Supplemental Methods

Molecular dynamics simulations

Apolipoprotein E4

The simulation dataset for Apolipoprotein E4 was generated similar to Stuchell-Brereton *et al.* 2023[54]. Briefly, the NMR structure of an ApoE3-like protein (PDB: 2L7B)[81] was used as a starting point. Mutations reverting this structure to the sequence of ApoE2 (112C, 158C), ApoE3 (112C,158R), ApoE4 (112R, 158R), and ApoE3-Christchurch (112C, 158R, 136S), and each mutation underwent 20 rounds of adaptive sampling using the FAST algorithm to explore the three distances pairs: R92 and S263, G182 and A241, and S223 and A291. All datasets were clustered into a shared model using backbone RMSD to a minimum difference of 3.5Å, yielding 18,182 centers. Each cluster center was solvated in a dodecahedron box with a 1.0 nm pad from the largest observed cluster center containing 0.1M sodium chloride. Each center was energy minimized and equilibrated by starting simulations at 20K and heating to 300K over a period of 2ns before a final NPT equilibration at 300K of 0.4ns. Each structure was launched on Folding @ home twice using different initial velocities and each trajectory reached 100ns, yielding an aggregate simulation time of 3.61 ms. All simulations were performed in the amber 03 force field with TIP3P water, hydrogen mass partitioning at 300K and a timestep of 4 fs. FAST simulations were performed using GROMACS and Folding @ home simulations were performed using OpenMM. Simulations were clustered using distance based clustering using the 5 FRET probe positions and 10 additional residue pairs as features, to generate a coarse model containing 8000 cluster centers. A Markov State Model was generated using a 2 ns lag time and ENSPARA's row normalization builder.

Aβ40

Simulations of Aβ40 were acquired from Robustelli *et al.* Briefly, an extended conformation of Aβ40 was simulated in the following force fields: a99SB*-ILDN with TIP3P, C22* with TIP3P-CHARMM, C36m with TIP3P-CHARMM, a03ws with TIP4P/2005 interactions, a99SB with TIP4P-Ew with the Head-Gordon vdW and dihedral modifications (a99SB-UCB), a99SB-ILDN with TIP4P-D, and a99SB-disp with a modified TIP4P-D water. Simulations were run at 300K in NPT ensemble on Anton hardware with a 2.5-fs time step for a total of ~30 μs.

For simulations generated during this manuscript, we used either amber03 or amber99sb-ws force fields with TIP3P water. Simulations were started from the top 10 divergent structures of Aβ40 found in the Robustelli *et al.* simulations. Each structure was solvated in a cubic box with box lengths of 12.307 nm which was determined by solvating the fully unfolded Aβ40 with a 1 nm pad, 0.1M sodium chloride, and virtual sites for hydrogens. Each structure was energy

minimized and allowed to equilibrate for 1 ns before starting production runs. Three 250ns long replica simulations with independent velocities were started from each pose, for a total of 7.5 μs of aggregate simulation time using a 4 fs timestep.

Clustering was performed using the distance between every 5th residue as an input feature to generate 250 unique cluster centers. MSMs were generated using a lagtime of 5 ns (Anton datasets) or 2 ns.

T4 Lysozyme

Simulations of T4 Lysozyme were initialized using PDB structure 5LZM, solvated in a cubic box with edges extending 1.8 nm beyond the edge of the protein with TIP3P water and 0.1M sodium chloride. Virtual sites were included for hydrogens. Structures were energy minimized for 1000 steps and equilibrated for 1 ns prior to production runs. For initial unbiased simulations, 5 replica production runs were performed for 5 μs each with each run having differing initial velocities.

Metadynamics simulations were performed using PLUMED with the metad restraint using a pace of 500, gaussian height of 0.3, and gaussian widths of 0.05 for a total of 250 ns. Biases were placed on the distance between residue 44 and 150 using both CA-CA distance the terminal side-chain atoms. 4 divergent structures were taken from the minimal 44-150 distances observed in the metadynamics simulation, resolvated in a cubic box with a 1.8 nm pad of TIP3P water and 0.1M sodium chloride, and re-energy minimized and re-equilibrated. Each pose was run with 5 replicates with differing initial velocities for 1 μs per each replica.

Adaptive sampling simulations exploring the transition between 5LZM and the alternate state of lysozyme identified by metadynamics were performed using FAST. Briefly, 10 40 ns long simulations were started from either 5LZM or the alternate pose of lysozyme. These simulations were clustered, a MSM was built, and 10 states with a minimal backbone RMSD to the target state were chosen to restart 40 ns simulations from. We iterated between clustering and simulation until states were identified with a backbone RMSD of <2Å. All simulations were performed at 300K with GROMACS 2020.

Coarse grained models were built on the initial unbiased simulations from 5LZM or a combination of the initial unbiased simulations from 5LZM, the 4 differing alternate states, and the FAST simulations observing the transition between the alternate state of lysozyme and

5LZM. For both models, clustering was performed based on RMSD of the backbone to either 500 centers, or a final RMSD of 2.5 Å, whichever was greater. Markov State Models were generated using ENSPARA's normalize method with a lagtime of 2 ns.

Post-simulation modeling of smFRET

Direct simulation of dye emission events was achieved by building a MSM for both donor and acceptor dyes. We model each dye conformation from the MSM onto each labeling position in the protein MSM, discarding any dye positions that resulted in steric clashes with the protein. Next, for each state in the protein MSM, we simulate dye emission events similar to previous methods [40,41]. Briefly, we choose a random dye starting position for both the acceptor and donor dyes from the MSM based on their equilibrium probability. Next, we calculate the probability that the donor dye can undergo radiative decay (p_{rad}, emit a donor photon), transfer energy to the acceptor (p_{RET}, emit an acceptor photon), non-radiatively decay (p_{nonrad} no observed photon), or remain excited (p_{remain}).

$$p_{rad} = 1 - e^{(-k_{rad} * \Delta t)} \quad \text{Equation 2}$$

$$p_{RET} = 1 - e^{(-k_{RET} * \Delta t)} \quad \text{Equation 3}$$

$$p_{nonrad} = 1 - e^{(-k_{nonrad} * \Delta t)} \quad \text{Equation 4}$$

$$p_{remain} = 1 - p_{rad} - p_{RET} - p_{nonrad} \quad \text{Equation 5}$$

Where Δt is the timestep of the Monte Carlo which is the same as the dye MSM lagtime (2 ps), k_{rad} is the rate of radiative decay, k_{RET} is the rate of energy transfer, and k_{nonrad} is the rate of nonradiative decay, given by the following:

$$k_{rad} = \frac{Q_D}{\tau_D} \quad \text{Equation 6}$$

$$k_{RET} = \frac{1}{\tau_D} * \left(\frac{R_0}{r}\right)^6 \quad \text{Equation 7}$$

$$k_{nonrad} = \frac{1}{\tau_D} - k_{rad} \quad \text{Equation 8}$$

Where Q_D is the donor fluorescence yield in the absence acceptor, τ_D the donor lifetime in the absence acceptor, r the distance between the donor and acceptor emission centers, and R_0 the Förster radius, given by the following:

$$R_0^6 = \frac{2.07}{128\pi^5 N_A}\frac{\kappa^2 Q_D}{n^4}J \qquad \text{Equation 9}$$

Where N_A is Avogadro's number, n the refractive index of the solution, J the donor-acceptor spectral overlap integral, and κ^2 the dipole orientation factor between the two dyes. In this equation, we hold all values constant except for κ^2 which we calculate according to:

$$\kappa^2 = (\cos(\theta_{DA}) - 3\cos(\theta_D)\cos(\theta_A))^2 \qquad \text{Equation 10}$$

Where θ_{DA} is the angle between the donor dipole moment and the acceptor dipole moment, θ_D the angle between the donor dipole moment ($\hat{d}$) and the vector connecting the donor emission center to the acceptor emission center ($\hat{r}$), and θ_A the angle between the acceptor dipole moment ($\hat{a}$) and $\hat{r}$.

After we calculated the emission outcome probabilities, we choose a random outcome weighted by the respective probability of occurring. If the donor dye remains excited, we allow both dyes to update their positions based on the probability of transitioning states from the dye MSM, recalculate potential emission outcomes, and choose another dye outcome. We repeat this process until the donor dye is no longer excited, recording both the number of Monte Carlo steps required to reach the emission event (dye lifetime) as well as the outcome. To enable efficient computation, we pre-calculate the lifetimes and outcomes for each protein center, repeating each Monte Carlo simulation 1000 times (ApoE) or 5000 times (Lysozyme, Aβ40) to scale with the respective numbers of cluster centers that each protein MSM has.

For the dye point cloud method dye molecules of interest were modeled onto the protein at the appropriate labeling positions. We do this using a rotamer library approach based on prior work[26]. Briefly, dyes attached to the appropriate label and linker were simulated free in solution to determine all the potential dye configurations. All resulting simulation frames were aligned based on the backbone and the center of fluorescence emission from the dye was saved as a single point to generate a point cloud of all potential emission centers. Next, the point cloud is modeled onto the protein labeling position of interest and all points that would result in a steric clash are discarded. Finally, we generate a distance probability distribution which describes the distance between all potential configurations of the donor and acceptor dyes. We determine the photon color by choosing a random donor-acceptor distance, assessing the probability of transfer based on the Förster relationship (Equation 1), and

choose whether the donor was a photon or acceptor based on the established probability. We specify an R_0 of 5.6 nm for Alexa488 and Alexa594.

To account for protein conformational sampling during the measurement window, we recolor an experimental photon time course from experiments performed on ApoE. Using our MSM, we generate a synthetic trajectory that matches the length of the experimental photon burst. The trajectory starts from a random state in the MSM based on the equilibrium probability of that state, and the synthetic trajectory is built based on the probability of that state transitioning to any other state in the MSM. The length of the trajectory is determined based on the length of the experimental photon burst and rescaled to account for simulations being faster than experiment. In our calculations, we use a time-factor of 10,000. Each time an experimental photon is recorded, we note the corresponding frame in the synthetic trajectory and decide whether the photon was a donor or acceptor photon based on the schema outlined above (dye emission simulation or point cloud approximation). We repeat this for each observed photon in the experimental burst and return a total FRET efficiency for the burst as the ratio of observed acceptor photons and total observed photons. This entire process is repeated, generating new synthetic trajectories from new starting states, for each observed molecule in the ApoE experiment resulting in >14,000 observations.

Code used to run dye modeling and smFRET calculations are available on github: https://github.com/bowman-lab/enspara. MSMs of proteins and dyes, as well as example code for running smFRET calculations and generating dye MSMs is available on OSF: https://osf.io/82xtd/?view_only=b7f354e86eb144a69d9d047b42e21a9f.

Analysis/Software

Simulations generated during this manuscript were performed in GROMACS or OpenMM as noted. Structure viewing was performed in PyMOL. Trajectory analysis was performed using MDtraj. Clustering and MSMs were created using ENSPARA. All graphs were generated using Matplotlib.

Supplementary Figures

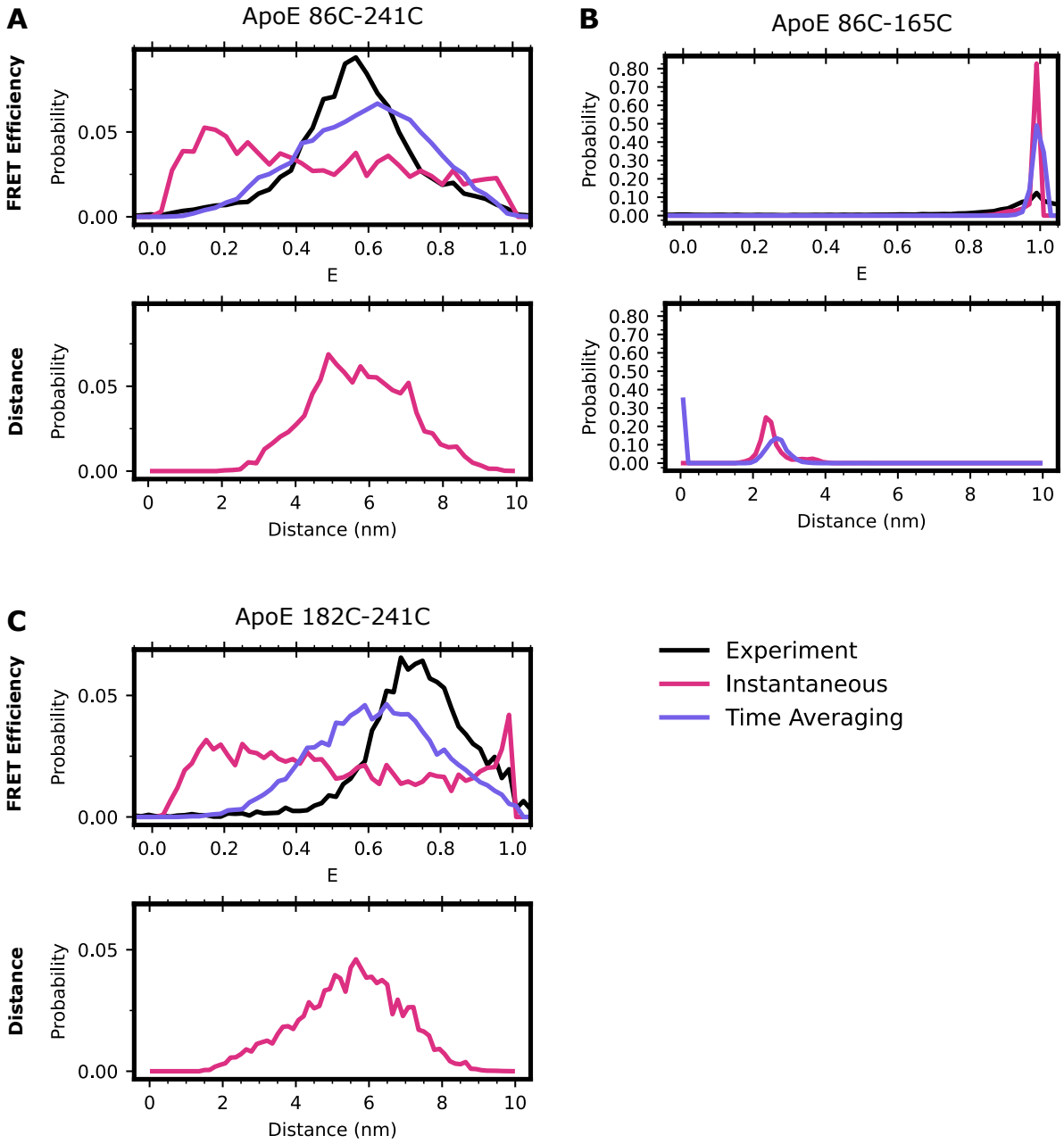


Figure S1: Accounting for time averaging significantly alters the apparent structural distribution from our model and increases agreement with experiments. Top, smFRET histograms for experimental (black), instantaneous simulation (red), or time averaged simulation (purple), bottom the inter-dye distances for apolipoprotein E. Labeled positions are A) 86-241, B) 86-165, or C) 182-241. In all cases, labeling is performed with Alexa 488 and Alexa 594 using maleimide chemistry.

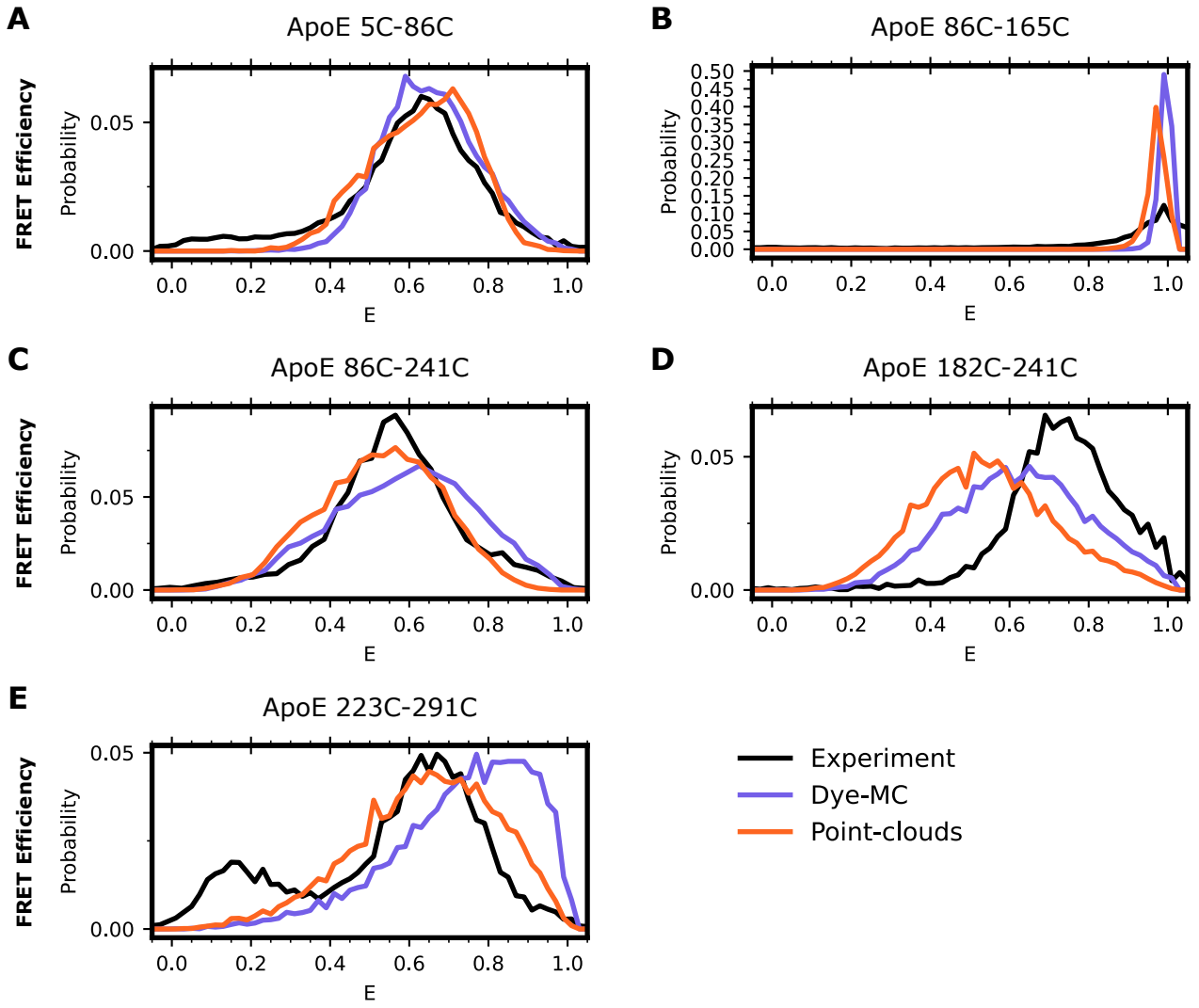
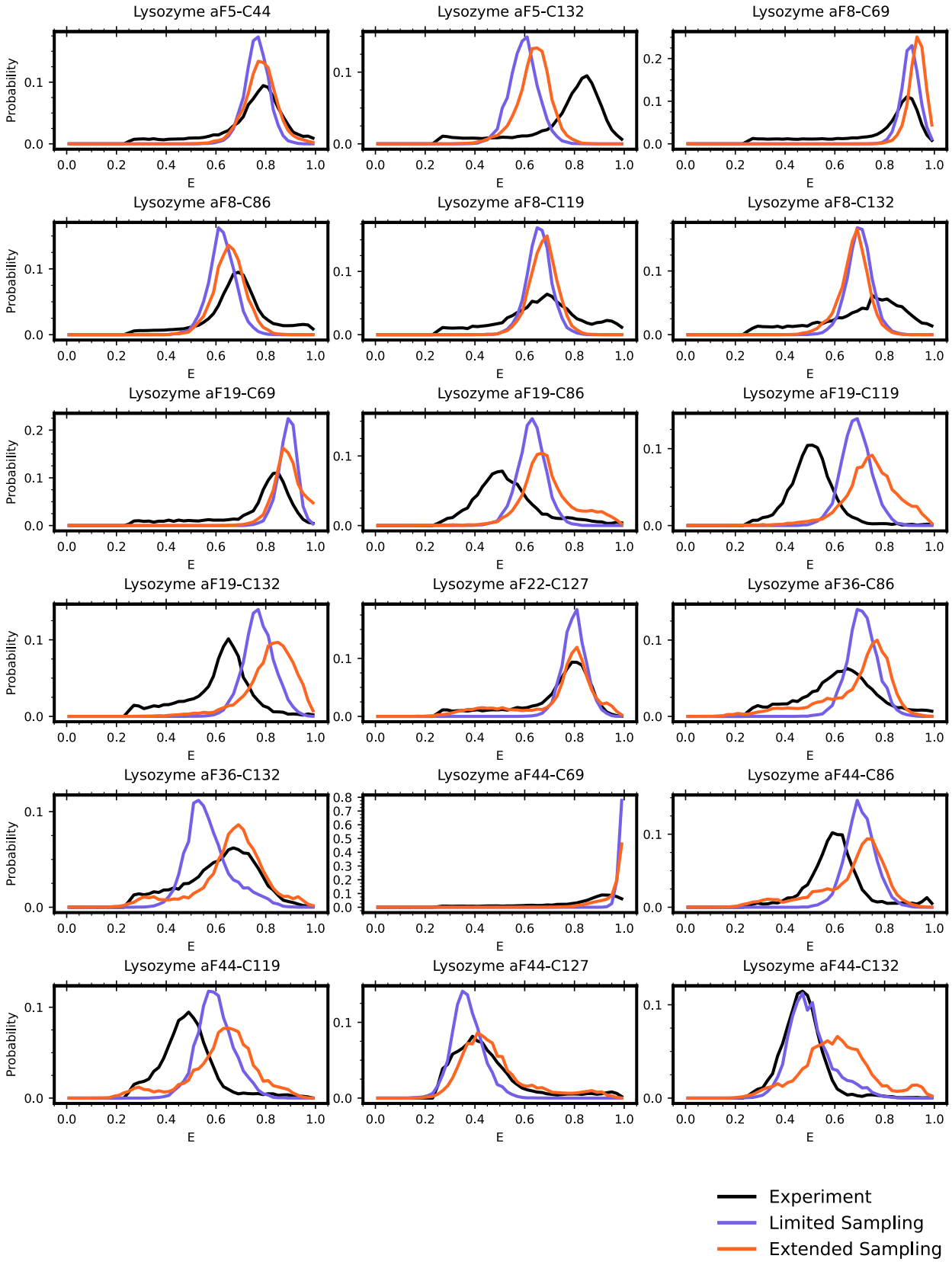


Figure S2: Treating dyes as point clouds yields comparable results to accounting for dye dynamics. FRET efficiencies for apolipoprotein E labeled with Alexafluor 488 and Alexafluor 594 at positions A) 5-86, B) 86-165, C) 86-241, D) 182-241, and E) 223-291. The black trace is the experimental distribution, in purple is accounting for time averaging while accounting for dye-dynamics (Dye-MC), and in orange is treating dyes as a point cloud with no dynamics, using a constant R0 of 5.6.



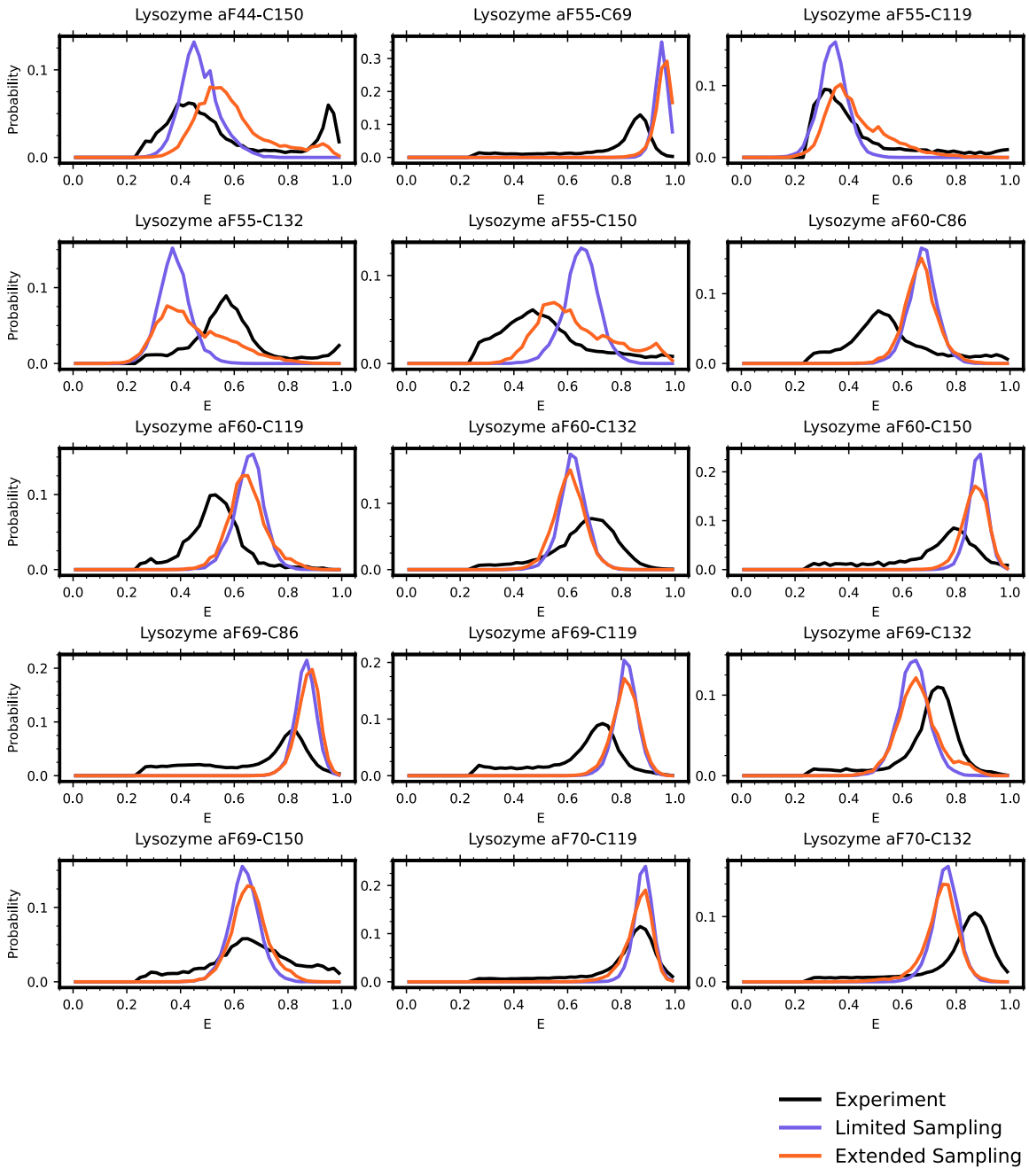


Figure S3: Extended sampling of lysozyme yields improved agreement with experiment. FRET efficiency distributions for various lysozyme probe positions. In black is the experimental trace, donor only counts (E < 0.25) have been removed for comparison purposes as simulated FRET has total labeling. In purple is the distribution from our initial simulation runs which only sample crystal-like poses. In orange is a model of Lysozyme which includes the novel state. Simulations run in amber03 force field using TIP3P water. All calculated FRET was performed while accounting for both dye and protein dynamics.

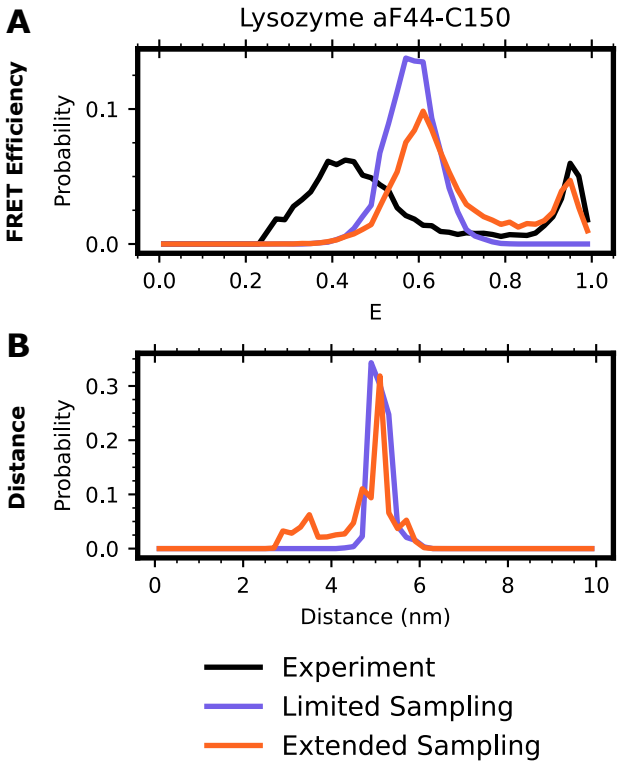


Figure S4: Limited sampling of lysozyme using charmm36m fails to recapitulate the third state of Lysozyme. FRET efficiency distributions for lysozyme 44-150. The black trace is the experimental distribution with donor only counts (E < 0.25) removed for clarity. in red is the equilibrium distance distribution from simulation accounting for added dye-distances, and in purple is the effect of time averaging on the red trace. Simulations run in charmm36m with TIP3P water.