Inverse protein folding in 3D hexagonal prism lattice under HPC model

Alireza Hadj Khodabakhshi, Ján Maňuch*, Arash Rafiey and Arvind Gupta
School of Computing Science, Simon Fraser University
8888 University Dr, Burnaby BC, V5A 1S6 Canada

Abstract

The inverse protein folding problem is that of designing an amino acid sequence which has a prescribed native protein fold. This problem arises in drug design where a particular structure is necessary to ensure proper protein-protein interactions. In Gupta et al. (2007), a tubular structures for 3D hexagonal prism lattice were introduced and their stability was formally proved for simple instances under the HP model of Dill. In this paper, we generalize the design of tubular structures to allow for much larger variety of designable structures by allowing branching of tubes. Our generalized design could be used to approximate given 3D shapes in the considered lattice. Although the generalized tubular structures are not stable under the HP model we can prove that a simple instance of generalized tubular structures is structurally stable (all native folds have the designed shape) under a refined version of HP model, called HPC model. We conjecture that all generalized tubular structures are structurally stable under the HPC model.

1 Introduction

It has long been known that protein interactions depend on their native three-dimensional fold and understanding the processes and determining these folds is a long standing problem in molecular biology. The most significant force acting on protein folding are hydrophobic interactions (see Dill (1990) for details). This led Dill to introduce the Hydrophobic-Polar model Dill (1985). Here the 20 amino acids from which proteins are formed are replaced by two types of monomers: hydrophobic or polar, depending on their affinity to water. To simplify the problem, the protein is laid out on vertices of a lattice with each monomer occupying exactly one vertex and neighboring monomers occupying neighboring vertices. The free energy is minimized when the maximum number of non-consecutive hydrophobic monomers are adjacent in the lattice. Therefore, the “native” folds are those with the maximum number of such HH contacts. Even though the HP model is the simplest model for protein folding, computationally it is an NP-hard problem, cf. Crescenzi et al. (1998) for two- and Berger and Leighton (1998) for three-dimensional square lattices.

Another significant force in the folding process of the proteins are disulfide bridges between two cysteine monomers which play an important role in the stability of the protein structures Jaenicke (1991). In Hadj Khodabakhshi et al. (2008) we extended the HP model by considering a third type of monomers, cysteines, and incorporating disulfide bridges between two cysteines into the energy model. This model is called the hydrophobic-polar-cysteine (HPC) model. The cysteine monomers in the HPC model act as hydrophobic Naganoa et al. (1999), but in addition two neighboring cysteines can form a disulfide bridge to further reduce the energy of the fold.

In many applications such as drug design, we are interested in the complement problem to protein folding: inverse protein folding or protein design. The hardness of the inverse protein folding under the HP model is still unknown but is conjectured to be NP-hard. A major challenge in protein design is to avoid proteins that have multiple native folds. We say that a protein is stable if its native fold is unique. Furthermore, we say that a protein is structurally stable if all its native folds define the same mapping from the set vertices of the lattices to the set of amino acids, i.e., if all native folds appear to be identical and only differ by the peptide
connections. In Gupta et al. (2005) a new version of the inverse protein folding problem was considered: instead of a target fold, a target structure (a connected set of lattice vertices) is given, and the goal is to design a sequence which would (preferably uniquely) fold into a structure (picked from a rich class of “constructible” structures) “close” to the target structure. The 2D square lattice was used and it was shown that all designed proteins fold into corresponding constructible structures. It was also “formally” shown that the proteins for the simplest (but arbitrary long) constructible structures fold uniquely, and conjectured that the same holds for all constructible structures. Design of stable proteins of arbitrary lengths in the HP model was also studied in Aichholzer et al. (2003) (for 2D square lattice) and in Li et al. (2005) (for 2D triangular lattice), motivated by a popular paper of Hayes (1998).

![figure](image.png)

Figure 1: An example of (a) hexagonal prism lattice; (b) a tubular structure built with 3 tubes. Hydrophobic (polar) monomers are depicted with black (white) beads.

In Gupta et al. (2007), the 3D lattice (hexagonal prism lattice, cf. Figure 1(a)) was used to design a class of tubular structures and their corresponding proteins. It was shown that each protein folds into the corresponding tubular structure, and that the proteins for the smallest tubular structures (with up to two tubes) are structurally stable under the HP model, however in the two tubes case, the paper missed one case in which protein for the tubular structure with two tubes fold into a very similar structure, and hence, it is not completely structurally stable. An example, of a tubular structure is shown in Figure 1(b). The shortcoming of this design is that it only allows to chain tube in linear fashion which severely limits ability of design to approximate given shapes.

In this paper, we generalize the design introduced in Gupta et al. (2007) by adding a new building block: a “connector”. The hydrophobic core of the connector consists of 2 layers of two adjacent hexagons. The connector can be attached to 4 tubes (one per top/bottom of each hexagon). We call these structures generalized tubular structures. An example, with 3 tubes attached to the connector is shown in Figure 2. Such design is sufficiently robust to approximate any given shape.

We show that a generalized tubular structure is one of the native folds of its protein under the HPC model. We conjecture that the proteins of the generalized tubular structures are structurally stable, i.e., it is possible to choose hydrophobic monomers which are cysteines in all generalized tubular structures such that their proteins fold uniquely into designed structures. We are able to prove this formally for infinite subclass of the simple structures (consisting of one connector and three tubes, cf. Figure 2) under the assumption that each of three tubes is sufficiently long. In addition, similar to Gupta et al. (2007), we assume that our proteins are closed chains of monomers, i.e., that the beginning and the end of the sequence are adjacent in the lattice. Note that generalized tubular structures from this subclass are not structurally stable under the HP model, thus our results show that adding disulfide bridges into the energy model helps to stabilize our designed proteins.

Despite the tremendous amount of work on protein design for 2D lattices, as far as we know, this is the first general design of arbitrary long stable proteins for the 3D lattice. Given that 3D is the realistic setting, we believe that this work could eventually help in designing proteins with applications to drug design and nanotechnology.
2 Preliminaries

In this section we will review the HPC model and introduce some terminology used in the paper.

2.1 Hydrophobic-polar-cysteine model

In HPC model, proteins are chains of monomers where each monomer is either hydrophobic-non-cysteine, cysteine or polar. Such a chain is represented as a string $p = p_1 p_2 \ldots p_{|p|}$ in \{0, 1, 2\}*, where “0” represents a polar monomer (depicted in figures as empty circles), “1” a hydrophobic-non-cysteine (depicted as black squares) and “2” a cysteine monomer (depicted as black triangles). We use $H$ to represent a monomer which could be either 1 or 2 (depicted in figures as a black circle). The proteins are folded onto the regular lattice. A fold of a protein $p$ is embedding of a path of length $n$ into the lattice.

In our 3D HPC model we use the hexagonal prism lattice as a lattice structure. The vertices adjacent to a vertex are called the neighbors of that vertex. As depicted in Figure 1(a), each vertex has 5 neighbors: 3 horizontal neighbors lying in the same hexagonal grid and 2 vertical neighbors lying above and below the vertex in the parallel hexagonal grids.

A protein will fold into a conformation with the minimum free energy, also called a native fold. The energy function in the HPC model consists of two parts: hydrophobic interactions and disulfide bridges. The hydrophobic monomers which are not consecutive in the protein but are adjacent in the lattice form contacts. Each contact contributes $-1$ to the total energy. The cysteine monomers act as hydrophobic for this part of energy function. In addition to hydrophobic interactions each pair of cysteines which are not consecutive in the protein but are adjacent in the lattice can form a disulfide bridge to further reduce the energy of the fold. Unlike the hydrophobic interactions in which a hydrophobic monomer can take part in several contacts, each cysteine can only participate in one disulfide bridge. Therefore, the number of disulfide bridges contributing to the energy of a fold is equal to the number of pairs in the maximum matching in the graph of potential disulfide bridges. Each disulfide bridge contributes $-1$ to the total energy\(^1\). Hence, a fold with the lowest free energy corresponds to a fold with the largest number of HH contacts and disulfide bridges.
2.2 Stability and structural stability

Note that there might be several native folds for a given protein. A protein with a unique native fold is called stable protein. Every protein and its fold define a partial mapping from the lattice vertices to the set \{0, 1, 2\}. We say that two folds of the same protein are structurally similar if they define the same mapping (up to translations and rotations). If all native folds of a given protein are similar to each other, the protein is called structurally stable. Note that all native folds of a structurally stable protein have exactly the same shape (from outside they appear as the same fold). For instance, the string \(t = (0100110010)^6\) is structurally stable, but not stable. Figure 3 depicts both native folds of this string. It is easy to see that the mappings defined by \(t\) and its two folds are identical, i.e., the folds are structurally similar.

2.3 Terminology

A lattice vertex containing an \(X \in \{0, 1, 2\}\) monomer is called an \(X\)-vertex. An \(H\)-vertex is either a 1-vertex or a 2-vertex. A neighbor of a vertex \(v\) which is an \(X\)-vertex is called \(X\)-neighbor.

Consider a fold \(F\). A path in \(F\) is a sequence of vertices \((x_1, x_2, \ldots, x_k)\) such that consecutive vertices are connected by peptide bonds. We say that \(F\) contains an occurrence of substring \(w_1, w_2, \ldots, w_k\) if there is a path \((x_1, x_2, \ldots, x_k)\) in \(F\) such that \(x_i\) is a \(w_i\)-vertex.

We number hexagonal grids of the lattice (also referred to as planes) with integer numbers, and denote the \(i\)-th hexagonal grid by \(P_i\). Consider vertex \(x \in P_i\). We denote the vertical neighbor of \(x\) in \(P_{i+1}\) (above \(x\)) by \(x^1\), and recursively, the vertical neighbor of \(x^j\) in \(P_{i+j+1}\) by \(x^{j+1}\). Similarly, we denote the neighbor of \(x\) in \(P_{i-1}\) by \(x^{-1}\), and the neighbor of \(x^{-j}\) in \(P_{i-j-1}\) by \(x^{-j-1}\).

Let \(G_x\) be the graph of all \(H\)-vertices in \(P_i\) which are reachable from \(x\) by a path of \(H\)-vertices in \(P_i\). Let \(G\) be a set of vertices in \(P_i\). Then for \(j \geq 1\), let \(G^j\) be the graph of all vertices in \(P_{i+j}\) which have a neighbor in \(G^{j-1}\), and \(G^{-j}\) be the graph of all vertices in \(P_{i-j}\) which have a neighbor in \(G^{j+1}\), i.e., \(G^j, j \neq 0\), are vertical copies (translations) of the set \(G\). Note that \(G_x\) is a planar graph (as \(P_i\) is as well). The degree of a vertex in \(G_x\) is called a plane degree. Let \(B_x\) be the boundary cycle of \(G_x\), i.e., the set of vertices of \(G_x\) which lie on the outer face of \(G_x\). An \(H\)-component in a fold \(F\) is a maximal set of \(H\)-vertices for which there is a path of \(H\)-vertices between any pair of them.

Let \(C\) be an \(H\)-component that lies in the planes \(P_{j+1}\) to \(P_{j+r}\). Let layer \(C_i\) be a set of all vertices of \(C\) in plane \(P_{j+i}\). We say that layers \(C_i\) and \(C_k\) are the same (\(C_i\) is a subset of \(C_k\) if \(C_{k-i} = C_k \cap C_{i-k} \subseteq C_k\)). In this case we write \(C_i \simeq C_k\). The plane containing \(C_i\) will be denoted by \(P(C_i)\).

2.4 Saturated folds

The proteins used in Gupta et al. (2005) and the proteins we will use in our design have a special property. The number of possible with respect to the number of hydrophobic “I” and contacts and disulfide bridges

\[\text{Note that our results are not based on the assumption that the energy contributions of a hydrophobic bond and a disulfide bridge are equal, they are valid for any ratio between these two energies, as long as they are both negative.}\]
of their native folds is maximal cysteine “2” monomers contained in the protein. The following useful observation characterizes native folds of such proteins.

**Observation 1 (Saturated folds).** Let \( p \in 0\{0,1,2\}^\ast 0 \) be a protein, and \( F \) be the fold of \( p \). If for every H-vertex \( v \), three out of five edges incident with \( v \) are contacts and in addition if \( v \) is a cysteine it belongs to a maximum matching in the graph of potential disulfide bridges, then (a) \( F \) is a native fold of \( p \); and (b) any other native fold of \( p \) satisfies these properties. We will call a fold satisfying these properties a saturated fold.

The proof of the observation follows by a simple argument that any hydrophobic vertex \( v \) can have at most three contacts since it is connected to exactly two neighbors with a peptide bond and furthermore any cysteine monomer can be involved in at most one disulfide bridge. Note that not every protein has a saturated fold.

### 3 Generalized tubular structures and their proteins

![Figure 4: Illustrations of (a) a tube with a hydrophobic core of height 8 — the wavy lines at the top and dashed lines at the bottom represent loops; (b) a connector.](image)

The first basic building block of our generalized tubular structures is a **tube**, depicted in Figure 4(a). Tubes were the only building block of tubular structures introduced in Gupta et al. (2007). A tube consists of 6 identical “alpha helix”-like subfolds of the substring \( p_n = (H00H)^n \) forming a \( 2 \times 2n \) vertical zig-zag pattern (“plate”).

The plates are connected to each other with 6 short loops (3 at the top and 3 at the bottom), each consisting of only two polar monomers. Thus, the hydrophobic core is completely surrounded by polar monomers, i.e., the fold is saturated. The complete protein string for the tube is \( t_n = (0p_n0)^6 \). We assign the first and the second H monomer of one of the plates of each tube with cysteine monomers 2. We represent the fold of \( t_n \) by \( T_n \). The height of the hydrophobic core of the tube \( T_n \) is 2n.

The second building block of our generalized tubular structures is a **connector**, depicted in Figure 4(b). A connector can be formed by overlapping two very short tubes (with height of hydrophobic core 2). Two tubes or a tube and a connector can be connected to one protein structure in two different ways:

1. In the first way, one top loop of the first tube is overlapped with a bottom loop of the second tube/connector, vice versa, and the peptide bonds between two polar monomers of each loop are disconnected. This way of connecting two H-components is called vertical connection. Tubes in Figure 5(a) are vertically connected. Tubes \( T_1 \) and \( T_2 \) in Figure 2 are vertically connected to the connector.

2. In the second way, called horizontal connection, the tubes or the tube and the connector are placed beside each other such that they have H-vertices in exactly one common plane \( P_1 \) and exactly two H-vertices of the first H-component are connected to two H-vertices of the other H-component each through one 0-vertex. Tubes in Figure 5(b) are horizontally connected and tube \( T_3 \) in Figure 2 is horizontally attached to the connector.
Figure 5: Two possible connections of two tubes: (a) vertical; and (b) horizontal. The dashed lines shows the part which are folded differently.

Note that both connections define the same protein sequence, thus in the HP model any generalized structure with at least two building blocks is not structurally stable. To stabilize the structures in the HPC model, the two cysteines of the attached tube to the existing structure should be directly adjacent to the connection, cf. Figure 2. The class of generalized tubular structures contains all structures obtained by interconnecting several connectors and tubes such that no space violation occurs. Since the folds of generalized tubular structures are saturated, by Observation 1, they are native folds to corresponding proteins (which can be easily reconstructed from the folds).

4 On stability of generalized tubular structures

In what follows we will show that the protein of one basic generalized tubular structure: the structure built from one connector and three tubes, cf. Figure 2, is structurally stable. We will assume that three tubes $T_{k_1}, T_{k_2}, T_{k_3}$ used to construct this structure are sufficiently long. In particular, we will assume that $k_1, k_2, k_3 \geq 712$. We conjecture that this structure is structurally stable also for other values of $k_1, k_2, k_3$ and that all generalized tubular structures are structurally stable. Let $q$ be the protein string of this structure and $Q$ be its original fold.

Definition 1 (Sparse protein). We say that a protein is sparse if does not contain HHH as a substring and does not start or end with H.

4.1 Types of $H$-vertices

Let $F$ be a saturated fold of a sparse protein. Then each $H$-vertex has at exactly three contacts, i.e., it has at least three $H$-neighbors and the remaining two neighbors are connected (via a peptide bond) and at most one of the two is an $H$-vertex. We can classify every $H$-vertex $x$ of $F$ to one of the five types based on the position of its $0$-neighbor(s), cf. Figure 6:

(a) $\text{vh-type: } x$ has one vertical $0$-neighbor (on top or below) and one horizontal $0$-neighbor (in the same hexagonal grid);
(b) $\text{vv-type: } x$ has two vertical $0$-neighbors;
(c) $\text{hh-type: } x$ has two horizontal $0$-neighbors;
(d) $\text{h-type: } x$ has one horizontal $0$-neighbor and one $H$-neighbor (in the same hexagonal grid);
(e) $\text{v-type: } x$ has one vertical $0$-neighbor and one $H$-neighbor (in the same hexagonal grid).

Figure 6: Five types of possible neighborhood of an $H$-vertex $x$: $S$-vertices: (a) $\text{vh}$, (b) $\text{vv}$, (c) $\text{hh}$; and $D$-vertices: (d) $\text{h}$ and (e) $\text{v}$.
Let $v, y$ be $s$-connected if there is a path $y_1, y_2, \ldots, y_k, y$ in the lattice such that $v_1$ is an $s$-vertex. If $x$ is a $u$-vertex and $y$ is a $v$-vertex, this path is called an $usu$-connection. In particular, we will be interested in $H0H$-connections and $(S \approx D)$-connections. A $usu$-connection with end points $x$ and $y$ is called internal, if $x$ and $y$ are in the same $H$-component, and otherwise it is called external. We say that two $usu$-connections with end points at $x$ and $y$, respectively, are parallel if $x(y)$ is directly above/below $x'(y')$, i.e., $x = x' = y = y'$, for some integers $i, j$, and all vertices between $x$ and $x'(y$ and $y'$) are $H$-vertices. Note that it is also possible that $x$ and $y'$ are $u$-vertices and $x'$ and $y$ are $v$-vertices.

We have the following observations:

**Observation 2.** Let $F$ be an arbitrary saturated fold of $q$. Then $F$ contains 6 $H0H$-connections, 52 $S$-vertices, the number of $D$-vertices is 4 modulo 6 and it contains 36 $(S \approx D)$-connections. Furthermore, $q(F)$ does not contain $HHH$, $000$, $H0HOH$ and $H0HH$, but it does contain one occurrence of $20100101$.

**Observation 3.** Let $F$ be a saturated fold of a sparse protein. Then every $H$-vertex of $F$ is either a $vh$-vertex, $vv$-vertex, $hh$-vertex, $h$-vertex or $v$-vertex. Furthermore, any neighboring $0$-vertex and $H$-vertex are connected by a peptide bond.

**Lemma 4.** Let $F$ be a saturated fold of a sparse protein with no $H0HH$ as a substring. Then no $v$-vertex in $F$ can connect directly to an $h$-vertex.

**Proof.** Consider a $v$-vertex $x$. Without loss of generality assume that its 0-neighbor is $x^1$. Assume to the contrary that $x$ connects to an $h$-vertex. Two cases are possible: first, $x$ connects to its horizontal $h$-neighbor $z$ cf., Figure 7(a). Then $z^1, x^1, z$ form the substring $H0HH$, a contradiction. Second, $x$ connects $x^{-1}$ which is an $h$-vertex. Let $z$ be the horizontal 0-neighbor of $x^{-1}$. Then $z^1, z, x^{-1}, x$ form the substring $H0HH$, a contradiction cf., Figure 7(b).

**Lemma 5.** Let $F$ be a saturated fold of a sparse protein. No $v$-vertex can connect to an $h$-vertex via two 0-vertices.

**Proof.** Consider a $v$-vertex $x$. Without loss of generality assume that its 0-neighbor is $x^1$. Assume to the contrary that $x^1$ connects to an $h$-vertex via one 0-vertex $y$. If $y$ is a horizontal neighbor of $x^1$ then it would connect down to an a vertex which is not an $h$-vertex. Hence, $y$ must be $x^2$. Vertex $x^2$ should connect to an $h$-vertex, hence it cannot connect to $x^3$. Assume it connects to one of its horizontal neighbor $z$. Since $z$ is an $h$-vertex, $z^{-1}$ is an $H$-vertex. However, this a contradiction, as $x^1$ would have to connect to three vertices: $x, x^2$ and $z^{-1}$ Figure 7(c).

Lemmas 4 and 5 imply the following claim.

**Lemma 6.** Let $F$ be a saturated fold of a sparse protein with no $H0HH$ as a substring. Any occurrence of substring $(00HH)^k$ in $F$ contains either only $v$-vertices or only $h$-vertices.
4.2 Types of $H$-components

In this section we study all possible $H$-components that can arise in saturated folds of $q$. We first classify all $H$-components to three categories and then study which of these can appear in saturated folds of $q$.

Let $F$ be a saturated fold of a sparse protein and $C$ an $H$-component in $F$. Assume that $C$ lies in the planes $P_s, \ldots, P_e$. Note that any $H$-vertex of plane degree one in the first or last layer of $C$ is adjacent to at least three $0$-vertices, a contradiction. Hence, we have the following observation.

Observation 7. Let $F$ be a saturated fold of a sparse protein and let $C$ be an $H$-component in $F$. Then all vertices of the first or last layer of $C$ have plane degree 2 or 3.

The following definition defines several types of $H$-components.

Definition 3 (Tube, simple tube, 2-layer $H$-component, wall, and complex $H$-component). A tube is an $H$-component in which all layers are identical and each layer contains only vertices of plane degree two (a cycle). A simple tube is a tube with only one hexagon in each layer. A 2-layer $H$-component is an $H$-component with two identical layers which have no vertex with plane degree 1 and at least one vertex with plane degree 3. A wall is an $H$-component with all its layers identical and each layer a single path. Finally, a complex $H$-component is an $H$-component $C$ such that there is some $i$ for which $C_i$ and $C_{i+1}$ are different.

We have the following observations.

Observation 8. Any $H$-component $C$ in a saturated fold of a protein is one of the following three types: a tube, a 2-layer $H$-component or a complex $H$-component.

4.3 Different types of complex $H$-components

In what follows we further classify different types of complex $H$-components which can occur in saturated folds of sparse proteins with at most six occurrences of substring $H0H$. 

Figure 7: Case analysis showing that a $vh$-vertex cannot directly (a) and (b); or via two $0$-vertices (c) connect to an $h$-vertex.

Figure 8: One layer of (a) the smallest non-simple tube; (b) the smallest non-simple tube without occurrences of $H0H$; and (c) the smallest non-simple tube with one occurrence of $H0H$ per layer.

Observation 9. Let $F$ be a saturated fold of a sparse protein. If $F$ contains a tube then the height (number of layers) of this tube is at least 2. One layer of the smallest non-simple tube is depicted in Figure 8(a). It contains two occurrences of $H0H$ per layer, i.e., at least 4 such occurrences. One layer of the smallest non-simple tube with no occurrences of $H0H$ is depicted in Figure 8(b). One layer of the smallest tube with two occurrences of $H0H$ per layer is depicted in Figure 8(c).
4.3.1 Complex H-components with a vv-vertex

![Diagram of a complex H-component with a vv-vertex. The arrows are pointing at six vv-vertices.](image)

Figure 9: Part of a complex H-component with a vv-vertex. The arrows are pointing at six vv-vertices.

**Lemma 10.** Let $F$ be a saturated fold of a sparse protein with no occurrences of substrings H0HH and H0HOH and at most six occurrences of substring HOH. Consider a complex H-component $C$ of $F$ containing a vv-vertex. Then $C$ has 6 vv-vertices forming a hexagon, lies in two layers which are almost identical, except for the six vv-vertices which are replaced with 0-vertices in the other layer, and neither layer contains a vertex of plane degree 1. We will call such a complex H-component, a vv-H-component. A vv-H-component contains 6 occurrences of H0H.

**Proof.** Any vv-vertex must be adjacent to at least two other vv-vertices in its plane, otherwise, there is a 0-vertex connected to three H-vertices (with a peptide bond), or we get a substring H0HOH which cannot occur in $F$. Therefore, any set of vv-vertices in a plane forms a graph with no vertices of plane degree 1. Each vv-vertex on the boundary of this graph is adjacent to one non-vv-vertex which creates a distinct H0H substring. Since there are only 6 occurrences of H0H in $F$, the boundary of this graph must be a hexagon, i.e., $C$ contains exactly 6 vv-vertices $x_1, \ldots, x_6$ located on a single hexagon, cf. Figure 9. Furthermore, $C$ does not contain a vertex with plane degree 1. Assume to the contrary that $v$ is a vertex with plane degree 1 and let $k$ be the smallest number such that $v^k$ is a vertex with plane degree more than 1 (note that such a $k$ exists). Let $w$ be a horizontal H-neighbor of $v^k$. Now, the path $(w, w^{-1}, v^{k-1})$ is an H0H-connection which is different from the H0H-connections containing the vv-vertex of $F$, a contradiction.

For $i = 1, \ldots, 6$, let $y_i$ be the non-vv horizontal neighbor of $x_i$. Consider $y_1$. One of its vertical neighbors is an H-vertex while the other is a 0-vertex, cf. Figure 9. Without loss of generality assume $y_1^2$ is an H-vertex. Let $z_1$ be the horizontal neighbor of $y_1$ which is closer to $y_2$. Since $C$ does not contain any vertex of plane degree 1, all the horizontal neighbors of $y_1^2$ except $x_1^1$ are H-vertices. In addition, $y_1^2$ must be a 0-vertex otherwise, $F$ would contain the substring H0HH, a contradiction. It follows that $z_1$ is an H-vertex and $z_1^{-1}$ and $z_1^2$ are 0-vertices otherwise, we get additional H0H-connections, a contradiction.

Next, we show that $y_2^2$ is an H-vertex. Let $z_2$ be the common neighbor of $z_1$ and $y_2$. Clearly, $z_2$ is an H-vertex otherwise we get another H0H-connection, a contradiction. Similarly, $z_2^{-1}$ and $z_2^2$ are 0-vertices and $z_2^1$ is an H-vertex. It follows that $y_2^2$ is an H-vertex. By similar arguments, we can show that for every $i = 1, \ldots, 6$, $y_i^1$ is an H-vertex and $y_i^{-1}$ and $y_i^2$ are 0-vertices. Since there is no other occurrence of H0H in $F$, it is easy to see that the whole H-component lies in two layers (the layers containing $y_i$’s and $y_i^1$’s) which are almost identical with exception that 6 vv-vertices in lower layer replaced with 0-vertices in the upper layer.

Note that a vv-H-component is essentially a 2-layer H-component which is missing vertices of one hexagon in one of the two layers.

4.3.2 Complex H-components without a vv-vertex

**Lemma 11.** Let $F$ be a saturated fold of a sparse protein with no H0HH as a substring. Let $C$ be a complex H-component of $F$ without a vv-vertex and $C_1, \ldots, C_r$ its layers. Let $V^{2,2,2}$ be the set of all H-vertices in $F$
with plane degree 2 such that both its horizontal H-neighbors have plane degree 2 as well.

(a) For \( k \geq 1 \), let \( C'_k \) be a subset of \( C_k \) consisting of components of \( C_k \) which are intersecting \( C_1 \). Let \( s \) be the smallest integer such that layer \( C'_{s} \) is different from \( C_1 \). Then \( s > 2 \) and \( C'_{s} \) is a collection of paths where each path is a subset of \( C_1 \cap V^{2,2,2} \).

(b) For \( k \leq r \), let \( C''_k \) be a subset of \( C_k \) consisting of components of \( C_k \) which are intersecting \( C_r \). Let \( e \) be the largest integer such that layer \( C''_{e} \) is different from \( C_r \). Then \( e < r - 1 \) and \( C''_{e} \) is a collection of paths where each path is a subset of \( C_r \cap V^{2,2,2} \).

**Proof.** We prove only part (a) of the claim, part (b) follows by symmetry. Since there is no vv-vertex in \( C \), \( C_2 \) (and hence, also \( C'_2 \)) is a superset of the \( C_1 \). We show that these two layers are identical. To the contrary assume that \( C'_2 \) contains a vertex \( w \) such that its vertical neighbor in the plane \( P(C_1) \) is a 0-vertex. Since \( C'_2 \) is intersecting \( C_1 \), there must be a shortest path connecting \( w \) to some vertex \( u \) of \( C'_1 \). Note that \( u \in C'_s \) and \( u^{-1} \in C_1 \). Let \( v \) be the neighbor of \( u \) on this path, i.e., \( v \) is an H-vertex in \( C_2 \) and \( v^{-1} \) is a 0-vertex. Since the plane degree of \( u^{-1} \) is at least 2, its horizontal neighbors other than \( v^{-1} \) are H-vertices. Since \( C_1 \sim C'_2 \), all horizontal neighbors of \( u \) are H-vertices, i.e., \( u \) is a v-vertex. Therefore, \( u^{-1} \) is a 0-vertex. Furthermore, since there is no vv-vertex in \( F \), \( v^{-1} \) is an H-vertex, cf. Figure 10(a). Since \( u \) is a D-vertex, it is connected to one of its H-neighbors, say \( z \). Then, \( v, u^{-1}, u, z \) form the substring \( H0HH \), a contradiction. Hence, \( C_1 \sim C'_2 \).

Let \( s \) be the smallest integer such that \( C'_s \) is different from \( C_1 \). Since \( C_1 \sim C'_2 \), it follows that \( s > 2 \). Next, we show that \( C'_s \) is a subset of \( C_1 \sim C'_{s-1} \). Assume the contrary. Since (projection of) \( C'_s \) is intersecting \( C_1 \sim C'_{s-1} \), there exists an H-vertex \( v \) and its horizontal H-neighbor \( u \) in \( C'_s \) such that \( v^{-1} \) is a 0-vertex and \( u^{-1} \) is a H-vertex in \( C'_{s-1} \). Since \( C'_{s-1} \sim C'_{s-2} \sim C_1 \), the plane degree of \( u^{-1} \) is 2 and \( v^{-1} \) is an H-vertex, cf. Figure 10(b). Hence, \( u^{-1} \) is a D-vertex, i.e., it is connected to some H-vertex \( z \). Then \( v, v^{-1}, u^{-1}, z \) form the substring \( H0HH \), a contradiction.

Finally, note that any vertex with plane degree 3 in \( C'_{s-1} \) must have a 0-neighbor in the plane \( P(C_s) \), as otherwise it would have five H-neighbors. Since \( C'_s \) is a subset of \( C_1 \cap V^2 \), where \( V^2 \) is the set of all H-vertices in \( F \) with plane degree two, \( C'_s \) is a collection of paths.

Finally, let us prove that each path in \( C'_s \) lies in \( V^{2,2,2} \). Assume the contrary. Then the end point \( v \) of such a path in \( C'_s \) has a 0-neighbor \( u \) such that \( u^{-1} \) is an H-vertex of plane degree 3 in \( C'_{s-1} \). Hence, \( u^{-1} \) is a v-vertex and we have an occurrence of \( H0HH \) (cf. Figure 10(c)), a contradiction. \( \square \)

**Lemma 12.** Let \( F \) be a saturated fold of a sparse protein with no occurrences of the substring \( H0HH \), and at most six occurrences of the substring \( H0H \). Let \( C \) be a complex H-component of \( F \) without a vv-vertex and \( C_1, \ldots, C_r \) be its layers. Let \( s > 2 \) (\( e < r - 1 \)) be the smallest (largest) integer such that \( C_{s} \) (\( C_{e} \)) is different from \( C_1 \) (\( C_r \)). Then both \( C_s \) and \( C_e \) contain a single path, and each of the layers \( C_1, \ldots, C_s, C_e, \ldots, C_r \) is connected. Furthermore, each complex H-component creates at least four occurrences of the substring \( H0H \) in \( F \), two between layers \( C_{s-1} \) and \( C_s \) and other two between layers \( C_e \) and \( C_{e+1} \).

**Proof.** Let \( C'_k, C''_k \) be the sets and \( s, e \) the integers defined in Lemma 11. By this lemma, both \( C'_s \) and \( C''_e \) are collections of paths. Each path in \( C'_s \) and \( C''_e \) creates two new occurrences of substring \( H0H \). Therefore, the total number of paths in \( C'_s \) and \( C''_e \) is either 2 or 3.

First, assume that \( C'_s \) and \( C''_e \) contain 2 paths in total. It is enough to show that for every \( k = 2, \ldots, s \), \( C'_k = C_k \) and for every \( k = e, \ldots, r - 1 \), \( C''_k = C_k \). Assume that there is \( l \in \{2, \ldots, s\} \) such that \( C'_l \neq C_l \) and...
The lemma follows.

for every \( k \) one path. Thus, the change from \( C \) to \( C' \) consists of one path, as any change would introduce a new occurrence of \( \text{H0H} \). Similarly, for any \( l < k < e \), there is only one \( \text{H}-\)component of \( C_k \) intersecting \( K \). Now, the layer \( C_{e-1} \) contains two paths and \( C_e \) only one path. This implies that \( s = e \) and \( e = e \). The lemma follows.

Second, assume that \( C'_s \) and \( C''_e \) contain 3 paths in total. Without loss of generality assume that \( C'_s \) contains 2 paths and \( C''_e \) has only 1 path. This will create 6 occurrences of \( \text{H0H} \) in \( F \). Therefore, as before, \( C_{e-1} \) contains two paths and \( C_e \) only one path, a contradiction.

\[\begin{array}{c}
\includegraphics[width=0.5\textwidth]{figure11.png}
\end{array}\]

Figure 11: A complex \( \text{H}-\)component: the case when layers \( C_s \) and \( C_e \) are identical.

**Observation 13.** Let \( F \) be a saturated fold of a sparse protein with no occurrences of the substrings \( \text{H0HH} \) and \( \text{H0H0H} \), and at most six occurrences of the substring \( \text{H0H} \). Let \( C \) be a complex \( \text{H}-\)component without a \( \text{v} \)-\( \text{v} \)-vertex. Let \( s > 2 \) (\( e < r - 1 \)) be the smallest (largest) integer such that \( C_s \) (\( C_e \)) is different from \( C_1 \) (\( C_r \)). Then \( s \neq e \), i.e., the middle part of a complex \( \text{H}-\)component without a \( \text{v} \)-\( \text{v} \)-vertex (layers \( C_s, \ldots, C_e \)) contains at least 4 \( \text{S} \)-vertices.

**Proof.** By Lemma 12, \( C_s \) consists of one path \( P \). First, notice that if \( s = e \) then the end point \( u \) of the path \( P \) in \( C_s = C_e \) belongs to two different occurrences of \( \text{H0H} \). If these two occurrences share a \( 0 \)-\( \text{v} \)-vertex \( v \) then \( v \) connects to three vertices, a contradiction. Otherwise, we have an occurrence of substring \( \text{H0H0H} \), cf. Figure 11, again a contradiction. \( \square \)

### 4.3.3 Basic complex \( \text{H}-\)component

**Definition 4** (Basic complex \( \text{H}-\)component). Let \( F \) be a saturated fold of a sparse protein with no \( \text{H0HH} \) as a substring. Let \( C \) be a complex \( \text{H}-\)component of \( F \) without a \( \text{v} \)-\( \text{v} \)-vertex with layers \( C_1, \ldots, C_r \). Let \( s \) be the smallest integer such that \( C_s \) is different from \( C_1 \) and let \( e \) be the largest integer such that \( C_e \) is different from \( C_r \). If \( C_s \) is a path and for any \( i \in s + 1, \ldots, e \), \( C_i \) is identical to \( C_s \) then we call \( C \) a **basic complex \( \text{H}-\)component**.

Note that a basic complex \( \text{H}-\)component consists of three parts stacked vertically on each other: (1) a tube or 2-layer \( \text{H}-\)component; (2) a wall; and (3) a tube or 2-layer \( \text{H}-\)component.

**Observation 14.** Let \( F \) be a saturated fold of a sparse protein with no \( \text{H0HH} \) as a substring. Any basic complex \( \text{H}-\)component of \( F \) contains at least 20 \( \text{S} \)-vertices (the lower and upper part at least 8 each and the wall at least 4) and at least 4 occurrences of substring \( \text{H0H} \).

### 4.3.4 Appendix \( \text{H}-\)components

In this subsection, we show that if a complex \( \text{H}-\)component \( C \) without \( \text{v} \)-\( \text{v} \)-vertices is not basic, then its layers change exactly four times, i.e., it consists of five parts stacked on top of each other: (1) a 2-layer \( \text{H}-\)component or a tube; (2) a wall; (3) a pseudo 2-layer \( \text{H}-\)component with exactly one vertex with plane degree 1 in each of two layers; (4) another wall; and (5) a 2-layer \( \text{H}-\)component or a tube. The part in the middle (3) will be
called an appendix, and such a complex H-component will be called an appendix H-component. An example of an appendix H-component is in Figure 12(a). Let us start with the formal definition of an appendix H-component.

![Figure 12](a) An example of an appendix H-component and the six occurrences of H0H contained in it. (b) Illustration what happens if C_{m-1} is not a subset of C_m.

**Definition 5** (Appendix H-component). Let F be a saturated fold of a sparse protein with no occurrence of the substring H0HH. Let C be a complex H-component of F without a vv-vertex with layers C_1, \ldots, C_r. Let s be the smallest integer such that C_s is different from C_1 and let e be the largest integer such that C_e is different from C_r. Assume that both C_s and C_e contain only one path, and that there is an integer s < m < e - 1 such that C_s \simeq C_{s+1} \simeq \cdots \simeq C_{m-1}, C_m \simeq C_{m+1}, C_{m+2} \simeq C_{m+3} \simeq \cdots \simeq C_e, and either C_s is a subset of C_e or C_e is a subset of C_s and both of them are subsets of C_m. Furthermore, assume that C_m has exactly one vertex with plane degree 1 and this vertex is an end point of the paths in C_s and C_e. Such a complex H-component will be called an appendix H-component and the layers C_m and C_{m+1} will be called an appendix. Consider a path in C_m (C_{m+1}) starting at the vertex with plane degree 1 and ending before the first vertex with plane degree 3. These paths in C_m and C_{m+1} will be called the arm of the appendix.

Note that an appendix without its arm is a proper 2-layer H-component.

**Lemma 15.** Let F be a saturated fold of a sparse protein with no occurrence of the substring H0HH, and at most six occurrences of the substring H0H. Every non-basic complex H-component without a vv-vertex in F is an appendix H-component.

**Proof.** Consider a complex H-component C in F without vv-vertices with layers C_1, \ldots, C_r. Assume that C is not a basic complex H-component. Let s (e) be the smallest (largest) integer such that C_s (C_e) is different from C_1 (C_r). By Lemma 12, both C_s and C_e contain only one path. Let m be the smallest integer such that s < m < e and C_m is different from C_s.

First, we will prove that C_s is a subset of C_m. Since C_s \simeq C_{m-1}, C_{m-1} is a path P = (p_1, \ldots, p_t). Assume to the contrary that C_m is not a superset of C_{m-1}. Let p_i (p_j) be the first (last) vertex on path P such that p_i (p_j) is a 0-vertex. Clearly, i \neq j, hence, we have two new occurrences of H0H in addition to four described in Lemma 12. There are no other occurrences of H0H. Therefore, C_m \simeq C_{m+1} \simeq \cdots \simeq C_e, i.e., C_m is a path. Thus, there is a path in C_m connecting paths (p^1_1, p^1_{t+1}) and (p^2_1, p^2_{t+1}). Let (q_1, \ldots, q_\ell) be a shortest such path. Then q_t = p^1_t for some t \in \{1, \ldots, i - 1\} and q_{\ell+1} does not lie on P, i.e., it is a 0-vertex. Then the paths p_i, q_1^1, q_\ell forms another occurrence of H0H, a contradiction, cf. Figure 12(b).

Let m’ be the largest integer such that s < m’ < e and C_{m’} is different from C_e. By symmetry, we have that C_{m’} is a superset of C_e. Obviously, m \leq m’ + 1. We will show that m \leq m’, i.e., that there are at least two changes between layers C_s and C_e. Assume to the contrary that m = m’ + 1. Then C_s \simeq C_{m-1} \simeq C_{m’} \simeq C_m and C_e \simeq C_{m+1} \simeq C_{m’} \simeq C_m, i.e., C_{m’} \simeq C_m. However, this is a contradiction with the fact that C is not a basic complex H-component, since we have C_s \simeq \cdots \simeq C_{m-1} \simeq \cdots \simeq C_{m’} \simeq \cdots \simeq C_e.

Since there are at least two changes from layer C_s to layer C_e and each change will introduce at least one new occurrence of H0H, each of the two changes can create only one occurrence of H0H and there are
no other changes. Therefore, there is exactly one vertex $z$ in $C_m$ which is a horizontal neighbor of some $p_i$ such that $z^{-1}$ is a 0-vertex. If $i \neq 1, \ell$ then we get an occurrence of H0HH. Hence, $C_m$ extends the copy of path $P$ in the plane $P(C_m)$ at one of its ends. Similarly, $C_m'$ extends a copy of the path in layer $C_{m'} + 1$ at one of its ends. Furthermore, since there are no other changes $C_{m} \approx C_{m+1} \approx \cdots \approx C_{m'}$.

It remains to show that $m' = m + 1$ and that $C_m$ has exactly one vertex with plane degree 1. The extended part of $C_m$ ($C_{m'}$) does not have a vertex of plane degree one because otherwise it will be an H-vertex with three 0-neighbors. The number of vertices with odd plane degree in $C_{m'}$ is even. Since there is only one vertex with plane degree one in $C_m$ ($C_{m'}$), there is an odd number of vertices with plane degree 3, which implies there is at least one such a vertex, say $w \in C_m$. Now, if $m' > m + 1$ then $w^1 \in C_{m+1}$ has five H-neighbors, a contradiction. Second, if $m' = m$ then $z$ is a v-vertex, a contradiction. Hence, $m' = m + 1$, i.e., the complex H-component $C$ has a pseudo 2-layer H-component between two walls. It follows that $C$ is an appendix H-component. 

The following observation follows by a careful examination of Figure 12(a).

**Observation 16.** Let $F$ be a saturated fold of a sparse protein with no occurrences of the substrings H0HH and HHHH. Let $C$ be an appendix H-component and $C_s, C_m$ and $C_0$ be the layers after the first, after the second and before the last change, respectively. Then $m \geq s + 2$ and $e \geq m + 3$. Each wall (layers $C_s, \ldots, C_{m-1}$ and $C_{m+2}, \ldots, C_e$) contains at least 4, the arm of appendix of $C$ at least 4 and the appendix without arm at least 10 $S$-vertices. Thus layers $C_s, \ldots, C_e$ contain at least 22 $S$-vertices.

### 4.4 Counting in one plane

Consider a set $S$ of vertices in a hexagonal plane. Set $S$ naturally induces a graph in the plane in which any two neighboring vertices are connected by an edge. In the following S will represent both the set of vertices and the graph induced by this set. Assume that each vertex of $S$ has a degree at least 2. We say that $S$ is complete if every vertex which lies inside the boundary of $S$, denoted as $B(S)$, is in $S$ as well. Let $K_{O}(S)$ be the number of hexagons which lie inside the boundary $B(S)$. $K_2(S)$ the number of vertices of degree 2 of $S$ and $K_3(S)$ the number of vertices of degree 3. Our goal is to lower bound $K_3(S)$ by some function of $K_2(S)$. We will do that in two steps.

**Lemma 17.** Let $S$ be any set of vertices in a hexagonal plane such that each vertex of $S$ has a degree at least 2. We have $K_3(S) \leq 2K_{O}(S) - 2c$, where $c$ is the number of connected H-components of $S$.

**Proof.** First, assume that $S$ is a complete 2-connected set. We proceed by induction on $K_{O}(S)$. If $K_{O}(S) = 1$ then the lemma trivially holds. There must be a hexagon $H$ in $S$ sharing at least two sides with the boundary $B(S)$ such that all its boundary sides form a single path $P$. Consider a set $S'$ obtained from $S$ by removing inner vertices of path $P$. Set $S'$ contains all hexagons contained in $S$ besides $H$. Thus $S'$ is a complete 2-connected set and the number of hexagons $K_{O}(S')$ is $K_{O}(S) - 1$. At the same time, $S'$ must have two vertices of degree 3 less than $S$ (end points of $P$ become vertices of degree 2 and other vertices on $P$ which were removed when constructing $S'$ must have had degree 2). By induction hypothesis, $K_3(S) - 2 = K_3(S') \leq 2K_{O}(S') - 2 = 2(K_{O}(S) - 1) - 2$. This implies that $K_3(S) \leq 2K_{O}(S) - 2$.

Second, assume that $S$ is just a 2-connected set. Let $\bar{S}$ be a set constructed from $S$ by adding all vertices which lie inside the boundary $B(S)$. Note that $B(\bar{S}) = B(S)$ and $\bar{S}$ is complete. Furthermore, the number of vertices of degree 3 of $\bar{S}$ could only increase when adding vertices to $S$. Therefore, $K_3(S) \leq K_3(\bar{S}) \leq 2K_{O}(\bar{S}) - 2 = 2K_{O}(S) - 2$.

Third, assume that $S$ is connected and let $S_1, \ldots, S_l$ be 2-connected H-components of $S$. Contracting every 2-connected H-component to a single vertex we obtain a tree $T$. Every vertex of $T$ of degree 1 or higher than 3 must be a contracted vertex and the number of contracted vertices is $l$. Let $n_d$ be the number of all vertices of degree $d$ and let $n_d'$ the number of all contracted vertices of degree $d$. Note that for $d = 1$ and $d \geq 4$, $n_d' = n_d$ and that $\sum_{d \geq 3} n_d' = l$. Set $S$ has three types of vertices of degree 3: (i) vertices of degree 3 from 2-connected H-components; (ii) vertices of degree 3 created by edges attached to 2-connected H-components; and (iii) $n_3 - n_3'$ of vertices of degree 3 which are not part of any 2-connected H-component.
Note that a contracted vertex of degree \( d \) in \( T \) corresponds to \( d \) vertices of degree 3 of type (ii). Therefore,

\[
K_3(S) = \sum_{i=1}^{l} K_3(S_i) + \sum_{d \geq 1} d \cdot n'_d + n_3 - n'_3 = \sum_{i=1}^{l} K_3(S_i) + 2l + \sum_d (d-2)n_d. \tag{1}
\]

It can be easily shown by induction that for any tree, \( \sum_d (d-2)n_d = -2 \). We know that the lemma holds for every 2-connected \( H \)-component, i.e., for every \( i = 1, \ldots, l \), \( K_3(S_i) \leq 2K_\bigcirc(S_i) - 2 \). Plugging these two facts into formula (1) for \( K_3 \) we obtain

\[
K_3(S) \leq 2 \sum_{i=1}^{l} K_\bigcirc(S_i) - 2l + 2l - 2 = 2K_\bigcirc(S) - 2.
\]

Finally, by summing the bound for each connected \( H \)-component of \( S \), we obtain the desired bound for any \( S \).

\[\square\]

Figure 13: Example of a pseudohexagon with sides 3,3,2,4,2.

**Lemma 18.** Let \( S \) be any set of vertices in a hexagonal plane such that each vertex of \( S \) has a degree at least 2. We have \( K_\bigcirc(S) \leq \frac{1}{12}(K_2(S)^2 + K_2(S) - 30) \).

**Proof.** First, assume that \( S \) is complete and 2-connected, and that the boundary does not have two consecutive concave angles, i.e., the boundary forms a pseudohexagon, cf. Figure 13. We will show that claim holds for any such a pseudohexagon by induction on \( K_2(S) \), which is now equal to the sum of its six sides (measured in the number of hexagons on the particular side). It is easy to verify that the claim holds in the base case when there are two neighboring sides equal to one. Indeed, in this case hexagonal shape is formed by a linear chain of \( t \) hexagons and the number of vertices of degree 2 is \( 2t + 4 \). Assume it is not a base case and let \( s \) be the shortest side of the hexagonal shape \( S \). Observe that the neighboring sides to \( s \) are longer than 1. Consider a hexagonal shape \( S' \) obtained from \( S \) by removing a row of hexagons on the side \( s \). The number of hexagons \( K_\bigcirc(S') \) is \( K_\bigcirc(S) - s \) and since side \( s \) was prolonged by 1, while the neighboring sides shortened by 1, \( K_2(S') = K_2(S) - 1 \). By induction hypothesis, \( K_\bigcirc(S) - s = K_\bigcirc(S') \leq \frac{1}{12}(K_2(S')(K_2(S') + 1) - 30) = \frac{1}{12}(K_2(S)(K_2(S) - 1) - 30) \). Since \( s \) is the shortest side of \( S \), \( K_2(S) \geq 6s \), and hence

\[
K_\bigcirc(S) \leq s + \frac{1}{12}(K_2(S)(K_2(S) - 1) - 30) \\
\leq \frac{1}{12}K_2(S) + \frac{1}{12}(K_2(S)^2 - K_2(S) - 30) = \frac{1}{12}(K_2(S)^2 + K_2(S) - 30).
\]

Second, assume that \( S \) is complete and 2-connected. We will transform \( S \) to a new set \( S' \) by repeating the following process until possible: if there are two or three consecutive concave angles on the boundary add the vertices of the hexagon they are part of, to \( S \). It is easy that this process must stop (we will never go outside of any hexagonal shape enclosing \( S \)). Note that in each step \( K_\bigcirc \) increases by 1 and \( K_2 \) either stays the same or decreases by 1. Thus \( K_\bigcirc(S) \leq K_\bigcirc(S') \) and \( K_2(S') \leq K_2(S) \). Since \( S' \) is a hexagonal shape and complete, the lemma holds for it. Thus it holds for \( S \) as well: \( K_\bigcirc(S) \leq K_\bigcirc(S') \leq \frac{1}{12}(K_2(S')^2 + K_2(S') - 30) \leq \frac{1}{12}(K_2(S)^2 + K_2(S) - 30) \).

Third, assume that \( S \) is 2-connected, but not complete. Let \( \tilde{S} \) be the completion of \( S \) as in the proof of Lemma 17. Note all vertices of degree 2 in \( \tilde{S} \) are on the boundary \( B(S) = B(S) \) and they must be vertices of

\[\text{2Note that this is not a tight bound. We conjecture that the following bound holds } K_\bigcirc(S) \leq \frac{1}{12}(K_2(S)^2 - 6K_2(S) + 12).\]
degree 2 in \( S \) as well. Hence, \( K_\circ(S) = K_\circ(\tilde{S}) \) and \( K_2(S) \geq K_2(\tilde{S}) \). Since \( \tilde{S} \) is complete and 2-connected, it satisfies the lemma. It follows that \( S \) satisfies the lemma as well.

Finally, we prove that any set \( S \) satisfies the lemma by induction on the number of 2-connected H-components. Let \( S' \) be a 2-connected H-component of \( S \) with at most one edge to \( S - S' \). Clearly, such an H-component exists. If \( S' \) is not connected to \( S - S' \), let \( S'' = S - S' \). Otherwise, let \( P = (x, \ldots, y) \) be the path such that \( x \) is the only vertex of \( P \) in \( S' \), all inner vertices \( I(P) \) of \( P \) have degree 2 and \( y \) has degree 3. Then let \( S'' = S - S' - I(P) \). Note that \( K_\circ(S) = K_\circ(S') + K_\circ(S'') \) and \( K_2(S) \geq K_2(S') + K_2(S'') - 2 \). Furthermore, \( S'' \) satisfies the lemma by induction hypothesis and \( S' \) as well, since it is a 2-connected set. Easy calculations and the fact that \( K_2(S'), K_2(S'') \geq 6 \) show that \( S \) satisfies the lemma as well. \( \blacksquare \)

**Corollary 19.** Let \( S \) be any set of vertices in a hexagonal plane such that each vertex of \( S \) has a degree at least 2. We have \( K_3(S) \leq \frac{1}{6}(K_2(S)^2 + K_2(S) - 30) - 2c \), where \( c \) is the number of connected H-components of \( S \).

### 4.5 Limiting certain types of connections and vertices

In this subsection we limit certain types of connections and vertices that occur in a saturated fold \( F \) of \( q \). We first prove that there are at most 4 \( v \)-vertices in \( F \).

**Lemma 20.** Let \( F \) be a saturated fold of \( q \) and assume it contains a complex H-component \( C \) without a \( vv \)-vertex. Let \( s \) be the smallest integer such that \( C_s \) is different from \( C_1 \) and let \( e \) be the largest integer such that \( C_e \) is different from \( C_r \). Let \( w_1 \) be the length of the path in \( C_s \) and \( w_2 \) the length of the path in \( C_e \). Then \( w_1 + w_2 \leq 40 \).

**Proof.** First, note that \( w_1 \) and \( w_2 \) are well-defined, as by Lemma 15, \( C_s \) and \( C_e \) contain only one path. Let \( (p_1, \ldots, p_{w_1}) \) be the path in \( C_s \). Obviously, vertices \( p_1^{-1}, \ldots, p_{w_1}^{-1} \) are \( v \)-vertices. Let \( p_0^{-1} (p_1^{-1}_{w_1+1}) \) be the other neighbor of \( p_1^{-1} (p_{w_1}^{-1}) \). Both, \( p_0^{-1} \) and \( p_{w_1+1}^{-1} \), are \( vh \)-vertices, otherwise we have an occurrence of substring \( HH0H \). Similarly, all vertices, \( p_0^{-1} + 1, p_1^{-1} + 1, \ldots, p_{w_1+1}^{-1} \), are \( vh \)-vertices. Therefore, in layers \( C_1 \) and \( C_s \) we have at least \( w_1 + 4 \) \( S \)-vertices. Similarly, in layers \( C_e \) and \( C_r \) we have at least \( w_2 + 4 \) \( S \)-vertices. Hence, by Observation 13, \( C \) contains at least \( w_1 + w_2 + 12 \) \( S \)-vertices. Since \( q \) contains 52 \( S \)-vertices, the claim follows. \( \blacksquare \)

**Lemma 21.** Let \( F \) be a saturated fold of \( q \). No \( v \)-vertex can be part of substring \((00HH)^{356} \). Consequently, there are at most four \( v \)-vertices in \( F \).

![Figure 14: An example of extending the wall’s end in layer eliminating vertices with horizontal degree 1.](image)

**Proof.** By Lemma 12, each complex H-component introduces at least 4 occurrences of \( H0H \), and hence, there is at most one complex H-component in \( F \). Assume to the contrary that the substring \((00HH)^{356} \) contains a \( v \)-vertex. By Lemma 6, the substring contains only \( v \)-vertices. Let \( P_1, \ldots, P_k \) be all hexagonal planes containing these \( v \)-vertices and let \( S_i \) be the set of H-components in the plane \( P_i \) which contain at least one of these \( v \)-vertices and let \( S \) be the union of \( S_1, \ldots, S_k \). Since every H-component is either a tube, a 2-layer H-component, a complex H-component with six \( vv \)-vertices, a basic complex H-component or an appendix complex H-component, we have the following observations:

- The set \( S \) contains only layers of 2-layer H-components, complex H-components with \( vv \)-vertices, the lower and upper parts of a complex H-components without \( vv \)-vertices if they are 2-layer H-components and layers of the appendix of an appendix H-components. Since all these layers come in identical pairs with exception of a \( vv \)-H-component in which 2-layers differ in 6 vertices, we will consider only one layer in the pair. From each pair select only one layer, for the \( vv \)-H-component select the layer with \( vv \)-vertices. Let \( J \subseteq \{1, \ldots, k\} \) be the set of the selected layers and let \( M = \bigcup_{i \in J} S_i \). We have \( K_2(M) \leq K_2(S) \) and \( K_3(M) \geq \frac{1}{2}K_3(S) \).
• All vertices have horizontal degree 2 or 3 with exception of the wall and (possibly) appendix of a complex H-component without vv-vertices. The layer of a wall without appendix contains two vertices with horizontal degree 1, but no vertex with horizontal degree 3, hence, it is not included in \( M \). On the other hand, a layer containing the appendix contains exactly one vertex with horizontal degree 1. Let us extend the path ending in this vertex in its layer until we join another H-vertex, see an example in Figure 14. There is always a way to do this which introduces at most 4 new vertices with horizontal degree 2, and eliminates at least one such vertex. Let \( M' \) be the set \( M \) extended by these elements and \( S'_i \) either \( S_i \) or \( S_i \) extended by these elements if \( S_i \) was the H-component containing the appendix. Hence, since there is at most one complex H-component and it contains at most two layers with appendix, we have \( K_2(M') \leq K_2(M) + 3 \) and \( K_3(M') \geq K_3(M) \).

By Corollary 19, we have

\[
K_3(M) \leq K_3(M') = \sum_{i \in J} K_3(S'_i) \leq \frac{1}{6} \sum_{i \in J} (K_2(S'_i)^2 + K_2(S'_i)) - 7k
\leq \frac{1}{6} (K_2(M')^2 + K_2(M')) - 7 \leq \frac{1}{6} (K_2(M)^2 + 7K_2(m)) - 5. \tag{2}
\]

It remains to upper bound the number of vertices with horizontal degree 2. Such vertices are either vh-vertices or h-vertices. By Observation 2, there is at most 52 vh-vertices. If we examine all possible H-components, we can see that h-vertices are in the inner layers of tubes or in the last (first) layer of the lower (upper) part of the complex H-components which are directly attached to the walls. However, the H-component in the inner layer of tube contains only vertices with horizontal degree 2, hence, it does not belong to \( S \). Since we have at most one complex H-component, by Lemma 20, we have at most 40 h-vertices which are in \( S \). At most half of these vertices are in \( T \), hence, \( K_2(M) \leq (52 + 40)/2 = 46 \). By (2), we have

\[
K_3(S) \leq 2K_3(M) \leq \frac{1}{3} (46^2 + 7 \times 46) - 10 < 711.
\]

Since every v-vertex has horizontal degree 3, by the assumption, we have \( K_3(S) \geq 2 \times 356 = 712 \), a contradiction.

4.5.1 \((S \cong h)\)-connections

Corollary 22. Let \( F \) be a saturated fold of \( q \). Then \( F \) contains 36 \((S \cong h)\)-connections.

Proof. By Observation 2, \( F \) contains 36 \((S \cong D)\)-connections. Each D-vertex in such a connection is part of the substring \((00HH)^{356}\), hence, by Lemma 21, is an h-vertex.

![Figure 15](image-url)

Figure 15: (a-c) Illustration of an external horizontal \((S \cong h)\)-connection. Contradictory cases: (a) the case when \( v = u \), (b) the case where \( x \) and \( y \) are on the same hexagon. The only possible configuration in (c). (d) Illustration of a vertical external \((S \cong h)\)-connection.

We define two types of \((S \cong h)\)-connections. Assume that S-vertex \( x \) is \((S \cong h)\)-connected to \( y \). We say that this \((S \cong h)\)-connection is horizontal if \( x \) and \( y \) are on the same plane (cf. Figure 15(c)) and it is vertical if \( x \) and \( y \) are on two consecutive planes (cf. Figure 15(d)).
Lemma 23. Let $F$ be a saturated fold of $q$. Let $x$ be a vh-vertex and $y$ be an h-vertex in two different components $W_1$ and $W_2$. If $x$ and $y$ are $(S \asymp h)$-connected, then the $(S \asymp h)$-connection $(x, u, v, y)$ is either horizontal or vertical. Furthermore, if the $(S \asymp h)$-connection $(x, u, v, y)$ is vertical, it creates also an H0H-connection between $x$ and a vertical neighbor of $y$. Finally, if $H$-components $W_1$ and $W_2$ are non-complex, there is at most one parallel $(S \asymp h)$-connection with $(S \asymp h)$-connection $(x, u, v, y)$ and in the vertical case the components share only one layer.

Proof. Let $x$ be on plane $P_x$. Without loss of generality assume that $x^1$ is a 0-vertex and let $w$ be the horizontal 0-neighbor of $x$. Clearly, $u$ is either $x^1$ or $w$. We consider each case separately.

Case 1 $(u = w)$. If $v = u^1$ then $y$ must be a horizontal neighbor of $v$ and thus, $u$ is adjacent to the H-vertex $y^1$, a contradiction (cf. Figure 15(a)). Furthermore, if $v = u^1$ then $y = x^{-1}$ and it follows that $x$ and $y$ are in the same H-component, a contradiction. Therefore, $v$ is a horizontal neighbor of $u$ and $y$ is a horizontal neighbor of $v$. Note that $y$ must be the horizontal neighbor of $v$ that is not on the same hexagon with $x$ otherwise, $x$ and $y$ would be in the same H-component, a contradiction, cf. Figure 15(b). Hence, $x$ and $y$ are on the same plane (horizontal $(S \asymp h)$-connection), cf. Figure 15(c). Next, assume that $(x^i, u^i, v^i, y^i)$ and $(x^j, u^j, v^j, y^j)$ are two parallel $(S \asymp h)$-connections with $(x, u, v, y)$. Obviously, $i, j < 0$, and let $i < j$. Since $(x, u, v, y)$ and $(x^i, u^i, v^i, y^i)$ are parallel connections, all vertices between $x$ and $x^i$ $(y$ and $y^i)$ are H-vertices, i.e., neither $x^i$ nor $y^i$ is an vh-vertex. If the H-components they are contained in are non-complex, they must be D-vertices, a contradiction.

Case 2 $(u = x^1)$. By a similar argument as used in the first case, we can show that $v \neq u^1$. Therefore, $v$ is a horizontal neighbor of $u$. If $v$ is above an H-neighbor of $x$ then $y$ is the H-neighbor of $x$, i.e., $x$ and $y$ are in the same component, a contradiction. Hence, $v = w^1$. It follows that $y$ is a horizontal neighbor of $v$ and it is on plane $P_{i+1}$, cf. Figure 15(d). This is a vertical $(S \asymp h)$-connection. Furthermore, in this setting $(y^1, w, x)$ form an H0H-connection. Second, note that $y$ is an S-vertex. If the H-component containing $y^1$ is non-complex, then it is a vh-vertex, i.e., $y^{-1}$ is 0-vertex and the two H-components can share only one layer. Consequently, there is at most one parallel $(S \asymp h)$-connection to $(x, u, v, y)$.

4.5.2 H0H-connections

Definition 6 (Horizontal and vertical H0H-connections). We say that an H0H-connection is horizontal, vertical if both peptide edges of the connection are horizontal, vertical, respectively.

We have the following simple observation.

Observation 24. Let $F$ be a saturated fold of a sparse protein of length at least 5. Then every H0H-connection connecting two different H-components is either horizontal or vertical.

Proof. Assume that H0H-connection $(x, y, z)$ is neither horizontal nor vertical. Without loss of generality, assume that the edge $(x, y)$ is vertical, let $y = x^1$, and $(y, z)$ is horizontal. If $z^{-1}$ is a 0-vertex then we have a closed path of length 4. If $z^{-1}$ is an H-vertex then $x$ and $y$ belong to the same H-component.

Lemma 25. Let $F$ be a saturated fold of $q$ and let $C$ be an H-component of $F$. Assume that $C_i$ is a layer in $F$ that does not contain any vertex of plane degree 1. Then there is no H0H-connection with both end points in $C_i$. Consequently, there is no internal H0H-connection in a tube or a 2-layer H-component.

![Diagram of H0H-connection](image)

Figure 16: Horizontal H0H-connection $(x, z, y)$: (a) the case where $y^{-1}$ is 0-vertex, (b) the case where $y^1$ is 0-vertex.
Proof. To the contrary assume that \(x\) and \(y\) have a common horizontal \(0\)-neighbor \(z\). We remark that H-component \(C\) cannot be a \(\text{w-H}\)-component since such an H-component already contains 6 H0H-connections which are different type than \((x, z, y)\). Clearly one of the vertical neighbors of \(x\) has to be a \(0\)-vertex otherwise \(F\) contains an occurrence of \(\text{H0H}\) as a substring. Without loss of generality assume that \(x^1\) is a \(0\)-vertex. Similarly, one of the vertical neighbors of \(y\) has to be a \(0\)-vertex. First assume that \(y^1\) is a \(0\)-vertex, cf. Figure 16(a). Note that in this case, layers \(C_{i-1}, C_i\) and \(C_{i+1}\) are different which cannot happen in any H-component of \(F\). Therefore, \(x\) and \(y\) are in different H-components, a contradiction.

Second assume that \(y^1\) is a \(0\)-vertex. It follows that \(y^1\) is an H-vertex. Note that \(x^{-2}\) and \(y^{-2}\) are \(0\)-vertices, otherwise \(F\) would contain \(\text{H0HH}\) as a substring cf. Figure 16(b). Moreover, all horizontal neighbors of \(y^1, y^{-2}, x^1\) and \(x^{-2}\), except \(z^1\) and \(z^{-1}\) are \(0\)-vertices, otherwise \(F\) would contain an occurrence of the substring \(\text{H0HH}\). Next consider the H0H connection \((x, z, y)\). One of the vertices \(x\) and \(y\) has to connect to a D-vertex \(w\) via two \(0\)-vertices \(u\) and \(v\). By Lemma 21, \(w\) must be an h-vertex. It is easy to see that \(u = x^1\) and \(v = x^2\). Now \(w\) must be a horizontal neighbor of \(x^2\) which is not possible.

Corollary 26. Let \(F\) be a saturated fold of \(q\). Then the smallest non-simple tube contains 7 hexagons and 36 S-vertices, cf. Figure 8(b).

\[
\begin{tikzpicture}
\node at (0,0) {\text{(a)}};
\node at (3,0) {\text{(b)}};
\node at (6,0) {\text{(c)}};
\node at (0,1) {\text{W_1}};
\node at (0,-1) {\text{W_2}};
\node at (3,1) {\text{W_1}};
\node at (3,-1) {\text{W_2}};
\node at (6,1) {\text{W_1}};
\node at (6,-1) {\text{W_2}};
\node at (0,0) {x};
\node at (0,1) {y};
\node at (0,-1) {z};
\node at (3,1) {u};
\node at (3,-1) {v};
\node at (6,1) {w};
\node at (6,-1) {w'};
\node at (1.5,0) {\text{W_1}};
\node at (1.5,1) {\text{W_2}};
\node at (1.5,-1) {\text{W_2}};
\node at (4.5,1) {\text{W_1}};
\node at (4.5,-1) {\text{W_2}};
\node at (7.5,1) {\text{W_1}};
\node at (7.5,-1) {\text{W_2}};
\node at (0.5,0) {x'};
\node at (1.5,0) {y'};
\node at (2.5,0) {z'};
\node at (4.5,0) {x'};
\node at (5.5,0) {y'};
\node at (6.5,0) {z'};
\node at (8.5,0) {x'};
\node at (9.5,0) {y'};
\node at (10.5,0) {z'};
\node at (0.5,1) {v'};
\node at (1.5,1) {u'};
\node at (2.5,1) {w'};
\node at (4.5,1) {v'};
\node at (5.5,1) {u'};
\node at (6.5,1) {w'};
\node at (8.5,1) {v'};
\node at (9.5,1) {u'};
\node at (10.5,1) {w'};
\node at (0.5,-1) {v'};
\node at (1.5,-1) {u'};
\node at (2.5,-1) {w'};
\node at (4.5,-1) {v'};
\node at (5.5,-1) {u'};
\node at (6.5,-1) {w'};
\node at (8.5,-1) {v'};
\node at (9.5,-1) {u'};
\node at (10.5,-1) {w'};
\draw (0,0) -- (0,1);
\draw (0,0) -- (0,-1);
\draw (3,0) -- (3,1);
\draw (3,0) -- (3,-1);
\draw (6,0) -- (6,1);
\draw (6,0) -- (6,-1);
\draw (0,1) -- (3,1);
\draw (0,-1) -- (3,-1);
\draw (3,1) -- (6,1);
\draw (3,-1) -- (6,-1);
\draw (0,1) -- (1.5,1);
\draw (0,-1) -- (1.5,-1);
\draw (3,1) -- (4.5,1);
\draw (3,-1) -- (4.5,-1);
\draw (6,1) -- (7.5,1);
\draw (6,-1) -- (7.5,-1);
\draw (0.5,0) -- (1.5,0);
\draw (1.5,0) -- (2.5,0);
\draw (2.5,0) -- (3.5,0);
\draw (3.5,0) -- (4.5,0);
\draw (4.5,0) -- (5.5,0);
\draw (5.5,0) -- (6.5,0);
\draw (6.5,0) -- (7.5,0);
\draw (0.5,1) -- (1.5,1);
\draw (1.5,1) -- (2.5,1);
\draw (2.5,1) -- (3.5,1);
\draw (3.5,1) -- (4.5,1);
\draw (4.5,1) -- (5.5,1);
\draw (5.5,1) -- (6.5,1);
\draw (6.5,1) -- (7.5,1);
\draw (0.5,-1) -- (1.5,-1);
\draw (1.5,-1) -- (2.5,-1);
\draw (2.5,-1) -- (3.5,-1);
\draw (3.5,-1) -- (4.5,-1);
\draw (4.5,-1) -- (5.5,-1);
\draw (5.5,-1) -- (6.5,-1);
\draw (6.5,-1) -- (7.5,-1);
\end{tikzpicture}
\]

Figure 17: Situation when two non-complex H-components are connected with a horizontal H0H-connection: (a) \(x\) is connected to a h-vertex \(w\) away from the other H-component; (b) \(w\) belongs to the other H-component. (c) An example of two non-complex H-components connected with a vertical H0H-connection.

Lemma 27. Let \(F\) be a saturated fold of \(q\). Consider an H0H-connection \((x, y, z)\) connecting two non-complex H-components \(W_1\) and \(W_2\). If this connection is horizontal then at least one of the two H-components is a tube with more than two layers, they share only one plane and they are configured as in Figure 17(b). If this connection is vertical then they do not share any plane (cf. Figure 17(c)).

Proof. First, assume that \((x, y, z)\) is a horizontal H0H-connection. It is easy to see that \(W_1\) and \(W_2\) make another horizontal H0H-connection \((x', y', z')\), where \(x' \in W_1\) and \(y' \in W_2\), cf. Figure 17(a). Since in the string \(q\), every H0H continues with O0HH on one end, at least one of the vertices \(x\) or \(z\) must connect to a D-vertex, say \(w\), via two \(0\)-vertices, say \(u\) and \(v\). Without loss of generality, let it be \(x\). Obviously, \(x\) is an vh-vertex. Without loss of generality, assume that \(u = x^1\). By Lemma 21, \(w\) must be an h-vertex, therefore, \(w\) is a horizontal neighbor of \(x\). Now, if \(v = u^1\) then \(u\) will be adjacent to the H-vertex \(w^{-1}\) (cf. Figure 17(a)), a contradiction. Hence, \(v\) is a horizontal neighbor of \(u\) and it is easy to see that \(v = y^1\) and \(w = z^1\). The configuration of parts of two H-components is depicted in Figure 17(b). Since the h-vertex \(w\) belongs to \(W_2\), \(W_2\) must be a tube with height more than 2 layers and since these two H-components are non-complex, they can only share one plane.

Second, assume that \((x, y, z)\) is a vertical H0H-connection. Obviously, the two H-components do not share any plane, and all H0H-connections between them are vertical. An example of configuration in which two non-complex H-component are vertically H0H-connected is depicted in Figure 17(c).

4.6 Limiting the possible configurations of complex H-components

In this subsection we show that only a limited number of configurations are possible for a complex H-component. This will greatly simplify our analysis in the later sections. In the following arguments we say that a path has length \(k\) if it contains \(k\) vertices.
Lemma 28. Let $F$ be a saturated fold of $q$. Then $F$ does not contain any $vv$-$H$-component.

Proof. Let $C$ be a $vv$-$H$-component. Consider any of the $H0H$-paths in $C$ for example $(x_1, x_1', y_1')$, cf. Figure 9. Notice that this path has to continue with substring $(00HH)^{k_1}$ at one end. By Lemma 21, all $H$-vertices in this substring are $h$-vertices, i.e., either $y_1'$ or $x_1^{-1}$ has to connect via two $0$-vertices to an $h$-vertex. It is easy to check that none of these connections is possible, a contradiction.

Lemma 29. Let $F$ be a saturated fold of $q$ and let $C$ be a complex $H$-component in $F$ with layers $C_1, C_2, \ldots, C_r$. Layer $C_1$ (and similarly $C_r$) consists of either one hexagon, or of two hexagons sharing one edge, or of two hexagons connected by a path (cf. Figure 18 for all three cases).

![Figure 18: Possible configurations for the upper and lower part of a complex $H$-component.](image)

Proof. By Lemma 28, $C$ does not contain any $vv$-vertex. We prove the claim for $C_r$; the proof for $C_1$ follows by symmetry. By Observation 7 and Lemma 25, $C_r$ does not contain any horizontal $H0H$-connection. By Lemma 11, layers $C_{r-1}$ and $C_r$ are the same, and hence, for every vertex with plane degree 3 in $C_r$, we have two $v$-vertices (the vertex itself and its vertical neighbor in $C_{r-1}$). Therefore, $C_r$ cannot contain more than 2 vertices of plane degree 3 because otherwise we get more than 4 $v$-vertices, a contradiction by Lemma 21. Consequently, $C_r$ must have one of the following three topologies: (i) a cycle, (ii) two cycles sharing a path; (iii) two cycles connected by an edge. We will show that all cycles must be simple hexagons, i.e., we only have possibilities depicted in Figure 18.

![Figure 19: (a) The second smallest cycle without $H0H$ occurrences. (b) The smallest possible layer $C_1$ of a complex $H$-component with the lower part being a 2-layer $H$-component containing a large cycle.](image)

The smallest possible layer of an $H$-component with no vertices of plane degree 1 or 3 and no horizontal $H0H$-connections other than a simple hexagon is a cycle containing 7 hexagons inside, cf. Figure 19(a) and the smallest such layer with exactly two vertices of plane degree 3 and containing inside at least three hexagons is depicted in Figure 19(b). We prove that $C_r$ cannot be the cycle in Figure 19(a) and any larger cycle by computing a lower bound on the number of $S$-vertices in $F$. Clearly, $C$ will have more $S$-vertices if $C_r$ has two vertices of degree 3, i.e., contains two cycles as in Figure 19(b).

Assume the contrary. We will consider two cases: $C$ is either a basic or an appendix complex $H$-component.

Case 1. Let $C$ be a basic complex $H$-component. Note that the number of $S$-vertices in $C$ is minimized when the wall width is maximized and the wall height is minimized. The lower part of $C$ is either a simple tube or the second smallest tube similar to $C_r$. Figure 20(a)-(b) depicts these configurations with the smallest number of $S$-vertices. The width of the wall can be at most 4 and 16 in the first and second configurations, respectively. However, in both of these configurations the number of $S$-vertices is at least 44 which happens...
when the height of the wall is 2. In addition, notice that \( C \) only contains 4 \( \text{H0H} \)-connections, therefore, \( F \) must contain another \( \text{H} \)-component which brings the total number of \( S \)-vertices up to at least \( 44 + 12 = 56 \), a contradiction with Observation 2.

**Case 2.** Let \( C \) be an appendix \( \text{H} \)-component and let \( w_1 \) and \( w_2 \) be the lower and the upper wall width of \( C \), respectively cf. Figure 20(c). Similar to case 1, the lower part of \( C \) is either a simple tube or the second smallest tube. If it is the second smallest tube then the minimum number of \( S \)-vertices will be \((18 + 2) \cdot 2 \) (vertices in lower and upper part) + 22 (vertices in appendix and wall ends) = 62, a contradiction. Hence, assume that the \( C_1 \) consists of one hexagon. Note that \( w_1 \leq 4 \) and \( w_2 \leq 16 \). The minimum number of \( S \)-vertices in different layers of \( C \) is as follows:

- **vertices in \( C_r \):** 18
- **vertices in the first layer of upper part:** \( 18 - w_2 \)
- **vertices at the ends of two walls:** 8
- **vertices of the appendix:** the appendix without the arm contains at least 10 \( S \)-vertices, the arm on its ends contain 4 and if the walls have different widths, then on the side of the shorter wall the arm has additional \( |w_2 - w_1| \) \( S \)-vertices. Hence, in total appendix has at least \( 14 + |w_2 - w_1| \) \( S \)-vertices.
- **vertices of the first layer of the upper part:** \( 6 - w_1 \)
- **vertices in \( C_1 \):** 6

Hence, the total number of \( S \)-vertices is at least \( 70 - w_1 - w_2 + |w_2 - w_1| \). Now, if \( w_1 \leq w_2 \) then the minimum number of \( S \)-vertices is \( 70 - w_1 - w_2 + |w_2 - w_1| = 70 - 2w_1 \geq 62 \), a contradiction with Observation 2. If \( 4 \geq w_1 > w_2 \) it is \( 70 - w_1 - w_2 + |w_2 - w_1| = 70 - 2w_2 \geq 62 \), a contradiction as well.

**Lemma 30.** Let \( F \) be a saturated fold of \( q \) and let \( C \) be a complex \( \text{H} \)-component in \( F \). Then the lower and upper part of \( C \) are simple tubes.

**Proof.** By Lemma 29, the upper (lower) part of \( C \) is either a simple tube or one of two layer \( \text{H} \)-components depicted in Figure 18(b)-(c). Suppose the contrary that one of the parts is not a simple tube, say the upper part. First notice that either of the 2-layer \( \text{H} \)-components in Figure 18(b) and Figure 18(c) contain four \( v \)-vertices. Therefore, by Lemma 21, \( C \) cannot have an appendix, and the lower part must be a simple tube as well. Therefore, \( C \) contains only four \( \text{H0H} \) connections, and hence, \( F \) must contain at least one other \( \text{H} \)-component \( T \). Furthermore, \( T \) has to be a simple tube because if it is a 2-layer \( \text{H} \)-component, a complex \( \text{H} \)-component or a large tube then \( F \) would contain more than 4 \( v \)-vertices, more than 6 \( \text{H0H} \)-connections (Lemma 12) or more than 52 \( S \)-vertices (by Observation 9, the smallest non-simple tube has 28 \( S \)-vertices, and by Observation 14, the smallest basic complex \( \text{H} \)-component has 20 \( S \)-vertices, but since the upper part of our basic complex \( \text{H} \)-component contains a large cycle, the smallest number of \( S \)-vertices in it is at least by \( 18 - 6 = 12 \) more in the top-most layer), respectively.

Next we consider two cases for the shape of the upper part of \( C \):

**Case 1.** Assume that the upper part of \( C \) is a connector. By Lemma 11, the width of the wall is 2. Now independent of the height of the wall in \( C \) the number of \( \text{D} \)-vertices modulo 6 in \( F \) is 2, a contradiction.

**Case 2.** Assume that the upper part of \( C \) consists of two hexagons connected by a path, cf. Figure 18(c). The wall part of \( C \) can either attach to one of the hexagons or the path \( P \) connecting the two hexagons, cf. Figure 21. Similar to the previous case if the width of the wall is 2 the number of \( \text{D} \)-vertices modulo 6 in \( F \) is 2 independent of height or the location of the wall, a contradiction. Furthermore, if the wall is attached to one of the hexagons then by Lemma 11, the width of the wall can be at most 3. Figure 21(a) depicts this configuration with wall width equal to 3. Let \( x \), \( y \) and \( z \) be the vertices on the last layer of the wall. Each of the vertices \( x^1 \) and \( z^1 \) must connect to a \( \text{D} \)-vertex via a peptide bond. The only \( \text{D} \)-vertex in their neighborhood is \( y^1 \) thus, \( x^1 \) and \( z^1 \) must both connect to \( y^1 \) which is not possible. Using a similar argument we can show that the width of the wall cannot be 3 for the case where it is attached to the path \( P \). Since the lower part is a simple tube the only case remaining for analysis is the configuration in which the wall is attached to \( P \) and its width is 4. By Lemma 11, the smallest length of \( P \) is 6 and by Observation 13, the smallest height of the wall is 4. Note that such an \( \text{H} \)-component would have 40 \( S \)-vertices (28 upper
Figure 20: (a) A basic complex H-component with the second smallest tube as upper part and a simple tube as the lower part. (b) A basic complex H-component with the second smallest tube as upper and lower part. (c) An appendix H-component with the second smallest tube as upper part and a simple hexagon as lower part.

part, 4 wall and 8 lower part), and with the extra H-component at least 52 S-vertices. Increasing either the length of the path or the height of the wall would increase this number hence, Figure 21(b) depicts the only possible configuration of the complex H-component. We show that this configuration is also impossible by determining the maximum number of (S × h)-connections possible. Note that at most 12 internal (S × h)-connections are possible across the vertices of C and T, respectively. Therefore, by Corollary 22 we need to create at least 12 connections between the S-vertices of C and h-vertices of T. However, at least 10 of these (S × h)-connections must be horizontal because each vertical (S × h)-connection create an H0H-connection, by Lemma 23. Since a horizontal (S × h)-connection between C and T is possible only when the H-vertices in the connection are on the same plane, C and T must have at least 5 connections per plane which is easy to see it is not possible given the shape of C.

Lemma 31. Let F be a saturated fold of q and let C be a complex H-component in F. The width of the wall in C is either 2 or 4.

Proof. By Lemma 28, C does not contain any w-w-vertex and by Lemma 30, its lower and upper part are simple tubes. Assume that the lower wall starts at layer Cs of C. First observe that the wall width cannot be 1 or 5 otherwise, we get an H-vertex with three 0-neighbors or a 0-vertex with three H-neighbors, respectively, both contradictions.

Therefore, it is enough to show that the wall width cannot be 3. Let x, y, z be the path of the wall in layer Cz (attached to the tube H-component). Note the number of D-vertices in this layer and above is odd. Since they have to form pairs, y−1 has to connect to y, and hence, x and z have to connect to x1 and z1, respectively. Let us look at patterns of vertical connections between consecutive layers of a tube. It can be shown by induction (from the top of the tube) that only the patterns depicted in Figure 22(a) are
Figure 21: Examples of complex H-components with a 2-layer H-component consists of two hexagons connected by a path as the upper part: (a) wall is attached to one of the hexagons; (b) wall is attached to the path connecting hexagons.

possible. However, the pattern required to realize connections $xx^1$ and $zz^1$, depicted in Figure 22(b) cannot be obtained, a contradiction.

Figure 22: (a) All possible patterns (up to rotation) for vertical connections between two consecutive layers of a simple tube. The "x" means vertical connection is not present, arrow means it is present. (b) Pattern required to connect to the last layer of a simple tube which is connected to a path of length 3.

4.7 There is no appendix H-component

Consider an appendix H-component $C$ in a saturated fold $F$ of $q$. By Lemma 30, the upper and lower part of $C$ are simple tubes. Let $C_a$ and $C_{a+1}$ be the layers of $C$ that contain the appendix part. Observe that $C_a$ (and similarly $C_{a+1}$) contains an odd number of vertices of plane degree 3 (such vertices correspond to v-vertices in $C$). Therefore, $C$ contains $4k - 2$ v-vertices for some positive integer $k$. By Lemma 21, $F$ contains at most 4 v-vertices hence, the appendix part of $C$ contains one hexagon.

Observation 32. Let $F$ be a saturated fold of $q$. Let $C$ be an appendix H-component in $F$. Then $C$ contains exactly 2 v-vertices.

Lemma 33. Let $F$ be a saturated fold of $q$. Then $F$ does not contain any appendix H-component.

Proof. Assume that $F$ contains an appendix H-component $C$. First we show that $F$ can only contain simple tubes. By Observation 16 and Corollary 26, $F$ cannot contain a non-simple tube, otherwise we have too many S-vertices. If $F$ contains another complex H-component, then we have at least 10 H0H substrings, which is not possible. If it contains a 2-layer H-component, then $F$ contains at least 6 v-vertices, two in $C$ and 4 in the 2-layer H-component, a contradiction. Hence, all other H-components of $F$ are simple tubes. Let $N_t$ be the number simple tubes.

Let $w_1 (h_1)$ be the width (height) of the lower wall of $C$ and $w_2 (h_2)$ the width (height) the upper wall. Let $a$ be the lengths of the arm. We will calculate the number of D-vertices modulo 6 and the number of S-vertices in $C$ and $F$. The lower (upper) part of $C$ (a tube) contains $w_1 \text{ mod } 6 (w_2 \text{ mod } 6)$ D-vertices modulo 6, the lower (upper) wall $(w_1 - 2)h_1 ((w_2 - 2)h_2)$, the arm of appendix $w_1 - 1 + w_2 - 1$ and the remaining part of the appendix, by Observation 32, 2 D-vertices. That is

$$2(w_1 + w_2) + (w_1 - 2)h_1 + (w_2 - 2)h_2 \text{ mod } 6$$
D-vertices modulo 6 in $C$, and since all other H-component are simple tubes, the same number in $F$. The number of S-vertices is $12 - w_1 (12 - w_2)$ in the lower (upper part), $2h_1 (2h_2)$ in the lower (upper) wall, $2a - w_1 - w_2 + 2$ in the arm and at least 10 in the remaining part of the appendix. That is at least $36 + 2(h_1 + h_2 + a - w_1 - w_2)$ S-vertices in $C$, and $36 + 2(h_1 + h_2 + a - w_1 - w_2) + 12N_t$ in $F$.

By Lemma 31, both $w_1$ and $w_2$ are either 2 or 4, hence, we will consider the following 3 cases (without loss of generality, we assume that $w_1 \leq w_2$).

Case 1. $w_1 = w_2 = 2$. By the above formula, the number of D-vertices modulo 6 in $F$ is 2, a contradiction with Observation 2.

In the remaining two cases, we will first show that $C$ is the only H-component, i.e., that $N_t = 0$.

Case 2. $w_1 = w_2 = 4$, cf. Figure 23(a). By the above formula, the number of D-vertices in $F$ is $4 + 2(h_1 + h_2) \mod 6$. Since, by Observation 2, this number is 4, we have $h_1 + h_2 \equiv 0 \mod 3$. Since, by Observation 16, $h_1, h_2 \geq 2$, we must have $h_1 + h_2 \geq 6$. Also note that $a \geq 5$. Hence, the number of S-vertices is at least $36 + 2 \times 3 + 12N_t$. Since this number should be 52, we have $N_t = 0$.

Case 3. $w_1 = 2$ and $w_2 = 4$, cf. Figure 23(b). The number of D-vertices modulo 6 in $F$ is $2h_2 \mod 6$. Hence, $h_2 \geq 3$. And since $w_2 = 4$, we have again $a \geq 5$. Therefore, the number of S-vertices in $F$ is at least $36 + 2 \times 5 + 12N_t$. Hence, again $N_t = 0$.

We will determine the maximum number of $(S \times h)$-connections in $F$. Notice that in any of the configurations the S-vertices in the wall except for the end vertices on the first and the last layers cannot connect to any h-vertex, so there are at most 4 $(S \times h)$-connections involving the S-vertices of the wall H-components. Furthermore, the S-vertices in the appendix part and its arm can only connect to the h-vertices in the wall that are in the same plane with them otherwise, we get additional H0H-connections, a contradiction. Therefore, there can be at most 4 $(S \times h)$-connections involving the S-vertices of the appendix part and its arm. The last $(S \times h)$-connections that we can get in $F$ are through vH-vertices of the lower and upper tubes, which are 16 in the first configuration and 18 in the second configuration. Two more $(S \times h)$-connections are possible in the second configuration through vH-vertices $x$ and $y$ in Figure 23(b). Therefore, in total $F$ can contain at most 28 $(S \times h)$-connections, a contradiction, by Corollary 22.

No other type of possible H-components can introduce 6 occurrences of H0H, hence, a saturated fold of $F$ contains at least two H-components. On other hand, since any of possible H-components has at least 12 S-vertices, we have the following corollary.

**Corollary 34.** Any saturated fold of $q$ has at least 2 and at most 4 H-components.

In what follows we will analyze all three possibilities. But first, let us have a closer look at tubes.
4.8 Tubes

Lemma 35. Let $F$ be a saturated fold of $q$. Any tube in $F$ has either 12 or at least 36 $S$-vertices.

Proof. Obviously, any cycle in a hexagonal plane has at least 6 vertices, i.e., a smallest possible tube will have at least 12 $S$-vertices. Furthermore, by Lemma 25, there is no $H0H$ with both ends in the same tube. The smallest cycle larger than a hexagon such that no two non-adjacent vertices are at distance two contains 7 hexagons inside. Thus, the second smallest tube has 36 $S$-vertices (18 in the top and 18 in the bottom layers).

Lemma 36. Let $F$ be a saturated fold of $q$. Two $H0H$-connected tubes in $F$ are both simple and furthermore, they make exactly two $H0H$-connections.

Figure 24: (a) A shortest possible collection of paths connecting the parts of cycle of $T_2$ that make 6 horizontal $H0H$-connections with a simple tube $T_1$. (b) Six vertical $H0H$-connections between two simple tubes.

Proof. Let $T_1$ and $T_2$ be two tubes in $F$. By Corollary 26, one of them, assume $T_1$, must be a simple tube. First note that if $T_2$ is not a simple tube it must make 6 $H0H$-connections with $T_1$ since $F$ cannot have another $H$-component. Assume that there is an $H0H$-connection $(x, y, z)$ such that $x$ and $z$ are $H$-vertices in $T_1$ and $T_2$, respectively. By Observation 24, there are two cases:

Horizontal $H0H$-connection $(x, y, z)$. By Lemma 27, $T_1$ and $T_2$ share only one plane $P_i$ and create at least two $H0H$-connections as depicted in Figure 17(b). We will show that $T_2$ must also be a simple tube. Assume the contrary. Since $T_2$ has at least 36 $S$-vertices, there are no other $H$-components in $F$, and hence, $T_2$ must make 6 $H0H$-connections with $T_1$. Moreover, since $T_1$ and $T_2$ share a plane $P_i$, no vertex of $T_1$ can be directly above/below any vertex of $T_2$, i.e., all the $H0H$-connections are horizontal and they are on plane $P_i$. Therefore, the only way how to make 6 horizontal $H0H$-connections is when the large cycle $C_2$ of $T_2$ on plane $P_i$ contains 3 parts depicted in Figure 24(a) with thick lines. The shortest collection of 3 disjoint paths which do not create $H0H$-connection and connecting these parts to one cycle is shown with dashed lines. Note that $C_2$ would contain at least 30 vertices and hence, $T_2$ would have more than 60 $vh$-vertices, a contradiction.

Vertical $H0H$-connections. Assume that $x$, $y$ and $z$ are on three consecutive planes $P_i$, $P_{i+1}$ and $P_{i+2}$, respectively. In this case, $T_1$ and $T_2$ do not share any plane and hence, all the $H0H$-connections between them must be vertical and in the same planes. Note that if $T_2$ is not a simple tube it can vertically overlap (overlap its vertical projection) with $T_1$ on at most 3 edges creating at most 4 $H0H$-connections, a contradiction since there are no other $H$-components in $F$. Clearly, two simple tubes could vertically overlap either on 1 or 6 edges creating 2 or 6 $H0H$-connections, respectively. It is enough to show that two simple tubes cannot make 6 vertical $H0H$-connections. Assume the contrary. Figure 24(b) depicts two $H0H$-connected simple tubes with 6 $H0H$-connections. Note that no pair of $H0H$-connections in this configuration can connect through two $0$-vertices. Therefore, $F$ does not contain the substring $H0H00H0H$ which is in $q$, a contradiction.

4.9 2 $H$-components

Lemma 37. Let $F$ be a saturated fold of $q$. Fold $F$ cannot have only 2 $H$-components.
Proof. Assume there are two $H$-components in $F$.

Case 1. Assume they are both tubes. By Lemma 36, they can only make two $H0H$-connections, a contradiction.

Case 2. Assume they are both 2-layer $H$-components. By Lemma 21, we have no occurrence of substring $(00HH)^k$, a contradiction.

Case 3. Assume they are both basic complex $H$-components. Then we have 8 occurrences of $H0H$, a contradiction.

Case 4. Assume one $H$-component is a tube $T$ and the other a 2-layer $H$-component $C$. By Lemma 21 and Lemma 35, there are only two configurations with 52 $S$-vertices. The first configuration consists of a connector and a tube that is the second smallest tube with 7 hexagons inside of its boundary on each layer $(16 + 36 = 52)$. The second configuration consists of a 2-layer $H$-component with two simple tubes connecting by a path of length 11 and a simple tube $(40 + 12 = 52)$. Note that in both configurations the two $H$-components must make 6 $H0H$-connections. In the first configuration it is easy to see that at most two horizontal $H0H$-connections can be created between the tube and connector. Therefore, all the $H0H$-connections must be vertical. However, in this case a $v$-vertex of the 2-layer $H$-component will be part of an $H0H$-connection creating the substring $H0HH$, a contradiction. We show that the second configuration is not possible by showing that the maximum number of $(S \times h)$-connections in $F$ is less than 36. Notice that at most two $h$-vertices of each side of the tube can make $(S \times h)$-connections with $vh$-vertices of the 2-layer $H$-component and hence, we can obtain at most 12 external $(S \times h)$-connections between the 2-layer $H$-component and the tube. Considering the 12 internal $(S \times h)$-connections in the tube, the total number of $(S \times h)$-connections in this configuration is at most 24, a contradiction by Corollary 22.

Figure 25: A schematic vertical projection of the wall ($W$) and possible locations of a tube (1, 2,...): (a) shows possible locations of the tube when the tube and the wall are 0-connected and (b) when they are 00-connected (up to symmetry).

Case 5. Assume one $H$-component is a tube and the other a basic complex $H$-component. Obviously, the tube must be a simple tube. By Lemma 30, the lower and upper parts of the complex $H$-component are both simple tubes. Let $w$ be the width of the wall and $h$ its height. The number of $S$-vertices is $12 + 24 - 2w + 2h = 52$, so we have $h = w + 8$. On the other hand, the number of $D$-vertices modulo 6 is $2w + (w - 2)h$. By Lemma 31, $w$ is either 2 or 4. For, $w = 4$ the number of $D$-vertices modulo 6 is 2, a contradiction. Thus, the only possibility is $w = 2$ and $h = 10$. Note that if the tube does not connect (through one or two 0-vertices) to an end of the wall then, a substring $(00H)^9$ is created which does not occur in $q$. Hence, the tube has to connect to both sides of the wall. Figure 25 depict a schematic vertical projection of the wall ($W$) and possible locations of a tube (numbered positions). It also shows possible connections between the wall and a tube through one and two 0-vertices, respectively. Notice that if the wall connects to tube through two 0-vertices the first two connections have to be horizontal. If the third connection is vertical then we get the configuration in Figure 25(a) in one layer above or below. Clearly, the only way that a tube can be connected to both sides of the wall is when it is in position 2 in Figure 25(a). Notice that in this case tube is connected to both sides of wall through one 0-vertex creating an $H0H$-connection on each end. Furthermore, there will be at least one parallel $H0H$-connection on each side and in total at least 4 additional $H0H$-connection, a contradiction.

Case 6. Assume one $H$-component is a 2-layer $H$-component $W$, and the other a basic complex $H$-component $C$. We show that the maximum number of $(S \times h)$-connections in this configuration is less than 36. First
we count the internal \((S \asymp h)\)-connections in \(C\). All \(h\)-vertices in \(F\) appear inside the wall and the lower and the upper part of \(C\). The \(S\)-vertices of the wall except for the \(S\)-vertices on its first and last layers cannot connect to any \(h\)-vertex. Therefore, there are at most 4 \((S \asymp h)\)-connections with the \(S\)-vertices of the wall. There are at most \(12 - w\) \(S\)-vertices in the upper (lower) part of \(C\), where \(w\) is the wall width. Since \(w \geq 2\), there are at most 20 internal \((S \asymp h)\)-connections with the \(S\)-vertices of these parts. Therefore, there has to be at least 12 external \((S \asymp h)\)-connections between \(C\) and \(W\). It is easy to verify that at most two \(h\)-vertices of each side of \(C\) can \text{00}-connect to an \(S\)-vertex of \(W\). Hence, \(W\) has to \text{00}-connect to \(C\) from each side. However, one can easily show that for this to happen \(W\) must have at least 28 \(S\)-vertices in each layer which is a contradiction, since \(W\) has at most \(52 - 20 = 32\) \(S\)-vertices.

\[\square\]

### 4.10 3 \(H\)-components

**Lemma 38.** Let \(F\) be a saturated fold of \(q\). Then \(F\) cannot contain 3 \(H\)-components where none of them is a complex \(H\)-component.

**Proof.** Since the second smallest tube has 36 \(S\)-vertices, all tubes must be simple. Note that \(F\) does not contain a complex \(H\)-component and by Lemma 21, \(F\) can contain at most one 2-layer \(H\)-component, hence, to obtain 52 \(S\)-vertices, \(F\) must have two tubes \(T_1\) and \(T_2\), and a 2-layer \(H\)-component \(W\) with two hexagons connected by a path of length 5 in each layer.

By Lemma 25, there is no \(H0H\)-connection with both ends in \(W\). Therefore, at least one of the tubes, say \(T_1\), must \(H0H\)-connect to \(W\). Furthermore, notice that \(S\)-vertices of \(T_1\) and \(T_2\) can only provide 24 \((S \asymp h)\)-connections, so we need to create 12 external \((S \asymp h)\)-connections between the \(S\)-vertices of \(W\) and \(h\)-vertices of \(T_1\) and \(T_2\). By Lemma 23, these connections are either horizontal or vertical. If \(W\) and one of the tubes are vertically \((S \asymp h)\)-connected then we have configuration in Figure 26(a). Notice that although in this configuration two \((S \asymp h)\)-connections are created between the tube and \(W\), we lose two \((S \asymp h)\)-connections across the tube. Therefore, there are 12 horizontal \((S \asymp h)\)-connections between the tubes and \(W\). The only way to create these connections is depicted in Figure 26(b).

Furthermore, since \(T_1\) and \(W\) are \(H0H\)-connected, by Lemma 27, none of the \(h\)-vertices of \(T_1\) is on the same plane as the \(v_h\)-vertices of \(W\), and hence, they cannot make any horizontal \((S \asymp h)\)-connections. Therefore, all of the 12 \((S \asymp h)\)-connections must be made between \(W\) and \(T_2\). This requires that \(W\) connects to \(T_2\) from every side which is not possible since the path connecting two hexagons of \(W\) has length only 5.

\[\square\]

**Lemma 39.** Let \(F\) be a saturated fold of \(q\). Then \(F\) cannot contain 3 \(H\)-components where one of them is a complex \(H\)-component.

**Proof.** Assume that \(F\) contains a complex \(H\)-component \(B\). By Lemmas 28 and 33, \(B\) must be a basic complex \(H\)-component. By Lemma 30, \(B\) does not have a 2-layer part. Therefore, the number of \(S\)-vertices and the number of \(D\)-vertices modulo 6 of \(B\) are \(24 - 2w + 2h\) and \(2w + (w - 2)h\) mod 6, respectively where \(h\) is the height and \(w\) is the width of the wall of \(B\). By Lemma 31, two values are possible for \(w\): \(w = 2\) or \(w = 4\). We will consider each case separately.

**Case 1.** \((w = 4)\) Since \(F\) can contain at most one 2-layer \(H\)-component, one of the three \(H\)-components in \(F\) must be a tube \(T\). Furthermore, \(B\) has at least 20 \(S\)-vertices, therefore, the third \(H\)-component can have
at most 20 $S$-vertices. Hence, it can be either another tube $T_2$, a connector $C$ or a 2-layer $H$-component $W$ that consists of two hexagons connected by one edge in each layer. The values for $h$ are 6, 4 and 2 when the third $H$-component is $T_2$, $C$ or $W$, respectively. For $h = 6, 4$ the number of $D$-vertices modulo 6 is 2, a contradiction. Therefore, the only possible configuration is the one in which the third $H$-component of $F$ is $W$ and $h = 2$.

The basic complex $H$-component $B$ is depicted in Figure 27(a). It has 20 $S$-vertices, out of which 8 are part of $H0H$-connections. Notice that only one of the two $S$-vertices involved in an $H0H$-connection (such as $x$ and $y$) can 00-connect to an h-vertex, otherwise $F$ will contain the substring $HH0H0H0H0H$ which does not occur in $q$. Therefore, the maximum number of possible $(S \times h)$-connections with $S$-vertices vertices of $B$ and $T$ is $16 + 12 = 28$. Hence, we need to create 8 external $(S \times h)$-connections with the $S$-vertices of $W$ and h-vertices of $B$ or $T$. Figure 27(b) depicts the only possible configuration to make 8 of such connections. Notice that in this configuration the $H$-components are far away to make any $H0H$-connections with each other so the total number of $H0H$-connections possible is 4, a contradiction.

Case 2. ($w = 2$) The number of $D$-vertices modulo 6 of $B$ is 4 independent of the value of $h$. Therefore, the only possibility for the other two $H$-components in $F$ is that they are both simple tubes, say $T$ and $T'$. To have right number of $S$-vertices in $F$ the height $h$ must be 4.

Note that an $H$-vertex from one side of the wall cannot connect to an $H$-vertex from the other side of the wall through one or two 0-vertices. Therefore, if the wall is not connected to any vertices of $T$ or $T'$ through one or two 0-vertices, then the two $H0H$-connections on the same side of wall has to connect through a subsequence containing only $S$-vertices. This creates a substring which does not occur in $q$, a contradiction. Therefore, at least one vertex on each side of the wall must connect to a tube.

First, we show that the wall cannot 00-connect to a tube. To the contrary assume that a vertex $v$ of tube $T$ is connected to a vertex $x$ of the wall through a 0-vertex $w$. Vertex $x$ cannot be located on the first or the fourth level of the wall otherwise, $F$ would contain the substring $H0H0H$, a contradiction. Assume that $v$ is in the hexagon that touches that wall. In this case we get another $H0H$-connection between other side of the wall and $T$ in the same plane. This situation repeats in the plane above or below. Hence, there are at least 4 new $H0H$-connections, a contradiction. Now, assume that $v$ is not on the hexagon that touches the wall. The vertex $v$ is a $vh$-vertex otherwise, $F$ would contain a substring $H0HH$. Without loss of generality assume $v^1$ is a 0-vertex. One of the vertices $v$ or $x$ must 00-connect to an h-vertex. It is easy to verify that it cannot be $v$. Therefore, assume that $x$ connects to an h-vertex of $T'$ through 0-vertices $y$ and $y'$. The only position of $T'$ is shown in Figure 28. However, in this configuration the right side of the wall cannot
connect to neither of the tubes, a contradiction. Therefore, each side of the wall is 00-connected to a vertex of a tube.

![Diagram](image1.png)

**Figure 28:** H0H-connections between the tube $T$ and a wall of complex H-component with $w = 2$ and $h = 4$.

Notice that it is not possible to 00-connect both sides of the wall to the same tube and hence, one side of the wall is 00-connected to $T$ while the other side is 00-connected to $T'$, e.g., Figure 29(a).

There are two ways to 00-connect a tube to the wall, cf. Figure 30. Note that we need to have two more H0H-connections in $F$. First, we show that no H0H-connections can be made between $B$ and one of the tubes, say $T$. Since $T$ cannot H0H-connect to the wall, it would have to connect to the lower or the upper part of $B$. This is not possible given the relative position of wall of $B$ and $T$ depicted in Figure 30(a). If the relative position of the wall of $B$ and $T$ is as depicted in Figure 30(b), there is only one possible configuration which is depicted in Figure 29(b). However, this configuration contains the substring H0H0H, a contradiction.

Therefore, the H0H-connections must be made between $T$ and $T'$. The gray hexagons in Figure 30 depicts the possible positions for $T'$. Clearly, $T'$ cannot 00-connect to the other side of the wall in any of these positions, a contradiction.

![Diagram](image2.png)

**Figure 29:** (a) One possible attachment of two tubes to the wall of complex H-component. (b) H0H-connections of tube $T$ and basic complex H-component $B$.

4.11 4 H-components

So far we have proved that any saturated fold $F$ of $q$ must have exactly four H-components. In this section we prove that the fold $F$ is structurally similar to the designed fold, i.e., that $q$ is structurally stable. First, we show that the H-components in $F$ are the same as the H-components in the designed fold. Note that, so far, we did not use the fact that some of the hydrophobic monomers are cysteines in the proofs. Hence, the
Figure 30: A schematic vertical projections showing possible configurations of a tube $T$ connected to the wall $B$ of the complex $H$-component through two $0$-vertices. Gray hexagons represent the locations of $T'$ that can $H0H$-connect to $T$ and is not too far from the wall.

The following result is valid also in the HP model.

**Theorem 40.** Let $F$ be a saturated fold of $q$, then $F$ has three simple tubes and a connector. This is true even if the HP model is considered (and all cysteines are replaced with ordinary hydrophobic monomers).

**Proof.** Since the smallest $H$-component other the tube with one hexagon contains at least 16 $S$-vertices and $F$ contains exactly four $H$-components, $F$ must have three simple tubes and one $H$-component other than a tube. The three tubes together have 36 $S$-vertices, therefore, the forth $H$-component in $F$ must have 16 $S$-vertices. The only $H$-component with 16 $S$-vertices is the connector. Therefore, the $H$-components in $F$ are the same as the $H$-components in the designed fold.

Next, we prove that in the HPC model the $H$-components in $F$ must connect the same way as in the designed fold.

In Lemma 36, two tubes in $F$ can connect with at most two $H0H$-connections. We will show the same for a tube and connector.

**Lemma 41.** Let $F$ be a saturated fold of $q$. A tube and a connector in $F$ can create at most two $H0H$-connections.

**Proof.** Assume that the connector $W$ and a tube $T$ are $H0H$-connected. By Observation 24, this connection is either horizontal or vertical. If the connection is horizontal, by Lemma 27, $W$ and $T$ share only one plane, cf. Figure 17(b). Obviously, all other $S$-vertices of $W$ and $T$ are too far from each other to create more $H0H$-connections than the two depicted in the figure.

Second, assume there is a vertical $H0H$-connection between $W$ and $T$. Then $W$ and $T$ do not share any plane and $H0H$-connections are created if an edge of $W$ is directly above/below an edge of $T$. If $W$ and $T$ vertically overlap on more than one edge, then there is a $D$-vertex of $W$ directly above/below a vertex of $T$, which would create a substring $H0HH$ in $F$, a contradiction. Hence, $W$ and $T$ vertically overlap on only one edge, and hence, create exactly two $H0H$-connections.

**Lemma 42.** Let $F$ be a saturated fold of $q$. Assume that a connector $C$ and a tube $T$ are horizontally ($S \approx h$)-connected in $F$. Then there are at most two external ($S \approx h$)-connections between them and $T$ is missing at least two internal ($S \approx h$)-connections.

**Proof.** Assume that vertex $x$ of $C$ is horizontally ($S \approx h$)-connected to vertex $y$ of $T$. By Lemma 23, we have the configuration depicted in Figure 15(d). Vertex $x^{-1}$ is an $S$-vertex of $C$ and it cannot be part of a parallel ($S \approx h$)-connection, because $y^{-1}$ is an $S$-vertex as well. Also note that $S$-vertex $y^{-1}$ of $T$ cannot be part of internal ($S \approx h$)-connection. Since horizontal neighbors of $y^{-1}$ and $x$ are $H$-vertices we have another $H0H$-connection between these two neighbors and we lose another internal ($S \approx h$)-connection. Similarly, there is at most one ($S \approx h$)-connection between $C$ and $T$ parallel to this $H0H$-connection. Considering the layout of $C$ and $T$, it is clear that they cannot ($S \approx h$)-connect at any other point. Hence, the claim follows.
Observation 43. Let $F$ be a saturated fold of $q$. Assume that two tubes $T_1$ and $T_2$ are $(S \times h)$-connected. Then the number of missing internal $(S \times h)$-connections in $T_1$ and $T_2$ minus the number of external $(S \times h)$-connections between them is at least zero.

Lemma 44. Let $F$ be a saturated fold of $q$. Assume that two tubes $T_1$ and $T_2$ are $H0H$-connected. Then the number of missing internal $(S \times h)$-connections in $T_1$ and $T_2$ minus the number of external $(S \times h)$-connections between them is at least two.

Proof. If $T_1$ and $T_2$ are vertically $H0H$-connected then at most one endpoint of each of two $H0H$-connections is $00$-connected to an $h$-vertex, since there is no $HH00H0H0HH$ in $q$. Therefore, we lose at least two internal $(S \times h)$-connections and gain no external $(S \times h)$-connections between $T_1$ and $T_2$.

If $T_1$ and $T_2$ are horizontally $H0H$-connected we have the configuration depicted in Figure 17(b). Vertices $x, x', z, z'$ are $S$-vertices of the tubes which cannot be part of internal $(S \times h)$-connections, hence we lose at least four of $(S \times h)$-connections. Furthermore, all possible external $(S \times h)$-connections between $T_1$ and $T_2$ are $(x, u, v, w)$, $(z, z^{-1}, y^{-1}, x^{-1})$, $(x', x'^1, y'^1, z'^1)$ and $(z', z'^{-1}, y'^{-1}, x'^{-1})$. However, first two and last two cannot be present at the same time, otherwise we have $HH00H0H0HH$ in $q$. Hence, there are at most two such connections.

Lemma 45. Let $F$ be a saturated fold of $q$. The tubes in $F$ have more than 3 layers.

Proof. Assume that one of the tubes, say $T_1$, has two or three layers. We prove this lemma by counting the number of possible $(S \times h)$-connections in $F$. If $T_1$ has 2 layers, then it does not contain any internal $(S \times h)$-connections, since it has no $h$-vertices. If it has 3 layers then it contains 6 $h$-vertices, but since they are connected to each other with a peptide bond and there are only two occurrence of substring $0H0H0H0H0$ in $q$ which occur in the connector, at most one in each pair can be involved in an $(S \times h)$-connection. Hence, there is at most 3 internal $(S \times h)$-connections. There should be 36 $(S \times h)$-connections in $F$, and the remaining two tubes have at most 24 internal $(S \times h)$-connections. Hence, $F$ must contain at least 9 external $(S \times h)$-connections. By Lemma 42, any external vertical $(S \times h)$-connection eliminates at least one internal $(S \times h)$-connection. Hence, there has to be at least 9 external horizontal $(S \times h)$-connections.

Consider an external horizontal $(S \times h)$-connection $(x, u, v, y)$ connecting $H$-components $W_1$ and $W_2$, cf. Figure 15(c). By Lemmas 36 and 41, any pair of $H$-components in $F$ can create at most two $H0H$-connections, i.e., at least three pairs of $H$-components are $H0H$-connected. Since these pairs cannot be horizontally $(S \times h)$-connected, there are at most three pairs of horizontally $(S \times h)$-connected $H$-components. Hence, by Lemma 22, there are at most 6 horizontal $(S \times h)$-connections, a contradiction.

We can proceed by proving the following lemma.

Lemma 46. Let $F$ be a saturated fold of $q$. Any $H$-component in $F$ must be $H0H$-connected to at least one other $H$-component.

Proof. By Lemmas 35 and 41, there are at most two $H0H$-connections between any two $H$-components of $F$. Since $F$ contains 6 $H0H$-connections it is enough to show that there is no cycle of length 3 of $H0H$-connected $H$-components. Let $H$-components $W_1, W_2, W_3$ form such a cycle. By Lemmas 27, 36 and 41, two $H0H$-connected $H$-components are either in the configuration depicted in Figure 17(b) or Figure 17(c), i.e., they share exactly one plane or they share no planes and there is one plane in between them. Assume that $W_1$ is the topmost $H$-component in planes $P_i, P_{i+1}, \ldots, P_j$. If both $W_2$ and $W_3$ share one plane with $W_1$ (or none of them shares a plane with $W_1$) then they share at least two layers, i.e., they cannot be $H0H$-connected. Hence, assume that $W_2$ shares plane $P_i$ with $W_1$, i.e., it is located in planes $P_i, P_{i+1}, \ldots$ and $W_3$ does not share any plane, i.e., it is located in planes $P_{i-2}, P_{i-3}, \ldots$. Then $W_2$ and $W_3$ can share zero or one plane only if $W_2$ has either one or three layers. Obviously, the first case is not possible. In the second case, $W_2$ must be a tube, but by Lemma 45, it cannot have 3 layers, a contradiction.

We will proceed by proving the following important lemma:

Lemma 47. Let $F$ be a saturated fold of $q$. Two tubes in $F$ cannot be $H0H$-connected. Consequently, two tubes cannot be vertically $(S \times h)$-connected.
Proof. To the contrary assume that two tubes are H0H-connected. By Lemma 42, we need at least two external (S ≍ h)-connections, and by Lemma 42 and Observation 43, there are at most two horizontal (S ≍ h)-connections between the connector C and a tube T1.

Figure 31 shows a schematic vertical projection of the horizontal (S ≍ h)-connections between T1 and C. Notice that T1 and C cannot be H0H-connected. Therefore, by Lemma 46, T1 must H0H-connect to another tube T2. We will show that T2 cannot be H0H-connected to C. Assume the contrary. The tube T2 must be located in one of the three numbered positions in Figure 31.

Figure 32 depicts configurations for all three positions of T2. Clearly, in the first configuration T2 cannot make any H0H-connections with C (cf. Figure 32(a)). Consider vertex v in Figure 32(b) depicting the second configuration. It is H0H-connected to v−2 which is part of an (S ≍ h)-connection. Since there is no substring HH00H00HH in q, v cannot be part of any (S ≍ h)-connection. Therefore, we lose one more internal (S ≍ h)-connection in T2 which needs to be replaced by an external (S ≍ h)-connection between C and a tube. By Lemma 42, any external vertical (S ≍ h)-connection eliminates at least one internal (S ≍ h)-connection, therefore, the replaced connection must be a horizontal (S ≍ h)-connection. Clearly, T2 cannot make any horizontal (S ≍ h)-connections with C and furthermore, T1 cannot make any new horizontal (S ≍ h)-connections with C. Hence, T3 must make at least one horizontal (S ≍ h)-connection. By Lemma 46, T3 must H0H-connect to at least one other component. If it H0H-connects to C then it could not horizontally (S ≍ h)-connect to any other component. Therefore, T3 must H0H-connect to T1 or T2. In this case we lose at least two additional internal (S ≍ h)-connections which cannot be replaced by any external horizontal (S ≍ h)-connections. Finally, we show that also the third configuration is contradictory. Consider the v-vertex v in Figure 32(c). If it is 00-connected to w or x it follows that v is a part of the substring (00HH)k, a contradiction by Lemma 21. Therefore, v is 00-connected to u. However in this case, F contains the substring HH00H00HH which does not occur in q, a contradiction.

It follows that T2 and C are not H0H-connected. Therefore, by Lemma 46, T3 must H0H-connect to C and to create 6 H0H-connections in F, T3 must also H0H-connect to T1 or T2. However, in this case we lose at least two additional internal (S ≍ h)-connections which by Lemma 42, must be replaced by horizontal (S ≍ h)-connections between C and a tube. Clearly, T3 cannot make such connections with C. Furthermore, T1 cannot make any horizontal (S ≍ h)-connections with C. Thus, T2 must make two horizontal (S ≍ h)-connections with C. Let Pj and Pi+1 be the planes in which C lies. Without loss of generality assume that T2 is above T1. Since C and T1 make horizontal (S ≍ h)-connections the top most layer of T1 lying in the plane Pj is above Pi+1. Let P1 be the lowest plane in which T2 lies. Since T1 and T2 are H0H-connected, l ≥ j > i + 1. Therefore, C and T2 do not share any layer and hence, cannot be (S ≍ h)-connected, a contradiction.

Corollary 48. Let F be a saturated fold of q. All tubes in F must be H0H-connected to the connector.

Observation 49. Let F be a saturated fold of q. Any H00H-connection between two tubes is horizontal.

Proof. We consider three cases. If the connection is between two h-vertices then clearly all edges of the connection must be horizontal. Second the case when the connection is between h- and S-vertices follows by Lemma 47. Finally, if the connection is between two S-vertices, we lose two internal (S ≍ h)-connections which can be only replaced by horizontal (S ≍ h)-connection between connector and a tube. By Corollary 48, this is not possible.
So far we have shown that all tubes must H0H-connect to C. We prove the final theorem.

**Theorem 50.** The protein string \( q \) is structurally stable.

**Proof.** Let \( F \) be a saturated fold of \( q \). By Theorem 40 and Corollary 48, \( F \) contains three simple tubes which are H0H-connected to a connector \( C \). Note that there are no H0H-connections between tubes. Since the substring \( t = 10100102002 \) is in \( q \), it must appear also in \( F \). The substring \( t \) contains two H0H-connections that are 00-connected. We show that these H0H-connections are vertical and they belong to two tubes \( T_1 \) and \( T_2 \) where \( T_1 \) is connected to the top and \( T_2 \) is connected to the bottom of \( C \). To the contrary, assume that one of the H0H-connections \( (u, v, w) \) in \( t \) is horizontal, where \( u \) and \( w \) are H-vertices in \( C \) and \( T_1 \), respectively, and \( v \) is a 0-vertex. Note that \( C \) and \( T_1 \) make another H0H-connection \( (u', v', w') \) where \( u' \) and \( v' \) are horizontal neighbors of \( u \) and \( v \) respectively. Vertex \( u \) or \( w \) (respectively, \( u' \) or \( w' \)) must 00-connect to an h-vertex. It is easy to see that \( w \) \((w') \) cannot 00-connect to an h-vertex and the only h-vertex that \( u \) \((u') \) can 00-connect to is \( w^1 \) \((w'^1) \). Therefore, \( w \) must 00-connect to an H0H-connection.

Two configurations are possible in this case. In the first configuration \( w \) is 00-connected to \( w' \), cf. Figure 33(a), and hence, exactly one of the pairs of vertices \((u, w^1) \) or \((u', w'^1) \) contains 2-vertices. Since \( T_1 \) makes H0H-connections only with \( C \) and every 2-vertex is either a part of H0H-connection or is 00-connected to an H0H-connection, \( w^1 \) \((respectively, w'^1) \) cannot be paired with a 2-vertex, a contradiction. In the second configuration \( w \) is 00-connected to \( u^{-1} \) and \( C \) is vertically H0H-connected to another tube \( T_2 \) at \( u^{-1} \) and its horizontal neighbor (cf. Figure 33(b)). Note that \( T_3 \) must connect to the hexagon of \( C \) the does not contain \( u \) and \( u^{-1} \) otherwise, \( F \) would contain the substring \((00H)^6 \), a contradiction. Therefore, \( T_1 \) is too
Figure 33: Two possible configurations that contain the substring $t = 10100102002$, given that one of the H0H-connections in $t$ is horizontal.

far from $T_2$ and $T_3$ to 00-connect to either of them. Hence, $w^1$ is $p$-connected to $w'^1$ by a path $p$ which lies completely in $T_1$ and its 0-vertices (0-vertices surrounding $T_1$). Consequently, $p$ does not contain any H0H as a substring. Since H0H-connection $(w, v, u)$ is 00-connected to H0H-connection $(u^{-1}, u^{-2}, u^{-3})$, based on the properties of $q$, it follows that exactly one of the pairs $(u, w^1)$ or $(u', w'^1)$ contains 2-vertices, depending on the direction of the substring $t$. Clearly, $w^1$ (respectively, $w'^1$) cannot be paired with any other 2-vertex, a contradiction. Therefore, both H0H-connections in $t$ are vertical.

Figure 34: The only possible configuration that contains the substring $t = 10100102002$, given that the H0H-connections in $t$ are vertical.

Let $(u, u^1, u^2)$ be one of the H0H-connections in $t$ where $u$ and $u^2$ are H-vertices in $C$ and $T_1$, respectively, and $u^1$ is a 0-vertex. Without loss of generality assume that $T_1$ is connected to the top of $C$. Note that $T_1$ and $C$ make another vertical H0H-connection $(v, v^1, v^2)$, where $v$ is a horizontal neighbor of $u$. Clearly, $u$ cannot 00-connect to an h-vertex, therefore, it must 00-connect to another vertical H0H-connections. The only possibility is that $u^1$ is 00-connected to $u^{-1}$. Therefore, $C$ vertically H0H-connect to another tube $T_2$ at the vertex $u^{-1}$ and one of its horizontal neighbor. If this connection is $(v^{-1}, v^{-2}, v^{-3})$ then $F$ would contain another occurrence of the substring $t$ through vertices $v^2, v^1, v, *, *, v^{-1}, v^{-2}, v^{-3}$, a contradiction. It follows that $T_1$ and $T_2$ are H0H-connected to $C$ as in the original fold. It is easy to see that the last tube $T_3$ must horizontally H0H-connect to the other side of $C$ as in the original fold (cf. Figure 34).

Finally, notice that $T_1$, $T_2$ and $T_3$ are far away from each other to make any 00-connections. Therefore, the pair of H0H-connections in each tube are $p$-connected by a path $p$ that lies completely in that tube and its 0-vertices. This implies that the length of the tubes must be the same as the length of the tubes in the original fold and hence, $q$ is structurally stable.  

33
5 Conclusions

In this paper building on the work done in Gupta et al. (2005, 2007); Hadj Khodabakhshi et al. (2008) we solve the shape-approximating inverse protein folding problem on the HP model in 3D for designing tubular proteins by providing two basic building blocks: a tube and a connector, which can be interconnected to approximate any given shape. We showed that a simple subclass of the structures built in this way is structurally stable in the HPC model. Showing that all these structures are structurally stable is a very challenging problem. The first task in solving this problem is to choose which of the hydrophobic monomers are cysteines. The second is to prove that all native folds are structurally similar to the designed one. This gets more difficult with the higher number of building blocks (tubes and connectors) used, as each additional building block adds two special substrings to the protein sequence, and thus increases a variety and the number of possible H-components in the fold. One possible simplification of this problem is to consider only generalized tubular structures in which the tubes are not touching (such as the generalized tubular structure proved to be stable in this paper). This structures in this subclass could still approximate any given shape, although more roughly, i.e., the approximations would fill the shapes but would leave about 50% of space (between tubes) inside the shapes empty.

While the techniques presented here will not allow for the direct construction of proteins, they represent a starting point for this process. In particular, we believe that our techniques can be used to form the basis of an actual protein — we specify, at each point of the chain whether a cysteine, other hydrophobic or polar monomer is required and a designer can use this information to choose amino acids from set of all 20 amino acids. The choice of actual amino acid would depend on other desired molecular interactions and finer details about the protein structure.

References


