Electronic Supplementary Material (ESI) for Analytical Methods.

Supplementary Information

ForSDAT: An Automated Platform for Analyzing Force Spectroscopy Measurements

*Tal Duanis-Assaf, Yair Razvag and Meital Reches**

Institute of Chemistry and The Center for Nanoscience and Nanotechnology, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

*E-mail: meital.reches@mail.huji.ac.il

Histogram Bin Size Selection for Force Distribution Baseline Detection

The sizes of bins in the histogram can affect the accuracy of baseline shift and noise evaluation. If the bins are too small, the histogram will not display distinct peaks. However, larger bins cause over smoothed histograms (Fig. S2), leading to an over estimate of the noise amplitude. Moreover, over smoothed histograms may cause the smaller peaks to be overshadowed by larger ones, which would hamper the ability to detect multiple plateaus in the signal.

To determine the best binning method for force-distribution baseline detection, we simulated 100 force curves and analyzed the baseline using four bin size selection methods: constant bin size (10pN), square root of the number of data points ($\sqrt{N}$), Sturges rule and Freedman-Diaconis rule (F-D)[1].

To evaluate all 4 binning methods, we compared the baseline shift and noise amplitude detected by each method to the simulated value (Fig. S1). While all four methods successfully detect the baseline in most cases, the Sturges and Freedman-Diaconis rules generated a slightly wider distribution and more outliers, whereas the square root and 10pN binning methods produced similar results. When comparing the noise estimate, the 10pN and square root methods produced relatively similar results. Both the Freedman-Diaconis and Sturges rules produced less accurate noise estimates, featuring a wider distribution.

These results suggest that under the tested parameters (scanner speed, sampling rate and SNR) both the Sturges and Freedman-Diaconis rules are inappropriate for the force-distribution baseline detection method. The square root method and constant bin size generated similar results. It should be noted that

the bin size should be smaller than the noise level, to obtain accurate noise estimates. The 10pN method is optimized for our experimental setup and systemic noise, and it is advised to optimize the bin size before analysis. Nevertheless, if systemic noise is unknown, judging from simulation results, the square root binning method seems adequate.

References

1. Freedman, D. and Diaconis, P., Zeit. Wahr. ver. Geb., 1981, **57**, 453–476.

Method		Measured Runtime
Tail analysis		Less than 1 ms
Force distribution	1 population Gaussian series	100-200 ms
	3 populations Gaussian series	200-400 ms
	5 populations Gaussian series	500-1000 ms

Table 1 - Measured runtimes of 4 different baseline detection methods. Tail analysis and force distribution fitted with a Gaussian series of 1st, 3rd or 5th order. All runtimes were measured on a home laptop equipped with a processor with a maximal clock speed of 2.4 GHz.

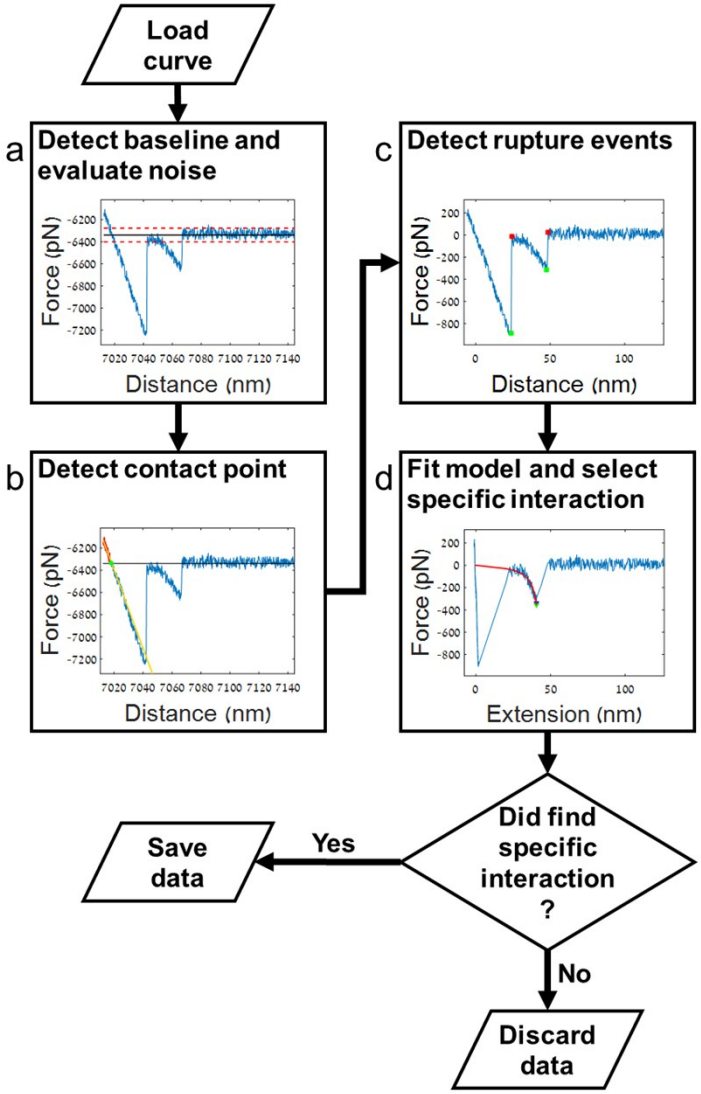

Fig. 1: The process of analyzing a single force distance curve. The analysis algorithm comprises four main steps: baseline detection (a), contact point detection (b), rupture event detection (c) and model fitting and specific interaction selection (d). At the end of the process, if a specific interaction is detected the data is saved. The black line represents the baseline and dashed lines represent the domain governed by random thermal noise around the baseline (a). The contact point (●) is evaluated from the intersection of the baseline (black) and the linear regression line (gold) of the contact domain (red) (b). Discontinuities in the curve represent rupture events, the squares ■ and ■ represent the detected points of rupture start and resting point of the cantilever respectively (c). ▼ represents the detected specific interaction rupture event and the red curve represents the worm-like-chain model fitting (d).

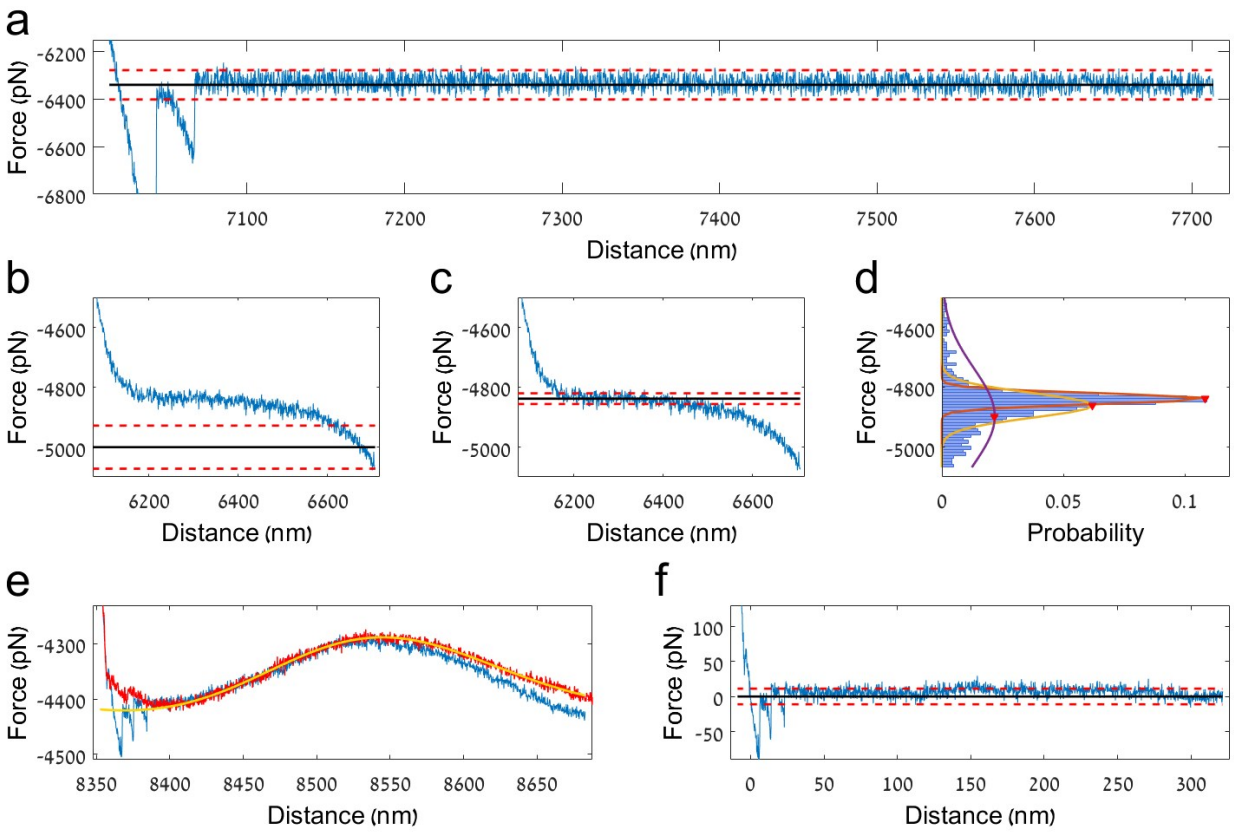

Fig. 2: Baseline correction and noise evaluation. The traditional method for baseline detection using tail analysis. The baseline shift and noise level are evaluated by calculating the mean value and standard deviation of the force within a window on the tail of the curve (a). Tail analysis fails to evaluate the baseline shift and noise when there is a disturbance to the baseline in the far distances (b). By plotting a histogram of the force values and fitting a 3rd order gaussian series (d), the baseline shift and noise level were evaluated from the mean value and standard deviation of the gaussian peak with the highest mean force (c). For curves with a non-linear baseline (e) baseline correction was done by fitting a 2nd order fourier series (gold) to the approach segment (red) of the force curve, and subtracting that function from the retract segment (teal) to flatten the curve. The curve's baseline shift and noise were then evaluated using tail analysis (f)

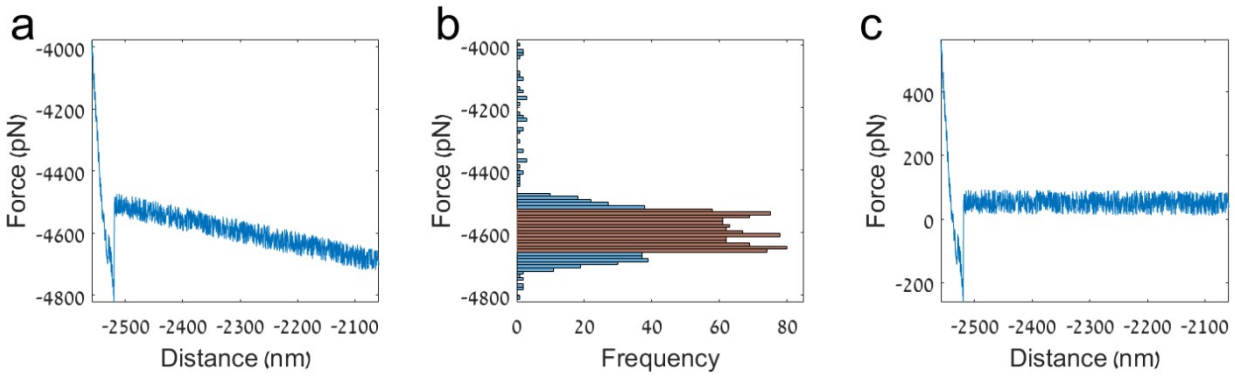

Fig. 3: Tilted baseline correction using the force distribution method. Simulated force distance curve exhibiting thermal drift in the baseline (a). The force histogram displays a plateau (b). Baseline drift slope was calculated from the mean frequency in the plateau, the scanner speed and data sampling rate. The fitted line was then subtracted from the curve to correct the baseline shift and drift (c).

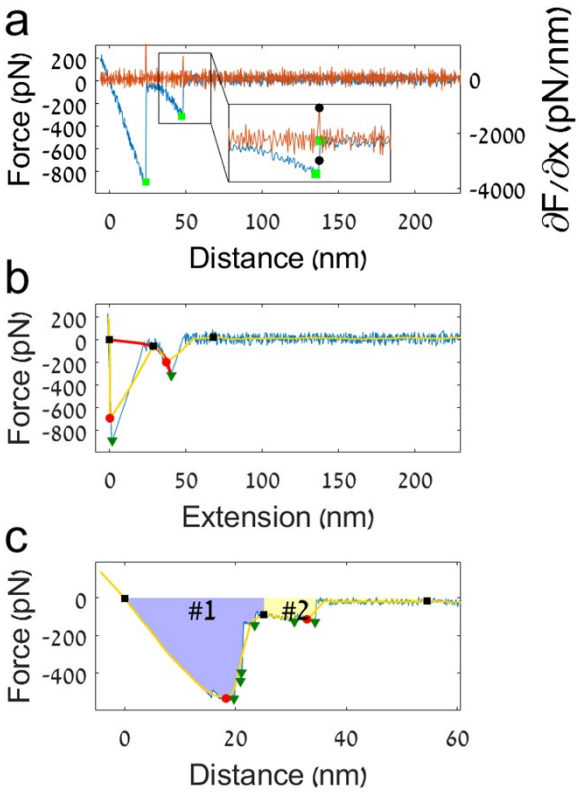

Fig. 4: Rupture and specific interaction detection. Rupture detection was performed by calculating the slope vector – ∂F/∂x (red). Rupture events are represented by discontinuities in the curve with a positive slope. The inner pane in panel a shows a magnification of the curve around the second rupture event. The peak of the slope vector (●) corresponds to the highest jump in the force (lower ●), which represents the discontinuity in the force. Each discontinuity corresponds to a positive slope peak. The start and end positions of the rupture (■) are found iteratively, by following the force vector forward and backward from the discontinuity position as long as the slope remains above a certain threshold (a). The curve was smoothed using a large window (gold) smoothed peaks and valleys are marked as ● and ■ respectively. Each smoothed peak represents a bundle of interactions (b & c). If the specific interaction candidate is the only rupture event (▼) within that bundle it is considered well separated from the non-specific interactions. The WLC fit of the specific interaction is plotted as a red line (b). Panel c shows the area of each smoothed peak as colored area under the curve. Each of these peaks contains several rupture events, which are therefore considered non-specific interactions.

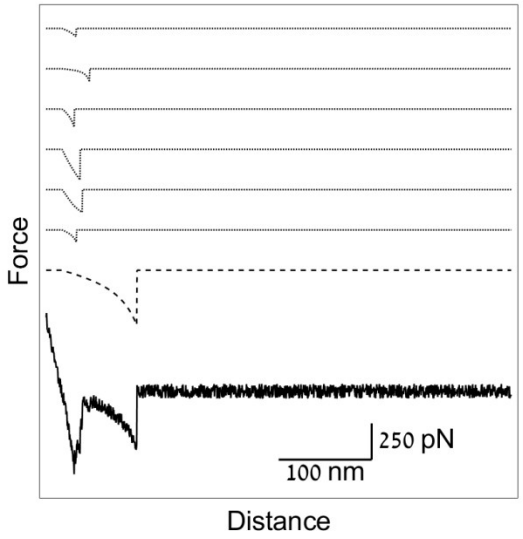

Fig. 5: Typical simulated force distance curve. The simulated curve (black line) is comprised of a random number of non-specific interactions (dotted lines), and randomly occurring specific interaction (dashed line) topped with uniformly distributing random noise.

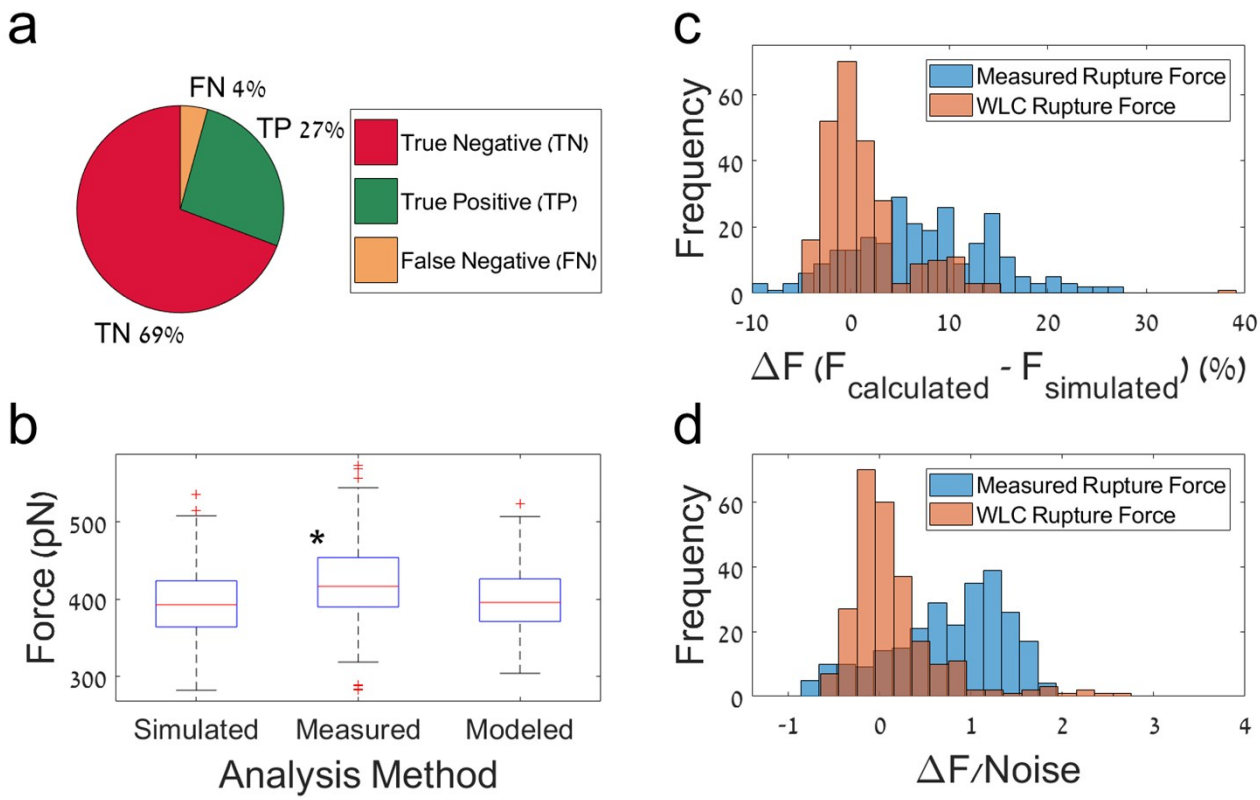

Fig. 6: Process success rate evaluation when analyzing a simulated data set. The automatic analysis of the simulated curves had a success rate of 96% (a). Our results showed a statistically significant difference between the measured force (the force delta between the start and end points of a rupture event) and the simulated value, whereas the force calculated using the WLC model showed no significant difference from the simulated force. The force calculated using the WLC model compared to the simulated force showed a difference which rarely exceeded 5% (c) and a single amplitude of the noise (d), while the measured force showed a wider distribution around 5% (c) and mostly between one and two amplitudes of the noise (d). The asterisk in b represents significantly different mean value in comparison to the simulated force as determined by one-way ANOVA followed by post hoc Tuckey test.

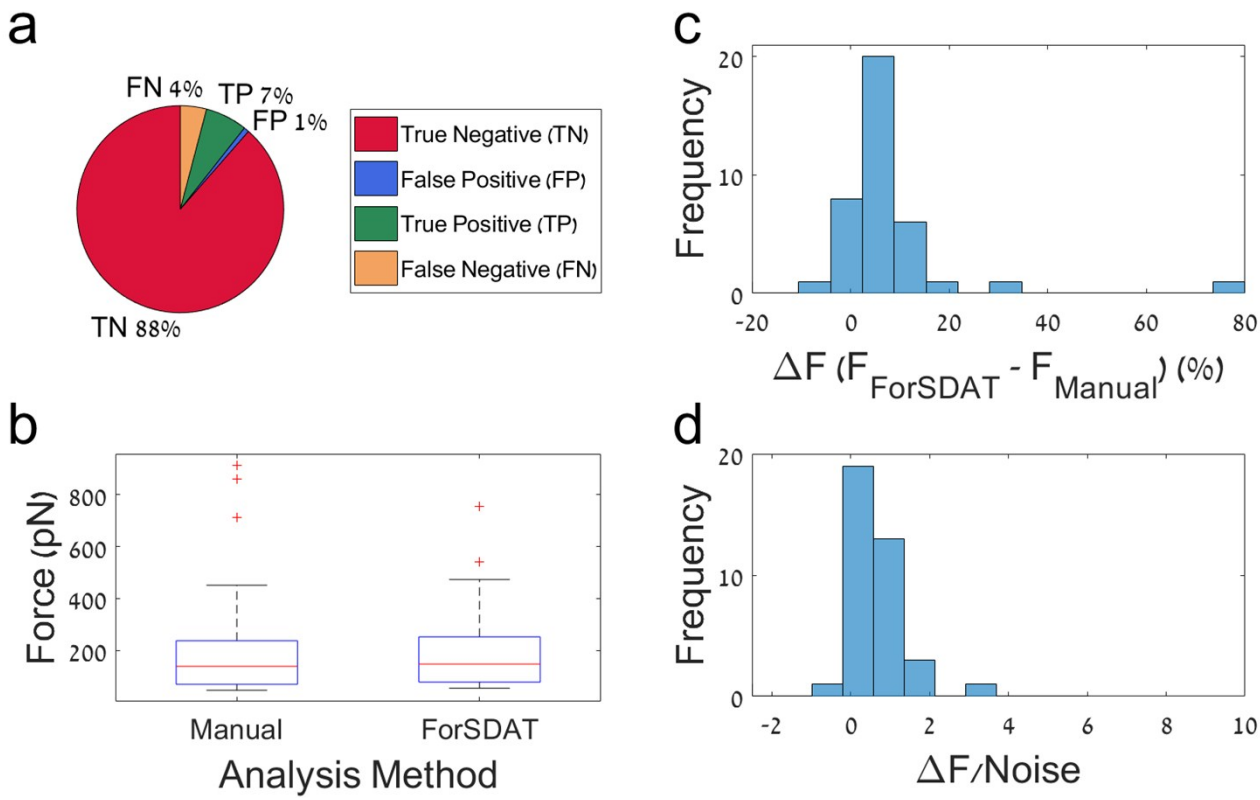

Fig. 7: Process success rate evaluation when analyzing an experimentally obtained data set analyzed manually using commercial software. The results of the automatic processing in comparison to the manual analysis show success in 95% of the processed curves (a). Our results showed no statistically significant difference between the force calculated using the WLC model by ForSDAT and the manually analyzed value calculated (b) as determined by one-way ANOVA. The automatically calculated force was compared to the force calculated by commercial software, displaying a difference around 10% (c), the difference was no greater than the range of the noise domain in almost all the curves (d).

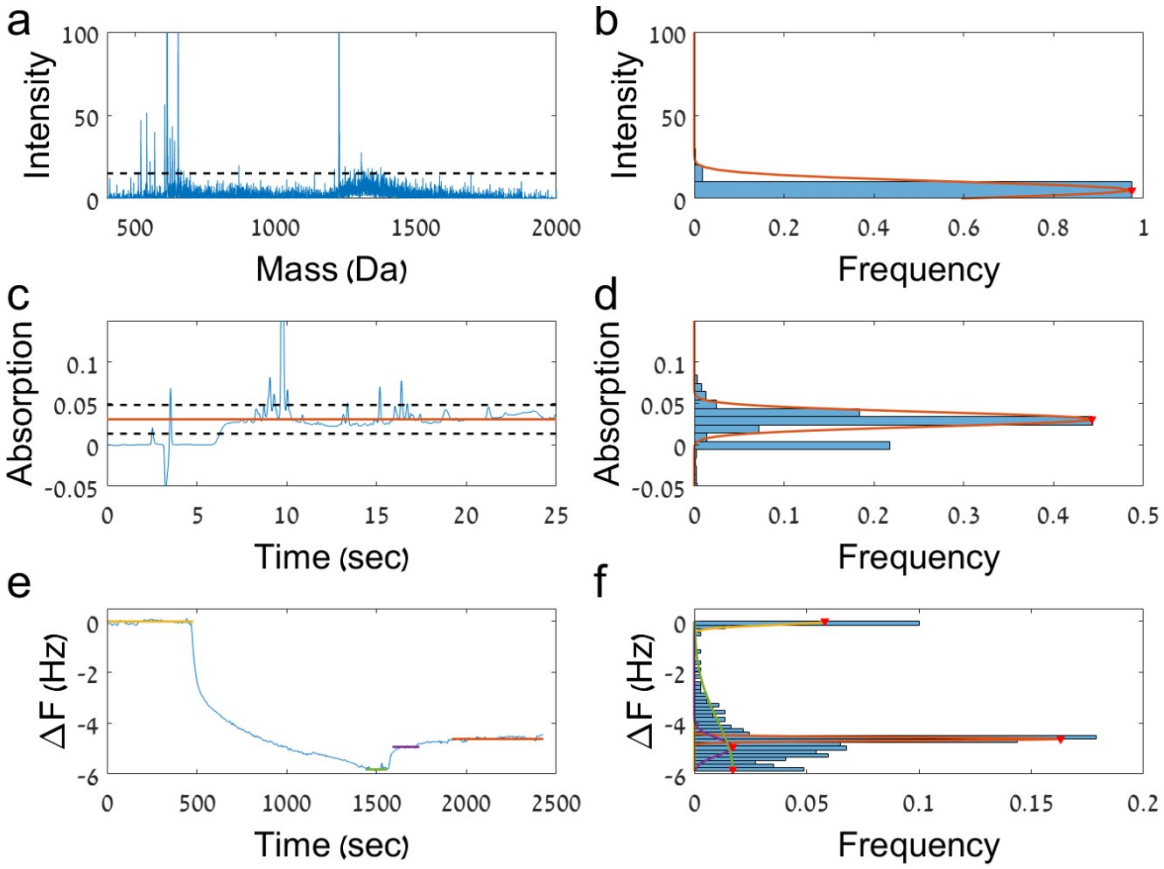

Fig. 8: Applying the Y-distribution baseline detection methods to data obtained from other methods. A mass spectrum (a) noise can be evaluated by fitting the intensity histogram with a gaussian (b). High pressure liquid chromatogram (c) baseline and noise level can be evaluated by fitting the absorption histogram with a gaussian (d). The baseline and plateaus of a quartz crystal microbalance (QCM) ΔF over time curve (e) can be evaluated by fitting the ΔF histogram with a gaussian series (f). QCM data was taken from Maity & Nir et al. (2014) [52].

Task	Component Name	Operation
Data Adjustment	FDCurveOOMAdjuster	Adjusts the units of the signal data
	TipHeightAdjuster	Transforms the distance signal to extension (tip-sample separation) by subtracting cantilever bending. Transforms the curve into a force vs. extension curve.
	DataSmoothingAdjuster	Smooths the force signal using any one of the following methods: sgolay - Stavinsky-Golay filter moving - Moving Average lowess - locally weighted scatterplot smoothing loess - locally estimated scatterplot smoothing rlowess - Robust LOWESS rloess - Robust LOESS movmedian - Moving Median gaussian - Gaussian
Oscillating baseline correction	LongWaveDisturbanceAdjusterBeta	Corrects the oscillation abstract based on Fourier series. Fitting to either approach or retract segments is supported.
Baseline	SimpleBaselineDetector	Tail analysis baseline alignment
	HistogramBaselineDetector	Force distribution baseline alignment
Contact point	ContactPointDetector	Contact point alignment
Rupture events	RuptureDetector	Rupture event detection based on df/dz peaks
Molecular chain fitting	WLCLoadFitter	Fits the loading domain of a rupture event with the worm-like chain (WLC) model
	FJCLoadFitter	Fits the loading domain of a rupture event with the freely-joint chain (FJC) model
	PolynomialLoadFitter	Fits the loading domain of a rupture event with a polynomial function
Interaction window	InteractionWindowSMIFilter	Selects only rupture events within the interaction window
Single Molecular specific interaction selection	BaselineThresholdSMIFilter	Differentiates specific interaction from non-specific interactions based on thresholding
	SmoothingSMIFilter	Differentiates specific interaction from non-specific interactions based on rupture events correlation to peaks in the smoothed signal

Table S1 – List of available components in ForSDAT.

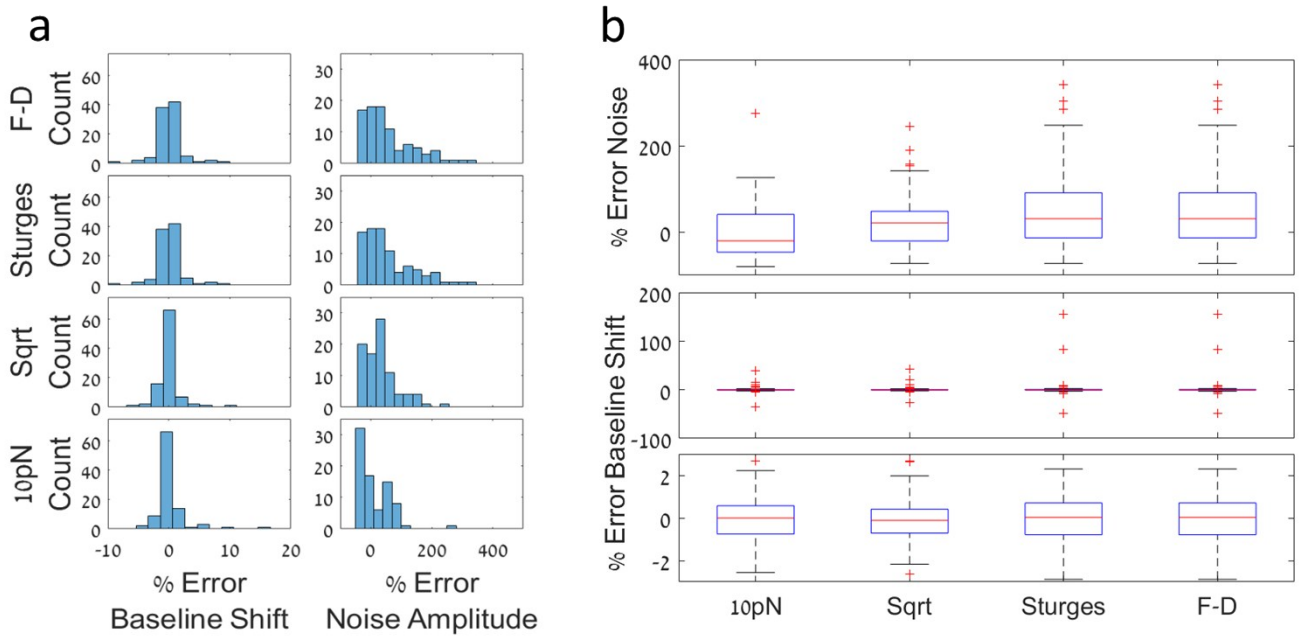

Fig. S1: Bin size selection methods for force-distribution histograms. (a) The left panel shows histogram view of the distribution of the error in baseline shift compared to the simulated baseline shift in percentage. The right side histograms show the distribution of the error in noise amplitude compared to the simulated noise amplitude in percentage. (b) The middle panel shows box plots of the error in baseline shift compared to the simulated baseline shift in percentage. The bottom panel shows the zoomed in data. The top panel shows box plots of the error in noise amplitude compared to the simulated noise amplitude in percentage.

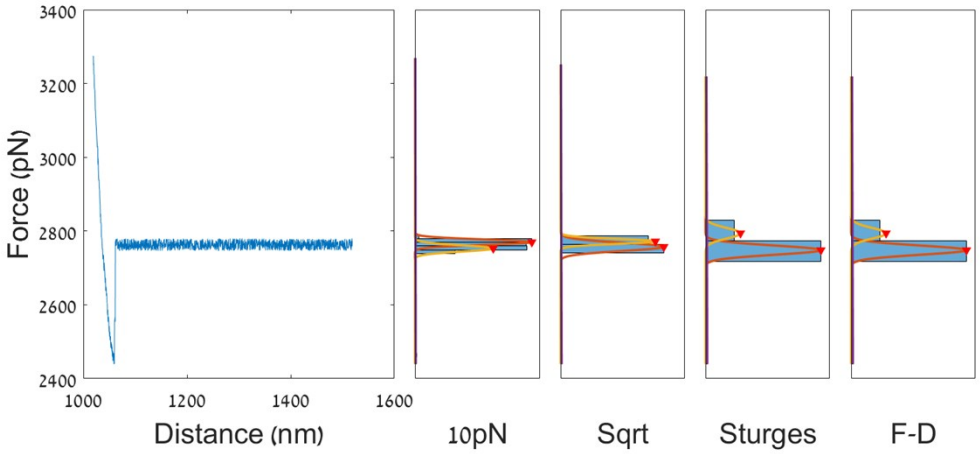

Fig. S2 – Force distribution histograms bin size selection methods. On the left a typical simulated FDC. On the right force distribution histograms generated using four different bin size selection methods.