JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

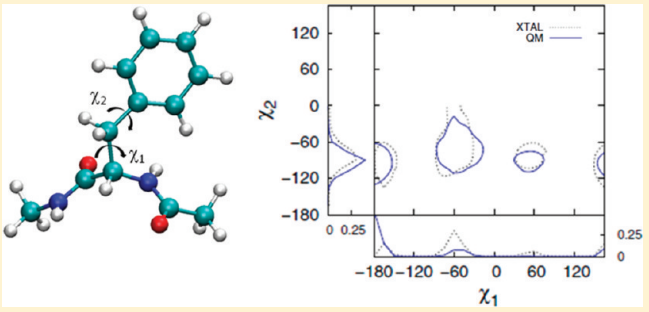
Intrinsic Energy Landscapes of Amino Acid Side-Chains

Xiao Zhu, Pedro E.M. Lopes, Jihyun Shim, and Alexander D. MacKerell, Jr.*

Department of Pharmaceutical Sciences, University of Maryland, School of Pharmacy, 20 Penn Street HSFII, Baltimore, Maryland 21201, United States

Supporting Information

ABSTRACT: Amino acid side-chain conformational properties influence the overall structural and dynamic properties of proteins and, therefore, their biological functions. In this study, quantum mechanical (QM) potential energy surfaces for the rotation of side-chain χ_1 and χ_2 torsions in dipeptides in the alphaR, beta, and alphaL backbone conformations were calculated. The QM energy surfaces provide a broad view of the intrinsic conformational properties of each amino acid side-chain. The extent to which intrinsic energetics dictates side-chain orientation was studied through comparisons of the QM energy surfaces with χ_1 and χ_2 free energy surfaces from probability distributions obtained from a survey of high resolution crystal structures. In general, the survey probability maxima are centered in minima of the QM surfaces as expected for sp^3 (or sp^2 for χ_2 of Asn, Phe, Trp, and Tyr) atom centers with strong variations between amino acids occurring in the energies of the minima indicating intrinsic differences in rotamer preferences. High correlations between the QM and survey data were found for hydrophobic side-chains except Met, suggesting minimal influence of the protein and solution environments on their conformational distributions. Conversely, low correlations for polar or charged side-chains indicate a dominant role of the environment in stabilizing conformations that are not intrinsically favored. Data also link the presence of off-rotamers in His and Trp to favorable interactions with the backbone. Results also suggest that the intrinsic energetics of the side-chains of Phe and Tyr may play important roles in protein folding and stability. Analyses on whether intrinsic side-chain energetics can influence backbone preference identified a strong correlation for residues in the alphaL backbone conformation. It is suggested that this correlation reflects the intrinsic instability of the alphaL backbone such that assumption of this backbone conformation is facilitated by intrinsically favorable side-chain conformations. Together our results offer a broad overview of the conformational properties of amino acid side-chains and the QM data may be used as target data for force field optimization.



INTRODUCTION

The primary sequence of a protein largely dictates the ensemble of secondary, tertiary, and quaternary interactions, which work in concert with environmental factors to modulate a protein's biological functions. Central to the role of the primary sequence in determining tertiary and quaternary structure are the properties of the amino acid side-chains. As observed in the first globular proteins for which 3D structures were resolved, hemoglobin, and myoglobin,[1,2] side-chain interactions between adjacent helices stabilize the helix-bundle topology and allows formation of the functional tetramer in hemoglobin.[3] Accordingly, understanding the physical properties of side-chains, including their conformational properties, is essential to understanding the relationship of structure and function in proteins.

The conformations of amino acid side-chains are influenced by both their intrinsic conformational energies and by interactions with the surrounding environment. While conformations of side-chains in folded structures can be determined using NMR and X-ray crystallography, quantification of the relative contributions of intrinsic versus environmental factors to the observed conformations of amino acids using experimental methods alone is difficult. Alternatively, the intrinsic energetics may be investigated using computational methods, which include quantum mechanics (QM). QM data can be used to help investigate the energy landscape of side-chains in the absence of environmental effects, allowing the intrinsic contributions to observed conformational preferences of side-chains to be determined.

To date, a range of studies have used computational methods to investigate the conformational properties of amino acid side-chains. Side-chain conformations in bovine pancreatic trypsin inhibitor (BPTI) were extensively analyzed by Gelin & Karplus[4] using empirical potential energy calculations on model dipeptides more than three decades ago. Previous studies have linked intrinsic energies to occurrences of rotamers in protein structures.[5,6] Of particular interest was the study of Butterfoss and Hermans. In their work, it was shown that the observed rotamer frequencies for Met and Lys correlate well with the estimated populations obtained by Boltzmann weighting of the torsion energies derived from QM calculation

Received: February 8, 2012

   dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

on the model compounds ethylmethyl sulfide and butane, respectively. However, the above studies involved calculations of the χ_1 and χ_2 dihedrals using empirical force field evaluations of dipeptides or QM calculations on model compounds that lack the peptide backbone. In the present study, we extend those efforts by employing QM methods to provide a more accurate and complete picture of the relationship between intrinsic energetics and conformational properties of side-chains in proteins.

In this work, backbone-dependent energy landscapes are calculated for the amino acid side-chain χ_1 and χ_2 dihedral angles using QM methods on dipeptides. Performing the calculations on dipeptide representations of the amino acids yields profiles for the intrinsic conformational energies where only local interactions of the side-chain with the adjacent peptide bonds of the backbone can occur. The QM results are then compared with observed distributions for χ_1 or χ_2 obtained from a survey of the RCSB Protein Data Bank (PDB)[7] to determine the degree by which intrinsic energies contribute to the prevalence of side-chain conformations in crystal structures. In addition, results from the presented QM calculations can serve as a set of reference data for the refinement of side-chain torsion parameters in empirical force fields.[8,9]

METHODS

QM Calculations. Molecular mechanics calculations were used to generate starting geometries for the χ_1/χ_2 1- and 2-dimensional (2D) QM potential energy surface (PES) calculations. N-acetylated and N′-methylamidated dipeptides were first constructed with the program CHARMM[10] and optimized with the CHARMM22/CMAP force field[11,12] with the following methodology. Backbone ϕ/ψ angles were restrained at the alphaR (−60.0/−45.0), beta (−120.0/120.0), or alphaL (63.5/34.8) conformations. Initial geometries were generated with χ_1 restrained from 0 to 360° in 15° increments with χ_2 harmonically restrained at 180° with a force constant of 10^4 kcal/mol/rad^2. In order to limit the dimensionality of the conformational space for the longer amino acids the remaining χ torsions were restrained to the following values based on a previous rotamer study:[13] χ_3 = 0° for Gln, 180° for Glu, 67° for Met, $\chi_{3,4}$ = 180° for Lys, and $\chi_{3,4,5}$ = 180° for Arg. A harmonic force constant of 10^4 kcal/mol/rad was used in all cases. Each generated conformation was energy minimized for 500 steps of conjugate-gradient (CONJ) followed by 500 steps of adopted basis Newton−Raphson (ABNR), each with a convergence criteria of the gradient <10^{-5} kcal/mol/Å. Coordinates of the optimized geometries were then used to initiate the QM χ_1 and χ_2 scans.

QM PES calculations were performed starting with the geometries from the MM calculations. With Val 1D torsion scans were performed about χ_1, while 2D χ_1 and χ_2 scans were performed for the remaining studied amino acids. 2D scans involved a series of 1D torsion scans for χ_2 with constrained ϕ, ψ, and χ_1 kept at its initial on-grid values of 15° increments. χ_3, χ_4, and χ_5 were constrained to values listed above for the larger side-chains. In the individual 1D scans χ_2 was incremented in 15° steps from the −180° conformation for a full 360° rotation about the torsion. Calculations were performed using the GAUSSIAN 03[14] program with the 6-31G*[15] and 6-31+G*[16] basis sets, with the latter used for Glu and Asp, at the MP2[17] level of theory. During the scans all nonconstrained degrees of freedom were allowed to relax to the default convergence criteria. Because several conformations contain steric clashes that led to problems with convergence, reverse scans were performed and data from the forward and reverse scans combined by selecting the lowest energy conformations to yield the complete 1D χ_2 surfaces. In a number of cases, the QM optimization failed to converge. This was generally caused by "flat" local energy regions such that the minimizer algorithm "hovers" around the true minimum but never achieves the convergence criteria. In such cases, the average of the coordinates from the last three optimization steps was used as the input geometry for an additional QM energy minimization. The use of the average structure was designed to move the molecule out of the quasi-minimum geometry, thereby allowing the molecule to relax into the correct minimum. Single point energy calculations of the optimized geometries were performed with the program QCHEM[18] at the RIMP2/cc-pVTZ[19,20] level of theory. The final RIMP2 energies were extracted for analysis. The energy surface for each side-chain was offset to its respective global minimum. With redundant points removed, this procedure yields 576 data points for each backbone conformation, yielding a total of 1728 conformations per residues except Val, for which 3 × 24 = 72 conformations were obtained. The energies and conformations from the QM calculations for all the studied amino acids may be access from the MacKerell web page at http://mackerell.umaryland.edu/MacKerell_Lab.html.

Protein Structure Survey. All protein structures were obtained from the RCSB Protein Data Bank[7] and only structures with a resolution ≤1.5 Å were considered. A sequence similarity filter was applied using the RCSB's PDB query functionality with a 50% BlastClust[21] sequence identity threshold to avoid systematic bias due to the presence of redundant proteins. Any residue with a side-chain that had one or more missing atoms was discarded. Initially, the database was processed with the program REDUCE,[52] which corrects for side-chain orientation of His, Asn, Asp, and Gln based on hydrogen-bonding contacts. To further improve the quality of the data included in the analysis χ entries from amino acids that contain at least one atom with B-factor >40 and with occupancy ≠1.0 were removed. In addition, amino acid residues with a B-factor of 0.0 were discarded as that value is typically assigned due to the unavailability of the B-factors. To minimize the possibility of errors from incorrect B-factors amino acids that involve one or more atoms with B-factor <1 were removed. Furthermore, only the first occurrence of duplicate residue entries, which were rare, was included. As hydrogens are not visible in protein X-ray structures only χ_1 data is available for Cys, Ser, and Thr, though the 2D QM surfaces were obtained for these amino acids. PDB entries with nonstandard residue names were also removed. The final survey results were also parsed as a function of secondary structure (alphaR, beta, and alphaL) based on secondary structure assignment determined by the program STRIDE[22] for alphaR and beta or based phi/psi values for alphaL as described below. A PDB survey statistics is presented in Table S1 of Supporting Information.

Definitions of Dihedral Angles. The χ torsion angles were defined as follows: χ_1 = N−Cα-Cβ-Xγ and χ_2 = Cα-Cβ-Xγ-Xδ where X is a side-chain-dependent atom type. In Asn, χ_2 is defined as Cα-Cβ-Cγ-Nδ. A rotamer is a conformer of the side-chain defined by a specific combination of torsion angles. Thus, each torsion angle was classified into three generic orientations annotated as gauche$^-$ (**m**) at −60°, gauche$^+$ (**p**) at +60°, and anti (**t**) at ±180°, a nomenclature employed by Lovell et al.[23] for the construction of rotamer libraries. For Asn

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

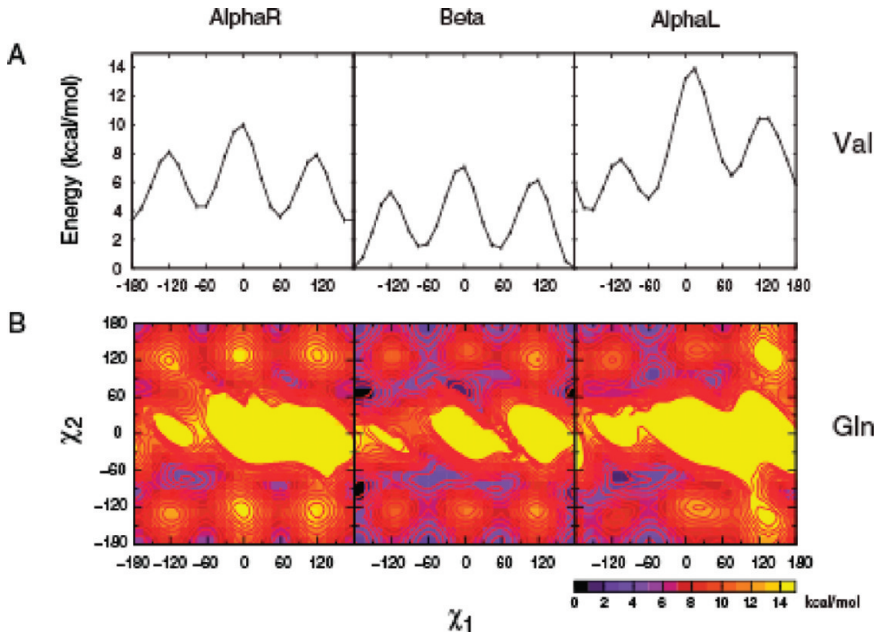


Figure 1. Val χ_1 and Gln 2D χ_1, χ_2 QM energy surfaces for the three backbone conformations based on the dipeptide model. Energies are offset to the global minima for all three backbone conformations.

and Trp, where χ_2 involves an sp^3-sp^2 type bond, rotamers definitions are trans (**t**) at 180°, cis (**c**) at 0°, gauche$^-$ (**m**) at −90°, and gauche$^+$ (**p**) at +90° with window sizes of ±45° used to define χ_2. For the symmetric side-chains of Asp/Phe/Tyr, the χ2 nomenclatures of gauche (**g**) at ±90° and χ_2 = trans (**t**) at 0° or ±180° was used to define and compare rotamer populations. The boundaries of each of these two rotamers were defined with a window size of ±45°. To present our QM results, we adhered to the generic m, p, and t definitions.

Definitions of Backbone Conformation. Backbone ϕ/ψ in the QM calculations were defined as AlphaR (−60,-45), Beta (−120,120), and AlphaL (63.5, 34.8). To obtain comparable crystal distributions, backbone ϕ/ψ angles were collated using helix/strand definitions determined from the program STRIDE.[22] Backbone dihedrals in the $\phi\psi$ ranges of [0,150] and [−30,120], respectively, were considered for comparisons with the alphaL data. Note that this definition does not identify alphaL helicies (that is, four or more consecutive residues in the alphaL backbone conformation) but rather just residues with the alphaL backbone conformation.

Comparisons of QM and Protein Survey Data. To compare the QM and survey data the calculated QM energies for each conformation were converted to probability values p_{ij} conforming to a Boltzmann distribution

$$p_{ij} = \frac{e^{-\Delta E_{ij}/kT}}{\sum_{\text{backbone}} \sum_{ij} e^{-\Delta E_{ij}/kT}} \quad (1)$$

where the relative potential energies are calculated as

$$\Delta E_{ij}^{\text{local}} = E_{ij} - E_{\text{min}}^{\text{bkb}} \quad (2)$$

for a given $(i,j) = (\chi_1,\chi_2)$ with $\chi \in \{-180, -165, -150, ..., 165\}$ where $E_{\text{min}}^{\text{bkb}}$ is the global minimum of the three energy surfaces associated with the studied backbone conformations. $E_{\text{min}}^{\text{local}}$, the local minimum on the individual backbone surfaces, was used in place of $E_{\text{min}}^{\text{bkb}}$ to remove the backbone-dependency in comparisons of individual backbone types. A temperature T of 300 K was used in the calculations.

The crystal distributions for χ_1 and χ_2 angles were calculated from the selected protein structures using a bin size of 15° centered on each QM χ_1 and χ_2 grid point. Populations of each of the rotamers were calculated by summing the probabilities at intervals $\chi \in [-120, 0)$ for **m**, $\chi \in [0, 120)$ for **p**, and $\chi \in [-180, -120) \cup [120, 180)$ for **t** rotamers. Absent from these populations is backbone dependence as the probability distributions were calculated independently for each backbone type.

The 1D and 2D overlap coefficients (OC) for two probability distributions were calculated over the sampled grid points using eq 3 as previously described.[24−28]

$$\text{OC} = \frac{\sum p_m \cdot p_n}{\sqrt{\sum (p_m)^2 \cdot \sum (p_n)^2}} \quad (3)$$

The OC is indicative of how well two probability distributions overlap and serves as a metric to compare probability distributions derived from the QM surfaces and the survey data distributions.

Secondary Structure Propensities. The propensity (or likelihood) for a side-chain to assume a particular secondary structure was calculated from the crystallographic survey using the following definition:

$$P_{\text{conf}}^{\text{AA,survey}} = \frac{A_{\text{conf}}/A_{\text{tot}}}{N_{\text{conf}}/N_{\text{tot}}} \quad (4)$$

where the propensity for residue A to be in conformation "conf" is the fraction of this type of residue in a certain secondary structure conformation normalized over the fraction of all residues in this conformation.[29] To allow for comparisons an analogous approach was applied for the determination of propensity based on the QM-calculated intrinsic energies. The QM-based propensity score is a sum over the normalized Boltzmann weighted probabilities for a defined backbone conformation conf as in eq 5

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Relative QM Energies of Model Residues at Exact sp^3 Rotamer Conformations[a]

	alphaR									global minimum		
residue/rotamer	tt	tm	tp	mt	mm	mp	pt	pm	pp	$\chi1$	$\chi2$	energy
arg	9.99	8.06	5.75	14.87	12.22	17.74	14.19	17.00	15.93	−150	−45	4.08
asn	11.76	6.54	6.81	3.10	3.85	10.68	2.64	8.91	5.97	60	120	0.66
asp	24.59	22.65	22.62	8.22	11.00	5.52	1.21	1.91	7.30	45	−180	0.00
cys	6.70	4.80	5.98	3.52	1.73	5.20	5.00	6.02	2.44	−60	−60	1.73
gln	5.36	10.70	7.18	5.04	5.23	14.95	5.35	10.15	13.01	−90	−75	2.62
glu	19.83	21.11	21.36	17.14	16.98	2.55	17.82	1.99	16.95	90	−60	0.00
hsd	4.42	10.26	9.05	8.31	6.85	7.33	11.23	6.06	8.40	−165	150	3.23
hse	12.86	5.90	6.32	4.15	3.29	8.29	3.17	8.44	5.18	60	120	0.00
hsp	8.43	17.99	16.27	21.65	24.38	24.07	20.39	21.83	24.37	−165	150	6.98
ile	4.40	7.68	3.25	3.98	3.65	6.05	4.68	6.20	9.05	−180	60	3.25
leu	5.66	5.58	2.26	2.99	5.44	5.76	5.34	8.79	8.70	−180	60	2.26
lys	10.44	7.63	4.56	15.01	11.93	17.88	14.29	17.42	16.64	165	60	3.60
met	4.51	16.47	5.21	5.34	7.32	4.03	6.01	12.06	10.44	−60	75	2.97
phe	7.63	6.92	5.62	5.62	2.91	7.68	10.34	6.46	6.40	−60	−75	2.65
ser	9.33	6.06	7.48	3.84	3.52	8.23	4.67	10.14	2.55	60	60	2.55
thr	10.92	8.09	8.73	4.05	4.08	8.62	5.54	11.14	3.58	45	60	3.46
trp	13.03	6.95	6.40	6.85	3.75	8.14	13.41	6.71	6.00	−60	−90	2.50
tyr	7.36	6.95	5.63	5.58	2.92	7.50	10.41	6.28	6.30	−60	105	2.53
val	3.34				4.30				3.60	165		3.34
avg (alphaR)	8.81											
	beta									global minimum		
residue/rotamer	tt	tm	tp	mt	mm	mp	pt	pm	pp	$\chi1$	$\chi2$	energy
arg	8.07	7.86	8.22	6.59	3.04	7.01	7.37	8.09	7.04	−90	30	1.19
asn	1.26	2.85	4.01	5.12	3.44	4.20	7.26	5.72	4.01	−60	−30	0.64
asp	10.63	14.13	13.64	15.84	15.47	13.85	19.75	14.47	26.46	−165	−180	10.38
cys	2.32	0.34	0.71	2.83	1.56	1.63	5.39	5.12	1.56	−180	−75	0.00
gln	1.83	4.37	0.70	2.04	4.59	7.51	3.69	7.02	9.97	−15	75	0.00
glu	18.61	16.89	13.13	19.93	17.80	11.06	21.91	10.76	28.46	−60	75	7.54
hsd	6.64	3.81	2.37	1.31	5.40	5.77	5.07	6.33	7.84	45	−120	0.00
hse	2.27	3.16	3.98	5.93	3.14	3.75	7.78	5.69	5.14	−180	−135	0.88
hsp	14.86	15.42	14.57	3.47	12.83	13.81	8.18	13.10	13.95	−60	165	3.36
ile	1.90	4.25	1.45	0.45	0.00	2.37	2.21	3.89	5.99	−60	−60	0.00
leu	2.53	2.62	0.38	0.00	2.26	3.01	3.91	6.89	6.25	−60	−180	0.00
lys	8.08	8.72	8.49	7.02	1.72	6.46	7.72	8.37	7.84	−45	−60	1.27
met	2.28	13.25	0.34	2.48	6.94	4.90	3.52	10.42	11.03	−180	45	0.00
phe	5.33	2.17	0.93	3.74	2.07	3.99	8.14	5.29	6.10	−180	−90	0.00
ser	3.14	3.04	3.80	4.91	3.83	3.16	7.00	6.28	0.69	60	45	0.00
thr	4.45	5.04	5.37	5.16	4.22	3.07	7.04	6.77	0.79	60	45	0.00
trp	10.50	3.08	2.38	9.15	3.05	4.46	11.71	6.25	5.47	−180	−105	0.00
tyr	5.41	2.07	0.86	3.62	2.17	3.97	8.09	5.27	6.08	−180	−90	0.00
val	0.00				1.53				1.44	−180		0.00
avg (beta)	6.33											
	alphaL									global minimum		
residue/rotamer	tt	tm	tp	mt	mm	mp	pt	pm	pp	$\chi1$	$\chi2$	energy
arg	7.69	8.75	8.26	9.34	12.48	13.04	6.62	12.56	0.00	60	60	0.00
asn	10.67	7.11	9.43	0.00	3.99	6.23	14.18	15.66	3.54	−60	−180	0.00
asp	33.67	40.34	30.90	10.26	14.01	11.61	27.66	26.27	28.80	−60	−165	9.79
cys	9.81	5.17	9.75	2.46	1.66	1.82	8.23	9.35	5.71	−60	−75	1.47
gln	6.84	16.18	8.10	2.71	3.91	9.44	6.41	14.66	7.67	−105	−75	1.03
glu	29.11	51.56	23.27	22.41	15.15	12.46	27.60	21.53	41.69	−90	60	7.41
hsd	10.30	13.73	8.41	10.77	4.32	4.92	9.49	10.29	10.82	60	−105	1.16
hse	12.66	9.38	10.79	1.35	4.55	6.04	17.05	13.59	4.22	−75	−135	1.34
hsp	9.14	14.47	9.82	19.35	16.27	17.18	7.99	11.52	9.93	75	−120	0.00
ile	10.43	18.39	8.65	6.92	6.95	9.14	5.74	11.24	6.60	−30	−180	4.40
leu	13.58	13.30	5.18	2.82	4.87	5.09	10.26	10.46	6.57	−60	−180	2.82
lys	8.09	9.80	8.21	9.41	13.05	13.88	7.27	11.88	0.00	60	60	0.00
met	6.45	24.68	5.43	4.53	7.46	3.16	6.73	25.91	9.65	−75	60	2.23
phe	11.44	12.78	7.79	6.26	2.42	4.85	14.78	10.91	6.39	−45	−75	1.52

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. continued

	alphaL									global minimum		
residue/rotamer	tt	tm	tp	mt	mm	mp	pt	pm	pp	$\chi 1$	$\chi 2$	energy
ser	12.38	2.31	11.76	3.56	5.18	5.04	10.32	10.05	4.45	−165	−60	1.34
thr	16.08	6.12	15.52	6.39	8.40	7.98	11.30	11.11	5.09	−165	−60	4.56
trp	14.13	10.85	10.04	7.70	2.95	5.72	23.62	12.26	4.21	−45	−75	1.69
tyr	11.52	12.81	7.82	6.50	2.48	4.76	14.59	10.80	6.48	−45	−75	1.57
val	4.09				4.86				6.49	−150		4.09
avg (alphaL)	10.77											

[a]Energies in kcal/mol correspond to the exact rotamer conformations (i.e., $\chi_1/\chi_2 = -60$, 180, or 60°) and offset with respect to the global minimum for all three surfaces of each residue. While symmetry in χ_2 dihedrals involving sp^2 atom centers in Asp, Phe, and Tyr is present, energies for all sp^3 rotamer geometries, as defined in the Methods section, are shown.

$$P_{\text{conf}}^{\text{AA,QM}} = \sum_{\text{conf}} p_{ij} \tag{5}$$

where p_{ij} is the Boltzmann factor normalized over the sum of all backbone conformations, as described in eq 1.

RESULTS AND DISCUSSION

In the present study, blocked dipeptides were used as model compounds for computing the intrinsic conformational properties about the χ_1 and χ_2 dihedral angles of the side-chains using QM calculations. Dipeptides have been widely used in both experimental and theoretical work to characterize the behavior of the amino acid side-chains.[30] 1D or 2D energy surfaces were obtained for three backbone conformations for side-chains in their physiologically relevant protonation states. Pro was not studied due to it being a constrained imino acid. As the intrinsic pK_a of His is around 6.5, which is close to physiological pH, the charged state (Hsp) was included in addition to the two neutral protonation states, δ-His (Hsd) and ε-His (Hse). We define the intrinsic conformational properties as those pertaining to the chemical connectivity of the side-chains and their interactions with the adjacent peptide bonds in the backbone in the absence of any interactions with the surrounding environment (i.e., in the gas phase). To understand the contribution of the intrinsic properties to the conformations of side-chains seen in proteins, the QM calculated energies were converted to probabilities based on a Boltzmann distribution and compared to PDB survey data obtained as part of the present work. Accordingly, it should be emphasized that in this study computed conformational potential energies are effectively being compared to conformational free energies from the survey data.

Energy Landscape of Amino Acid Side-Chains. Step one was the generation of a library of ab initio QM energy profiles for the amino acid side-chains using blocked dipeptide models. Full 1D χ_1 or 2D χ_1, χ_2 QM energy surfaces allowed the energy landscape for the rotation of the side-chain moieties to be mapped. The 15° increment for the surfaces was selected as a compromise between resolution and computational cost. Likewise, fixed backbone geometries were used on the basis that given the computational demand of the QM calculations, it is necessary to make the assumption that a single backbone conformation is representative of a type of secondary structure or backbone conformation in the case of alphaL. Example 1D surfaces for Val and 2D surfaces for Gln are shown in Figure 1. In each case the energy surfaces are presented for the alphaR, beta and alphaL backbone conformations offset to the global minima for all (All) conformations, in which the energy of the global minima for all three surfaces was used as the offset. The energy surfaces for the remaining studied amino acids are included as part of Figure S1 of the Supporting Information. Table 1 presents the relative energies for each of the **t**, **m**, or **p** conformations for all the studied dipeptides.

In general, the expected energy minima **m**, **p**, and **t** for sp^3 carbons and, in the case of Phe/Trp/Tyr/Asn χ_2, minima at **t** and ±90° for sp^2 carbons are reflected in the 2D energy maps. Due to the symmetry of their side-chains, the **m** and **p** rotamers for Asp, Phe and Tyr are structurally equivalent. However, results from the full continuous 360° scans of Asp/Phe/Tyr's side-chains are included in the QM energy surfaces. As shown in Figure S1 of Supporting Information, the two halves on the χ_2 space are not entirely symmetric, as is expected for symmetric moieties. Although minute, the differences are an indication of structural changes during QM geometry optimizations that are not fully relaxed due to the longer side-chains being trapped in local minima and present a potential limitation in direct comparisons of QM data with crystallographic surveys. We include the full data to highlight this limitation. However, for further analysis and comparisons with the crystallographic survey, only the (−180,0) range data was used for χ_2 for Asp/Phe/Tyr.

The extended (beta) conformation yields energies that are systematically more favorable than alphaR and alphaL (Table 1), which are evident in the Val energy surfaces shown in Figure 1. This is consistent with the alphaR and alphaL energies being 2 or more kcal/mol higher than that of the extended C5 conformation in the alanine dipeptide.[12,31] However, with Asp, Glu, and Hse the global minimum occurs in alphaR, while with Arg, Lys, Asn, and Hsp, it occurs in the alphaL backbone conformation. Thus, interactions of the polar and charged side-chains with the local backbone can intrinsically alter the ordering of the relative energies of the backbone conformations. Interactions with the backbone also lead to deep energy minima on the surfaces; such minima, which in certain cases do not correspond to conformations observed in the PDB survey, mask the presence of relatively shallower but energetically relevant minima as well as leading to high-energy regions. This effect is most obvious in the charged residues Asp and Glu, where deep energy minima are located on Asp's χ_1 = **m** and **p** and Glu's **mp** and **pm** regions of the alphaR backbone conformation (Figure S1, Supporting Information). The Asp QM surfaces contains low-energy conformations at χ_2 = 0/180° (or **t**). Analysis of the structure of the Asp dipeptide shows a favorable hydrogen bond between the acid group and the peptide bond stabilizing the χ_2 = **t** conformation (Supporting Information Figure S2A). In Hsp, the energy surface is perturbed due to highly favorable interactions with the backbone at the χ_1 = **p** region of the AlphaL backbone (Figure S1, Supporting Information). This is an inherent property of

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the dipeptide representations of side-chains in vacuum where the effects of the solvent as well as the protein environment are omitted. Their presence, however, indicates the importance of environmental contributions to the conformational properties of side-chains observed in protein structures as will be discussed below. Together these data reflect the presence of intrinsic differences in rotamer preferences for the different amino acids.

Protein Databank Survey. Analyses of dihedral distributions of protein side-chains have been the subject of numerous studies[23,32−34] and, as a result, there exists several compilations of protein survey data, including the recent compilation by Shapovalov and Dunbrack.[35] However, because of the accelerated increase of high-resolution structures deposited in the Protein Databank, we performed a new survey. The survey identified 2106 nonredundant (similarity <50%) high resolution (≤1.5 Å) PDB entries. The final data set is summarized in Table 2, with the results separated into distributions for the alphaR, beta, and alphaL backbone conformations as well as for all backbone conformations. To test the statistical significance of the data set, it was randomly split into 5 sets and the populations of the different rotamers compared (Table S2, Supporting Information). The results show the populations to be nearly identical for the 5 sets with the standard deviations of the populations being 0.02 or less. This indicates that the size of the sample is an adequate representation of χ distributions in protein crystal structures. Presented in Figure 2 are 1D Val and 2D Gln χ distributions from the crystallographic survey. Also presented are 1D distributions along χ_1 and χ_2 degrees of freedom for Gln, which are appended to the respective axes of the 2D surface. Comparison of the energy surfaces in Figure 1 and the distributions in Figure 2 show the location of the QM-calculated minima to generally correspond to the survey maxima, respectively, as the probability distributions for both χ_1 and χ_2 have the expected minima corresponding to the **t**, **m**, and **p** rotamers. However, significant differences are also present as expected since the intrinsic conformational energies of the side-chains do not solely dictate the conformations sampled in the heterogeneous environments encountered in protein structures. In the following section, details of the QM and survey differences will be presented allowing for an improved understanding of the relative contribution of intrinsic and environmental factors to be obtained.

Table 2. Relative SASA for Side-Chains in the alphaR, beta, and alphaL Backbone Conformations

	average SASA	SD[a]	*N*
all	32.63	0.08	428089
alphaR	30.74	0.08	139629
beta	18.46	0.23	98692
alphaL	60.08	0.74	20192

[a]SD: standard deviation calculated from 5 random and non-overlapping subsets.

Relationship between Intrinsic Energetics and Rotamer Populations in Protein Structures. To allow for comparison of the potential energy surfaces with data from the PDB survey, the QM energies were converted to probabilities based on a Boltzmann distribution using eqs 1 and 2. The resulting QM probabilities were then overlaid onto the PDB distributions producing "combined QM/PDB probability" plots including the 1D χ_1 and χ_2 summed representations along the *x* and *y* axes, respectively. For comparisons of data in different backbone types, we offset each QM surface to the minimum of respective backbone conformations in order to remove the relative energy contributions originating from the backbone itself. Figure 3 shows the combined QM/PDB probability plots for Thr and Gln. The corresponding plots for all the amino acids can be found in Figure S3 of the Supporting Information. In general, the heterogeneity in PDB rotamer distributions is apparent. For example, rotamer **t** is nearly absent in Thr χ_1 whereas **m** and **p** occur at approximately the same probabilities. The short and negatively charged side-chain of Asp is predominantly in the χ_1 = **t, m** and, interestingly, the χ_2 = 0°

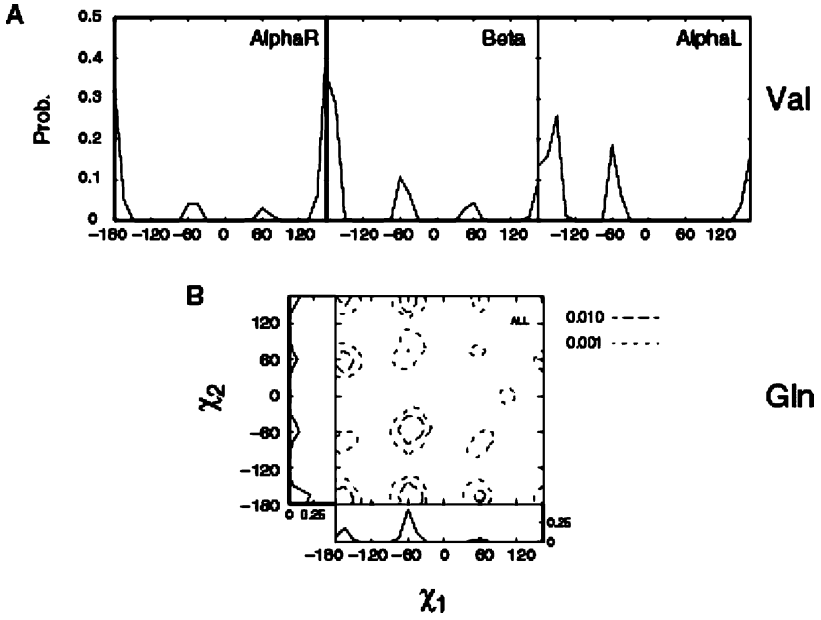


Figure 2. (A) Val χ_1 and (B) Gln 2D χ_1, χ_2 PDB probability distributions. For Val the data is for the individual backbone conformations such that the sum of the probabilities for each surface equal 1. For Gln, only the All surface is presented along with χ_1 and χ_2 1D distributions along the respective axes.

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

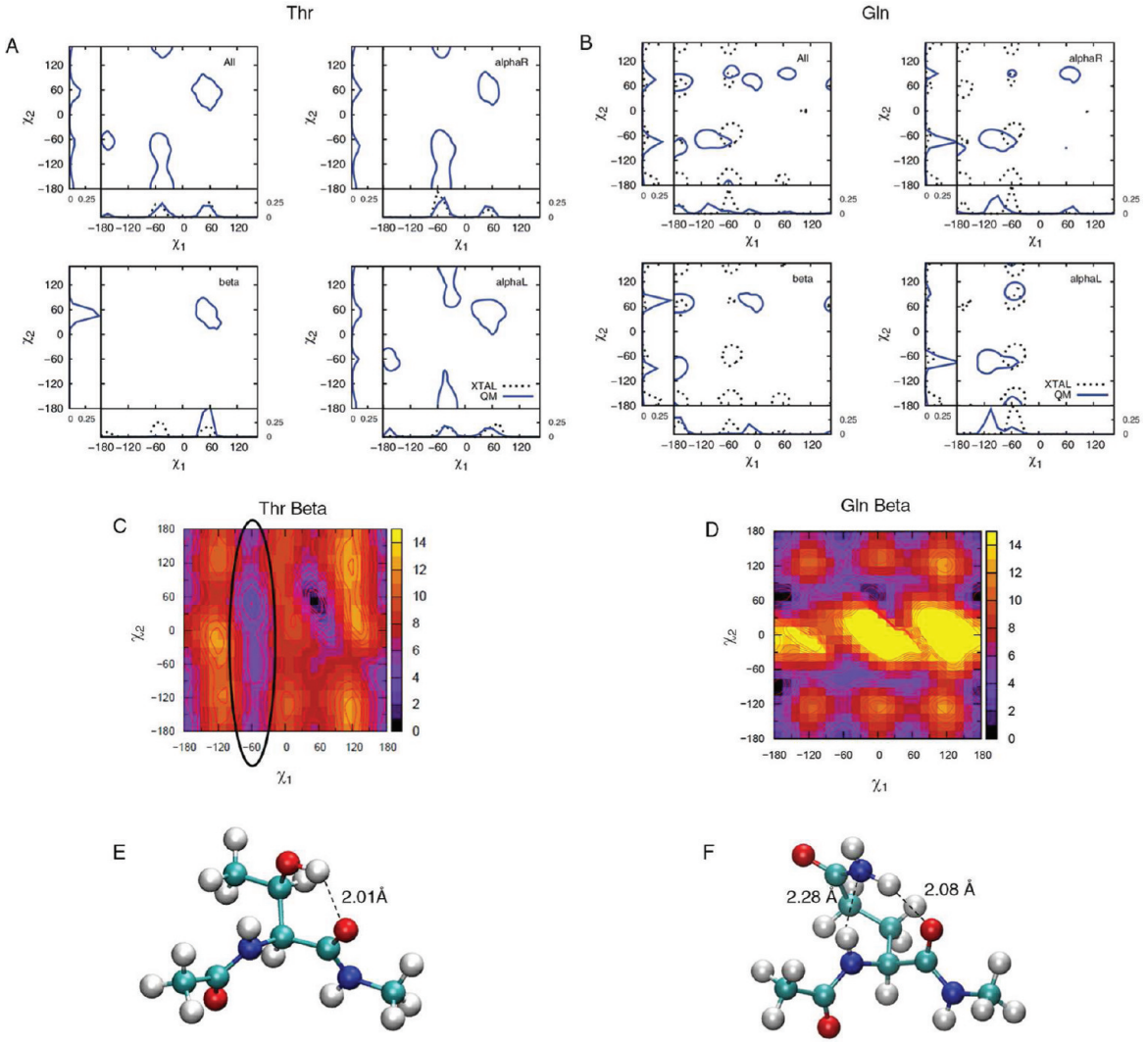


Figure 3. Combined QM/PDB survey probability plots for A) Thr and B) Gln. Crystal distributions for Thr χ_2 are absent as they involve a proton. The probabilities for the three backbone conformations are over those individual surfaces such that the sum of the probabilities for each surface equal 1. 2D QM energy surfaces containing the global minimum is shown for each residue (C and D). Structures of model dipeptides Thr (60°, 45°) (E) and Gln (−15°, 75°) (F) at their global minima show short hydrogen-bonding interactions with the backbone. 2D-plots are made with probability contours of 0.005.

(fully eclipsed) geometries. The χ_2 torsion of the His and Trp side-chains also exhibit a weak but noticeable population at 0°. In Asn/Phe/Tyr/Trp, where χ_2 involves rotation about an sp^2 atom, a notable population at 0° occurs when χ_1 = **m**. Fully eclipsed conformations for these residues are generally not energetically favorable and their presence is discussed in conjunction with intrinsic energetics in the following section.

Along with the combined QM/PDB probability plots for Thr and Gln in Figure 3, shown are the QM 2D energy surface for the beta backbone conformation and selected side-chain conformations for the dipeptides that correspond to deep minima on the energy surfaces. Comparison of the beta energy surface and the corresponding probability distribution shows the difficulty in simply converting the QM potential energy surfaces to probabilities and using those directly for analysis. In the beta surface of Thr there is a deep minimum around **pp** associated with a strong hydrogen bond between the hydroxyl and the backbone carbonyl as indicated in Figure 3C. Similarly, deep minima are found around the off-rotamer location $\chi_1\chi_2$ = (−15°, **p**) for Gln (Figure 3D). Structures corresponding to other minima for Gln and Arg are shown in Figure 4 to further exemplify the type of strong electrostatic interactions that can occur in the QM surface of residues with long side-chains. As calculation of the QM probabilities is based on the Boltzmann distribution of energies relative to the global minimum, the presence of these deep minima can obscure the presence of other local minima in the QM surface. This effect is emphasized by the significantly different rotamer populations of selected amino acids when they are calculated for the QM data offset to the global minima (Table S3 of the supporting data) and when the survey probability of the rotamers are determined over all the backbone conformations (Table S4 of the Supporting Information). Such limitations need to be considered when comparing the QM and PDB distributions. For example, the beta QM energy surface of Thr contains an extensive low-energy valley at χ_1 = **m** (circled in Figure 3C), which is not accurately reflected in the QM probability distribution. However, sampling of that region is evident in the PDB survey data in the 1D χ_1 plots in Figure 3A.

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

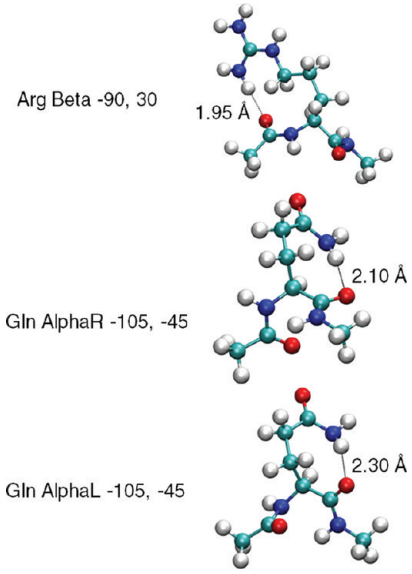


Figure 4. Structures of Arg and Gln showing favorable interactions with the backbone resulting in nonrotameric minimum in QM surfaces.

Quantitative comparisons of the QM and survey probability distributions were achieved using OC values calculated from eq 4. This was performed in both 1 and 2 dimensions for all the amino acids with the results shown in Figure 5. In the figure the amino acids were ordered along the x axis based on physicochemical properties starting from charged, to polar, to polar cyclic, and to hydrophobic. The degree of overlap between the QM and PDB distributions range from near zero for Hsp and Met to higher values for Asn, Ile, and Trp, with the largest OC values occurring with Leu, Phe and Tyr. In general, we observe that simultaneous consideration of the χ_1 and χ_2 yields overall poor agreement with polar or charged side-chains such as Arg, Asp, Glu, His (Hsp/Hsd/Hse), Lys, Asn, and Gln, while hydrophobic side-chains such as Phe, Tyr, and Ile are located in the upper spectrum in terms of correlations with survey distributions. In the case of polar side-chains, which can form favorable intramolecular interactions with the backbone, smaller OC values indicate that their experimentally observed conformations are stabilized by the protein or solution environment. For hydrophobic residues, high OC values indicate a dominance of intrinsic over environmental contributions to their orientations in protein structures. Because hydrophobic interactions are an important feature in stabilizing the core of proteins,[36] this data suggests that intrinsic energetics of hydrophobic residues may make an important contribution to the folding and stabilization of proteins (see below). Furthermore, the QM calculated χ_1 distributions correlate better with crystal occurrences than χ_2. This is expected as χ_1 modulates the orientation of moieties that interact directly with the backbone in contrast to the functional groups defined by the χ_2 dihedral that are removed from the backbone such that their conformations are more strongly influenced by interactions with their environment. Simply put, the further away from the backbone the less interactions with the backbone, and therefore the lower the contribution of the intrinsic conformational properties to the sampling of rotamer

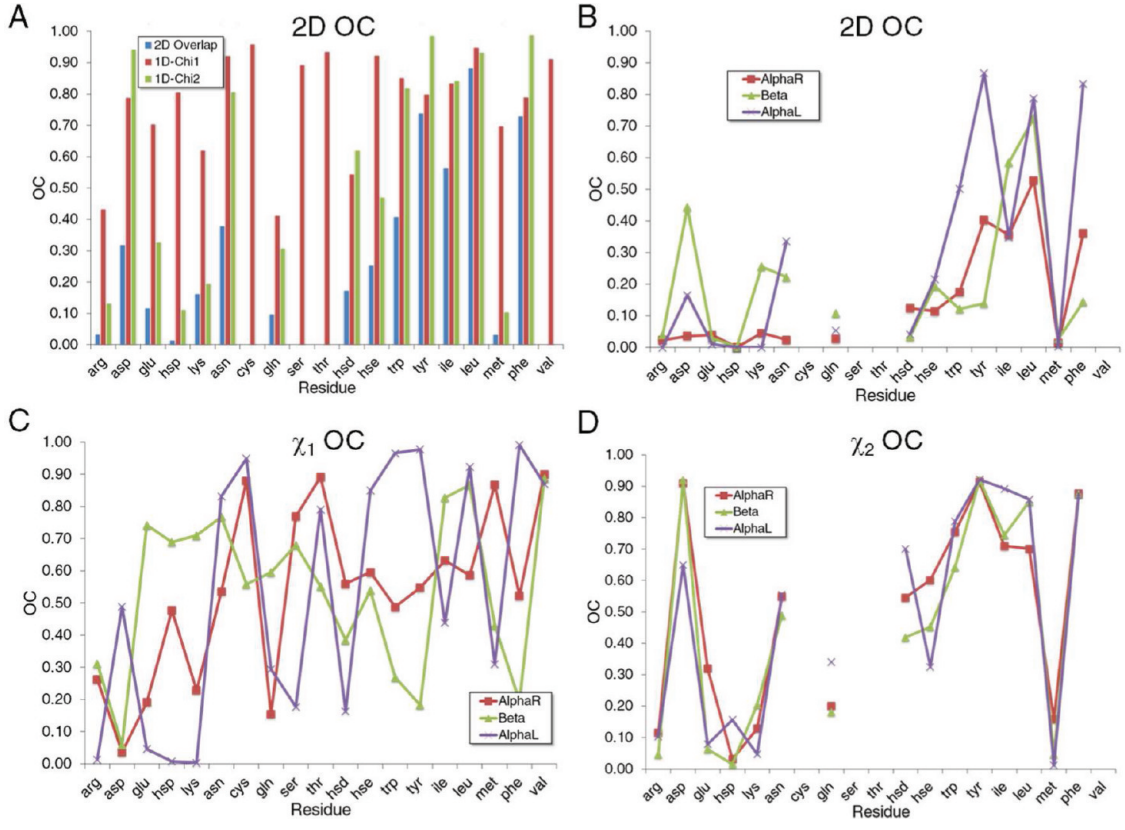


Figure 5. Overlap coefficients (OC) for (A) all surfaces associated with the three backbone conformations and for the three individual backbone conformations for the (B) 2D χ_1, χ_2, (C) 1D χ_1, and (D) 1D χ_2 OC values as a function of residue type. Residues are arranged along the x axis based on physicochemical properties.

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

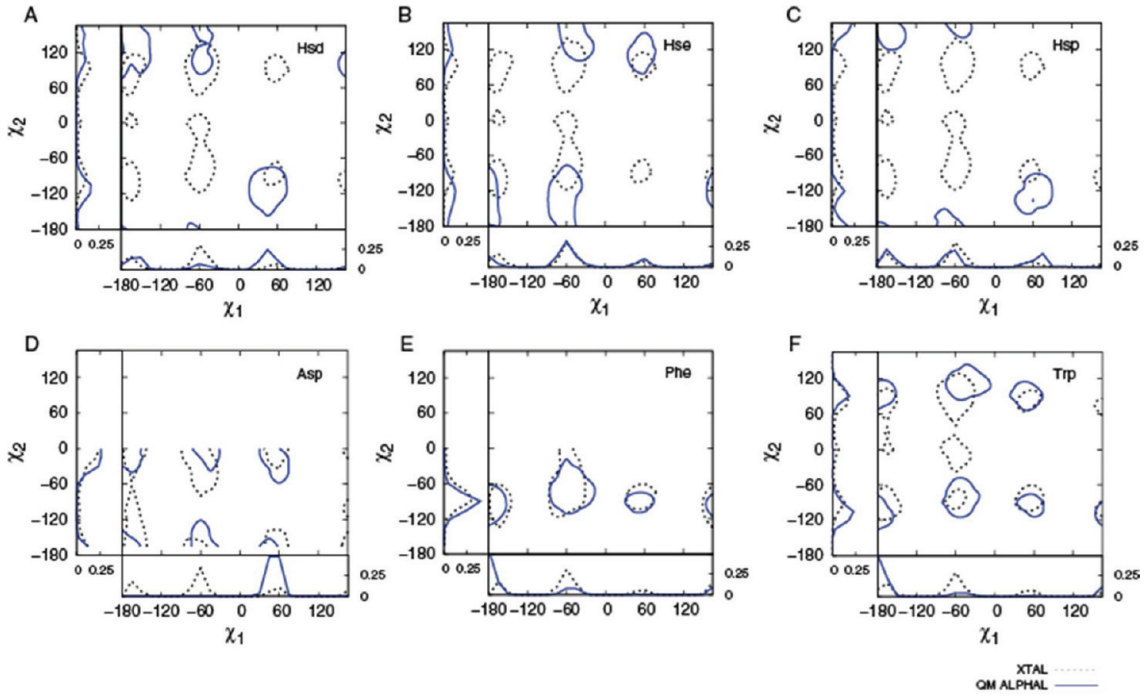


Figure 6. Combined QM/PDB survey probability plots for (A) Hsd, (B) Hse, (C) Hsp, (D) Asp, (E) Phe, and (F) Trp. QM (solid) and crystallographic survey (dashed) probabilities are shown. Probability contour lines are of 0.005. See Supporting Information for backbone-dependent plots.

populations in protein 3D structures. Taken together, these data demonstrate that the extent that environmental effects shift sampled conformations away from the intrinsically favored conformations is highly side-chain dependent. Specific cases will be discussed in the following paragraphs.

With a pK_a of 6.5, the imidazole ring can exist in either the neutral (Hsd or Hse) or +1 charged state under physiological conditions. Therefore, all three protonation states of the His side-chain were considered in the QM calculations and each was compared to the crystallographic His distributions. From the OC graphs (Figure 5) and combined QM/PDB probability plots (Figure 6) the neutral species, Hse, is in better agreement with the PDB survey data than Hsd. While speculative, this suggests that the Hse species occurs at a higher probability in crystal structures than the δ-His counterpart. The charged species Hsp shows little 2D overlap with the crystal distribution. In general, the moderate correlation of the His survey and QM distributions, especially evident in the 2D OC values, is likely associated with the often specialized role of His in proteins. For example, His residues are often involved in well-defined noncovalent interactions associated with chemical catalysis[37−39] or ligating metals[40] where it acts as a general acid or base. Such interactions may be expected to dominate the conformation of the side-chain, overcoming the intrinsic orientational contribution. This appears to be particularly true with Hsp, where its charged nature is suggested to lead to the environment having a larger role in dictating its conformation in proteins.

Intrinsic energies of both Phe and Tyr show high correlation with crystal distributions yielding 2D-OC values of 0.73, χ_1OC of 0.80, and χ_2OC of 0.98 for Phe with the Tyr OC values nearly identical to those of Phe, consistent with the structural and chemical similarity of their side-chain moieties. The high correlation between intrinsic energies and crystal distributions is an indication that the protein environment has a minimal effect on the orientation of the side-chains. To further investigate this we overlaid all occurrences of Phe within the selected protein structures based on side-chain atoms (Cβ and beyond) and identified aliphatic and aromatic carbon atoms within 5.0 Å of the Phe side-chain non-hydrogen atoms. A 3D occupancy map was then constructed at a grid resolution of 1.0 Å^3 and normalized over the total number of atoms that satisfy the distance criteria. This yielded a 3D probability density map (Figure 7) representative of the overall shape of the

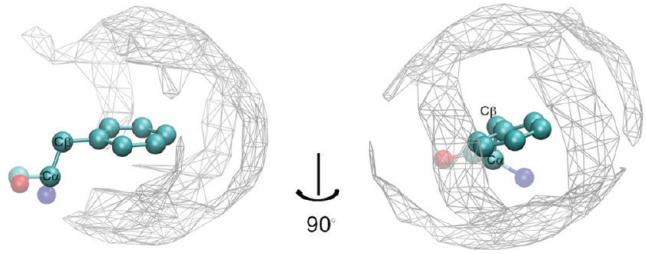


Figure 7. 3D probability distribution of aliphatic and aromatic carbons around the Phe side-chain atoms derived from the crystallographic survey. Densities were calculated by normalizing each occupancy point over the total number of atoms within 5 Å of the phenyl ring's non-hydrogen atoms and displayed with a cut off density of 0.001. An arbitrarily chosen backbone conformation (transparent) is displayed for convenience. Calculations were performed using the MDAnalysis[53] toolkit.

environment in which the phenyl moieties are embedded. As is evident in Figure 7 the environment surrounding the Phe side-chains forms an approximately isotropic cage-like hydrophobic distribution around the phenyl ring. This is consistent with the high OC values such that the environment generally does not offer specific interactions that dictate the conformational properties of the phenyl ring in protein structures. These

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

results suggest a model where the side-chain conformations of Phe and Tyr are dominated by the intrinsic energetics, with the surrounding protein structure accommodating those conformations. For example, an important role of three Phe residues in controlling both the folding rate and the stability of the Villin Head Piece has been documented.[41,42] While this was attributed to aromatic−aromatic interactions the intrinsic conformational properties of that residue may contribute to the entropy barrier to folding as well as the entropic contribution to protein stability.

For Thr χ_1 = **m** is highly populated in protein structures for all backbone types (Figure 3A). However, based on intrinsic energies, alphaR and alphaL backbones produce energetically accessible **m** regions but not beta. As discussed above, due to intramolecular hydrogen bonding, a deep energy well at $[\chi_1, \chi_2]$ = [45°, 60°] or **pp** is present in the beta backbone conformation. It is therefore clear that the protein environment must compete for these interactions. To investigate this, all Thr residues in the χ_1 = **m** orientation were obtained from the selected crystal structures and further divided into alphaR and beta backbone classes. For each of the two classes we calculated the occurrence of hydrogen-bond donor/acceptor atoms (O, N, S) within 5 Å of the side-chain oxygen atom, excluding O or N atoms in the peptides bonds covalently linked adjacent to the side-chain (self) and normalized that data to the total number of hydrogen-donor atoms that satisfy the distance criterion. No significant sampling of well-defined regions in the vicinity of the side-chain was found in the resulting distribution (not shown). We then further classified neighboring atoms based on their origin: water, backbone (nonself), or side-chains. Of particular interest was the distribution obtained from nonself backbone atoms. Shown in Figure 8 is an overlay of resulting 3D distributions of backbone hydrogen-bond donor/acceptor atoms around Thr for χ_1 = **m** summed over the alphaR and beta backbone conformations. Analysis of the probability distributions show well-defined densities where the Thr hydroxyl can act as both a hydrogen donor and acceptor. The χ_1 = **m** conformation in the beta backbone, which intrinsically interacts favorably with its own backbone, appears to find strong coordinating interactions with N/O atoms of the neighboring protein backbone. This would mean that the intrinsic conformational preference of Thr for χ_1 = **m** is effectively overcome by hydrogen bond interactions with other backbone groups. Interestingly, this type of interaction is also present in alphaR, although intrinsic energetics alone was sufficient to explain its occurrence. This indicates a more "passive" role of the environment in dictating the side-chain orientations of Thr in alphaR helices. An example of the type interaction is shown in Figure 8C where the ability of a selected Thr hydroxyl to act as both a hydrogen bond donor and an acceptor in interactions with the peptide backbone is evident. This observation is a well-defined example of how the protein environment can stabilize specific side-chain rotamers and provides a new insight into the complexity of structural properties of Thr residues.

Effects of Local Environment. The effect of local environment on rotamer population has previously been discussed in a number of studies.[4,43−45] It has been observed that the protein and aqueous environments provide stabilizing interactions to the side-chain and selects from a pool of intrinsically low-energy orientations.[4] In fact, it has been proposed by Schrauber et al.[43] that in some cases individual side-chains can adapt high-energy orientations as a compromise

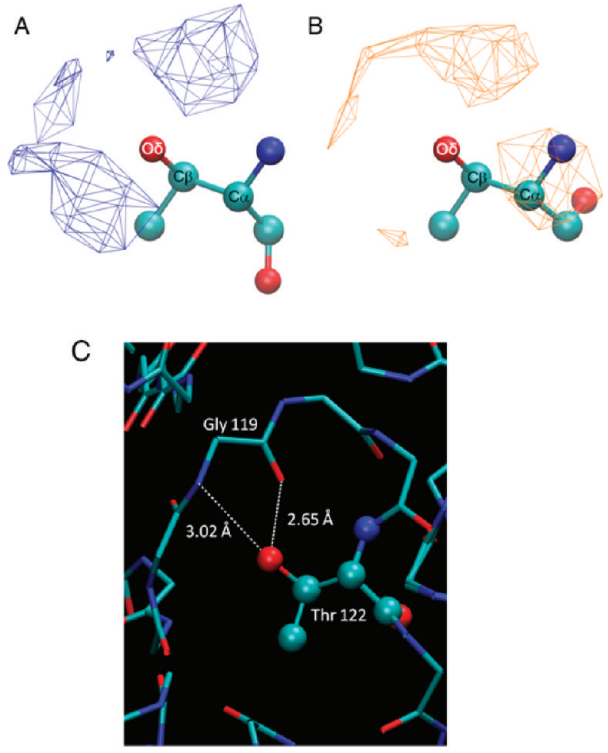


Figure 8. 3D Probability distribution for nonself backbone N/O atoms around the Thr side-chain Oδ at χ_1 = m derived from crystallographic survey. Densities were calculated by normalizing each occupancy point by the total number of atoms within 5 Å of O_δ of Thr and displayed with a cutoff density of 0.001. Shown are Thr structures in the alphaR (A) and beta (B) backbone conformations along with corresponding N/O densities represented as wireframes. (C) Thr122 of 1BYI forms hydrogen bonds with the backbone atoms from Gly119. Calculations were performed using the MDAnalysis[53] toolkit.

to allow the protein as a whole to access low free energy conformations. However, Petrella and Karplus[5] stipulated that for Val, Ile, Leu, Phe, Trp, and Met, the presence of off-rotamers can be attributed to internal energetics of the isolated side-chain. From the database survey, in addition to the probability peaks at locations corresponding to the energy minima associated with the sp^3 Cα and Cβ atoms, off-rotamers at χ_2 = 0° are present in His and Trp (Figure 6). Although the QM energy surfaces for these residues indicate presence of local minima at χ_2 = 0° converted QM probabilities do not contain significant populations for His or Trp in this conformation suggesting a dominant role of environmental effects. However, intrinsic effects are playing a role in the observed off-rotamer conformations of His and Trp. Previous computational studies[45] have attributed the occurrence of this Trp off-rotamer to C−H moieties on the indole ring acting as hydrogen bond donors interacting with hydrogen bond acceptors in the surrounding environment. In the case of His, the χ_2 = 0° population appeared only in the QM probability distribution for Hse in the χ_1 = **m** rotamer associated with the Beta backbone, albeit at relatively low levels (Figure S3, Supporting Information). In this conformation a hydrogen bond between C═O and C_δ-H of the side-chain is formed, therefore stabilizing this orientation (Figure S2B, Supporting Information). The presence of shallow local minima at or near χ_2 = 0° in the QM energy surfaces across all three protonation states of histidine (Figure S1, Supporting Information) indicate that

dx.doi.org/10.1021/ci300079j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

similar interactions are also present in Hsd and Hsp. However, the presence of deep energy wells near (**t**,−120°) of Hse/beta and (**m**,−120°) and (**p**,120°) of Hse/alphaR, (**p**,−120°) and (**t**,120°) of Hsd/beta, and (**p**,−120°) of Hsp/AlphaL leads to the χ_2 orientation at 0° not being significantly populated upon conversion of the QM surface to a probability distributions, as discussed above. In fact, nearly all the probability peaks observed in the crystallographic survey have corresponding local minima in Hsd/Hse QM energy surfaces, though they are obscured in the QM probability surfaces due to deep energy minima. This data suggests that interactions of the His side-chain with the environment strongly compete for the deep energy minima, as indicated by the weak 2D OC value (Figure 5a). In fact, close examination of the QM energy surfaces (Figure S1, Supporting Information) reveals energy minima near the specific off-rotamer location (**m**,0°). To understand the structural implications of the local protein environment in crystal occurrences of these off-rotamers we calculated the shortest distances of O, N, or S atoms with respect to the Nδ or Nε of the imidazole ring. A histogram was constructed using a bin size of 0.2 Å. In Figure 9, the distribution of O/N/S atom

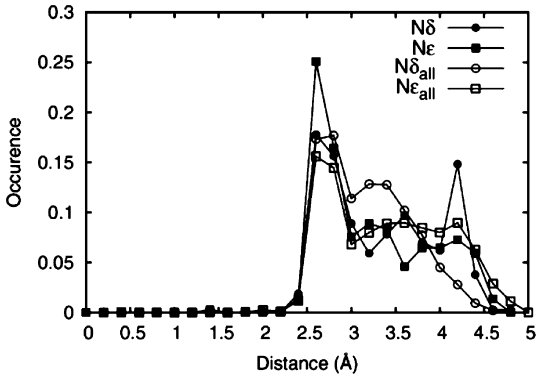


Figure 9. Distribution of O, N, or S atoms around the N atoms of the His side-chains. Distances of the nearest atoms to each Nδ (closed circles) or Nε (closed squares) of His side-chain that are in the $\chi_2 = 0° \pm 10°$ conformation (371 data points) were binned to 0.2 Å and compared with those for all (10101 data points) His entries (open circles and squares for Nδ and Nε, respectively).

distances are shown across all His residues versus those across conformations corresponding to the above-defined off-rotamer. It is clear that both nitrogens participate in strong hydrogen-bonding interacions, as seen in probability spikes in the 2.5−3.5 Å range. However, a marked rise in probability of closely interacting (2.5−3.0 Å) hydrogen-bonding partners around off-rotameric Nε, but not Nδ, provides evidence that stabilization of Nε is important at these conformations. The above-presented evidence provides a link between the occurrences of off-rotamers and intrinsic conformational as well as environmental effects.

Backbone Dependency of Intrinsic Conformational Energetics. It has been previously discussed that the prevalence of rotamers is dependent on backbone conformation.[44] To investigate the extent by which local backbone conformation impacts the contribution of intrinsic energetics to side-chain conformation, the crystallographic data was partitioned into three backbone classes corresponding to the three backbone conformations used in the QM calculations. Figure 5 includes the backbone-dependent OC values for all the studied amino acids. With the 2D OC values, alphaR offers the least agreement with the survey results with an average 2D OC over all the studied amino acids of 0.15 versus 0.21 and 0.28 for beta and alphaL, respectively. One explanation could be that the beta and alphaL conformations offer more potential for side-chain-self-backbone interactions, leading to larger intrinsic contributions. The intrinsic contribution is largest for Asp, Lys, Gln, and Ile when they are in the beta backbone conformation. For Asn, Trp, Tyr, Leu, and Phe, higher OC values are seen at the alphaL orientation, with values of 0.34, 0.50, 0.87, 0.79, and 0.83, respectively. These data indicate that the extent to which intrinsic energetics dictates conformational sampling is backbone dependent.

Secondary Structure Propensity. The preference for a residue to assume a particular backbone conformation has been a popular topic in structural bioinformatics (reviewed in ref 46). Early work by Blout et al.[47] established the first correlation between amino acid sequence and secondary structure using synthetic homopolymers. Such a correlation has led to the development of numerous secondary structure prediction methods that can be divided into two approaches: physicochemical and probabilistic. With the increasing number of available protein structures, the probabilistic approach has seen significant improvement[46] in terms of accuracy since the first propensity scales derived from observed frequencies were published (reviewed in ref 48). Results obtained in the present study allow us to investigate the intrinsic conformational energies of the side-chains, which contribute to their physicochemical properties, with their secondary structure propensity. To address this for each residue in backbone conformation "conf" the ratio of the sum of Boltzmann probabilities of the $\chi_{1,2,conf}$ energy surface is calculated relative to the sum of Boltzmann weights for all three backbone conformations (eq 5). This yields the probability of each residue occupying the different types of secondary structure based on the intrinsic conformational energies alone. These probabilities were then compared with analogous values from the survey data calculated using eq 4. To remove the likelihood of bias from strong side-chain-backbone electrostatic interactions contributing to the QM results side-chains with charged moieties that can interact directly with the local backbone (ie. Asp, Glu, Hsp) were excluded from the analysis. A linear regression model was built for each of the three backbone conformations (Figure 10) with the data presented in Table S5 of Supporting Information. For alphaR or beta, there is negligible correlation between the crystallographic and intrinsic propensities, indicating the intrinsic conformational properties to not play a role in conformational preference. However, a correlation is found for alphaL with a slope of 0.47 and R^2 of 0.82, indicating that the intrinsic energetics are contributing to the propensity of the different amino acids to assume that backbone conformation in protein structures. This result is consistent with a previously established relationship[49] between *J*-coupling constants of blocked residues (dipeptides) in solution and occurrences in coils.

Analyses of the results for individual residues reinforce the above observation. Selected amino acids that have the high propensities for alphaL in the survey, Asn, Lys, and Arg, all have the global minimum in alphaL (Table 1). In contrast, some of the amino acids with the least favorable average energies over all the alphaL minima relative to the alphaR and beta minima, Thr and Ile, have some of the lowest alphaL experimental propensities. These results suggest a model where intrinsically favorable interactions of the side-chain with the backbone in

dx.doi.org/10.1021/ci300079j | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

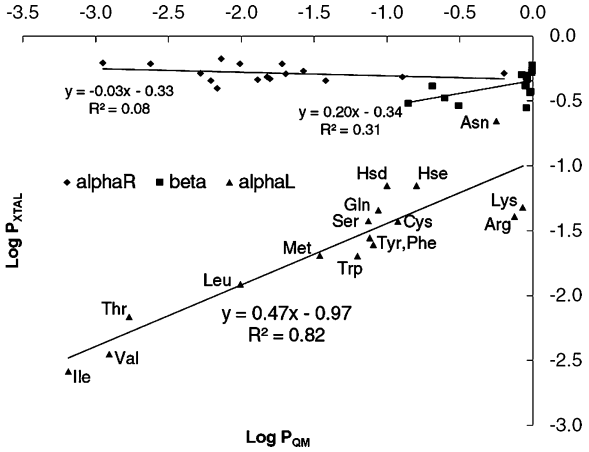


Figure 10. Correlation of secondary structure propensities calculated from crystallographic survey and from intrinsic energies. Charged residues Asp, Glu, and Hsp were omitted from the analysis.

the alphaL conformation lead to enhanced sampling of that backbone conformation in proteins. Interestingly, only short (less than 5 residues) left-handed helical regions of polypeptides are found in nature and are typically functionally important components of the protein.[50] Along with the suggestion that neighboring residue effects[51] play a role in these regions, the present observation of the role of intrinsic conformational energies in sampling of the alphaL backbone conformation offers another example of how primary sequence can play an important role in protein function by favoring biologically important conformations. It is important to bear in mind that, due to the alphaL backbone definition used in this study, no direct relationship can be drawn between our alphaL and true left-handed alpha helices. However, our results provide an insight as to how a relatively unusual backbone conformation such as alphaL can be stabilized.

Solvent Accessibility. Additional analysis of the crystallographic survey data involved the solvent accessible surface area (SASA) for the side-chains in different backbone types (Table 2). The alphaL set clearly stands out as being more solvent-exposed than alphaR or beta. As presented above the average 2D OC value for alphaL, 0.28, was higher than the values for beta and alphaR, with OCs of 0.15 and 0.21, respectively, and side-chain type is indicated to influence assumption of the alphaL conformation. These observations suggest a simple model where the additional impact of the intrinsic conformational energies in alphaL is due to a decrease in the number of specific interactions of the side-chains with the surrounding protein.

SUMMARY

Presented are the intrinsic energy landscapes for amino acid side-chains calculated using QM calculations on representative dipeptide models. The data offers insights into the role of intrinsic energetics on side-chain sampling of χ_1 and χ_2 torsions. The locations of local minima are in good agreement with observed conformations in protein structures. However, the extent to which they agree, as measured by OC values, depends on the type of amino acid as well as the backbone conformation. Strong electrostatic interactions with the backbone resulting in deep energy minima on QM surfaces contribute to the low OC values seen in charged or polar side-chains. However, the protein environment can effectively compete with the intrinsic interactions, as evidenced by the absence of observed rotamer populations in the protein survey data that correspond to deep minima within the QM surfaces. Information is also provided that support previously discussed occurrences of off-rotamers, with new evidence suggesting that intrinsic energetics of His and Trp lead to the stabilization of local minima corresponding to observed off-rotamer populations. Another interesting observation was the high 2D-OC values for Leu, Phe, and Tyr, indicating a dominant role of intrinsic energetics in their conformational properties. While speculative, based on the known role of these amino acids in protein folding the present results suggest that their intrinsic conformational properties make an entropic contribution to both protein folding rates and protein stability. Finally, a correlation for the propensities of side-chains to occur in the alphaL backbone conformation is found between intrinsically determined preferences and those occurring in proteins. This observation suggests that the identity of a side-chain plays a more prominent role in influencing assumption of the alphaL conformation as compared to the alphaR or beta backbone conformations.

The presented QM calculations will be important in empirical force field optimization. At the time of this writing, there is no published data on high level ab initio calculations of the complete energy profiles associated with side-chain χ_1 and χ_2 torsions for the relevant amino acids residues. Accordingly, the calculated QM energy surfaces will be of utility as target data for the optimization of protein force fields.

ASSOCIATED CONTENT

Supporting Information

Figures and tables referred to in the paper as well as complete QM surfaces and overlap plots with crystallographic survey data. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: 410-706-7442. Fax: 410-706-5017. Email: alex@outerbanks.umaryland.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was supported by the University of Maryland Computer-Aided Drug Design Center and NIH grants GM072558 and GM051501 and NSF grant CHE-0823198, and we thank Dr. Edwin Pozharski for providing insight on subtle aspects of PDB structures.

REFERENCES

(1) Kendrew, J. C. Structure and function in myoglobin and other proteins. *Fed. Proc.* **1959**, *18* (2,Part 1), 740−751.

(2) Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **1960**, *185* (4711), 416−422.

(3) Fermi, G.; Perutz, M. F.; Shaanan, B.; Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. *J. Mol. Biol.* **1984**, *175* (2), 159−174.

(4) Gelin, B. R.; Karplus, M. Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry (Mosc)* **1979**, *18* (7), 1256−1268.

(5) Petrella, R. J.; Karplus, M. The energetics of off-rotamer protein side-chain conformations. *J. Mol. Biol.* **2001**, *312* (5), 1161−1175.
(6) Butterfoss, G. L.; Hermans, J. Boltzmann-type distribution of side-chain conformation in proteins. *Protein Sci.* **2003**, *12* (12), 2719−2731.
(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78* (8), 1950−1958.
(9) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474−6487.
(10) Brooks, B. R.; Brooks, C. L., 3rd; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.
(11) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.
(12) MacKerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400−1415.
(13) Sutcliffe, M. J.; Hayes, F. R.; Blundell, T. L. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Eng.* **1987**, *1* (5), 385−392.
(14) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, Jr., J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford CT, 2004.
(15) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
(16) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta (Berlin)* **1973**, *28*, 213−222.
(17) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618−622.
(18) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., 3rd; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F., 3rd; Kong, J.; Krylov, A. I.; Gill, P. M.; Head-Gordon, M. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8* (27), 3172−3191.
(19) Woon, D. E.; Dunning, T. H. Gaussian-basis sets for use in correlated molecular calculations 0.5. Core-valence basis-sets for boron through neon. *J. Chem. Phys.* **1995**, *103* (11), 4572−4585.
(20) Weigend, F.; Häser, M. RI-MP2: First derivatives and global consistency. *Theor. Chem. Acc.: Theory, Comput., Model. (Theor. Chim. Acta)* **1997**, *97* (1), 331−340.
(21) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.
(22) Heinig, M.; Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **2004**, *32* (WebServer issue), W500−W502.
(23) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins* **2000**, *40* (3), 389−408.
(24) Manders, E. M. M.; Verbeek, F. J.; Aten, J. A. Measurement of co-localization of objects in dual-colour confocal images. *J. Microsc.* **1993**, *169*, 375−382.
(25) Bernard, D.; Coop, A.; MacKerell, A. D., Jr. Quantitative conformationally sampled pharmacophore for delta opioid ligands: reevaluation of hydrophobic moieties essential for biological activity. *J. Med. Chem.* **2007**, *50* (8), 1799−1809.
(26) Rais, R.; Acharya, C.; Tririya, G.; MacKerell, A. D.; Polli, J. E. Molecular switch controlling the binding of anionic bile acid conjugates to human apical sodium-dependent bile acid transporter. *J. Med. Chem.* **2010**, *53* (12), 4749−4760.
(27) Rais, R.; Acharya, C.; MacKerell, A. D.; Polli, J. E. Structural determinants for transport across the intestinal bile acid transporter using C-24 bile acid conjugates. *Mol. Pharmaceutics* **2010**, *7* (6), 2240−2254.
(28) Acharya, C.; Coop, A.; Polli, J. E.; MacKerell, A. D., Jr. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput.-Aided Drug Des.* **2011**, *7* (1), 10−22.
(29) Shortle, D. Composites of local structure propensities: Evidence for local encoding of long-range structure. *Protein Sci.* **2002**, *11* (1), 18−26.
(30) Hermans, J. The amino acid dipeptide: small but still influential after 50 years. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (8), 3095−3096.
(31) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27−85.
(32) James, M. N.; Sielecki, A. R. Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.* **1983**, *163* (2), 299−361.
(33) Chandrasekaran, R.; Ramachandran, G. N. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int. J. Protein Res.* **1970**, *2* (4), 223−233.
(34) Dunbrack, R. L., Jr.; Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **1997**, *6* (8), 1661−1681.
(35) Shapovalov, M. V.; Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19* (6), 844−858.
(36) Lins, L.; Brasseur, R. The hydrophobic effect in protein folding. *FASEB J.* **1995**, *9* (7), 535−540.
(37) Warshel, A.; Naray-Szabo, G.; Sussman, F.; Hwang, J. K. How do serine proteases really work? *Biochemistry (Mosc.)* **1989**, *28* (9), 3629−3637.

(38) Du, Z.; Shemella, P. T.; Liu, Y.; McCallum, S. A.; Pereira, B.; Nayak, S. K.; Belfort, G.; Belfort, M.; Wang, C. Highly Conserved Histidine Plays a Dual Catalytic Role in Protein Splicing: A pKa Shift Mechanism. *J. Am. Chem. Soc.* **2009**, *131* (32), 11581−11589.
(39) Inouye, M.; Dutta, R. *Histidine Kinases in Signal Transduction*; Academic Press: Amsterdam, 2003.
(40) Harding, M. M. The architecture of metal coordination groups in proteins. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2004**, *60* (Pt 5), 849−859.
(41) Frank, B. S.; Vardar, D.; Buckley, D. A.; McKnight, C. J. The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. *Protein Sci.* **2002**, *11* (3), 680−687.
(42) Vugmeyster, L.; Ostrovsky, D.; Khadjinova, A.; Ellden, J.; Hoatson, G. L.; Vold, R. L. Slow motions in the hydrophobic core of chicken villin headpiece subdomain and their contributions to configurational entropy and heat capacity from solid-state deuteron NMR measurements. *Biochemistry (Mosc.)* **2011**, *50* (49), 10637−10646.
(43) Schrauber, H.; Eisenhaber, F.; Argos, P. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **1993**, *230* (2), 592−612.
(44) Dunbrack, R. L., Jr.; Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat. Struct. Biol.* **1994**, *1* (5), 334−340.
(45) Petrella, R. J.; Karplus, M. The role of carbon-donor hydrogen bonds in stabilizing tryptophan conformations. *Proteins* **2004**, *54* (4), 716−726.
(46) Burkhard, R. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **2001**, *134* (2−3), 204−218.
(47) Blout, E. R.; Lozé, C. d.; Bloom, S. M.; Fasman, G. D. The dependence of the conformations of synthetic polypeptides on amino acid composition. *J. Am. Chem. Soc.* **1960**, *82*, 3787−3789.
(48) Schulz, G. E.; Schirmer, R. H. Prediction of secondary structure from the amino acid sequence. In *Principles of Protein Structure*; Cantor, C. R., Ed.; Springer-Verlag: New York, 1979.
(49) Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (5), 1272−1277.
(50) Novotny, M.; Kleywegt, G. J. A survey of left-handed helices in protein structures. *J. Mol. Biol.* **2005**, *347* (2), 231−241.
(51) Avbelj, F.; Baldwin, R. L. Origin of the neighboring residue effect on peptide backbone conformation. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (30), 10967−10972.
(52) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285* (4), 1735−1747.
(53) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32* (10), 2319−2327.