THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

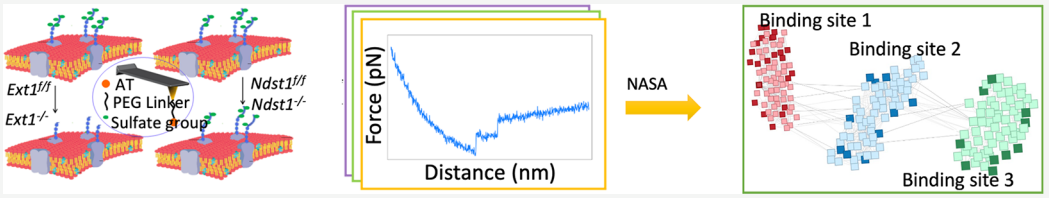
Details of Single-Molecule Force Spectroscopy Data Decoded by a Network-Based Automatic Clustering Algorithm

Huimin Cheng, Jun Yu, Zhen Wang, Ping Ma, Cunlan Guo,* Bin Wang, Wenxuan Zhong,* and Bingqian Xu*

Cite This: *J. Phys. Chem. B* 2021, 125, 9660−9667

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



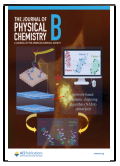
ABSTRACT: Atomic force microscopy-single-molecule force spectroscopy (AFM-SMFS) is a powerful methodology to probe intermolecular and intramolecular interactions in biological systems because of its operability in physiological conditions, facile and rapid sample preparation, versatile molecular manipulation, and combined functionality with high-resolution imaging. Since a huge number of AFM-SMFS force−distance curves are collected to avoid human bias and errors and to save time, numerous algorithms have been developed to analyze the AFM-SMFS curves. Nevertheless, there is still a need to develop new algorithms for the analysis of AFM-SMFS data since the current algorithms cannot specify an unbinding force to a corresponding/each binding site due to the lack of networking functionality to model the relationship between the unbinding forces. To address this challenge, herein, we develop an unsupervised method, i.e., a network-based automatic clustering algorithm (NASA), to decode the details of specific molecules, e.g., the unbinding force of each binding site, given the input of AFM-SMFS curves. Using the interaction of heparan sulfate (HS)−antithrombin (AT) on different endothelial cell surfaces as a model system, we demonstrate that NASA is able to automatically detect the peak and calculate the unbinding force. More importantly, NASA successfully identifies three unbinding force clusters, which could belong to three different binding sites, for both $Ext1^{f/f}$ and $Ndst1^{f/f}$ cell lines. NASA has great potential to be applied either readily or slightly modified to other AFM-based SMFS measurements that result in "saw-tooth"-shaped force−distance curves showing jumps related to the force unbinding, such as antibody−antigen interaction and DNA−protein interaction.

INTRODUCTION

Single-molecule force spectroscopy (SMFS), including optical tweezer, magnetic tweezer, and atomic force microscopy (AFM), has been developed as a powerful methodology to probe the details of intermolecular and intramolecular interactions and thus to obtain mechanistic insights of the systems that cannot be probed using traditional methods in bulk systems. SMFS is particularly useful in studying biological systems, such as protein folding/unfolding, nucleic acid structures, protein−drug interactions, and cellular surface receptor−ligand interactions,[1−17] due to its multifunctionalities. Compared with optical and magnetic tweezers, the AFM-based SMFS (AFM-SMFS) is advantageous in several aspects, such as facile and rapid sample preparation, molecular manipulation with an AFM tip by tethering the target molecules on the AFM tip, and spatial discrimination of SMFS with the high-resolution imaging.[3,14,18,19] It makes the AFM-SMFS superior in the study of biomolecular interactions, especially in complicated physiological conditions. For example, in an early study, the unbinding force of five different avidin−biotin pairs was probed using AFM-SMFS.[2] Recently, AFM-SMFS was used to image single human protease-activated receptor-1 (PAR1), a typical G protein-coupled receptor, in proteoliposomes. Moreover, AFM-SMFS was used to simultaneously quantify the dynamic binding strength of PAR1 to different ligands under physiologically relevant conditions.[20]

A typical AFM-SMFS force−distance curve provides unbinding events for intermolecular studies. The change of rupture force with a loading rate and the shape of rupture force distribution can provide parameters of the energy landscape of the intermolecular reaction, such as zero-force dissociation rate constant, the distance to the transition state, and the height of the energy barrier.[18,21] The accuracy of the parameters requires the statistical analysis of a large amount of the AFM-SMFS

Received: April 20, 2021
Revised: August 6, 2021
Published: August 23, 2021



https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667

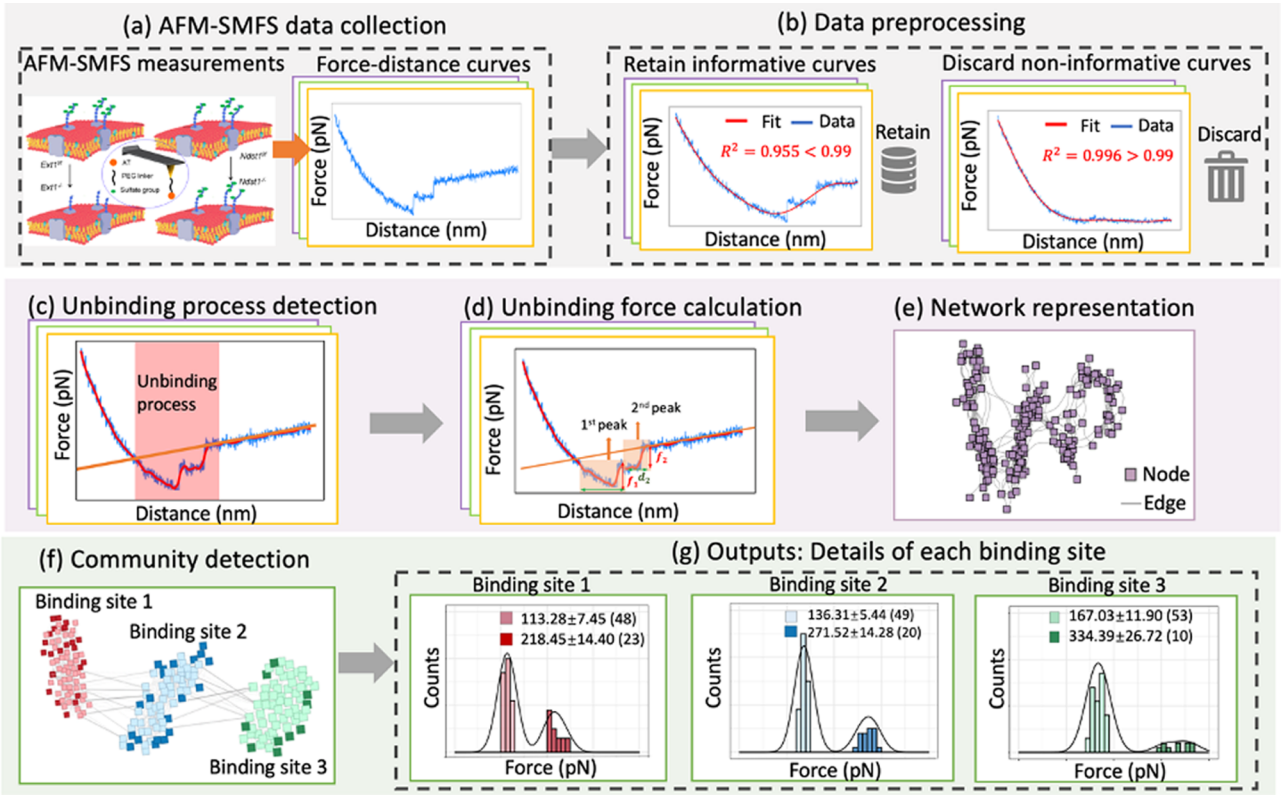


Figure 1. Workflow of NASA. (a) AFM-SMFS data collection. In this case, the interaction of HS−AT on different endothelial cell surfaces is studied. An AFM tip is modified with AT to probe the HS−AT interaction. (b) Data preprocessing. In this procedure, NASA retains informative curves and removes noninformative curves. (c) Unbinding process detection. (d) Unbinding forces calculation. (e) Network representation. (f) Community detection. (g) Outputs of NASA: details of each binding site, including the histogram of unbinding forces, and the average force ± standard deviation.

curves, which suffers from the subjectivity of manual methods. To avoid these biases and human errors and to save time, established algorithms are employed to analyze the AFM-SMFS curves. These algorithms include Hooke,[22] OpenFovea,[23] AtomicJ,[24] FRAME,[25] FC_analysis,[26] FEATHER,[27] ForSDAT,[28] etc. However, some of these algorithms lack automatic curve selection functions or automatic threshold selection functions to distinguish specific interactions from nonspecific interactions, while others are supervised methods that require human-labeled data. In sum, all aforementioned methods can not automatically specify an unbinding force to a corresponding/each binding site.

To tackle these challenges, herein, we present our proposed algorithm, i.e., a network-based automatic clustering algorithm (NASA). Network-based machine learning methods have shown surprising effectiveness to characterize complex systems and have gained increasing popularity in recent years.[29−32] Despite its popularity, this work is the first to leverage this powerful tool in studying the single-molecule AFM-SMFS data. Given the input (i.e., AFM-SMFS curves), NASA first extracts the unbinding forces from the curves, constructs a network depicting the relationship between the unbinding forces, and then detects the community structure of this network. Each community corresponds to a binding site, of which the details can be decoded by investigating the features (e.g., average unbinding force) in this community.

Previously, we have conducted an AFM-SMFS study of heparan sulfate (HS)−antithrombin (AT) interaction on different endothelial cell surfaces. Using an AT-functionalized AFM tip, we revealed that the AT interacts with endothelial HS on the cell surface through multiple binding sites.[33] Here, we apply this new NASA method to decode the complex HS−AT interaction on different cell surfaces under physiological conditions with more experimental data. We demonstrate the capabilities of NASA for a diverse set of experimental data (i.e., the single-molecule interaction of HS with AT on four endothelial cell membrane surfaces), confirming the accuracy of our manually statistical results. More importantly, we show that NASA is capable of identifying and characterizing the specific unbinding force for each binding site, which has not been achieved by conventional analysis methods. Such biophysical details on the AT−HS binding modes on the cell surface will help us to have a deeper understanding of the function of HS in both physiological and pathological processes. More importantly, this provides a radical method that can be applied to any system to reveal more accurate single-molecule interaction properties.

METHODS

Architecture of NASA. Figure 1 shows the workflow of NASA. Given the input data (i.e., force−distance curves resulted from AFM-SMFS measurements), as shown in Figure 1a, NASA first filters out noninformative curves, as shown in Figure 1b. Here, the noninformative curve is defined as a curve that does not show any unbinding event, thus providing no information about the binding site. After this data preprocessing, for each curve, NASA then automatically detects the unbinding process during which the unbinding event happens, as shown in Figure

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667

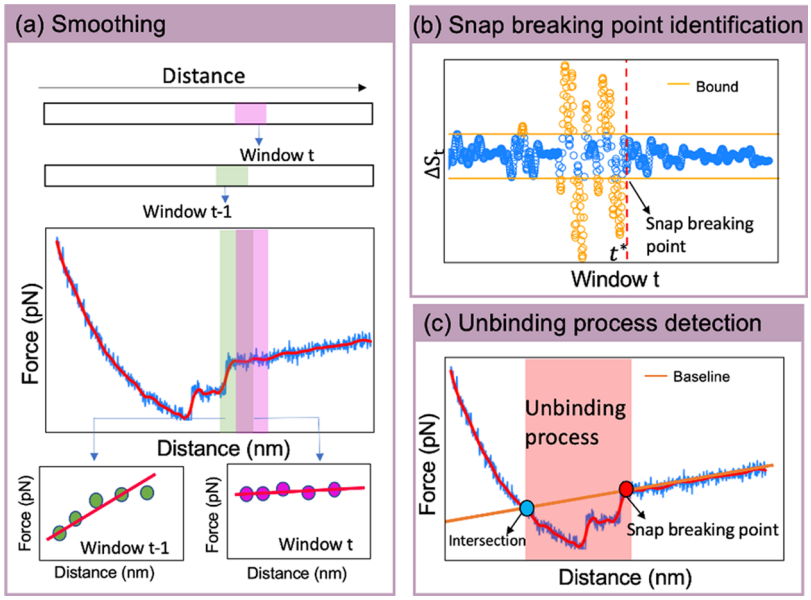


Figure 2. (a) Example of the fitted smooth curve. The red curve is the smooth curve and the blue line is the original curve. There are two window examples, i.e., window t and window $t - 1$. In window $t - 1$ and window t, there are five data points (in green) and five data points (in pink), respectively. The red line fitted to five points in each window is the fitted linear regression line. (b) Illustration of snap breaking point detection. The point t^* is the snap breaking point. (c) Illustration of unbinding process detection. The solid red dot is the snap breaking point detected by (b). The baseline intersects with the force−distance curve at the solid blue dot. The unbinding process starts at the solid blue dot and ends at the solid red dot.

1c. This unbinding process provides information about the AT−HS unbinding forces. To extract this information, NASA further identifies all unbinding forces from the curves in two steps. Step 1: Identify all peaks, which indicate AT−HS unbinding events on the cell surface during the unbinding process. For example, two peaks are extracted in Figure 1d. Step 2: Calculate the unbinding force f_i of each peak i. For example, in Figure 1d, with the calibrated AFM tip spring constant, we get two unbinding forces for two peaks.

To specify the unbinding force for each binding site, NASA employs network-based methods. More specifically, NASA constructs a network, where nodes represent unbinding forces, and edges represent the relationship between unbinding forces, as shown in Figure 1e. NASA then clusters nodes, as shown in Figure 1e, into different communities by leveraging a community detection algorithm.[34] Using the aforementioned detected community information, we redraw the network in a new layout so that nodes in the same community are close to each other, while nodes in different communities are far away from each other, as shown in Figure 1f. Different communities represent different types of binding sites. By characterizing each community, we get the details (e.g., the mean unbinding force) of each binding site, as shown in Figure 1g.

Three characteristics distinguish NASA from other methods: (1) NASA is an unsupervised method that learns patterns from the data without human intervention. (2) NASA employs the network-based method in an innovative way, which allows us to distinguish each binding site and investigate their details. (3) NASA is an automatic tool that avoids time-consuming and error-prone manual efforts.

AFM-SMFS Data Preprocessing. When conducting the AFM-SMFS measurement experiments, we use an AFM tip to investigate the unbinding force between the AT on the AFM tip and the HS on the cell surface. Usually, unbinding events are observed on SMFS curves, as shown in the peaks in the right panel of Figure 1a. Nevertheless, in some cases, the force−distance curves may only measure a very gentle tip-bare membrane interaction, i.e., the tip barely touches the membrane. An example of such curves is shown in the right panel of Figure 1b, from which we observe that (i) this curve is rather smooth and (ii) no unbinding event is present. Such curves do not express any information about the unbinding force. To filter out these noninformative curves, NASA works in the following steps.

The first step is to fit the kth degree polynomial regression model (see the Supporting Information, SI 1.1, for details), i.e., $y = a + b_1x^1 + ... + b_kx^k$ for each force−distance curve. Herein, y denotes the force and x denotes the distance. We employ the Bayesian information criterion (BIC)[35,36] (see SI 1.1 for details) to determine the appropriate polynomial order k. The second step is to calculate the corresponding R-squared (R^2)[37] value (see SI 1.1 for details) for the selected model. The R^2 value measures how well the model fits the force−distance curve. The third step is to exclude curves with R^2 greater than a threshold. The threshold is determined by a data-driven method (see SI 1.5.1), which is general for all unbinding processes, but the specific value may vary with different systems due to the different details of the force−distance curves. Particular to the system presented in this paper, the threshold is determined to be 0.990. In other words, a curve with R^2 less than 0.99 is labeled as an informative curve, and thus is retained. Otherwise, the curve is labeled as a noninformative curve, and thus is discarded.

Unbinding Process Detection. After filtering out those noninformative force−distance curves, we detect the unbinding process with abrupt changes in force.[38,39] The detection of the unbinding process depends on the detection of the baseline, i.e., the zero-force background line. Usually, the baseline is flat if we operate on a hard surface. Nevertheless, in this work, we operate on a living cell, which is very soft. A

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667

sloped baseline is thus very common in practice,[40] requiring us to consider the baseline to measure the rigidity of the cell membrane.

To detect the baseline, we first smooth the curve by employing the local regression method[41] (see SI 1.2.2 for details), which can smooth out the noise while preserving the main trend. The red curve in Figure 2a shows an example of the fitted smooth curve. Our following analysis is based on the fitted smooth curve, i.e., the y in the following analysis denotes the value in the fitted smooth curve.

After this denoising procedure, we then aim to detect the snap breaking point, which starts the baseline. Before the snap breaking point, the local slope of the force−distance curve changes drastically, while after the snap breaking point, the local slope of the force−distance curve seldom changes. Motivated by this observation, we develop a sliding window method to calculate the local slope and detect the snap breaking point. A sliding window method is a standard statistical technique to detect local pattern.[42,43] Herein, a window consists of a proportion of all data points on the smoothed force−distance curve, the proportion is denoted as $h \in [0,1]$.

The window slides from left to right over the force−distance curve. For example, in Figure 2a, we show two consecutive windows, i.e., window $t-1$ (highlighted in green) and window t (highlighted in pink). We train a linear regression model in each window. For example, the red lines are the fitted linear regression line using the corresponding data points. If h is too big, e.g., 0.5, the pattern will be "smoothed out", and we cannot accurately detect the snap breaking point. If h is too small, e.g., 0.001, it will be too sensitive to accurately detect the snap breaking point. By conducting extensive studies on different force−distance curves, we suggest that the appropriate value for h is between 0.05 and 0.1 (see SI 1.5.2). In this work, we set $h = 0.05$.

Suppose the number of windows is T, we thus fit T regression models, and we have T corresponding regression coefficients (i.e., slopes) $\{S_1, ..., S_T\}$. Intuitively, when the window moves after the snap breaking point, the slope does not change dramatically. We then utilize the change of the slopes to detect the snap breaking point. More specifically, we consider the slope change $\Delta S_t = S_t - S_{t-1}$ between two consecutive windows t and $t-1$. If $|\Delta S_t| < s\,\mathrm{d}(\Delta S_t)$, i.e., ΔS_t falls within the upper and lower orange bounds, as shown in Figure 2b, we accept window t as a nonchange point. Otherwise, if ΔS_t falls outside the bounds, the t is considered as a change point. As shown in Figure 2b, blue points are nonchange points, and orange ones are change points. If t^* is a change point while $\{t^*+1, ..., T\}$ are all nonchange points, t^* is identified as the snap breaking point. With the snap breaking point t^* (e.g., the solid red dot in Figure 2c) identified, we then fit a linear regression model (see SI 1.2.1 for details) for all data after the snap breaking point. The fitted line is the estimated baseline, e.g., the solid orange line in Figure 2c.

Extending the baseline to an intersection with the smooth force−distance curve, we extract the unbinding process between the intersection and the snap breaking point. As shown in Figure 2c, the force−distance curve in the red shaded area is the unbinding process.

Unbinding Forces Calculation. Identifying the unbinding events, i.e., the peaks during the unbinding process, is an essential step for calculating the unbinding forces. First, we detect the boundary points (shown in the green stars in Figure 3) between two unbinding events. Near the boundary point, there are three possible changes of the first derivative of y, and

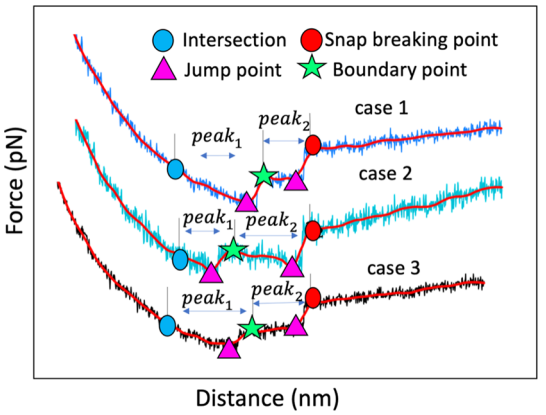


Figure 3. Illustration of unbinding event detection. The blue dot denotes the intersection between the baseline and the force−distance curve. The red dot denotes the snap breaking point. The pink triangle denotes the jump point. The green star denotes the boundary point, which separates two adjacent peaks.

we denote the first derivative of y as y'. (i) y' changes from positive to zero, as shown in case 1 in Figure 3. (ii) y' changes from positive to negative, as shown in case 2 in Figure 3. (iii) y' changes from a higher positive value to a lower positive value, as shown in case 3 in Figure 3. All three cases suggest that the boundary point is a change point when y' changes and $y'' < 0$. Hence, we first employ a change point detection method (see SI 1.3 for details) to detect all change points of y', and then identify the change points at which $y'' < 0$ as the boundary points.

Second, we detect the jump point (shown in the pink triangles in Figure 3) representing the "snap-off" of the interaction bonding in each unbinding event. The force difference between the jump point and the boundary point or the snap breaking point is calculated as the unbinding force. Near the jump point, there are also three possible changes of the first derivative y': (i) y' changes from zero to positive, as shown in case 1 of Figure 3. (ii) y' changes from negative to positive, as shown in case 2 of Figure 3. (iii) y' changes from a lower positive value to a higher positive value, as shown in case 3 of Figure 3. All three cases have a common fact: the jump point is a change point when y' changes and $y'' > 0$. Hence, we first employ a change point detection method (see SI 1.3 for details) to detect all change points of y', and then identify the change points at which $y'' > 0$ as the jump points.

After finding the boundary point and jump point, we calculate the force difference in the raw data between the jump point and boundary point as unbinding force. Finally, we obtain all unbinding forces f_i, $i = 1, ..., n$, from the force−distance curves, where n is the number of unbinding forces. To ensure that the unbinding forces we include represent the specific AT−HS unbinding events, forces less than 40 pN are excluded because the typical noise level of 10−40 pN is determined from the baseline fluctuations.

Network Representation. Previous methods calculate the unbinding forces from force−distance curves. Nevertheless, it is hard to distinguish different binding sites from the massive unbinding forces. Here, we consider that one binding site is corresponding to a certain unbinding force for a single-

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667

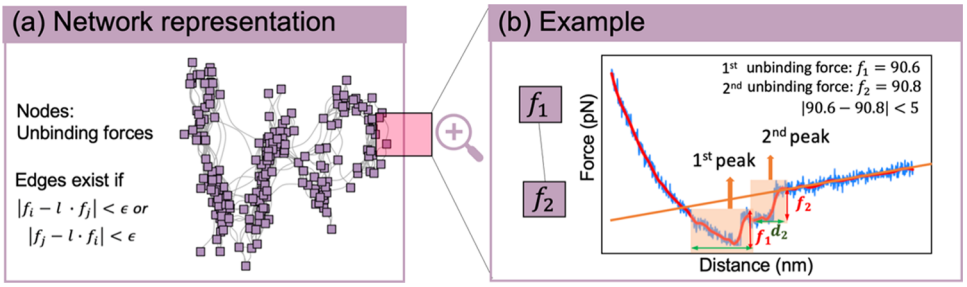


Figure 4. (a) Network for cell line Ext1$^{f/f}$. (b) Example of two nodes, i.e., f_1 and f_2, in the network shown in (a). f_1 measured from the first peak is 90.6. The f_2 measured from the second peak is 90.8. Since $|90.6 - 90.8| < 5$, f_1 and f_2 are connected.

molecular AT−HS interaction. If we have two AT−HS interactions from the same binding site, the unbinding force would be twice the one AT−HS interaction. Different binding sites have different unbinding forces, which do not have integer multiple relationship. Taking all of the abovementioned considerations into account, we employ a network representation method to shed light on the relationship between unbinding forces and then provide information for different binding sites.

For each cell line, we construct an undirected network G, where each node represents an unbinding force, as shown in Figure 4a. We draw an edge between two unbinding forces f_i and f_j if they have an approximate multiplicative relationship, i.e., there exists a positive integer l satisfying the following condition,

$$|f_i - l \cdot f_j| < \epsilon \text{ or } |f_j - l \cdot f_i| < \epsilon$$

where ϵ is a predefined tolerance parameter that specifies the maximum acceptable variation. Two unbinding forces f_i and f_j are defined to be similar if $|f_i - f_j| < \epsilon$. If ϵ is too large, e.g., 100, we will draw an edge between two forces with a significant difference, e.g., $f_1 = 100$ and $f_2 = 190$. If ϵ is too small, e.g., 0.1, even two forces shown in Figure 4b will not have an edge; we will thus lose too much information. In this study, we set $\epsilon = 5$, by taking all of the abovementioned considerations into account.

For example, as shown in Figure 4b, f_1 and f_2 are connected because they are similar. In this case, these two unbinding forces may come from the same binding site. If two unbinding forces f_i and f_j are connected with $l = 2$, and this means that either f_i is almost twice the f_j or f_j is almost twice the f_i. In this case, these two unbinding forces may also come from the same binding site. Thus, if two nodes are connected, they have a high probability of coming from the same binding site.

Community Detection. In a network, a community is a group of nodes that have dense edges within the community and sparse edges between communities. In this work, an edge is constructed between two nodes (forces) if they are likely to come from the same binding site. Thus, nodes (forces) in the same community are likely to come from the same binding site. Community detection aims to find out the communities.[44,45] In this work, we employ the fast-greedy modularity optimization algorithm[46] (see SI 1.4.2 for details) for community detection. This algorithm clusters nodes into different communities, each of which may correspond to a binding site. By characterizing each community, e.g., the mean unbinding force, we present the details of each binding site.

RESULTS AND DISCUSSION

We selected a set of AFM-SMFS curves of different cell lines Ext1$^{f/f}$, Ext1$^{-/-}$, Ndst1$^{f/f}$, and Ndst1$^{-/-}$ to proceed with the NASA analysis, which yielded 442, 857, 650, and 613 force−distance curves, respectively. Among these collected curves, NASA identifies 215, 209, 147, and 168 curves as informative curves for these four cell lines. By analyzing these informative curves, NASA extracted 376, 306, 246, and 212 unbinding forces for the four cell lines. For each cell line, NASA leveraged these unbinding forces to construct a network.

In summary, we constructed four networks for these four cell lines. Table 1 presents the summary statistics of these four networks, including the number of nodes, the number of edges, and the clustering coefficient. The clustering coefficient (see SI 1.4.1 for details) measures the degree to which nodes in a network tend to cluster together.[35] The number of nodes ranges from 212 to 376, while the number of edges is between 476 and 14 689. The clustering coefficients of the four networks are all greater than 0.5, suggesting the existence of communities.

Table 1. Summary Statistics for Four Networks of Four Different Cell Lines

cell line	Ext1$^{f/f}$	Ext1$^{-/-}$	Ndst1$^{f/f}$	Ndst1$^{-/-}$
number of nodes	376	306	246	212
number of edges	14 689	1730	8746	476
clustering coefficient	0.75	0.90	0.66	0.90

By employing the fast-greedy modularity optimization algorithm,[46] we clustered nodes into different communities for each network. Figure 5 visualizes four networks (for four different cell lines). For each network, nodes in the same color come from the same community. It is observed that the optimal numbers of communities in networks of both the Ndst1$^{f/f}$ and Ext1$^{f/f}$ cell lines are three, as shown in Figure 5. This observation suggests that there are three possible binding sites for Ndst1$^{f/f}$ and Ext1$^{f/f}$ cell lines. These results are consistent with the conclusions in our experimental study.[33]

Then, we investigated the details of each community. Figure 5 shows the average force ± standard deviation (sample size) of each community. Note that Figure 5 shows the results after deleting outlier nodes. Details of outlier node deletion can be found in SI 1.2.1. Figure 5a shows the unbinding force histogram in the red community of the Ext1$^{f/f}$ cell surface. We can observe that the unbinding forces can be further divided into two groups. The forces in the dark red group (i.e., right group) are almost twice the light red one (i.e., left one). Other communities show an analogous pattern, as shown in Figure

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667

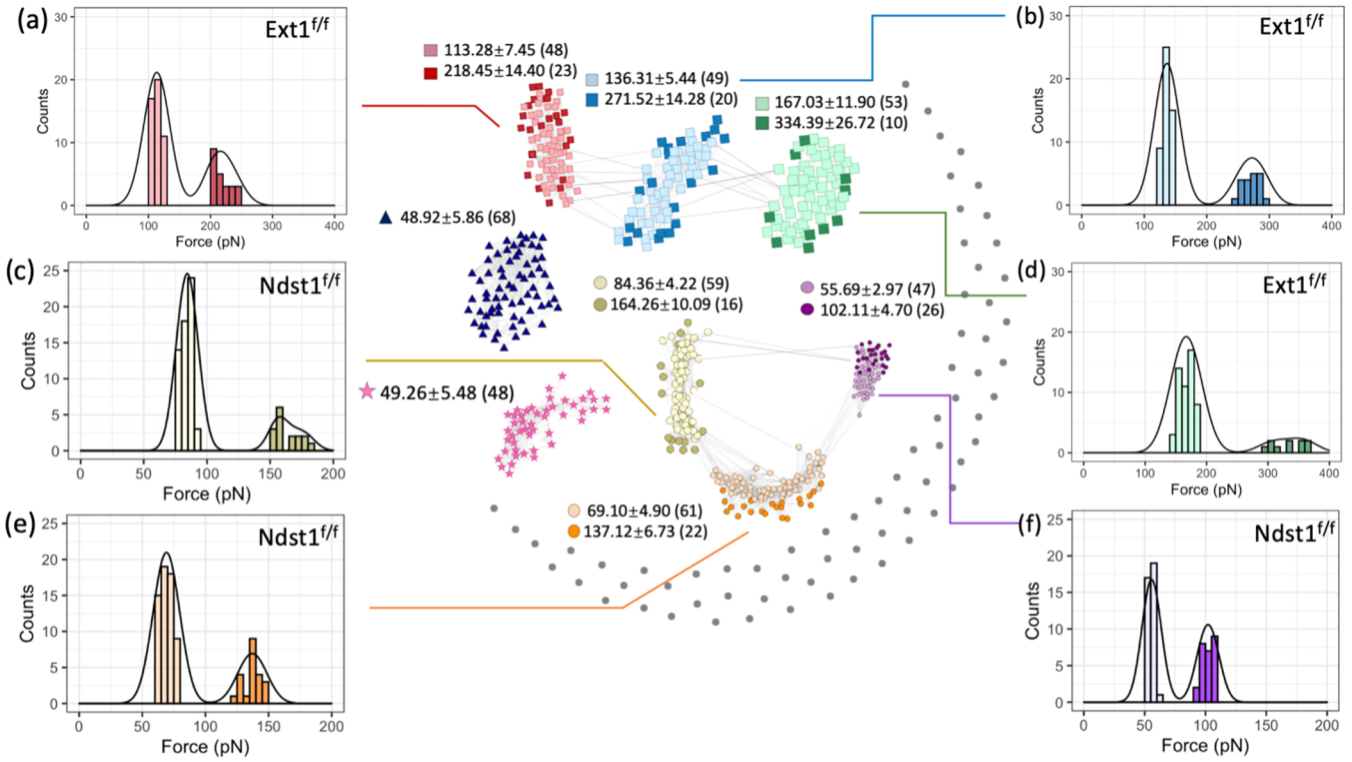


Figure 5. Visualization of four networks of four different cell lines. We use different symbols to distinguish four cell lines: Ext1$^{f/f}$ (square), Ext1$^{-/-}$ (triangle), Ndst1$^{f/f}$ (circle), and Ndst1$^{-/-}$ (star). In Ext1$^{f/f}$ and Ndst1$^{f/f}$ cell lines, we use a lighter color to represent the interaction force between a single AT and a single HS molecule and the darker color to denote the interaction force between two AT with two HS molecules. The number annotated beside the colored symbol is, average force ± standard deviation, and the number in the parentheses is the sample size. We only display parts of the noise and outlier nodes, which are colored in gray. (a–f) Force distributions for each community of Ndst1$^{f/f}$ and Ext1$^{f/f}$ cell lines.

5b–f. Herein, we focus on the analysis results of Ndst1$^{f/f}$ and Ext1$^{f/f}$ cell lines. More analysis results of Ext1$^{-/-}$ or Ndst1$^{-/-}$ could be found in SI 1.2.2.

These observations suggest the following conclusions. (1) There are three binding sites between AT and endothelial HS on Ndst1$^{f/f}$ and Ext1$^{f/f}$ cell lines. (2) In each binding site, the first group corresponds to the case when a single molecule from this site is measured. (3) In each binding site, the second group corresponds to the case when two molecules are measured.

We then conducted a manual analysis of the same data set used in this paper. First, manually extracted unbinding forces and NASA output unbinding forces have similar distribution (see SI 1.7.1), confirming that NASA can calculate unbinding forces accurately. Second, we conducted a manual analysis for these unbinding forces using our previous method[33] (see SI 1.7.1). In addition, we compared NASA with other clustering methods, i.e., PCA-based clustering, autoencoder-based clustering, K-means, and density-based spatial clustering of applications with noise (DBSCAN) (see SI 1.7.2). Briefly, the NASA method can group the data and identify three binding sites, while manual analysis and other clustering methods cannot provide these details.

CONCLUSIONS

We have developed a new algorithm, NASA, for unsupervised automatic analysis of AFM-SMFS data. We apply the new algorithm to analyze our previous AT–HS interaction on different endothelial cell lines and are able to automatically detect the peak and calculate the unbinding force under physiological conditions. Moreover, we demonstrate that NASA is able to group the data into three clusters and thus identify three corresponding binding sites for both Ext1$^{f/f}$ and Ndst1$^{f/f}$ cell lines. It should be noted that currently, the exact structure of binding sites is not clear, which may be addressed in the future by combining other complementary techniques such as molecular dynamics simulation. NASA has great potential to be applied either readily or slightly modified to other AFM-based SMFS measurements that result in "saw-tooth"-shaped force–distance curves showing jumps related to the force unbinding, such as antibody–antigen interaction and DNA–protein interaction. Application of our method to other AFM-based SMFS data is beyond the scope of the current publication and will be studied in our future work.

There are also inevitable limits to our method. First, the unbinding forces we analyze seem to be well separated, making it relatively easy to group into communities. Other biomolecular interactions may not have unbinding forces that are as well separated, so in our ongoing work, we are trying to extend NASA to analyze unbinding forces with overlapping distributions. Second, in this paper, we treat unbinding forces less than 40 pN as noise, which is deduced from 1 nm movement of a cantilever in the self-thermo vibration with energy k_BT, which is reasonable due to the soft, flexible, and complex cell environment at room temperature. The noise threshold may need to be adjusted for other systems since this assumption may not be suitable for other AFM-SMFS data collected from other interactions.

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660–9667

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c03552.

(SI 1.1) Methods used in data preprocessing, i.e., polynomial regression and BIC; (SI 1.2) methods used in unbinding process detection, i.e., linear regression and local regression; (SI 1.3) methods used in unbinding force calculation, i.e., change point detection; (SI 1.4) methods used in network representation, i.e., the clustering coefficient and fast-greedy modularity optimization algorithm; (SI 1.5) how to select parameters of NASA; (SI 1.6) sensitivity analysis of NASA to different levels of noise; (SI 1.7) results of comparing NASA with other methods; (SI 1.8) additional community detection results of NASA for Ext1$^{-/-}$ and Ndst1$^{-/-}$ cell lines; and (SI 2) experimental details, including chemicals and materials, modification of the AFM tip with the PEG linker and AT, cell culture and fluorescence imaging, and AFM measurement (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Cunlan Guo − *College of Chemistry and Molecular Sciences, Wuhan University, Wuhan 430072, P. R. China; Single Molecule Study Laboratory, College of Engineering, University of Georgia, Athens, Georgia 30602, United States;* orcid.org/0000-0001-7706-5230; Email: cunlanguo@whu.edu.cn

Wenxuan Zhong − *Big Data Analytics Lab, Department of Statistics, University of Georgia, Athens, Georgia 30602, United States*; Email: wenxuan@uga.edu

Bingqian Xu − *Single Molecule Study Laboratory, College of Engineering, University of Georgia, Athens, Georgia 30602, United States;* orcid.org/0000-0002-7873-3162; Email: nanoxu@uga.edu

Authors

Huimin Cheng − *Big Data Analytics Lab, Department of Statistics, University of Georgia, Athens, Georgia 30602, United States*

Jun Yu − *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, P. R. China*

Zhen Wang − *Big Data Analytics Lab, Department of Statistics, University of Georgia, Athens, Georgia 30602, United States*

Ping Ma − *Big Data Analytics Lab, Department of Statistics, University of Georgia, Athens, Georgia 30602, United States*

Bin Wang − *Single Molecule Study Laboratory, College of Engineering, University of Georgia, Athens, Georgia 30602, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c03552

Author Contributions

B.X. and W.Z. conceived the research. C.G. and B.W. performed the experiment and collected the data. P.M. and W.Z. supervised the theoretical simulation. H.C., J.Y., and P.M. carried out the simulation. H.C., C.G., W.Z., and B.X. cowrote the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the US National Science Foundation, the U.S. National Institutes of Health, and the National Natural Science Foundation of China for funding this work (US NSF ECCS 1231967 and 2010875 (B.X.), DMS 1903226, 1925066, US NIH R01GM1222080 (P.M. and W.Z.), and NSFC 21974102, 21705019 (C.G.)).

■ REFERENCES

(1) Lee, G. U.; Chrisey, L. A.; Colton, R. J. Direct measurement of the forces between complementary strands of DNA. *Science* **1994**, *266*, 771−773.

(2) Moy, V. T.; Florin, E.-L.; Gaub, H. E. Intermolecular forces and energies between ligands and receptors. *Science* **1994**, *266*, 257−259.

(3) Rief, M.; Oesterhelt, F.; Heymann, B.; Gaub, H. E. Single molecule force spectroscopy on polysaccharides by atomic force microscopy. *Science* **1997**, *275*, 1295−1297.

(4) Oesterhelt, F.; Oesterhelt, D.; Pfeiffer, M.; Engel, A.; Gaub, H. E.; Müller, D. J. Unfolding pathways of individual bacteriorhodopsins. *Science* **2000**, *288*, 143−146.

(5) Hinterdorfer, P.; Dufrêne, Y. F. Detection and localization of single molecular recognition events using atomic force microscopy. *Nat. Methods* **2006**, *3*, 347−355.

(6) Neuman, K. C.; Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* **2008**, *5*, 491−505.

(7) Li, Y.; Shi, X.; Liu, H.; Yi, S.; Zhang, X.; Fang, X. Study of the effect of atorvastatin on the interaction between ICAM-1 and CD11b by live-cell single-molecule force spectroscopy. *Sci. China Chem.* **2010**, *53*, 752−758.

(8) Wei, W.; Sun, Y.; Zhu, M.; Liu, X.; Sun, P.; Wang, F.; Gui, Q.; Meng, W.; Cao, Y.; Zhao, J. Structural insights and the surprisingly low mechanical stability of the Au−S bond in the gold-specific protein GolB. *J. Am. Chem. Soc.* **2015**, *137*, 15358−15361.

(9) Baumann, F.; Bauer, M. S.; Milles, L. F.; Alexandrovich, A.; Gaub, H. E.; Pippig, D. A. Monovalent Strep-Tactin for strong and site-specific tethering in nanospectroscopy. *Nat. Nanotechnol.* **2016**, *11*, 89−94.

(10) Vera, A. M.; Carrión-Vázquez, M. Direct identification of protein-protein interactions by single-molecule force spectroscopy. *Angew. Chem., Int. Ed.* **2016**, *55*, 13970−13973.

(11) Lei, H.; He, C.; Hu, C.; Li, J.; Hu, X.; Hu, X.; Li, H. Single-molecule force spectroscopy trajectories of a single protein and its polyproteins are equivalent: a direct experimental validation based on a small protein NuG2. *Angew. Chem., Int. Ed.* **2017**, *56*, 6117−6121.

(12) Sun, Y.; Di, W.; Li, Y.; Huang, W.; Wang, X.; Qin, M.; Wang, W.; Cao, Y. Mg2+-Dependent high mechanical anisotropy of three-way-junction pRNA as revealed by single-molecule force spectroscopy. *Angew. Chem., Int. Ed.* **2017**, *56*, 9376−9380.

(13) Milles, L. F.; Schulten, K.; Gaub, H. E.; Bernardi, R. C. Molecular mechanism of extreme mechanostability in a pathogen adhesin. *Science* **2018**, *359*, 1527−1533.

(14) Bernardi, R. C.; Durner, E.; Schoeler, C.; Malinowska, K. H.; Carvalho, B. G.; Bayer, E. A.; Luthey-Schulten, Z.; Gaub, H. E.; Nash, M. A. Mechanisms of nanonewton mechanostability in a protein complex revealed by molecular dynamics simulations and single-molecule force spectroscopy. *J. Am. Chem. Soc.* **2019**, *141*, 14752−14763.

(15) Wang, H.; Gao, X.; Li, H. Single molecule force spectroscopy reveals the mechanical design governing the efficient translocation of the bacterial toxin protein RTX. *J. Am. Chem. Soc.* **2019**, *141*, 20498−20506.

(16) Cuellar-Camacho, J. L.; Bhatia, S.; Reiter-Scherer, V.; Lauster, D.; Liese, S.; Rabe, J. rP.; Herrmann, A.; Haag, R. Quantification of multivalent interactions between sialic acid and influenza A virus spike proteins by single-molecule force spectroscopy. *J. Am. Chem. Soc.* **2020**, *142*, 12181−12192.

(17) Merkel, R.; Nassoy, P.; Leung, A.; Ritchie, K.; Evans, E. Energy landscapes of receptor−ligand bonds explored with dynamic force spectroscopy. *Nature* **1999**, *397*, 50−53.
(18) Krieg, M.; Fläschner, G.; Alsteens, D.; Gaub, B. M.; Roos, W. H.; Wuite, G. J.; Gaub, H. E.; Gerber, C.; Dufrêne, Y. F.; Müller, D. J. Atomic force microscopy-based mechanobiology. *Nat. Rev. Phys.* **2019**, *1*, 41−57.
(19) Müller, D. J.; Dufrene, Y. F. Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. *Nat. Nanotechnol.* **2010**, *3*, 269−277.
(20) Alsteens, D.; Pfreundschuh, M.; Zhang, C.; Spoerri, P. M.; Coughlin, S. R.; Kobilka, B. K.; Müller, D. J. Imaging G protein−coupled receptors while quantifying their ligand-binding free-energy landscape. *Nat. Methods* **2015**, *12*, 845−851.
(21) Giudice, C. L.; Dumitru, A. C.; Alsteens, D. Probing ligand-receptor bonds in physiologically relevant conditions using AFM. *Anal. Bioanal. Chem.* **2019**, *411*, 6549−6559.
(22) Sandal, M.; Benedetti, F.; Brucale, M.; Gomez-Casado, A.; Samorì, B. Hooke: an open software platform for force spectroscopy. *Bioinformatics* **2009**, *25*, 1428−1430.
(23) Roduit, C.; Saha, B.; Alonso-Sarduy, L.; Volterra, A.; Dietler, G.; Kasas, S. OpenFovea: open-source AFM data processing software. *Nat. Methods* **2012**, *9*, 774−775.
(24) Hermanowicz, P.; Sarna, M.; Burda, K.; Gabryś, H. AtomicJ: an open source software for analysis of force curves. *Rev. Sci. Instrum.* **2014**, *85*, No. 063703.
(25) Partola, K. R.; Lykotrafitis, G. FRAME (Force Review Automation Environment): MATLAB-based AFM data processor. *J. Biomech.* **2016**, *49*, 1221−1224.
(26) Dinarelli, S.; Girasole, M.; Longo, G. FC_analysis: a tool for investigating atomic force microscopy maps of force curves. *BMC Bioinf.* **2018**, *19*, No. 258.
(27) Heenan, P. R.; Perkins, T. T. FEATHER: automated analysis of force spectroscopy unbinding and unfolding data via a Bayesian algorithm. *Biophys. J.* **2018**, *115*, 757−762.
(28) Duanis-Assaf, T.; Razvag, Y.; Reches, M. ForSDAT: an automated platform for analyzing force spectroscopy measurements. *Anal. Methods* **2019**, *11*, 4709−4718.
(29) Barabási, A. L. The network takeover. *Nat. Phys.* **2012**, *8*, 14−16.
(30) Bassett, D. S.; Sporns, O. Network neuroscience. *Nat. Neurosci.* **2017**, *20*, 353−364.
(31) Mocanu, D. C.; Mocanu, E.; Stone, P.; Nguyen, P. H.; Gibescu, M.; Liotta, A. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. Commun.* **2018**, *9*, No. 2383.
(32) Variano, E. A.; McCoy, J. H.; Lipson, H. Networks, dynamics, and modularity. *Phys. Rev. Lett.* **2004**, *92*, No. 188701.
(33) Guo, C. L.; Fan, X.; Qiu, H.; Xiao, W. Y.; Wang, L. C.; Xu, B. Q. High-resolution probing heparan sulfate-antithrombin interaction on a single endothelial cell surface: single-molecule AFM studies. *Phys. Chem. Chem. Phys.* **2015**, *17*, 13301−13306.
(34) Newman, M. E. Detecting community structure in networks. *Eur. Phys. J. B* **2004**, *38*, 321−330.
(35) Holland, P. W.; Leinhardt, S. Transitivity in structural models of small groups. *Comp. Group Stud.* **1971**, *2*, 107−124.
(36) Sin, C.-Y.; White, H. Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **1996**, *71*, 207−225.
(37) Rencher, A. C.; Schaalje, G. B. *Linear Models in Statistics*. John Wiley & Sons, 2008.
(38) Dudko, O. K.; Hummer, G.; Szabo, A. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 15755−15760.
(39) Johnson, K. C.; Thomas, W. E. How Do We Know when Single-Molecule Force Spectroscopy Really Tests Single Bonds? *Biophys. J.* **2018**, *114*, 2032−2039.
(40) Newton, R.; Delguste, M.; Koehler, M.; Dumitru, A. C.; Laskowski, P. R.; Mueller, D. J.; Alsteens, D. Combining confocal and atomic force microscopy to quantify single-virus binding to mammalian cell surfaces. *Nat. Protoc.* **2017**, *12*, 2275−2292.
(41) Loader, C. Smoothing: Local Regression Techniques. In *Handbook of Computational Statistics*; Springer, 2012; pp 15−58.
(42) Cleveland, W. S.; Devlin, S. J.; Grosse, E. Regression by local fitting: methods, properties, and computational algorithms. *J. Econometrics* **1988**, *37*, 87−114.
(43) Hansun, S. In *A New Approach of Moving Average Method in Time Series Analysis*, 2013 Conference on New Media Studies (CoNMedia); IEEE, 2013; pp 1−4.
(44) Cheng, H.-M.; Ning, Y.-Z.; Yin, Z.; Yan, C.; Liu, X.; Zhang, Z.-Y. Community detection in complex networks using link prediction. *Mod. Phys. Lett. B* **2018**, *32*, No. 1850004.
(45) Liu, X.; Cheng, H.-M.; Zhang, Z.-Y. Evaluation of community detection methods. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1736−1746.
(46) Newman, M. E. J. Detecting community structure in networks. *Eur. Phys. J. B* **2004**, *38*, 321−330.

https://doi.org/10.1021/acs.jpcb.1c03552
J. Phys. Chem. B 2021, 125, 9660−9667