DISCRETE ISOPERIMETRY AND PROTEIN FOLDING

SHIVAM NADIMPALLI

Abstract. We introduce the reader to the broad field of isoperimetry and briefly survey discrete isoperimetric inequalities. We then cast protein folding in the HP model as an isoperimetric problem, and obtain new bounds on the conformation energy of linear polymers. In addition, we solve Istrail's bipole packing problem on the two-dimensional triangular and three-dimensional cubic lattices. We prove a related problem, bipole labelling, to be **NP**-hard, and propose a generalization of the bipole packing problem to consider bipoles of varying lengths.

Advisor: Sorin Istrail
Title: Julie Nguyen Brown Professor of Computational & Mathematical Sciences

Reader: Richard Schwartz
Title: Chancellor's Professor of Mathematics

Contents

1. Hello!

One of the earliest problems to be considered in geometry is the *isoperimetric problem*, dating back to the ancient Greeks. The problem can be stated as follows: Among all closed curves in the plane enclosing a fixed area, which curve minimizes the perimeter?

According to legend, the Phoenician queen Dido founded the city of Carthago on a piece of land obtained from the local king. As she got only as much land as she could surround by an oxhide, she cut the hide into thin strips and then encircled an entire hill nearby. A satisfactory proof that the circle is indeed the optimal solution to this problem, however,

Date: May 2, 2019. Brown University, Providence, RI.

was only obtained in the 19^{th} century. Today, isoperimetry is a mainstay of modern geometry, with connections to graph theory, partial differential equations, and probability theory.

This paper attempts to relate the notion of isoperimetry to the *protein folding problem*, one of the holy grails of computational biology. In addition to being of tremendous importance to drug design and medicine, the protein folding problem also carries broad implications for combinatorial optimization, discrete geometry, and biophysical simulations. We argue that protein folding is fundamentally an isoperimetric problem, and will then consider several combinatorial problems related to protein folding in the HP model.

1.1. **Outline**

This paper is organized as follows:

- In Section 2, we introduce the reader to isoperimetric inequalities. We present a proof of the classical isoperimetric inequality in $\mathbb{R}^n$, and briefly survey discrete isoperimetric inequalities.
- In Section 3, we introduce protein folding and cast it as an isoperimetric problem.
- In Section 4, we resolve Istrail's *biplane conjecture* on the two-dimensional triangular and three-dimensional cubic lattices, and prove a related problem, Bipole-Labelling, to be **NP**-complete.
- In Section 5, we indicate some potential directions for future work. A generalization of the bipole packing problem to consider bipoles of varying lengths is introduced and modeled as a linear program.

Standard mathematical notation is used throughout. All graphs are assumed to be connected and undirected unless stated otherwise.

This work is part expository and part exploratory, with some original findings in Sections 3 and 4. The scrambled and sometimes bizarre mixture of topics presented is not an accident. While we made some attempts at a narrative, this paper is ultimately just a collection of loosely connected topics that the author found interesting.

2. Classical and Discrete Isoperimetric Inequalities

Isoperimetry deals with questions such as "Given the size of a set, how small can its boundary be?" or "Given the boundary of a set, how large can its area be?". For example, in $\mathbb{R}^2$, circular discs are best. In Section 2.1, we give a proof of the classical isoperimetric inequality for $\mathbb{R}^n$, and then briefly discuss discrete isoperimetric inequalities in Section 2.2.

2.1. **The Classical Isoperimetric Inequality**

The classical isoperimetric inequality informally states that "spheres minimize boundary for fixed area in Euclidean space". Though this fact for $\mathbb{R}^2$ was known to the ancient Greeks, a satisfactory proof was not obtained until the 19$^{\text{th}}$ century. The famous geometer Jakob Steiner gave a proof, using a geometric construction, which, starting from a curve different from a circle, leads to a curve with the same length but with strictly greater area. His colleague Dirichlet, then in Berlin, tried without success to convince him that this does not suffice as a proof unless it shows the existence of a solution.

Many proofs of the classical isoperimetric inequality are now known. The following proof uses Brunn-Minkowski theory, and a comprehensive introduction to the subject can be found in [Sch93].

Definition 2.1 (Minkowski sum)**.** Given $A, B \subset \mathbb{R}^n$, the *Minkowski sum* of A and B is defined as $A \oplus B = \{a + b \mid a \in A, b \in B\}$.

Definition 2.2 (parallelepipeds)**.** Let $I_1, \ldots, I_n \subset \mathbb{R}$ be open, bounded intervals. The set $I_1 \times \ldots \times I_n \subset \mathbb{R}^n$ with the I_j parallel to the coordinate axes is said to be a parallelepiped.

Lemma 2.3. *Let $I = I_1 \times \ldots \times I_n$ and $J = J_1 \times \ldots \times J_n$ be two parallelepipeds in $\mathbb{R}^n$. Then $I \oplus J$ is a parallelepiped in $\mathbb{R}^n$.*

Proof. This follows from the fact that $I \oplus J = (I_1 + J_1) \times \ldots \times (I_n + J_n)$. □

Theorem 2.4 (Brunn-Minkowski Inequality)**.** *Let I and J be two bounded, open subsets of $\mathbb{R}^n$. Then*

$$|I|^{\frac{1}{n}} + |J|^{\frac{1}{n}} \leq |I \oplus J|^{\frac{1}{n}}.$$

The proof of this theorem consists of three parts: first, we will prove the inequality for parallelepipeds, then we will prove the inequality for disjoint finite unions of parallelepipeds, and finally, using an elementary result from Lebesgue theory, will prove the inequality in the general case.

Proof. Let $I = I_1 \times \ldots \times I_n$ and $J = J_1 \times \ldots \times J_n$ be two parallelepipeds in $\mathbb{R}^n$. Let $l(I_k)$ and $l(J_k)$ denote the length of I_k and J_k respectively. We then have

$$\begin{aligned}
\frac{|I|^{\frac{1}{n}} + |J|^{\frac{1}{n}}}{|I \oplus J|^{\frac{1}{n}}} &= \frac{\prod_{k=1}^n l(I_k)^{\frac{1}{n}} + \prod_{k=1}^n l(J_k)^{\frac{1}{n}}}{\prod_{k=1}^n (l(I_k) + l(J_k))^{\frac{1}{n}}} \\
&= \prod_{k=1}^n \left(\frac{l(I_k)}{l(I_k) + l(J_k)} \right)^{\frac{1}{n}} + \prod_{k=1}^n \left(\frac{l(J_k)}{l(I_k) + l(J_k)} \right)^{\frac{1}{n}} \\
&\leq \frac{1}{n} \sum_{k=1}^n \frac{l(I_k)}{l(I_k) + l(J_k)} + \frac{1}{n} \sum_{k=1}^n \frac{l(J_k)}{l(I_k) + l(J_k)} \\
&= 1
\end{aligned}$$

where the inequality follows from the AM-GM inequality. We conclude that the result holds for parallelepipeds.

Next, let I and J be disjoint finite unions of parallelepipeds, i.e. $I = \bigcup_{k=1}^{p} I_k$ and $J = \bigcup_{k=1}^{q} J_k$ where I_k and J_k are parallelepipeds as above. We proceed via induction on $p+q$. The base case when $p+q=2$ has already been established above. Suppose $p+q \geq 3$, with $p > 1$. As the I_k are disjoint, there exists a hyperplane H separating I_1 and I_2. Let H^+ and H^- be the two halfspaces resulting from H, and let $I^+ = I \cap H^+$ and $I^- = I \cap H^-$. Note that I^+ and I^- are also finite unions of parallelepipeds, i.e. $I^+ = \bigcup_{k=1}^{p^+} I_k^+$ and $I^- = \bigcup_{k=1}^{p^-} I_k^-$ where $p^+ < p$ and $p^- < p$ as P separates at least I_1 and I_2.

Now, we can find a hyperplane Q parallel to P separating J into J^+ and J^- such that

$$\frac{|I^+|}{|I|} = \frac{|J^+|}{|J|}.$$

This is guaranteed by the intermediate value theorem: simply start with Q parallel to P and to the "left" of J, and slide it right until it is to the "right" of J. The left hand side of the above expression is a fixed fraction between 0 and 1, and the expression on the right varies continuously from 0 to 1 during the above sliding procedure. Consequently, we are guaranteed a position where equality holds. Note that J^+ and J^- are also disjoint finite unions of parallelepipeds, i.e. $J^+ = \bigcup_{k=1}^{q^+} J_k^+$ and $J^- = \bigcup_{k=1}^{q^-} J_k^-$ where $q^+ \leq q$ and $q^- \leq q$ as Q may not separate two parallelepipeds of J.

Next, note that

$$\frac{|I^+|}{|I|} = \frac{|J^+|}{|J|} \implies \frac{|I^-|}{|I|} = \frac{|J^-|}{|J|}.$$

As $p^+ + q^+ < p+q$ and $p^- + q^- < p+q$, and so by the inductive hypothesis, we have

$$|I^+ \oplus J^+|^{\frac{1}{n}} \geq |I^+|^{\frac{1}{n}} + |J^+|^{\frac{1}{n}}$$

$$|I^- \oplus J^-|^{\frac{1}{n}} \geq |I^-|^{\frac{1}{n}} + |J^-|^{\frac{1}{n}}$$

Now, as P and Q are parallel, we have $P^+ \oplus Q^+ = (P \oplus Q)^+$. We then have $I^+ \oplus J^+ \subset P^+ \oplus Q^+ = (P \oplus Q)^+$, and similarly, $I^- \oplus J^- \subset P^- \oplus Q^- = (P \oplus Q)^-$. As $P \oplus Q$ is a hyperplane, we can conclude that $I^+ \oplus J^+$ and $I^- \oplus J^-$ are disjoint. It follows that

$$|I \oplus J| \geq |I^+ \oplus J^+| + |I^- \oplus J^-| \geq \left(|I^+|^{\frac{1}{n}} + |J^+|^{\frac{1}{n}}\right)^n + \left(|I^-|^{\frac{1}{n}} + |J^-|^{\frac{1}{n}}\right)^n.$$

In particular, using the fact that $\frac{|I^+|}{|I|} = \frac{|J^+|}{|J|}$ and $\frac{|I^-|}{|I|} = \frac{|J^-|}{|J|}$, we have

$$\begin{aligned}
|I \oplus J| &\geq \left(|I^+|^{\frac{1}{n}} + |J^+|^{\frac{1}{n}}\right)^n + \left(|I^-|^{\frac{1}{n}} + |J^-|^{\frac{1}{n}}\right)^n \\
&= |I^+| \left[1 + \left(\frac{|J|}{|I|}\right)^{\frac{1}{n}}\right]^n + |I^-| \left[1 + \left(\frac{|J|}{|I|}\right)^{\frac{1}{n}}\right]^n \\
&= |I| \left[1 + \left(\frac{|J|}{|I|}\right)^{\frac{1}{n}}\right]^n \\
&= \left(|I|^{\frac{1}{n}} + |J|^{\frac{1}{n}}\right)^n
\end{aligned}$$

which completes the induction.

Finally, let I and J be two bounded, open sets in $\mathbb{R}^n$. A standard result in Lebesgue theory states that there exist two sequences $\{I_n\}_{n\in\mathbb{N}}$ and $\{J_n\}_{n\in\mathbb{N}}$ where I_n, J_n are finite unions of disjoint parallelepipeds such that $I_n \subset I$ and $J_n \subset J$ for all $n \in \mathbb{N}$, and $|I_n| \to |I|$, $|J_n| \to |J|$ as $n \to \infty$. It follows that $I_n \oplus J_n \subset I \oplus J$ for all $n \in \mathbb{N}$, and hence we have

$$|I \oplus J|^{\frac{1}{n}} \geq |I_n \oplus J_n|^{\frac{1}{n}} \geq |I_n|^{\frac{1}{n}} + |J_n|^{\frac{1}{n}}$$

for all $n \in \mathbb{N}$. Taking the limit as $n \to \infty$, we get the desired result. □

Next, we make precise what we mean by the *boundary* or the *surface area* of a set in $\mathbb{R}^n$. Informally, the surface area is the rate at which the volume increases when we add a small ball to the object.

Definition 2.5 (surface area)**.** Let $K \subset \mathbb{R}^n$ be bounded with smooth boundary. The surface area of K, written $|\partial K|$, is defined as

$$|\partial K| = \lim_{\epsilon \to 0} \frac{|K \oplus \epsilon B_n| - |K|}{\epsilon}.$$

Finally, we can use the Brunn-Minkowski inequality to prove the classical isoperimetric inequality in $\mathbb{R}^n$.

Theorem 2.6 (Isoperimetric Inequality for $\mathbb{R}^n$)**.** *For any n-dimensional body K with volume $|K|$ and surface area $|\partial K|$, we have*

$$\frac{|K|^{n-1}}{|\partial K|^n} \leq \frac{|B_n|^{n-1}}{|\partial B_n|^n}.$$

Proof. By the Brunn-Minkowski inequality, we have

$$|K \oplus \epsilon B_n| \geq \left(|K|^{\frac{1}{n}} + |\epsilon B_n|^{\frac{1}{n}}\right)^n$$

but as $|\epsilon B_n|^{\frac{1}{n}} = (\epsilon^n |B_n|)^{\frac{1}{n}} = \epsilon |B_n|^{\frac{1}{n}}$, we have

$$|K \oplus \epsilon B_n| \geq \left(|K|^{\frac{1}{n}} + \epsilon |B_n|^{\frac{1}{n}}\right)^n = |K| \left(1 + \epsilon \left(\frac{|B_n|}{|K|}\right)^{\frac{1}{n}}\right)^n \geq |K| \left(1 + n\epsilon \left(\frac{|B_n|}{|K|}\right)^{\frac{1}{n}}\right)$$

where the last inequality above was obtained by throwing out all but the first two terms of the Maclaurin series of $(1+x)^n$.

We then have

$$|\partial K| = \lim_{\epsilon \to 0} \frac{|K \oplus \epsilon B_n| - |K|}{\epsilon} \geq \lim_{\epsilon \to 0} \frac{|K| \left(1 + n\epsilon \left(\frac{|B_n|}{|K|}\right)^{\frac{1}{n}}\right) - |K|}{\epsilon} = n|K|^{\frac{n-1}{n}} |B_n|^{\frac{1}{n}}$$

For an n-dimensional ball, we have $|\partial B_n| = n \times |B_n|$, which gives us

$$\frac{|\partial K|}{|\partial B_n|} \geq \left(\frac{|K|}{|B_n|}\right)^{\frac{n-1}{n}}$$

which can be rearranged to obtain the desired inequality. □

2.2. **Discrete Isoperimetric Inequalities**

Given a graph $G = (V, E)$, the *vertex boundary* of a subset $S \subset V$ is the set vertices of connected to at least one vertex in S, and the *edge boundary* of S is the set of edges crossing from S to its complement. The question, then, of the smallest possible boundary a set given its cardinality is an interesting combinatorial question in its own right, and is a natural analogue of classical isoperimetry. Furthermore, good estimates of these boundaries are also very useful for several applications; we briefly mention some:

- Lower bounds on the vertex boundary show how fast a neighbourhood of a set has to grow when the allowed distance from the set increases, and this leads to concentration of measure results for Lipschitz functions on the graph [MS86].
- Lower bounds on edge boundary directly provide upper bounds on the mixing time of random walks on the graph [Jer03].
- The notion of isoperimetry is closely related to that of *expansion* and can be used to characterize expander graphs, which have several applications in mathematics and theoretical computer science [Spi15].

The main player in this section will be the n-dimensional boolean hypercube Q_n. We will also briefly discuss isoperimetric results on other graphs. Parts of the following section are closely based on chapters from [Har04].

We start by formally defining what the vertex and edge boundaries of a set of nodes in a graph are.

Definition 2.7 (vertex boundary)**.** Given a graph $G = (V, E)$ and $S \subset V$, the *vertex boundary* of S is given by

$$\partial S = S \cup \{v \in V \mid (u, v) \in E \text{ for some } u \in S\}.$$

Definition 2.8 (edge boundary)**.** Given a graph $G = (V, E)$ and $S \subset V$, the *edge boundary* of S is given by

$$\partial_E S = \{(u, v) \in E \mid u \in S, v \in V \backslash S\}.$$

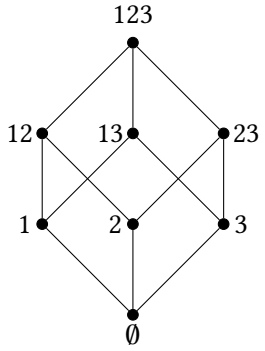

FIGURE 1. Q_3

2.2.1. *Isoperimetry on Q_n*

The *n-dimensional hypercube* is the n-dimensional analogue of the square ($n = 2$) and cube ($n = 3$). In what follows, $\mathscr{P}(X)$ denotes the *power set* of X (i.e. the set of all subsets of X).

Definition 2.9 (n-dimensional hypercube)**.** The n-dimensional hypercube Q_n is the graph with vertex set $V = \mathscr{P}(X)$ where $X = \{1, \ldots, n\}$, and edge set $E = \{(x, y) \mid \exists\, i \in X \text{ s.t. } x = \{i\} \cup y \text{ or } y = \{i\} \cup x \text{ for } x, y \in V\}$.

We will usually write $X = [n]$ instead $\{1, \ldots, n\}$. Thus, two vertices $x, y \in Q_n$ are adjacent if $|x \Delta y| = 1$ where Δ denotes the symmetric difference. For convenience, we will often ignore brackets and commas while denoting vertices of Q_n. In other words, $\{1, 2, 3\}$ can be written as 123. Figure 2.2.1 illustrates Q_3.

Alternatively, the n-dimensional hypercube can also be thought of as the graph with the set of 0-1 sequences of length n as its vertex set, with two vertices adjacent if and only if the corresponding 0-1 sequences differ at exactly one position.

Let $X^{(r)} = \{x \in \mathscr{P}(X) \mid |x| = r\}$, and $X^{(\leq r)} = \{x \in \mathscr{P}(X) \mid |x| \leq r\}$; we can define similar sets for "$\geq$", "$<$", and "$>$" instead of "$\leq$".

Now, given $0 \leq m \leq 2^n$, what $A \subset Q_n$ with $|A| = m$ minimizes $|\partial A|$? This question was answered rather beautifully by Harper in [Har66], who proved that the minimizers of $|\partial A|$ are the first m elements of $\mathscr{P}(X)$ taken in the *simplicial* order.

Definition 2.10 (simplicial order)**.** For $x, y \in \mathscr{P}(X)$, we have $x < y$ in the *simplicial order* if either $|x| < |y|$ or $|x| = |y|$ and $\min(x \Delta y) \in x$. Let I_m denote the first m elements of Q_n in the simplicial order.

FIGURE 2. Simplicial order on Q_n

We will use the notion of *compressions* to prove Harper's theorem. Informally, we will *compress* A to obtain a set A' such that $|A'| = |A|$ and $|\partial A'| \leq |\partial A|$, and A' is "closer" to I_m than A. All of this is made formal below.

Definition 2.11 (i-sections)**.** Let $A \subset \mathscr{P}(X)$ and let $1 \leq i \leq n = |X|$ where $X = [n]$. The *i-sections* on A are given by $A_{i-} = \{x \in \mathscr{P}(X\backslash\{i\}) \mid x \in A\} \subset Q_{n-1}^{(i-)}$ and $A_{i+} = \{x \in \mathscr{P}(X\backslash\{i\}) \mid x \cup \{i\} \in A\} \subset Q_{n-1}^{(i+)}$ where $Q_{n-1}^{(i-)}$ and $Q_{n-1}^{(i+)}$ are copies of Q_{n-1} labelled by sets of $\mathscr{P}(X\backslash\{i\})$.

Definition 2.12 (i-compression)**.** Given a set $A \subset \mathcal{P}(X)$ where $X = [n]$, we define the *i-compression* of A to be $C_i(A)$, given by its i-sections $C_i(A)_{i-}$ and $C_i(A)_{i+}$ where $C_i(A)_{i-}$ is the set of first $|A_{i-}|$ elements of $\mathcal{P}(X\backslash\{i\})$ in the simplicial order, and $C_i(A)_{i+}$ is the set of first $|A_{i+}|$ elements of $\mathcal{P}(X\backslash\{i\})$ in the simplicial order.

Definition 2.13 (i-compressed set)**.** We say that $A \subset \mathcal{P}(X)$ where $X = [n]$ is *i-compressed* if $C_i(A) = A$.

The following example should make the above definitions clear.

Example 2.14. The simplicial order on Q_3 is $\{\emptyset, 1, 2, 3, 12, 13, 23, 123\}$. Let $X = [4]$, and $A = \{1, 14, 23, 123\} \subset Q_4$. Then $A_{2-} = \{1, 14\}$, and $A_{2+} = \{3, 13\}$. Moreover, we have $C_2(A)_{2-} = \{\emptyset, 1\}$ and $C_2(A)_{2+} = \{\emptyset, 1\}$ which gives us $C_2(A) = \{\emptyset, 1\} \cup \{2, 12\} = \{\emptyset, 1, 2, 12\}$.

Theorem 2.15 (Harper's Theorem)**.** *Let $A \subset Q_n$ with $|A| = m$ and let $I_m \subset Q_n$ be the first m elements of Q_n in the simplicial order. Then $|\partial A| \geq |\partial I_m|$.*

Proof. We proceed via induction on n. The case when $n = 1$ is trivial. Let $C = C_i(A)$ be an i-compression of A. Note that $|C| = |C_i(A)_{i-}| + |C_i(A)_{i+}| = |A_{i-}| + |A_{i+}| = |A|$. We will attempt to show that $|\partial C| \leq |\partial A|$. In particular, this would follow if we could show $|(\partial C)_{i-}| \leq |(\partial A)_{i-}|$ and $|(\partial C)_{i+}| \leq |(\partial A)_{i+}|$ from an argument similar to that used to obtain the above expression.

Now, note that $(\partial A)_{i-} = \partial(A_{i-}) \cup A_{i+}$. The reverse inclusion is clear, and in order to see the forward inclusion, note that if $x \in (\partial A)_{i-}$ such that $x \notin A_{i+}$, then $x \cup \{i\} \notin A$ and $x \cup \{j\} \in A$ for some $j \in [n]$. But then $x \cup \{j\} \in A_{i-} \implies x \in \partial(A_{i-})$. Similarly, we have $(\partial C)_{i-} = \partial(C_{i-}) \cup C_{i+}$. By the inductive hypothesis, we have $|\partial(C_{i-})| \leq |\partial(A_{i-})|$. We also have $|C_{i+}| = |A_{i+}|$. Now, note that if C_{i-} is an initial segment of the simplicial order on

$\mathscr{P}(X\backslash\{i\})$, then so are $\partial(C_{i-})$ and C_{i+}. In particular, they must be nested and so we must have either $\partial(C_{i-}) \subset C_{i+}$ or $C_{i+} \subset \partial(C_{i-})$. Consequently, $(\partial C)_{i-} = \partial(C_{i-})$ or $(\partial C)_{i-} = C_{i+}$. In either case, we have $|(\partial C)_{i-}| \leq |(\partial A)_{i-}|$.

An identical argument gives $|(\partial C)_{i+}| \leq |(\partial A)_{i+}|$, and so we have $|\partial C| \leq |\partial A|$ as desired.

Next, define a sequence $A = A_0, A_1, \ldots$ as follows: if A_j is i-compressed for all $i \in [n]$, then terminate the sequence. Otherwise, there must exist some i for which A_j is not i-compressed; define $A_{j+1} = C_i(A_j)$. But why must this process terminate? For $x \in A_j$, let $f(x)$ denote the position of x in the simplicial order on Q_n. Then, as $\sum_{x\in A_j} f(x)$ is a decreasing function in j, the above process must terminate, say at some A_k.

Now, if A_k were an initial segment of simplicial order, then the result would follow immediately. This, however, is not always the case (for example, $\{\emptyset, 1, 2, 12\} \subset Q_3$ is i-compressed for all $i \in [3]$). However, Lemma 2.16 tells us that if a set is i-compressed for all $i \in [n]$ and is not an initial segment of simplicial order, then there are precisely two possibilities for it. In either case, the boundary has greater size than the corresponding initial segment of simplicial order with same size, completing the proof. □

Lemma 2.16. *Let $A \subset \mathscr{P}(X)$ be i-compressed for all $i \in [n]$ where $X = [n]$. If A is not an initial segment of simplicial order, then*

- *if n is odd, then $A = X^{(<n/2)}\backslash\{\{\frac{n+3}{2}, \frac{n+5}{2}, \ldots, n\}\} \cup \{\{1, 2, \ldots, \frac{n+1}{2}\}\}$.*
- *if n is even, then $A = X^{(<n/2)} \cup \{x \in X^{(n/2)} \mid 1 \in x\}\backslash\{\{1, \frac{n}{2}+2, \frac{n}{2}+3, \ldots, n\}\} \cup \{\{2, 3, \ldots, \frac{n}{2}+1\}\}$.*

Proof. Note that as A is not an initial segment of simplicial order, there exist $x, y \in \mathscr{P}(X)$ with $x < y$ in the simplicial order such that $x \notin A$ and $y \in A$. Now, note that if there exists an $i \in [n]$ such that $i \notin x$ and $i \notin y$, then $x, y \in C_i(A)_{i-}$, which contradicts the fact that B is i-compressed. Similarly, we cannot have $i \in [n]$ such that $i \in x$ and $i \in y$. Thus, for each $i \in [n]$, either $i \in x$ or $i \in y$, i.e. $y = X\backslash x$.

Consequently, if $z < y$ and $z \notin A$, then $z = x$ (as $z = X\backslash y$). Similarly, if $x < z$ and $z \in A$, then $z = y$. In particular, this implies that $z \in A$ whenever $z < x$ and $z \notin A$ if $z > y$. If $x < z < y$, then we cannot have $z \in A$ as $x < z$, and we also cannot have $z \notin A$ as $z < y$, a contradiction. Thus, $y = X\backslash x$ is the immediate successor of x in the simplicial order, which gives us $A = \{z \in \mathscr{P}(X) \mid z \leq y\}\backslash x$.

If n is odd, then $|x| + 1 = |y|$, i.e. x is the last set of $X^{(n-1/2)}$ in simplicial order. If n is even, then $|x| = |y|$, and we note that the only sets satisfying the above conditions are $x = \{1, \frac{n}{2}+2, \frac{n}{2}+3, \ldots, n\}$ and $y = \{2, 3, \ldots, \frac{n}{2}+1\}$. This completes the proof of the lemma. □

Having obtained a vertex isoperimetric inequality on Q_n, we turn our attention towards edge isoperimetry on Q_n. Suppose we want to minimize $|\partial_E(A)|$ given $A \subset Q_n$. Do initial segments of simplicial order again minimize edge boundary? Consider the case when

$n = 3$ and $|A| = 4$. Then $I_4 \subset Q_3$ has $|\partial_E(I_4)| = 6$, but $|\partial_E(Q_2)| = 4$ where Q_2 is the co-dimension 1 *subcube* of Q_3.

Experiments suggest that subcubes are minimizers of the edge boundary, but which vertices do we choose if $|A|$ is not a power of 2? Define the following order on $\mathscr{P}(X)$:

Definition 2.17 (binary order)**.** For $x, y \in \mathscr{P}(X)$, we have $x < y$ in the *binary order* if $\max(x \Delta y) \in y$. Equivalently, $x < y$ in the binary order if $\sum_{i \in x} 2^i < \sum_{i \in y} 2^i$.

Definition 2.18 (i-binary compression)**.** For $A \subseteq Q_n$ and $1 \leq i \leq n$, the *i-binary compression* of A is $B_i(A) \subset Q_n$ given by its i-sections $B_i(A)_{i-}$ and $B_i(A)_{i+}$ where $B_i(A)_{i-}$ is the set of first $|A_{i-}|$ elements of $\mathcal{P}(X \backslash \{i\})$ in the binary order, and $B_i(A)_{i+}$ is the set of first $|A_{i+}|$ elements of $\mathcal{P}(X \backslash \{i\})$ in the binary order.

For example, the binary ordering on Q_3 is $\{\emptyset, 1, 2, 12, 3, 13, 23, 123\}$. The following result states that the initial segments of binary order are minimizers of edge boundary on Q_n. This result is also known as the Theorem of Harper, Lindsey, Bernstein and Hart, and was initially solved with a coding theory application in mind.

Theorem 2.19 (edge isoperimetric inequality for Q_n)**.** *Let $A \subset Q_n$ and C be the first $|A|$ elements of Q_n in the binary order. Then $|\partial_E A| \geq |\partial_E C|$.*

Proof. Once again, we proceed via induction on n. The base case when $n = 1$ is trivial. Let $B = B_i(A)$ be a i-binary compression of A. We will attempt to show that $|\partial_E B| \leq |\partial_E A|$. Note that for A and B, we have $\partial_E A = \partial_E A_{i-} \cup \partial_E A_{i+}$ and $\partial_E B = \partial_E B_{i-} \cup \partial_E B_{i+}$. Now, $|\partial_E B_{i-}| \leq |\partial_E A_{i-}|$ and $|\partial_E B_{i+}| \leq |\partial_E A_{i+}|$ by the inductive hypothesis. Also, $|B_{i+}| = |A_{i+}|$, $|B_{i-}| = |A_{i-}|$. Finally, as B_{i+} and B_{i-} are nested, we get that $|\partial_E B| \leq |\partial_E A|$.

Define a sequence $A = A_0, A_1, \ldots$ as we did in the proof for Theorem 2.15, except using i-binary compressions instead of i-compressions. Once again, this sequence must terminate (say at A_k); however, this set need not be an initial segment of the binary order. For example, $\{\emptyset, 1, 2, 3\}$ in Q_3 is i-binary compressed for all $i \in [3]$, but is not an initial segment of binary order. Lemma 2.20 allows us to conclude, however, that if A_k is not an initial segment of binary order, then $|\partial_E A_k| \geq |\partial_E C|$ where C is as in the statement of the theorem. □

Lemma 2.20. *Let $A \subset Q_n$ be i-binary-compressed for all $i \in [n]$ but not an initial segment of the binary order. Then $A = Q_{n-1} \cup \{n\} - \{\{1, 2, 3, \ldots, (n-1)\}\}$.*

Proof. As before, we have $x \notin A$ and $y \in A$ for some $x < y$ in the binary order. Once again, we must have $y = X \backslash x$, and so $A = \{z \mid z \leq y \text{ IN THE BINARY ORDER}\} \backslash \{x\}$. As x and y are consecutive in the binary order, x must be the last point with $n \notin x$ and y is the first point with $n \in y$. □

We conclude this section by noting that in both proofs the extremal sets formed a nested family.

2.2.2. *Other Isoperimetric Problems*

In general, very few exact isoperimetric inequalities in discrete spaces are known. Some of these are listed in Table 1.

Graph	Edge-optimal shapes	Vertex-optimal shapes
Q_n	Subcubes	Hamming balls
$(\mathbb{Z}^d, l_1)$	Cubes	Cross-polytopes
$(\mathbb{Z}^d, l_\infty)$	Zonotopes	Cubes

TABLE 1. Some known edge-optimal and vertex-optimal shapes

An approximation to the size of the edge boundary (and consequently, a lower bound on the size of the vertex boundary) can be obtained via Cheeger's inequality, which bounds the size of the edge boundary of a graph using the second-largest eigenvalue of the adjacency matrix of the graph.

Theorem 2.21 (Cheeger's inequality)**.** *Let $G = (V, E)$ be a finite, connected, d-regular graph and let λ_2 be the second-largest eigenvalue of its adjacency matrix. Then we have*

$$\frac{d - \lambda_2}{2} \leq h(G) \leq \sqrt{2d(d - \lambda_2)}$$

where $h(G) = \min_{S \subset V : |S| \leq \frac{n}{2}} \frac{|\partial_E(S)|}{|S|}$ is the edge expansion *of G.*

We omit the proof of Cheeger's inequality due to its length; an accessible proof can be found in [Spi15].

3. The Protein Folding Problem

Proteins are large biomolecules comprising of one or more residues of amino acids, and perform a large variety of functions within the cell. The function of a protein, however, is intimately linked to its three-dimensional structure or *fold*. Moreover, experiments suggest that this *fold* of a protein is dictated entirely by its one-dimensional sequence.

Thus, the *protein folding problem* asks to determine the optimal fold of a protein given its one-dimensional amino-acid sequence. Unfortunately, under any reasonably biophysical model, this problem is **NP**-hard. A natural simplification of this problem, then, would be to isolate the major driving forces behind the folding process. This led to the introduction of the *hydrophobic-hydrophilic* or the HP model for protein folding [Dil85].

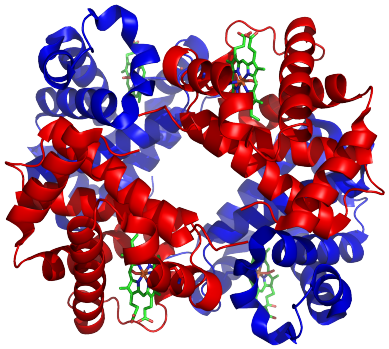

FIGURE 3. Structure of human hemoglobin

The HP model is based off the axiom that interactions between hydrophobic amino acid residues are the major driving force behind protein folding. Consequently, we can model a protein as a self-avoiding walk consisting of two kinds of residues, namely *hydrophobic* (H) and *hydrophilic* (P) residues. The objective, then, is to find a conformation of the HP-polymer that maximizes the number of H-H contacts.

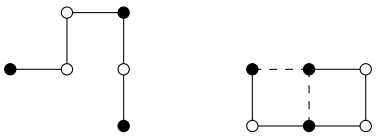

FIGURE 4. Two Conformations of HPPHPH on $\mathbb{Z}^2$

For convenience, we usually restrict the folding to a graph where the residues are placed at nodes and adjacent residues in the polymer must be connected by an edge in the graph. Figure 4 shows two conformations of a HP polymer on the two-dimensional square grid.

Unfortunately, even when restricted to lattices, protein folding in the HP model remains computationally intractable [HI97][Cre+98]. Several approximation algorithms for HP folding, however, do exist for graphs such as $\mathbb{Z}^2$, $\mathbb{Z}^3$, and the three-dimensional face-centered cubic (FCC) lattice. In order to evaluate the performance of these approximation algorithms, upper bounds on the maximum possible number of H-H contacts are desirable.

For example, given a HP polymer on $\mathbb{Z}^2$ with n hydrophobic residues, an easy upper bound can be obtained by noting that each non-start or end H residue can make at most 2 contacts, but as we double count each contact, we can have at most $\frac{2n}{2} = n$ contacts. A slightly better bound can be obtained for linear polymers on $\mathbb{Z}^2$ by noting that $\mathbb{Z}^2$ is bipartite; therefore,

	Parity	Isoperimetric
$H^2(P^2H)^7H$	11	10
$P^2HP^2(H^2P^4)^3H^2$	8	8

TABLE 2. Comparison of the parity and isoperimetric bounds on $\mathbb{Z}^2$

if we were to index the residues in a HP polymer, then hydrophobic contacts can only be made between residues differing in parity.

Note, however, that protein folding in the HP model can be modeled as an isoperimetric problem by attempting to minimize the number of *missed* contacts. In other words, given a fixed HP polymer, we wish to arrange the H residues so as to minimize the number of H-P and H-W contacts where W denotes an empty lattice point or node.

In other words, we wish to minimize the *edge boundary* of the set of H residues, where the boundary is defined as above. As the edge-optimal shapes on $\mathbb{Z}^2$ are squares, we can obtain upper bounds on the number of H-H contacts made by linear polymers by packing them into squares. A comparison between the parity bound and our isoperimetric bound is given in Table 2.

This section is very much a work-in progress, and we intend to update this document with a detailed write-up of our isoperimetric bound in the near future. We conclude by noting that our isoperimetric bound generalizes the results of, and in the case of $\mathbb{Z}^2$, is identical to their upper bound on the number of H-H contacts obtained by [GP13].

4. Bipole Packing and Related Problems

In the previous section, we considered proteins as *linear* polymers of hydrophobic and hydrophilic residues. This, however, is almost never the case with real-world proteins which consist of a few to a few dozen branched polymers held together by a backbone.

A natural first order approximation to real-world folding, then, would be to consider proteins to be made of a *backbone* of hydrophilic residues, together with several *bipoles* each of which is one hydrophobic residue connected to a backbone hydrophilic residue.

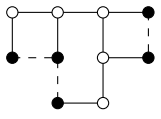

FIGURE 5. Side-chain Model for HP Folding

This model for folding is called the *side-chain model* and was introduced by Istrail [IL09]. In this section, our primary focus will not be the proteins in the side-chain model, but

the side-chains (i.e. bipoles) themselves). We will study several combinatorial problems focused on bipoles and their assemblies.

4.1. **Bipole Packing**

Assemblies or *packings* of bipoles occur commonly in nature as micelles, vesicles, and perhaps most notably as the phospholipid bilayer found in cell membranes. A natural question to ask, then, is whether bipoles naturally arrange into bi-layered or *biplanar* configurations? In other words, is the biplane the optimal arrangement of bipoles?

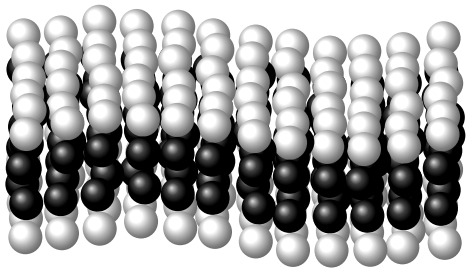

FIGURE 6. Phospholipid bilayer in cell membranes

Istrail's *biplane conjecture* states that this is indeed the case for a reasonable class of graphs, namely space-filling lattices. In this section, we will prove Istrail's biplane conjecture on the 2D triangular and 3D cubic lattices.

Definition 4.1 (bipole)**.** A *bipole* on a graph $G = (V, E)$ is an ordered pair of vertices (u, v) such that $(u, v) \in E$. Given a bipole $(u, v) \in E$, we will refer to the first vertex u as the *hydrophobic residue*, and the second vertex v as the *hydrophilic residue*.

Definition 4.2 (BPP)**.** An instance (G, n) of the BIPOLE-PACKING-PROBLEM (BPP) consists of a graph $G = (V, E)$ and an integer $n \leq |V|/2$. A *feasible solution* to BPP(G, n) is a pair $(\mathcal{H}, \mathcal{P})$ where $\mathcal{H}, \mathcal{P} \subset V$ and $|\mathcal{H}| = |\mathcal{P}| = n$ such that:

- $(\mathcal{H}_i, \mathcal{P}_i)$ is a bipole for $i \in \{1, \ldots, n\}$.
- No vertex appears more than once in $\mathcal{H} \cup \mathcal{P}$.

An *optimal solution* to BPP(G, n) is a feasible solution $(\mathcal{H}, \mathcal{P})$ such that $|\{E(G(\mathcal{H}))\}|$ is maximized over all feasible solutions, where $G(\mathcal{H}) \subset G$ is the subgraph induced by $\mathcal{H}$.

We will often refer to a feasible solution to BPP as a *packing*, the sets $\mathcal{H}$ and $\mathcal{P}$ as the sets of hydrophobic and hydrophilic residues respectively, and $E(G(\mathcal{H}))$ as the set of *contacts* among hydrophobic residues, i.e. any edge between two hydrophobic residues is a contact. The objective of BPP, therefore, is to find a configuration of bipoles on a graph that maximizes the number of hydrophobic contacts.

We will now show that BPP is **NP**-hard via a reduction from **NP**-complete problem Densest-k-Subgraph (DkS), a natural generalization of the Max-Clique problem.

Definition 4.3 (DkS)**.** An instance of (G, k) of Densest-k-Subgraph (DkS) is a graph $G = (V, E)$ and an integer $k < |V|$. A *feasible solution* to DkS(G, k) is a subgraph $H \subset G$ such that $|V(H)| = k$, and an *optimal* solution is a feasible solution H that maximizes $|E(H)|$ over all feasible solutions.

Theorem 4.4. BPP *is* **NP***-hard.*

Proof. Consider the following reduction from DkS(G, k) to BPP: Create $G' \cong G$, and for each vertex $v \in V(G)$, add a new vertex $v' \in V(G')$ and edge $(v, v') \in E(G')$. This is illustrated in Figure 4.1.

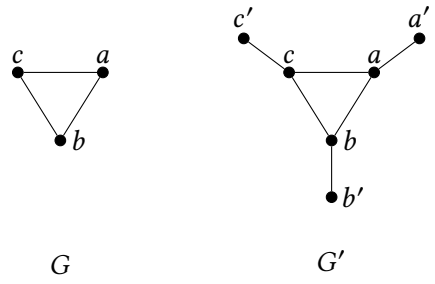

Figure 7. Reduction gadget for Theorem 4.4

- Given an optimal solution $H \subset G$ to DkS(G, k), it's clear that setting $\mathcal{H} = \{v \in G' \mid v \in H\}$ with the corresponding $\mathcal{P}$ residues at the added vertices is an optimal solution to BPP(G', k).
- Conversely, given an optimal instance $\mathcal{H}$ of BPP(G', k), note that if any of the added vertices v' contain a $\mathcal{H}$ residue, we may *flip* the bipole (v', v) so as to obtain a the $\mathcal{H}$ residue on a vertex contained in the original graph. As each of the added vertices are only connected to one other vertex, this operation cannot decrease the number of $\mathcal{H} - \mathcal{H}$ contacts. Again, it's easy to see that $H = \mathcal{H}$ is an optimal solution to DkS, since lifting a better solution of DkS to BPP would contradict the optimality of $\mathcal{H}$.

Clearly this reduction runs in time polynomial in the size of G, and so we conclude that BPP is **NP**-hard. □

Although BPP is **NP**-hard on arbitrary graphs, we shall soon obtain exact solutions on some particularly nice graphs, namely certain two and three dimensional lattices.

4.1.1. BPP *on* $\mathbb{Z}^2$ *and* $\mathbb{T}^2$

An exact solution for BPP on the two-dimensional square grid, namely $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$, was obtained in [IL09], which we reproduce below.

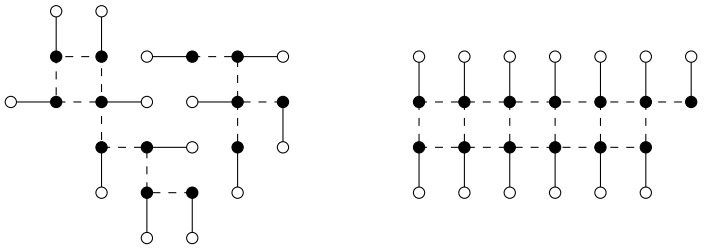

FIGURE 8. Packings of $n = 13$ bipoles on $\mathbb{Z}^2$

Theorem 4.5 (Biplane Conjecture on $\mathbb{Z}^2$)**.** *The optimal packing of bipoles on the two-dimensional square lattice is a bi-planar arrangement.*

Proof. For any packing $(\mathcal{H}, \mathcal{P})$ of n-bipoles, consider drawing lines parallel to the x and y-axes through each $h \in \mathcal{H}$. Note that each hydrophobic residue h divides the lattice into four quadrants $Q_1(h), Q_2(h), Q_3(h), Q_4(h)$ as shown in Figure 9.

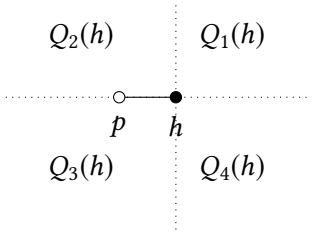

FIGURE 9. Quadrants generated by a hydrophilic residue.

Each quadrant $Q_i(h)$ includes its boundary lines, and for each $h \in \mathcal{H}$, let $q_i(h)$ be the number of hydrophobic residues in $\mathcal{H}$ besides h. We claim that for each $i \in \{1, \ldots, 4\}$, there exists a vertex h_i such that $q_i(h_i) = 0$, i.e. $Q_i(h_i)$ contains no hydrophobic residues besides h_i.

Indeed, suppose, for some fixed $1 \leq i \leq 4$, there is no $h_i \in \mathcal{H}$ such that $q_i(h_i) = 0$. This implies that $q_i(h) > 0$ for all $h \in \mathcal{H}$; pick $h \in \mathcal{H}$ such that $q_i(p)$ is minimized. It follows that there must exist some hydrophilic residue $h' \in Q_i(h)$ and $h \neq h'$, but note that $q_i(h') < q_i(h)$, contradicting our choice of h.

Now, each $h \in \mathcal{H}$ can make at most three contacts, as of its four adjacent vertices, one is occupied by a hydrophilic residue. Moreover, each $Q_i(h_i)$ results in one fewer contact. Consequently, the maximum number of contacts in any arrangement of n bipoles is $3n-4$, but as we double counted each contact, we get an upper bound of $\lfloor \frac{3n-4}{2} \rfloor$ on the number of contacts. Note, however, that in a bi-planar arrangement, each hydrophobic residue makes exactly 3 contacts except for the four residues corresponding to the corners of the bilayer (which lost 4 contacts for even n, and 5 contacts for odd n). It follows that the biplane meets the above upper bound. □

Next, we will consider BPP on the two-dimensional triangular lattice, $\mathbb{T}^2$. This lattice offers the densest packing of spheres—both on-lattice as well as off-lattice—in two-dimensional space, and thus, in a sense, is the best approximation to two-dimensional off-lattice BPP.

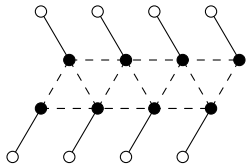

FIGURE 10. BPP on $\mathbb{T}^2$

Note that the main idea behind the proof of Theorem 4.5 carries over to $\mathbb{T}^2$: We can show that there exist $\{h_i\}_{i\in[6]}$ such that each of them *misses* one potential hydrophobic contact. As each vertex in $\mathbb{T}^2$ has degree 6, each hydrophobic residue can make at most 5 hydrophobic contacts, resulting in an upper bound of $\frac{5n-6}{2} = 2.5n-3$ on the number of hydrophobic contacts for any packing on $\mathbb{T}^2$. A biplane, however, makes $\frac{4n-6}{2} = 2n-3$ contacts.

Examining the two expressions above provides some intuition about BPP on $\mathbb{T}^2$: While a hydrophobic residue *can* make 5 contacts on $\mathbb{T}^2$, the optimal configuration prefers 4 contacts *on average*. In particular, this suggests that every time we have a hydrophobic residue making 5 contacts, there must exist one that makes 3 or fewer contacts.

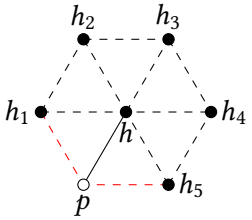

FIGURE 11. Collisions in BPP on $\mathbb{T}^2$

Indeed, as $\mathbb{T}^2$ is not bipartite (unlike $\mathbb{Z}^2$), every time a hydrophobic residue makes 5 contacts, it induces two *collisions* as shown in Figure 11: the bipole (h,p) makes 5 contacts, but the residues h_1 and h_5 can make at most 4 contacts as one of their neighboring vertices is occupied by p, the hydrophilic residue corresponding to the bipole (h,p).

Definition 4.6 (collision)**.** Let $(\mathcal{H},\mathcal{P})$ be a feasible solution to BPP(G,n), and let (h_1,p_1) be a bipole. A hydrophobic contact (h_1,h_2) is a *collision* if $(h_2,p_1)\in E$. Moreover, we say that the collision (h_1,h_2) is *induced* by h_1 and *absorbed* by h_2.

Definition 4.7 (collision count)**.** To each hydrophobic residue $h\in\mathcal{H}$ in a feasible solution $(\mathcal{H},\mathcal{P})$ to BPP(G,n) we associate a number C_h which is the number of collisions induced by h minus the number of collisions absorbed by h.

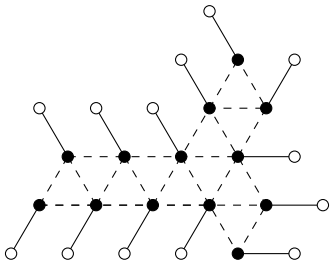

FIGURE 12. Non-biplanar optimal structure on $\mathbb{T}^2$ for $n = 13$ bipoles

More concretely, in Figure 11, the collisions are indicated with red dashed lines. Note that h induces 2 collisions (namely (h, h_1) and (h, h_5)), and absorbs 0 collisions; consequently, $C_h = 2$. The notion of a collision has appeared before: [Aga+97] use this notion to develop a set of local rules for protein folding in the HP model on $\mathbb{T}^2$ and the face-centered cubic lattice.

Theorem 4.8 (Biplane Conjecture for $\mathbb{T}^2$)**.** *The optimal packing of bipoles on the two-dimensional triangular lattice is a bi-planar arrangement.*

Proof. Let $(\mathcal{H}, \mathcal{P})$ be a feasible solution to BPP$(\mathbb{T}^2, n)$. Now, $\sum_{h \in \mathcal{H}} C_h = 0$, as every conflict induced by a residue is absorbed by another. Furthermore, if a hydrophobic residue h makes 5 contacts, then $C_h = 2$ as seen in Figure 11, and every residue h that makes 4 contacts has $C_h \geq 0$. Consequently, as $\sum_{h \in \mathcal{H}} C_h = 0$, we must have at least as many hydrophobic residues making 3 or fewer contacts as we have hydrophobic residues making 5 contacts.

It follows that the total number of contacts made by n bipoles is $\leq 4n$, and by an argument similar to that in the proof for Theorem 4.5, we have an upper bound of $\frac{4n-6}{2} = 2n - 3$ on the number of contacts. We conclude by noting that the biplane meets this bound. □

We note that non-trivial, non-biplanar solutions to BPP do exist on $\mathbb{T}^2$; these solutions, however, are always *locally* biplanar.

4.1.2. BPP *on* $\mathbb{Z}^3$

BPP$(\mathbb{Z}^3, n)$ was first studied in [CI99] where upper bounds showing that the biplane was within 94% of optimal were obtained. In [IL13], the biplane conjecture was verified via simulations for values up to $n = 150$. Here, we prove the optimality of the biplane via a rather elegant and surprisingly elementary argument.

Theorem 4.9 (Biplane Conjecture for 3D Cubic)**.** *The optimal packing of bipoles on the three-dimensional cubic lattice is a bi-planar arrangement.*

Proof. We are given n HP-bipoles on the three-dimensional cubic lattice, and we want to find the configuration that maximizes H-H adjacency. Suppose the optimal arrangement

has $k \geq 2$ layers, and let n_i be the number of H-residues in the i^{th} layer. We then have $n_1 + n_2 + \ldots + n_k = n$.

It follows from Theorem 8 in [BL91] that given a set of nodes $A \subset \mathbb{Z}^2$ (the integer grid), the edge boundary of A has size least $2 \times \lceil 2\sqrt{|A|} \rceil$.

From this result, in the i^{th} layer of any arrangement of the bipoles, we lose *at least* $2 \times \lceil 2\sqrt{n_i} \rceil$ H-H contacts. In total, we lose at least:

$$2 \times (\lceil 2 \times \sqrt{n_1} \rceil + \ldots + \lceil 2 \times \sqrt{n_k} \rceil)$$

H-H contacts.

However, note that given two numbers $a, b \geq 0$, we have $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$. This in turn implies:

$$\begin{aligned}
2 \times (\lceil 2 \times \sqrt{n_1} \rceil + \ldots + \lceil 2 \times \sqrt{n_k} \rceil) &\geq 2 \times \lceil 2\sqrt{n_1} + \ldots 2\sqrt{n_k} \rceil \\
&\geq 2 \times \lceil 2(\sqrt{n_1} + \ldots + \sqrt{n_k}) \rceil \\
&\geq 2 \times \lceil 2(\sqrt{n_1 + n_2} + \ldots \sqrt{n_k}) \rceil \\
&\vdots \\
&\geq 2 \times \lceil 2(\sqrt{n_1 + \ldots n_k}) \rceil \\
&= 2 \times \lceil 2 \times \sqrt{n} \rceil
\end{aligned}$$

Thus, any arrangement of n bipoles on at least two layers loses at least $2 \times \lceil 2 \times \sqrt{n} \rceil$ contacts, giving us the following bound:

$$\#\text{ hydrophobic contacts} \leq \frac{5n - 2 \times \lceil 2 \times \sqrt{n} \rceil}{2}$$

A biplanar structure meets this bound does meet this bound, and so we're almost done—all that remains is to show that the biplane is the better than any one-layered arrangement, which we omit. □

4.2. **Bipole Labelling**

We have attempted to pack or arrange bipoles so as to maximize the number of hydrophobic contacts. What if, instead of free-floating bipoles, their positions were "fixed" up to flipping the orientation of a bipole? We can therefore define the following problem:

Definition 4.10 (Bipole-Labelling)**.** Given a perfect matching $\mathcal{M}$ of a graph $G = (V, E)$, orient each edge in $\mathcal{M}$ as a bipole so as to maximize the number of hydrophobic contacts. The set of potential contacts is given by C.

This problem can be thought of, in some sense, as an inverse to BPP; alternatively, it can also be thought of as a constrained variant of BPP. Similar problems arise in various scenarios in statistical physics, most prominently in the context of spin glasses and as spin

labelling problems [SM99]. Below, we show that Bipole-Labelling is intractable on arbitrary graphs via a reduction from the **NP**-complete problem Max-2-SAT.

Definition 4.11 (Max-2-Sat)**.** Given a boolean formula in conjunctive normal form with exactly two literals per clause, find an assignment to the variables that maximizes the number of satisfied clauses.

Proposition 4.12. *The* Bipole-Labelling *problem is* **NP***-complete.*

Proof. Given an instance of Bipole-Labelling, we can nondeterministically check all possible labelings of the bipoles, and choose the one that has the maximum number of contacts in nondeterministic polynomial time. Thus Bipole-Labelling is in **NP**. Given an instance of Max-2-SAT in n variables $(x_i)_{i\in[n]}$ for some $n \in \mathbb{N}$, set up a bipole configuration assigning an unlabelled bipole $(x_i, \overline{x}_i)$ for all $i \in [n]$. For $i, j \in [n]$:

(1) Given clause (x_i, x_j), place potential contacts between (x_i, x_j), $(x_i, \overline{x}_j)$, and $(\overline{x}_i, x_j)$.

(2) Given clause $(x_i, \overline{x}_j)$, place potential contacts between $(\overline{x}_i, \overline{x}_j)$, $(x_i, \overline{x}_j)$, and (x_i, x_j).

(3) Given clause $(\overline{x}_i, \overline{x}_j)$, place potential contacts between $(\overline{x}_i, \overline{x}_j)$, $(x_i, \overline{x}_j)$, and $(\overline{x}_i, x_j)$.

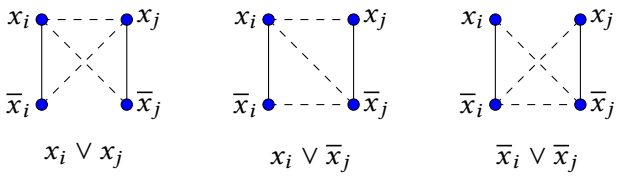

Figure 13. Reduction from Max-2-SAT to Bipole-Labelling

We claim that a maximal-contact labelling of the bipoles corresponds to a maximal assignment for the Boolean formula where the H-residue of a bipole corresponds to the equivalent variable being assigned 1. Indeed, as each bipole can only have one H-residue, only one of $x_i, \overline{x}_i$ can be assigned true (and vice versa). Moreover, in each of the three contact-edges placed in each of the cases above, only one edge can ever be an H-H edge, and every H-H contact is in one-to-one correspondence with a satisfied clause.

Clearly the above reduction can be performed in time polynomial in the size of the input, and so Bipole-Labelling is **NP**-complete. □

5. Conclusion

To summarize, we were able to obtain non-trivial bounds on linear polymers via discrete isoperimetric inequalities. We were also able to prove Istrail's biplane conjecture on the 2D triangular and 3D cubic lattices, and proved Bipole-Labelling to be **NP**-hard. Below, we indicate some possible directions for future work.

5.1. **BPP on the FCC Lattice**

While BPP has been solved for the three-dimensional cubic lattice, Istrail's biplane conjecture remains open on the three-dimensional triangular or the *face-centered cubic* (FCC) lattice.

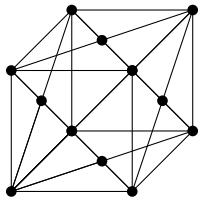

FIGURE 14. Unit cell of the FCC Lattice

The FCC provides the densest packing of spheres in three-dimensional space, famously proven by Thomas Hales using a computer-assisted proof [Hal+15], and therefore would approximate real-world protein folding much more closely than $\mathbb{Z}^3$. In [CI99], counting arguments for BPP on the FCC lattice were given, but the optimality of the bilayer remains an open problem.

We are uncertain whether a layer-by-layer decomposition argument as in the proof for Theorem 4.9 would work for the FCC, in particular because the FCC can be decomposed in two different ways: either as a stack of $\mathbb{Z}^2$ layers or as a stack of $\mathbb{T}^2$ layers both with some offset. We believe that the first step towards solving BPP on the FCC lattice would actually be to "fix" our proof for BPP on $\mathbb{Z}^3$, which only considers contacts within a layer and ignores contacts across layers.

5.2. **Variants of Bipole Packing**

While BPP packing and the side-chain model allow for a coarse approximation to branched, real-world proteins, considering "bipoles" of varying lengths would provide even better approximations to real-world folding, and is a natural generalization of BPP.

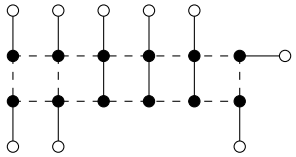

FIGURE 15. An instance of Generalized BPP

In other words, instead of just considering packings of dominos (i.e. bipoles), we can consider packings of polyominoes (i.e. "generalized" bipoles). Figure 5.2 illustrates a packing of a mixture of side chains of length 2 as well as length 3. A natural question to ask is

whether near-biplanar solutions (such as the packing in 5.2) are still optimal packings for such mixtures? What other optimal packings or arrangements exist?

As discussed in [IL13], BPP admits a formulation as a linear program. This LP formulation can easily be extended to consider bipoles of varying lengths. Simulations for certain mixtures of side-chains on graphs such as $\mathbb{Z}^2, \mathbb{Z}^3$ would be the natural next step. In addition to the aforementioned linear-programming algorithm, Monte Carlo methods could also be of use.

Two other variants of BPP that could be of interest are listed below:

- How does changing our potential function (i.e. considering P-P and/or H-P contacts in addition to just H-H contacts) affect BPP?
- On what graphs is the optimal packing far from a biplane, assuming one can be defined?

5.3. **Algorithms for Bipole Labelling**

Although Bipole-Labelling is **NP**-hard on arbitrary graphs, it may be tractable on certain "nice" graphs (lattices, for example). Moreover, formulating Bipole-Labelling as a quadratic program should result in an approximation algorithm for Bipole-Labelling on arbitrary graphs; similar algorithms exist for problems such as Max-2-SAT, Max-Cut, etc.

Acknowledgements

First and foremost, I would like to thank my advisor Sorin Istrail for all the advice, encouragement, and freedom that he provided over the past year. None of the work contained in this thesis would have been possible without his guidance and support. In addition, I am grateful to Richard Schwartz for agreeing to be my reader, and would like to thank the following people for brief conversations related to this thesis: Terrence George, Tianrun (Alice) Cheng, and Xufan Zhang.

More informally, I would like to thank my family, friends, and teachers for everything else.

References

[Aga+97] Richa Agarwala et al. "Local Rules for Protein Folding on a Triangular Lattice and Generalized Hydrophobicity in the HP Model". In: *Journal of Computational Biology* 4.3 (1997), pp. 275–296. DOI: 10.1089/cmb.1997.4.275.

[BL91] Béla Bollobás and Imre Leader. "Edge-isoperimetric inequalities in the grid". In: *Combinatorica* 11.4 (Dec. 1991), pp. 299–314. ISSN: 1439-6912. DOI: 10.1007/BF01275667.

[CI99] John H. Conway and Sorin Istrail. *Mathematics of Self-Assembly*. Tech. rep. 1999.

[Cre+98] Pierluigi Crescenzi et al. "On the Complexity of Protein Folding (Extended Abstract)". In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. STOC '98. Dallas, Texas, USA: ACM, 1998, pp. 597–603. ISBN: 0-89791-962-9. DOI: 10.1145/276698.276875. URL: http://doi.acm.org/10.1145/276698.276875.

[Dil85] Ken A. Dill. "Theory for the folding and stability of globular proteins". In: *Biochemistry* 24 (1985), pp. 1501–1509.

[GP13] Emanuele Giaquinta and Laura Pozzi. "An Effective Exact Algorithm and a New Upper Bound for the Number of Contacts in the Hydrophobic-Polar Two-Dimensional Lattice Model". In: *Journal of Computational Biology* (2013). DOI: 10.1089/cmb.2012.0266.

[Hal+15] Thomas C. Hales et al. "A formal proof of the Kepler conjecture". In: *CoRR* abs/1501.02155 (2015). arXiv: 1501.02155. URL: http://arxiv.org/abs/1501.02155.

[Har04] L. H. Harper. *Global Methods for Combinatorial Isoperimetric Problems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511616679.

[Har66] L.H. Harper. "Optimal numberings and isoperimetric problems on graphs". In: *Journal of Combinatorial Theory* 1.3 (1966), pp. 385–393. ISSN: 0021-9800. DOI: https://doi.org/10.1016/S0021-9800(66)80059-5. URL: http://www.sciencedirect.com/science/article/pii/S0021980066800595.

[HI97] William Hart and Sorin Istrail. "Robust Proofs of NP-Hardness for Protein Folding: General Lattices and Energy Potentials". In: *Journal of computational biology : a journal of computational molecular cell biology* 4 (Feb. 1997), pp. 1–22. DOI: 10.1089/cmb.1997.4.1.

[IL09] Sorin Istrail and Fumei Lam. "Combinatorial Algorithms for Protein Folding in Lattice Models: A Survey of Mathematical Results". In: *Communications in Information and Systems* 9.4 (2009), pp. 303–346.

[IL13] Sorin Istrail and Kshitij Lauria. "Combinatorial Algorithms For Bipole Self-Assembly On Lattices With Applications To The Lipid Bilayer". In: (2013). URL: cs.brown.edu/research/pubs/theses/ugrad/2013/lauria.pdf.

[Jer03] Mark Jerrum. *Counting, Sampling and Integrating: Algorithm and Complexity*. Jan. 2003. DOI: 10.1007/978-3-0348-8005-3.

[MS86] Vitali D Milman and Gideon Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Berlin, Heidelberg: Springer-Verlag, 1986. ISBN: 0-387-16769-2.

[Sch93] Rolf Schneider. *Convex Bodes: The Brunn-Minkowski Theory*. Cambridge University Press, 1993.

[SM99] Jean-Francois Sadoc and Rémy Mosseri. *Geometrical Frustration*. Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press, 1999. DOI: 10.1017/CBO9780511599934.

[Spi15] Daniel A. Spielman. *Spectral Graph Theory*. 2015.