A') & A' \neq A, \\ \mathcal{X}_0(A) + \mathbf{v}n & A' = A. \end{cases}$$

We claim that $\mathcal{X}_n \in X_\Gamma$ for all sufficiently large n and $\gamma(\mathcal{X}_n)$ tends to a limit in $\mathcal{S}^{\text{vert}\Gamma}$ as $n \to \infty$. The nondegeneracy conditions for each vertex in $\text{vert}\Gamma \setminus S_A$ hold for each $n \geq 0$. Furthermore the pose assigned by $\gamma(\mathcal{X}_n)$ to each vertex of this kind is independent of $n \geq 0$. There are four different types of vertices in S_A, and the verification of the claim for each of them is slightly different. We will spare the reader the detailed calculations and merely quote the final results, which should be more or less clear on geometrical grounds. Suppose A, A', A'' are the three distinct

atoms which are elements of bonds in a, where $(a, b, A') \in S_A$. The limiting pose will be denoted by $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. In all four cases $\mathbf{e}_0 = \mathcal{X}_0(A')$ (it never changed as n varied.) Since we always have $\mathbf{e}_2 = \mathbf{e}_3 \times \mathbf{e}_1$ we will only specify $\mathbf{e}_1, \mathbf{e}_3$. The four cases are as follows.

(1) $b = \{A, A'\}$ and $\alpha_1(a) = A$. $\mathbf{e}_3 = \mathbf{v}$, and $\mathbf{e}_1$ is the unit vector in the direction of the component of $\mathcal{X}_0(A'') - \mathcal{X}_0(A')$ perpendicular to $\mathbf{v}$, i.e.

$$\mathbf{e}_1 = \frac{\mathcal{X}_0(A'') - \mathcal{X}_0(A') - \mathbf{v}[\mathbf{v} \cdot (\mathcal{X}_0(A'') - \mathcal{X}_0(A'))]}{\|\mathcal{X}_0(A'') - \mathcal{X}_0(A') - \mathbf{v}[\mathbf{v} \cdot (\mathcal{X}_0(A'') - \mathcal{X}_0(A'))]\|}.$$

(2) $b = \{A, A'\}$ and $\alpha_1(a) = A'$. $\mathbf{e}_3$ and $\mathbf{e}_1$ are the same as in case (1).

(3) $b = \{A', A''\}$ and $\alpha_1(a) = A'$. $\mathbf{e}_3 = [\mathcal{X}_0(A'') - \mathcal{X}_0(A')]/\|\mathcal{X}_0(A'') - \mathcal{X}_0(A')\|$, and $\mathbf{e}_1$ is the unit vector in the direction of the component of $\mathbf{v}$ perpendicular to $\mathcal{X}_0(A'') - \mathcal{X}_0(A')$, i.e.

$$\mathbf{e}_1 = \frac{\mathbf{v} - \mathbf{e}_3[\mathbf{e}_3 \cdot \mathbf{v}]}{\|\mathbf{v} - \mathbf{e}_3[\mathbf{e}_3 \cdot \mathbf{v}]\|}.$$

(4) $b = \{A', A''\}$ and $\alpha_1(a) = A''$. $\mathbf{e}_3$ and $\mathbf{e}_1$ are the same as in case (3).

The fact that $\mathcal{X}_n \in X_\Gamma$ for all sufficiently large n, and the existence of these limiting poses depends on the fact that $\mathbf{v} \notin \mathrm{dir}(a, b, A')$. Hence our claim is proved. To finish our proof of the necessity of condition (1) we show that our assumption that (1) is violated leads to a contradiction. Since $\mathcal{X}_n \in X_\Gamma$ for all sufficiently large n, we have that $(\phi \circ \gamma)(\mathcal{X}_n) \in \mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$ for all sufficiently large n. Since $\gamma(\mathcal{X}_n)$ converges in $\mathcal{S}^{\mathrm{vert}\Gamma}$ as $n \to \infty$ we see that $(\phi \circ \gamma)(\mathcal{X}_n)$ converges in $\mathcal{S} \times G_p^{\mathrm{edge}\Gamma}$. We claim that this limit is in $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$. The pose of the root vertex converges to something in $\mathcal{S}$. None of the bond lengths change at all, and G_2 is closed, so the only possible problem is that some of the bond angle cosines might approach 1 or -1 (recall G_1 is diffeomorphic to $(-1, 1)$). Bond angle cosines for edges of type 1 between vertices not in S_A are independent of n and are in $(-1, 1)$. Edges of type 1 involving at least one vertex from S_A can only be an edge connecting a vertex in case (2) to a vertex in case (3), in fact involving the same angle a. In case (2) $\mathbf{e}_3 = \mathbf{v}$ and in case (3) $\mathbf{e}_3$ is parallel to $\mathcal{X}_0(A'') - \mathcal{X}_0(A')$, and we have required that the angle between them is not 0 or π. Thus the limit of $(\phi \circ \gamma)(\mathcal{X}_n)$ is in $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$. However if $\phi \circ \gamma$ defines a homeomorphism between X_Γ and $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$, then it must be true that $\mathcal{X}_n$ converges to something in X_Γ. However it clearly diverges. This contradiction shows that condition (1) is necessary.

Condition (2) requires that all the poses based at atom A associated with the coordinate system corresponding to Γ be connected through a sequence of coordinate transformations of types 1 or 2. It is not difficult to see why this condition is necessary if $\phi \circ \gamma$ is to define a bijection between X_Γ and $\mathcal{S} \times G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$. Suppose for some $A \in \mathrm{vert}\mathcal{G}$ the subgraph Γ_A is not connected, i.e. there exist $v_0 = (a_1, b_1, A), v_n = (a_2, b_2, A) \in \mathrm{vert}\Gamma_A$ which are distinct, and such that there is no path in Γ_A connecting v_0 and v_n. Since Γ is a tree there is a unique path (of length n) in Γ connecting these two vertices, but some vertex $v_k = (a', b', A')$ on this path must fail to lie in Γ_A. Thus $A' \neq A$. Consider the path in Γ starting at v_0 and ending at v_n; we may choose v_k, $1 < k < n$, to be the last vertex on this path which does not lie in $\mathrm{vert}\Gamma_A$. The next vertex v_{k+1} on this path after v_k must be in Γ_A, hence $v_{k+1} = (a', b', A)$ and these two vertices must be connected

by the edge $e = \{(a', b', A'), (a', b', A)\} \in \text{edge}\Gamma$ of type 0. Now suppose $\mathcal{X} \in X_\Gamma$, $\mathcal{E} = \gamma(\mathcal{X})$ and $\phi(\mathcal{E}) = (E, \mathcal{A})$. For each $1 \leq j \leq n$ define

$$\omega_j = \begin{cases} 1 & \text{if } v_{j-1} \text{ is the parent and } v_j \text{ is the child,} \\ -1 & \text{if } v_j \text{ is the parent and } v_{j-1} \text{ is the child,} \end{cases}$$

where the orientation of the edge $\{v_{j-1}, v_j\}$ is defined by the root of Γ. We know that $\mathcal{X}(A) = \mathcal{E}(v_0)\mathbf{u}_1 = \mathcal{E}(v_n)\mathbf{u}_1$. From the condition defining $\mathcal{A}$ we see that

$$\mathcal{E}(v_n) = \mathcal{E}(v_0)\mathcal{A}(\{v_0, v_1\})^{\omega_1}\mathcal{A}(\{v_1, v_2\})^{\omega_2} \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}.$$

Thus

$$\mathcal{E}(v_n)\mathbf{u}_1 = \mathcal{E}(v_0)\mathcal{A}(\{v_0, v_1\})^{\omega_1}\mathcal{A}(\{v_1, v_2\})^{\omega_2} \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 = \mathcal{E}(v_0)\mathbf{u}_1,$$

and hence

$$\mathbf{u}_1 = \mathcal{A}(\{v_0, v_1\})^{\omega_1}\mathcal{A}(\{v_1, v_2\})^{\omega_2} \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1.$$

We call this the *loop equation*. The presence of this constraint suggests that the mapping $\phi \circ \gamma$ is not onto $\mathcal{S} \times G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$, which of course contradicts the assumption that it is a bijection. To demonstrate this we need to find another $\mathcal{A}' \in G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$ which does not satisfy the loop equation, and hence is not in $\mathcal{S} \times G_\Gamma = (\phi \circ \gamma)(X_\Gamma)$. Suppose $\mathcal{A}(\{v_k, v_{k+1}\}) \in G_0$ corresponds under the map $T_0^{-1} \colon G_0 \to (0, \infty)$ to a bond length $l > 0$. For all $e \in \text{edge}\Gamma$ define

$$\mathcal{A}'(e') = \begin{cases} \mathcal{A}(e') & e' \neq e = \{v_k, v_{k+1}\}, \\ T_0(l+1) & e' = e. \end{cases}$$

All the edges e' on the path $(v_0, v_1, \ldots, v_n)$ after $e = \{v_k, v_{k+1}\}$ must be of type 1 or 2, hence $\mathcal{A}'(e')\mathbf{u}_1 = \mathcal{A}(e')\mathbf{u}_1 = \mathbf{u}_1$. Also $T_0(l+1)\mathbf{u}_1 = T_0(l)\mathbf{u}_1 + \mathbf{u}_4$. Therefore

$$\begin{aligned} \mathcal{A}'(\{v_k, v_{k+1}\})^{\omega_{k+1}} \cdot \cdots \cdot \mathcal{A}'(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 &= T_0(l+1)\mathbf{u}_1 = T_0(l)\mathbf{u}_1 + \mathbf{u}_4 \\ &= \mathcal{A}(\{v_k, v_{k+1}\})^{\omega_{k+1}} \cdot \cdots \cdot \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 + \mathbf{u}_4. \end{aligned}$$

Consequently

$$\begin{aligned} &\mathcal{A}'(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}'(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 \\ &\quad = \mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 + \mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{k-1}, v_k\})^{\omega_k}\mathbf{u}_4 \\ &\quad = \mathbf{u}_1 + \mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{k-1}, v_k\})^{\omega_k}\mathbf{u}_4 \end{aligned}$$

$\mathcal{A}'$ will satisfy the loop equation only if this last vector vanishes; however it cannot since the matrices $\mathcal{A}(e')$ are all invertible. Thus condition (2) is necessary.

To see that (3) is necessary for $G_\Gamma = G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$, suppose (3) is not true, i.e. there exists $(a, b_1, A_1) \in \text{vert}\Gamma$ such that $a = \{b_1, b_2\}$, $A_1 \neq \alpha_1(a)$, and the edge connecting $(a, b_1, \alpha_1(a))$ and (a, b_1, A_1) is not in $\text{edge}_0\Gamma$. Suppose $b_2 = \{A_2, \alpha_1(a)\}$. The vertex (a, b_1, A_1) is connected to the rest of Γ *only* through edges of type 2. By varying the wedge angle coordinate associated to one of these edges we can perform a hinge-bending deformation of the molecule about the bond b_1, which will rotate the $\mathbf{e}_1$ axis of the pose at (a, b_1, A_1) out of the plane determined by the angle a. Since Γ is a tree, there exist a unique path $(v_0, v_1, \ldots, v_n)$ in Γ, where $v_0 = (a, b_1, A_1)$ and $\pi_3(v_n) = A_2$. We can arrange this because by condition (1), which we have already shown to be necessary, there is a vertex v_n of Γ such

that $\pi_3(v_n) = A_2$. Define ω_j to be 1 if v_{j-1} is the parent and v_j is the child, and -1 if the reverse is true. As usual we have

$$\mathcal{X}(A_1) = \mathcal{E}(v_0)\mathbf{u}_1$$
$$\mathcal{X}(A_2) = \mathcal{E}(v_n)\mathbf{u}_1 = \mathcal{E}(v_0)\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1$$

In the above we have $\mathcal{X} \in X_\Gamma$, $\mathcal{E} = \gamma(\mathcal{X})$ and $\phi(\mathcal{E}) = (E, \mathcal{A})$. The constraint $\mathcal{A} \in G_\Gamma$ implies that the following equations hold.

$$\mathbf{u}_2^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] > 0$$
$$\mathbf{u}_3^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] = 0.$$

As we noted earlier, the edge $\{v_0, v_1\}$ must be of type 2, so $\mathcal{A}(\{v_0, v_1\})^{\omega_1} = T_2(e^{i\varphi})$. For $\mathcal{A} \in G_\Gamma$ the wedge angle φ assumes a definite value φ_0 for which the two above relations hold. We will show that even an infinitesimal variation of φ from the value φ_0 will lead to a violation of the second equation, and hence to an element of $G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$ which is not in G_Γ. $\mathbf{u}_3^T T_2(e^{i\varphi}) = (0, \sin\varphi, \cos\varphi, 0)$ and $\frac{d}{d\varphi}\mathbf{u}_3^T T_2(e^{i\varphi}) = (0, \cos\varphi, -\sin\varphi, 0) = \mathbf{u}_2^T T_2(e^{i\varphi})$. Therefore for all φ sufficiently near φ_0 we have

$$\frac{d}{d\varphi}\mathbf{u}_3^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1]$$
$$= \mathbf{u}_2^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] > 0.$$

Thus condition (3) is necessary.

To see that (4) is necessary, suppose (4) is not true, i.e. there exists $a \in \pi_1\mathrm{vert}\Gamma$ such that $a = \{b_1, b_2\}$ and the edge connecting $(a, b_1, \alpha_1(a))$ and $(a, b_2, \alpha_1(a))$ is not in $\mathrm{edge}_1\Gamma$. As before let $b_1 = \{\alpha_1(a), A_1\}$ and $b_2 = \{\alpha_1(a), A_2\}$. Since we now know condition (3) is necessary we may assume without loss of generality that $(a, b_1, \alpha_1(a)) \in \mathrm{vert}\Gamma$. There are two cases to be considered.

(1) $\{(a, b_1, \alpha_1(a)), (a, b_1, A_1)\} \notin \mathrm{edge}_0\Gamma$.
(2) $\{(a, b_1, \alpha_1(a)), (a, b_1, A_1)\} \in \mathrm{edge}_0\Gamma$.

The difficulty with both of these alternatives is the same, and in fact is identical to the problem with the negation of condition (3). We use the same argument as above with the following modifications. In case (1) the path must start at $v_0 = (a, b_1, \alpha_1(a))$, and as before the first edge $\{v_0, v_1\}$ must be of type 2. Everything else in the argument works exactly as before. Finally in case (2) there are two types of edges emanating from each of the vertices $(a, b_1, \alpha_1(a))$ and (a, b_1, A_1), of types 0 and 2. If we connect both of them to the same vertex in Γ_{A_2} via paths in Γ and one of those paths starts with an edge of type 0 (it must be the edge $\{(a, b_1, \alpha_1(a), (a, b_1, A_1)\}$) then the other path must start with an edge of type 2 (since paths cannot double back on themselves). Thus at least one of the paths starts with an edge of type 2 and the argument can be completed as in case (1). Thus condition (4) is necessary.

To prove that condition (5) is necessary if G_Γ is to be all of $G_0^{\mathrm{edge}_0\Gamma} \times G_1^{\mathrm{edge}_1\Gamma} \times G_2^{\mathrm{edge}_2\Gamma}$ we assume that condition (5) is false, i.e. there is a bond $b \in \pi_2(\mathrm{vert}\Gamma)$, where $b = \{A_1, A_2\}$, such that for all $a \in \pi_1(\mathrm{vert}\Gamma)$ we have $\{(a, b, A_1), (a, b, A_2)\} \notin \mathrm{edge}_0\Gamma$. Since $b \in \pi_2(\mathrm{vert}\Gamma)$, there exists $a \in \pi_1(\mathrm{vert}\Gamma)$ such that $b \in a$; without loss of generality assume that $\alpha_1(a) = A_1$. By condition (4) (now known to be necessary) we have that $(a, b, A_1) \in \mathrm{vert}\Gamma$. By condition (1) there is a vertex of Γ

based at A_2 so we can find a path $(v_0, v_1, \ldots, v_n)$ in Γ such that $v_0 = (a, b, A_1)$ and $\pi_3(v_n) = A_2$. The first edge $\{v_0, v_1\}$ of this path cannot be of type 0, so it must be of type 2 or 1. If it is of type 2, then the next edge cannot be of type 0. Thus we can delete from our path any initial edges of type 2 so that the first edge is of type 1 (since we must eventually get to a different atom A_2). Suppose that this deletion process had already been performed on the original path $(v_0, v_1, \ldots, v_n)$, so that $\{v_0, v_1\}$ is of type 1. Define as usual ω_j to be 1 if v_{j-1} is the parent and v_j is the child, and -1 if the reverse is true. Let $\mathcal{X} \in X_\Gamma$, $\mathcal{E} = \gamma(\mathcal{X})$ and $\phi(\mathcal{E}) = (E, \mathcal{A})$. We have that

$$\mathcal{X}(A_1) = \mathcal{E}(v_0)\mathbf{u}_1$$
$$\mathcal{X}(A_2) = \mathcal{E}(v_n)\mathbf{u}_1 = \mathcal{E}(v_0)\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1$$

The vector $\mathcal{X}(A_2) - \mathcal{X}(A_1)$ should be parallel to $\mathbf{e}_3$ of the pose at v_0, i.e.

$$\mathbf{u}_2^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] = 0$$
$$\mathbf{u}_3^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] = 0$$
$$\mathbf{u}_4^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] > 0.$$

$\mathcal{A}(\{v_0, v_1\}) = \mathcal{A}(\{v_0, v_1\})^{\omega_1} = T_1(\cos\theta)$, where $\theta = \theta_0$ when $\mathcal{A} \in G_\Gamma$. We will show that an infinitesimal variation of θ from the value θ_0 leads to an $\mathcal{A}$ which is not in G_Γ. $\mathbf{u}_2^T T_1(\cos\theta) = (0, -\cos\theta, 0, \sin\theta)$, so $\frac{d}{d\theta}\mathbf{u}_2^T T_1(\cos\theta) = (0, \sin\theta, 0, \cos\theta) = \mathbf{u}_4^T T_1(\cos\theta)$. Therefore for all θ sufficiently near θ_0 we have

$$\frac{d}{d\theta}\mathbf{u}_2^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1]$$
$$= \mathbf{u}_4^T[\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_1 - \mathbf{u}_1] > 0.$$

Thus variation of θ causes $\mathcal{X}(A_2) - \mathcal{X}(A_1)$ to have a nonzero component in the $\mathbf{e}_1$ direction (in the pose at v_0), which means that the variant $\mathcal{A}$ cannot be in G_Γ. Thus condition (5) is necessary.

To see that condition (6) is necessary, suppose it is false, i.e. there exist vertices $(a_1, b, A), (a_2, b, A) \in \text{vert}\Gamma$ which are not connected in Γ by a path consisting entirely of edges of type 2. Because of condition (2), which is already known to be necessary, they are connected by a path $(v_0, v_1, \ldots, v_n)$ in Γ_A (consisting of edges of types 1 and 2). Hence $v_0 = (a_1, b, A)$ and $v_n = (a_2, b, A)$. By deleting initial edges of type 2 if necessary we may assume the first edge $\{v_0, v_1\}$ is of type 1. Using the same notation as has been used in the last two proofs we see that a consequence of $\mathcal{A} \in G_\Gamma$ is the following set of relations.

$$\mathbf{u}_2^T\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_4 = 0$$
$$\mathbf{u}_3^T\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_4 = 0$$
$$\mathbf{u}_4^T\mathcal{A}(\{v_0, v_1\})^{\omega_1} \cdots \mathcal{A}(\{v_{n-1}, v_n\})^{\omega_n}\mathbf{u}_4 = 1.$$

From this point on the argument is exactly like that in the proof of the necessity of condition (5). Thus condition (6) is necessary. □

4.3. **Proof of Sufficiency.**

Proof. Now we assume (Γ, r) is a rooted tree subgraph of $AL^2(\mathcal{G})$ which satisfies conditions (1)-(6) in the statement of the theorem. By the theorem of section 3.1 it follows from condition (1) that γ is a smooth embedding, and $\mathcal{S}_\Gamma = \gamma(X_\Gamma)$ is a

smooth embedded submanifold of $\mathcal{S}^{\mathrm{vert}\Gamma}$. Since ϕ is a diffeomorphism we have that $\phi(\mathcal{S}_\Gamma)$ is an embedded submanifold of $\mathcal{S}\times G_p^{\mathrm{edge}\Gamma}$. From the arguments of section 3.2 the set $\phi(\mathcal{S}_\Gamma)$ is of the form $\mathcal{S}\times G_\Gamma$, where $G_\Gamma \subset G_0^{\mathrm{edge}_0\Gamma}\times G_1^{\mathrm{edge}_1\Gamma}\times G_2^{\mathrm{edge}_2\Gamma}$. Thus it suffices to prove that $G_\Gamma = G_0^{\mathrm{edge}_0\Gamma}\times G_1^{\mathrm{edge}_1\Gamma}\times G_2^{\mathrm{edge}_2\Gamma}$, since then G_Γ will be a submanifold of $G_p^{\mathrm{edge}\Gamma}$, and thus $\phi(\mathcal{S}_\Gamma)=\mathcal{S}\times G_\Gamma$ as manifolds, and not just as sets, so that the codomain restricted map $\phi\circ\gamma$ is a left G_a-equivariant diffeomorphism.

Now suppose $E\in\mathcal{S}$ and $\mathcal{A}\in G_0^{\mathrm{edge}_0\Gamma}\times G_1^{\mathrm{edge}_1\Gamma}\times G_2^{\mathrm{edge}_2\Gamma}$. We want to show that there is a conformed pose assignment $\mathcal{E}\in\mathcal{S}_\Gamma$ such that $\phi(\mathcal{E})=(E,\mathcal{A})$. Let $\mathcal{E}=\psi(E,\mathcal{A})$, where $\psi=\phi^{-1}$ was constructed in section 2.3. We need to show that $\mathcal{E}\in\mathcal{S}_\Gamma$. Let $A\in\mathrm{vert}\mathcal{G}$, and let $\mathrm{vert}\Gamma_A=\{(a,b,A)\in\mathrm{vert}\Gamma \mid b\in\mathrm{vert}L^1(\mathcal{G}), a\in \mathrm{vert}L^2(\mathcal{G})\}$. Now we claim that the origin $(\pi_1\circ\mathcal{E})(a,b,A)$ of the pose assigned to every vertex $(a,b,A)\in\mathrm{vert}\Gamma_A$ is the same. This is true because by condition (2) all the vertices in $\mathrm{vert}\Gamma_A$ are connected by a path consisting of edges in $\mathrm{edge}_1\Gamma\cup\mathrm{edge}_2\Gamma$, and the associated coordinate transformations are in G_1 or G_2, which do not move the origin. Thus we can define $\mathcal{X}(A)$ to be the common origin of all these poses.

We must now show that $\mathcal{X}\in X_\Gamma$. Suppose $a\in\pi_1(\mathrm{vert}\Gamma)$, where $a=\{b_1,b_2\}$, $A=\alpha_1(a)$, $b_1=\{A,A_1\}$, and $b_2=\{A,A_2\}$. We must check the two nondegeneracy conditions for $\mathcal{X}$ with respect to the angle a. By condition (4) Γ contains the edge e of type 1 connecting the vertices (a,b_1,A) and (a,b_2,A). Since $\mathcal{A}(e)\in G_1$ we can write $\mathcal{A}(e)=T_1(c)$ for some $c\in(-1,1)$. By condition (5), corresponding to b_i there exists a_i such that Γ contains the edges e_i of type 0 connecting (a_i,b_i,A_i) and (a_i,b_i,A), $i=1,2$. Since $\mathcal{A}(e_i)\in G_0$ we can write $\mathcal{A}(e_i)=T_0(l_i)$ for some $l_i\in(0,\infty)$, $i=1,2$. For each $i=1,2$ we can apply condition (6) to show that either $a_i=a$ or (a,b_i,A) is connected to (a_i,b_i,A) by a path consisting of edges of type 2. If $a_i=a$ define $z_i=1$. If $a_i\neq a$ then each of the oriented edges of type 2 is assigned by $\mathcal{A}$ an element of G_2. Since $\mathcal{E}(v_c)=\mathcal{E}(v_p)\mathcal{A}((v_p,v_c))$ whenever (v_p,v_c) is an oriented edge of Γ connecting two vertices of Γ, by forming a suitable product of these elements of G_2 or their inverses we see that there exists $z_i\in S^1$ such that $\mathcal{E}(a_i,b_i,A)=\mathcal{E}(a,b_i,A)T_2(z_i)$, $i=1,2$. We also have that $\mathcal{E}(a_i,b_i,A_i)=\mathcal{E}(a_i,b_i,A)T_0(l_i)$, $i=1,2$, and $\mathcal{E}(a,b_2,A)=\mathcal{E}(a,b_1,A)T_1(c)$. We illustrate this situation with the following diagram.

$$\begin{array}{ccccc}
(a_1,b_1,A_1) & \xrightarrow{T_0(l_1)} & (a_1,b_1,A) & & (a_2,b_2,A) \xrightarrow{T_0(l_2)} (a_2,b_2,A_2)\\
 & & T_2(z_1)\vdots & & T_2(z_2)\vdots \\
 & & (a,b_1,A) & \xrightarrow{T_1(c)} & (a,b_2,A)
\end{array}$$

Since $\mathcal{X}(A)=\mathcal{E}(a_i,b_i,A)\mathbf{u}_1$, $\mathcal{X}(A_i)=\mathcal{E}(a_i,b_i,A_i)\mathbf{u}_1=\mathcal{E}(a_i,b_i,A)T_0(l_i)\mathbf{u}_1$, and $T_0(l_i)\mathbf{u}_1=\mathbf{u}_1+\mathbf{u}_4 l_i$, we have that $\mathcal{X}(A_i)=\mathcal{X}(A)+\mathcal{E}(a_i,b_i,A)\mathbf{u}_4 l_i$, and thus $\|\mathcal{X}(A_i)-\mathcal{X}(A)\|=\|\mathcal{E}(a_i,b_i,A)\mathbf{u}_4\|l_i=l_i>0$. That is the first nondegeneracy condition. But since $T_2(z_i)\mathbf{u}_4=\mathbf{u}_4$ we have that $\mathcal{E}(a_i,b_i,A)\mathbf{u}_4=\mathcal{E}(a,b_i,A)T_2(z_i)\mathbf{u}_4=\mathcal{E}(a,b_i,A)\mathbf{u}_4$. Thus $\mathcal{X}(A_i)-\mathcal{X}(A)=\mathcal{E}(a,b_i,A)\mathbf{u}_4 l_i$. From the elementary relation $T_1(c)\mathbf{u}_4=\mathbf{u}_2\sqrt{1-c^2}+\mathbf{u}_4 c$, we see that $\mathcal{E}(a,b_2,A)\mathbf{u}_4=\mathcal{E}(a,b_1,A)T_1(c)\mathbf{u}_4=\mathcal{E}(a,b_1,A)\mathbf{u}_2\sqrt{1-c^2}+\mathcal{E}(a,b_1,A)\mathbf{u}_4 c$. Hence

$$\begin{aligned}
&[\mathcal{E}(a,b_1,A)\mathbf{u}_4]\cdot[\mathcal{E}(a,b_2,A)\mathbf{u}_4]\\
&\qquad=[\mathcal{E}(a,b_1,A)\mathbf{u}_4]\cdot\{[\mathcal{E}(a,b_1,A)\mathbf{u}_2]\sqrt{1-c^2}+[\mathcal{E}(a,b_1,A)\mathbf{u}_4]c\}=c.
\end{aligned}$$

Therefore $|[\mathcal{X}(A_1) - \mathcal{X}(A)] \cdot [\mathcal{X}(A_2) - \mathcal{X}(A)]| = |c| l_1 l_2 < l_1 l_2$. This is the second nondegeneracy condition. Thus $\mathcal{X} \in X_\Gamma$.

It remains to show that $\gamma(\mathcal{X}) = \mathcal{E}$. We must show that $\gamma(\mathcal{X})(v) = \mathcal{E}(v)$ for all $v \in \text{vert}\Gamma$. If we adopt the notation of the previous argument then by symmetry and condition (3) it is sufficient to prove this for the cases $v = (a, b_1, A)$ and $v = (a_1, b_1, A_1)$ (under the assumption that $a_1 = a$). First we will verify that $\gamma(\mathcal{X})(a, b_1, A) = \mathcal{E}(a, b_1, A)$. Let $\mathcal{E}(a, b_1, A) = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. We must verify that each of these follow the prescription for a conformed pose.

Checking $\mathbf{e}_0$. By definition of $\mathcal{X}$ we have $[\gamma(\mathcal{X})(a, b_1, A)]\mathbf{u}_1 = \mathcal{X}(A) = \mathbf{e}_0$.

Checking $\mathbf{e}_3$. Define $\mathbf{U} = \mathcal{X}(A_1) - \mathcal{X}(A)$. In the above we computed $\|\mathbf{U}\| = l_1$ and $\mathbf{U} = \mathcal{X}(A_1) - \mathcal{X}(A) = \mathbf{e}_3 l_1$. Thus we see that $[\gamma(\mathcal{X})(a, b_1, A)]\mathbf{u}_4 = \mathbf{U}/\|\mathbf{U}\| = \mathbf{e}_3$.

Checking $\mathbf{e}_1$. Define $\mathbf{V} = \mathcal{X}(A_2) - \mathcal{X}(A)$. In the above we computed that $\|\mathbf{V}\| = l_2$ and $\mathbf{V} = \mathcal{X}(A_2) - \mathcal{X}(A) = \mathcal{E}(a, b_2, A)\mathbf{u}_4 l_2$. It follows from this and more of our calculations above that

$$\begin{aligned}\mathbf{V}/\|\mathbf{V}\| &= \mathcal{E}(a, b_2, A)\mathbf{u}_4 = \mathcal{E}(a, b_1, A)\mathbf{u}_2\sqrt{1 - c^2} + \mathcal{E}(a, b_1, A)\mathbf{u}_4 c \\ &= \mathbf{e}_1\sqrt{1 - c^2} + \mathbf{e}_3 c.\end{aligned}$$

Therefore $\mathbf{e}_3 \cdot \mathbf{V} = \mathbf{e}_3 \cdot \mathbf{e}_1\|\mathbf{V}\|\sqrt{1 - c^2} + \mathbf{e}_3 \cdot \mathbf{e}_3\|\mathbf{V}\|c = \|\mathbf{V}\|c$, and $\mathbf{e}_1\|\mathbf{V}\|\sqrt{1 - c^2} = \mathbf{V} - \mathbf{e}_3\|\mathbf{V}\|c = \mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})$. Consequently

$$[\gamma(\mathcal{X})(a, b_1, A)]\mathbf{u}_2 = \frac{\mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})}{\|\mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})\|} = \mathbf{e}_1.$$

Checking $\mathbf{e}_2$. Thus $\gamma(\mathcal{X})(a, b_1, A)$ and $\mathcal{E}(a, b_1, A)$ agree in their first, second, and fourth components. Since the last three components (in both poses) form a positively oriented orthonormal basis of V, we see that these two poses must agree in the third component as well. Therefore $\mathcal{E}(a, b_1, A)$ is a conformed pose.

Now assume $a_1 = a$. We want to show that $\mathcal{E}(a, b_1, A_1) = (\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ is a conformed pose, i.e. $\gamma(\mathcal{X})(a, b_1, A_1) = \mathcal{E}(a, b_1, A_1)$. We have already seen that $(\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3) = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)T_0(l_1)$. Refer to Figure 5.

Checking $\mathbf{e}'_0$. We have $[\gamma(\mathcal{X})(a, b_1, A_1)]\mathbf{u}_1 = \mathcal{X}(A_1) = \mathcal{X}(A) + \mathcal{E}(a, b_1, A)\mathbf{u}_4 l_1 = \mathbf{e}_0 + \mathbf{e}_3 l_1 = \mathbf{e}'_0$.

Checking $\mathbf{e}'_3$. $[\gamma(\mathcal{X})(a, b_1, A_1)]\mathbf{u}_4 = -\mathbf{U}/\|\mathbf{U}\| = -\mathbf{e}_3 = \mathbf{e}'_3$.

Checking $\mathbf{e}'_1$. $[\gamma(\mathcal{X})(a, b_1, A_1)]\mathbf{u}_2$ is the unit vector in the direction of $\mathbf{V} - \mathbf{e}'_3(\mathbf{e}'_3 \cdot \mathbf{V}) = \mathbf{V} - \mathbf{e}_3(\mathbf{e}_3 \cdot \mathbf{V})$, which is the same as $\mathbf{e}_1 = \mathbf{e}'_1$.

Checking $\mathbf{e}'_2$. The third components of $\gamma(\mathcal{X})(a, b_1, A_1)$ and $\mathcal{E}(a, b_1, A_1)$ must agree, since both are part of positively oriented orthonormal bases.

Therefore $\gamma(\mathcal{X}) = \mathcal{E}$. This finishes the proof that $G_\Gamma = G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$, and hence the proof of the main theorem. □

5. Examples of 3D Molecules

5.1. **Amino Acids.** In the previous sections we have developed a formalism for describing the conformations of molecules by attaching a number (an internal coordinate) to each edge of a Z-tree (or a GZ-tree) $\Gamma \subset AL^2(\mathcal{G})$ for the molecule. In this way molecules attain a definite shape in their mathematical description (which in chemistry courses is almost never true). These three dimensional (3D) molecules (i.e. labelled rooted GZ-trees) can then be manipulated like building blocks, with the result that larger and more complex structures may be built up. Thus we now

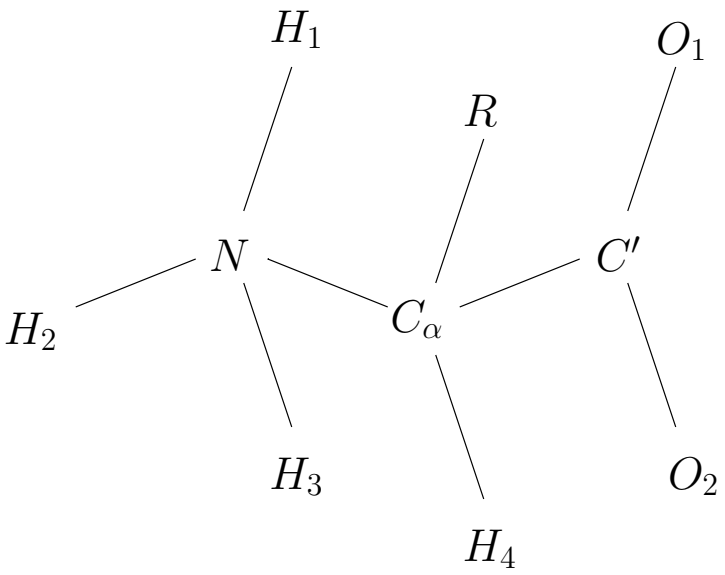

FIGURE 9. Molecular graph for a model amino acid in zwitterion form. R is an abbreviation for the side chain group. We have drawn this diagram in an angular way to remind the reader that this is a three dimensional molecule.

have in hand the most basic mathematical tool we need for a geometric study of structural molecular biology.

In the next three sections we will give concrete examples of labelled Z-trees for particular molecules. We restrict ourselves to Z-trees instead of using GZ-trees because of the close relationship with Z-matrices (which makes our 3D molecules easily understandable by existing visualization programs) and because Z-trees are easier to describe in tabular or list formats (see the discussion below). Thus these sections are addressed primarily to mathematicians, since chemists are well-aware that these molecules can be specified using Z-matrices. However, chemists may be interested to see how our formulation in terms of Z-trees works out in situations they already understand in other ways.

Our examples are canonical ones from biochemistry. The reader will not be assumed to know anything about the chemistry or biology. In Figure 9 we have the molecular graph $\mathcal{G}$ for a model amino acid. Amino acids are small molecules consisting of between 10 and 27 atoms, which are the building blocks of proteins, which are the workhorses of living cells.

In Figure 9 we have named atoms by the type of their chemical element, with a subscript to distinguish two atoms of the same chemical element. So H_2 means the second Hydrogen atom, not a molecule of two Hydrogen atoms as it would in chemistry courses. The reader will notice the "atom" R in Figure 9; but there is no chemical element with the symbol R. This stands for an entire molecular fragment called the "side chain". There are about 20 different types of amino acids which appear in natural proteins, and these types are distinguished by the identity of the side chain. (One of the amino acids called *Proline* has a side chain which attaches both at C_α and at N; we could give a Z-tree for proline, but will not.) For

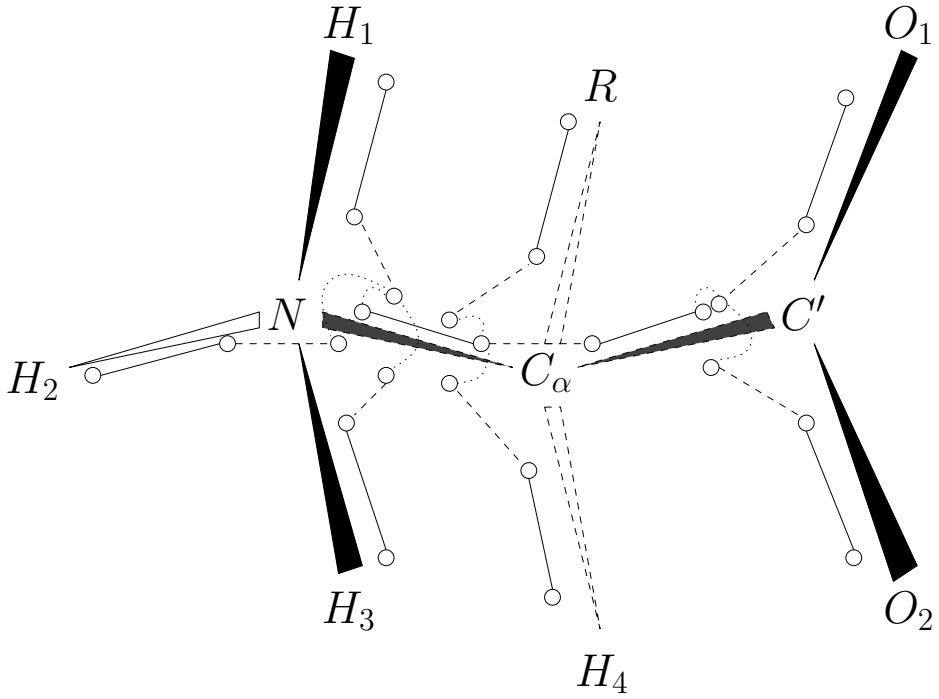

FIGURE 10. An embedded Z-tree for an amino acid. We use the embedding into three dimensional space discussed in section 2.2 (see also Figure 3). Vertices are denoted by small open circles. Edges of type 0 are denoted by solid lines, edges of type 1 by dashed lines, and edges of type 2 by dotted curves. There are $24 = 3(10) - 6$ edges in any Z-tree for a molecule with 10 atoms.

specificity, the reader might substitute a Hydrogen atom for the R, and this would yield the amino acid *Glycine*. But nothing about specifying Z-trees depends on the exact nature of the side chains, so we will keep the R as a dummy atom or group (a common chemical practice). It will eventually be clear how to enlarge the Z-tree for this model amino acid into a Z-tree for a specific type of amino acid.

A casual glance at an organic chemistry book might lead the reader to doubt if we have our amino acid right, since electrically neutral Nitrogen usually forms three bonds, not four. In fact, in the form of the amino acid we are considering the Nitrogen has a positive charge. Likewise electrically neutral Oxygen prefers to form two bonds, not one. In our molecular graphs we do not distinguish single bonds from double bonds; such chemical details are not in focus here. However, for the curious reader, both the atoms O_1 and O_2 carry about half of a negative charge. These details have to do with how the electrons are distributed in the space surrounding the positions of the nuclei. We will focus only on the positions of the nuclei. We pay some attention to the presence or absence of a covalent bond between two nuclei, but do not elaborate any further on the electron distribution.

The atom names C_α and C' conform to standard conventions in the study of proteins [23], [9]. A protein is formed from many copies of the above amino acid by linking the C' atom of one to the N atom of the next one via a covalent bond called

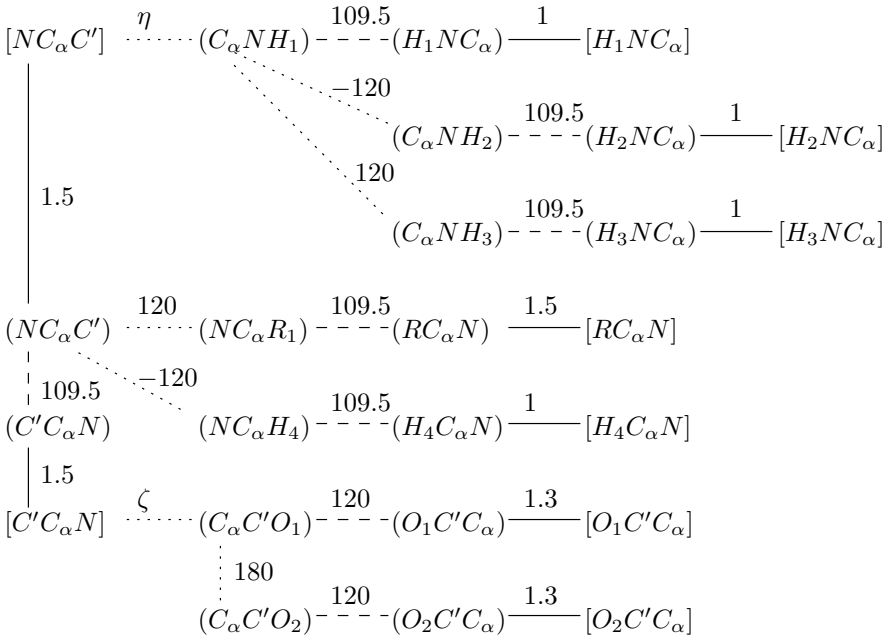

FIGURE 11. An abstract view of the same Z-tree pictured in Figure 10. We use a notation for the vertices which is explained in the text. Edges of type $0, 1, 2$ are indicated by solid, dashed, or dotted lines; they are labelled with internal coordinates in angstroms, degrees, and degrees respectively.

the *peptide bond*. In the process the atoms O_2 from the first amino acid and H_2 and H_3 from the second amino acid combine to form a water molecule, which loses all attachment to the pair of linked amino acids. This process is called a *condensation* reaction, or more specifically a *dehydration* reaction. By repeated dehydrations a long chain of amino acids, i.e. a protein, is formed.

A Z-tree for the amino acid is shown in Figure 10 in embedded form. This form is difficult to draw for complex molecules since it is essentially three dimensional, but hopefully this example shows how the structure of the Z-tree resembles that of the underlying molecular graph. In order to specify Z-trees for more complex situations we will employ other representations. However, this involves some new notation. Suppose A_1, A_2, A_3 are distinct vertices of $\mathcal{G}$. Then by the symbol $(A_1A_2A_3)$ we mean the triple (a, b, A_2), where $b = \{A_1, A_2\}$, $b' = \{A_2, A_3\}$, and $a = \{b, b'\}$. By the symbol $[A_1A_2A_3]$ we mean the triple (a, b, A_1). Thus the angle a is that determined by the three atoms, where $\alpha_1(a) = A_2$. The first two atoms determine the bond b. $(A_1A_2A_3)$ or $[A_1A_2A_3]$, i.e. parentheses or brackets, determine the base atom, either A_2 or A_1 respectively. Each open circle in Figure 10, representing a vertex in the Z-tree, can be assigned a symbol by the above rule. These symbols can then be connected with line segments, and we use a solid, dashed, or dotted line if the edge is of type 0, 1, or 2 respectively. An abstract view of the same Z-tree Γ as is pictured in Figure 10 can be seen in Figure 11.

Recall that the Z-tree Γ can be built up step by step by adding "branches". One starts with the "trunk", Γ_1, with four vertices and three edges; it is visible in

the first column on the left in Figure 11. Then one successively adds "branches" consisting of three new vertices and three new edges, starting at some vertex, the vertex of attachment, which was part of the previous level. These branches can be seen as (mostly) horizontal extensions of Γ_1 in Figure 11. The same information can be conveyed in *list* form.

$$
\begin{aligned}
[NC_\alpha C'] &\overset{1.5}{\text{———}} (NC_\alpha C') \overset{109.5}{\text{– – – –}} (C'C_\alpha N) \overset{1.5}{\text{———}} [C'C_\alpha N] \\
[NC_\alpha C'] &\overset{\eta}{\cdots} (C_\alpha NH_1) \overset{109.5}{\text{– – – –}} (H_1NC_\alpha) \overset{1}{\text{———}} [H_1NC_\alpha] \\
(C_\alpha NH_1) &\overset{-120}{\cdots} (C_\alpha NH_2) \overset{109.5}{\text{– – – –}} (H_2NC_\alpha) \overset{1}{\text{———}} [H_2NC_\alpha] \\
(C_\alpha NH_1) &\overset{120}{\cdots} (C_\alpha NH_3) \overset{109.5}{\text{– – – –}} (H_3NC_\alpha) \overset{1}{\text{———}} [H_3NC_\alpha] \\
(NC_\alpha C') &\overset{120}{\cdots} (NC_\alpha R) \overset{109.5}{\text{– – – –}} (RC_\alpha N) \overset{1.5}{\text{———}} [RC_\alpha N] \\
(NC_\alpha C') &\overset{-120}{\cdots} (NC_\alpha H_4) \overset{109.5}{\text{– – – –}} (H_4C_\alpha N) \overset{1}{\text{———}} [H_4C_\alpha N] \\
[C'C_\alpha N] &\overset{\zeta}{\cdots} (C_\alpha C'O_1) \overset{120}{\text{– – – –}} (O_1C'C_\alpha) \overset{1.3}{\text{———}} [O_1C'C_\alpha] \\
(C_\alpha C'O_1) &\overset{180}{\cdots} (C_\alpha C'O_2) \overset{120}{\text{– – – –}} (O_2C'C_\alpha) \overset{1.3}{\text{———}} [O_2C'C_\alpha]
\end{aligned}
$$

The first line is the trunk Γ_1. Subsequent lines describe the attached branches. Each one begins with a vertex already given on a previous line, the vertex of attachment. The order of the lines after the first is not important as long as the first vertex on each of these lines is given on a previous line. This list presentation is flexible enough to describe other more general types of graphs, not just Z-trees.

Since Γ is a Z-tree, rather than some more general type of subgraph of $AL^2(\mathcal{G})$, it admits of an even more compact presentation. Since the trunk Γ_1 is associated with a single angle (see Figure 3), it can be reconstructed from any of its vertices. Also each added branch is associated with a single new atom. The edges are added in the following order: first an edge of type 2, then an edge of type 1, and finally an edge of type 0. The last vertex must be based at the new atom. Given these rules, each branch can be uniquely reconstructed if we give its vertex of attachment, and the name of the new atom. Thus a more condensed *tabular* form of the above Z-tree might be

root vertex $[NC_\alpha C']$		bond length 1.5	bond angle 109.5	bond length 1.5
attachment vertex	new atom	wedge angle	bond angle	bond length
$[NC_\alpha C']$	H_1	η	109.5	1
$(C_\alpha NH_1)$	H_2	-120	109.5	1
$(C_\alpha NH_1)$	H_3	120	109.5	1
$(NC_\alpha C')$	R	120	109.5	1.5
$(NC_\alpha C')$	H_4	-120	109.5	1
$[C'C_\alpha N]$	O_1	ζ	120	1.3
$(C_\alpha C'O_1)$	O_2	180	120	1.3

Chemists will no doubt recognize the resemblence of the above to a Z-matrix specification of an amino acid. Thus we remark here (for the chemists) on the relation between the two descriptions, which are equivalent. In regard to atom

connectivity, the first three lines of a Z-matrix contain the same information as the first line of our tabular Z-tree specification. In regard to subsequent lines of a Z-matrix, which are in one to one correspondence to our Z-tree branch specifications, we have the following summary.

wedge type	Z-tree form		Z-matrix form			
dihedral	$[A_1A_2A_3]$	A_4	A_4	A_1	A_2	A_3
improper	$(A_1A_2A_3)$	A_4	A_4	A_2	A_1	A_3

In the Z-matrix data structure the ordered quadruple $A_4 \quad A_1 \quad A_2 \quad A_3$ means that new atom A_4 is bonded to A_1, which is (usually) in turn bonded to A_2, and A_3 is another atom which will (usually) be bonded to either A_1 or A_2. Which of these two options is the case will have been specified earlier in the Z-matrix, so the Z-tree and the Z-matrix specifications are equivalent in total, although not line by line. We have made caveats above when describing the meaning of an ordered quadruple in Z-matrices because Z-matrices are actually not restricted to follow the covalent bonding network of the molecule in the consistent and orderly manner that we have suggested. We will not discuss this more general concept of Z-matrix. Z-matrices are practical tools for describing molecules, and we now see how they are related to the underlying Z-trees. We prefer to use Z-trees, despite the extra mathematical "overhead", because they facilitate a deeper study of molecular conformation.

In order to describe a 3D amino acid we must label each edge of its Z-tree with the appropriate internal coordinate values. These are also shown in Figure 11. These numbers are also reported in the above list and tabular formats. We have filled the improper wedge angles with ideal tetrahedral values at N and C_α, and with the ideal planar value at C'. The bond angles also reflect these ideal geometries. The bond lengths are approximate for the types of atoms and the types of bonds between them. There are two unspecified dihedral angles η and ζ, which reflect the fact that amino acids are flexible molecules. Not all choices for these angles are energetically equivalent, but they are much less constrained than the other internal coordinates. The reader may have wondered why, in specifying a Z-tree, we attached $(C_\alpha NH_2)$ and $(C_\alpha NH_3)$ to $(C_\alpha NH_1)$ rather than directly to $[NC_\alpha C']$. The answer is that with this arrangement we can rotate the entire *amino group* (i.e. the $NH_1H_2H_3$ rigid assembly) by changing the single angle η. Likewise with our Z-tree the *carboxyl group* (i.e. $C'O_1O_2$) can be rotated by changing the single angle ζ. This benefit compensates us for a decided loss of symmetry.

The mathematical reader might be wondering at this stage why we care about these specific coordinate values, such as 1.5 angstroms or 109.5 degrees. The actual measured values deviate from these numbers slightly anyway. Mathematical structures do not typically contain at the outset restrictions on the values of numerical parameters without some exploration into the reasons why such restrictions are important. These values can be derived by solving the equations of quantum chemistry, but this does not help very much. Our belief is that specific geometrical constructions such as occur in biological macromolecular structures more or less constrain the values of these parameters to the near vicinity of their measured values, up to a scaling factor on all the bond lengths. By replacing the measured bond angles by ideal values we are hinting at the possibility that these geometric constructions might remain feasible in the idealized case. But these issues will have to be pursued elsewhere. For the time being we simply take these values as examples.

Another issue is whether the labelled Z-tree is just for looks, or is it good for something? We would like to think that it is a mathematical structure to be computed with. This means among other things that we should be able to check if it is correct (well formed) without using a computer program like RasMol to draw it. (Incidently, the freeware RasMol can draw molecules in Mopac Z-matrix format [11] using the command 'load mopac ⟨filename⟩'. Do a websearch on the keyword 'RasMol'.) The bond lengths can be checked directly based on the types of chemical elements involved and some knowledge of the bond order (i.e. single, double, aromatic, etc.). This sort of information can be found in organic chemistry textbooks. Bond angles and improper wedge angles are related to the type of *electronic hybridization* at each atom. Bond angles of 109.5 and improper wedge angles of ± 120 are characteristic of (ideal) sp^3 hybridization. Bond angles of 120 and improper wedge angles of ± 180 are characteristic of (ideal) sp^2 hybridization. From the Z-tree we have given one can discern that the atoms N and C_α are sp^3 hybridized and C' is sp^2 hybridized. Another aspect which is crucial to get right is the *chirality*, i.e. that aspect of the molecular conformation at an atom with at least four distinguishable *substituents*, i.e. groups bonded to it, which is reversed in its mirror image. This aspect is coded in the signs of the improper wedge angles, which depend on the choice of the root vertex in the Z-tree. In our example the root vertex is $[NC_\alpha C']$, and the atom C_α has four distinguishable substituents (provided R is not a Hydrogen atom), and hence is called a *stereocenter*. The choice of signs below, i.e.

$$(NC_\alpha C')\cdot\overset{120}{\cdots}\cdot(NC_\alpha R)$$
$$(NC_\alpha C')\cdot\overset{-120}{\cdots}\cdot(NC_\alpha H_4)$$

means that we have given the "L" form of the amino acid, since this is exclusively the form which appears in biosynthesized proteins. The "D" form is the mirror image, and can be obtained by switching the signs on the two improper wedge angles shown. Both chirality and hybridization can then be associated with particular patterns in the labelled Z-tree.

Another extremely important issue concerns the values assumed by the free dihedral angles η and ζ, as well as the details of the structures of the side chains. Sometimes a particular labelled Z-tree might represent an "impossible" molecule because certain atoms which are not bonded to each other are much too close to each other in space. Each type of chemical element has a *van der Waals radius*, and a sphere of that radius is considered to be centered at the nuclear position. If two atoms are not bonded (Hydrogen bonds are considered bonds for this purpose) then the interiors of the two spheres cannot intersect. This condition is called *van der Waals exclusion*. Checking a labelled Z-tree to see if it is *sterically allowed*, i.e. satisfies van der Waals exclusion for each nonbonded pair of atoms, is obviously a very extensive calculation, especially for larger molecules. But one could hope to check the condition on smaller molecules and then combine those components in such a way that the conditions will automatically be satisfied. This issue will be explored in future work, but not considered any further in this work.

5.2. **Nucleotides.** Just like amino acids are the building blocks for proteins, nucleotides are the building blocks of nucleic acids like DNA and RNA. Even though

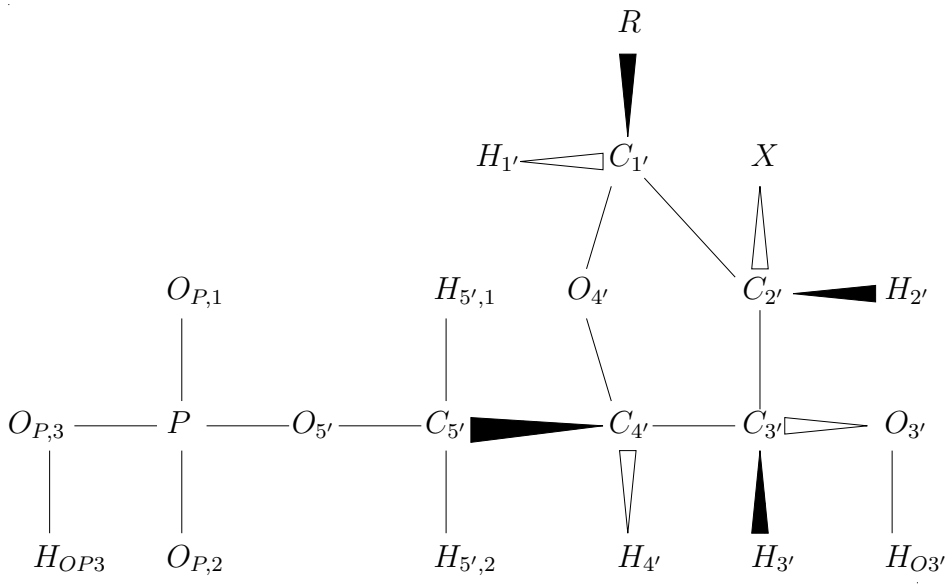

FIGURE 12. Molecular graph for a model nucleotide. R is an abbreviation for the base, of which there five different types; bases in nucleotides are analogous to side chains in amino acids. X stands for a Hydrogen atom in deoxyribonucleotides, and an OH group in ribonucleotides.

nucleic acids are the cellular information storage medium and proteins are fabricated by living cells based on these stored instructions, we describe nucleotides after amino acids because of their greater structural complexity. A molecular graph $\mathcal{G}$ for a model nucleotide is shown in Figure 12. We call it a model because the base (analogous to the side chain in amino acids) is denoted by R, and because a group X, whose identity depends on whether the nucleotide occurs in DNA (deoxyribonucleic acid, where $X = H_{2',2}$) or in RNA (ribonucleic acid, where $X = O_{2'}\text{——}H_{O2'}$), is also abstracted.

We will not burden the reader with the structures of the five types of bases which occur either in DNA or RNA, any more than we concerned ourselves with the structures of the 20 side chains in amino acids. DNA or RNA are formed from many copies of the above nucleotide, where the $O_{P,3}\text{——}H_{OP3}$ group of one nucleotide combines with the $H_{O3'}$ of another nucleotide to form a water molecule, and the two nucleotides become linked by a bond between the P of one and the $O_{3'}$ of the other. This bond completes the *phosphodiester linkage* between the two nucleotides. Thus the sequence of atoms $P, O_{5'}, C_{5'}, C_{4'}, C_{3'}, O_{3'}$ combine to form a backbone of a nucleic acid much like the N, C_α, C' atoms combine to form the backbone of a protein. The odd practice of putting primes on the indices of the atom names is to conform to standard nomenclature. The atom indices in the bases have no primes in standard notation [25].

Connecting the backbone and the base is the *furanose ring* comprised of atoms $C_{1'}, C_{2'}, C_{3'}, C_{4'}, O_{4'}$, covalently bonded in a five membered cycle. It is impossible

for a five membered ring to have tetrahedral bond angles [13]. Hence the conformation assumed by the furanose ring is a compromise between several different types of stresses, which are too complicated to discuss in detail here (for an introduction with references, see [42]). The main point which must be emphasized is that the conformations assumed by this ring are not determined by geometric criteria (at least this has yet to be demonstrated, see [2]), but by the criterion of *minimum energy*. This energy minimization also produces important but not especially geometrically natural conformations of the exocyclic (not part of the ring) substituents to the ring atoms. These effect the angle at which the base projects outward from the backbone, and hence the ability of the nucleotide units to fit together into a compact double helix. This energy function (the Born-Oppenheimer potential) is very complicated to define, and is very difficult to compute accurately for most molecular systems with 10 or more nuclei. Molecular mechanics energy functions are rough approximations to the Born-Oppenheimer potential involving elementary functions with adjustable parameters not computed from first principles. Thus we wish to avoid a direct energy minimization. It is not our purpose here to survey all the subtleties of furanose ring conformations. (There are still unresolved mathematical issues here!) We wish to illustrate how our system for conformational specification can be applied to this ring without becoming too bogged down in details. Thus the expert should forgive various inaccuracies in our given conformation.

We can give a labelled Z-tree for the nucleotide in list form as follows.

$$
\begin{aligned}
&[PO_{5'}C_{5'}] \overset{1.6}{\text{———}} (PO_{5'}C_{5'}) \overset{109.5}{\text{– – – –}} (C_{5'}O_{5'}P) \overset{1.4}{\text{———}} [C_{5'}O_{5'}P] \text{ trunk}\\
&[PO_{5'}C_{5'}] \overset{\alpha'}{\cdots\cdots} (O_{5'}PO_{P,1}) \overset{109.5}{\text{– – – –}} (O_{P,1}PO_{5'}) \overset{1.5}{\text{———}} [O_{P,1}PO_{5'}] \text{ phosphate}\\
&(O_{5'}PO_{P,1}) \overset{120}{\cdots\cdots} (O_{5'}PO_{P,2}) \overset{109.5}{\text{– – – –}} (O_{P,2}PO_{5'}) \overset{1.5}{\text{———}} [O_{P,2}PO_{5'}] \text{ phosphate}\\
&(O_{5'}PO_{P,1}) \overset{-120}{\cdots\cdots} (O_{5'}PO_{P,3}) \overset{109.5}{\text{– – – –}} (O_{P,3}PO_{5'}) \overset{1.6}{\text{———}} [O_{P,3}PO_{5'}] \text{ linker}\\
&[O_{P,3}PO_{5'}] \overset{\zeta'}{\cdots\cdots} (PO_{P,3}H_{OP3}) \overset{109.5}{\text{– – – –}} (H_{OP3}O_{P,3}P) \overset{1}{\text{———}} [H_{OP3}O_{P,3}P] \text{ linker}\\
&[C_{5'}O_{5'}P] \overset{-\beta}{\cdots\cdots} (O_{5'}C_{5'}C_{4'}) \overset{109.5}{\text{– – – –}} (C_{4'}C_{5'}O_{5'}) \overset{1.5}{\text{———}} [C_{4'}C_{5'}O_{5'}] \text{ backbone}\\
&[C_{4'}C_{5'}O_{5'}] \overset{-\gamma}{\cdots\cdots} (C_{5'}C_{4'}C_{3'}) \overset{\theta(\tau_4)}{\text{– – – –}} (C_{3'}C_{4'}C_{5'}) \overset{1.52}{\text{———}} [C_{3'}C_{4'}C_{5'}] \text{ backbone}\\
&[C_{3'}C_{4'}C_{5'}] \overset{-\delta}{\cdots\cdots} (C_{4'}C_{3'}O_{3'}) \overset{\theta(\tau_3)}{\text{– – – –}} (O_{3'}C_{3'}C_{4'}) \overset{1.4}{\text{———}} [O_{3'}C_{3'}C_{4'}] \text{ backbone}\\
&[O_{3'}C_{3'}C_{4'}] \overset{\epsilon'}{\cdots\cdots} (C_{3'}O_{3'}H_{O3'}) \overset{109.5}{\text{– – – –}} (H_{O3'}O_{3'}C_{3'}) \overset{1}{\text{———}} [H_{O3'}O_{3'}C_{3'}] \text{ linker}\\
&(O_{5'}C_{5'}C_{4'}) \overset{120}{\cdots\cdots} (O_{5'}C_{5'}H_{5',1}) \overset{109.5}{\text{– – – –}} (H_{5',1}C_{5'}O_{5'}) \overset{1.1}{\text{———}} [H_{5',1}C_{5'}O_{5'}] \text{ ethylene}\\
&(O_{5'}C_{5'}C_{4'}) \overset{-120}{\cdots\cdots} (O_{5'}C_{5'}H_{5',2}) \overset{109.5}{\text{– – – –}} (H_{5',2}C_{5'}O_{5'}) \overset{1.1}{\text{———}} [H_{5',2}C_{5'}O_{5'}] \text{ ethylene}\\
&(C_{3'}C_{4'}C_{5'}) \overset{-\phi(\tau_4)}{\cdots\cdots} (C_{3'}C_{4'}O_{4'}) \overset{\tau_4}{\text{– – – –}} (O_{4'}C_{4'}C_{3'}) \overset{1.45}{\text{———}} [O_{4'}C_{4'}C_{3'}] \ 4' \text{ arm}\\
&(C_{3'}C_{4'}O_{4'}) \overset{-\phi(\tau_4)}{\cdots\cdots} (C_{3'}C_{4'}H_{4'}) \overset{\theta(\tau_4)}{\text{– – – –}} (H_{4'}C_{4'}C_{3'}) \overset{1.1}{\text{———}} [H_{4'}C_{4'}C_{3'}] \ 4' \text{ arm}\\
&[O_{4'}C_{4'}C_{3'}] \overset{-\nu_4}{\cdots\cdots} (C_{4'}O_{4'}C_{1'}) \overset{\tau_0}{\text{– – – –}} (C_{1'}O_{4'}C_{4'}) \overset{1.42}{\text{———}} [C_{1'}O_{4'}C_{4'}] \ 4' \text{ arm}\\
&[C_{1'}O_{4'}C_{4'}] \overset{-\nu_0+\phi(\tau_1)}{\cdots\cdots} (O_{4'}C_{1'}R) \overset{\theta(\tau_1)}{\text{– – – –}} (RC_{1'}O_{4'}) \overset{1.5}{\text{———}} [RC_{1'}O_{4'}] \ 4' \text{ arm}
\end{aligned}
$$

$$[C_{1'}O_{4'}C_{4'}]\overset{-\nu_0-\phi(\tau_1)}{\cdots\cdots}(O_{4'}C_{1'}H_{1'})\overset{\theta(\tau_1)}{----}(H_{1'}C_{1'}O_{4'})\overset{1.1}{\text{———}}[H_{1'}C_{1'}O_{4'}]\ 4'\text{ arm}$$

$$(C_{4'}C_{3'}O_{3'})\overset{-\phi(\tau_3)}{\cdots\cdots}(C_{4'}C_{3'}C_{2'})\overset{\tau_3}{----}(C_{2'}C_{3'}C_{4'})\overset{1.52}{\text{———}}[C_{2'}C_{3'}C_{4'}]\ 3'\text{ arm}$$

$$(C_{4'}C_{3'}C_{2'})\overset{-\phi(\tau_3)}{\cdots\cdots}(C_{4'}C_{3'}H_{3'})\overset{\theta(\tau_3)}{----}(H_{3'}C_{3'}C_{4'})\overset{1.1}{\text{———}}[H_{3'}C_{3'}C_{4'}]\ 3'\text{ arm}$$

$$[C_{2'}C_{3'}C_{4'}]\overset{-\nu_2-\phi(\tau_2)}{\cdots\cdots}(C_{3'}C_{2'}X)\overset{\theta(\tau_2)}{----}(XC_{2'}C_{3'})\overset{1.4}{\text{———}}[XC_{2'}C_{3'}]\ 3'\text{ arm}$$

$$[C_{2'}C_{3'}C_{4'}]\overset{-\nu_2+\phi(\tau_2)}{\cdots\cdots}(C_{3'}C_{2'}H_{2'})\overset{\theta(\tau_2)}{----}(H_{2'}C_{2'}C_{3'})\overset{1.1}{\text{———}}[H_{2'}C_{2'}C_{3'}]\ 3'\text{ arm}$$

The root vertex of Γ is $[PO_{5'}C_{5'}]$. This graph contains of course an enormous amount of information which is not so easy to assimilate, so we will discuss several features of the labelled Z-tree we have chosen. The labelling values could always be obtained from molecular mechanics software, but this would yield very little understanding and we will not use this approach.

First consider the backbone, although it only becomes complete when several nucleotides are linked together. The bond lengths shown are approximate, taken from the CHARMM forcefield, [30]. The bond angles along the backbone have been set at the tetrahedral value of $109.5^\circ = \cos^{-1}(-\frac{1}{3})$ for simplicity, even though experimental values are known [36]. On the part of the backbone which is also part of the furanose ring, the bond angles are affected by other factors which we will address below. Also improper wedge angles have been chosen according to ideal sp^3 hybridization, even though that is also not exactly true. We have named the backbone dihedrals β, γ, δ according to standard conventions. In usual DNA strucutes (i.e. so called B-DNA) we have the average values $\beta = 136$, $\gamma = 38$, and $\delta = 139$ (see [8]). However, the value of δ is coupled to the conformation of the furanose ring in a manner which we will discuss below. The orientation of the entire phosphate group is controlled by α', and that of its linker OH bond by ζ'. The orientation of the other linker OH bond is controlled by ϵ'. These names are related to, but not identical with, the standard linker backbone dihedral names, α, ϵ, ζ, in the chain polymers DNA and RNA [42]. Since we are focussing on a single nucleotide rather than a chain of them linked together, we will not discuss the angles α, ϵ, ζ any further; we also will not discuss specific values for $\alpha', \epsilon', \zeta'$.

Because a Z-tree is a tree subgraph, it cannot contain cycles. Thus even though $\mathcal{G}$ has a closed ring, Γ cannot. This means that one must decide how one wants to describe the ring, and which conformational variables should be omitted. We have chosen the $\{C_{1'}, C_{2'}\}$ bond as the part of the ring not traversed by Γ. Thus we break the ring into two arms off the backbone, one (the $4'$ arm) terminating in the base R and the other (the $3'$ arm) terminating in the group X. This approach would allow us if we wished to attach the base (described by its own Z-tree, unspecified here) and to specify the orientation of this base with respect to the rest of the nucleotide in terms of a torsion angle χ about the $\{C_{1'}, R\}$ bond. It is consistent with standard conventions that this edge of type 2 have the parent vertex $[RC_{1'}O_{4'}]$. Our approach would then have the angle $-\chi$ labelling that edge. Thus our choice of Z-tree is designed to allow the standardized conformational angles to appear explicitly as labels of edges, even after the base is attached.

Now let us concentrate on the five membered furanose ring, and momentarily forget also about the exocyclic substituents of the ring atoms. Thus we have a

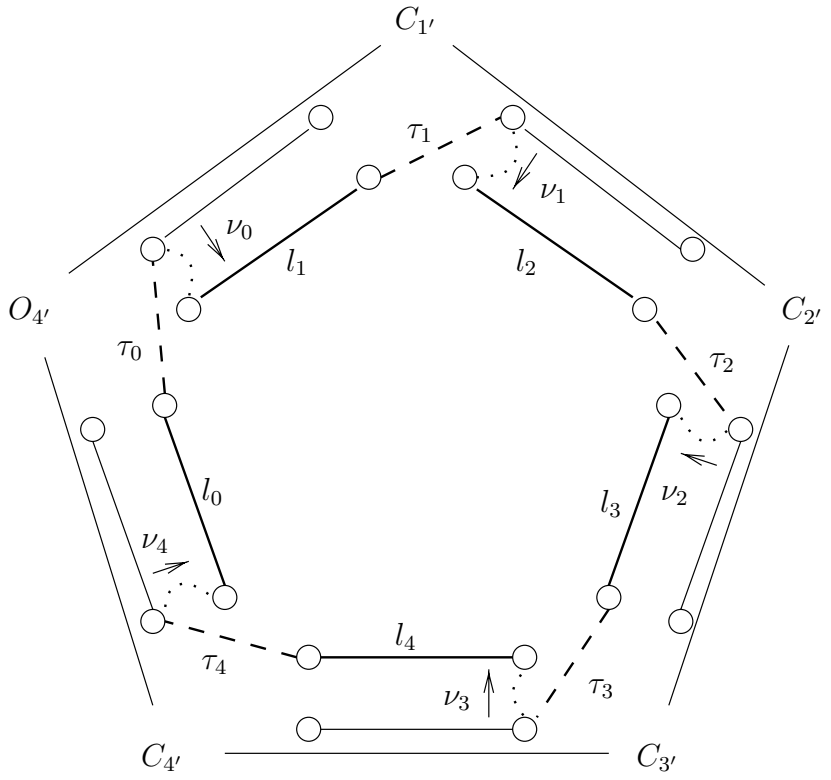

FIGURE 13. Conformational constraints in a five membered ring arise from the fact that the composition of all the coordinate transformations in a path once around the ring is the identity.

system pictured in Figure 13. The cyclic graph in the figure is not part of Γ, and its orientation is as indicated, i.e. clockwise is the positive direction. Following IUPAC conventions [25], we call the dihedral angles in the ring by the names ν_j, $j = 0, 1, \ldots, 4$. These are defined as in the figure. This means that $T_2(e^{i\nu_j})$ transforms the pose associated to the parent vertex to the pose associated with the child vertex. There are five bond lengths l_j, five bond angles τ_j, and five torsion angles ν_j, $j = 0, 1, \ldots, 4$, for a total of 15 parameters, subject to six equations of constraint. The constraint equations can be written as follows (see Figure 13).

$$\begin{aligned}\mathbf{1} = {} & T_0(l_0)T_1(\cos\tau_0)T_2(e^{i\nu_0})T_0(l_1)T_1(\cos\tau_1)T_2(e^{i\nu_1})T_0(l_2)T_1(\cos\tau_2)T_2(e^{i\nu_2}) \\ & \cdot T_0(l_3)T_1(\cos\tau_3)T_2(e^{i\nu_3})T_0(l_4)T_1(\cos\tau_4)T_2(e^{i\nu_4}).\end{aligned}$$

Since all these matrices are in the six dimensional Lie group G_p, we obtain six equations, called the *ring closure* equations. However, using the relation $T_2(e^{i\nu_3})T_0(l_4) = T_0(l_4)T_2(e^{-i\nu_3})$, the first column of the above equation yields

$$\begin{aligned}\mathbf{u}_1 = {} & T_0(l_0)T_1(\cos\tau_0)T_2(e^{i\nu_0})T_0(l_1)T_1(\cos\tau_1)T_2(e^{i\nu_1})T_0(l_2)T_1(\cos\tau_2)T_2(e^{i\nu_2}) \\ & \cdot T_0(l_3)T_1(\cos\tau_3)T_0(l_4)T_2(e^{-i\nu_3})T_1(\cos\tau_4)T_2(e^{i\nu_4})\mathbf{u}_1\end{aligned}$$

$$= T_0(l_0)T_1(\cos\tau_0)T_2(e^{i\nu_0})T_0(l_1)T_1(\cos\tau_1)T_2(e^{i\nu_1})T_0(l_2)T_1(\cos\tau_2)T_2(e^{i\nu_2})$$
$$\cdot\, T_0(l_3)T_1(\cos\tau_3)T_0(l_4)\mathbf{u}_1.$$

This gives a system of three equations, since the first component of the column vector on the right hand side is always 1. We will call these equations the *loop equations*. The other three equations of the ring closure system can be written down explicitly using formulae (18) and (19) on page 73 of [3]. Since $\cos\theta = \frac{1-x^2}{1+x^2}$, and $\sin\theta = \frac{2x}{1+x^2}$, these equations can be converted into multivariate polynomial equations, and algebraic geometry tools can be applied (see [15] and references therein). If the bond lengths $l_0, \ldots, l_4$ are assumed fixed and given then this system determines a four dimensional manifold, with interesting properties that we cannot discuss here. We will call this manifold the *ring manifold*. The author has not succeeded in finding where in the mathematical or chemical literature a systematic study of the ring manifold has been carried out (see however [41]).

Most of the points of this four dimensional ring manifold describe ring conformations which are unrealistic because the bond angles or torsion angles are not in the proper ranges, as dictated by energetic considerations. A simpler situation to study first is where the atom $O_{4'}$ is replaced by $C_{0'}$, and all the bond lengths are equal. This is the case in the molecule *cyclohexane*. In that case it turns out that there is a distinguished one dimensional submanifold, diffeomorphic to the circle, sitting inside the ring manifold on which the energy is nearly constant and away from which the energy increases markedly. There is no effective energetic barrier to motion along this one dimensional manifold; such motion is called *pseudorotation*. We will call this one dimensional submanifold (which we will not carefully define) of the ring manifold the *pseudorotation manifold* in the case of cyclohexane. It is an interesting problem to come up with a geometric characterization of the pseudorotation manifold, i.e. a characterization which does not depend on all the details of the actual potential energy function (see [2]). But the author is not aware if this problem has been solved. There are only two ring conformations of cyclohexane where all the bond angles are equal, both of which are planar [13]. None of the ring conformations on the pseudorotation manifold correspond to a planar ring, i.e. they are all *puckered*. This is a result of the influence of the exocyclic substituents of the ring atoms. If all other influences are equal torsion angles resulting in a *staggered* arangement of the substituents is of lower energy than one where the substituents are *eclipsed* (as one looks down the common bond of the wedge). Bond angle bending strain and torsional strain are both important parts of the balance on the pseudorotation manifold (see Table I of [28]). Thus it is difficult to come up with simple geometric descriptions of realistic puckered ring conformations (points on the ring manifold), even in cyclohexane.

In the furanose ring of nucleotides the situation is even more complicated. Besides the fact that the bond lengths are no longer equal there is the important aspect that the exocyclic substituents vary from one ring atom to the next. $O_{4'}$ has no exocyclic substituents at all; $C_{1'}$ has the base attached to it, and this causes distortions in the geometry of the ring. $C_{3'}$ and possibly $C_{2'}$ have bulky OH substituents which are less tolerant of eclipsed conformations than the exocyclic hydrogen atoms of cyclohexane. All five bond angles differ in their tolerance of bending strain. Despite these differences, part of the "pseudorotation manifold" remains of low energy; there are two energy local minima corresponding to the

conformations usually called $C_{2'}$ *endo* and $C_{3'}$ *endo*, connected over a relatively low energy barrier by an one dimensional steepest descent path which fairly closely follows the ideal pseudorotation manifold [28]. There is very little resistence to pseudorotational motion at either of the two local minima. Furthermore, this flexibility of the furanose ring is biologically important since among other things it helps the double helix structure of DNA to adjust to local deformations of its structure due to the particular sequence of the bases (see page 38ff of [8]). There is little chance of reducing this complicated behavior to simple geometric rules.

Although it is tricky to parameterize the one dimensional path on the ring manifold connecting the $C_{2'}$ endo and $C_{3'}$ endo endpoints, it is much easier to do this approximately, allowing the path to deviate somewhat from the ring manifold. The following approximate formulae can be found in Table 6 of [27].

$$\begin{aligned}\tau_j &= a'_j + b'_j \cos(2P + 8\pi j/5),\\ \nu_j &= a_j + b_j \nu_m \cos[P + 4\pi(j-2)/5], \quad j = 0, 1, \ldots, 4.\end{aligned}$$

The parameter P is called the *phase*, and as it is varied the conformation moves around the pseudorotation manifold. The phase value for the typical $C_{2'}$ endo conformation is about $P = 170$ degrees; for $C_{3'}$ endo it is about $P = 10$ degrees [42]. The parameter ν_m is called the *torsional puckering amplitude*; when it is zero the ring conformation should be planar (notice that the standard deviations of a_j listed below typically exceed the average value of a_j). The puckering amplitude is constant in an ideal pseudorotational motion. The typical puckering amplitude for furanose rings in deoxyribonucleotides in the $C_{2'}$ endo conformation is about $\nu_m = 36$ degrees (see Table 4 of [27]). The part of the pseudorotation manifold which survives in nucleotides can be expressed as $0 \leq P \leq 180$ degrees, with $\nu_m = 36$ degrees. The transition state between the two endpoint conformations occurs near $P = 90$ degrees (the so-called $O_{4'}$ endo conformation), but the energy of the transition state is sufficiently greater than that of the endpoints so that it is rarely seen in crystals. The specific parameter values are given in the following table.

j	l_j	av. a'_j	b'_j	std. dev. a'_j	av. a_j	b_j	std. dev. a_j
0	1.45	107.8	2.3	0.7	0.13	1.027	0.2
1	1.42	106.0	1.7	0.7	−0.18	1.010	0.3
2	1.52	102.6	1.7	0.8	0.09	0.976	0.2
3	1.52	102.9	0.7	0.7	0.29	1.015	0.3
4	1.52	104.9	1.7	0.7	−0.20	1.026	0.3

These are derived from averages of structures determined by X-ray crystallography, [27]. The bond lengths are taken from Figure 4.13, page 70 of [42], and are specific to the $C_{2'}$ endo conformation of deoxyribonucleotides.

Using Lagrange multipliers one could find the point on the ring manifold which minimizes the deviation of the ten angles $\tau_j, \nu_j, j = 0, \ldots, 4$ from the values predicted by the above formulae, with $P = 170$ and $\nu_m = 36$. This would involve solving a system of 16 equations in 16 unknowns, and would result in a level of accuracy which is far in excess of that of our description of the other parts of the nucleotide. Thus our approach will be to use the above formulae as a substitute for

solving the ring closure equations. Thus we have the values

$$\begin{aligned} \tau_0 &= 110.0 \qquad \nu_0 = -17.2 \\ \tau_1 &= 105.9 \qquad \nu_1 = 32.5 \\ \tau_2 &= 101.0 \qquad \nu_2 = -34.5 \\ \tau_3 &= 102.5 \qquad \nu_3 = 25.7 \\ \tau_4 &= 106.0 \qquad \nu_4 = -5.3 \end{aligned}$$

These compare well with tabulated averages (see Figure 4-13 of [42], and [36]). This will fix the conformation of the ring for a typical deoxyribonucleotide.

Since the endocyclic bond angles deviate from the tetrahedral values, it no longer makes sense to assign tetrahedral values for the exocyclic bond and improper wedge angles. These angles do vary somewhat during pseudorotation, but they are also influenced by many factors such as the identity of X and the orientation of the base R, which we are glossing over. Thus we will instead use a simple geometrically motivated approximation to compute these angles. In this approach we assume that the measure τ of the endocyclic bond angle $A_1 - C - A_2$ is given (A_1, A_2 are ring atoms) and that the other two substituents X_1, X_2 to the shared ring carbon atom C arrange themselves so that the all the bond angles $A_1 - C - X_1$, $A_1 - C - X_2$, $A_2 - C - X_1$, $A_2 - C - X_2$, and $X_1 - C - X_2$ have the same measure. Let this measure be denoted by $\theta(\tau)$, since it is a function of τ. It is an amusing exercise in trigonometry to show that

$$\cos\theta(\tau) = \frac{-2}{1 + \sqrt{1 + 16\frac{1-\cos\tau}{\sin^2\tau}}}.$$

Let $\phi(\tau)$ denote the (absolute value of the) common measure of the exocyclic improper wedge angles

$$\begin{aligned} &(A_1CX_1)\cdot\cdot\cdot\cdot\cdot(A_1CA_2)\cdot\cdot\cdot\cdot\cdot(A_1CX_2) \\ &(A_2CX_1)\cdot\cdot\cdot\cdot\cdot(A_2CA_1)\cdot\cdot\cdot\cdot\cdot(A_2CX_2). \end{aligned}$$

Then further trigonometric work yields the formula

$$\phi(\tau) = \pi - \frac{1}{2}\cos^{-1}\left(\frac{\cos\theta(\tau)}{1+\cos\theta(\tau)}\right).$$

These formulae, together with the above values of the endocyclic bond angles $\tau_j, j = 1, \ldots, 4$ yield the following values.

$$\begin{aligned} \theta(\tau_1) &= 110.2 \qquad \phi(\tau_1) = 119.1 \\ \theta(\tau_2) &= 111.1 \qquad \phi(\tau_2) = 117.9 \\ \theta(\tau_3) &= 110.8 \qquad \phi(\tau_3) = 118.3 \\ \theta(\tau_4) &= 110.2 \qquad \phi(\tau_4) = 119.1 \end{aligned}$$

Clearly the deviations from tetrahedral geometry are not large in this scheme. These determine a particular (idealized) conformation of the exocyclic substituents of the furanose ring.

It remains to explain how the backbone conformation is coupled to the ring conformation. This occurs because ν_3 is closely related to δ. The relationship is clearer if we inspect Figure 14. Since $T_2(e^{i\phi})T_0(l) = T_0(l)T_2(e^{-i\phi})$, and since the angle connecting $(C_{3'}C_{4'}C_{5'})$ to $(C_{3'}C_{4'}O_{4'})$ is $-\phi(\tau_4)$ (according to the above

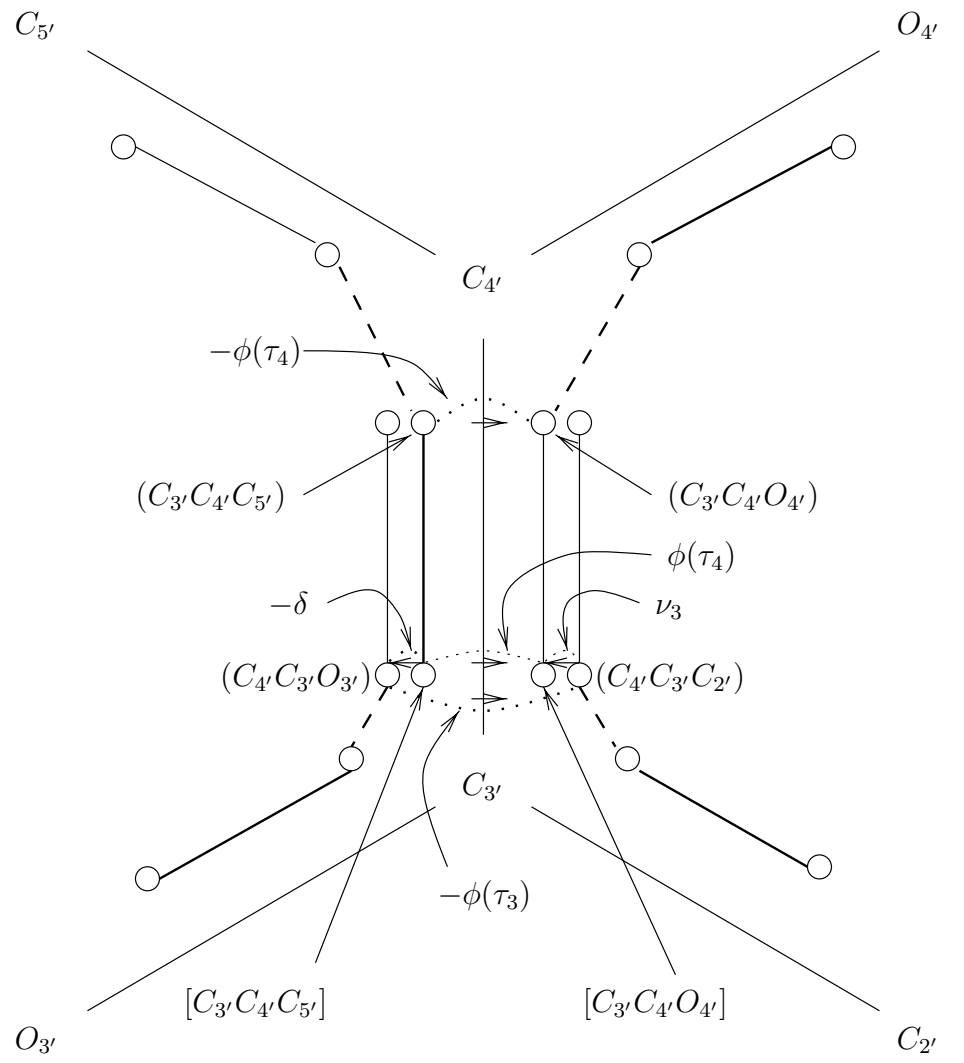

FIGURE 14. Detail of Γ near the bond $\{C_{4'}, C_{3'}\}$. Heavier lines denote edges in Γ whereas lighter lines denote edges in $AL^2(\mathcal{G})$. The labelled values of edges of type 2 are associated with the indicated orientations.

list of the Z-tree Γ), then the angle connecting $[C_{3'}C_{4'}C_{5'}]$ to $[C_{3'}C_{4'}O_{4'}]$ is $\phi(\tau_4)$. Since $T_2(e^{i\phi})T_2(e^{i\phi'}) = T_2(e^{i(\phi+\phi')})$, we have that $-\delta - \phi(\tau_3) = -\nu_3 + \phi(\tau_4)$, or $\delta = \nu_3 - \phi(\tau_3) - \phi(\tau_4)$ (mod 360°). With the above values we compute $\delta = 148.3$ degrees, as compared to the average value of $\delta = 139$ in B-DNA that we quoted earlier. This discrepancy is not serious given all of our approximations.

Because of the demands of describing both a backbone and a ring strucure, we attached the $4'$ arm to the vertex $(C_{3'}C_{4'}C_{5'})$ instead of to the vertex $(C_{5'}C_{4'}C_{3'})$. This allows a much simpler relationship between the ring conformation and the

backbone conformation to be derived. Also we attached the substituents at $C_{1'}$ and at $C_{2'}$ in an unusual manner. Rather than using improper wedge angles to specify the substituents' orientation relative to some reference angle we have used two torsion angles with more complex labels of the form $\nu_0 \pm \phi(\tau_1)$ and $\nu_2 \pm \phi(\tau_3)$. This means we are taking the endocyclic bond angle as the reference, even though we cannot add the third atom of that angle in the usual manner because it would cause the atom to be added too many times. Thus one checks the chirality at the atoms $C_{4'}$, $C_{1'}$, and $C_{2'}$ in a different manner than was discussed in section 5.1.

The nucleotide is therefore seen to be rather a complex building block (we have simplified its complexity!). Despite its complexity, we aspire to a mathematical understanding of the geometry of polynucleotides, and their transformations. The Z-tree given here is just a necessary starting point for such a study. Directions in which this study has been pursued nonrigorously (i.e. without proving theorems) and computationally by chemists can be found in [42].

5.3. **Glucose.** The third major category of small molecules which can be strung together to form large polymeric molecules essential for living organisms is that of *monosaccharides*. In this section we will describe a labelled Z-tree for the sugar *glucose*. This molecule possesses a ring of six atoms in which bond angle bending and torsional strain are quite small, and hence an approximate ring conformation (the "chair") can be obtained with much less effort than was the case for the furanose ring in nucleotides. The geometric study of more exact ring conformations of glucose is mathematically quite interesting (see [12], [13], and [15]), but beyond the purview of this work. Glucose (as well as other sugar monomers) has the capability to be linked into chains (polysaccharides) in a variety of ways, leading both to unbranched polymers (such as *amylose* and *cellulose*) and to branched polymers (such as *amylopectin* and *glycogen*). Thus we will choose our Z-tree with these modes of polymerization in mind.

In Figure 15 we present a molecular graph $\mathcal{G}$ for glucose. It is very close to a planar projection of an actual conformation of this molecule, but the shape is difficult to discern from this figure. Each OH group is a potential site where linkages can be made between two glucose monomers. For example the disaccharide *lactose* is formed when the H_{O4} atom of one glucose monomer combines with the O_1——H_{O1} group of another glucose monomer to form a water molecule, and a new bond between the C_1 atom of the second monomer and the O_4 atom of the first monomer forms. This new bond, which links the two monomers to form a lactose molecule is called the $1 \to 4$ *glycosidic* bond. It is analogous to the peptide bond in proteins, and the phosphodiester linkage in nucleic acids. However, in contrast to proteins or nucleic acids, two glucose monomers can also be connected by a $1 \to 6$ glycosidic bond to form *gentiobiose*. Furthermore, these disaccharides have available OH groups attached to carbons C_4 or C_6, as sites for further attachments. In this way complex branched polymers could be formed. These polymers are used by living cells to store chemical energy and for structural fibers. Unlike amino acids or nucleotides, sugar monomers do not have a variety of side chains; the OH groups are the "side chains", but there is only one type. However these side chains can appear in two different positions relative to the ring, *axial* or *equatorial*. Glucose has all its OH groups in an equatorial position. If the OH group on C_4 in glucose is changed from equatorial to axial position, then the sugar is called *galactose*. The atom C_1 is called the *anomeric* carbon. If the OH group on the anomeric carbon

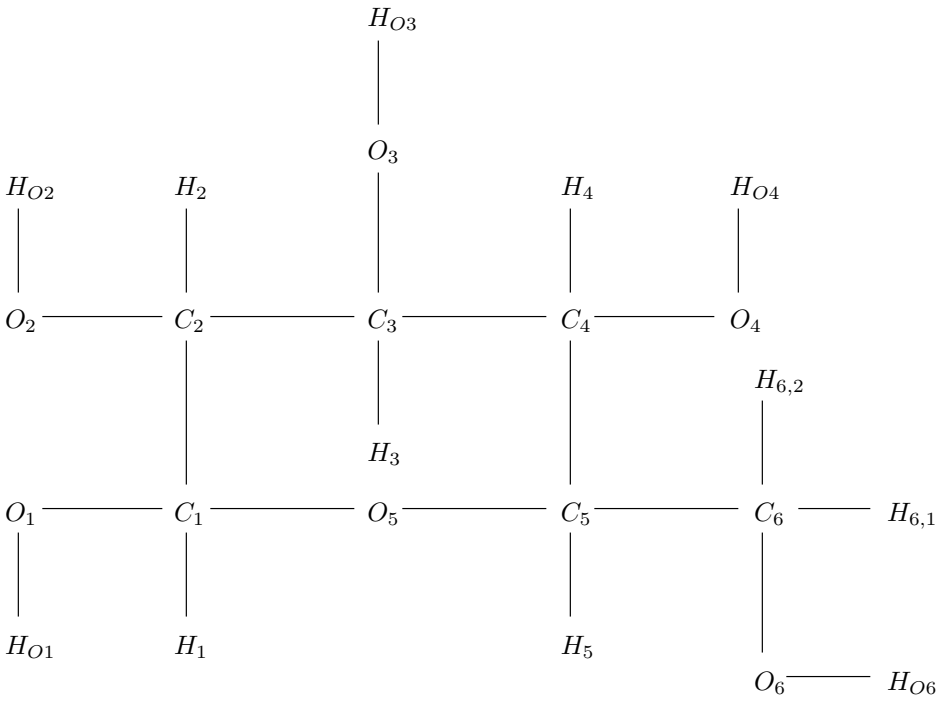

FIGURE 15. Molecular graph for glucopyranose.

is equatorial (resp. axial) then the sugar is called β (resp. α) glucose. Glucose can exist in a structure which does not have a ring. If we wish to be certain that we are discussing the ring form of glucose, we call it *glucopyranose*.

The following is a labelled Z-tree for β-D-glucopyranose, which is a particular form of glucose.

$$[O_5C_1C_2]\overset{1.5}{\text{———}}(O_5C_1C_2)\ \overset{109.5}{\text{– – – –}}\ (C_2C_1O_5)\overset{1.5}{\text{———}}[C_2C_1O_5]\ \text{trunk}$$
$$[C_2C_1O_5]\cdot\overset{-\nu_1}{\cdots}\cdot(C_1C_2C_3)\ \overset{109.5}{\text{– – – –}}\ (C_3C_2C_1)\overset{1.5}{\text{———}}[C_3C_2C_1]\ \text{backbone}$$
$$[C_3C_2C_1]\cdot\overset{-\nu_2}{\cdots}\cdot(C_2C_3C_4)\ \overset{109.5}{\text{– – – –}}\ (C_4C_3C_2)\overset{1.5}{\text{———}}[C_4C_3C_2]\ \text{backbone}$$
$$[C_4C_3C_2]\cdot\overset{-\nu_3}{\cdots}\cdot(C_3C_4C_5)\ \overset{109.5}{\text{– – – –}}\ (C_5C_4C_3)\overset{1.5}{\text{———}}[C_5C_4C_3]\ \text{backbone}$$
$$[C_5C_4C_3]\overset{-\nu_4+120}{\cdots}(C_4C_5C_6)\ \overset{109.5}{\text{– – – –}}\ (C_6C_5C_4)\overset{1.5}{\text{———}}[C_6C_5C_4]\ \text{backbone}$$
$$(O_5C_1C_2)\cdot\overset{-120}{\cdots}\cdot(O_5C_1H_1)\ \overset{109.5}{\text{– – – –}}\ (H_1C_1O_5)\overset{1}{\text{———}}[H_1C_1O_5]\ \text{1-substituents}$$
$$(O_5C_1C_2)\cdot\overset{120}{\cdots}\cdot(O_5C_1O_1)\ \overset{109.5}{\text{– – – –}}\ (O_1C_1O_5)\overset{1.4}{\text{———}}[O_1C_1O_5]\ \text{1-substituents}$$
$$[O_1C_1O_5]\cdot\overset{-\chi_1}{\cdots}\cdot(C_1O_1H_{O1})\ \overset{109.5}{\text{– – – –}}\ (H_{O1}O_1C_1)\overset{1}{\text{———}}[H_{O1}O_1C_1]\ \text{1-substituents}$$
$$(C_1C_2C_3)\cdot\overset{120}{\cdots}\cdot(C_1C_2H_2)\ \overset{109.5}{\text{– – – –}}\ (H_2C_2C_1)\overset{1}{\text{———}}[H_2C_2C_1]\ \text{2-substituents}$$
$$(C_1C_2C_3)\cdot\overset{-120}{\cdots}\cdot(C_1C_2O_2)\ \overset{109.5}{\text{– – – –}}\ (O_2C_2C_1)\overset{1.4}{\text{———}}[O_2C_2C_1]\ \text{2-substituents}$$

$$[O_2C_2C_1]\cdot^{-\chi_2}\cdot(C_2O_2H_{O2}) \overset{109.5}{----} (H_{O2}O_2C_2)\overset{1}{\text{———}}[H_{O2}O_2C_2] \text{ 2-substituents}$$

$$(C_2C_3C_4)\cdot^{-120}\cdot(C_2C_3H_3) \overset{109.5}{----} (H_3C_3C_2)\overset{1}{\text{———}}[H_3C_3C_2] \text{ 3-substituents}$$

$$(C_2C_3C_4)\cdot^{120}\cdot(C_2C_3O_3) \overset{109.5}{----} (O_3C_3C_2)\overset{1.4}{\text{———}}[O_3C_3C_2] \text{ 3-substituents}$$

$$[O_3C_3C_2]\cdot^{-\chi_3}\cdot(O_3C_3H_{O3}) \overset{109.5}{----} (H_{O3}O_3C_3)\overset{1}{\text{———}}[H_{O3}O_3C_3] \text{ 3-substituents}$$

$$(C_3C_4C_5)\cdot^{120}\cdot(C_3C_4H_4) \overset{109.5}{----} (H_4C_4C_3)\overset{1}{\text{———}}[H_4C_4C_3] \text{ 4-substituents}$$

$$(C_3C_4C_5)\cdot^{-120}\cdot(C_3C_4O_4) \overset{109.5}{----} (O_4C_4C_3)\overset{1.4}{\text{———}}[O_4C_4C_3] \text{ 4-substituents}$$

$$[O_4C_4C_3]\cdot^{-\chi_4}\cdot(C_4O_4H_{O4}) \overset{109.5}{----} (H_{O4}O_4C_4)\overset{1}{\text{———}}[H_{O4}O_4C_4] \text{ 4-substituents}$$

$$[C_5C_4C_3]\cdot^{-\nu_4-120}(C_4C_5H_5) \overset{109.5}{----} (H_5C_5C_4)\overset{1}{\text{———}}[H_5C_5C_4] \text{ 5-substituent}$$

$$[C_6C_5C_4]\cdot^{-\chi_5}\cdot(C_5C_6O_6) \overset{109.5}{----} (O_6C_6C_5)\overset{1.4}{\text{———}}[O_6C_6C_5] \text{ 6-substituents}$$

$$(C_5C_6O_6)\cdot^{-120}\cdot(C_5C_6H_{6,1}) \overset{109.5}{----} (H_{6,1}C_6C_5)\overset{1}{\text{———}}[H_{6,1}C_6C_5] \text{ 6-substituents}$$

$$(C_5C_6O_6)\cdot^{120}\cdot(C_5C_6H_{6,2}) \overset{109.5}{----} (H_{6,2}C_6C_5)\overset{1}{\text{———}}[H_{6,2}C_6C_5] \text{ 6-substituents}$$

$$[O_6C_6C_5]\cdot^{-\chi_6}\cdot(C_6O_6H_{O6}) \overset{109.5}{----} (H_{O6}O_6C_6)\overset{1}{\text{———}}[H_{O6}O_6C_6] \text{ 6-substituents}$$

All the bond angles are 109.5 degrees and the improper wedge angles are ± 120 degrees, reflecting ideal tetrahedral geometry. The lengths of all the bonds in the six membered *pyranose* ring have been set at 1.5 angstroms for simplicity. A set of six endocyclic torsion angles $\nu_0, \nu_1, \ldots, \nu_5$ can be defined in a manner directly analogous to the definitions in the furanose ring. The ring closure equations can be written down in the same way; but this time if we assume the six bond lengths and six bond angles are given then we have six equations in the six unknowns $\nu_0, \ldots, \nu_5$. This system of equations has been extensively studied (see [41], [12], [13], [31], and [15]). There is an isolated solution, $(\nu_0, \nu_1, \nu_2, \nu_3, \nu_4, \nu_5) = (-60, 60, -60, 60, -60, 60)$, called the *chair* conformation. There is also a one parameter family of solutions (at which the system is degenerate) containg various combinations of *boat* and *twist-boat* conformations. We will assign our endocyclic torsion angles to their values in the chair conformation. The pyranose ring then becomes essentially rigid.

It is natural to choose $(O_5C_1C_2)$ to be the root vertex. The C_5——O_5 bond is not traversed by our Z-tree. Rather we have made O_5 part of the trunk. This is to accomodate the standard definitions of conformational angles in polysaccharides, and to simplify as much as possible the process of formation of Z-trees for such polysaccharides. The chirality at each carbon can be read off the improper wedge angles except for C_5. At C_5 we use two torsion angles with the labels $-\nu_4 \pm 120$ to attach the exocyclic substituents. The definitions of all the side chain torsion angles $\chi_1, \ldots, \chi_6$ conform to the conventions laid out in [24]. These angles are not free, since staggered conformations are preferred energetically over eclipsed conformations. Hence these angles should be chosen from the set $\{-60, 60, 180\}$. Once these choices are made, the labelled Z-tree is specific and describes glucose precisely as a 3D molecule.

6. Some Graph Theoretical Problems

Our discussion of examples emphasizes the utility of our scheme for the geometric description and analysis of 3D molecules. In this section we will discuss some theoretical questions which are surrounding this formalism.

As a result of our main theorem GZ-trees emerge as a class of graphs which is important and about which unanswered questions motivated by applications are plentiful. Our main theorem can be viewed as a theorem in combinatorial graph theory, namely if a tree subgraph Γ of $AL^2(\mathcal{G})$ satisfies the conditions (1)-(6) (listed in the statement of the main theorem) then it has $3N-6$ edges, where $N \geq 3$ is the number of vertices in $\mathcal{G}$. This conclusion follows from counting dimensions between the two diffeomorphic manifolds X_Γ and $\mathcal{S} \times G_0^{\text{edge}_0\Gamma} \times G_1^{\text{edge}_1\Gamma} \times G_2^{\text{edge}_2\Gamma}$. Our proof however does not use combinatorial methods, and it would be interesting to find a direct graph theory proof of this result (not invoking diffeomorphisms).

Since GZ-trees represent a chemically natural class of internal coordinate systems, it is desirable that they be classified. Suppose Γ_1 and Γ_2 are two GZ-trees in $AL^2(\mathcal{G})$. We say Γ_1 is *equivalent* to Γ_2, and write $\Gamma_1 \sim \Gamma_2$, if $\pi_1 \text{vert}\Gamma_1 = \pi_1 \text{vert}\Gamma_2$. This is clearly an equivalence relation. By condition (3) it is clear that whenever $\Gamma_1 \sim \Gamma_2$ we can choose $r \in \text{vert}\Gamma_1 \cap \text{vert}\Gamma_2$. In this case we also have $X_{\Gamma_1} = X_{\Gamma_2}$. By our main theorem then there is a diffeomorphism between the internal coordinate spaces $G_{\Gamma_1} \cong G_{\Gamma_2}$. If $r_1 \in \text{vert}\Gamma_1$ is not equal to r then the diffeomorphism ϕ constructed in section 2.3 should be replaced by ϕ_1. Applying our main theorem again we get a diffeomorphism $(\phi \circ \gamma)^{-1} \circ (\phi_1 \circ \gamma)$ from $\mathcal{S} \times G_{\Gamma_1}$ to itself. Thus whenever $\Gamma_1 \sim \Gamma_2$, regardless of the choice of root vertex, we get a diffeomorphism of $\mathcal{S} \times G_{\Gamma_1}$ onto $\mathcal{S} \times G_{\Gamma_2}$ which is left G_a-equivariant, and hence a diffeomorphism $G_{\Gamma_1} \cong G_{\Gamma_2}$. An example of such a pair of GZ-trees is suggested by Figure 14. It shows a close up of a portion of $AL^2(\mathcal{G})$ in a region where the Z-tree Γ and the cyclic subgraph from Figure 13 are in close proximity. The rectangles

$$\begin{array}{ccccc}
(C_{3'}C_{4'}C_{5'}) & \xrightarrow{T_0(l_4)} & [C_{3'}C_{4'}C_{5'}] & \overset{T_2(e^{-i\delta})}{\cdots} & (C_{4'}C_{3'}O_{3'}) \\
T_2(e^{-i\phi(\tau_4)}) \;\vdots & & T_2(e^{i\phi(\tau_4)}) \;\vdots & & T_2(e^{-i\phi(\tau_3)}) \;\vdots \\
(C_{3'}C_{4'}O_{4'}) & \xrightarrow{T_0(l_4)} & [C_{3'}C_{4'}O_{4'}] & \overset{T_2(e^{-i\nu_3})}{\cdots} & (C_{4'}C_{3'}C_{2'})
\end{array}$$

commute, if we agree that the vertical edges of type 2 are oriented from top to bottom, and the horizontal from left to right. The leftmost and rightmost vertical edges and the two top horizontal edges are in Γ. However, it is natural to consider a different subgraph Γ' inside $AL^2(\mathcal{G})$, where the two top horizontal edges (from Γ) are replaced in Γ' by the two bottom horizontal edges. It is clear that $\Gamma \sim \Gamma'$. The labels of the edges of Γ' will provide an equivalent set of internal coordinates as the labels of the edges of Γ. However, Γ' is not a Z-tree! A clearer example can be seen in Figure 16. The middle figure shows a GZ-tree which is not a Z-tree. The first and third figures are distinct Z-trees, which are nevertheless equivalent. It would be interesting to classify GZ-trees (or Z-trees) up to equivalence. Is there a GZ-tree which is not equivalent to any Z-tree? If Γ is a GZ-tree (or a Z-tree) is there an algorithm for finding all the GZ-trees (or Z-trees) which are equivalent to Γ?

Since GZ-trees are important, it is worthwhile to examine other sets of graph theoretical axiom sets which are equivalent to the conditions (1)-(6) we have given.

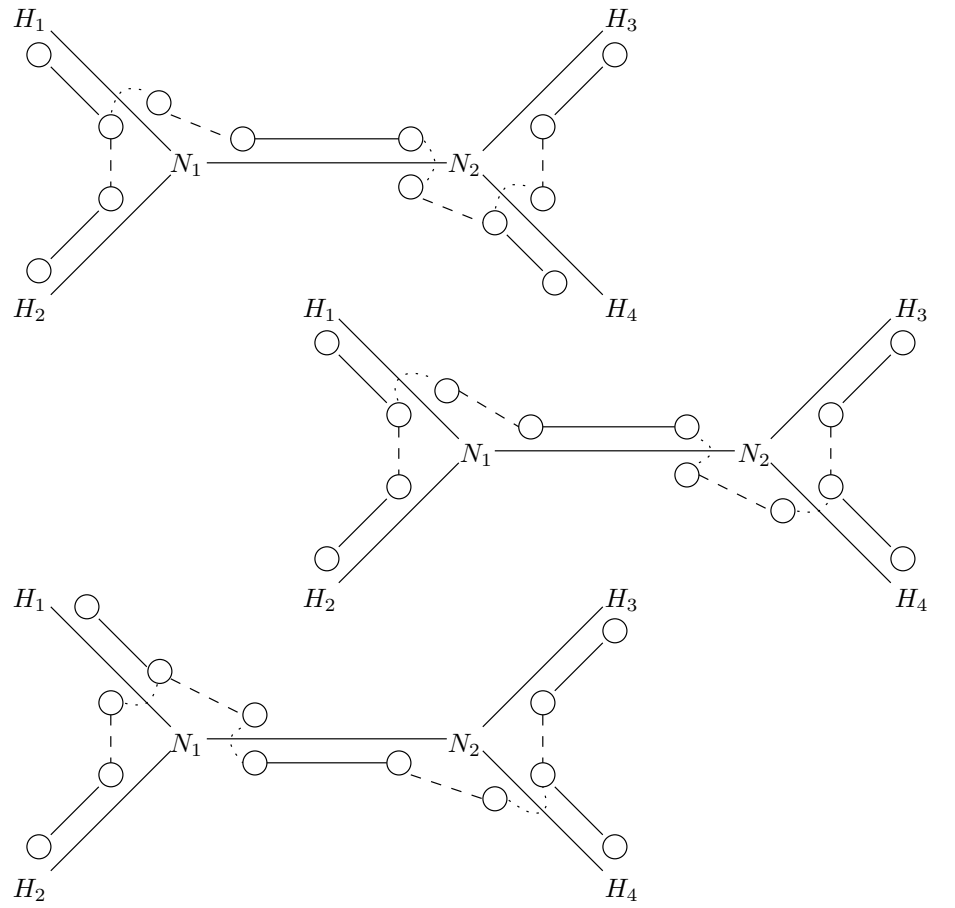

FIGURE 16. Three GZ-trees for the molecule N_2H_4. The first and third are Z-trees, and the second is not a Z-tree. All three GZ-trees are equivalent. The first and the second differ only by a single edge of type 0.

For example, is it true that if a tree subgraph satisfies condition (1) and has $3N-6$ edges then it must be a GZ-tree? Another aspect of our conditions (1)-(6) is that we can easily test a subgraph to see if it is a GZ-tree. But Z-trees lend themselves to a list presentation. Is there a generalization of the list presentation for Z-trees that can express the most general GZ-tree? For example, if we allow more than one trunk (consists of edges of type 0, 1, and 0) and also allow *swivels* (three edges, of types 2, 1, and 2), then we seem to be able to present in list form any GZ-tree that we have looked at thus far. The second GZ-tree in Figure 16 has two trunks which are connected by a *linker*, which is an ordinary branch (consists of edges of types 2, 1, and 0) followed by swivel. We wonder if an arbitrary GZ-tree can be decomposed as a family of Z-trees which are connected by linkers. Linkers appear naturally when one studies polymerization of 3D molecules. We intend to study this process in more detail in subsequent work.

As we pointed out in section 3.1 our main theorem allows us to describe a large open subset X_Γ of all the molecular configurations using a subgraph Γ of $AL^2(\mathcal{G})$. It would be of interest to cover all the exceptional configurations by considering several such subgraphs. For which molecular graphs $\mathcal{G}$ (if any) is this possible?

References

[1] R. Abraham, J.E. Marsden, *Foundations of Mechanics*, second edition, Benjamin/Cummings, London, 1978.

[2] W.J. Adams, H.J. Geise, L.S. Bartell, *Structure, Equilibrium Conformation, and Pseudorotation in Cyclopentane. An Electron Diffraction Study*, J. Am. Chem. Soc., **92**, no. 17, 5013–5019, 1970.

[3] S.L. Altmann, *Rotations, Quaternions, and Double Groups*, Clarendon Press, Oxford, 1986.

[4] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

[5] J. Baker, A. Kessi, B. Delley, *The generation and use of delocalized internal coordinates in geometry optimization*, J. Chem. Phys., **105**, no. 1, 192–212, 1 Jul 1996.

[6] A.T. Balaban (Ed.), *Chemical Applications of Graph Theory*, Academic Press, London, 1976.

[7] A.T. Balaban (Ed.), *From Chemical Topology to Three-Dimensional Geometry*, Plenum Press, New York, 1997.

[8] G.M. Blackburn, M.J. Gait, (eds.), *Nucleic Acids in Chemistry and Biology*, IRL Press, Oxford, 1990.

[9] C. Brandon, J. Tooze, *Introduction to Protein Structure*, Garland Publishing, Inc., New York, 1991.

[10] T. Carrington, Jr., *Vibrational Energy Level Calculations*, in the Encyclopedia of Computational Chemistry, (P.v.R. Schleyer, ed.), Volume 5, 3157–3166, John Wiley and Sons, Chichester, 1998.

[11] *Cerius*^{2} *Users Guide, Quantum Mechanics, Quantum Chemistry, ADF, Gaussian, MOPAC*, Molecular Simulations, Inc., San Diego, 1997.

[12] G.M. Crippen, T.F. Havel, *Distance Geometry and Molecular Conformation*, Research Studies Press, Taunton, Somerset, England, 1988.

[13] G.M. Crippen, *Exploring the Conformation Space of Cycloalkanes by Linearized Embedding*, J. Computational Chem., **13**, no. 3, 351–361, 1992.

[14] A. Dress, A. Dreiding, H. Haegi, *Classification of Mobile Molecules by Category Theory*, in Symmetries and Properties of Non-Rigid Molecules: A Comprehensive Survey, J. Maruani and J. Serre (eds.), Studies in Physical and Theoretical Chemistry, Vol. 23, 39–58, Elsevier, Amsterdam, 1983.

[15] I.Z. Emiris, B. Mourrain, *Computer Algebra Methods for Studying and Computing Molecular Conformations*, Algorithmica, **25**, 372–402, 1999.

[16] E. Estrada, N. Guevara, I. Gutman, L. Rodriguez, *Molecular connectivity indices of iterated line graphs. A new source of descriptors for QSPR and QSAR studies*, SAR and QSAR in Environmental Research, **9**, no. 3-4, 229–240, 1998.

[17] J.B. Foresman, A. Frisch, *Exploring Chemistry with Electronic Structure Methods*, second edition, Gaussian, Inc., Pittsburgh, 1996.

[18] N. Gō, H.A. Scheraga, *Ring Closure and Local Conformational Deformations of Chain Molecules*, Macromolecules, **3**, no. 2, 178–187, 1970.

[19] M.S. Gordon, *Applications of Approximate Molecular Orbital Theory to Organic Molecules*, Ph.D. Dissertation (directed by J.A. Pople), Department of Chemistry, Carnegie-Mellon University, 1968.

[20] A. Guichardet, *On rotation and vibration motions of molecules*, Ann. Inst. Henri Poincaré-Physique théorique, **40**, 329–342, 1984.

[21] T.F. Havel, I. Najfeld, *Applications of geometric algebra to the theory of molecular conformation 2. The local deformation problem*, Theochem- J. Mol. Struc., **336**, 175–189, 1995.

[22] W.J. Hehre, L. Radom, P.R. v. Schleyer, J.A. Pople, *Ab initio Molecular Orbital Theory*, Wiley, New York, 1986.

[23] IUPAC-IUB Commission on Biochemical Nomenclature, *Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains*, Biochemistry, **9**, 3471–3479, 1970.

[24] IUPAC-IUB Joint Commission on Biochemical Nomenclature, *Symbols for the Specifying the Conformation of Polysaccharide Chains*, Eur. J. Biochem., **131**, 5–7, 1983.

[25] IUPAC-IUB Joint Commission on Biochemical Nomenclature, *Abbreviations and Symbols for the Description of Conformations of Polynucleotide Chains*, Eur. J. Biochem., **131**, 9–15, 1983.
[26] N.P. Landsman, *Mathematical Topics Between Classical and Quantum Mechanics*, Springer, New York, 1998.
[27] H.P.M. de Leeuw, C.A.G. Haasnoot, C. Altona, *Empirical Correlations Between Conformational Parameters in β-D-Furanoside Fragments Derived from a Statistical Survey of Crystal Structures of Nucleic Acid Constituents*, Israel J. Chem., **20**, 108–126, 1980.
[28] M. Levitt, A. Warshel, *Extreme Conformational Flexibility of the Furanose Ring in DNA and RNA*, J. Am. Chem. Soc., **100**, no. 9, 2607–2613, 1978.
[29] H. Lodish, D. Baltimore, A. Berk, S.L. Zipursky, P. Matsudaira, J. Darnell, *Molecular Cell Biology*, Scientific American Books, Inc., New York, 1995.
[30] A.D. MacKerell Jr., N. Foloppe, *All-Atom Empirical Force Field for Nucleic Acids*, J. Computational Chem., **21**, 86–120, 2000. See also the website: http://www.pharmacy.umaryland.edu/˜alex/research.html
[31] D. Manocha, Y. Zhu, W. Wright, *Conformational analysis of molecular chains using nano-kinematics*, Computer Applications in the Biosciences, **11**, no. 1, 71–86, 1995.
[32] C.K. Matthews, K.E. Van Holde, *Biochemistry*, second edition, Benjamin/Cummings, Menlo Park, 1996.
[33] A.K. Mazur, R.A. Abagyan, *New Methodology for Computer-aided Modelling of Biomolecular Structure and Dynamics 1. Non-cyclic Structures, 2. Local Deformations and Cycles*, J. Biomolecular Struct. Dynamics, **6**, no. 4, 815–845, 1989.
[34] P.G. Mezey, *Potential Energy Hypersurfaces*, Studies in Physical and Theoretical Chemistry, Vol. 53, Elsevier, Amsterdam, 1987.
[35] A. Neumaier, *Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure*, SIAM Rev., **39**, no. 3, 407–460, 1997.
[36] See the website, http://ndb-mirror-2.rutgers.edu/NDB/archives/proj/valence/index.html
[37] D. Plavšić, M. Šoškić, Z. Daković, I. Gutman, A. Graovac, *Extension of the Z Matrix to Cycle-Containing and Edge-Weighted Molecular Graphs*, J. Chem. Inf. Comput. Sci., **37**, 529–534, 1997.
[38] P. Pulay, G. Fogarasi, *Geometry Optimization in Redundant Internal Coordinates*, J. Chem. Phys., **96**, no. 4, 2856–2860, 15 Feb 1992.
[39] C.C. Pye, R.A. Poirier, *Graphical Approach for Defining Natural Internal Coordinates*, J. Comp. Chem., **19**, no. 5, 504–511, 1998.
[40] J.R. Quine, *Helix Parameters and Protein Structure using Quaternions*, Theochem- J. Mol. Struc., **460**, 53–66, 1999. (See also the web site: http://www.math.fsu.edu/˜quine for information on *Discrete Frenet Frames*).
[41] R. Randell, *A Molecular Conformation Space*, 125–140, and *Conformation Spaces of Molecular Rings*, 141–156, in MATH/CHEM/COMP 1987, R.C. Lacher (ed.), Studies in Physical and Theoretical Chemistry, Vol. 54, Elsevier, Amsterdam, 1988.
[42] W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984.
[43] N. Trinajstić, *Chemical Graph Theory*, Vol. I and II, CRC Press, Boca Raton, 1983.
[44] D.B. West, *Introduction to Graph Theory*, Prentice Hall, Upper Saddle River, NJ, 1996.
[45] E.B. Wilson, J.C. Decius, P.C. Cross *Molecular Vibrations*, McGraw-Hill, New York, 1955.

Department of Mathematics, University of South Carolina
E-mail address: dix@math.sc.edu