Performance of the Uniform Closure Method for open knotting as a Bayes-type classifier

Emily Tibor$^1$, Elizabeth M. Annoni$^2$, Erin Brine-Doyle$^2$, Nicole Kumerow$^2$, Madeline Shogren$^2$, Jason Cantarella$^3$, Clayton Shonkwiler$^4$, and Eric J. Rawdon$^2$

$^1$Department of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA
$^2$University of St. Thomas, St. Paul, MN 55105, USA
$^3$Department of Mathematics, University of Georgia, Athens, GA 30602, USA
$^4$Department of Mathematics, Colorado State University, Fort Collins, CO 80523, USA

November 19, 2020

Abstract

The discovery of knotting in proteins and other macromolecular chains has motivated researchers to more carefully consider how to identify and classify knots in open arcs. Most definitions classify knotting in open arcs by constructing an ensemble of closures and measuring the probability of different knot types among these closures. In this paper, we think of assigning knot types to open curves as a classification problem and compare the performance of the Bayes MAP classifier to the standard Uniform Closure Method. Surprisingly, we find that both methods are essentially equivalent as classifiers, having comparable accuracy and positive predictive value across a wide range of input arc lengths and knot types.

Keywords: Open knot, polygonal knot, Uniform Closure Method

1 Introduction

Many physical systems contain long, open chain-like objects with two free ends, e.g. DNA, RNA, and other proteins. Anyone who has ever packed away a string of lights, a garden
hose, or headphone cables knows that these sorts of objects tend to be entangled. However, it is not clear how to measure this entanglement mathematically.

Several techniques have been proposed [5, 14]. In most (the exception being knotoids [40]), entanglement of open chains is measured by generating an ensemble of closures and measuring their knot types using traditional tools from mathematical knot theory.

There are a number of proposals for how to generate the closures (for a review, see [5, 14]) which vary in computation speed. In the simplest of cases (such as when the endpoints are close to each other or lie on the surface of the convex hull), all of the techniques generally agree. However, in more ambiguous situations, such as when the endpoints are well inside the convex hull, the different techniques can disagree.

These ambiguous situations cannot be neglected. In proteins, for example, it has been proposed that the borders of the “knotted cores” (i.e. the shortest subchains realizing a particular knot type) are rich with intra-chain contacts, provide structural stability, and may be correlated with their active sites [11]. The determination of which subchains determine the knotted cores requires that one computes the knotting in all subchains [1, 10, 24, 28, 29, 38, 42], which necessarily include the difficult-to-classify subchains. These papers show that the “core” knotting and linking has been preserved within classes of proteins which have the same function in different organisms. The proteins within these classes are separated by hundreds of millions of years of evolution, suggesting that their entanglement is critical to their function [12, 38].

While trying to compare different approaches to building ensembles of closures for open arcs, we were struck by the similarity to a problem in machine learning: design a classifier which predicts the knot type of a closed polygon $P$ given a subarc $A$ of that polygon (for some existing machine learning approaches to knot classification, see [20, 22, 25, 27, 41]). In this paper, we will focus on equilateral closed polygons chosen from the uniform probability measure [6, 8] on such polygons. For polygons $P$ of various knot types randomly sampled from this measure, we will compare the performance of the classical Bayes maximum a posteriori probability (MAP) classifier to the performance of the Uniform Closure Method, which is one of the most-used definitions for measuring entanglement in open chains [5, 10, 14, 19, 23, 35, 38, 38].

The Uniform Closure Method is defined as follows. For a given open chain and a given direction (seen as point on the unit sphere $S^2$), one can create a closed knot by extending rays in that direction from the two endpoints and closing the knot at infinity. Then the Uniform Closure Method “knot type” of the open chain is the probability distribution of closed knot types obtained over all points on $S^2$. A close relative of this method was initially introduced in [31], with the small difference that the authors closed the knot by extending line segments from the endpoints to random points on a large sphere encompassing the chain. Note that in practice, one needs only to have the rays extend beyond the convex hull of the chain, and then the configuration can be closed via a straight line segment, as seen on the

\footnote{This probability distribution is not meant to be a good approximation of the distribution of shapes of biological macromolecules (for instance, it ignores steric effects and bending stiffness), but it is a mathematically natural object of study.}
Figure 1: Examples of closures used by PU (left) and PR (right) associated with a subchain of the 98-edge KnotPlot +75 knot \cite{37} with 11 edges removed. Note that the configurations analyzed in this paper are random, and thus are not as elegant as the configuration shown here.

Our main experimental finding is that the Uniform Closure Method is almost exactly as good at classifying subarcs as the Bayes MAP classifier.

**Definition 1.1.** Given a $k$-edge arc $A$ and any $n > k$, the Bayes classifier with maximum a posteriori probability (MAP) decision rule for the knot type of $A$ is

$$
\text{BC}(A, n) := \text{the most common knot type among closed equilateral n-gons with subarc } A.
$$

If several knot types are equally common, the classifier selects among them at random. If there are no closed $n$-gons with subarc $A$ the classifier is undefined.

It is a standard theorem that the Bayes MAP classifier minimizes the probability of misclassification \cite{13}, so BC is the maximum accuracy classifier for the knot type of $P$. If we can sample from closed equilateral $n$-gons with subarc $A$ by choosing random $(n - k)$-edge arcs to close $A$ (see Figure 1B), we can approximate BC as follows:

**Definition 1.2.** Given a $k$-edge arc $A$ and any $n > k$, the random closure classifier of the knot type of $A$ is

$$
\text{PR}(A, n) := \text{the most common knot type among 100 random closed equilateral n-gons with subarc } A.
$$

As one would expect, we could build improved versions of the classifier PR by increasing the number of samples or (potentially) by choosing closures from a low discrepancy sequence in the space of possible closures rather than at random. However, in practice, computer time is limited and the metric structure of the space of closures is not well-understood. So
we investigated the performance of PR as a reasonable proxy for the performance of the maximum accuracy classifier BC.

The posterior distribution on knot types defined by the Uniform Closure Method has its own MAP classifier which is, like BC, impractical to compute. It is typically approximated as follows:

**Definition 1.3.** Given a \( k \)-edge arc \( A \), the *uniform closure classifier* of the knot type of \( A \) is

\[
PU(A) := \text{the most common knot type among 100 closures of } A \text{ by parallel rays from the endpoints of } A \text{ (see Figure 1)}.
\]

The directions of the rays are a fixed set chosen by the applet [1] to cover the sphere evenly. We weight the knot type associated with each direction by the relative size of the spherical Voronoi cell of the corresponding ray, and define the “most common” knot type to be the one with the largest total weight.

PU is clearly important for several reasons. The entanglement measured by PU has shown itself to be biologically relevant (again, see [5, 14, 23, 24, 35]). PU is considerably faster to compute than PR because it deals with \( k + 2 \) edge polygons instead of \( n \) edge polygons. And perhaps most usefully, given an open arc \( A \), PU does not require one to fix a length \( n - k \) for the closing arc.

The posterior distribution on knot types given by the Uniform Closure Method appears very different from the “true” posterior distribution of random arc closures. Because the Bayes MAP classifier has the maximum possible accuracy, PU cannot be more accurate than BC, and it is very unlikely to be more accurate than PR, which approximates BC. Indeed, we might expect PU to be much less accurate than PR because the posterior distributions are so different.

We now argue informally that, under plausible assumptions, the distribution of knot types in the posterior distributions used by PR and PU may not be so different. If one believes that knots in random polygons tend to be localized, then the \( (n - k) \)-edge section of a random polygon \( P \) joining the ends of a \( k \)-edge subarc \( A \) containing a local knot is likely to stay well away from the small knot in \( A \). In this case, if the \( (n - k) \)-edge arc is short enough that it does not introduce additional knotting, the knot type of \( P \) is likely to be the same as that of a knot formed by joining the ends of \( A \) “to infinity” with parallel rays. If this argument is right, at least for some \( n \), PU and PR may perform similarly.

We can go a little further. In probability problems with a large number of degrees of freedom, concentration of measure often leads the distribution of some relatively coarse labeling function on the space to be highly concentrated on its expected value or mode. Therefore, we might guess that almost all closures of \( A \) (whether by random arcs or rays) have the same knot type. If so, then the following classifiers might be almost as accurate as PR and PU at much lower computational cost.

\[\text{We do this because this matches the implementation of approximations to the Uniform Closure Method used elsewhere in the literature. Choosing evenly-spaced points provides a better approximation to the true posterior distribution than choosing random points.}\]

\[\text{Again, this is an intuitive argument and is not meant as a proof!}\]
Definition 1.4. Given a $k$-edge arc $A$ and any $n > k$, the single **random closure classifier** of the knot type of $A$ is
\[
\text{SR}(A, n) := \text{the knot type of one random closed equilateral } n\text{-gon with subarc } A.
\]

Definition 1.5. Given a $k$-edge arc $A$, the **single uniform closure classifier** of the knot type of $A$ is
\[
\text{SU}(A) := \text{the knot type of the closure of } A \text{ by parallel rays from the endpoints of } A \text{ whose direction is chosen randomly from the 100 fixed directions used in PU.}
\]

We now have two hypotheses about accuracy to test: when $n$ is not too much larger than $k$, PR and PU should be approximately equally accurate (and both close to the maximum possible accuracy), and for reasonably large $n$ and $k$, even SR and SU should be close to the maximum possible accuracy. We will see data below which supports both of these hypotheses. We will also see support for a much stronger result: the accuracies of all four classifiers are very close even when we restrict our attention to the tiny subspace of quite rare and complicated random knots.

However, when studying the performance of classifiers, accuracy is not the only important quantity. So we also compare the positive predictive value\(^4\) (PPV) of a variety of predictions made by each of our four classifiers from unknots to 9-crossing knots. Here we see that the (easier to compute) PU is just as good as PR, but that there are large differences between the PPV of these two and the PPV of SU and SR. Interestingly, though SR may seem “more natural” than SU\(^5\), the observed PPV for SU is substantially higher for every predicted knot type than the PPV of SR.

Our experiments are computationally expensive, so we could not choose a wide range of $n$ and $k$. Instead, we focused on random closed 100-edge equilateral polygons, analyzing the performance of our four classifiers at recovering the knot type of a particular polygon given the partial information in a subarc of length $k$. We computed accuracy and PPV for all $k$ from 1 to 100, finding as expected that all of our classifiers performed better when given more information.

## 2 Experimental procedure

### 2.1 Generation of random closures for PR and SR

In the random closure method, we must sample 100 random $n$-gons containing the initial $k$-edge arc $A$. It is a classical observation\(^6\) that the uniform probability distribution\(^6\) on equilateral $n$-edge polygons can be written as the probability distribution of $(n - k)$-edge

\(^4\)The chance that a prediction of knot type $K$ is accurate.

\(^5\)At least to some of the authors of this paper!

\(^6\)That is, the Hausdorff measure on the subspace of $(S^2)^n$ consisting of edge directions which sum to $\vec{0}$. 

5
open arcs and $k$-edge open arcs\footnote{The distribution of a $m$-edge arc is the uniform distribution of edge directions in $(S^2)^m$.} conditioned on the hypothesis that the end-to-end distances of the two arcs agree.

Since the end-to-end distance $\ell$ of $A$ is fixed, we can therefore construct a sample from the conditional distribution of closed equilateral $n$-gons containing $A$ by constructing a random $(n-k)$-edge arc $B$ with the end-to-end distance $\ell$ and joining its ends to $A$. Constructing such a $B$ is equivalent to sampling a random closed $(n-k+1)$-gon with $n-k$ edges of length 1 and one edge of length $\ell$. To do so, we used \texttt{plcurve} \cite{cantarella2007random}, which implements an algorithm of Cantarella and Shonkwiler \cite{cantarella2011random}. We note that $A$ and $B$ can be rotated independently around the vector joining their endpoints, and that we must choose the rotation angle uniformly to get the correct posterior distribution of $n$-gons. We see an example on the right of Figure 1.

2.2 Determination of knot types

Throughout this paper, the knot types of the polygons were determined using a combination of Ewing and Millett’s HOMFLYPT \cite{ewing2021knottype, millett2017knottype} polynomial software \cite{millett2016knottype} and Hoste and Thistlethwaite’s \texttt{knotfind} program (which is a part of the Knotscape software \cite{hoste2002knotscape}). Thistlethwaite’s \texttt{unraveller} program \cite{thistlethwaite2010unraveller} was used to simplify projections, easing the burden in computing the knot types. Using these programs, we were able to identify knot types of closed polygons as long as the knot type has a crossing number of 16 or less (implicitly assuming that the crossing number of a composite knot type is the sum of the crossing numbers of the factor knots, which is conjectured but has not been proven).

For all but a handful of knot types (those for which the HOMFLYPT polynomials match for chiral pairs), we can distinguish between the chiralities of the knot types as well. For chiral knot types, we designated one of the pair as the + version and one as the − version, based on the writhe of the standard diagram for the knot, or the spatial writhe of the ropelength-minimization \cite{cantarella2004random} if the writhe of the standard diagram for the knot is zero.

2.3 Sampling procedure for test data

We first generated 5,000,000 random 100-edge equilateral polygons using the \texttt{plCurve} library from Cantarella’s lab \cite{cantarella2007random}. We computed the knot type of each polygon as above, obtaining the distribution of knot types in Table 1.

Since the frequency of different knot types varies by several orders of magnitude, it is interesting to compare the performance of the classifiers both on the overall data set and on particular knot types (even very rare ones). The reason for this is that a classifier which was particularly good at recognizing unknots might have a very good overall accuracy score (because unknots are very common), but perform very poorly at recognizing the complicated knots that one is presumably most interested in.

So we took 100 samples from each of the knot types $0_1$, $+3_1$, $4_1$, $+5_1$, $+5_2$, $+6_1$, $+6_2$, $6_3$, $+3_1\# + 3_1$, $+3_1\# - 3_1$, $+7_1$, $+7_2$, $+8_2$, $+8_8$, $+8_{16}$, $+8_{19}$, $+8_{20}$, $+8_{21}$, $+9_2$, $+9_{15}$, $+9_{21}$, $+9_{36}$, $+9_{43}$, $+9_{44}$, and $+9_{46}$. These knot types were chosen to provide a variety of different
crossing numbers, and a combination of alternating and non-alternating, composite and prime, and chiral and amphichiral knot types. Note that we need to include chirality in the computations to avoid false-positives. For the chiral knot types, we always analyzed the configurations with the positive version of the knot type. The results for the other chirality were similar (up to computational error). The amphichiral prime knots analyzed are $0_1$, $4_1$, and $6_3$. The knots $+3_1 \# - 3_1$ and $+3_1 \# + 3_1$ are the composite granny and square knots, respectively. The knots $+8_{19}$, $+8_{20}$, $+8_{21}$, $+9_{43}$, $+9_{44}$, and $+9_{46}$ are non-alternating.

### 2.4 Experiments measuring accuracy of classifiers

For each 100-edge polygon, there were 9900 different open, connected subchains of various lengths. We ran all four classifiers on each subchain, recording their predictions as accurate when they matched the knot type of the overall closed polygon. We also computed the PPV. We recorded the accuracy for each classifier by subchain length and knot type. We recovered an overall accuracy score for all samples and all knots by reweighting each polygon by the probability of its knot type. Our overall accuracy scores are shown in Figure 2. Note that each of the data points is based on 1,000,000 closures (100 samples per knot type, 100 subchains of the given length per sample, 100 closures per subchain).

We can check that the data makes sense by considering the small $k$ limit. Every ray closure of a small arc is unknotted, so PU and SU always classify short arcs as unknots. Since about 70% of the random 100-gons are unknots, these classifiers are accurate about 70% of the time, which is the (best possible) Bayes accuracy. The PR classifier constructs 100 completely random equilateral 100-gons and returns the most common knot type observed. This most common knot type is $0_1$ virtually all of the time, so PR is also very close to Bayes accuracy for small $k$. On the other hand, SR is only accurate for small $k$ when two

---

Table 1: Percentage of the 5,000,000 randomly generated equilateral 100-gons which form the knot types in our sample.

| Knot     | %       | Knot     | %       | Knot     | %       |
|----------|---------|----------|---------|----------|---------|
| $0_1$    | 70.6613 | $+3_1 \# - 3_1$ | 0.7738 | $+9_2$   | 0.0024  |
| $+3_1$   | 8.4252  | $+7_1$   | 0.0234  | $+9_{15}$| 0.0035  |
| $4_1$    | 3.2827  | $+7_2$   | 0.0499  | $+9_{21}$| 0.0028  |
| $+5_1$   | 0.4994  | $+8_2$   | 0.0113  | $+9_{36}$| 0.0023  |
| $+5_2$   | 0.8546  | $+8_8$   | 0.0201  | $+9_{43}$| 0.0107  |
| $+6_1$   | 0.2066  | $+8_{16}$| 0.0039  | $+9_{44}$| 0.0162  |
| $+6_2$   | 0.2358  | $+8_{19}$| 0.0263  | $+9_{46}$| 0.0054  |
| $6_3$    | 0.2694  | $+8_{20}$| 0.0518  |          |         |
| $+3_1 \# + 3_1$ | 0.3865 | $+8_{21}$| 0.0293  |          |         |

---

8Because we analyze all subchains of a given polygon, there is some worry about self-correlations within our data set. We did an experiment where we analyzed only a single subchain per closed polygon and the results were similar to what we report below, albeit with more noise in the graphs.
Figure 2: The accuracy of each classifier is shown as a function of subchain length $k$ for a random 100-gon of any knot type including unknots (left plot), and a random 100-gon known to be nontrivially knotted (right plot).

independent random 100-gons happen to have the same knot type. The probability that both samples agree and are unknots is roughly $0.7^2 = 0.49$. Other possible coincidences of knot type are unlikely, but possible; the total probability of agreement is about 52%, consistent with the results we see for the red SR curve on the left hand side of the left plot in Figure 2.

A striking feature of Figure 2 is the essentially equal accuracy of PU and PR for all $k$, whether over the entire sample (left plot) or just the knotted curves (right plot). As $k$ approaches $n$, we see the reassuring result that all classifiers approach 100% accuracy; that is, they essentially always yield the correct knot type when given the entire knot as input!

It might be somewhat disconcerting that the red SR classifier outperforms our approximation PR of the Bayes MAP classifier for small $k$ in the right hand plot. Is it not the case that the MAP classifier always has the maximum possible accuracy? The solution to this apparent paradox is that the inputs on the right-hand plot are chosen only from knotted polygons, so PR is approximating the wrong Bayes MAP classifier. The correct Bayes MAP classifier would predict the most likely nontrivial knot ($\pm 3_1$) and be accurate $\sim 8.42\%$ of the time for small $k$; better than both SR and PR.

We found that the accuracy of all four classifiers remained comparable if we restricted our attention to curves of any single (nontrivial) knot type. Figure 3 shows some example accuracy plots for the $+3_1$, $4_1$ and $+9_{15}$ knots. The only significant difference in performance was observed for the SR classifier on unknots, which started at the expected $\sim 70\%$ accuracy and gradually improved to 100% as $k$ increased.

We can see from Figure 3 that our classifiers require a larger fraction of the curve to recognize knot type when the knot type of the host is more complicated. Table 2 shows the number of segments required to reach 50% accuracy for each of our classifiers for the non-trivial knot types studied here. We note that the PU and PR classifiers are consistently

\footnote{We note that when $k = n - 1$, SR and PR are always correct. In theory, SU or (less likely) PU could be wrong if the ray closure directions pass back through the body of the knot in just the wrong way. But this does not seem to occur often enough to be visible in our experimental data.}
Figure 3: With the exception of the unknot, our four classifiers have comparable accuracy across all \( k \) when the input set was restricted to curves of a given knot type. We observed this effect in all the knot types we tested.

very slightly faster at recognizing knots (though only by a few segments). One way to think about the number of segments required to recognize a knot is that it is a qualitative measure of something like an average size of the knot in a random configuration. This size generally increases with crossing number. However, the 8-crossing non-alternating knots (\(+8_{19}, +8_{20}, +8_{21}\)) are a few segments smaller than the 8-crossing alternating knots (\(+8_2, +8_8, +8_{16}\)), behaving more like 7-crossing alternating knots. The same is true for the 9-crossing non-alternating knots (\(+9_{43}, +9_{44}, +9_{46}\)) which compare to 8-crossing alternating knots. This phenomenon is shared with other “geometric” measures of knot complexity: generally non-alternating knot types have higher probabilities in random knot studies [15] and lower knot energy values [3, 26, 32] than alternating knot types with the same crossing number.

2.5 Experiments measuring positive predictive value

In the last section, we analyzed the accuracy of each of our classifiers and found them to be comparable in almost all cases. However, accuracy is not the only important measure of a classifier. In particular, one might be interested in the positive predictive value (PPV), which, as discussed in the Introduction, measures the probability that a knot type detected in a subchain actually matches the knot type of the host polygon. Like accuracy values, higher
Table 2: Minimum number of edges at which SU, PU, SR, and PR reach 50% classification accuracy.

| Knot  | SU | PU | SR | PR |
|-------|----|----|----|----|
| $+3_1$ | 75 | 73 | 79 | 73 |
| $4_1$ | 82 | 79 | 84 | 80 |
| $+5_1$ | 86 | 83 | 87 | 84 |
| $+5_2$ | 87 | 84 | 88 | 85 |
| $+6_1$ | 90 | 87 | 90 | 88 |
| $+6_2$ | 90 | 87 | 90 | 87 |
| $6_3$ | 90 | 87 | 90 | 88 |
| $+3_1\#+3_1$ | 90 | 88 | 90 | 89 |
| $+3_1\#-3_1$ | 90 | 88 | 90 | 88 |
| $+7_1$ | 92 | 89 | 91 | 90 |
| $+7_2$ | 92 | 89 | 92 | 89 |
| $+8_2$ | 94 | 92 | 93 | 92 |

PPV values represent better performance. Unlike the accuracy, which is always bounded by the Bayes accuracy, the PPV can be arbitrarily close to 1.

We start by considering the mean PPV for each method as a function of $k$. Unlike accuracy, this mean is not weighted by the frequency with which the classifier makes a given prediction: it is simply the total PPV of all outputs of the classifier, divided by the number of possible outputs. It is important to note that this number of possible outputs changes with $k$, at least for the SU and PU classifiers; for small numbers of segments, ray closure simply cannot generate any very complicated knots. SR, by contrast, can make any prediction, even for $k = 1$, as can PR.\(^{10}\)

Figure 4 shows the mean PPV of all predictions produced by each classifier for each value of $k$. This reveals two surprises. First, PU and PR are again essentially tied! Second, we see that SU is significantly better than SR for almost all $k$.

We see a clear trend in the data that predictions of more complicated knotting are more likely to be correct. This is especially true for the MAP classifiers PU and PR, which seem very unlikely to see a complicated knot a preponderance of the time unless the knot is really present in the host polygon.

In fact, the low PPV for predictions of trefoil knots tell us that many other host knots have arcs which are classified as trefoils, even by the high standards of the MAP classifiers PU and PR. Some of this effect is doubtless due to the presence of composite knot types in our data such as $+3_1\#+3_1$. However, we suspect that other knots in the same family, such as $+5_1$, also contain subarcs classified as the trefoil knot $[7, 9, 17, 23, 30, 33, 35]$.\(^{10}\)

\(^{10}\)Although it is overwhelmingly probable that PR will predict the unknot in this case, as we discussed above.
Figure 4: The mean PPV of the predictions produced by each classifier on our entire data set (100 knots of each type), for all predictions (left) and for predictions of nontrivial knots (right). Since the SU and PU classifiers cannot predict a nontrivial knot for very small $k$, we start the right hand plot at $k = 10$.

Figure 5: The PPV of the PU and PR classifiers was essentially the same, and increased with the complexity of the predicted knot type, eventually getting close to 1. The data is noisier for predictions of very complicated knots such as $+9_{15}$ because this prediction was made very rarely.
3 Discussion

In this paper, we found a surprising near-equivalence between the performance of the (approximate) Bayes MAP classifier PR and the uniform closure MAP classifier PU despite what seemed to be large differences in the posterior distributions sampled by each method. The methods were comparable both in terms of accuracy and PPV, so experimenters should feel confident using them interchangeably, at least for $n \approx 100$.

In the process, we noticed that the SU classifier (which is $100\times$ faster than the standard PU method) has comparable accuracy. Therefore, SU may be an acceptable substitute for PU or PR when speed is more important than PPV.

Our experiments leave open the question of whether these phenomena persist when $n$ is much larger. As discussed above, it seems very plausible that for very large $n$, PR would predict additional knotting that PU cannot. Whether this is desirable seems to be up to the experimenter.

It would be interesting to compare PU to a different measurement of “surplus knotting”. Suppose $\mathcal{P}(n, K)$ is the probability of finding knot type $K$ in the prime decomposition of the knot type of a random polygon of $n$ edges and $\mathcal{P}(n, K, A)$ was the corresponding probability for a random polygon of $n$ edges containing $A$ as a subarc. Intuitively, a measure of surplus knotting would compare $\mathcal{P}(n, K)$ to $\mathcal{P}(n, K, A)$ as $n \to \infty$. It seems reasonable to conjecture that this might be a better approximation of what PU is detecting.

4 Acknowledgments

Rawdon was supported by NSF DMS #1115722, #1418869, and #1720342. Annoni was supported by NSF DMS #1115722. Kumerow and Shogren were supported by NSF DMS #1115722 and #1418869. Brine-Doyle and Tibor were supported by NSF DMS #1418869. Cantarella and Shonkwiler were supported by the Simons Foundation (#524120 to JC, and #354225 and #709150 to CS). Tibor, Annoni, Brine-Doyle, Kumerow, and Shogren completed their work while students at the University of St. Thomas.

References

[1] Bob Allanson. Martin’s polyhedra. http://members.ozemail.com.au/~llan/mpol.html, 2013.

[2] Ted Ashton, Jason Cantarella, and Harrison Chapman. plcurve. http://www.jasoncantarella.com/wordpress/software/plcurve/, 2012.

[3] Ted Ashton, Jason Cantarella, Michael Piatek, and Eric J. Rawdon. Knot tightening by constrained gradient descent. Experiment. Math., 20(1):57–90, 2011.
[4] Daniel Böliger, Joanna I. Sulikowska, Hsiao-Ping Hsu, Leonid A. Mirny, Mehran Kar-\ndar, Jose N. Onuchic, and Peter Virnau. A stevedore’s protein knot. *PLoS Comput.\nBiol.*, 6(4):e1000731, 2010.

[5] Erin Brine-Doyle, Madeline Shogren, Emily Vecchia, and Eric J. Rawdon. Open knot-
ting. In Philipp Reiter, Simon Blatt, and Armin Schikorra, editors, *New Directions in Geometric and Applied Knot Theory*. De Gruyter, 2018.

[6] Jason Cantarella, Bertrand Duplantier, Clayton Shonkwiler, and Erica Uehara. A fast\ndirect sampling algorithm for equilateral closed polygons. *J. Phys. A*, 49(27):275202,\n2016.

[7] Jason Cantarella, Allison Henrich, Elsa Magness, Oliver O’Keefe, Kayla Perez, Eric J.\nRawdon, and Briana Zimmer. Knot fertility and lineage. *J. Knot Theory Ramifications*,\n26(13):1750093, 2017.

[8] Jason Cantarella and Clayton Shonkwiler. The symplectic geometry of closed equilateral\nrandom walks in 3-space. *Ann. Appl. Probab.*, 26(1):549–596, 2016.

[9] Harrison Chapman. On the structure and scarcity of alternating knots. Preprint,\n*arXiv:1804.09780 [math.GT]*, 2018.

[10] Pawel Dabrowski-Tumanski, Pawel Rubach, Dimos Goundaroulis, Julien Dorier, Pi-
otr Sulkowski, Kenneth C. Millett, Eric J. Rawdon, Andrzej Stasiak, and Joanna I.\nSulkowska. KnotProt 2.0: a database of proteins with knots and other entangled struc-
tures. *Nucleic Acids Research*, 47(D1):D367–D375, 12 2018.

[11] Pawel Dabrowski-Tumanski, Andrzej Stasiak, and Joanna I. Sulikowska. In search of\nfunctional advantages of knots in proteins. *PLOS ONE*, 11(11):1–14, 11 2016.

[12] Pawel Dabrowski-Tumanski and Joanna I. Sulikowska. Topological knots and links in\nproteins. *Proc. Natl. Acad. Sci. USA*, 114(13):3415–3420, 2017.

[13] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern\nRecognition*, volume 31 of *Applications of Mathematics*. Springer, New York, 1996.

[14] Julien Dorier, Dimos Goundaroulis, Eric J. Rawdon, and Andrzej Stasiak. Open knots.\nIn Colin Adams, Erica Flapan, Allison Henrich, Louis H. Kauffman, Lewis D. Ludwig,\nand Sam Nelson, editors, *Encyclopedia of Knot Theory*. Chapman and Hall/CRC Press,\n2020. To appear.

[15] Chaim Even-Zohar. Models of random knots. *J. Appl. Comput. Topol.*, 1(2):263–296,\n2017.

[16] Bruce Ewing and Kenneth C. Millett. Computational algorithms and the complexity of\nlink polynomials. In Michel Boileau, Michel Domergue, Yves Mathieu, and Kenneth C.\nMillett, editors, *Progress in Knot Theory and Related Topics*, volume 56 of *Travaux en\nCours*, pages 51–68. Hermann, Paris, 1997.
[17] Alessandro Flammini and Andrzej Stasiak. Natural classification of knots. *Proc. R. Soc. A*, 463(2078):569–582, 2007.

[18] Peter J. Freyd, David N. Yetter, Jim Hoste, William B. R. Lickorish, Kenneth C. Millett, and Adrian Ocneanu. A new polynomial invariant of knots and links. *Bull. Amer. Math. Soc. (N.S.)*, 12(2):239–246, 1985.

[19] Dimos Goundaroulis, Eriz Lieberman Aiden, and Andrzej Stasiak. Chromatin is frequently unknotted at the megabase scale. *Biophys. J.*, 118(9):2268–2279, 2020.

[20] Sergei Gukov, James Halverson, and Fabian Ruehle. Learning to unknot. Preprint, arXiv:2010.16263 [math.GT], 2020.

[21] Jim Hoste and Morwen Thistlethwaite. Knotscape. http://www.math.utk.edu/~morwen/knotscape.html, 2009. Program for computing topological information about knots.

[22] Mark C. Hughes. A neural network approach to predicting and computing knot invariants. *J. Knot Theory Ramifications*, 29(3):2050005, 2020.

[23] David A. B. Hyde, Joshua Henrich, Eric J. Rawdon, and Kenneth C. Millett. Knotting fingerprints resolve knot complexity and knotting pathways in ideal knots. *J. Phys.: Condens. Matter*, 27:354112, 2015.

[24] Michal Jamroz, Wanda Niemyska, Eric J. Rawdon, Andrzej Stasiak, Kenneth C. Millett, Piotr Sułkowski, and Joanna I. Sułkowska. Knotprot: a database of proteins with knots and slipknots. *Nucleic Acids Research*, 43(D1):D306–D314, 2015.

[25] Vishnu Jejjala, Arjun Kar, and Onkar Parrikar. Deep learning the hyperbolic volume of a knot. *Phys. Lett. B*, 799:135033, 2019.

[26] Denise Kim and Rob Kusner. Torus knots extremizing the Möbius energy. *Experiment. Math.*, 2(1):1–9, 1993.

[27] Jesse S. F. Levitt, Mustafa Hajij, and Radmila Sazdanovic. Big data approaches to knot theory: Understanding the structure of the Jones polynomial. Preprint, arXiv:1912.10086 [math.GT], 2019.

[28] Marc L. Mansfield. Are there knots in proteins? *Nat. Struct. Biol.*, 1:213–214, 1994.

[29] Marc L. Mansfield. Fit to be tied. *Nat. Struct. Biol.*, 4:166–167, 1997.

[30] Kenneth C. Millett. Knots in knots: A study of classical knot diagrams. *J. Knot Theory Ramifications*, 25(9):1641013, 2016.

[31] Kenneth C. Millett, Akos Dobay, and Andrzej Stasiak. Linear random knots and their scaling behavior. *Macromolecules*, 38(2):601–606, 2005.
[32] Kenneth C. Millett and Eric J. Rawdon. Energy, ropelength, and other physical aspects of equilateral knots. *J. Comput. Phys.*, 186(2):426–456, 2003.

[33] Kenneth C. Millett and Alex Rich. More knots in knots: A study of classical knot diagrams. *J. Knot Theory Ramifications*, 26(8):1750046, 2017.

[34] Józef H. Przytycki and Paweł Traczyk. Invariants of links of Conway type. *Kobe J. Math.*, 4(2):115–139, 1987.

[35] Eric J. Rawdon, Kenneth C. Millett, and Andrzej Stasiak. Subknots in ideal knots, random knots, and knotted proteins. *Sci. Rep.*, 5:8928, 2015.

[36] Lord Rayleigh. On the problem of random vibrations, and of random flights in one, two, or three dimensions. *Philosophical Magazine*, 37(220):321–347, 1919.

[37] Robert G. Scharein. KnotPlot. http://www.knotplot.com, 1998. Program for drawing, visualizing, manipulating, and energy minimizing knots.

[38] Joanna I. Sulkowska, Eric J. Rawdon, Kenneth C. Millett, Jose N. Onuchic, and Andrzej Stasiak. Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. USA*, 109(26):E1715–E1723, 2012.

[39] Morwen Thistlethwaite. unraveller, 2009. Software to simplify knot diagrams.

[40] Vladimir Turaev. Knotoids. *Osaka J. Math.*, 49(1):195–223, 2012.

[41] Olafs Vandans, Kaiyuan Yang, Zhongtao Wu, and Liang Dai. Identifying knot types of polymer conformations by machine learning. *Phys. Rev. E*, 101:022502, 2020.

[42] Peter Virnau. Protein knot server. http://knots.mit.edu/. Server that allows users to check for knots in PDB entries or other uploaded structures.
A Appendix

The accuracies of the PR, PU, SR, and SU classifiers for each of the knot types in our data set, shown as a function of subchain length.

Subchains of 0 knots

Subchains of +3 knots

Subchains of 4 knots

Subchains of +5 knots

Subchains of +6 knots
The PPV of the PR, PU, SR, and SU classifiers for each of the knot types in our data set, shown as a function of subchain length.
Prediction is $6_3$

Prediction is $+3_1$ $#$ $+3_1$

Prediction is $+3_1$ $#$ $-3_1$

Prediction is $+7_1$

Prediction is $+7_2$

Prediction is $+8_2$

Prediction is $+8_8$

Prediction is $+8_{16}$

Prediction is $+8_{19}$

Prediction is $+8_{20}$
Prediction is $+8_{21}$

Prediction is $+9_2$

Prediction is $+9_{15}$

Prediction is $+9_{21}$

Prediction is $+9_{36}$

Prediction is $+9_{43}$

Prediction is $+9_{44}$

Prediction is $+9_{46}$