A Geometric Measure Theory Approach to Identify Complex Structural Features on Soft Matter Surfaces

Enrique Alvarado,[†,§] Zhu Liu,[‡,§] Michael J. Servis,[‡] Bala Krishnamoorthy,[*,¶] and Aurora E. Clark[*,‡]

†*Department of Mathematics and Statistics, Washington State University, Pullman, Washington 99164, United States*

‡*Department of Chemistry, Washington State University, Pullman, Washington 99164, United States*

¶*Department of Mathematics and Statistics, Washington State University, Vancouver, Washington 98686, United States*

§*These authors contributed equally to this work.*

E-mail: kbala@wsu.edu; auclark@wsu.edu

Abstract

The structural features that protrude above or below a soft matter interface are well-known to be related to interfacially mediated chemical reactivity and transport processes. It is a challenge to develop a robust algorithm for identifying these organized surface structures, as the morphology can be highly varied and they may exist on top of an interface containing significant interfacial roughness. A new algorithm that employs concepts from geometric measure theory, algebraic topology, and optimization, is developed to identify candidate structures at a soft matter surface, and then using a probabilistic approach, to rank their likelihood of being a complex structural feature. The algorithm is tested for a surfactant laden water/oil interface, where it is robust to identifying protrusions responsible for water transport against a set identified by visual inspection. To our knowledge, this is the first example of applying geometric measure theory to analyze the properties of a chemical/materials science system.

Introduction

Surfaces of soft matter phases are host to a variety of scientifically important phenomena that include chemical reactivity and transport. Often these processes are sensitive to surface deformation, including surface buckling, capillary roughness, and the formation of locally organized regions that facilitate the process of interest. A quintessential example lies within the realm of transport across a soft matter interface, for example surfactant laden water/oil surfaces that are employed within solvent extraction based separations. Depending upon solution conditions, deformations as large as micelles or as small as a handful of water molecules have been implicated within solute transport mechanisms. Micelles are proposed to form via a vesicle budding process, whereby a region of high curvature forms at a planar interface, peeling off the surface as it forms a spherical micelle.[1–5] At much smaller length scales, water protrusions have been identified as a means of transferring water and other solutes between phases, where surface active amphiphiles create extremes in surface roughness.[6–8]

These extremities, once disengaged from the interface, release molecular components from the surface of one phase into the bulk of the other. Lastly, structures referred to as water fingers have been identified to accompany bare ion transport between high and low dielectric phases. The water finger is composed of H_2O molecules that trail the remaining solvation shell of the ion, leading back to the aqueous interface, before the ion is fully transferred to the low dielectric phase.[9–16]

The importance of surface structures to soft matter interfacial chemistry has its roots within the atomistic simulation literature, where molecular-level detail augments knowledge obtained from experimental observation. This is particularly relevant given the challenges of experimental characterization of liquid interfaces. Spectroscopic measurements are generally limited to spatial and temporal averages. This complicates precise structural interrogation of features across the heterogeneous interface. Even for simulation studies that have access to instantaneous configurations, systematic classification of interfacial features is not simple. Importantly, most reported simulation studies of hierarchical interfacial structure, e.g., protrusions, water fingers or budding micelles, analyze their data using visual inspection or by a measure of surface curvature that is dependent on surface feature homogeneity and size relative to their molecular constituents.[1,2,5–11,17] Not only does this limit quantitative and statistical studies of the behavior of surface structures but it also can cause interpretations that fall prey to lack of objectivity or are biased by individual experiences or expectations of those doing the analysis. A means of automatically identifying surface structures from simulation trajectories is lacking, and indeed there are significant technical challenges to differentiate background interfacial surface fluctuations from collectively organized surface structures.

Toward this end we present a new method that employs concepts from geometric measure theory (GMT), algebraic topology, and optimization. GMT is an area of pure mathematics that studies questions of area-minimizing surfaces, soap bubble conjectures, regularity of solutions to energy minimization problems, and so on.[18] In this context, the *flat norm*

(FN) has been used to define a distance between generalized hypersurfaces called currents lying inside a Euclidean space. In our application to Chemistry, which is the first such instance as far as we are aware, we consider the minimization of a generalized area of a complicated surface (in x, y, z) representing a soft matter interface, as we "flatten" the surface toward the horizontal plane with the same x, y dimensions. Using this framework we identify relevant volumes lying between the complicated surface and flat surface at various scales by minimization of the flat norm function. Furthermore, we automatically identify the volumes that represent structural features of the surface that are potentially relevant to the interfacial chemistry. The algorithm that employs the flat norm, labeled `GMTChem.FN`, is scale-independent, and its utility is demonstrated for a test system that has previously been shown to have hierarchically organized surface protrusions—the tri-n-butyl phosphate (TBP)-laden water/hexane interface. The framework is applicable to many different chemical systems and represents an important new tool to study interfacial structure, dynamics, and reactivity.

Theoretical Background and Algorithm Workflow

Background

The `GMTChem.FN` framework utilizes a method called the *multiscale simplicial flat norm* (`MSFN`) that produces smoother versions of a triangular surface lying inside a tetrahedral volume mesh. This method has its roots in the application of the flat norm to image denoising.[19,20] More recently, the flat norm has been used to define an average of shapes in a fairly general setting.[21] The `MSFN` algorithm takes in three inputs: a triangular surface mesh T, a tetrahedral volume mesh K which contains T as a subcomplex, and a scale parameter $\lambda \geq 0$. The algorithm solves a minimization problem (Equation 1) to identify a tetrahedral volume S inside K and returns a "flatter", i.e., smoother, version of T, given by $T - \partial S^{\lambda}$. The smaller the scale λ, the smoother $T - \partial S^{\lambda}$ will be. In this paper, we denote by T the input to the

flat norm algorithm that we are trying to smooth, which could be a surface, a triangular mesh, or a curve.

We first introduce the continuous analogue of `MSFN` called the *multiscale flat norm* (`MFN`). The definition applies to generalized n-dimensional hypersurfaces, but our focus will be on the case of T being a surface in $\mathbb{R}^3$. Let T be an oriented surface in $\mathbb{R}^3$. The `MFN` of T with scale $\lambda \geq 0$ is computed as

$$\mathbf{F}^{\lambda}(T) = \min_{S}\{\text{Area}(T - \partial S) + \lambda \cdot \text{Vol}(S)\}, \tag{1}$$

where we minimize over all oriented 3-dimensional volumes S. A 3-dimensional oriented volume S^* that attains the minimum in the above computation is a *flat norm minimizer* for T *with scale* λ. We refer to the flat norm minimizers for T with scale λ simply as *minimizers* and denote them as S^{λ}. The smoothed version of T for a minimizer S^{λ} is $T - \partial S^{\lambda}$, and is called the *flat surface* of T *with scale* λ, or simply the *flat surface*. In practice we usually consider a discrete set of values for λ within a range $(0, L)$, denoted $\{\lambda_i\}_{i=1}^{N}$ for some positive integer N. Hence the compact notation of S^i and $T - \partial S^i$ is utilized for the minimizers and the corresponding flat surfaces.

We use the case of T being a curve in $\mathbb{R}^2$ to further illustrate the concepts and definitions (Figure 1). Corresponding to the computation in Equation (1) for surfaces, here we compute $\mathbf{F}^{\lambda}(T) = \min_S\{\text{Length}(T - \partial S) + \lambda \cdot \text{Area}(S)\}$, minimizing over all 2-dimensional oriented domains S of $\mathbb{R}^2$. One could intuitively think of the flat norm computation as minimizing the total cost of erasing the curve T completely in two steps, with each step incurring its own respective cost. The first step is to erase a portion of T with the boundary ∂S of a 2-dimensional region S, the cost of which is the area of S. The second step is to erase the curve that is left after the first step, i.e., $T - \partial S$, having a cost of the length of the curve erased in this step.

With the scale parameter $\lambda = \lambda_1$ (e.g., $\lambda_1 = 1$) the minimizer S^1 is the collection of two brown areas shown in Figure 1. The portion of T that ∂S^1 erases in the first step consists of two parts of the input curve that coincide with (parts of) ∂S^1. This step contributes a cost of $\lambda \text{Area}(S^1)$. The resulting flat curve (an analogue of the flat surface) is shown in green, and represents the smoothed version $T - \partial S^1$ of T in Figure 1. Note that while this step removed the sharper "bumps" in T, it also added the bottom portions of ∂S^1 not coinciding with T. The second step is to erase the entire green curve which incurs as cost the length of $T - \partial S^1$. Hence the total cost of erasing the input curve T is $\text{Length}(T - \partial S^1) + \lambda \text{Area}(S^1)$.

We now consider the cost of erasing T when the scale parameter $\lambda = \lambda_2 \ll \lambda_1$ (e.g.,$\lambda = \lambda_1$ and $\lambda_2 = 0.001$). Note that the first step becomes much cheaper, and hence the minimizer S^2 consists of the two large areas made of the *union* of the pair of brown and pink areas in Figure 1. We can afford the much larger cost $\text{Area}(S^2)$ compared to $\text{Area}(S^1)$ in the first step. The resulting flat curve $T - \partial S^2$ is shown in red, and is much "flatter" that the intermediate green curve $T - \partial S^1$. The second step incurs the extra cost of $\text{Length}(T - \partial S^2)$ for erasing the red curve. As illustrated by this example, we capture "features" of T at all scales by constructing the flat curves at multiple scale values $\lambda_i \in (0, L]$, $i = 1, \ldots, N$ ($L > 1$, typically).

While the definition of `MFN` is given in the continuous setting, we usually employ a discretized version of the input and ambient space, i.e., of T as well as choices of S, to perform the computations in practice. A natural discretization considers T and choices of S as *simplicial complexes*, i.e., meshes. Ibrahim et al.[22] proposed a simplicial version of `MFN` called the `MSFN`. For the setting of our interest where T is a finite surface in $\mathbb{R}^3$, or when T is a finite curve in $\mathbb{R}^2$ (more generally, in *codimension* 1), the `MSFN` can be computed efficiently by solving a linear program.

In this case, the first step is to create the surface of interest T in $\mathbb{R}^3$, e.g., a soft matter interface. For a liquid/liquid interface, the instantaneous surface (the layer of molecules in direct contact with the opposing immiscible phase) can be represented by the Willard-

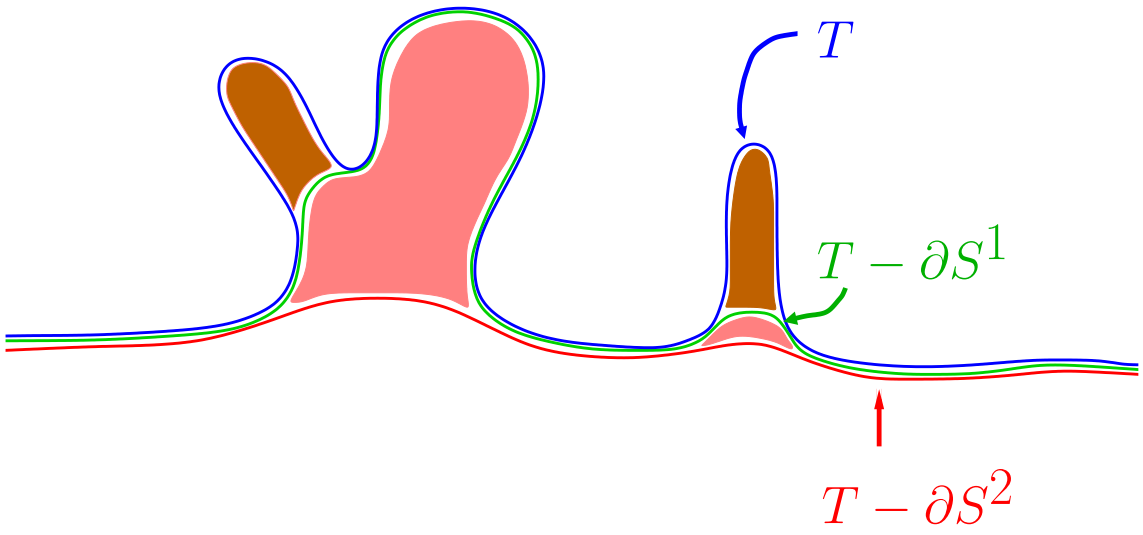

Figure 1: Curve T in blue, representing a 1-dimensional analog of a soft-matter interface, together with minimizers of the flat norm for two values of $\lambda = \lambda_1$ and $\lambda = \lambda_2 \ll \lambda_1$. The flat curve $T - \partial S^1$ (for $\lambda = \lambda_1$) is shown in green, and the flat curve $T - \partial S^2$ (for $\lambda = \lambda_2$) is shown in red. Note that the resulting curve get flatter as λ decreases.

Chandler interface,[23,24] as is done in this work (see Section). We discretize this interface into a triangular surface mesh T, which is a 2-dimensional simplicial complex. Because the flat norm computation makes use of volumetric data, the ambient space around T is discretized into a tetrahedral volume mesh K, which is a 3-dimensional simplicial complex. When generating K, the discretization is performed so as to preserve the triangles from T—by ensuring the triangles in T are faces of tetrahedra included in K. More formally, we ensure T is a *subcomplex* of K, and do so using the method of constrained Delaunay tetrahedralization[25] implemented in the open source software TetGen[26] (see Supplementary Information for details).

While the minimizer S^λ constructed by the `MSFN` algorithm is expected to identify meaningful structural features on the complicated surface T, the results can be quite sensitive to the choice of the scale parameter λ. As λ decreases, the volume of S^λ increases until eventually, $S^\lambda = K$. Initially it may seem possible to have a unique value for λ at which S^λ contains all of the relevant features of the interface, and nothing more. However, we could

not identify such values of λ for any of the interfaces studied here. One would expect that when λ is small enough, S^λ would contain essentially all relevant features. Yet we observed that many of the features in S^λ were either too small or had broad or ill-defined geometries. In part, this is due to the large range of morphologies of the structural features at the liquid/liquid interface used to develop the `GMTChem.FN` algorithm. Hence we consider a broad range of values for λ, and further require that a candidate feature has large enough volume across a large number of the λ values considered (vide infra).

The scale parameter λ captures curvature of the surface T in the following intuitive sense. Imagine moving a ball of radius $1/\lambda$ with its center lying on T (rather than the entire ball "rolling on" the surface). Portions of T and ambient space that lie inside the union of balls, i.e., copies of the ball as its center moves across T, are smoothed out by the flat norm algorithm. Going back to Figure 1, consider the space covered by a 2-dimensional disc of unit radius ($\lambda_1 = 1$) as its center is moved along T, the blue curve. The two areas shown in brown, which form the minimizer S^1, are completely covered by the union of balls and hence are erased. When λ is much smaller (e.g., $\lambda_2 \ll 1$), the ball has a much larger size and hence smooths out most of the curve T along with the shaded areas in brown and pink.

Motivated by this analogy, we compare volumes of the connected components of minimizers S^λ to volume of the ball of radius $1/\lambda$. Rather than directly label each component as a feature or not, we adopt a probabilistic approach. The components are studied over a range of λ values as well as a range of values for the ratio of volume of component to volume of the $1/\lambda$-ball. We then classify at which of these values a particular component is "alive" as a feature, and compute the fraction of pairs of values as the probability of the component being a bona fide surface structural feature.

`GMTChem.FN` Algorithm Overview

The `GMTChem.FN` algorithm workflow consists of five steps (Figure 2). For a frame of simulation data indexed by t (obtained from molecular dynamics, Monte Carlo, etc.) the soft-

matter surface is first modeled as a triangular mesh. We use the Willard-Chandler interface $T(t)$ in our studies. For each t, we *embed* $T(t)$ into a 3-dimensional complex $K(t)$ such that $T(t)$ is a portion of the boundary of $K(t)$ and the triangles in $T(t)$ are preserved. We then run the `MSFN` algorithm N times on each pair $(T(t), K(t))$ with the scale parameter λ taking a range of decreasing values $\{\lambda_i\}_{i=1}^{N}$. The range of λ values and N will depend on the application at hand.

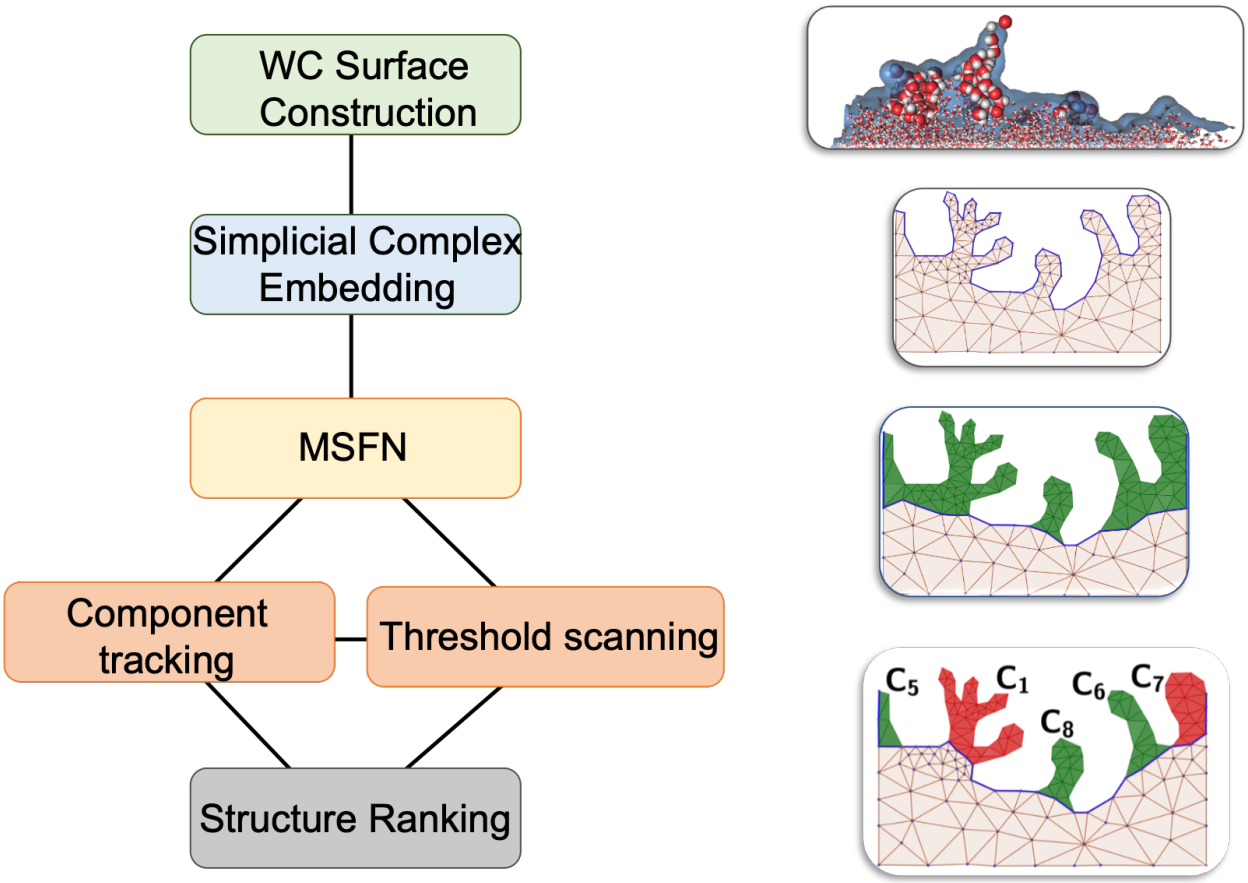

Figure 2: Workflow for structure identification by the `GMTChem.FN` algorithm.

For each λ_i the `MSFN` algorithm outputs the minimizer $S^i(t)$ which is then partitioned into its connected components $C^i(t) = \{C_k^i\}_{k=1}^{n_i}$, where n_i is the number of connected components of $S^i(t)$. We compute two quantities for each connected component C_k^i: its volume $\mathrm{Vol}(C_k^i)$ and its ratio $R(C_k^i)$ defined as the ratio of $\mathrm{Vol}(C_k^i)$ and volume of a ball of radius $1/\lambda_i$. If $R(C_k^i)$ is above a particular threshold r, we say that component C_k^i is *alive at scale* λ_i

and threshold r. Although not required, it is insightful to have a reference chemical system that has an interface without any relevant features that provides a basis for selecting the range of thresholds $\{r_j\}_{j=1}^{M}$ to be considered. For example, in the test system that contains water/hexane/TBP, one may first employ the algorithm on water/hexane to identify the range of thresholds $\{r_j\}_{j=1}^{M}$.

A component that is alive at a specific λ_i and threshold r_j is considered a candidate structural feature, and is given a *component label* that keeps track of the merging and growing behavior of the component as lambda decreases. Note that smaller components merge into larger ones as λ becomes smaller. Each frame indexed by t is examined over all λ_i to obtain the complete list of connected components $C(t) = \bigcup_i C^i(t)$. All the components in $C(t)$ are then labeled so as to track their growing and merging behavior as λ decreases. The labeling procedure is illustrated in Figure 3. Each connected component is assigned at most one component label per lambda value that enables tracking the evolution its volume and ratio as a function of λ. We denote the set of all component labels at a frame of data t as $\mathcal{C}(t)$, and the set of all component labels ranging over all frames of data as $\mathcal{C}$.

Given the range of lambda values $\{\lambda_i\}_{i=1}^{N}$ and ratio thresholds $\{r_j\}_{j=1}^{M}$, for each component label C in $\mathcal{C}$ we construct an $N \times M$ matrix $A_C = [a_{ij}]$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. We set $a_{ij} = 1$ if component C is alive at λ_i and ratio threshold r_j, and set $a_{ij} = 0$ otherwise. Assuming lambda and ratio cutoff values are chosen uniformly at random, the overall probability that component label C is alive is given by

$$P(C \text{ is alive}) = \frac{\sum a_{ij}}{NM}\,. \tag{2}$$

Subsequently, we rank all component labels in $\mathcal{C}$ by sorting the corresponding list of probabilities from largest to smallest values.

The following sections describe relevant portions of the code in detail. The technical aspects of embedding the Willard-Chandler Surface into a 3-dimensional simplicial complex

are described in Supporting Information (Section S1) for clarity.

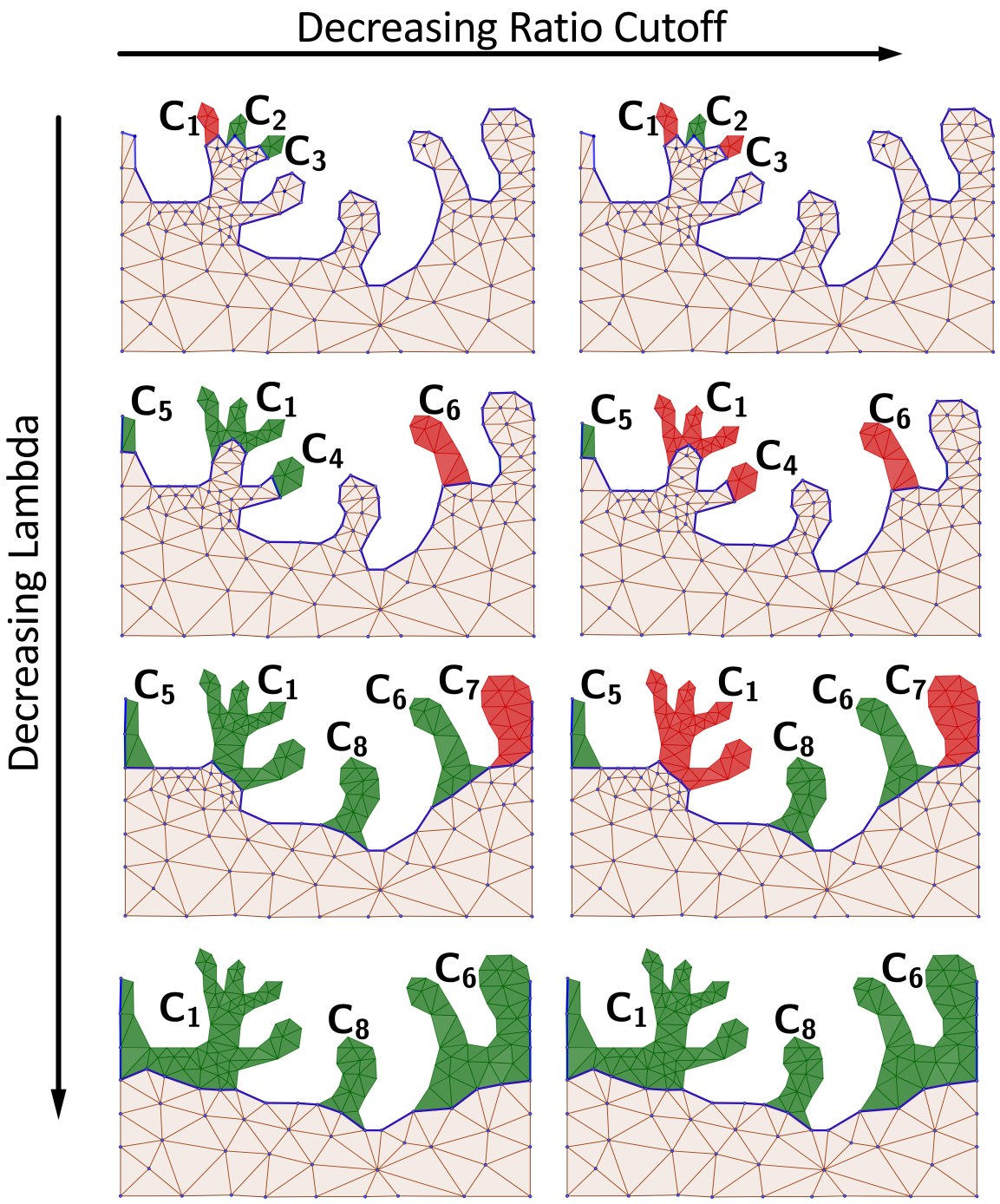

Figure 3: Connected components and component labels illustrated in 2D for four λ values and two ratio cutoffs. Connected components are colored green or red, with ones colored red being alive.

Multiscale Simplicial Flat Norm Algorithm (MSFN)

In the next step of the GMTChem.FN software, the input is the Willard-Chandler 2-dimensional simplicial complex, $T(t)$ that is embedded in its respective $K(t)$, along with a decreasing sequence of positive values, $\{\lambda_i\}_{i=1}^N$ for some particular positive integer N. This portion of the algorithm is applied to each $T(t)$ and $K(t)$ independently, to simplify notation, we denote $T(t)$ and $K(t)$ as T and K. In Figure 2, the picture corresponding to the *Simplicial Complex Embedding* portion of the workflow shows T in blue, embedded in light brown colored region K. The MSFN algorithm is computed on T with each scale λ_i and outputs a sequence of minimizers $\{S^i\}_{i=1}^N$ together with the corresponding flat surfaces $\{T - \partial S^i\}_{i=1}^N$. Note that each S^i consists of an ordered set of tetrahedra of K, and each $T - \partial S^i$ consists of an ordered set of triangles of K.

As an example, for Figure 3 the MSFN algorithm is run on T for the following five decreasing positive values of lambda $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$. For now, focusing on the left-most column in Figure 3, going from the top frame to the bottom fourth frame, the regions colored in green/red are the four minimizers $S^1, ..., S^4$ corresponding to each $\lambda_1, ..., \lambda_4$, respectively. In particular for λ_1, the minimizer S^1 is equal to the union of the 3 disconnected components. For λ_2, the minimizer S^2 is equal to the union of the 4 disconnected components. For λ_3, the minimizer S^3 is equal to the union of the 5 disconnected components, and for λ_4 the minimizer S^4 is equal to the 3 disconnected components. The frame that corresponds to λ_5 is not shown because $S^5 = K$, which would have made all of K in that fifth frame, green. As we go further along the pipeline, the notation in Figure 3, and the distinctions of the green and red colors will be made apparent.

Component Tracking

This step of the GMTChem.FN algorithm is carried out independently for each frame t of simulation data. For simplicity the parameter t is dropped within the notation. The input to this step is the ordered set of minimizers $\{S^i\}_{i=1}^N$ generated by N applications of the

`MSFN` algorithm, as described in Section . The output of this step is an ordered set of *component labels*, $\{C_k\}$. Each component label $C_k = (C^1_{k(1)}, \ldots, C^N_{k(N)})$ is an ordered set of *connected components*, where each $C^i_{k(i)}$ is a connected component of the minimizer S^i, for $i = 1, \ldots, N$. What follows is a description of how a component label describes the *merging* and *growing* behavior of its first nonempty connected component as a function of the decreasing sequence of lambda values $\{\lambda_i\}_{i=1}^N$. A more detailed description is presented in the Supporting Information (Section S2).

Connected Components

Running the `MSFN` algorithm on T for scale λ_i yields the minimizer S^i, which consists of a subset of tetrahedra from K. The *dual graph* G of S^i is constructed, where vertices of G correspond to tetrahedra in S^i, and an edge is present between vertices in G if the corresponding tetrahedra share a triangle. By this mapping of nodes with tetrahedra, the collection of connected components of G corresponds exactly to a collection $\{C^i_k\}_{k=1}^{n_i}$ of the *connected components* of S^i. Each C^i_k is a ordered set of oriented tetrahedra, and n_i denotes the number of connected components of S^i.

Returning to the example in Figure 3, the three connected components of the minimizer S^1 corresponding to $\lambda = \lambda_1$ are labeled $\{C^1_1, C^1_2, C^1_3\}$ (from left to right). Similarly, the four connected components of S^2 are $\{C^2_1, C^2_2, C^2_3, C^2_4\}$. Note that labels listed in Figure 3 are in fact modified versions of these labels that capture the growing and merging behavior of the components (see below for details). For S^3, the minimizer consists of 5 connected components $\{C^3_1, \ldots, C^3_5\}$. Note that there are 5 connected components here rather than 4, even though the two rightmost features share a vertex. Recall we defined connected components of S^λ by identifying the corresponding connected components of its dual graph. Hence two tetrahedra are connected if and only if they share a triangle. If they share only an edge or a vertex, they are considered to be not connected. Similarly, for the 2-dimensional simplicial complex K in our example, two triangles are connected if and only if they share an edge. In the case

of S^3, the two components share only a vertex, and hence are considered not connected.

Growing and Merging Behavior

A connected component of S^{i-1} either grows or merges into a connected component of S^i. A connected component C_k^{i-1} of S^{i-1} *grows into* component C_j^i of S^i if $C_k^{i-1} \subseteq C_j^i$ and any additional tetrahedra in C_j^i not in C_k^{i-1} are also not in S^{i-1}. In contrast, connected components $C_{k_1}^{i-1}, \ldots, C_{k_n}^{i-1}$ of S^{i-1} *merge* into component C_j^i of S^i if $\bigcup_{l=1}^{n} C_{k_l}^{i-1} \subseteq C_j^i$. When one goes from from S^1 to S^2 in the left column of Figure 3, one can see that components $\{C_1^1, C_2^1, C_3^1\}$ merge into C_2^2 (labeled C_1 in the Figure). Going from S^2 to S^3, C_1^2 and C_4^2 (labeled C_5 and C_6) grow into C_1^3 and C_4^3 (labeled C_5 and C_6), respectively. Also notice that C_2^2 and C_3^2 (labeled C_1 and C_4) merge into C_2^3 (labeled C_1). From S^3 to S^4 we see that C_3^3 (labeled C_8) grows into C_2^4 (labeled C_8, identical to C_3^3), C_1^3 and C_2^3 (labeled C_5 and C_1) merge into C_1^4 (labeled C_1), and finally, C_4^3 and C_5^3 (labeled C_6 and C_7) merge into C_3^4 (labeled C_6). Note that since $S^5 = K$, going from S^4 to S^5 the three components C_1^4, C_2^4, C_3^4 in S^4 (labeled C_1, C_8, C_6) merge into $C_1^5 = S^5 = K$.

Labeling Connected Components

Starting with the largest value of $\lambda = \lambda_1$, each connected component in S^1 is given as label a unique number in the range $[1, n_1]$, where n_1 is the number of components in S^1. Going to λ_2, if a component grows from S^1 to S^2, it is given the same label as before. If a new component appears in S^2, it is labeled with the next available number that has not been used yet. And if components $C_{k_1}^1, \ldots, C_{k_n}^1$ of S^1 merge into component C_j^2, we label C_j^2 with the smallest of the labels for $C_{k_1}^1, \ldots, C_{k_n}^1$. Continuing this process for all λ values, each *component label* is specified as $C_k = (C_{k(1)}^1, \ldots, C_{k(i)}^i, \ldots, C_{k(N)}^N)$ where the entry $C_{k(i)}^i$ denotes the connected component of S^i labeled as C_k. In general if a component label C_k appears for the first time in S^i for $i > 1$ and disappears when it merges with another component in S^j for $j > i$, then $C_k = (\emptyset, \ldots, \emptyset, C_{k(i)}^i, \ldots, C_{k(j-1)}^{j-1}, \emptyset, \ldots, \emptyset)$. In particular, there are no connected components

of S^h corresponding to C_k for any $h < i$ or $h \geq j$. The complete list of component labels for T is obtained by repeating this process for all components across all λ values, and is denoted by $\mathcal{C}$.

For the example in Figure 3, the following component labels are generated for the five λ values by this step of the `GMTChem.FN` algorithm:

$$C_1 = (C_1^1, C_2^2, C_2^3, C_1^4, K); \qquad C_2 = (C_2^1, \emptyset, \emptyset, \emptyset, \emptyset); \qquad C_3 = (C_3^1, \emptyset, \emptyset, \emptyset, \emptyset);$$
$$C_4 = (\emptyset, C_1^2, \emptyset, \emptyset, \emptyset); \qquad C_5 = (\emptyset, C_3^2, C_1^3, \emptyset, \emptyset); \qquad C_6 = (\emptyset, C_4^2, C_4^3, C_3^4, \emptyset);$$
$$C_7 = (\emptyset, \emptyset, C_5^3, \emptyset, \emptyset); \qquad C_8 = (\emptyset, \emptyset, C_3^3, C_2^4, \emptyset).$$

Structure Ranking

The input of this step of the `GMTChem.FN` algorithm is the collection $\mathcal{C}$ of all component labels over all frames of data along with scale parameter values $\{\lambda_i\}_{i=1}^N$ and the ratio cutoffs $\{r_j\}_{j=1}^M$. The output is a ranking of all component labels in $\mathcal{C}$ in terms of the number of times each individual component is *alive* over all scale parameter values and ratio cutoffs.

Alive components

A component label $C_k = (C_{k(1)}^1, \ldots, C_{k(i)}^i, \ldots, C_{k(N)}^N)$ is defined to be *alive* at scale λ_i and ratio r_j if the volume of $C_{k(i)}^i$ is greater than the volume of a molecule associated with the soft-matter surface, and if the ratio of the volume of $C_{k(i)}^i$ to the volume of a ball of radius $1/\lambda_i$ is strictly greater than the ratio r_j. For the example system of water/hexane/TBP the volume of a water molecule is chosen. Precisely, the component label C_k is alive at scale λ_i and ratio r_j if both

$$\text{Vol}(C_{k(i)}^i) \geq \text{Vol(molecule) and} \tag{3}$$
$$R(C_{k(i)}^i) > r_j, \tag{4}$$

where R is defined to be the ratio

$$R(C^i_{k(i)}) = \frac{\text{Vol}(C^i_{k(i)})}{\text{Vol}(\text{Ball of radius } \lambda_i^{-1})}\,. \tag{5}$$

In the example in Figure 3, two ratio cutoff values $r_1 > r_2$ are used. To determine how often (i.e., number of (λ_i, r_j) pairs) at which the component labels are alive, the volumes of corresponding connected components

$$\begin{aligned}
\text{Vol}(C_1) &= (\text{Vol}(C_1^1), \text{Vol}(C_2^2), \text{Vol}(C_2^3), \text{Vol}(C_1^4), \text{Vol}(K));\\
\text{Vol}(C_2) &= (\text{Vol}(C_2^1), 0, 0, 0, 0);\\
&\vdots\\
\text{Vol}(C_8) &= (0, 0, \text{Vol}(C_3^3), \text{Vol}(C_2^4), 0),
\end{aligned}$$

are divided by volumes of the corresponding λ_i^{-1}-balls, and compared to the r_j values to determine which labels satisfy the inequalities (3) and (4). In Figure 3, the left and right columns correspond to ratio cutoffs r_1 and r_2, respectively. Further, the components colored in red are ones that are alive at the respective λ_i and r_j values.

Each component label $C \in \mathcal{C}$ has a matrix $A_C = [a_{ij}]$ where $a_{ij} = 1$ if C is alive at λ_i and r_j, and 0 otherwise. One can think of A_C as a lookup table that records which λ_i and r_j values the component label is alive at. Referring back to Figure 3, to construct the first column of the lookup table A_{C_1} for component label C_1 corresponding to ratio r_1, we simply look at the first column. Since C_1 is colored red only for $\lambda = \lambda_1$, $a_{11} = 1$ and all other entries in the first column of A_C are 0. To construct the second column of A_{C_1} corresponding to ratio r_2, we look at the second column. In this case, C_1 is colored red for $\lambda = \lambda_1, \lambda_2, \lambda_3$ and therefore $a_{12} = a_{22} = a_{32} = 1$ and entries $a_{42} = a_{52} = 0$. This gives us our completed lookup

table for component label C_1: $A_{C_1} =$

$$\begin{array}{c} \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{array} \begin{array}{c} \begin{array}{cc} r_1 & r_2 \end{array} \\ \left[\begin{array}{cc} 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{array}\right] \end{array} .$$

Therefore, the probability that C_1 is alive for a uniformly random λ_i and r_j is given as

$$P(C_1 \text{ is alive}) = 4/10 = 0.4.$$

Repeating these steps for the remaining component labels $C_2, \ldots, C_8$ in Figure 3 yields the following probabilities:

$$P(C_2 \text{ is alive}) = 0; \quad P(C_3 \text{ is alive}) = 0.1; \quad P(C_4 \text{ is alive}) = 0.1; \quad P(C_5 \text{ is alive}) = 0;$$
$$P(C_6 \text{ is alive}) = 0.1; \quad P(C_7 \text{ is alive}) = 0.2; \quad P(C_8 \text{ is alive}) = 0\,.$$

It is proposed that component labels that are alive more frequently are better candidates for being identified as structural features on a soft-matter surface. Under this proposition, all component labels are ranked from the most to least likely candidates to be identified as structural features. In our example in Figure 3, the ranking for labels in $\mathcal{C}$ is

$$(C_1, C_7, C_3, C_4, C_6, C_2, C_5, C_8).$$

It is advisable to obtain an appropriate range of ratio cutoff values, which is obtained in this work by employing a *reference surface* that has few collectively organized structures.

Results

To test the applicability of our algorithm to identify an ensemble of surface structures at a soft-matter interface we use a complex liquid/liquid system that has been the topic of prior study: water/hexane/TBP, where the amphiphilic molecule TBP acts as a surfactant. In this system, protrusions of water and TBP form at the instantaneous liquid/liquid interface, and have been implicated as being the mechanism for water transport into the organic phase.[27] Details of the molecular dynamics simulation protocol, the composition and periodic box size, and construction of the Willard-Chandler interfaces for each snapshot are provided in the Supporting Information. To quantify the robustness of the surface structure algorithm, 119 protrusions were identified by visual inspection in 35 snapshots of the molecular dynamics trajectory. The visual identification is completely subjective, and as this is the first algorithmic approach to identify surface structures, it is entirely possible that chemical intuition does not always follow volumetric trends of features as identified by `GMTChem.FN`.

Analogous MD simulation data of water/hexane were used as the reference chemical system to determine the range of λ values to be employed within the algorithm. Using 80 frames of simulation data, the upper range of λ was chosen to be 0.5 as this value led to almost no flattening of the Willard-Chandler surfaces, while a lower value of $\lambda = 0.01$ was chosen because this led to complete flattening of the WC surfaces across all trajectory data (Figure S1). Fifty uniformly spaced values of lambda between these values were utilized. Within the reference data, the distribution of ratios for any component with volume of at least $V_{\min}$ was also examined. Looking at the maximum ratio of the distribution after removing the top $1, 2, \ldots, 10$ percent of ratios, the list of ratio cutoffs in Table 1 was obtained. A protrusion was defined to contain at least one water molecule. Given the density threshold of 95 percent and a Gaussian kernel radius of 2 Å used to construct the Willard-Chandler interfaces, the volume inside the 0.95 super-level-set of a Gaussian kernel of radius 2 Å in 3D was computed in order to estimate the correct volume lower bound, which is denoted by $V_{\min}$. Only those components whose volume is at least $V_{\min}$ are considered.

Table 1: The maximum ratios (see Equation (5)) after removing the top 0–10 percent of ratios from the water-hexane chemical system.

Ratio	0.0840	0.0945	0.1050	0.1155	0.1365	0.1575	0.1890	0.2520	0.3360	0.4935	1.0500
%	10	9	8	7	6	5	4	3	2	1	0

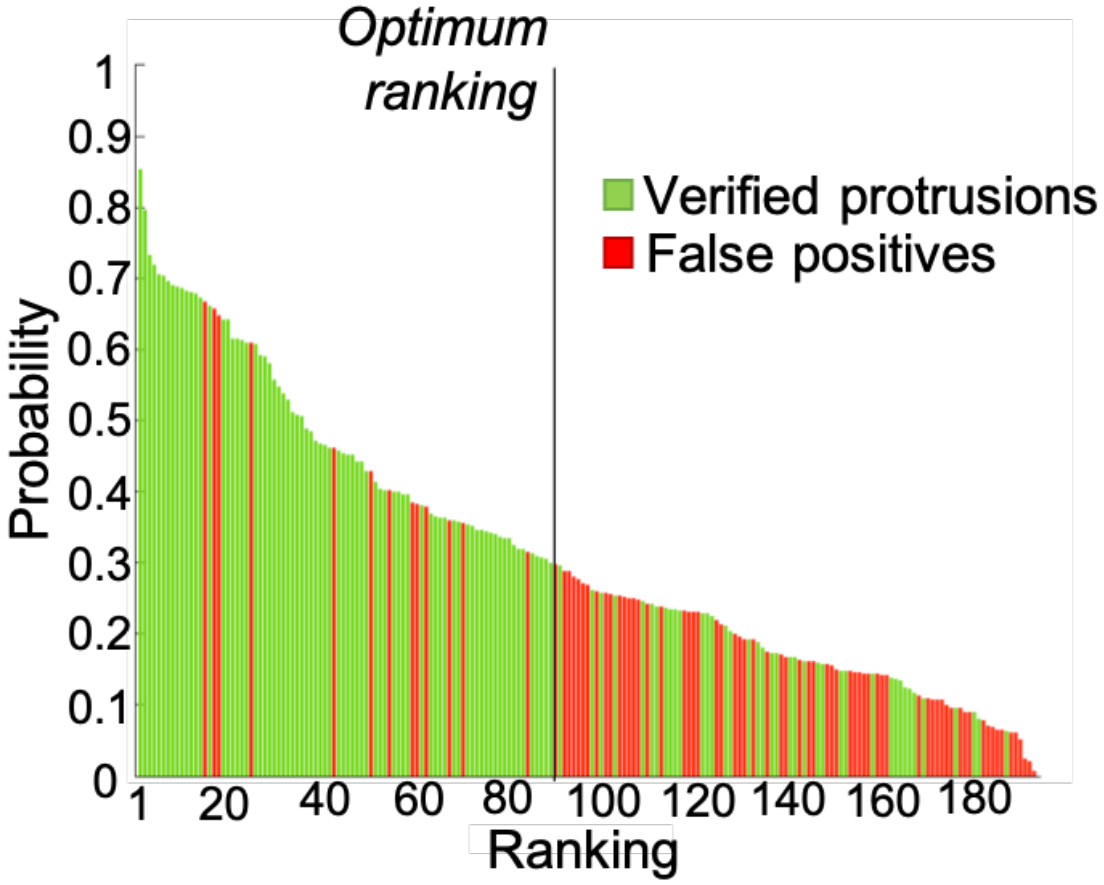

Figure 4: The probability vs. ranking of all component labels identified by the `GMTChem.FN` for 35 snapshots of the water/hexane/TBP interface. All protrusions identified by `GMTChem.FN` were cross referenced against 119 protrusions at the interface identified by subjective visual inspection, and those in agreement are labeled in green as "verified protrusions". "False positives" are those surface structures that do not appear to be protrusions by visual inspection, but may still be interesting surface structures. The optimum ranking threshold minimizes the total difference of verified protrusions and false positives, thus maximizing the identified protrusions, and is found to be $T^* = 90$.

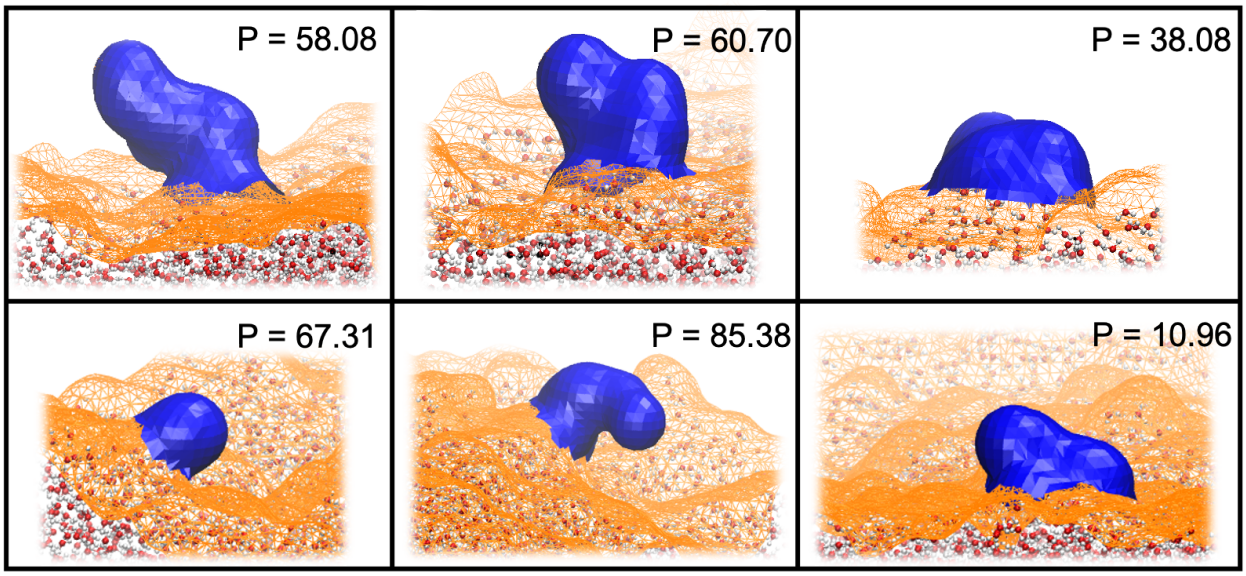

Figure 5: Examples of protrusion diversity identified by `GMTChem.FN` at the water/hexane/TBP interface, alongside their associated probability values (as percentages).

Using the `GMTChem.FN` algorithm on the 35 frames of water/hexane/TBP data, all of the protrusions identified by visual inspection exist within the total 195 component labels that were ranked. In Figure 4, we present the ranking of the component labels vs. respective probabilities. Notably, as the probability value of a component label decreases the likelihood of having a "false positive" for a protrusion (a feature that by visual inspection does not appear to be a protrusion) increases. At the same time, the concept of a "false positive" is somewhat misleading, as further study is needed across many chemical systems to understand the reactivity and chemical behavior of surface structures as a function of their size, morphology, and chemical composition. If we assume that the visually identified protrusions are the only surface structures of interest, then the optimum ranking threshold (T^*) that minimizes the total number of errors and maximizes the identified protrusions is found to be $T^* = 90$. Considering all component labels below a T^* ranking of 90, there are 13 overcounts (red bars to the left of the Optimum ranking line in Figure 4). Upon further analysis, these are observed to be features that occur on the periodic boundary of the simulation cell.

Indeed, the algorithm employed to create the Willard-Chandler surface did not obey the periodic boundary conditions of the simulation box, and as such the `GMTChem.FN` overcounted protrusions on these boundaries. Amending this issue is a topic of current algorithmic optimization. If the T^* ranking of 90 is chosen there are 43 undercounted surface structures that are visually identified as protrusions (green bars to the right of the Optimum ranking line in Figure 4). These protrusions have morphologies that consist primarily of broad features and we hypothesize their lower probability values are due to the fact that broad features are not removed as quickly as long and thin features that emerge at higher probabilities (Figure 5). Note that the entries in the probability matrices are uniformly weighted, and further improvement on the algorithm may be obtained by using varying weights. For example, if one would like to rank large components with high priority, small λ values may be weighted more than larger λ values. Finally, it is important to point out that the `GMTChem.FN` algorithm also identifies surface structures that were not initially identified by visual inspection. These surface features were missed by visual inspection and were difficult to identify because they resided in low basins, and were surrounded by high interfacial roughness (e.g., see the last feature in Figure 5 with P = 10.96).

Conclusion

Robust algorithms that automatically identify complex structural features at soft-matter surfaces have the potential to dramatically expand our ability to study interfacial reactivity and transport mechanisms. This includes understanding how surfactants can control the morphology of protrusions that move solutes across a phase boundary, or how interfacial composition modulates micelle formation. It is a significant challenge to create such an approach, as surface feature morphology can be highly varied and because the method must be able to identify structures that can exist on top of a surface that has significant natural roughness. To address this challenge, a new algorithm is developed which employs concepts from

geometric measure theory and algebraic topology. We utilize the *flat norm*, which minimizes a generalized area of a complicated surface (in x, y, z) representing a soft matter interface, and "flattens" the surface toward the horizontal plane with the same x, y dimensions. Using this framework, we identify relevant volumes lying between the complicated surface and the flat surface at various scales by minimization of the flat norm function. Subsequently, using a probabilistic approach those identified surface features are ranked by their likelihood of being a complex structural feature. The algorithm is tested for a surfactant-laden water/oil interface, where its ability to identifying protrusions is validated against a set of surface features identified by visual inspection. To our knowledge, this is the first example of applying geometric measure theory to analyze the properties of a chemical/materials science system. This approach is scale-independent and has potential application in many different chemical systems, and presents an important new tool to study interfacial structure, dynamics, and reactivity.

Acknowledgement

The authors acknowledge the Department of Energy, Basic Energy Sciences Separations program (DE-SC0001815) for funding. Krishnamoorthy acknowledges funding from the National Science Foundation through grants DBI-1661348 and DMS-1819229. This research used resources from the Center for Institutional Research Computing at Washington State University.

Supporting Information Available

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Yamamoto, S.; Hyodo, S.-A. Budding and fission dynamics of two-component vesicles. The Journal of Chemical Physics **2003**, 118, 7937–7943.

(2) Laradji, M.; Kumar, P. S. Dynamics of domain growth in self-assembled fluid vesicles. Physical Review Letters **2004**, 93, 198105.

(3) Laradji, M.; Sunil Kumar, P. Domain growth, budding, and fission in phase-separating self-assembled fluid bilayers. The Journal of Chemical Physics **2005**, 123, 224902.

(4) Lipowsky, R.; Brinkmann, M.; Dimova, R.; Haluska, C.; Kierfeld, J.; Shillcock, J. Wetting, budding, and fusion-morphological transitions of soft surfaces. Journal of Physics: Condensed Matter **2005**, 17, S2885.

(5) Allain, J.-M.; Amar, M. B. Budding and fission of a multiphase vesicle. The European Physical Journal E **2006**, 20, 409–420.

(6) Lian, Z.; Bu, W.; Schweighofer, K.; Walwark, D.; Harvey, J.; Hanlon, G.; Amoanu, D.; Erol, C.; Benjamin, I.; Schlossman, M. Nanoscale view of assisted ion transport across the liquid-liquid interface. Proceedings of the National Academy of Sciences of the United States of America **2018**, 201701389.

(7) Qiao, B.; Muntean, J.; Olvera de la Cruz, M.; Ellis, R. Ion Transport Mechanisms in Liquid-Liquid Interface. Langmuir **2017**, 33, 6135–6142.

(8) Servis, M.; Wu, D.; Shafer, J. The Role of Solvent and Neutral Organophosphorus Extractant Structure in their Organization and Association. Journal of Molecular Liquids **2018**, 253, 314–325.

(9) Kikkawa, N.; Ishiyama, T.; Morita, A. Molecular dynamics study of phase transfer catalyst for ion transfer through water-chloroform interface. Chemical Physics Letters **2012**, 19–22.

(10) Kikkawa, N.; Wang, L.; Morita, A. Microscopic barrier mechanism of ion transport through liquid-liquid interface. Journal of the American Chemical Society **2015**, 8022–8025.

(11) Kikkawa, N.; Wang, L.; Morita, A. Computational study of effect of water finger on ion transport through water-oil interface. The Journal of Chemical Physics **2016**, 145, 014702.

(12) Benay, G.; Wipff, G. Liquid–Liquid Extraction of Uranyl by TBP: The TBP and ions models and related interfacial features revisited by MD and PMF simulations. The Journal of Physical Chemistry B **2014**, 118, 3133–3149.

(13) Jayasinghe, M.; Beck, T. L. Molecular dynamics simulations of the structure and thermodynamics of carrier-assisted uranyl ion extraction. The Journal of Physical Chemistry B **2009**, 113, 11662–11671.

(14) Benjamin, I. Hydronium ion at the water/1, 2-dichloroethane interface: Structure, thermodynamics, and dynamics of ion transfer. The Journal of Chemical Physics **2019**, 151, 094701.

(15) Karnes, J. J.; Villavicencio, N.; Benjamin, I. Transfer of an erbium ion across the water/dodecane interface: Structure and thermodynamics via molecular dynamics simulations. Chemical Physics Letters **2019**, 737, 136825.

(16) Karnes, J. J.; Benjamin, I. Geometric and energetic considerations of surface fluctuations during ion transfer across the water-immiscible organic liquid interface. The Journal of Chemical Physics **2016**, 145, 014701.

(17) Tokman, M.; Lee, J. H.; Levine, Z. A.; Ho, M.-C.; Colvin, M. E.; Vernier, P. T. Electric field-driven water dipoles: nanoscale architecture of electroporation. PLoS One **2013**, 8.

(18) Morgan, F. Geometric Measure Theory: A Beginner's Guide, 4th ed.; Academic Press, 2008.

(19) Morgan, S. P.; Vixie, K. R. L^1TV computes the flat norm for boundaries. Abstract and Applied Analysis **2007**, 2007, Article ID 45153,14 pages.

(20) Vixie, K. R.; Clawson, K.; Asaki, T. J.; Sandine, G.; Morgan, S. P.; Price, B. Multiscale flat norm signatures for shapes and images. Applied Mathematical Sciences **2010**, 4, 667–680.

(21) Hu, Y.; Hudelson, M.; Krishnamoorthy, B.; Tumurbaatar, A.; Vixie, K. R. Median Shapes. Journal of Computational Geometry **2019**, 10, 322–388.

(22) Ibrahim, S.; Krishnamoorthy, B.; Vixie, K. R. Simplicial flat norm with scale. Journal of Computational Geometry **2013**, 4, 133–159.

(23) Willard, A. P.; Chandler, D. Instantaneous liquid interfaces. The Journal of Physical Chemistry B **2010**, 114, 1954–1958.

(24) Liu, Z.; Stecher, T.; Oberhofer, H.; Reuter, K.; Scheurer, C. Response properties at the dynamic water/dichloroethane liquid–liquid interface. Molecular Physics **2018**, 116, 3409–3416.

(25) Si, H. Constrained Delaunay tetrahedral mesh generation and refinement. Finite Elements in Analysis and Design **2010**, 46, 33–46.

(26) Si, H. TetGen, a Delaunay-Based Quality Tetrahedral Mesh Generator. ACM Transactions on Mathematical Software **2015**, 41, Available at http://tetgen.org.

(27) Servis, M. J.; Clark, A. E. Surfactant-enhanced heterogeneity of the aqueous interface drives water extraction into organic solvents. Physical Chemistry Chemical Physics **2019**, 21, 2866–2874.