A Self Consistent Approach to Rotamer and Protonation State Assignments (RAPA): Moving Beyond Single Protein Configurations

Mossa Ghattas[1,2], Prerna Gera[4], Steven Ramsey[2], Anthony Cruz Balberdy[4], Nathaniel Abraham[4], Daniel McKay[3], Tom Kurtzman[1,2,3]

[1] Ph.D. program in Chemistry, The Graduate Center, City University of New York, New York, USA
[2] Department of Chemistry, Lehman College, City University of New York, New York, USA
[3] Ventus Therapeutics, Inc. 4800 rue Levy, Montreal, Quebec H4R 2P7, Canada
[4] Ventus Therapeutics U.S. Inc. 100 Beaver St. Suite 201, Waltham, MA 02453, USA

1

Abstract

There are currently over 160,000 protein crystal structures obtained by X-ray diffraction with resolutions of 1.5Å or greater in the Protein Data Bank. At these resolutions hydrogen atoms do not resolve and heavy atoms such as oxygen, carbon, and nitrogen are indistinguishable. This leads to ambiguity in the rotamer and protonation states of multiple amino acids, notably asparagine, glutamine, histidine, serine, tyrosine, and threonine. When the rotamer and protonation states of these residues change, so too does the electrochemical surface of a binding site. A variety of computational tools have been developed to assign these states for these residues based on a crystal protein structure by evaluating the possible states and typically deciding on one single state for each residue. We posit that multiple rotamer and protonation states of residues are consistent with the resolved structure of the proteins and introduce a protonation and rotamer assignment tool that identifies an ensemble of rotamer and protonation states that are consistent with the X-ray scattering data of the protein. Here, we present a Rotamer and Protonation state Assignment (RAPA) tool that analyzes local hydrogen bonding environments in the resolved structures of proteins and identifies a set of unique rotamer and protonation states that are energetically consistent with the crystal structure. We evaluate all RAPA predicted states in unrestrained molecular dynamics simulations and find that there are multiple configurations for each protein which match the X-ray results with RMSDs of less than 1.0Å for the atoms with the lowest 90% B-factors. We find that for most protein systems (62 of 77) there are 8 or fewer possible states suggesting that there is no combinatorial explosion of accessible configurations for a majority of proteins. This suggests that investigating all energetically accessible rotamer and protonation states for most proteins is computationally feasible and that the selection of single states is arbitrary.

1. Introduction

Computational modeling of proteins in atomistic detail relies upon a precise set of coordinates of every protein atom. These are primarily obtained by X-ray crystallographic experiments, Nuclear Magnetic Resonance (NMR), or cryo-electron microscopy (Cryo-EM). Currently, there are over 160,000 protein crystal structures obtained by X-ray diffraction with resolutions of 1.5Å or greater in the Protein Data Bank (PDB)[1], which accounts for 89% of the available protein crystal structures. At these resolutions, oxygen, nitrogen, and carbon atoms cannot be reliably distinguished from each other and hydrogen atoms do not resolve. This leads to ambiguities in the assignment of protonation and rotamer states which is particularly relevant for asparagine (ASN), glutamine (GLN), and histidine (HIS) residues. For asparagine and glutamine residues, there are two possible rotamer states for the side chain amide group. For histidine side chains, the imidazole ring has two possible rotamer states and three potential protonation states (being protonated in the delta, epsilon, or both positions) leading to six possible states that are not readily distinguishable from experimental X-ray scattering data. Though less frequent there can also be ambiguity in the protonation states of the charged residues aspartate (ASP) and glutamate (GLU), similar ambiguity can be found in the positions of hydrogen atoms in hydroxy-containing residues serine (SER), threonine (THR), and tyrosine (TYR) as well as potential disulfide bridges between cysteine (CYS) residues.[2–4]

These ambiguities are a well-known limitation of experimental methods and a number of academic and commercial protein preparation methods have been developed to assign appropriate states for those residues in protein crystal structures.[5–31] Most of these methods focus on identifying a single *most-probable* configuration[a] or require user input to modify the method's proposed state assignments of the residues. Here, we posit that multiple configurations of a protein structure are consistent with the experimental data and introduce a Rotamer and Protonation state Assignment tool (RAPA) that enumerates a set of unique configurations that are energetically consistent with the experimentally resolved structure. In the context of structure-based drug discovery (SBDD), enumerating such a set of viable rotamer and protonation (RP) states is particularly important. A fundamental principle of SBDD is that a potential drug must be electrostatically complementary to the protein surface i.e. donating or accepting hydrogen bonds (h-bonds) and making hydrophobic contacts where appropriate. Alternate rotamer states of residues in a binding site change the positions of h-bond contacts and alternate protonation states change the character of interaction sites from donor to acceptor or vice-versa. Correspondingly, the chemical matter making complementary interactions changes with varying RP states. Most preparation tools assign a single state. In virtual screening applications this effectively limits the screen to identifying chemical compounds that are complementary to that specific state. Identifying a set of viable RP states expands the chemical space being explored to chemical compounds that are complementary to any viable

---

[a] We use the term "configuration" to denote a complete description of the positions of every atom in the protein.

2

configuration. This, in turn, increases the ability of virtual screening workflows to identify compounds with potentially better affinity, pharmaco-kinetic properties, binding specificities, and patentability.

RAPA uses a recursive approach to determine the protonation and/or rotamer states of all ambiguous[b] residues in the protein of interest. The tool works by estimating the energetics of the local hydrogen-bonding environment of all ambiguous residues in which all potential protonation and rotamer states of ambiguous neighboring residues are considered. The approach is described in detail in section **2.2**. Unlike other approaches, RAPA often outputs multiple energetically viable configurations of a protein structure that are consistent with the experimentally determined X-ray diffraction data. We apply RAPA to a subset of 77 crystal protein structures from the Database of Useful Decoys – Enhanced (DUDE).[32] For most of these proteins (62 of 77), RAPA identifies 8 or fewer energetically accessible configurations, nullifying the assumption that there could potentially be a combinatorial explosion of accessible configurations. This suggests that investigating all accessible RP states for most proteins is computationally feasible and that the selection of a single state, as most well used protein preparation methods do, is arbitrary. We evaluate the proposed configurations using molecular dynamics (MD) simulations in two ways: First, we evaluate the stability of each configuration using unrestrained simulations and determine whether the time averaged MD structure remains comparable to the crystal structure. We find that all the proposed configurations were consistent with the crystal structures having RMSDs less than 2.18Å for the lowest 90% of B-factor atoms and a majority had RMSDs less than 1.0 Angstrom. Second, for ASN, GLN, and HIS residues that are predicted to be stable in multiple rotamer states, we evaluate whether both predicted states are stable. We find that for most systems (62 of 77) there are multiple configurations that are consistent with the X-ray scattering data, evaluated to be energetically feasible, and have stable structures in MD simulations.

2. Methods

2.1 Selection of proteins for analysis and initial preparation

We applied RAPA to the subset of 77 protein structures (**Table S1**) from the DUDE dataset that have no missing loops, covalently-bound small molecules, or heme groups. Each protein was prepared using OpenEye Spruce[31] using default parameters. Ligand and water molecules were removed from all systems prior to being analyzed.

2.2 RAPA Program

2.2.1 RAPA Workflow

The RAPA workflow is diagrammed in **Figure 1.**

Step 1 - (**Initialization**) Step 1 assigns *known* status to all polar heavy atoms for which there is no ambiguity. Backbone nitrogen atoms, arginine, and tryptophan side chain nitrogen atoms are always h-bond donors. Backbone carbonyl oxygen atoms are always h-bond acceptors. Cysteine sulfur atoms are either protonated or deprotonated if determined to be involved in disulfide bond. All of these are assigned *known* status at initialization. All other sidechain heavy atoms that can act as donors or acceptors (N and O) are assigned *unknown* status. Residues with sidechains that contain an *unknown* heavy atom or for which there are possible alternate rotamer states are considered *ambiguous*.

Step 2 (Termination Check): If there are *ambiguous* residues present in the structure then the procedure continues to step 3. If there are no *ambiguous* residues, the structure considered fully determined and proceeds to STEP 7.

Step 3 (Environment Evaluation): Each *ambiguous* residue is sequentially investigated. For each residue, the energetics of all its possible RP states as well as all RP states of *proximal* neighbor residues are estimated (**Section 2.2.5**). If one RP state is found to be at least 1 kcal more favorable than all other possible states then that RP state for the residue is set and the status of the residue is switched from *ambiguous* to *known*. Otherwise, the status of the residue is maintained as *ambiguous*. In either case, the next residue is then investigated. Step 3 is concluded after all *ambiguous* residues have been investigated. Details of the energetic estimates are fully described in section 2.2.3. If a residue has no *proximal* neighbors then it is considered *fully solvated* and set as *known* and maintains the assignments as input.

Step 4: If the number of *ambiguous* residues changed in Step 3, the protocol returns to Step 2 and proceeds from there. Otherwise, the procedure advances to Step 5.

---

[b] We use the italicized term "*ambiguous*" to denote that the protonation/rotamer state is not yet determined as *known* by the RAPA protocol.

Step 5 (Branching): If this step is reached, then the protocol has determined that there is at least one residue with two or more RP states that are deemed to be consistent with the crystal structure. We term these as *"degenerate"* states which consist of the RP state with the lowest assessed energetic environment (Section 2.2.5) and all other RP states that have an environmental energy within one kcal of this lowest value. At this step, the first *degenerate* residue in sequence is identified. A copy of the system (*known and ambiguous status*) is made for each *degenerate* RP state of the residue in which that RP state is assigned as *known*. The tool then proceeds to step 6. We refer to this process as *branching* and each copy as a *branch*.

Step 6: A loop is initiated in which each branched configuration is processed through the protocol with each branch being independently processed initiating at step 2 to check if there are remaining *ambiguous* residues. Each new configuration continues the process until all residues are assigned *known* RP states with no remaining *ambiguous* residues in which case the configuration is entirely determined and output in PDB format (Step 7).

Step 7: The configuration is entirely determined and output into PDB format and the workflow for this *branch* is terminated.
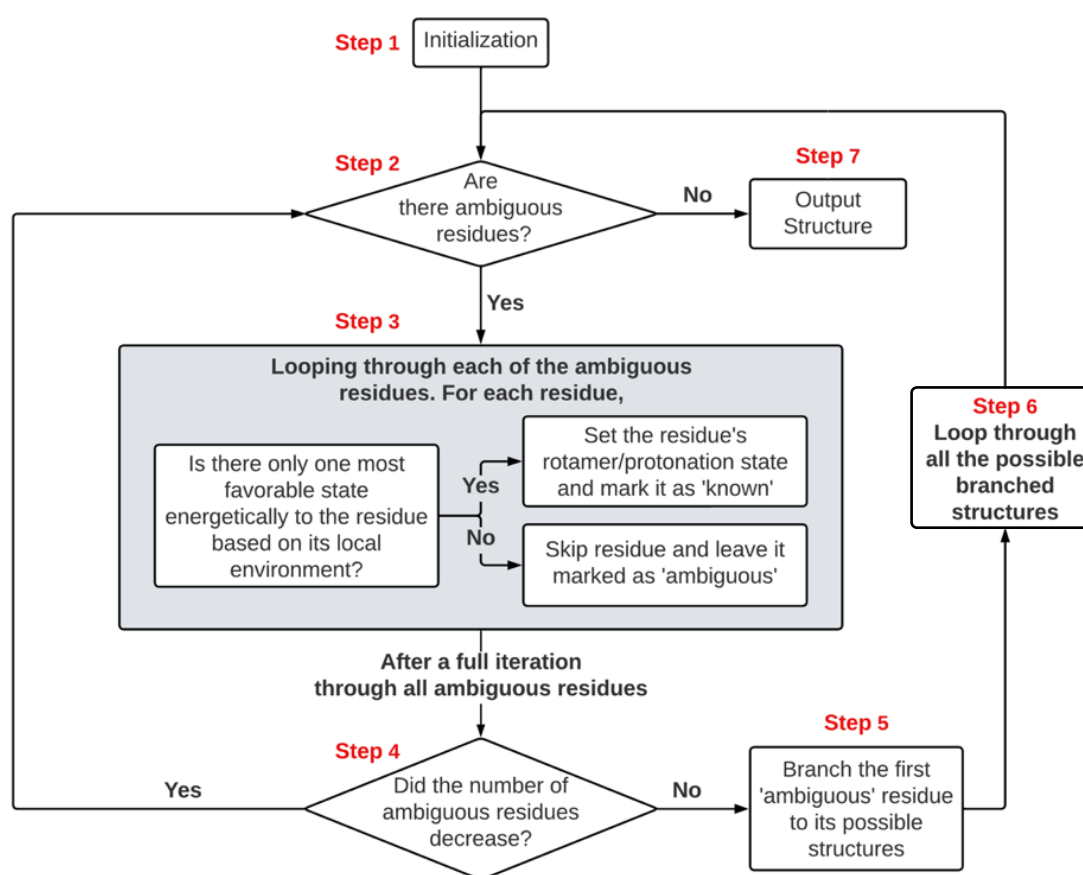

Figure 1. Flowchart demonstrating the sequence of operations performed by the RAPA methodology. *The process begins with an input protein structure, and it terminates with one or more RAPA states output in PDB format.*

2.2.2 Treatment of the Different Residue Types

Known Residues

The following residues are assigned as *known* in the initialization (STEP 1). Nonpolar residues such as Glycine (GLY), Alanine (ALA), Valine (VAL), Leucine (LEU), Isoleucine (ILE), Methionine (MET), Proline (PRO), and Phenylalanine (PHE) as they have no side chain donors or acceptors. Tryptophan (TRP) side chain nitrogen atoms are always considered protonated. Lysine (LYS) and Arginine (ARG) side chain nitrogen atoms are always assigned the

positively charged protonation state. Cysteine sulfur atoms are protonated unless they are within 2.3Å of a neighboring CYS sulfur atom in which case they are deprotonated set to be participating in a disulfide bond. In either case the CYS residue is assigned *known* status.

Histidine Residues

HIS residues are initially assigned *ambiguous* as there is uncertainty in both their protonation and rotamer states. Local energetics are evaluated for the six RP states which are all combinations of the 2 possible rotamer states and 3 protonation states (HIE, HID, and HIP).

Asparagine and Glutamine Residues

ASN and GLN residues are initially assigned as *ambiguous* as there is uncertainty in their rotamer states. Local energetics are evaluated for both possible rotamer states for these residues.

Aspartate and Glutamate Residues

ASP and GLU residues are initially assigned *ambiguous* in their protonation state. They are considered negatively charged (unprotonated) unless they are part of an acid dyad. If two ASP or GLU residues are proximal to each other, defined as having carboxylate oxygen atoms within 3.8Å, then the energetics of all possible singly protonated configurations is evaluated (Figure S2 in Supp. Info).

Serine, Tyrosine, and Threonine Residues

SER, THR, and TYR residues are initially assigned to be *ambiguous*. Their hydroxy groups can act as both donors and acceptors, however, there is uncertainty in the position of the proton which determines the directionality of the donor and acceptor interactions. For SER and THR residues local energetics are evaluated for 9 potential positions of the hydroxy hydrogen. The hydrogen atom with the shortest distance from the neighboring heavy atom is the position of the hydrogen atom evaluated energetically. The 9 possible positions include the three staggered positions corresponding to their optimal sp3 hybridization as well (±20° rotations) from each of these three positions (see **Figure 4**). The hydroxy groups of TYR residues are planar about the aromatic ring due to bond resonance. The energetic evaluations position the H atom at 2 distinct positions that are planar with the aromatic ring (180° rotation along the C-O axis).

2.2.3 Local Environment Energy Estimations

Energetics of h-bonds and electrostatic clashes are evaluated only for residues for which the distance between polar heavy atoms is *proximal* which we define here to be within 3.8Å.

H-bond energies are estimated using a lookup table based on heavy atom – heavy atom (hv–hv) distances and heavy atom – hydrogen – heavy atom (hv–h–hv) angles. The lookup table was constructed using a potential energy surface (PES) calculation generated using Schrodinger Jaguar[33] on a set of configurations of a water dimer with Oxygen-Oxygen (O-O) distances ranging from 2.5Å - 4.0Å with 0.1Å intervals and O⋯H-O angle ranging from 90°-180° with 0.1° intervals. Local h-bond energies for the protein use the hv-hv distance and hv-h-hv angle as a proxy for the O-O distance and O⋯H-O angle. Full details of the Jaguar calculations for constructing the look-up table are in the Supplemental information.

The energy of electrostatic clashes (e.g. between two h-bond donors or two h-bond acceptors) is simply estimated to be the negative of the energy of making a favorable h-bond interaction between donor acceptor pairs. This is illustrated in **Figure 2** for which the estimated energetics of the O-O clash (**Figure 2**, left) is simply estimated to be the inverse of the favorable Amino-Carbonyl interaction in the same position (**Figure 2**, right).

The energetics of a given RP state is estimated by summing up all local polar interactions as shown illustrated in **Figure 2**.
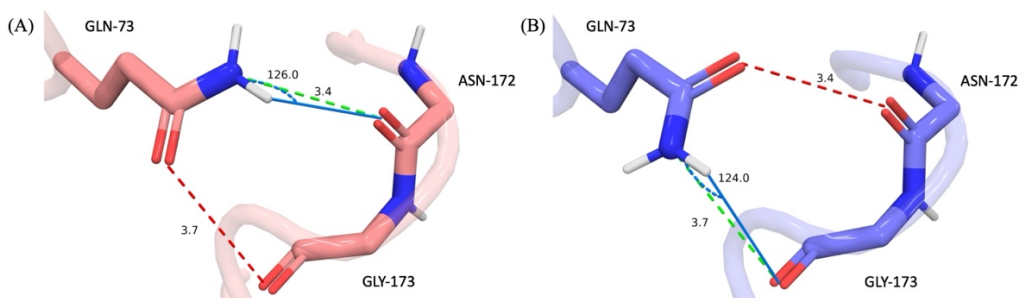
Figure 2. The interactions of beta-site amyloid cleaving enzyme GLN-73 (PDB entry 3L5D[34]) Here we evaluate the local energetics of two possible rotamer states of GLN-73. In each state, GLN-73 forms a h-bond interaction (favorable, green dashed lines) and an electrostatic clash (unfavorable, red dashed lines). The left configuration was found to have lower total energy as the h-bond is stronger than the electrostatic clash. However, the difference in local energies for the two rotamer states of GLN-73 (A and B) is estimated to be 0.76 kcal/mol. Thus, GLN-73 is considered degenerate and both RP states further evaluated in the protocol.

2.2.4 Evaluation of RP states for a residue

In step 3 of the workflow, each *ambiguous* residue is sequentially investigated such that the local energetics are tabulated for each potential RP state in each environment provided by all possible configurations of proximal residues.

As an example of this, consider an RP state from the human dopamine D3 receptor (D3R, PDB: 3PBL[35] that contains two residues ASN-1132 and SER-1117 (**Figure 3**) with the amide of the ASN interacting with the hydroxy of the SER. ASN-1132 has two possible rotamer states and the SER-1117 hydroxy group can interact with ASN-1132 as either a h-bond donor or acceptor resulting in four possible RP states.

Of these 4 possible configurations, two are unfavorable (**Figure 3b** and **Figure 3d**) containing electrostatic clashes and two are favorable forming h-bonds between the two residues (**Figure 3a** and **Figure 3c**). The two favorable configurations have energetics estimated to be within 0.85 kcal/mol and thus the two RP states of ASN-1132 are considered to be energetically *degenerate*.
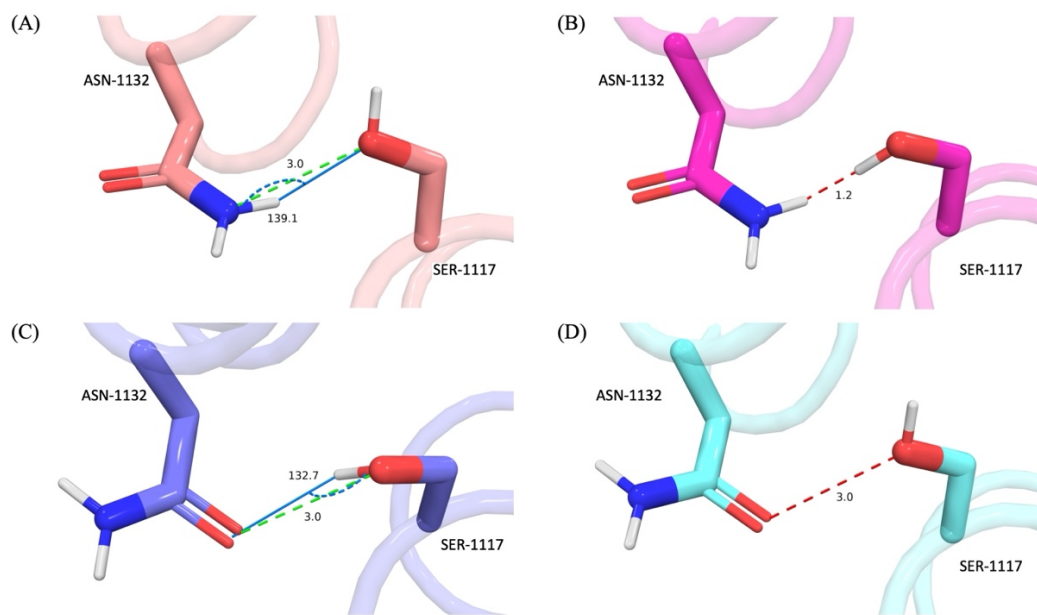
Figure 3. The interactions of Dopamine receptor 3 residues ASN-1132 and SER-1117 (PDB entry 3PBL[35]). *A-D show the 4 environments evaluated by RAPA. Those 4 environments arise from 2 rotamer states of ASN-1132 and 2 possible states of SER-1117. Electrostatic clashes shown as red dashed lines, h-bond interactions shown as green dashed lines. H-bond angles displayed as marine line and arc. The difference in local energetics between environments A and C was found to be 0.86 kcal/mol, thus, both environments are deemed energetically viable.*

When the hydroxy of SER-1117 donates a h-bond to the ASN-1132 amide oxygen, 9 possible positions of the hydroxy hydrogen are investigated in the energy assessment (**Figure 4**). The hydrogen position used for the energy evaluation is the position that results in the shortest h-bond distance to the oxygen acceptor. In this case, the hydrogen atom labeled P7 makes the shortest distance to the amide oxygen of ASN-1132 (**Figure 3c**).
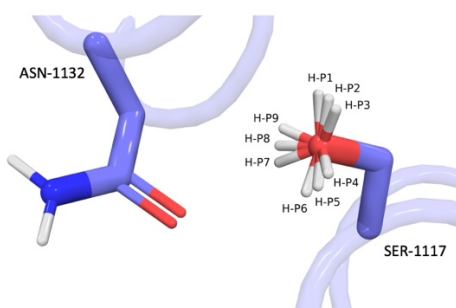

Figure 4. Determining the hydroxy directionality of SER-1117 when it donates a h-bond to ASN-1132 amide oxygen. *The 9 potential position of the hydrogen atom are labeled H-P1 to H-P9. In this example H-P7 is the shortest distance to the oxygen acceptor and is used in the energy evaluation.*

2.2.5 An example of branching to alternate configurations.

Here, we illustrate how the *branching* into separate configurations occurs in Step 5 when an RP state is *degenerate*. **Figure 5** shows the interactions of two *ambiguous* asparagine residues from Estrogen receptor β (PDB entry 2FSZ[36]). In every evaluation of Step 3, both rotamer states are considered *degenerate* as the configurations in Panel A and C are within 1 kcal of each other. Eventually, at Step 5, two configurations of the protein are branched. In the first branch, ASN-431 of chain A is set to *known* in the configuration shown in panel A. In the second branch, the configuration is set to be *known* with the configuration shown in panel C. Each branch is fed through the protocol separately. For the first branch, when ASN-457 of chain B is investigated, its two rotamer states are evaluated interacting with ASN-431 of chain A and the possible states are **Figure 5a** & **Figure 5b**. It is found that the energy of A is more than 1 kcal more favorable than B and for the first branch, ASN-431 (chain A) and ASN-457 (chain B) are *known* in the configuration shown in panel A. The same procedure for the second branch yields configuration of C as *known* for both residues. In this case, configurations representing both panels A and C are represented in the set of final configurations output by the protocol.
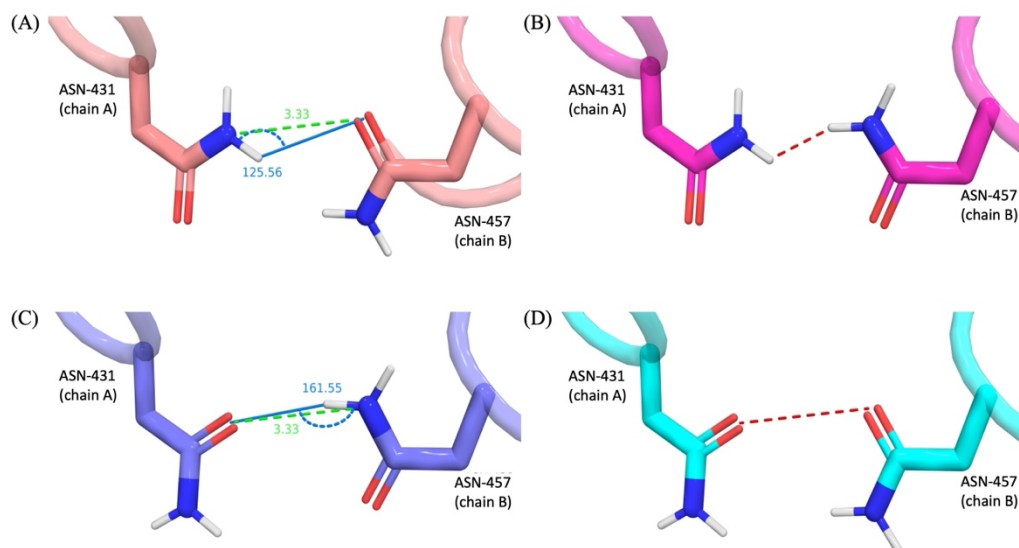
7

Figure 5. The potential interactions of two ambiguous asparagine residues from Estrogen receptor β (PDB entry 2FSZ[36]). *Here, each asparagine has two possible rotamer states yielding four possible configurations shown in the four panels. Electrostatic clashes shown as red dashed lines, h-bond interactions shown as green dashed lines. H-bond angles displayed as marine line and arc. The energy of each proximal interaction is estimated using the angle and distance in a pre-calculated lookup table. In this case, the configuration in Panel C is the lowest energy and the configuration in Panel A is within 1 kcal of this value so the residue remains ambiguous.*

2.3 Molecular Dynamics Simulations

We conducted two sets of MD simulations to evaluate RAPA proposed configurations. The first set (section 2.3.2.1) evaluated whether a proposed configuration complexed with the cognate ligand remained stable in unrestrained MD simulations. The second set of simulations (section 2.3.3.1) considered *ambiguous* residues for which there were two or more viable RP states and assessed each states stability to determine the consistency of RAPA proposed states. This was performed with simulations wherein the *ambiguous* side chains were allowed to rotate freely between each RP state.

2.3.1 Initial Preparation and Forcefield Parameterization

Small molecule ligands were parameterized using version 22.0 for both Amber[37] and Ambertools[38]. Atom types were assigned with *antechamber*[39], with AM1-BCC[40] as the charge model, and "General Amber Force Field 2" (GAFF2[41]) as the ligand force field. For the two sets of MD validations, *tleap*[38] was used to prepare and solvate the systems with protein force field ff99SB[42] and OPC[43] water in a rectangular polyhedral box with a 10Å buffer between the box edges and the protein.

2.3.2 Complex Stability Simulations

2.3.2.1 Molecular Dynamics Protocol

The MD preparation followed the protocol described in our previous work[44]. The systems were energetically minimized in a two-step process. The first minimization step was performed with 1500 steps of steepest descent with all protein atoms restrained harmonically using a force constant of 100 kcal/mol·Å^2. For the second minimization step, only the protein heavy atoms were restrained for 1500 steps. The system was heated from 0 to 300 K in a 240 ps NVT simulation with the protein heavy atoms restrained with the same force constant; the temperature was regulated by Langevin thermostat with collision frequency of 1 ps. This was followed by a 20 ns NPT equilibration simulation with the atom restraints uniformly reduced from 100 to 2 kcal/mol·Å^2 over 10 ns. The output from this was used as the input configuration for the production phase MD simulations. In the production phase, the temperature was regulated via a Langevin thermostat set to 300 K with a collision frequency of 2 ps. The pressure (1 atm) was maintained by isotropic position scaling with a relaxation time of 0.5 ps.

During the 10 ns production phase, the protein-ligand complex was simulated with restraints applied to the alpha carbons of three disparate residues. These restrained alpha carbons had low B-factors and were located at least 10 Å away from the crystallized ligand. We manually selected the three restraint sites from different alpha helices. If there were not enough alpha helices, we chose beta sheets, and if neither were available, we selected low B-factor residues from the protein core. The restraints applied had a cartesian weight of 2 kcal/mol·Å^2, which prevented the protein from rotating or translating. This protocol[45], which we will refer to as 3 distal site restraints (3DS), was designed to prevent global translation and rotation of the protein while allowing structural fluctuations.

2.3.2.2 Post-processing and RMSD analysis

For each MD simulation, we used CPPTRAJ[46] to generate the time-averaged structure of the 40,000 snapshots in the trajectory. The time-averaged structure refers to a representative structure in which the reported atom positions are the average position over all snapshots. The time-averaged MD structure was aligned to the crystal structure with CPPTRAJ and the RMSD was calculated from the distances between the backbone heavy atoms (Atom names in Amber: N, CA, and C) between these two aligned structures. In this analysis atoms that have no B-factor data were excluded. RMSDs are reported for subsets of included atoms with the lowest 50%, 70%, and 90% of B-factor values.

2.3.3 Rotamer States Simulations

2.3.3.1 Molecular Dynamics Protocol

The protocol of this MD set is the same as described in the previous section (section 2.3.2.1) except for the following: the 20 ns NPT equilibration reduces the atom restraints from 100 to 0.10 kcal/mol·Å^2 over the first 10 ns. In the 10ns production runs, cartesian restraints are applied on all protein heavy atoms except for the side chains of the *ambiguous* residues being evaluated, such as amide groups of ASN/GLN residues and the imidazole rings of HIS residues. The production run restraints strengths are set to of 0.1 kcal/mol·Å^2. This maintains the crystal 3D structure but allows these residues' amide groups and imidazole rings to rotate.

2.3.3.2 Molecular Dynamics Analysis

We use CPPTRAJ[46] to extract the dihedral angle distribution of the side chains of ASN, GLN and HIS across the trajectories. The dihedral angles in Amber atom names are CA-CB-CG-OD1 for ASN residues, CB-CG-CD-OE1 for GLN residues, and CA-CB-CG-ND1 for HIS residues. We fit each dihedral angle distribution to Gaussian functions and a Gaussian curve representing a rotamer state was considered significant if its population was more than 10% of the simulation. To examine the sampling and potential interconversion between possible rotamer states, the distributions of dihedrals were overlaid in a single plot for each residue.

2.3.4 Labeling Ambiguous Residues

When the procedure is completed for a given protein, one of three labels is assigned to each ASN, GLN, and HIS residue. A residue is labeled as *fully solvated* if it makes no h-bond interactions with the protein. If a residue makes one or more h-bond interactions with the protein and no alternate RP states have energies within 1 kcal of the minimum energy state, then it is labeled as *fixed*. Finally, a residue is labeled as *degenerate* if it has multiple RP states that are within 1 kcal of the minimum energy RP state.

3. Results

3.1 RAPA Proposed Alternate Configurations

In the 77 proteins investigated here, there were a total of 2274 ASN/GLN residues and 780 HIS residues. We find that the majority of these are *fixed* (84% of ASN/GLN and 89% of HIS) with only one energetically viable RP state. However, we find that a small percentage of the ASN/GLN residues (4%) and HIS residues (10%) have *degenerate* RP states. The remainder of the residues are *fully solvated* and make no intramolecular h-bond interactions (**Table 1**).

Table 1. RAPA labels of the ASN, GLN, and HIS residues in the 77 structures

Total # of residues	# *degenerate* residues	# *fixed* residues	# *fully solvated* residues
2274 ASN/GLN	87	1918	269
780 HIS	79	694	7

9

We found that 14 systems (18%) had one configuration and 68 (88%) of the systems had 32 or fewer configurations (**Figure 6**). All of the 9 systems with more than 32 configurations were homo-dimers, -trimers, and -tetramers and for each monomer the number of proposed configurations was 24 or fewer (**Figure 6**).
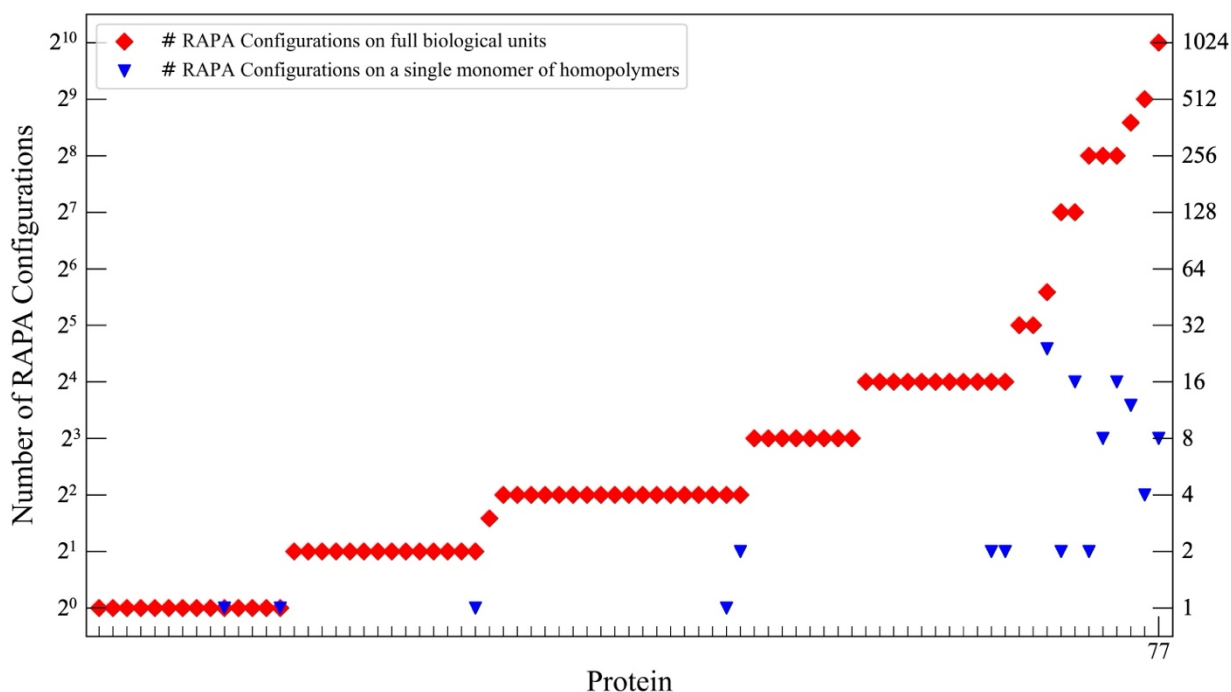

Figure 6. The total number of unique configurations determined by RAPA for 77 protein systems *when the full biological unit of the system is evaluated (red diamonds) versus when one single repeating monomer is considered for systems containing homopolymers (blue triangles).*

3.2 Molecular Dynamics Simulations of the RAPA Proposed Structures

3.2.1 RAPA state configurational stability

For each protein target that had 48 or fewer proposed configurations (**Figure 5**), we ran 10 ns MD simulation using the 3DS protocol (**Methods 2.3.2.1**) for each configuration. The restraints of the 3DS protocol[45] were constructed to prevent global rotation and translation of the protein while otherwise allowing protein structural fluctuations.

These simulations assessed whether the RAPA proposed configurations were stable such that these simulations did not result in large structural fluctuations and the fluctuations observed were consistent with the crystallographic data. All 469 MD simulations completed their entire 10ns simulations.

For each of the 469 simulations, we calculated the RMSD between the time-averaged structure and the crystal PDB structure as outlined in the methods sections. For all systems, the RMSD was 2.18 or under and 467/469 systems had RMSD values of under 2.0 (**Figure 7**). Fifty-six percent of the simulations had RMSD's of 1.0Å or lower including many systems with multiple configurations (**Figure 8**). RMSD data for the lowest 50% and lowest 70% of atoms is in the Supplemental Information (**Figures S4 and S5**).
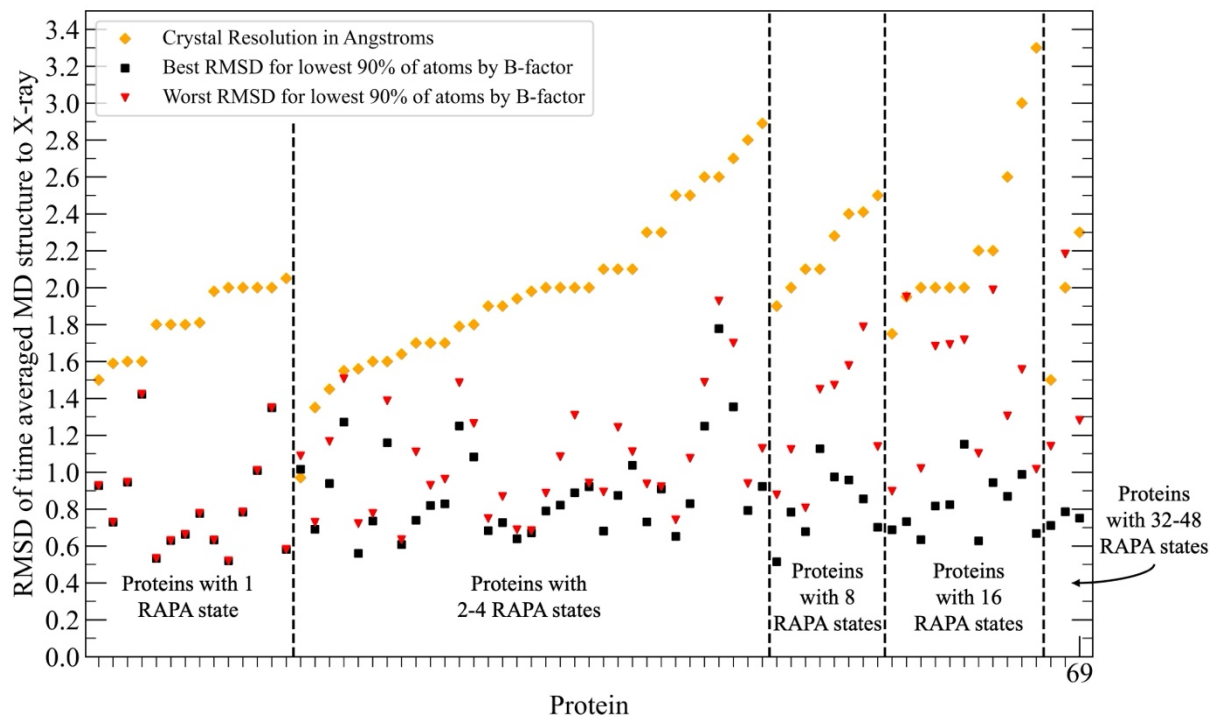
Figure 7. RMSD between time average MD structure and X-Ray coordinates for atoms with the lowest 90% of B-factors. *Best RMSD (black squares) and worst RMSD (red triangles) are shown for each of the 69 targets. X-Ray crystal structure resolution is shown for each target (yellow diamonds).*
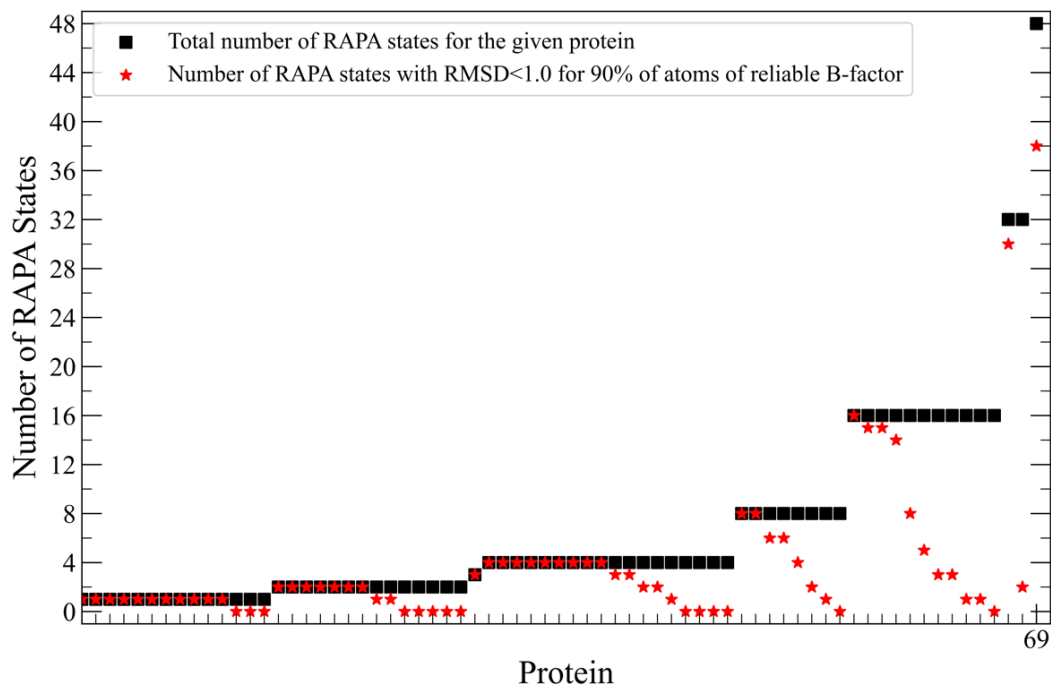

Figure 8. Number of configurations determined by RAPA for 69 systems. *The initial number of states (black squares) and number of those with RMSD less than 1Å for the lowest 90% of B-factors (red stars). For many systems there are multiple RAPA proposed configurations that are consistent with the X-ray structure.*

11

3.2.2 Sampling RAPA Rotamer States in Molecular Dynamics

In this section we evaluate the consistency of energetically *degenerate RP* states identified by RAPA with MD simulations. To assess this, for each system that was predicted to have *degenerate* rotamer states, we ran short (10 ns) MD simulations in which the rotamer states of *degenerate* residues were unrestrained and allowed to freely interchange. For each *degenerate* residue the rotamer states were arbitrarily labeled 'A' and 'B' and separate MD simulations were run in which the initial rotamer states were initialized in either 'A' or 'B'. In these simulations, we tracked the dihedral angles for each *degenerate* residue and characterized three possible outcomes:

1) **Interchange:** Both rotamer states are sampled. In these simulations, the rotamer states of both 'A' and 'B' were sampled regardless of whether the MD simulations had initial rotamer states corresponding to 'A' or 'B' (**Figure 9**). In this case, the RAPA prediction was considered consistent with the MD simulations as both rotamer states were sampled in the MD simulations.

2) **Only the initial state is sampled**: In these simulations, no transitions between rotamer states were observed in the MD simulations. Rotamer states were found to remain in the state they were initialized in (**Figure 10**). In these cases, both rotamer states are considered *stable* over the course of the MD simulation and we assess that the RAPA prediction of *degenerate* states is correct.

3) **Only one state is stable**: In these simulations, the rotamer states of the RAPA predicted *degenerate* residues switched to a single state and remained there for the entirety of the MD simulation (**Figure 11**). In these cases, the residues switched rotamer state during minimization, equilibration, or within the first nanosecond of the production simulation. In this case, the RAPA prediction of degenerate rotamer states is considered inconsistent with the MD simulation or incorrect.
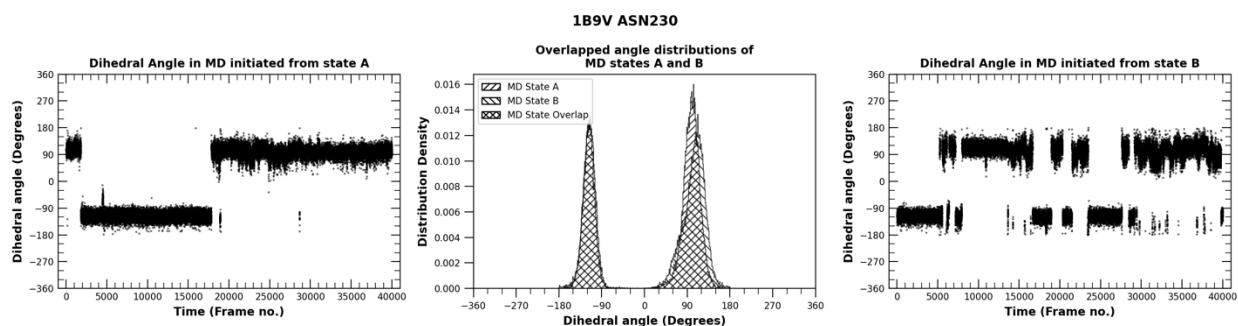

Figure 9. Rotamer States interchange: *Dihedral angle vs. time for MD simulations that started in state 'A' (left) and state 'B' (right). The middle figure has the distributions obtained from the simulations that were initiated in rotamer state 'A' (shaded in forward-slash lines) and state 'B' (shaded in back-slash lines lines). The data is for residue ASN-230 (PDB entry 1B9V[47]). RAPA predicted that these two states were degenerate which is consistent with the MD simulation. Data that appears cross-hatched in the middle figure is because the forward and backlash lines from the distributions of each simulation overlap. Figure S6 in Supplemental Information shows the interaction of the residue in its two rotamer states.*
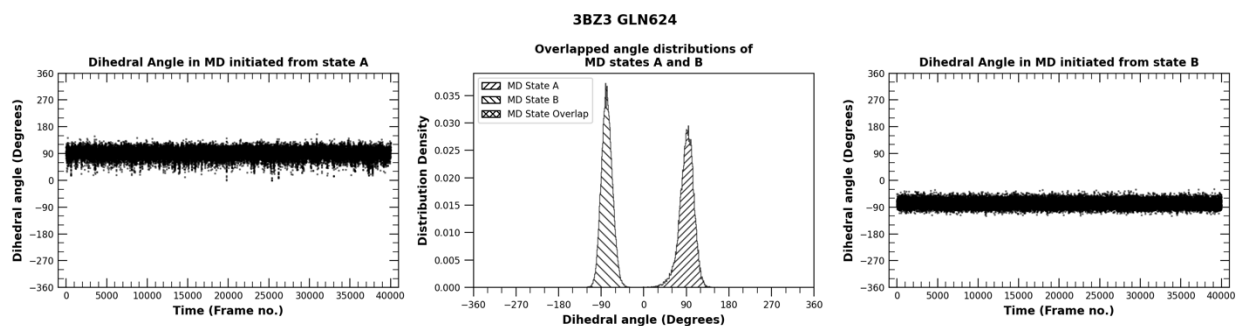
Figure 10. Only the Initial State is Sampled: *Dihedral angle vs. time for MD simulations that started in state 'A' (left) and state 'B' (right). The middle figure has the distributions obtained from the simulations that were initiated in rotamer state 'A' (shaded in forward-slash lines) and state 'B' (shaded in back-slash lines lines). The data is for residue GLN-624 (PDB entry 3BZ3[48]). RAPA predicted that these two states were degenerate which is consistent with the MD simulation. Figure S7 in Supplemental Information shows the interaction of the residue in its two rotamer states.*
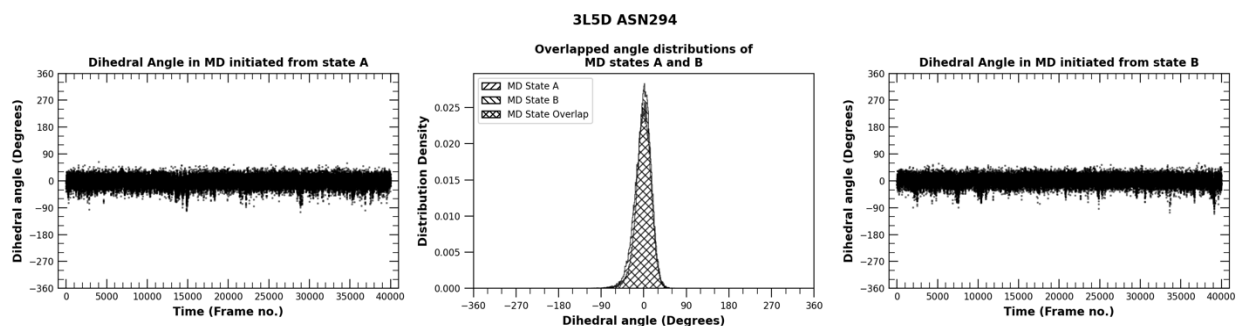

Figure 11. Only One State is Stable: *Dihedral angle vs. time for MD simulations that started in state 'A' (left) and state 'B' (right). The middle figure has the distributions obtained from the simulations that were initiated in rotamer state 'A' (shaded in forward-slash lines) and state 'B' (shaded in back-slash lines). The data is for residue ASN-294 (PDB entry 3L5D[34]). RAPA predicted that these two states were degenerate which is inconsistent with the MD simulation in which only one state is well sampled. Data that appears cross-hatched in the middle figure is because the forward and backlash lines from the distributions of each simulation overlap. Figure S8 in Supplemental Information shows the interaction of the residue in its two rotamer states.*

3.2.2.1 *Degenerate* ASN/GLN Rotamer consistency with molecular dynamics simulations

Of the 87 residues that RAPA predicted to be degenerate, 49 residues (56%) had two stable states in the MD simulations either sampling both states (24/49) or having no interchange between the two states (25/49) (**Table 2**). In these cases, the predictions were considered consistent with the MD simulations and the RAPA predictions were assessed to be correct. On the other hand, for the other 38 residues (44%), only one state was found to be stable in the dynamic simulations and the RAPA predictions were assessed to be incorrect (**Table 2**).

Table 2. MD Analysis of *degenerate* ASN and GLN residues

# of degenerate residues	# of degenerate residues correctly predicted by RAPA in MD	# of residues with outcome #1	# of residues with outcome #2	# of residues with outcome #3
51 ASN	22	13	9	29
36 GLN	27	11	16	9

3.2.2.2 *Degenerate* HIS Rotamer and Protonation State MD Analysis

Seventy-nine HIS residues were predicted to be *degenerate*. Of these, 73 (92%) were found to be stable in each of the multiple degenerate states predicted by RAPA in their respective MD simulations (**Table 3**). We note that we considered a protonation state to be unstable if the rotamer state moved at least 30 degrees away from the predicted stable state.

Table 3. MD Analysis of *degenerate* RAPA predicted HIS residues

HIS residues degenerate in …	# RAPA-predicted degenerate residues	# residues inherently *stable* in their *degenerate* states over the course of the MD simulations	# residues stable in only one of the RAPA predicted states
Two rotamer states	20	18	2
Two protonation states	16	14	2
Two states[a]	35	34	1
More than two states[a]	8	7	1[b]

[a] Corresponds to states that differ in both rotamer and protonation state.
[b] For HIS159 in 1D3G, three states were predicted to be *degenerate* however only two were stable in MD

4. Discussion

The inability of X-ray scattering experiments to resolve hydrogens and differentiate between oxygen, carbon, and nitrogen atoms leaves ambiguity as to the *correct* configuration of protein sidechains. Indeed, there may not be one correct configuration and there may be multiple configurations that are found in protein crystals. To address this a number of methods have been developed that attempt to identify the precise atomic configuration, however, most of these methods identify a single configuration. Here, we have introduced a protocol, RAPA, that attempts to identify all configurations of the protein that are energetically accessible and consistent with the resolved X-ray data. The protocol does this by examining the local h-bond networks of *ambiguous* side chains and quickly approximates the energetics of potential alternate configurations. When implemented computationally, the RAPA protocol takes several minutes to fully output a set of structures that are consistent with the X-ray structure and determined to be energetically *degenerate.*

We applied the RAPA protocol to 77 protein structures and identified that 63 of these had multiple potential configurations resulting in a total of 469 configurations. In short 10 ns MD simulations, 467 of these configurations maintained conformations close to the crystal structure with time averaged RMSDs within 2.0Å or less. These results suggest that the structures produced by RAPA are stable and consistent with their crystal structures. We also evaluated RAPA's predictions of *degenerate* rotamer states of ASN, GLN, and HIS residues in 10 ns MD simulations in which the rotamer states were unrestrained and found that 56% of the ASN/GLN and 92% of the HIS had two or more stable states matching these predictions.

The fundamental premise of structure based drug discovery is that a small molecule lead needs to be electrostatically complementary to the protein surface making h-bond or hydrophobic contacts where appropriate. When the rotamer or protonation states of *ambiguous residues* residues change, the positions or character of these contacts change and, correspondingly, the chemical matter that is complementary to the surface differs. Limiting computer aided drug discovery workflows to a single configuration of a protein effectively limits the search for viable leads to chemical matter that is complementary to that configuration. Conversely, if every possible configuration of these side chains is considered there ends up being a combinatorial explosion and investigating all possible configurations is intractable in a computer aided drug discovery workflow. RAPA addresses this problem by efficiently identifying only those configurations that are energetically accessible and discarding those that are not. Importantly, for 62 of the 77 systems investigated here, 8 or fewer energetically accessible configurations were identified and the largest number of such *degenerate* configurations was 24 per monomer. This suggests that investigating all configurations predicted by RAPA can be feasible in a drug discovery workflow context.

5. Acknowledgements

6. References

(1) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(2) Aparoy, P.; Kumar Reddy, K.; Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors. *Curr. Med. Chem.* **2012**, *19* (22), 3763–3778. https://doi.org/10.2174/092986712801661112.

(3) Kontoyianni, M. Docking and Virtual Screening in Drug Discovery. In *Proteomics for Drug Discovery*; Lazar, I. M., Kontoyianni, M., Lazar, A. C., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2017; Vol. 1647, pp 255–266. https://doi.org/10.1007/978-1-4939-7201-2_18.

(4) Wang, X.; Song, K.; Li, L.; Chen, L. Structure-Based Drug Design Strategies and Challenges. *Curr. Top. Med. Chem.* **2018**, *18* (12), 998–1006. https://doi.org/10.2174/1568026618666180813152921.

(5) Vriend, G. WHAT IF: A Molecular Modeling and Drug Design Program. *J. Mol. Graph.* **1990**, *8* (1), 52–56. https://doi.org/10.1016/0263-7855(90)80070-V.

(6) Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. Errors in Protein Structures. *Nature* **1996**, *381* (6580), 272–272. https://doi.org/10.1038/381272a0.

(7) Hooft, R. W. W.; Sander, C.; Vriend, G. Positioning Hydrogen Atoms by Optimizing Hydrogen-Bond Networks in Protein Structures. *Proteins Struct. Funct. Genet.* **1996**, *26* (4), 363–376. https://doi.org/10.1002/(SICI)1097-0134(199612)26:4<363::AID-PROT1>3.0.CO;2-D.

(8) Bass, M. B.; Hopkins, D. F.; Jaquysh, W. A. N.; Ornstein, R. L. A Method for Determining the Positions of Polar Hydrogens Added to a Protein Structure That Maximizes Protein Hydrogen Bonding. *Proteins Struct. Funct. Bioinforma.* **1992**, *12* (3), 266–277. https://doi.org/10.1002/prot.340120305.

(9) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793. https://doi.org/10.1006/jmbi.1994.1334.

(10) 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 1 1Edited by J. Thornton. *J. Mol. Biol.* **1999**, *285* (4), 1735–1747. https://doi.org/10.1006/jmbi.1998.2401.

(11) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogen Atoms 1 1Edited by J. Thornton. *J. Mol. Biol.* **1999**, *285* (4), 1711–1733. https://doi.org/10.1006/jmbi.1998.2400.

(12) Davis, I. W.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MOLPROBITY: Structure Validation and All-Atom Contact Analysis for Nucleic Acids and Their Complexes. *Nucleic Acids Res.* **2004**, *32* (Web Server), W615–W619. https://doi.org/10.1093/nar/gkh398.

(13) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: All-Atom Contacts and Structure Validation for Proteins and Nucleic Acids. *Nucleic Acids Res.* **2007**, *35* (Web Server), W375–W383. https://doi.org/10.1093/nar/gkm216.

(14) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. *MolProbity* : All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (1), 12–21. https://doi.org/10.1107/S0907444909042073.

(15) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33* (Web Server), W368–W371. https://doi.org/10.1093/nar/gki464.

(16) Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of p K s in Macromolecules. *Proteins Struct. Funct. Bioinforma.* **2006**, *63* (4), 928–938. https://doi.org/10.1002/prot.20922.

(17) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40* (W1), W537–W541. https://doi.org/10.1093/nar/gks375.

(18) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pK $_a$ Values. *Proteins Struct. Funct. Bioinforma.* **2005**, *61* (4), 704–721. https://doi.org/10.1002/prot.20660.

(19) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of p K_a Values for Protein–Ligand Complexes. *Proteins Struct. Funct. Bioinforma.* **2008**, *73* (3), 765–783. https://doi.org/10.1002/prot.22102.

(20) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *J. Chem. Theory Comput.* **2011**, *7* (7), 2284–2295. https://doi.org/10.1021/ct200133y.

(21) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537. https://doi.org/10.1021/ct100578z.

(22) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* **2007**, *35* (Web Server), W522–W525. https://doi.org/10.1093/nar/gkm276.

(23) Weichenberger, C. X.; Sippl, M. J. NQ-Flipper: Recognition and Correction of Erroneous Asparagine and Glutamine Side-Chain Rotamers in Protein Structures. *Nucleic Acids Res.* **2007**, *35* (Web Server), W403–W406. https://doi.org/10.1093/nar/gkm263.

(24) Weichenberger, C. X.; Sippl, M. J. NQ-Flipper: Validation and Correction of Asparagine/Glutamine Amide Rotamers in Protein Crystal Structures. *Bioinformatics* **2006**, *22* (11), 1397–1398. https://doi.org/10.1093/bioinformatics/btl128.

(25) Lippert, T.; Rarey, M. Fast Automated Placement of Polar Hydrogen Atoms in Protein-Ligand Complexes. *J. Cheminformatics* **2009**, *1* (1), 13. https://doi.org/10.1186/1758-2946-1-13.

(26) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminformatics* **2014**, *6* (1), 12. https://doi.org/10.1186/1758-2946-6-12.

(27) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75* (1), 187–205. https://doi.org/10.1002/prot.22234.

(28) Li, Y.; Roy, A.; Zhang, Y. HAAD: A Quick Algorithm for Accurate Prediction of Hydrogen Atoms in Protein Structures. *PLoS ONE* **2009**, *4* (8), e6701. https://doi.org/10.1371/journal.pone.0006701.

(29) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided Mol. Des.* **2013**, *27* (3), 221–234. https://doi.org/10.1007/s10822-013-9644-8.

(30) Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2017**, *57* (7), 1511–1516. https://doi.org/10.1021/acs.jcim.7b00190.

(31) Spruce 1.5.2.1. http://eyesopen.com.

(32) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

(33) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A High-performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *Int. J. Quantum Chem.* **2013**, *113* (18), 2110–2142. https://doi.org/10.1002/qua.24481.

(34) Zhu, Z.; Sun, Z.-Y.; Ye, Y.; Voigt, J.; Strickland, C.; Smith, E. M.; Cumming, J.; Wang, L.; Wong, J.; Wang, Y.-S.; Wyss, D. F.; Chen, X.; Kuvelkar, R.; Kennedy, M. E.; Favreau, L.; Parker, E.; McKittrick, B. A.; Stamford, A.; Czarniecki, M.; Greenlee, W.; Hunter, J. C. Discovery of Cyclic Acylguanidines as Highly Potent and Selective β-Site Amyloid Cleaving Enzyme (BACE) Inhibitors: Part I—Inhibitor Design and Validation. *J. Med. Chem.* **2010**, *53* (3), 951–965. https://doi.org/10.1021/jm901408p.

(35) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Won Han, G.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2010**, *330* (6007), 1091–1095. https://doi.org/10.1126/science.1197410.

(36) Wang, Y.; Chirgadze, N. Y.; Briggs, S. L.; Khan, S.; Jensen, E. V.; Burris, T. P. A Second Binding Site for Hydroxytamoxifen within the Coactivator-Binding Groove of Estrogen Receptor β. *Proc. Natl. Acad. Sci.* **2006**, *103* (26), 9908–9911. https://doi.org/10.1073/pnas.0510596103.

(37) D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, N. Forouzesh, G. Giambaşu, T. Giese, M.K. Gilson, H. Gohlke, A.W. Goetz, J. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, X. Wu,

Y. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, Q. Zhu, and P.A. Kollman. *Amber 2023*; University of California, San Francisco, 2023.

(38) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M. AmberTools. *J. Chem. Inf. Model.* **2023**, *63* (20), 6183–6191. https://doi.org/10.1021/acs.jcim.3c01153.

(39) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.

(40) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

(41) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. A Fast and High-Quality Charge Model for the next Generation General AMBER Force Field. *J. Chem. Phys.* **2020**, *153* (11), 114502. https://doi.org/10.1063/5.0019056.

(42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.

(43) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5* (21), 3863–3871. https://doi.org/10.1021/jz501780a.

(44) Olson, B.; Cruz, A.; Chen, L.; Ghattas, M.; Ji, Y.; Huang, K.; Ayoub, S.; Luchko, T.; McKay, D. J.; Kurtzman, T. An Online Repository of Solvation Thermodynamic and Structural Maps of SARS-CoV-2 Targets. *J. Comput. Aided Mol. Des.* **2020**, *34* (12), 1219–1228. https://doi.org/10.1007/s10822-020-00341-x.

(45) From Discussion with Dan McKay@Ventus Therapeutics and Chris Bayly @ Cadence, OpenEye

(46) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.

(47) Finley, J. B.; Atigadda, V. R.; Duarte, F.; Zhao, J. J.; Brouillette, W. J.; Air, G. M.; Luo, M. Novel Aromatic Inhibitors of Influenza Virus Neuraminidase Make Selective Interactions with Conserved Residues and Water Molecules in the Active Site 1 1Edited by I. A. Wilson. *J. Mol. Biol.* **1999**, *293* (5), 1107–1119. https://doi.org/10.1006/jmbi.1999.3180.

(48) Roberts, W. G.; Ung, E.; Whalen, P.; Cooper, B.; Hulford, C.; Autry, C.; Richter, D.; Emerson, E.; Lin, J.; Kath, J.; Coleman, K.; Yao, L.; Martinez-Alsina, L.; Lorenzen, M.; Berliner, M.; Luzzio, M.; Patel, N.; Schmitt, E.; LaGreca, S.; Jani, J.; Wessel, M.; Marr, E.; Griffor, M.; Vajdos, F. Antitumor Activity and Pharmacology of a Selective Focal Adhesion Kinase Inhibitor, PF-562,271. *Cancer Res.* **2008**, *68* (6), 1935–1944. https://doi.org/10.1158/0008-5472.can-07-5155.