JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Estimation of Hydrogen-Exchange Protection Factors from MD Simulation Based on Amide Hydrogen Bonding Analysis

In-Hee Park,[†] John D. Venable,[†] Caitlin Steckler,[†,§] Susan E. Cellitti,[†] Scott A. Lesley,[†,‡,§] Glen Spraggon,[†] and Ansgar Brock*[,†]

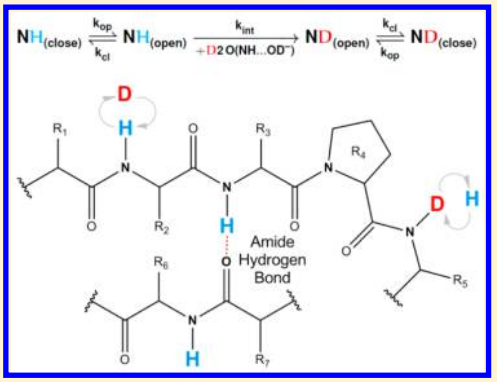
[†]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, United States

[‡]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California 92037, United States

[§]Joint Center for Structural Genomics, La Jolla, California 92037, United States

Supporting Information

ABSTRACT: Hydrogen exchange (HX) studies have provided critical insight into our understanding of protein folding, structure, and dynamics. More recently, hydrogen exchange mass spectrometry (HX-MS) has become a widely applicable tool for HX studies. The interpretation of the wealth of data generated by HX-MS experiments as well as other HX methods would greatly benefit from the availability of exchange predictions derived from structures or models for comparison with experiment. Most reported computational HX modeling studies have employed solvent-accessible-surface-area based metrics in attempts to interpret HX data on the basis of structures or models. In this study, a computational HX-MS prediction method based on classification of the amide hydrogen bonding modes mimicking the local unfolding model is demonstrated. Analysis of the NH bonding configurations from molecular dynamics (MD) simulation snapshots is used to determine partitioning over bonded and nonbonded NH states and is directly mapped into a protection factor (PF) using a logistics growth function. Predicted PFs are then used for calculating deuteration values of peptides and compared with experimental data. Hydrogen exchange MS data for fatty acid synthase thioesterase (FAS-TE) collected for a range of pHs and temperatures was used for detailed evaluation of the approach. High correlation between prediction and experiment for observable fragment peptides is observed in the FAS-TE and additional benchmarking systems that included various apo/holo proteins for which literature data were available. In addition, it is shown that HX modeling can improve experimental resolution through decomposition of in-exchange curves into rate classes, which correlate with prediction from MD. Successful rate class decompositions provide further evidence that the presented approach captures the underlying physical processes correctly at the single residue level. This assessment is further strengthened in a comparison of residue resolved protection factor predictions for staphylococcal nuclease with NMR data, which was also used to compare prediction performance with other algorithms described in the literature. The demonstrated transferable and scalable MD based HX prediction approach adds significantly to the available tools for HX-MS data interpretation based on available structures and models.



INTRODUCTION

Mapping of protein−protein and protein−small molecule interactions by hydrogen−deuterium exchange mass spectrometry (HX-MS) is now extensively used.[1,2] Compared to other chemical labeling methods, HX has the advantage of a uniform probe distribution (amide hydrogens) across the system under study and a labeling chemistry that is based on a naturally occurring exchange mechanism that does not perturb the structure and dynamics of the system via the introduction of large atoms or groups and/or electrostatic modification.[3] In addition, the use of MS allows sensitive analysis in even complex matrices. Expressions for the HX rate equations derived from the local folding/unfolding model are widely used.[4] The general description of amide hydrogen (NH) exchange assumes a pre-equilibrium between closed (i.e., exchange incompetent or folded) state and an exchange competent or open (i.e., unfolded) state. Only in the open state, exchange of NH with solvent hydrogen is possible (Figure 1). In the closed state, the amide hydrogen is protected from exchange by virtue of either being hydrogen bonded to other protein hydrogen bond (H-bond) acceptor atoms or exclusion from the solvent. In HX, the native exchange is visualized by dilution of a protein into a buffer containing other hydrogen isotopes like deuterium. This results in incorporation of deuterium into the protein by exchange of NH with deuterium (D) from the solvent. As deuterium is 1 Da (Da) heavier than

Received: April 2, 2015
Published: August 4, 2015

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

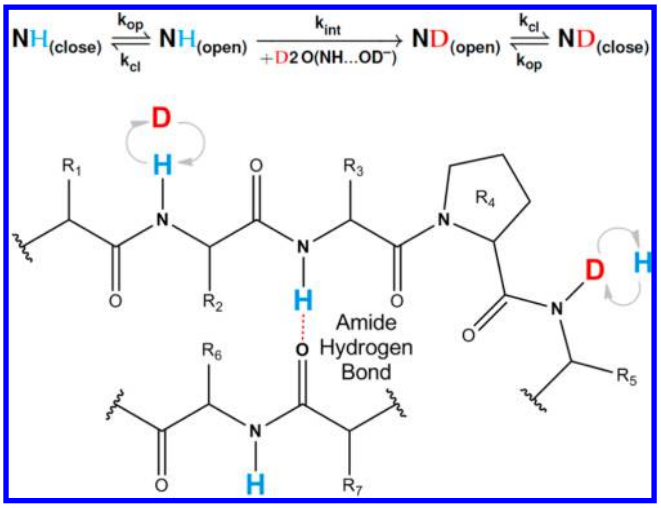


Figure 1. Illustration of hydrogen−deuterium exchange between amide hydrogen (NH) in blue and deuterium (D) from the solvent D_2O in red.

hydrogen the labeling of a protein can be followed by its change in mass using MS. Further, a change in pH and temperature can sufficiently stabilize the backbone amide labels to allow the fragmentation of the protein with acid proteases. This localization of the label to a specific fragment resolves the incorporation of the deuterium on the primary sequence level. From quantitation of the observed mass shift over time, local rate information can be extracted.[2]

HX: Can Probe Structural Change. Protein structure can be largely characterized by a protein's backbone NH hydrogen bonding (H-bonding) arrangement in the folded state. Changes in the folded state typically require partial or complete unfolding that is synonymous with breaking of backbone NH hydrogen bonds, which changes the exchange competence of the NH involved. Changes in the rate of HX are indicative of conformational changes or changes in the protein dynamics. This makes HX an ideal tool for studying protein−ligand interactions, protein folding, or the intrinsic stability of a protein at a specific condition such as pH, temperature, or denaturant concentration.[5] It is customary to express the observable exchange rate (k_{obs}) as product of an intrinsic chemical rate (k_{int}) and the inverse of a protection factor (PF) eq 1 as suggested by Englander and Kallenbach:[6]

$$k_{obs} = k_{int}\left(\frac{k_{op}}{k_{cl}}\right) = k_{int}/\mathrm{PF} \tag{1}$$

In eq 1, k_{int} represents the protein amide specific intrinsic chemical rate of the fully open state for a given pH, temperature, and set of bracketing amino acid side chains. The PF is defined as the ratio of the closing and opening rate constants (k_{cl}/k_{op}) as shown in the pre-equilibrium scheme I Figure 1. Using basic thermodynamic relationships the PF can be related to the opening free energy via eq 2.

$$\ln \mathrm{PF} = \ln k_{cl} - \ln k_{op} = \Delta G_{op}/RT \tag{2}$$

The maximum opening or local unfolding free energy ΔG_{op} is calculated to be 6.6−8.2 kcal/mol, from HX-NMR measurements at 30 °C.[7]

Prior Attempts to Build Computational Models of HX. The determinants of hydrogen exchange and the merits of the mechanistic models derived over the years based on solvent accessibility, solvent penetration, electrostatics, polarizability, packing density, structural dynamics, strength, and length of the amide hydrogen bond have been discussed in detail by relating single amide resolved exchange data to high resolution crystal structures.[4,8,9] The authors of these publications arrive at the conclusion that hydrogen bonding is one of the most important determinant of exchange and that the structural environment provides additional modulatory effects (e.g., burial, etc.). Other specific factors including hydrogen bond strength/length, electrostatics, and small atomic displacements are found to correlate poorly or not at all with single amide resolved exchange rates. Further, from the findings it is concluded that a successful HX prediction algorithm will need to be able to differentiate alternative pathways that lead to exchange competence and that interpretation of HX of individual amides in the detailed structural context is most meaningful in elucidating pathways and mechanism.[4]

Published attempts to model the PF either use a direct approach of parametrizing the open/closed (or unfolded/folded) state ratio as a function of H-bonding and SASA, or formulate the problem on the basis of a pseudoenergy function that indirectly incorporates the open/closed metric as summarized below.

Solvent accessibility suggests itself as a significant parameter in HX, which is a chemical labeling method that uses the solvent as reagent. In that regard, solvent-accessible-surface-area (SASA) of a residue has been one of the frequently used HX modeling metrics. The SASA is typically used as an approximation for the conformational entropy in conjunction with conventional free-energy calculation protocols instead of time-demanding normal-mode analysis.[10] However, given the complex geometric characteristics of amide hydrogen (NH) response to structural change, the SASA averaged over all atoms in a residue is only a crude estimate, while the SASA of the NH atom alone is not very sensitive (see below).

Liu et al.[11] published the DXCOREX algorithm using an empirical energy function based on the parametrization of energy and entropy terms as a function of SASA of polar and nonpolar atoms as derived from a limited base set of globular proteins. Their algorithm uses a statistical thermodynamics formulation enumerating an ensemble of native-like states by sequence-partitioning. This is accomplished by assignment of successive short sequence stretches called folding units to either a folded or unfolded state. Each microstate generated in this fashion is scored by summation of all constituent residues' SASA-based energy relative to the tripeptide model (Gly-X-Gly) representing the fully unfolded state of the corresponding residue. A PF is calculated as the ratio of the folded-ensemble-averaged probabilities over unfolded-ensemble-averaged probabilities.

Vendruscolo et al.[12] used Monte Carlo (MC) sampling with experimental constraints from NMR data to fit the parameters of a phenomenological expression of the PF to approximate the experimentally observed one. The PF was modeled as a function of the number of contacts with other residues N^c and the number of hydrogen bonding interactions of the amide hydrogen N^h using the expression $\ln \mathrm{PF} = \beta_c^{nb}N^c + \beta_h^{bond}N^h$. They found that $\beta_c^{nb} = 1$ and $\beta_h^{bond} = 5$, which corresponds to 0.6 kcal/mol (= RT) for a nonbonded interaction and 3 kcal/mol (= $5RT$) for a hydrogen bond. Their work is specifically designed to use HX-NMR data by utilizing (i) NMR constraints to guide the MC simulation of the structural ensemble and (ii) experimental PF as a scaling factor of the predicted PF to make it comparable to the experimental value.

Craig et al.[13] compare coarse-grained model[14] predictions of protections factors for human ubiquitin, chymotrypsin inhibitor

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

2, and staphylococcal nuclease (SNase) with NMR derived values. The work is unique in that it is to the authors' knowledge the only attempt of using coarse-graining to sample a large fraction of the structural ensemble in an attempt to improve HX prediction accuracy. Accessibility and hydrogen bonding criteria were developed in the work using number of native contacts and the change in the pairwise distance of Cβ atoms between snapshots, respectively. These criteria were used to define open and closed states and probabilities for residues being in an exchange competent and incompetent state were calculated using weighted histogram analysis[15] and integration over the global reaction coordinate. Key findings of the work were the requirement of a significant distortion of the local environment of a residue to produce an exchange-competent state and the ability to predict HX without explicit consideration of hydrogen bonding energy and geometry.

Kieseritzky et al.[16] describe an MD-based ensemble approach using various metrics for PF modeling via linear combinations similar to Vendruscolo et al.[12] They observed that backbone amide PFs show positive correlations with the number of contacts, H-bond occupancy and H-bond survival times. Further, inverse correlation with fluctuations of backbone atoms and H-bond lengths derived from MD simulation data were observed.

Recently, Petruk et al.[17] used MD simulations and a metric based on average SASA of NH and the number of H-bonds with water molecules as the basis for the binary decision if exchange would occur or not in the ERK2MAP kinase system. The authors were successful in explaining some of the observed differences between apo and holo dynamics in terms of their metrics.

Garcia and Hummer[18] applied MD simulations and ensemble averaging of the mean square displacement to Cytochrome c. They found that the opening and closing of backbone hydrogen bonds involved in secondary structure stabilization could be better understood by monitoring the amino group interactions with water through the NH−OW pair correlation function and the number of waters that occupy the first hydration shell of these atoms.

Ma and Nussinov[19] considered the average number of H-bonds of an NH with the peptide backbone (NH_β) in various Aß42 peptide structures (folded state) and the average number of H-bond of an NH with water (NH_{sol}) (unfolded state). They used these terms in the expression $(100 - P_{sol}) = C{\cdot}NH_{sol}/(NH_{sol} + NH_\beta)$ to estimate the exchange probability and concluded that it should be possible to identify major structural species of a polymorphic structural ensemble based on correlation between NMR data and prediction.

Sljoka and Wilson[20] used NMR ensemble based modeling using H-bonding as a quantitative rigidity/flexibility predictor together with the SASA of NH as a qualitative metric for HX prediction. Lastly, Resing et al.[21] empirically modeled exchange for ERK2 kinase helices by regression of the PF using SASA, hydrogen bond length, and a positional parameter according to $\log PF = a{\cdot}SASA + b/(\text{H-bond length}) + c(\text{distance from } \alpha \text{ helix ends})$.

NH: Specifically Able to Probe Structural Change. NMR and MS provide complementary technical capabilities in exchange experiments. The former provides single-residue information by default, although sequence coverage is often limited especially for larger systems because of a laborious labeling and resonance assignment process, whereas the latter can often be performed even in complex matrices but sequence resolution is moderate and largely dependent on the size and number of observable peptides.[22] Experimental residue-specific PF data from the Biological Magnetic Resonance Bank (BMBR) show clear anticorrelations between SASA of NH and experimental PFs (Supporting Figure S1) and have led many to suggest SASA might be a suitable determinant for HX. However, as evident from literature cited above there are no solid reference values for modeling purposes that would allow PF prediction on an absolute scale. Part of this is explained by the different SASA algorithms in use, such as the Richards and Lee method[23] or the Shrake and Rupley method,[24] that result in different SASA values. Further, it is impossible to discriminate the large number of buried NH typically found in proteins based on SASA alone as all of them have a zero SASA. Parameterizations of energy functions in terms of SASA are observed to be very sensitive to even small changes in SASA. Additionally, a pure SASA approach does not provide a clear threshold value of SASA of NH for representing the folded state of residues whose NHs are fully exposed at the protein surface, either.

To overcome the limitations of a SASA approach while retaining the specificity of NH as a probe atom, we have implemented an alternative metric on the basis of the NH-bond status alone; this seemed most consistent with what is known about the structural dynamics of proteins as summarized by Englander and Kallenbach.[6] Key observations are that hydrogen bonding makes HX slow and that HX chemistry is controlled by structure effects as follows: First, the HX rate is affected by H-bond (not only via backbone NH but also backbone CO or other side-chains) that blocks effective proton transfer to NH. Second, the HX rate may be slow without H-bond formation because burial of NH alone may be sufficient to retard the exchange, although such cases are the rare exception. Third, the mere proximity of water to NH is an insufficient requirement for exchange competence. Further, H-bonding contributions are one of the major determinants in the definition of structure[25] and energetics of protein folds.[26]

The stated observations suggest that exchange propensity should be a function of both the persistence and nature of the NH H-bond. Therefore, we hypothesized that an approach that quantifies the H-bonding relationships of NH on the basis of ensemble solution structures from MD simulations alone should be ideal for quantitative prediction of exchange propensities.[27−29]

In the following, we demonstrate a computational HX prediction method quantifying the interactions of NH with other internal residues (closed state of NH) and those with explicit waters (open state of NH) over an ensemble of structures of a protein in solution generated by MD simulation. First, it is shown that this MD-based method achieves high correlations between model predictions and experimental PFs for a comprehensive HX-MS data set collected for Fatty acid synthase thioesterase domain (FAS-TE),[30] which was studied in detail, under a wide range of experimental conditions. Second, application of the method to other published data sets including apo/holo proteins (ligand/receptor-bound including metal ligand), is made to demonstrate the general transferability/applicability and utility of the method and its utility in discriminating possible protein conformations based on comparison of experimental data with MD-based predictions. Comparison of MD-based predictions with published DXCOREX results[11] is used to show improved correlations for the MD method. Further, decomposition of experimental in-exchange curves into rate classes is performed and then compared with decompositions of computational predictions, demonstrating a means to improve the NH resolution of a typical fragmentation HX-MS experiment. Lastly, predictions of residue resolved protection factors for staphylococcal nuclease (SNase) are compared with NMR protection factors

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

and predictions from three other models to demonstrate the comparative performance of the algorithm.

METHODS

Computational HX Modeling Protocol. All the calculations and data analyses described here are implemented in Python 2.7.3 and R 3.0.1 (nls.lm package for nonlinear fitting and lattice package for multiple-factor data visualization).

Input. 3D structure coordinates of X-ray crystallography structures listed in Table 1 were used for MD simulations. Models for comparing wild-type versus mutant, apo versus holo, etc., were constructed by modifying available structures where required.

Table 1. HX-MS Data Set Used in the DXCOREX Model[11]

PDB	protein systems	seq	ref
1NFI	bound IκBα (chains: A, B, and F)	615	31
	free IκBα (chain: F)	213	
	pH = 7.5; temp = 298 K; time = 2 min		
2EYI	apo α-actin CH_2 domain	116	32
	pH = 2.5; temp = 277 K; time = avg(0.25, 0.5, 1, 2, 5, and 15 min)		
2NT1	apo GCase	497	33
2NSX	holo GCase with isofagomine	498	
	pH = 7.8; temp = 296 K; time = 0.8, 1.6, 5, 16.6, and 50 min		
1PU0	dimer superoxide dismutase WT	206	34
	dimer SOD G85R mutant	206	
	pH = 7.2; temp = 277 K; time = avg(0.25, 0.8, 2.5, and 8.3 min)		
1P38	MAPK p38	357	17
	pH = 7.5; temp = 298 K; time = 300 min		

Ensemble Generation by MD Simulation. Water solvated MD simulations to generate trajectories of proteins were carried out using Amber 11 (ff99SB force field). Initially, the protein was solvated using the tleap protocol of AmberTools with a 12 Å TIP3BOX explicit water model, followed by neutralization of the system by adding counterions. This was followed by solvent relaxation imposing Cartesian restraints on the protein, which were subsequently released for minimization of the entire system using the MPI version of sander protocol of Amber11 package. Temperature equilibration was performed for 20 ps heating at *NVT* from 0 to 300 K, pressure equilibration for 300 ps at *NPT* with weak coupling and SHAKE, and finally a production run at *NVT* for 50−100 ns. For rapid equilibration and production runs the parallel cuda-enabled pmemd protocol of Amber11 was executed on the workstation with the following specifications: Intel Xeon CPUs, Ubuntu 12.04 LTS (Precise) 64-bit OS equipped with two GPUs (NVIDA Tesla C2075). NH-bond analysis with explicit waters was performed using UCSF Chimera[35] (see below) after reimaging of the water molecules back into the central simulation box using ptraj protocol of AmberTools.

Amide Hydrogen (NH) Bond Statistics from MD Snapshots. Solute−solute NH-bonds and solute−solvent NH-bonds from MD snapshots were quantified at every 20 ps time step over 50−100 ns MD trajectories. The "closed/folding" propensity for a NH_j (j = residue index) was modeled from the number of snapshots showing H-bonding to solute no. $(NH{:}CO)_j$. Likewise, the "open/unfolding" propensity was modeled via the number of snapshots showing H-bonding to solvent no. $(NH{:}water)_j$. The difference no. $(NH{:}CO)_j$ − no. $(NH{:}water)_j$ was normalized by the total number of snapshots eq 3 and used as a representation of the overall "NH-bond statistics" ranging from −1 to 1.

$$\mathrm{NHstat}_j = \frac{\mathrm{no.(NH{:}\ protein)}_j - \mathrm{no.(NH{:}\ water)}_j}{\text{total no. of snapshots}} \tag{3}$$

More extended NH bond models were constructed using the definitions in Table 2 for counting of snapshots.

Protection Factor (PF) Modeling with NH-Bond Statistics. To map the range of −1 to 1 from eq 3 into a PF scale of 1 to PFmax a logistic growth function was employed. The logistics growth function $y = c/(1 + ab^x)$ provides approximately exponential weighing to the NH-bond statistics in eq 4. The three parameters a, b, and c were determined by imposing constraints: (i) for the upper bound of the PF $(x, y) = (1, \text{base})$ for maximum PF = base/2; (ii) a midpoint passing through $(x, y) = (0, \sqrt{\text{base}})$; and (iii) a lower bound to be set $(x, y) = (-1, 1)$ for minimum PF. The final form of the fitting function is then

$$\mathrm{PF}_j = \frac{\text{base}}{1 + \sqrt{\text{base}}\cdot(1/\sqrt{\text{base}})^{\mathrm{NHstat}_j}} \tag{4}$$

where the parameter "base" is the only adjustable parameter of the model. The value of base can be set by referring to the HX-NMR experimental PF values or derived on the basis of an optimal correlation.[7,22] For example, slowest exchanging NHs from NMR measurement suggest ΔG_{op} = 6.6−8.2 kcal/mol at 30 °C,[7] which corresponds to a base value range of 1×10^4 to 1×10^6 by eq 2.

Calculation of Peptide Deuteration from PF Modeling Results. Once PF_j (j = residue index) is calculated, deuterium incorporation (DI) for a peptide is estimated by summing contributions of exchangeable NH_j for each residue using eq 5. To exclude the N-terminus and the first backbone amide of a peptide, whose back-exchange rates are too fast to be observed in our fragmentation HX-MS experiments, the residue index runs from $j = m + 2$ (where m corresponds to the N-terminal residue of a given peptide) to $j = n$ (where n corresponds to the C-terminal residue of a given peptide):

$$\mathrm{DI}_{\mathrm{fp},t} = \sum_{j=m+2}^{n}\left(1 - \exp\left[\frac{-k_{\mathrm{int},j}}{\mathrm{PF}_j}\right]t\right) \tag{5}$$

In eq 5, t is time in units of either minutes or seconds; $k_{\mathrm{int},j}$ is an intrinsic chemical rate in matching units of inverse minutes or seconds. The applicable experimental conditions of pH and temperature as well as the protective effect due to the neighboring side-chains of an NH are captured in $k_{\mathrm{int},j}$, which also serves as a maximum upper-bound of each NH's exchange rate (Supporting Information 1).[36]

Table 2. Amide Hydrogen Bond Models

NH-bond option	no. (NH:protein) HX incompetent "closed" state counting solute−solute interaction snapshots	no. (NH:water) HX competent "open" state counting solute−solvent interaction snapshots
HB1	no. (NH:CO)	no. (NH:wat)
HB2	no. (NH:CO + NH:side-chain)	no. (NH:wat)
HB3	no. (NH:CO + NH:side-chain + C═O:side-chain)	no. (NH:wat + C═O:wat)

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

Table 3. Comparison of Correlation Coefficient (R) and Regression Equation for the DXCOREX Protein Set[a]

method	1NFI bound	1NFI free	1PU0 G85R	1PU0 WT	2EYI apo	2NT1 apo	2NSX holo	1P38
k_{int} only	0.51	0.81	0.92	0.96	0.72	0.70	0.70	0.87
R/R^2	$y = 0.5x + 9.5$	$y = 0.5x + 7.8$	$y = 1.1x + 4.4$	$y = 2.0x + 3.1$	$y = 0.8x + 2.3$	$y = 1.2x + 3.9$	$y = 1.2x + 3.9$	$y = 1.2x + 2.9$
HX model	0.80	0.92	**0.99**	**0.98**	0.70	**0.81**	**0.82**	0.88
R/R^2	$y = 0.8x + 2.7$	$y = 0.6x + 3.1$	$\boldsymbol{y = 0.8x + 0.9}$	$\boldsymbol{y = 1.4x + 0.4}$	$y = 0.6x + 1.4$	$\boldsymbol{y = 0.9x + 1.5}$	$\boldsymbol{y = 1.0 + 1.5}$	$y = 1.0x + 2.2$
DXCOREX	**0.96**		NA	0.84	**0.91**	0.75		**0.93**
R/R^2[b]	$\boldsymbol{y = 0.75x + 0.78}$			$y = 0.7x + 3.0$	$\boldsymbol{y = 1.1x + 0.2}$	$y = 0.9x + 2.0$		$\boldsymbol{y = 0.99x - 0.2}$

[a]Bold entries denote the highest R among three prediction methods. [b]R values calculated from R^2 as reported in the work of Liu et al.[11]

HX-MS FAS-TE Data Set. FAS-TE is an enzyme participating in the conversion of dietary carbohydrate to fat. It has been pursued as an anticancer target because increased expression of FAS is a hallmark of all major cancers. FAS-TE was chosen as the model protein for this study because of its prior in-house use as model system,[37] ready availability (see Supporting Information 3 for expression and purification), and availability of in-house structural data (PDB code: 4Z49).

An extensive data set varying pD (5, 6, 6.5, 7, 7.5, and 8), temperature (0 and 25 °C), and time (10, 30, 270, 810, 2430, 7290, 21 870, and 65 610 s) was collected on the HX behavior of FAS-TE for the purpose of accessing the predictive performance of the models over the typical range of conditions used in HX-MS. A detailed description of the standard in-exchange experimental method as well as the mass spectral data reduction procedures is found in Supporting Information 3. Peptide deuterium incorporations (DIs) for each condition was compiled into tables (DI^{raw}) together with the control values for fully deuterated control (DI^{fullD}). Back-exchange corrected DI numbers ($DI^{correct}$) were calculated by applying a back-exchange correction on the basis of the fully deuterated control as shown in eq 6, where no. ExNH is the number of observable NHs of a peptide. Further, all prolines were removed from consideration due to lack of NH.

$$DI_{fp}^{correct} = DI_{fp}^{raw}\left(\frac{\text{no. ExNH}}{DI_{fp}^{fullD}}\right) \quad (6)$$

The large amount of exchange data collected for FAS-TE necessitated automated data curation and cleaning procedures. The major consideration for filtering was consistency of observations. This was implemented by requiring less than 5% deviation between time point replicates and trending of the DI numbers for any given time course to increase with time for acceptability of the data.

Deconvolution of Experimental Peptide Exchange Rates into Rate Classes. To improve the exchange resolution beyond peptide resolution for more accurate comparisons of model and experiment, decomposition of peptide rates (eq 6) into the three rate classes (fast, medium, and slow) using a triexponential model (eq 7) was performed.[38] The constraints MaxDI = $A + B + C$ and positive values of all fitting parameters (A, B, and C representing the number of amides in each rate class and k_1, k_2, and k_3 the respective rate constants) were applied. Fitting was performed in R (R code available in Supporting Information 2) using nlsLM/nlsList functions for fitting the experimental in-exchange time points of each peptide to eq 7.

$$DI_{fp,t} = A(1 - e^{-k_1 t}) + B(1 - e^{-k_2 t}) + C(1 - e^{-k_3 t}) \quad (7)$$

RESULTS

To gauge the performance of our HX model based on MD snapshot (eq 3) counting and mapping of the statistics into a PF (eq 4), which was subsequently used for calculation of time-dependent DI numbers (eq 5), we first benchmarked the predicted DI numbers against those for a published DXCOREX protein set (Table 3). Performance benchmarking against a wider array of experimental conditions and time scales was then carried out in the context of extensive FAS-TE data, which we acquired. Based on the robust prediction results at the peptide level (Figure 2), we next decomposed (eq 7) deuterium uptake curves into HX rate classes (Figure 3) to demonstrate the ability to improve resolution.

DI Prediction for the DXCOREX Protein Set. Table 3 summarizes the correlation of calculated and experimental DIs obtained for HX modeling of published data using the method described (for base = 1×10^6 and NH-bond model HB3) together with correlations reported from published DXCOREX prediction results.[11] In most cases, our HX model shows higher correlation values than calculated from simple intrinsic rate predictions and those reported for DXCOREX calculations. Considering the diverse properties of these protein systems (apo/holo, wild-type/mutant, etc.; see Table 1) the observed strong correlations suggest robust predictive power of the method. The correlations of the intrinsic rate prediction are in fact surprisingly high and may be misleading given that they are calculated from the limited available data derived mostly from long in-exchange time points. This issue becomes clearer when the slope and abscissa intersect values are scrutinized. For an accurate prediction these values should be close to 1 and 0 respectively, which is clearly not the case for the majority of the k_{int} predictions.

HX-MS Experimental Results for FAS-TE. HX-MS data from the literature is limited in terms of the sequence coverage depth of the proteins studied as well as the range of experimental conditions (especially limited number of time points). As this limited the exploration of the predictive nature of the approach, it was decided to generate an extensive data set on FAS-TE spanning a large pD (= pH + 0.4) range and multiple temperatures. The hope was that this would provide the widest possible exchange range against which predictions could be made assuming the structural conformation is unaffected by a change in experimental conditions. FAS-TE was chosen for this purpose as it was readily available in-house, has diverse structural features, and is comparatively well behaved. The obtained sequence coverage depth of FAS-TE under optimal conditions is illustrated in Supporting Figure S2. The 283-residue protein was covered with 148 and 137 unique peptides at 25 and 0 °C respectively in searches of the raw tandem MS data against the protein sequence. The height of a histogram in the plots is the number of times any given residue of the protein was covered by a unique peptide observation and provides a measure of the coverage

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

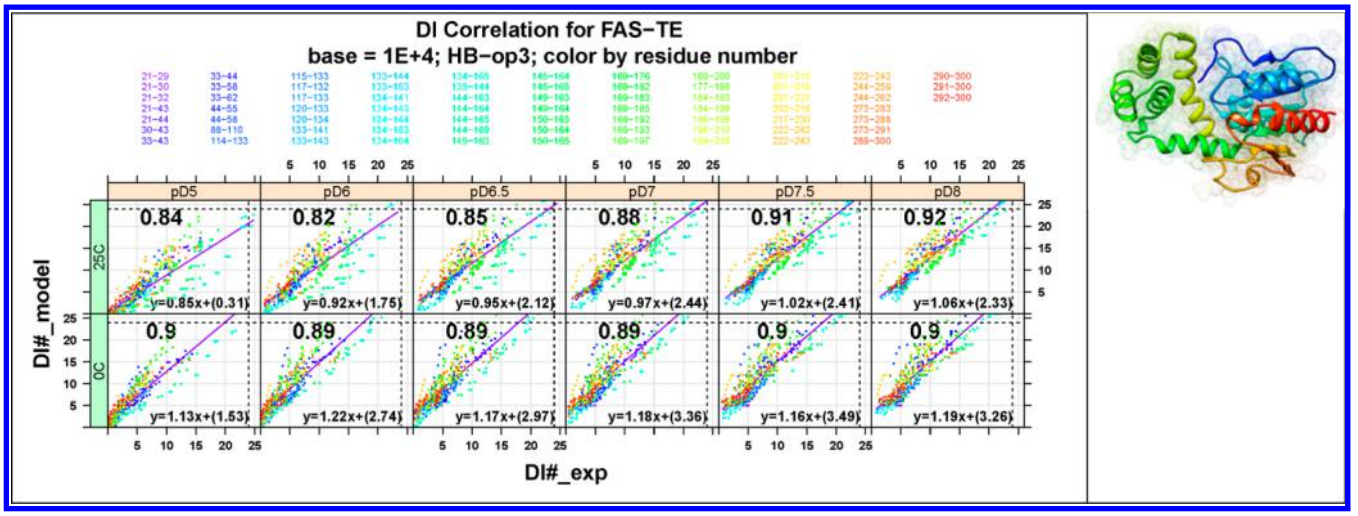


Figure 2. DI correlation for FAS-TE for all peptides (each color) at different time points (dots), temperature (each row), and pD conditions (each column), estimated by MD-based HX model (HB3). The correlation (correlation coefficient (R) see individual panels) from model predictions is improved for all conditions compared to the reference values based on intrinsic rate alone, which are 0.84, 0.77, 0.79, 0.82, 0.85, and 0.88 at 25 °C; 0.76, 0.68, 0.70, 0.73, 0.77, and 0.82 at 0 °C. Colors shown in the FAS-TE monomer ribbon structure with van der Waals radii (blue N-terminal to red C-terminal) are consistent with the residue index legend.

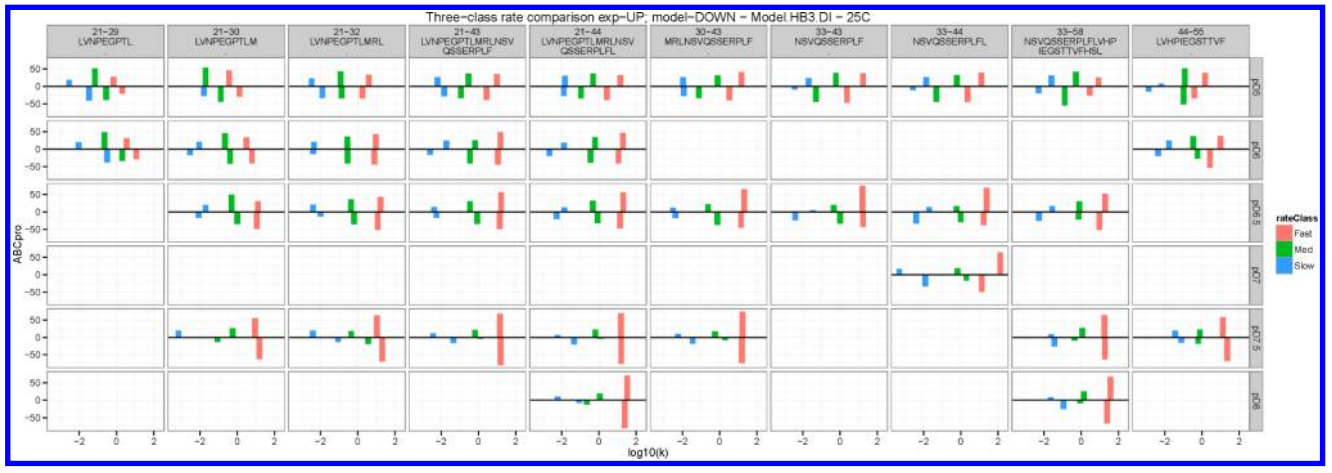


Figure 3. Three rate class decomposition of experimental (up) and predicted (down) deuterium incorporations for peptides of the N-terminal region (each column represents a fragmented peptide; each row represents different pD condition) of FAS-TE at 25 °C. Results for the full sequence are available in Figure S5.

density. The butterfly plots show the effects of automated data filtering to be substantial as coverage is significantly reduced post filtering. This is typical for HX-MS data sets as many of the peptide sequences derived from database searching cannot be confidently observed under all experimental conditions or reliable deuteration values are impossible to extract due to interference.

Figure 4 shows filtered in-exchange curves from the 25 °C data set for peptides selected to cover the protein sequence. Some curves show missing observations for some time points. These absences are either due to an inability to observe or extract the value or extracted values for the sample failing the filtering criteria. As expected, the magnitude of the observed exchange is strongly modulated by pH. A wide range of peptide exchange behavior is observed. For example, peptide 145−163 shows rapid saturation of exchange under almost all pD conditions. Peptide 169−200 reveals what looks to be mostly uniformly exchanging amides as indicated by the straight parallel lines, yet one of its sub sequence peptides, 184−198, reveals that the "true" exchange behavior in that sequence is much more complex.

DI Prediction for the FAS-TE. Comparison of all the experimental data for FAS-TE with predicted values is made in the form of correlation plots (Figure 2) capturing the experimental variables (pH, temperature, time) and modeling variables (k_{int} and base). The 25 °C temperature data (upper panels) shows a slight trend to higher correlation coefficient in the direction of increasing pD. Consistently higher correlations coefficients around 0.9 are obtained throughout all pD conditions at 0 °C (lower panels). The slope values show similar slight trending from significantly less than one at the lowest pD to close to one at the highest for the 25 °C series. For the 0 °C series consistent but elevated slope values around 1.17 (expected value 1) are observed for all pDs. Overall, correlations seem to improve with pH and temperature as judged by the trends in the correlation coefficients. This suggests systematic underestimation of the DI by the model at higher temperatures and low pD if it is assumed that experimental conditions do not modulate the structural ensemble of FAS-TE significantly.

On the other side the improvement in correlation at high pH and temperature can be interpreted as a consequence of the

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

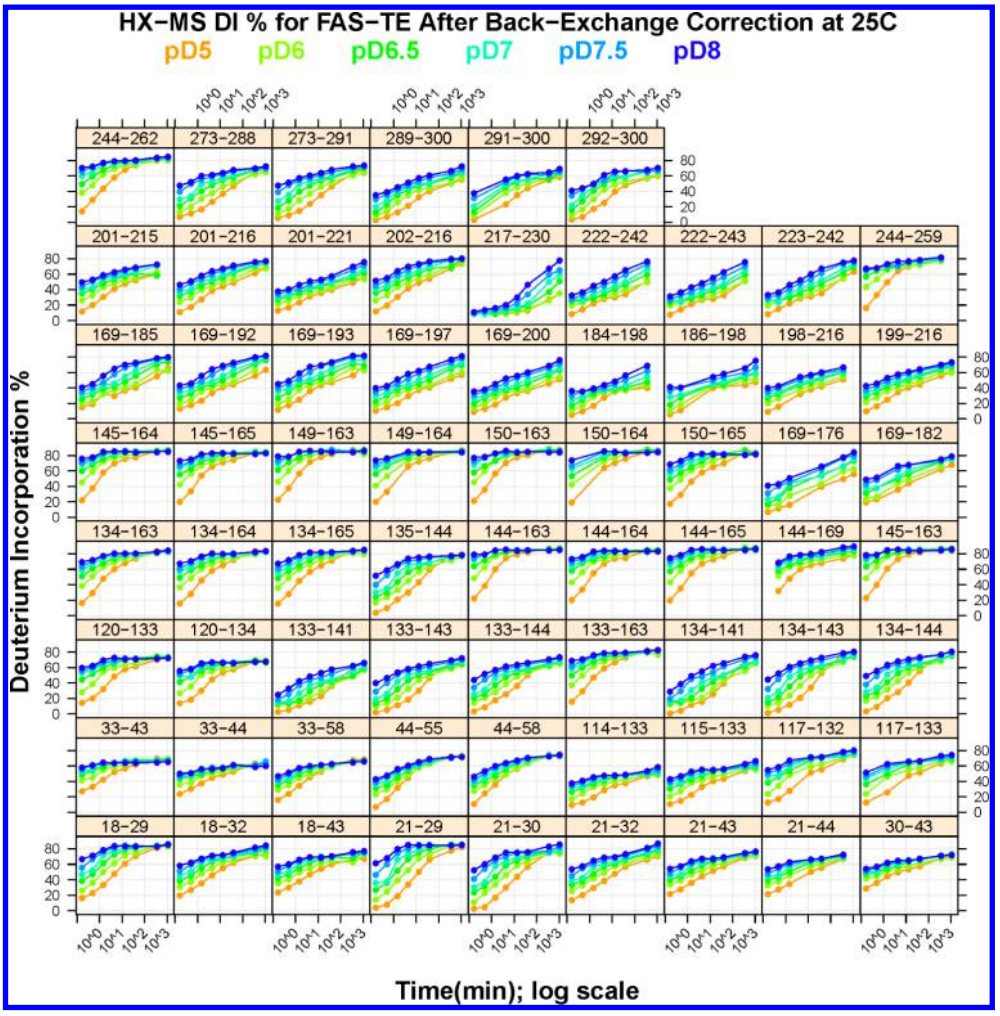


Figure 4. Comprehensive FAS-TE HX-MS experimental data at 25 °C (also see Supporting Figure S4 for 0 °C data). Each panel shows DI % as a function of time (10, 30, 270, and 810s; 40.5, 121.5, 864.5, and 1093.5 min) for six pD conditions (5.0, 6.0, 6.5, 7.0, 7.5, and 8.0).

banded nature of the predictions (the DI values saturate already at the shortest time points due to all PFs being 1; see Figure 4). For long in-exchange times many or most of the time points will be "accurately" predicted based on k_{int} as the number of saturated amides increases. This effect is magnified for predictions based on k_{int} alone as seen in Supporting Figure S3, so that at exchange saturation, correlation with predictions based on k_{int} alone will be quite accurate. This is also the reason why predictions based on k_{int} alone are in general meaningless (see also large *Y*-axis offsets in Supporting Figure S3) and that high correlation coefficients in DI space are not necessarily a meaningful measure of the predictive power of a model. Correlation of the peptide DI provides only a limited assessment of the accuracy of prediction of protection factors as peptide DI can be quite insensitive to the individual amides' PFs. The reasons for this are found in the summing over many PF_j that occur (eq 5), a comparatively large experimental error due to the inability to accurately correct for back-exchange (eq 6) and the large dynamic range of the PF from 1 to base (on the order of 1×10^4 to 1×10^6).

Decomposition of HX Rates for FAS-TE. Fortunately, the HX-MS experimental data collected for FAS-TE after data cleaning are of sufficient quality to allow decomposition into rate classes by fitting the experimental DI curves (Figure 4) with eq 7.[38] This increases the effective resolution of the experiment in amide and PF space as the DI curves calculated from predicted PFs on the basis of the HX model can be also decomposed in a similar fashion to the experimental ones. The improved resolution allows more detailed comparison between experiment and prediction and this should reveal more clearly if prediction actually represents the individual amides' contributions (and therefore PFs) accurately on an individual per amide basis or only in an average sense.

Figure 3 shows butterfly representations of overlapping and normalized comparisons of rate decompositions for experimental (up) and predicted (down) data for different pDs (for the full set see Supporting Figure S5 and S6 for 25 and 0 °C data, respectively). The three rates in eq 7 are plotted on the *X*-axis using a logarithmic scale as k_1 (fast in red), k_2 (medium in green), and k_3 (slow in sky blue). The *Y*-axis values are the relative number of amides in each rate class. For easier comparison, the values have been normalized and *A*, *B*, and *C* are plotted as the percentage of total amides of a peptide in a rate class. Each column represents a different peptide and each row the decomposition of that peptide at a different pD condition. Missing panels indicate that for a specific peptide at the specific temperature and pD condition, either no experimental data are available or the fitting was unsuccessful due to lack of data points or an inhomogeneous nature indicating a lack of trend.

Generally, a good agreement of the decomposition in terms of rate classes and the relative intensities of their contributions to

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

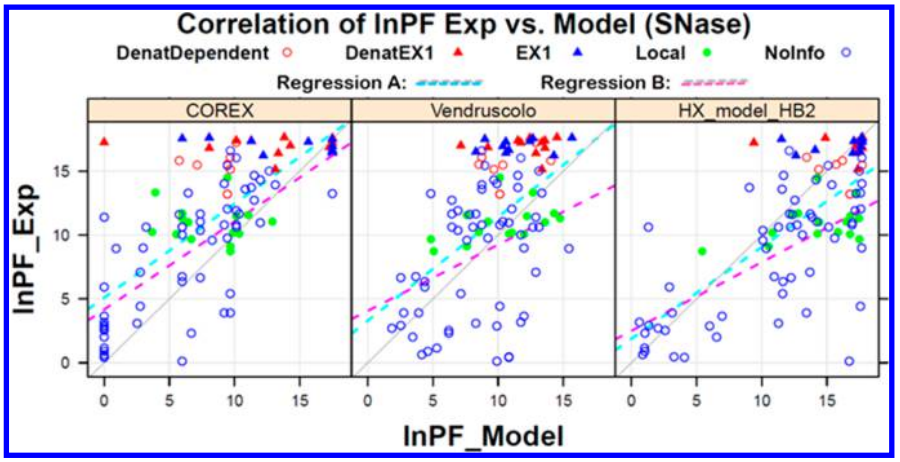


Figure 5. Correlation plots of experimental over predicted protection factors for SNase at pH 5.5 and 37 °C for three algorithms. Experimental and predicted values for COREX and Vendruscolo are replotted from ref 4 to compare with the model developed in this work (HB2, base = 1×10^8; see Table 2 for model definition). Color coding of data points follows categorization of exchange by ref 4. Regression A: full data set from NMR. Regression B: excluding denaturant dependent and/or EX1 classified values corresponding to experimental ln PF > 15.

Table 4. Regression Comparison of HX Prediction Models for SNase at pH 5.5 and 37 °C

model	COREX		Vendruscolo et al.		HB2, base = 1×10^8		Craig et al.	
data taken from	ref 7				this work		ref 12	
regression	A	B	A	B	A	B	A	B
regression line	$y = 0.73x + 5.10$	$y = 0.69x + 4.17$	$y = 0.81x + 3.29$	$y = 0.52x + 4.07$	$y = 0.72x + 1.87$	$y = 0.54x + 2.49$	$y = 0.68x + 1.8$[a]	
R	0.72	0.70	0.52	0.41	0.72	0.68	0.72	
$\langle \lvert \Delta \ln PF \rvert \rangle$	2.93	1.78	3.02	1.85	2.64	1.93	1.80[b]	

[a]Estimate from Figure 6 of ref 13. [b]Estimate from Figure 4 of ref 13.

the DI is observed for pDs of 6 and higher and trends between related peptides are consistent. This suggests that modeling accurately predicts individual protection factors whose contributions modulate the DI and that modeling can be used to assign individual amides in a peptide to a rate class, which effectively improves the resolution beyond the fragmentation level of the primary sequence.

Comparison with Other HX Prediction Algorithms. Protection factor prediction accuracy has been benchmarked in the literature by comparing predictions with experiment for model systems for which comprehensive sets of single amide resolved NMR data is available. Staphylococcal nuclease has been used as a preferred model protein and a comparison in the SNase system for three prediction algorithms is shown in Figure 5. The figure compares results from predictions of the COREX and Vendrusolo et al.'s algorithms as published in ref 4 with our model. A similar plot for an unknown subset of the data plotted in Figure 5 comparing predictions from the coarse-grained model developed in ref 13 with experiment can be found in that work. Regression parameters for all four models are summarized in Table 4 for ease of comparison. Inspection of the scatter patterns observed in the plots of Figures 5 and 6 panel C of ref 13 reveal pileups of predicted values at the scale extremes. In our model this is indicative of the amides in question being observed only in the closed state during the simulation. Therefore, these amides will be assigned the maximum PF (base/2). The issue cannot be resolved by increasing the base value of the model to allow for a larger maximum PF as this is equivalent to linearly rescaling of the ln PF_model axis in Figure 5. A more accurate treatment of the amides that are only observed in the closed state during simulation would be to eliminate them from consideration, as their expected PF is larger than what the scale allows. Indeed, improved correlation is observed if those amides are taken out of consideration ($R = 0.84$, $R^2 = 0.70$ vs $R = 0.72$, $R^2 = 0.51$ from Table 4). As this elimination would favor the predictive accuracy of our model over the others we did not make that adjustment and continued with the full set (all amides for which experimental values were reported) instead. A similar accumulation of data points occurs at the low end of the prediction scale in the COREX plot, which is traced back to the algorithms inability to predict random coil and surface exposed amides with any accuracy (see below).

Overall prediction performance of the models as assessed by the regression coefficient (R) and for all models as compiled in Table 4 seems comparable with the exception of Vendruscolo et al.'s, whose predictive ability appears to be poor. If the average of the absolute differences between predicted and experimental PF ($\langle \lvert \Delta \ln PF \rvert \rangle$) is used as a measure of prediction performance as suggested by Craig et al. then their coarse-grained model performs best (even so it is not fully clear if all data points are taken into consideration).[13] The coarse-grained model's average error factor of PF is about 6 ($\langle \lvert \Delta \ln PF \rvert \rangle = 1.8$); this is followed by an error factor of about twice that size for our model and another factor 2 larger error factor for the two remaining models. Removal of amides with experimental ln PF > 15 (regression B in Figure 5 and Table 4) results in slight reduction of the regression and correlation coefficients for COREX and our model and a substantial further degradation of the Vendruscolo model. Surprisingly, the opposite trend is observed for $\langle \lvert \Delta \ln PF \rvert \rangle$ values. All models show now comparable predictive performance based on the $\langle \lvert \Delta \ln PF \rvert \rangle$ measure. This suggests strong bias in this measure toward accurate prediction of high protection factors, which are those expected to show a comparatively larger error.

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

Potential correlation of prediction accuracy with alternate exchange pathways can be assessed from the coding of data points in Figure 5. Coding represents categorization of amide exchange behavior into denaturant dependence (red), the lack of denaturant dependence or local (green), if an EX2/EX1 transition is observable (triangle) at high pH, and uncategorized (blue circle).[4] As seen in the plots high protection factors (experimental ln PF values >15) are all categorized as denaturant dependent and/or showing an EX2/EX1 transition or both. If values with experimental ln PF > 15 are removed from consideration regression lines B in Figure 5 are obtained and overall slight derating of the regression parameters for all models is observed (Table 4). This is somewhat surprising as one would intuitively expect those values to be the most difficult ones to predict and therefore expect correlation to improve, which is not the case.

Supporting Figure S7 shows correlation plots of the data plotted in Figure 5 with amide PFs color coded by general structural environment. The structural environment is classified as class 1 for random coil/surface exposed loop (SASA of NH > 0 and no-secondary structure; red), class 2 exposed and structured (SASA of NH > 0 and secondary structure (alpha or beta); green), class 3 buried and structured (SASA of NH = 0 and secondary structure and more than 3 Å from protein surface; black), and class 4 all remaining (blue). The general underestimation of PFs by COREX already observed in Figure 5 (most data points above and to the left of the centerline) is now further differentiated. COREX does not predict random coil (class 1) and exposed amide PFs (class 2) with any accuracy, in addition buried and structured amide PFs (class 3) are underestimated too. The model by Vendruscolo strongly overestimates protection of exposed and structured amides (class 2) and underestimates somewhat class 3 amide PFs, which explains the overall poor predictive ability of this model. Our model overestimates buried and structured amide PFs (class 3) somewhat, but shows little structural bias overall.

DISCUSSION

Phenomenological HX Model Expressions. The main objective of the current work is to provide a comparatively simple means to predict deuterium incorporation levels of proteins that can be readily compared with values typically measured in fragmentation HX-MS experiments and thus provide guidance for enhanced structural and dynamic interpretation of results. We opted to pursue an approach based on MD simulation due to the ready availability of the tools and general applicability of MD to even large multiprotein systems and alternative environments (denaturing solutions, solid-state formulations, etc.). The literature suggests largely phenomenological modeling approaches for quantitative PF prediction from MD.[12,16] However, it is far from clear what factors or metric should be taken into consideration when optimizing a model for predictive accuracy.

Kieseritzky et al.,[16] following the phenomenological model of Vendruscolo et al.,[12] explored a wide range of PF elements/metrics, which are the Ns in equation ln PF = $\beta_c^{nb}N^c + \beta_h^{bond}N^h$ (see the Introduction for discussion) by optimizing the value of βs in the bacterial cytochrome C system. From their efforts, as well as Vendruscolo's, we concluded that (i) accurate PF modeling is quite challenging by MD or other methods, (ii) comparable results are achievable using a number of different metrics if optimization is performed, and (iii) optimization is likely required for individual proteins. The prediction accuracy of the MD based method for PFs seem to be worse if the comparison of the reported ΔG values[39] and casual inspection of the plot comparing computed and experimental NMR PFs for lysozyme is an indication (Supporting Figure S1 and S2 of ref 11). Despite the inaccuracies in the predicted PFs, a meaningful discrimination of structural models on the basis of calculated peptide in-exchange values/curves seems to be possible by the DXCOREX algorithm, which suggests even coarse PF predictions when used to predict in-exchange values at the peptide level might suffice to support structural interpretations.[11]

Hydrogen bonding is considered the characteristic feature of the folded state of proteins and, based on thermodynamic considerations, makes a significant contribution to the overall stability of proteins.[26] We surmised that PF prediction for backbone amides should be possible on the basis of intramolecular and intermolecular hydrogen bonding patterns observed during MD alone. Trying to avoid complex and protein specific optimization procedures attempts were made to derive a metric based on the number of snapshots an amide hydrogen is found forming a hydrogen bond with the protein backbone (or with additional protein H-bond acceptors in some modified models) and the number of snapshots were hydrogen bonding to water is observed. By analogy to the local unfolding model, amide hydrogen bonding to the protein backbone represents the "closed" or exchange incompetent state, whereas hydrogen bonding to solvent represents the "open" or exchange competent state in our model. After evaluation of various scale laws, it was found that the normalized difference of snapshots eq 3 could be mapped into a protection factor by a simple exponential function using a large base, where the large base value is equivalent to a maximum protection factor. For the purpose of convenience, we opted to use a logistic function eq 4 instead of a simple exponential. Our model is similar to the phenomenological expression used by Vendruscolo[12] as seen by comparing formulas in Table 5. The

Table 5. Comparison of Phenomenological Model Expressions Used in Reference 12 with the Current Work

ref 12	$PF_j = \exp(\beta_c N_j^c + \beta_h N_j^h)$
this work	$PF_j = \text{base}^{\left(\frac{-N_j^{NH:wat} + N_j^{NH:protein}}{N^{total}}\right)}$

simplicity of our model derives from the fact that β_c and β_h are predefined and do not need optimization as in the case of that previous work.

It is surprising that a simple model based on NH bonding state analysis has not been explored earlier, considering the analogy to the local unfolding model. Part of this might be due to the general attempt to derive models that parametrize the energetics/thermodynamics of the folding process through MD derivable quantities that positively correlate with exchange propensity and the early success of the lattice model formulation in explaining the general characteristics observed in the exchange behavior of globular proteins.[40] The lack of reliable estimates of the relative bond strength of intramolecular hydrogen bonds in proteins versus intermolecular hydrogen bonds to water, the absence of a clear energetic advantage of bonding to the backbone over bonding to solvent, and the inability to correlate amide hydrogen bond strength with exchange propensity might be other reasons why hydrogen bonding analysis has not been pursued extensively.[4,8,26,41]

Protein-Independent or Transferable HX Model. For calculation of the PF eq 4 is used. The functional form of eq 4 is

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

very close to that of the exponential function in Table 5 but allows the mapping of the full range of the exponent. Here "NHstat" ranges from −1 to 1 as defined through eq 3. Mapping into the PF scale of 1 (no protection) to base/2 (PF maximum) without further scaling or adjustment of the hydrogen bond statistics is facilitated by eq 4, which has only one parameter (the base of the exponential). The optimal value of base for the purpose of calculating peptide DI values and the universality or transferability of the value of base between proteins and/or experimental conditions needs to be explored. Supporting Figure S8 shows calculated DI percentage curves for hen egg-white lysozyme using different values of the base for the three models evaluated here (Table 2). Comparing the curves with the data plotted in Figure 1 of ref 42 suggests base values in the range of 1×10^5 to 1×10^6 for models HB1 and HB2 and a larger value of around 1×10^7 for HB3 as suitable. Correlation plots for the FAS-TE in-exchange data to be discussed below for base values of 1×10^4, 1×10^5, and 1×10^6 (Supporting Figure S9) produce the highest correlation coefficients for a base value of 1×10^4. The plots also indicate that the predictive power of the model is not very sensitivity to the value of base. This suggests that within the limitations inherent in our modeling approach a single base value is likely sufficient for peptide in-exchange prediction for different proteins and experimental conditions.

Comparative Assessment of Model Prediction Performance. The comparisons of predictions from COREX, Vendruscolo's, Craig's, and our model for SNase demonstrated that no single model provides a superior approach. A conformational sampling, structure partitioning/combination method was used in COREX; NMR restraint-guided Monte Carlo simulation was used in Vendruscolo's model; coarse-grain MD with umbrella sampling was used in Craig's model; and all-atom MD simulation with explicit water solvation was used in our model. Extensive sampling is important for accurate free energy calculation; however considering the results it is not clear if it is necessary to sample large-scale unfolding as the high energy barrier (i.e., low probability) makes its contribution to PF calculation mostly insignificant. It appears to be more important to sample various local unfolded states mediated by explicit water molecules accurately, as the majority of amides measured in SNase by NMR belongs to the buried/structured class and even the exposed amides keep local interaction with a structural motif resulting in moderate PF.[8] Therefore, extensive sampling covering more of the local unfolding space with more advanced explicit water model might be able to discriminate accurately among the relatively high PF values. Such an approach would also be in line with the assessment made from the experimentalist's side that structural detail at the individual amide level should be taken into account in the interpretation of hydrogen exchange.[4,8]

Our HX model was able to capture various locally unfolded states presumably because explicit water dynamics was included in our conformational sampling in comparison with wider sampling space methods such as COREX and the coarse-grained model. Building on this finding and the demonstrated distribution of experimental protection factors classified by Skinner et al.[8] by type of H-bond acceptor (exposed random coil, exposed on the structure, internal water, side-chain and backbone) one should be able to build more optimal open/closed state definitions. Delicate consideration of the structural environment can be incorporated into our model and it needs to be seen if enhanced discrimination and improved prediction accuracy can be achieved that way. Another avenue of exploration is provided by the increased understanding of protein solvation and that water around proteins can be divided into bulk water and protein-bound water (individually bound water in the cavity and hydration water on the surface).[43] Currently, if water is an H-bond acceptor, our model consider it as in the "open" state. Exploration of advanced water models and classification into protein bound water provide further opportunities for exploration as this provides a direct means of manipulating the hydrogen bond statistic calculation on which PF prediction of our model is based.

General Pitfalls of the Model and Computational Approach. Analysis of the snapshot statistics from MD for FAS-TE reveals non-hydrogen bonded conformations are a common occurrence though they do not dominate with the exception of an eight residue sequence (Figure S10) found buried inside the core. For the purpose of exchange prediction the non-hydrogen bonding snapshots for these residues were counted as closed or protected. In all other cases non-hydrogen bonded snapshots were not taken into consideration in calculating NHstat. From a modeling standpoint the question arises if improved prediction results could be obtained by accounting for these nonbonded snapshots by either counting them in the respective bonded or nonbonded pools or taking them into account through an expanded metric. We have seen little change in correlations between experiment and predictions when counting nonbonding snapshots to bonded or nonbonded pools in the various models (data not shown). This is not surprising, considering the general insensitivity of peptide exchange predictions toward the magnitude of any individual protection factor, so we would not predict an improvement in predictions from an expanded model, though this was not explored further.

As in all other MD based studies issues relating to the limited time scale of such simulations persist. The longest time scale of typical MD simulations as used here, were 50−100 ns. This limits representative sampling of the conformational space to conformers close to the starting model. This is in stark contrast to the typical experimental HX-MS time scales, which range from seconds to days. However, it should be noted that even during the long time scales explored in typical HX-MS experiments, EX2 kinetics is found to be descriptive of most of the observed exchange behavior. This can be taken as an indication that global unfolding or conformers that require large scale unfolding are largely negligible contributors to exchange. We would not expect our model to perform well for sequences showing significant EX1 kinetic behavior. Our assumption is that explicit water solvated MD simulation on the tens of nanosecond time scale will be appropriate to provide a description of the average of the relevant "local unfolding ensemble" (EX2) in contrast to global unfolding (EX1), which is outside the scope of the model. Further, the hope is that this average description will be accurate despite significant limitations arising from sparse sampling of a conformational restricted space around the fully folded conformer.[6] Extension of the MD simulation time scale alone is not expected to improve the prediction as the conformational space explored will likely not expand.[44] Specialized methods to overcome energetic barriers might remedy this.[13,45,46] For the purpose of PF prediction the weighing of contributions from MD simulations of various known stable conformations might be sufficient and should be explored in the future.

In this study the popular ff99SB Amber force field combined with a TIP3P water model was used to quantify solvent−solute interactions. Other force field/water model combinations providing altered solute−solvent potentials might result in different NHstat values (eq 3) and consequently different PFs (eq 4).

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

This might affect the correlation with experimental data via eq 5. It is again the opinion of the authors that the finer details of the force field/solvation model will not significantly affect the prediction results for reasons of the accuracy of the model and the general insensitivity of peptide exchange prediction to individual amide PFs as already mentioned. Further, others have explored the subject of force field solvation model combinations in other contexts and the combination used here was found to perform close to optimal.[47−49]

We have tried to minimize the impact of experimental measurement errors by stringent "data cleaning" applied to our FAS-TE HX-MS data set, which had to be done in automated fashion due to the size of the data set. It is clear that even with stringent data cleaning, systematic errors persist like those resulting from back-exchange correction, which is only approximate. Nevertheless, it is unlikely that experimental uncertainty limits the degree of correlation observed between experiment and prediction considering that PF prediction accuracy errors are approximately two logs. The plots of rate decompositions of experimental data and prediction in Figure 3 (see also Supporting Figure S5 for additional data) show good agreement for a large fraction of peptides and exchange conditions. This is taken as validation of the approach and indication that accuracy is sufficient to provide a means of improving the sequence resolution of HX-MS.

Potential Alternate Use of the Defined Energy Function. Equation 2 provides a means of relating our empirical energy function through the protection factor back to the change of the Gibbs free energy. As such, the described hydrogen-bonding analysis approach can be used to calculate $\Delta\Delta G$ values (eq 8), which can be used for example for computational mutagenesis to estimate the relative stability of mutants.

$$\begin{aligned}\Delta\Delta G(\mathrm{wt}-\mathrm{mut}) &= \Delta G(\mathrm{wt}) - \Delta G(\mathrm{mut}) \\ &= RT\sum_j \ln \mathrm{PF}_j(\mathrm{wt}) - RT\sum_j \ln \mathrm{PF}_j(\mathrm{mut}) \\ &= RT\sum_j \ln \frac{\mathrm{PF}_j(\mathrm{wt})}{\mathrm{PF}_j(\mathrm{mut})}\end{aligned} \tag{8}$$

We tested this approach in simulations of 10 mutants of hen white-egg lysozyme (PDB code: 4LYZ) for which extensive mutagenesis experimental data are available in the literature.[49] The resulting $\Delta\Delta G$ correlation coefficient was 0.87 (data not shown). This provides secondary confirmation that our empirical energy function captures the underlying physics well and shows the potential for expansion into other applications, which are currently under exploration.

CONCLUSION

Analysis of hydrogen bonding patterns from MD snapshots was demonstrated to be a suitable metric for the estimation of protection factors. The approach appears to be generic and translatable to other systems, as protein specific optimization procedures are not required. The presented data suggest that the empirical energy function based on exponential mapping of the hydrogen bond statistics into a protection factor captures the underlying physics accurately. This approach is easily implemented by others due to its simplicity and it is expected to be highly valuable in the interpretation of HX data on the basis of available structural data and models. Besides being simple the model lends itself to modification so that more accurate descriptions of the modulating structural environment and of the solvent can be taken into account. Lastly, as the protection factor has a direct thermodynamic interpretation, the approach is likely of value in other applications and can be extended to address other problems or complement other computation tools.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00185.

(1) Intrinsic rate calculation implemented in Python. (2) Deuterium incorporation curve fitting in R script. (3) FAS-TE HX-MS experimental procedure. Figure S1−S10; Table S1−S4; HX-MS experimental data from the literature (PDF)

Experimental data summary at 0 °C (CSV)

Experimental data summary at 25 °C (CSV)

AUTHOR INFORMATION

Corresponding Author

*E-mail: abrock@gnf.org. Phone: 858-812-1549.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Tong Liu for providing technical insight into the DXCOREX algorithm and Dr. Virgil Woods, who could not see this work come to fruition due to an untimely death, for motivation to pursue this research endeavor. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH), Award Number **U54 GM094586**. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ABBREVIATIONS

HX, hydrogen exchange; MS, mass spectrometry; NH, amide hydrogen; H-bond/HB, hydrogen bond; PF, protection factor; MD, molecular dynamics; DI, deuterium incorporation

REFERENCES

(1) Zhang, Q.; Willison, L. N.; Tripathi, P.; Sathe, S. K.; Roux, K. H.; Emmett, M. R.; Blakney, G. T.; Zhang, H. M.; Marshall, A. G. Epitope Mapping of a 95 kDa Antigen in Complex with Antibody by Solution-Phase Amide Backbone Hydrogen/Deuterium Exchange Monitored by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2011**, *83*, 7129−7136.

(2) Brock, A. Fragmentation Hydrogen Exchange Mass Spectrometry: A Review of Methodology and Applications. *Protein Expression Purif.* **2012**, *84*, 19−37.

(3) Mendoza, V. L.; Vachet, R. W. Probing Protein Structure by Amino Acid-Specific Covalent Labeling and Mass Spectrometry. *Mass Spectrom. Rev.* **2009**, *28*, 785−815.

(4) Skinner, J. J.; Lim, W. K.; Bedard, S.; Black, B. E.; Englander, S. W. Protein Hydrogen Exchange: Testing Current Models. *Protein Sci.* **2012**, *21*, 987−95.

(5) Maity, H.; Lim, W. K.; Rumbley, J. N.; Englander, S. W. Protein Hydrogen Exchange Mechanism: Local Fluctuations. *Protein Sci.* **2003**, *12*, 153−160.

(6) Englander, S. W.; Kallenbach, N. R. Hydrogen-Exchange and Structural Dynamics of Proteins and Nucleic-Acids. *Q. Rev. Biophys.* **1983**, *16*, 521−655.

(7) Fitzkee, N. C.; Torchia, D. A.; Bax, A. Measuring Rapid Hydrogen Exchange in the Homodimeric 36 kDa HIV-1 Integrase Catalytic Core Domain. *Protein Sci.* **2011**, *20*, 500−12.

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925

(8) Skinner, J. J.; Lim, W. K.; Bedard, S.; Black, B. E.; Englander, S. W. Protein Dynamics Viewed by Hydrogen Exchange. *Protein Sci.* **2012**, *21*, 996−1005.
(9) Milne, J. S.; Mayne, L.; Roder, H.; Wand, A. J.; Englander, S. W. Determinants of Protein Hydrogen Exchange Studied in Equine Cytochrome c. *Protein Sci.* **1998**, *7*, 739−45.
(10) Wang, J. M.; Hou, T. J. Develop and Test a Solvent Accessible Surface Area-Based Model in Conformational Entropy Calculations. *J. Chem. Inf. Model.* **2012**, *52*, 1199−1212.
(11) Liu, T.; Pantazatos, D.; Li, S.; Hamuro, Y.; Hilser, V. J.; Woods, V. L. Quantitative Assessment of Protein Structural Models by Comparison of H/D Exchange MS Data with Exchange Behavior Accurately Predicted by DXCOREX. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 43−56.
(12) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. Rare Fluctuations of Native Proteins Sampled By Equilibrium Hydrogen Exchange. *J. Am. Chem. Soc.* **2003**, *125*, 15686−15687.
(13) Craig, P. O.; Latzer, J.; Weinkam, P.; Hoffman, R. M.; Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Prediction of Native-State Hydrogen Exchange from Perfectly Funneled Energy Landscapes. *J. Am. Chem. Soc.* **2011**, *133*, 17463−72.
(14) Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−50.
(15) Bouzida, D.; Kumar, S.; Swendsen, R. H. Efficient Monte Carlo Methods for the Computer Simulation of Biological Molecules. *Phys. Rev. A: At., Mol., Opt. Phys.* **1992**, *45*, 8894−8901.
(16) Kieseritzky, G.; Morra, G.; Knapp, E. W. Stability and Fluctuations of Amide Hydrogen Bonds in a Bacterial Cytochrome c: A Molecular Dynamics Study. *JBIC, J. Biol. Inorg. Chem.* **2006**, *11*, 26−40.
(17) Petruk, A. A.; Defelipe, L. A.; Limardo, R. G. R.; Bucci, H.; Marti, M. A.; Turjanski, A. G. Molecular Dynamics Simulations Provide Atomistic Insight into Hydrogen Exchange Mass Spectrometry Experiments. *J. Chem. Theory Comput.* **2013**, *9*, 658−669.
(18) Garcia, A. E.; Hummer, G. Conformational Dynamics of Cytochrome c: Correlation to Hydrogen Exchange. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 175−191.
(19) Ma, B. Y.; Nussinov, R. Polymorphic Triple beta-Sheet Structures Contribute to Amide Hydrogen/Deuterium (H/D) Exchange Protection in the Alzheimer Amyloid beta 42 Peptide. *J. Biol. Chem.* **2011**, *286*, 34244−34253.
(20) Sljoka, A.; Wilson, D. Probing Protein Ensemble Rigidity and Hydrogen-Deuterium Exchange. *Phys. Biol.* **2013**, *10*, 056013.
(21) Resing, K. A.; Hoofnagle, A. N.; Ahn, N. G. Modeling Deuterium Exchange Behavior of ERK2 Using Pepsin Mapping to Probe Secondary Structure. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 685−702.
(22) Arrington, C. B.; Teesch, L. M.; Robertson, A. D. Defining Protein Ensembles with Native-State NH Exchange: Kinetics of Interconversion and Cooperative Units from Combined NMR and MS Analysis. *J. Mol. Biol.* **1999**, *285*, 1265−1275.
(23) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(24) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79*, 351−71.
(25) Baker, E. N.; Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97−179.
(26) Myers, J. K.; Pace, C. N. Hydrogen Bonding Stabilizes Globular Proteins. *Biophys. J.* **1996**, *71*, 2033−9.
(27) Jacobs, D. J. Ensemble-Based Methods for Describing Protein Dynamics. *Curr. Opin. Pharmacol.* **2010**, *10*, 760−9.
(28) Baldwin, R. L. In Search of the energetic Role of Peptide Hydrogen Bonds. *J. Biol. Chem.* **2003**, *278*, 17581−17588.
(29) Mitchell, J. B. O.; Price, S. L. On the Relative Strengths of Amide··· Amide and Amide··· Water Hydrogen-Bonds. *Chem. Phys. Lett.* **1991**, *180*, 517−523.
(30) Pemble, C. W. t.; Johnson, L. C.; Kridel, S. J.; Lowther, W. T. Crystal Structure of the Thioesterase Domain of Human Fatty Acid Synthase Inhibited by Orlistat. *Nat. Struct. Mol. Biol.* **2007**, *14*, 704−9.
(31) Truhlar, S. M. E.; Torpey, J. W.; Komives, E. A. Regions of IkBa that are Critical for its Inhibition of NF-kB·DNA Interaction Fold upon Binding to NF-kB. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 18951−18956.
(32) Full, S. J.; Deinzer, M. L.; Ho, P. S.; Greenwood, J. A. Phosphoinositide Binding Regulates Alpha-Actinin CH2 Domain Structure: Analysis by Hydrogen/Deuterium Exchange Mass Spectrometry. *Protein Sci.* **2007**, *16*, 2597−2604.
(33) Kornhaber, G. J.; Tropak, M. B.; Maegawa, G. H.; Tuske, S. J.; Coales, S. J.; Mahuran, D. J.; Hamuro, Y. Isofagomine Induced Stabilization of Glucocerebrosidase. *ChemBioChem* **2008**, *9*, 2643−2649.
(34) Molnar, K. S.; Karabacak, N. M.; Johnson, J. L.; Wang, Q.; Tiwari, A.; Hayward, L. J.; Coales, S. J.; Hamuro, Y.; Agar, J. N. A Common Property of Amyotrophic Lateral Sclerosis-Associated Variants Destabilization of the Copper/Zinc Superoxide Dismutase Electrostatic Loop. *J. Biol. Chem.* **2009**, *284*, 30965−30973.
(35) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−12.
(36) Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. Primary Structure Effects on Peptide Group Hydrogen Exchange. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 75−86.
(37) Cellitti, S. E.; Jones, D. H.; Lagpacan, L.; Hao, X.; Zhang, Q.; Hu, H.; Brittain, S. M.; Brinker, A.; Caldwell, J.; Bursulaya, B.; Spraggon, G.; Brock, A.; Ryu, Y.; Uno, T.; Schultz, P. G.; Geierstanger, B. H. In Vivo Incorporation of Unnatural Amino Acids to Probe Structure, Dynamics, and Ligand Binding in a Large Protein by Nuclear Magnetic Resonance Spectroscopy. *J. Am. Chem. Soc.* **2008**, *130*, 9268−81.
(38) Zhang, Z.; Smith, D. L. Determination of Amide Hydrogen Exchange by Mass Spectrometry: A New Tool for Protein Structure Elucidation. *Protein Sci.* **1993**, *2*, 522−31.
(39) Hilser, V. J.; Freire, E. Structure-based Calculation of the Equilibrium Folding Pathway of Proteins. Correlation with Hydrogen Exchange Protection Factors. *J. Mol. Biol.* **1996**, *262*, 756−72.
(40) Miller, D. W.; Dill, K. A. A Statistical Mechanical Model for Hydrogen Exchange in Globular Proteins. *Protein Sci.* **1995**, *4*, 1860−73.
(41) Eberhardt, E. S.; Raines, R. T. Amide-Amide and Amide-Water Hydrogen Bonds: Implications for Protein Folding and Stability. *J. Am. Chem. Soc.* **1994**, *116*, 2149−2150.
(42) Chung, E. W.; Nettleton, E. J.; Morgan, C. J.; Gross, M.; Miranker, A.; Radford, S. E.; Dobson, C. M.; Robinson, C. V. Hydrogen Exchange Properties of Proteins in Native and Denatured States Monitored by Mass Spectrometry and NMR. *Protein Sci.* **1997**, *6*, 1316−24.
(43) Chen, X.; Weber, I.; Harrison, R. W. Hydration Water and Bulk Water in Proteins have Distinct Properties in Radial Distributions Calculated from 105 Atomic Resolution Crystal Structures. *J. Phys. Chem. B* **2008**, *112*, 12073−80.
(44) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. O.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 139−44.
(45) Atzori, A.; Bruce, N. J.; Burusco, K. K.; Wroblowski, B.; Bonnet, P.; Bryce, R. A. Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics. *J. Chem. Inf. Model.* **2014**, *54*, 2764−2775.
(46) Pan, A. C.; Weinreich, T. M.; Shan, Y. B.; Scarpazza, D. P.; Shaw, D. E. Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice. *J. Chem. Theory Comput.* **2014**, *10*, 2860−2865.
(47) Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M. Explicit Water Models Affect the Specific Solvation and Dynamics of Unfolded Peptides While the Conformational Behavior and Flexibility of Folded Peptides Remain Intact. *J. Chem. Theory Comput.* **2010**, *6*, 3569−3579.
(48) Wickstrom, L.; Okur, A.; Simmerling, C. Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009**, *97*, 853−856.
(49) Shih, P.; Kirsch, J. F. Design and Structural-Analysis of an Engineered Thermostable Chicken Lysozyme. *Protein Sci.* **1995**, *4*, 2063−2072.

DOI: 10.1021/acs.jcim.5b00185
J. Chem. Inf. Model. 2015, 55, 1914−1925