Electronic Supplementary Material

Adaptive nanopores: A bioinspired label-free approach for protein sequencing and identification

Andrea Spitaleri[1,2,§], Denis Garoli[3,4,§], Moritz Schütte[5], Hans Lehrach[5,6], Walter Rocchia[1] (✉), and Francesco De Angelis[3] (✉)

[1] *CONCEPT Lab, Istituto Italiano di Tecnologia, Via Morego 30, Genova, I-16163, Italy*
[2] *Center for Omics Sciences, IRCCS San Raffaele Scientific Institute, Milan, Via Olgettina 58, Milano, I-20132, Italy*
[3] *Plasmon Nanotechnology Unit, Istituto Italiano di Tecnologia, Via Morego 30, Genova, I-16163, Italy*
[4] *AB ANALITICA s.r.l., Via Svizzera 16, I-35127 Padova, Italy*
[5] *Alacris Theranostics GmbH, Max-Planck-Strasse 3, D-12489 Berlin, Germany*
[6] *Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin, Germany*
[§] *Andrea Spitaleri and Denis Garoli contributed equally to this work.*

Supporting information to https://doi.org/10.1007/s12274-020-3095-z

Section S1 Molecular Dynamics simulation setup.

MD simulations were performed with the GROMACS software (http://www.gromacs.org/), version 4.6.1, using amber14SB as force field, together with the explicit TIP3P water model and standard parm99 parameterization for 60 Na^+ ions, which were added to neutralize the net charge of the system. The equilibration protocol consisted of a series of energy minimizations and short restrained MD runs in the NVT ensemble. This was followed by production runs in the NPT ensemble at 298 K and pressure of 1 atm. Periodic boundary conditions, together with the Particle Mesh Ewald method to treat long-range interactions and a 2 fs time step were used.

Initial positioning of the peptides

The center of mass of each linear peptide has been positioned in the center of mass of the pore. Then, the principal axis of the linear peptide was aligned along the principal axis of the pore. The set-up has been done using in house scripts based on the VMD program and the orient and la101psx packages as described in the web page: http://www.ks.uiuc.edu/Research/vmd/script_library/scripts/orient/

The set-up consisted in the following 4 steps:
1) overlap the two centers of mass of the peptide and the pore using move2com function;
2) align the principal axes using the align function;
3) rotate the peptide by 45° in order to get the different replicas using the rotatemol function;
4) save the coordinates of the peptide using the savemol function;

```
proc move2com { id0 id1 } {
      set sel0 [atomselect $id0 "all"]
      set sel1 [atomselect $id1 "all"]
      set com0 [measure center $sel0    weight mass]
      set com1 [measure center $sel1    weight mass]
      set dist [vecsub $com0 $com1]
      set newcoords {}
      set offset $dist
        foreach coord [$sel1 get {x y z}] {
         lappend newcoords [vecadd $coord $offset]
      }
      set $sel1 $newcoords
      set numatoms [$sel1 num]
      for { set i 0 } { $i<$numatoms } { incr i } {
        set t [atomselect $id1 "index $i"]
        set tt [lindex $newcoords $i]
        $t moveto $tt
      }
}
```

Address correspondence to Walter Rocchia, walter.rocchia@iit.it; Francesco De Angelis, francesco.deangelis@iit.it

```
proc align { id0 id1 } {
        package require Orient
        namespace import Orient::orient
      # donut
        set sel0 [atomselect $id0 "all"]
    # peptide
        set sel1 [atomselect $id1 "all"]
        set I [draw principalaxes $sel0]
        #$sel1 moveby [mvecinvert [measure center $sel0 weight mass]]
        set J [draw principalaxes $sel1]
        set B [orient $sel1 [lindex $J 2] [lindex $I 0]]
        $sel1 move $B
        set J [draw principalaxes $sel1]
}
proc rotatemol { id axis angle} {
        set sel1 [atomselect $id "all"]
        set matrix [transaxis $axis $angle]
        $sel1 move $matrix
}
proc savemol { id0 num } {
        set sel0 [atomselect $id0 "all"]
        set filename [molinfo $id0 get name]
        set file0 [exec echo $filename | awk -F. {{print $1}}]
        set fl $file0\_align$num.pdb
        puts $fl
        $sel0 writepdb $fl
}
```

Section S2 Details on the protein construct design

The N-terminal of each of the 10 monomers composing the SmPrx protein was prolonged with a 16aa-long sequence inspired to three-dimensional Leucine zipper structural motif. The length of the elongation was decided based on the internal diameter of the protein channel, in order to allow both for translocation and for the interaction of the added tips with the translocating polypeptide. The full GCN4 Leucine Zipper sequence is: RMKQLEDKVEELLSKNYHLENEVARLKKLVGER. Preliminary to choose which stretch to extract and use, the full Leucine Zipper structure was simulated via MD for 200ns in explicit solvent and its secondary structure stability, shown in Figure S1, was analyzed.

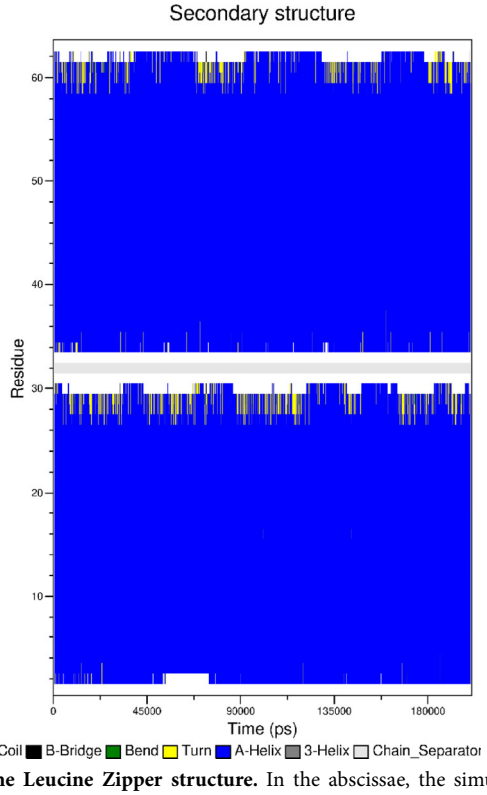

Figure S1 Secondary structure analysis for the Leucine Zipper structure. In the abscissae, the simulated time is represented in ps, while in the ordinates the residue numbers, gathering the two arms of the zipper, which are 34aa long each (PDB ID 2ZTA). As one can see, the helical structure is prevalent.

The selected stretch has the following sequence: LEDKVEELLSKNYHLE. To this sequence, two mutations have been further performed: the first residue was mutated to a Cysteine, to allow the formation of a stabilizing bond at the N-terminal of the tip. Furthermore, the Serine residue in the 10th position, which is facing outwards in the structure, was also mutated to Cysteine in order to realize an anchor point for the linkers of the dyes. So the final prepended sequence is: CEDKVEELLCKNYHLE. A 100ns-long MD simulation of the full structure was performed, specifically focusing on the analysis of the stability of the added tips (see Figure S2).

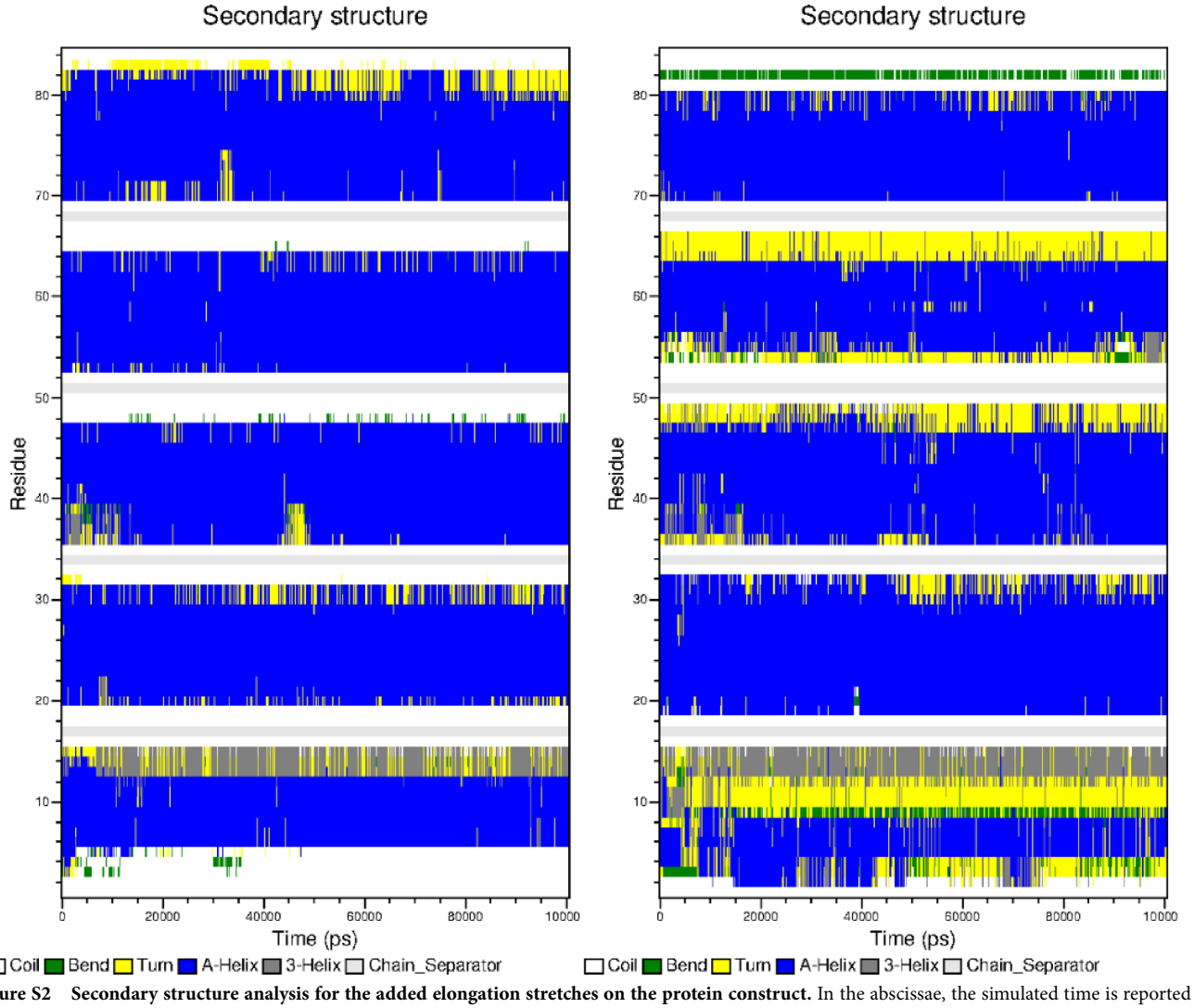

Figure S2 Secondary structure analysis for the added elongation stretches on the protein construct. In the abscissae, the simulated time is reported in ps, while in the ordinates the residue numbers, gathering the two arms of each tip in the system. As one can see, the helical structure (in blue) is still prevalent, although sometimes is a bit distorted (in yellow and grey). The constraint at the end of the tip is providing further stability to the system.

Section S3 Machine learning classification

Input data time, distance, k^2 and efficiency was provided per amino acid (of E,G,H,I,R,S,W,X,Y with X the control without an amino acid in the protein ring), for the seven FRET pairs (homo-FRET, 6 combinations of 425 - 647N) and the angle of orientation (of 0, 45, 90, 135, 180, 225, 270, 315 degrees). To prepare the data for the classification, we first calculated for each FRET pair the average across the angle of orientation. Classification was implemented using the sklearn python package (https://scikit-learn.org/stable/). Both classifiers were implemented using custom scripts using python 2.7.15.

Support vector machine classifier steps: 1) reading input data, 1a) for the tips, Fig.3, additional scaling was applied (X = preprocessing.scale(X)), while the FRET efficiencies are already scaled 0-1, 2) split into 75% training and 25% test set, 3) grid search to tune parameters with 'precision_micro' score (*clf = GridSearchCV(SVC(), tuned_parameters, cv=5,scoring='%s' % score)*), 4) fitting the model on the training set (*clf.fit(X_train, y_train)*), 5) predicting on the test set (*y_true, y_pred = y_test, clf.predict(X_test)*), 6) print results (*classification_report(y_true, y_pred)*).

Random forest classifier steps: 1) reading input data, 2) split into 75% training and 25% test set, 3) grid search to tune parameters with 'precision_micro' score (*rfc = RandomForestClassifier(); CV_rfc = GridSearchCV(rfc, param_grid, cv=5, scoring=score)*), 4) fitting the model (*CV_rfc.fit(X_train, y_train)*), 5) predicting on the test set (*y_true, y_pred = y_test, CV_rfc.predict(X_test)*), 6) print results (*classification_report(y_true, y_pred)*).

tSNE was plotted using R-3.6.0 and the R_package tsne.

Section S4 Dye and linker characterization, parameterization and setup for MD

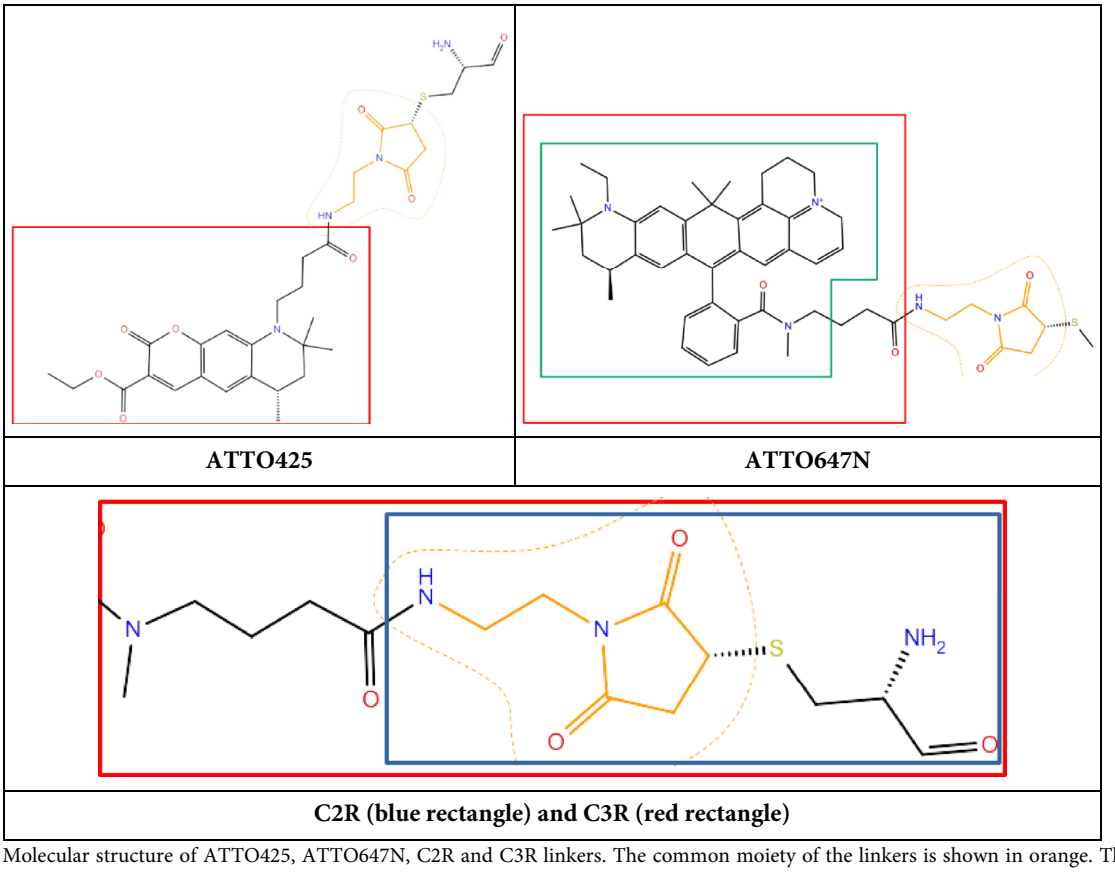

Figure S3 Molecular structure of ATTO425, ATTO647N, C2R and C3R linkers. The common moiety of the linkers is shown in orange. The dyes are shown in the red boxes. The scaffold from ATTO610 and ATTO488 exploited to build the ATTO647N parameters is shown in red box respectively. The green selection in ATTO647N is the scaffold used to calculate the RESP charges (see below).

The force field parameters for ATTO647N and ATTO425 used in the Molecular Dynamics simulations were from the AMBER-DYES database as reported in [1], in which the modular parameter of the dye/linker/amino acid combinations were calculated following the AMBER force field parametrization procedure. The files can be downloaded from www.github.com/t-/amber-dyes as reported in the same work. In particular, ATTO425 was already present in the AMBER-DYES database along with the linker C2R, whereas ATTO647N was added to the database following the AMBER force field parametrization procedure. The ATTO647N optimized structure was used to assign atom types and bonded parameters from the Generalized AMBER Force Field 33 (GAFF). The parameters were assigned by comparing known atom configurations from the GAFF force field to the configurations found in the optimized structure. The ATTO647N dye RESP charges were calculated from quantum QM calculations at B3LYP/6-31G* level of theory and the nonbonded parameters derived using ANTECHAMBER approach. The linker used for ATTO647N was C3R from the AMBER-DYES database.

Identification of dye and linker equilibrium locations and conformations

Due to the long relaxation times that would be required to find equilibrium locations and conformations for the dyes linked to the pore construct, which are expected to be in the order of milliseconds, we decided to use alternative simulative approaches, borrowed from the field of CADD (Computer Aided Drug Design) where similar problems are encountered and addressed with a good degree of success.

Molecular docking of the dyes on the modified SmPrxL

Molecular docking is a fast and approximate technique that performs a quite exhaustive exploration of the roto-translational space available to the small molecule, here the dye, trying to identify the most likely binding poses. Different candidate poses are traditionally ranked via heuristic scoring functions. Here, we docked the dyes onto the modified SmPrxl system in order to identify a number of possible preferred locations circumventing the problem of slow rearrangement dynamics. In this preliminary phase, the linker connecting the dye to the construct was only virtually taken into account, being replaced by limiting the space exploration only to positions that are compatible with its length. Preliminary to the molecular docking phase, we performed 1 μs of simulation of the modified SmPrxl (i.e. without dyes), we did K-means cluster analysis on the final part of the simulation and used the method of the most populate cluster as input for the successive docking calculations.

Molecular docking was performed with the GLIDE software tool (SCHRÖDINGER Suite)[2]. ATTO425 and ATTO647N (see Figure S3) dyes were docked onto the modified SmPrxl system in five independent runs and using a positional restraint between the cysteine side chain sulfur and the dye's carbonyl (distance restraint range between 5Å and of 8Å) in order to mimic the presence of the dye linker. In each stage, the poses were clustered using Daura algorithm, making a fitting on the donut backbone and calculating the RMSD on the dye heavy atoms.

Docking poses MD-based ranking. Conventional scoring functions have been replaced with a MD-based protocol that proved to be successful both in ranking different congeneric compounds based on their residence time [3] and in identifying the nearest-to-native small molecule pose in protein-drug binding.[4] In this protocol, the stability of each cluster medoid in terms of low RMSD with respect to the initial pose is performed based on the Scaled-MD approach (3 replicas), that we already used to evaluate the residence time of bound poses [3]. In Table S1 the results of the docking and ranking calculations are summarized.

Identification of the conformation of the linkers.

We performed Steered MD [5], slowly bringing the linker extreme atoms to a final target position which is compatible with a covalent bond with the anchoring atoms, namely the Carbonyl on the dye and the Nitrogen atom on the anchor modified cysteine, inserted by mutating the leucine-zipper like stretch at positions indicated in Table S1.

Table S1 Positioning of the dyes on the protein construct.

Anchor point	Ligand (dye)	Lower and upper distance restraints [Å]	N. Poses obtained/N. Poses output	N. Clusters with 2Å RMSD cutoff
C10-S	ATTO425	5-8	3/100	3
C402-S	ATTO425	5-8	59/100	10
C794-S	ATTO647N	5-8	8/100	2
C1186-S	ATTO647N	5-8	5/100	2
C1578-S	ATTO647N	5-8	42/100	15

Table S2 5ns long Steered MD simulations of the construct including the dyes (3 replicas). The "Docking" and "With Linker" columns represent the average RMSD values with respect to the initial docking pose, with and without the linker, respectively. The RMSD values are calculated on the heavy atoms of the dyes without the linker for the whole trajectory (Full5ns) and for only the last 1 ns of simulation (Last1ns), fitting on the backbone of the protein. The presence of the linker does not influence the pose defined by the docking.

DYE	Full5ns		Last1ns	
	Docking [Å]	With Linker[Å]	Docking[Å]	With Linker[Å]
C10 - ATTO425	7 ± 2	7 ± 1	8 ± 2	7 ± 2
C402 - ATTO425	6 ± 1	3± 0	7 ± 1	3 ± 0
C794 - ATTO647N	7 ± 1	6 ± 0	8 ± 1	7 ± 1
C1186 - ATTO647N	5 ± 1	3 ± 1	5 ± 1	4 ± 1
C1578 - ATTO647N	3 ± 0	2 ± 2	3 ± 0	2 ± 1

Section S5 analysis of MD simulations.

FRET efficiency analysis from MD simulations of pore construct in presence of the analyte

Molecular dynamics simulations were performed to evaluate the stability and dynamics of the SmPrxl-dyes system in the presence of the polypeptidic analyte. The FRET-related parameters, including distance, k^2, and the instantaneous FRET efficiency E(t), are extracted from the trajectory files by means of the *md2fret* post-processing tool.[6] This tool monitors the dye transition dipole moment, defined by the double atom pair as shown in figure S4. The distance is calculated on the basis of the centers of mass of the given atom pairs. In Figure S5 two examples of the output of this analysis are shown.

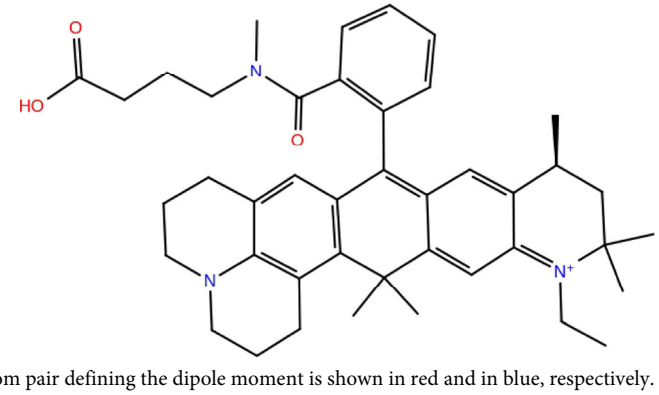

Figure S4 ATTO647N dye. The atom pair defining the dipole moment is shown in red and in blue, respectively.

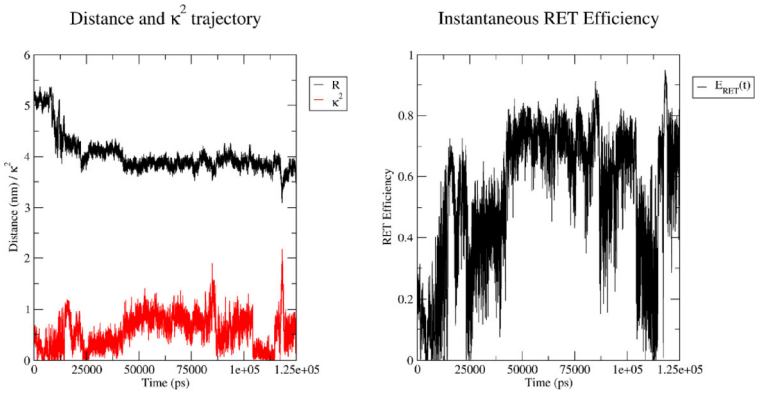

Figure S5 Left panel: Distance and k^2 of a donor-acceptor (ATTO425 and ATTO647N) pair dye as function of time. The distance and k^2 are shown in black and red lines respectively. Right panel: FRET Efficiency of the same donor-acceptor pair as function of time.

Evaluation of equilibrium configurations

Dedicated analysis was performed to assess whether the molecular dynamics simulations reached a steady state configuration, needed for a statistically valid analysis. To improve the sampling of the configuration space, we simulated 8 independent, 125ns-long, replicas per peptide. The replicas differ in the initial orientation of the central side chain. The average RMSD values of the last 75ns out of 125ns with respect to the starting structure (Table S3 and Fig. S6). Low standard deviation implies low fluctuations around the mean, supporting the hypothesis that it is an equilibrium conformation. We moreover calculated the distance between the center of mass of each dye (in Table S4 and S5, average and standard deviation are respectively reported) at the end of the X simulation (where no peptide is present in the nanopore's hole) to evaluate the R_0 as function of the time.

Table S3 RMSD values in nm with respect to the starting position. The values represent the average (Av column) and standard deviation (StdDev column) over the 8 MD simulations, considering the last 75ns out of 125ns of simulation length. The low StdDev values indicate that the dyes reached a steady state.

	T42-resid10		T42-resid402		T64-resid794		T64-resid1186		T64-resid1578	
	Av	StdDev	Av	StdDev	Av	StdDev	Av	StdDev	Av	StdDev
TRP	0.81	0.02	0.45	0.03	0.99	0.04	0.73	0.05	0.29	0.02
TYR	0.78	0.03	0.57	0.03	0.86	0.03	0.68	0.05	0.45	0.02
ARG	0.75	0.03	0.66	0.04	0.84	0.03	0.54	0.03	0.29	0.01
HIS	0.79	0.06	0.66	0.05	1.06	0.03	0.73	0.04	0.45	0.03
GLU	0.83	0.04	0.71	0.03	1.07	0.06	0.63	0.02	0.29	0.02
GLN	0.70	0.04	0.43	0.04	0.95	0.02	0.76	0.04	0.34	0.02
SER	0.66	0.02	0.61	0.03	0.99	0.04	0.61	0.06	0.31	0.01
ILE	0.79	0.03	0.51	0.08	0.85	0.03	0.48	0.02	0.28	0.02
GLY	0.66	0.02	0.63	0.05	1.18	0.04	0.69	0.02	0.31	0.01
X	0.95	0.35	0.52	0.15	1.12	0.38	0.74	0.21	0.25	0.06

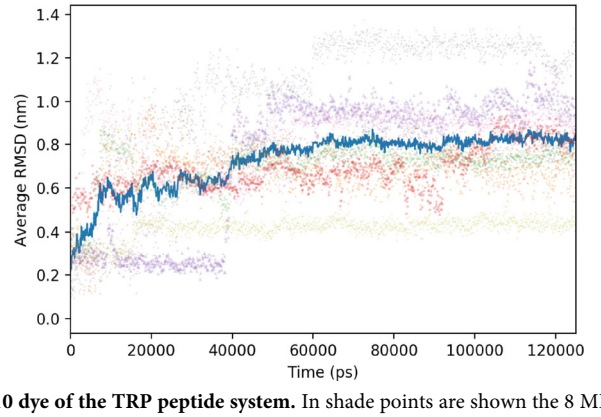

Figure S6 RMSD vs time of T42-resid10 dye of the TRP peptide system. In shade points are shown the 8 MD simulations, in the dark line the RMSD average.

Table S4 Distance matrix between dyes in the blank (no peptide) MD simulations. The values represent the average values of the last frame over the 8 MD replicas. Values are expressed in nm. The Förster-radius R_0 per dyes couple is R_0[ATTO425-ATTO647N] = 4.3 nm, R_0 [ATTO425-ATTO425] = 3.6 nm, R_0 [ATTO647N-ATTO647N] = 6.5 nm.

		T42	T42	T64	T64	T64
		resid10	resid402	resid794	resid1186	resid1578
T42	resid10	0				
T42	resid402	2.25	0			
T64	resid794	3.97	3.20	0		
T64	resid1186	4.24	4.53	2.30	0	
T64	resid1578	2.64	4.29	4.06	2.90	0

Table S5 Standard deviations of the distances between the dyes in the blank (no peptide) MD simulations. The values are calculated as the average values of the last frame over the 8 MD simulations. Values are expressed in nm.

		T42	T42	T64	T64	T64
		resid10	resid402	resid794	resid1186	resid1578
T42	resid10	0				
T42	resid402	0.35	0			
T64	resid794	0.40	0.49	0		
T64	resid1186	0.26	0.22	0.78	0	
T64	resid1578	0.22	0.20	0.63	0.16	0

Section S6 Multi-dye system efficiency estimate for the 2- and 4-channel systems

In the 4-channel system, we considered the cases homo-FRET(ATTO-425-ATTO-425), highest efficiency of (ATTO-425 1- Acceptor 1) and (ATTO-425 2- Acceptor 1), highest efficiency of (ATTO-425 1- Acceptor 2) and (ATTO-425 2- Acceptor 2), highest efficiency of (ATTO-425 1- Acceptor 3) and (ATTO-425 2- Acceptor 3).

In the 2-channel system, we constructed the superposed FRET signal following [7] as $E_{tot} = \frac{1}{1+\frac{1}{\sum_{i=1}^{i=N}\frac{E_i}{1-E_i}}}$. This was applied to the signal of the first ATTO-425 with the three 647N ($\boldsymbol{E}_{tot1}$) and to the second ATTO-425 with the three 647N ($\boldsymbol{E}_{tot2}$). Then the higher efficiency max($\boldsymbol{E}_{tot1}, \boldsymbol{E}_{tot2}$) was used.

References

[1] Graen, T. et al., AMBER-DYES: Characterization of Charge Fluctuations and Force Field Parameterization of Fluorescent Dyes for Molecular Dynamics Simulations, *J. Chem. Theory Comput.* **2014**, *10*, 5505−5512.

[2] Friesner R. A. et al., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes, *J. Med. Chem.* **2006**, 49, 21, 6177–6196.

[3] Mollica L. et al., Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Reports* **2015**, *5*, Article number: 11539.

[4] Spitaleri, A. et al. Fast dynamic docking guided by adaptive electrostatic bias: the MD-Binding approach, *J. Chem. Theory and Comp.* **2018**, *14*, 727–1736.

[5] Rocchia, W. et al. Enhanced Sampling Methods in Drug Design, in Physico-Chemical and Computational Approaches to Drug Discovery, RSC publishing, Chapt. 11, pp. 273-301

[6] Hoefling, M. and Grubmüller, H. In silico FRET from simulated dye dynamics, *Computer Physics Communications* **2013.** *184*, 841-852.

[7] Bunt G. and Wouters F.S., FRET from single to multiplexed signaling events, *Biophysical Reviews*, **2017**, 9(2):119-129