JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 2751−2760

pubs.acs.org/JCTC

Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design

Jack B. Maguire,[†,▽] Scott E. Boyken,[‡,§,▽] David Baker,[‡,§,∥] and Brian Kuhlman*[,⊥,#]

[†]Program in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

[‡]Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States

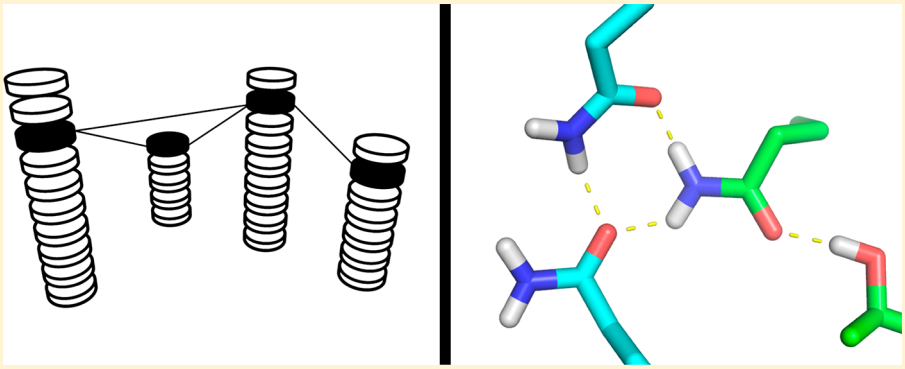
[§]Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States

[∥]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States

[⊥]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

[#]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

Supporting Information



ABSTRACT: Hydrogen bond networks play a critical role in determining the stability and specificity of biomolecular complexes, and the ability to design such networks is important for engineering novel structures, interactions, and enzymes. One key feature of hydrogen bond networks that makes them difficult to rationally engineer is that they are highly cooperative and are not energetically favorable until the hydrogen bonding potential has been satisfied for all buried polar groups in the network. Existing computational methods for protein design are ill-equipped for creating these highly cooperative networks because they rely on energy functions and sampling strategies that are focused on pairwise interactions. To enable the design of complex hydrogen bond networks, we have developed a new sampling protocol in the molecular modeling program Rosetta that explicitly searches for sets of amino acid mutations that can form self-contained hydrogen bond networks. For a given set of designable residues, the protocol often identifies many alternative sets of mutations/networks, and we show that it can readily be applied to large sets of residues at protein−protein interfaces or in the interior of proteins. The protocol builds on a recently developed method in Rosetta for designing hydrogen bond networks that has been experimentally validated for small symmetric systems but was not extensible to many larger protein structures and complexes. The sampling protocol we describe here not only recapitulates previously validated designs with performance improvements but also yields viable hydrogen bond networks for cases where the previous method fails, such as the design of large, asymmetric interfaces relevant to engineering protein-based therapeutics.

INTRODUCTION

Hydrogen bonds are essential for specifying biomolecular structure, and proteins often employ extensive networks of hydrogen bonds to preorganize catalytic active sites,[1−4] mediate interaction specificity,[5,6] and achieve structure and function with a high level of cooperativity.[7−9] Hydrogen bond networks at protein−protein interfaces help overcome desolvation costs associated with binding while providing polar groups that contribute to the solubility of the unbound monomers. The ability to accurately create new hydrogen bond networks is critical for many problems in protein design, and rational design approaches have successfully achieved networks that specify membrane protein interactions[10] and the coordination of functional metal cofactors;[11−15] however, developing general computational methods for this problem has been challeng-

Received: January 11, 2018
Published: April 13, 2018

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

ing.[16] This is in part because hydrogen bond strength is very sensitive to small perturbations in the relative positions of the atoms forming the hydrogen bond.[17,18] Designing buried hydrogen bonds at protein interfaces has been particularly difficult.[19,20]

A key challenge in designing hydrogen bond networks is ensuring that each polar group in a protein or complex has a hydrogen bond partner or is exposed to solvent. It has been estimated that the energetic cost of burying a hydrogen-bond donor or acceptor that does not have a hydrogen bond partner ("unsatisfied") is 5−6 kcal/mol,[21] which is comparable to the total free energy of unfolding for some proteins. The energy functions that are typically used for computational protein design,[22,23] including the Rosetta Energy Function,[24] are expressed as the sum of pairwise energies, which is important for computational efficiency and algorithmic compatibility.[25] However, networks of hydrogen bonds that span multiple residues are inherently not pairwise decomposable, and evaluating burial and hydrogen bond satisfaction cannot be achieved in a pairwise manner. Hence, conventional protein design algorithms are not well-suited for capturing and evaluating satisfied hydrogen bond networks.

Recently, to enable the computational design of hydrogen bond networks, we developed a sampling protocol (HBNet[26]) in the Rosetta software package[27] that explicitly searches through sequence space and side chain conformational space (rotamers) to find sets of amino acids that can form self-contained hydrogen bond networks. To maximize the number of potential networks that are identified for a given backbone conformation and set of residues, HBNet enumerates through all possible closed networks that can be created with a given rotamer library. We define a "closed network" to be one in which every buried polar group has a hydrogen bond partner. HBNet was experimentally validated by using it to design highly symmetric networks in the center of *de novo* designed coiled coils.[26] Since these initial results, we have begun to apply HBNet to other design problems, and we have found that it scales poorly to larger systems, especially cases where symmetry cannot be used to reduce search space. For larger rotamer libraries or larger numbers of residue positions, we observe that the exhaustive search employed by HBNet often does not complete after several hours, which precludes its use in design pipelines that also involve backbone sampling and docking. Here, we introduce a new Monte Carlo-based algorithm (MC HBNet) that makes it possible to rapidly sample and design viable hydrogen bond networks for larger design problems.

The MC HBnet protocol begins by building a graph in which each node in the graph represents a potential side chain conformation (rotamer) for an amino acid at a specified sequence position, and an edge is drawn between two nodes if a hydrogen bond is formed between the two rotamers. Hydrogen bond networks are then assembled by stochastically traversing the graph and outputting networks that do not leave any buried polar group without a hydrogen bond partner. We show that MC HBNet is able to recapitulate the networks of native protein−protein interfaces and that it can be robustly used with large rotamer libraries. Because MC HBNet can be used with a finer degree of side chain sampling than HBNet, we also show that it can find more favorable networks than HBNet in substantially shorter runtimes. These improvements will allow explicit hydrogen bond network design to be incorporated into more complex, multistage protocols such as *de novo* interface design and enzyme design and are general strategies that can be readily incorporated into modeling packages other than Rosetta.

METHODS

Side Chain Sampling and Identification of Hydrogen Bonds. MC HBNet begins by examining residue pairs and identifying which amino acid mutations and side chain rotamers at the two positions will allow the formation of one of more hydrogen bonds between the residues. All residue pairs within a user-defined set of packable positions are enumerated. The set of amino acids considered at each position are also defined by the user. The protein backbone is held fixed throughout the protocol, and side chain conformations are sampled using a backbone-dependent rotamer library.[28] The user can specify to only consider side chain conformations constructed from the most preferred side chain torsion angles (chi angles) for each rotamer ("base" rotamer), or the rotamer library can be expanded by introducing extra chi sampling. Extra chi sampling builds additional rotamers from base rotamers by varying the side chain's chi angles by an amount determined by statistical measurements of that chi angle's variance in high resolution crystal structures of naturally occurring proteins. The magnitude and frequency of extra chi sampling can be controlled by the user (Table 1).

Table 1. Definitions of Extra-Chi Sampling Levels[a]

label	chi 1	chi 2	chi 3	chi 4
Ø	0	0	0	0
$\chi1$	1	0	0	0
$\chi1\chi2$	1	1	0	0
$\chi1\chi2\chi3$	1	1	1	0
$\chi1\chi2\chi3\chi4$	1	1	1	1

[a]A 0 in any column means that there were no extra samples for that chi angle. A 1 in any column means that chi angle had extra samples at ±1 standard deviation defined by the Dunbrack backbone dependent rotamer library.[28]

After side chain coordinates are calculated for the rotamers being considered at each designable position, each rotamer pair from all the residue pairs are examined to determine if a hydrogen bond is being formed between the rotamers. Hydrogen bonds are detected using Rosetta's standard hydrogen bonding potential, which depends on the distance and relative orientation of the donor and acceptor groups.[17,29] Hydrogen bonds typically score between −0.5 and −1.5 Rosetta Energy Units (REU). For most of this work, we only consider interactions with an energy less than −0.5 as a hydrogen bond. When a hydrogen bond is detected between two rotamers, this information is saved in an interaction graph that is used during the sampling protocol. MC HBNet uses a new data structure called HbondGraph that includes nodes for each rotamer, as well as atom-level information for each hydrogen bond, enabling more efficient organization and lookup as compared to the graph used by the original implementation of HBNet, which has been described previously.[26]

HBondGraph Data Structure. HBondGraph creates a node for every candidate rotamer at each position (Figure 1). An edge is created between every pair of nodes whose corresponding rotamers form a hydrogen bond. Additionally, nodes store information about which other nodes in the HbondGraph are incompatible due to steric clashes between

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

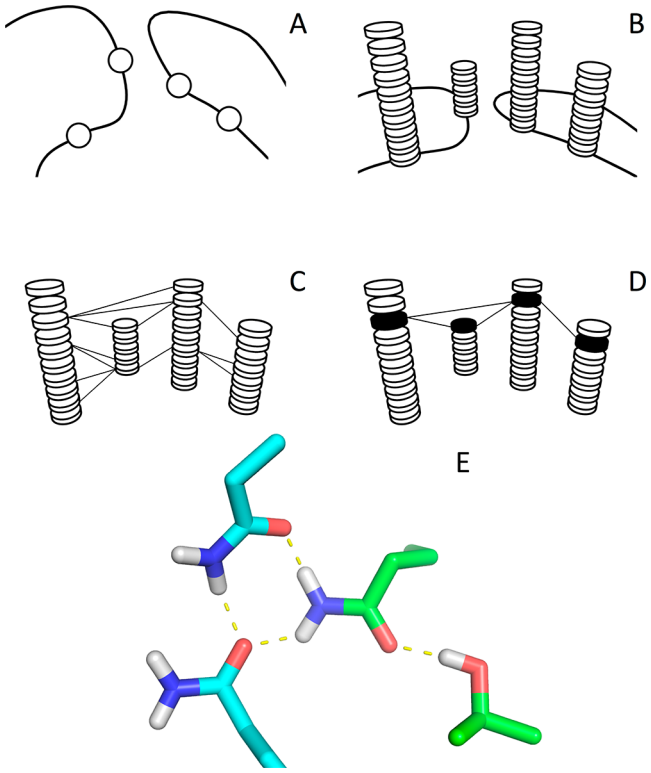


Figure 1. HBondGraph. (A) MC HBNet identifies every residue position that is being designed or repacked. (B) Each position is expanded into an array of graph nodes, one node for every side chain conformation being considered at that position. (C) Graph edges are created between every pair of nodes that form a hydrogen bond. (D) An example of what a hydrogen bond network looks like in this data structure and (E) a possible hydrogen bond network that this example might represent (cyan side chains are part of a different chain than green side chains).

their respective rotamers. A hydrogen bond network can be defined as a set of nodes that form a connected component in the HbondGraph without having any two nodes that represent rotamers that clash or occupy the same residue position.

Each node in the HbondGraph keeps track of every side chain polar atom index in its respective rotamer. MC HBNet strips this atom information for polar atoms that are already satisfied by the background (either implicitly by solvent exposure or explicitly by hydrogen bonds to the backbone or nonpackable side chains, which are held fixed; hydrogen bonds from side chain atoms to backbone atoms are scored and taken into account when evaluating satisfaction). Combining the individual lists of atoms from each node in the network produces an ad hoc checklist of atoms that need to be satisfied for a network to be accepted. For each network, MC HBNet tracks satisfaction by storing a list of all heavy (non-hydrogen) polar atoms that are buried and not satisfied ("heavy unsat"); if a heavy unsat becomes satisfied during network growth, that atom is removed from the list. Satisfaction can be rapidly evaluated because each edge in the HbondGraph stores the atom indices for the acceptor atom, donor atom, and hydrogen atom for every hydrogen bond represented by the edge (there may be multiple hydrogen bonds represented by one edge because a single pair of rotamers can only be connected by one edge but may form multiple hydrogen bonds).

Monte Carlo HbondGraph Traversal. The MC HBNet sampling protocol is composed of a user-defined number of trajectories. Each trajectory begins by randomly selecting a "seed" edge from the HbondGraph, and networks are grown stochastically by adding adjacent edges that lead to compatible nodes. Seed edges can be selected based on predefined starting criteria. An example starting criterion is HBNetStapleInterface,[26] which searches for networks that span across a protein−protein interface; when used with MC HBNet, the HBNetStapleInterface protocol requires that all seed edges must contain at least one node position that is at the interface. Efficiency is improved by restricting sampling to only start at seed edges relevant to the task at hand. Starting criteria can be specified by the user to customize the search for different design scenarios.

After the seed is selected, MC HBNet identifies all of the "candidate" nodes in the HbondGraph that are adjacent to either of the seed's nodes but do not conflict either through steric clashing or sharing a residue position (Figure 1). For each of these candidates, MC HBNet counts the number of HbondGraph nodes ("children") adjacent to the candidate that are compatible with both of the two seed nodes. The relative probability of selecting a candidate to be added to the network is proportional to the number of its compatible children plus one. One candidate is stochastically selected to be added to the network, and the network is registered as a result if it is determined to be absent of heavy unsats. This process is repeated until there are no more adjacent nodes in the HbondGraph that are compatible with every node in the network.

The process described above defines a single MC HBNet trajectory and will be repeated a user-defined number of times (the default trajectory setting is 10^4). MC HBNet trajectories are made faster by keeping track of the heavy unsatisfied polar atoms as the network grows. If the network has any heavy unsats at any point within a trajectory, the MC HBNet sampling protocol will shift its focus to only consider candidate nodes whose addition would result in the satisfaction of a heavy unsat. The trajectory is brought to a premature end if the HbondGraph contains no nodes that can achieve this task. This implementation is represented by the λ term in eq 1 and prevents MC HBNet from wasting time by exploring sample space where finding a fully satisfied network is impossible.

$$f(c_i) = \lambda(\beta + 1) \tag{1}$$

Equation 1 describes $f(c_i)$ as the relative probability of adding candidate c_i (all candidate nodes are shown in blue in Figure 2) to the network, and β is the number of compatible children c_i has. λ is 0 if all of the side chain polar atoms in c_i's root node (shown in black in Figure 2) are satisfied and there are nodes with unsatisfied side chain polar atoms in the current network; otherwise, λ is 1.

Additionally, MC HBNet will not register a network as a result if it contains any heavy-atom donors or acceptors that are buried and unsatisfied. This check reduces MC HBNet's runtime by approximately 40% by creating fewer false positives that need to be filtered out by a more computationally expensive satisfaction check that occurs at the end of the protocol, before networks are output. After all of the trajectories have completed, networks are ranked and prepared for output.

Ranking Results and Output. Networks are filtered and sorted, eliminating those that are redundant in primary amino acid sequence past a user defined threshold. Networks that contain at least one heavy (non-hydrogen) buried polar atom

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

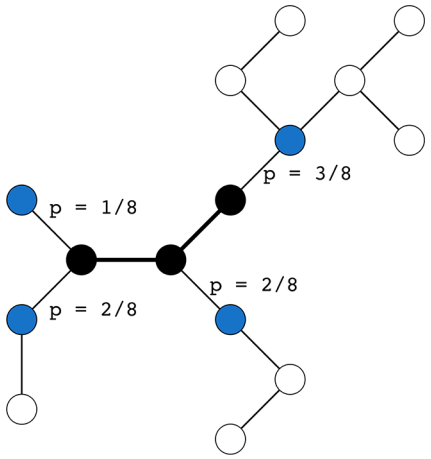


Figure 2. Monte Carlo growth of a network. Residues that are already part of the network are shown in black. Candidate residues are shown in blue, and downstream nodes are shown in white. All nodes and edges exist in the HBondGraph. Blue and white nodes may occur at the same residue positions as other blue and white nodes, but none may occur at the same residue position as a black node. Edges to candidate nodes are labeled with their probability (p) of being added to the network in the next round of Monte Carlo growth.

that is not either donating or accepting in a hydrogen bond are also eliminated. Buried polar hydrogen atoms that do not participate in hydrogen bonds are allowed but incur a penalty during sorting. Hydroxyl groups are only required to either donate or accept, but not both, to be considered satisfied (consistent with what is observed in experimentally determined structures[30]); however, there is an option for requiring that hydroxyl groups donate in order to be considered satisfied. Hydrogen bonds to the backbone are considered at this stage and are taken into account when evaluating satisfaction; native proteins often make use of hydrogen bonds from side chains to backbone atoms to preorganize structure, and networks that can extend to the backbone are captured by the protocols we present here. Users can also eliminate networks based on a custom criterion (e.g., minimum number of residues in the network, or number of intermolecular hydrogen bonds). The remaining hydrogen bond networks are sorted and ranked first by the number of buried unsatisfied polar hydrogen atoms (Num_Unsat_Hpol), then saturation, then HBNet Score.

Saturation. An early version of this metric was referred to as "connectivity," and it was shown that highly connected (saturated) networks were in close agreement with experimentally determined structures, whereas less connected networks were more easily displaced by water molecules.[26] We define saturation as the fraction of total hydrogen bonding capacity (given the polar atoms that comprise the network) that is met by the actual hydrogen bonds of the network. Higher values are better, with 1.0 implying that a network has reached its full hydrogen bond capacity. For every side chain in the network, each polar hydrogen atom on that side chain contributes 1 point and each lone pair contributes 1 point. Only one of the two lone pairs on a hydroxyl oxygen atom contributes a point because it is common for a hydroxyl to be an acceptor to only one hydrogen bond.[30] Saturation is calculated by dividing the sum of the points of all polar side chain atoms in the network by the sum of the points of all polar side chain atoms in the network residues (including atoms that do not participate in network hydrogen bonds). Saturation values can potentially be larger than 1.0 in the case of a hydroxyl participating in three hydrogen bonds or the case of bifurcated hydrogen bonds.

HBNet Score. HBNet Score is used to further discriminate between networks that are identical in Num_Unsat_Hpol and saturation by evaluating their energies using the full Rosetta energy function[24] within the same context: the network residues are placed onto a common "background" structure, which is the input structure with all packable residues mutated to alanine (except for any existing Gly, Pro, or disulfides, which are kept). HBNet Score is the difference in energy between the background structure with and without the network residues placed, normalized by the number of residues in the network.

Output. Once filtered and sorted, the networks are iteratively placed onto the input structure and output in order of ranking. Constraints are turned on to ensure that the hydrogen bonds of the network are maintained during downstream design; for design, the assumption is that there will be downstream design steps to optimize the space around the hydrogen bond network residues. Users can also opt to combine compatible networks together on the same output structure. Once the output structures are returned, any other part of Rosetta can be called to perform further design and analysis, or the structures can be output to disk.

Burial Calculations. Determining which polar atoms in the networks are buried versus solvent-exposed is challenging because the space around the hydrogen bond networks is often not yet designed, leaving large voids. The original implementation of HBNet used solvent-accessible surface area (SASA)[31,32] calculations with an increased probe radius.[26] In its current form, burial is precomputed by classifying each residue position as buried or not based on the number of neighboring residue positions that fall within a cone around the vector between its Cα and Cβ atoms;[33] this approach is advantageous because the precomputation is faster than the SASA calculations, and it is consistent for each input backbone, yielding the same classification independent of amino acid sequence and side chain conformation.

Benchmarks and Analysis. *Native Network Recovery.* A library of native protein crystal structures was generated by providing the Pisces web server[34,35] with the following conditions: sequence percentage identity $\leq$ 60; resolution $\leq$ 2.0 Å; R-factor $\leq$ 0.3; sequence length 40−10 000; non-X-ray entries excluded; CA-only entries excluded; cull PDB by entry; cull chains within entries set to "No." This library was pruned to only include structures that contain at least one protein−protein interface. HBNet was used to generate a list of unique native hydrogen bond networks within this pruned library by considering only native rotamers in each structure. This list went through a filter that removes networks that were comprised of at least one side chain that had a heavy atom with a B factor greater than 40 Å^2. A total of 2776 networks met these criteria and made the final list.

For every network in this list, we identified every hydrogen bond that had an energy $\leq$ −0.5 REU using Rosetta's ref2015 score function. For every extra-chi sampling level (Table 1), we checked to see if it produced a combination of side chains that rebuilt the native network in such a way in which every native hydrogen bond with an energy $\leq$ −0.5 REU was simultaneously present with an energy $\leq$ −0.4 REU. If all of the hydrogen bonds in the network could be simultaneously sampled, the network was deemed to be recovered. This was

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

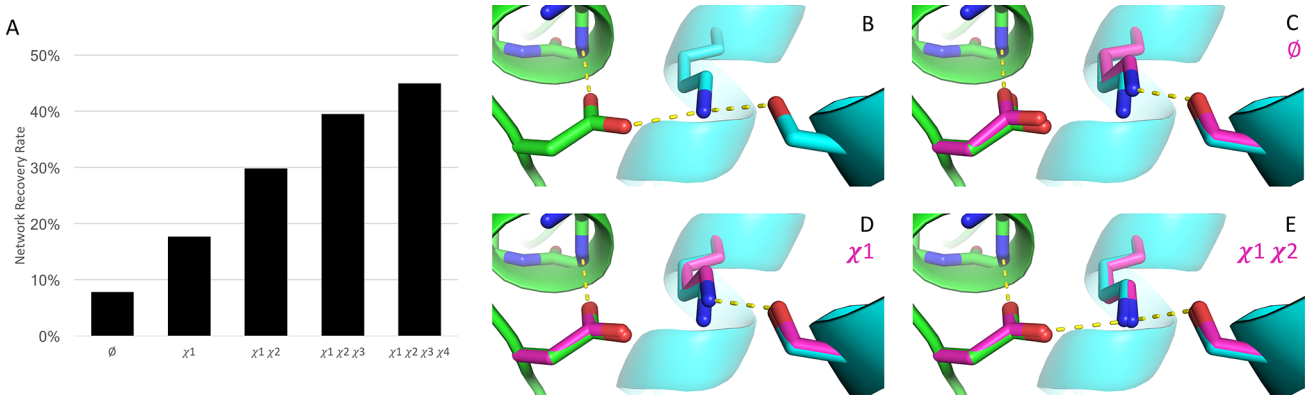


Figure 3. Native networks: impact of extra chi sampling. (A) Fraction of native networks that were sampled with different extra chi sampling levels (levels defined in Table 1). Chi sampling level increases from left to right. (B) An example native hydrogen bond network that is recovered at the $\chi1\chi2$ extra chi sampling level but not at $\chi1$ or Ø. (C–E) The lowest-RMSD rotamers (magenta) for the native network using the (C) Ø sampling level, (D) $\chi1$ sampling level, and (E) $\chi1\chi2$ sampling level.

repeated for all 2776 networks and with the extra-chi sampling levels displayed in Table 1.

Network Design Benchmarks. Four "motivating" design scenarios were chosen to compare the performances of HBNet and MC HBNet. These scenarios were chosen because they are similar to previous experiments we have run where HBNet did not perform adequately, hence motivating the development of the Monte Carlo protocol: (1) **Small interface, one-sided design** (PDB code 1YRK): All residue positions of the first chain were designed and all positions on the other chain were set to repack only (amino acid sequence fixed but rotamer conformations sampled), for a total of 40 packable positions. (2) **Medium interface one-sided design** (PDB code 1DPJ): All residue positions of the first chain were designed and all positions on the other chain were set to repack only, for a total of 121 packable positions. (3) **Large interface one-sided design** (PDB code 1GK9): All residue positions of the first chain were designed and all positions on the other chain were set to repack only, for a total of 342 packable positions. (4) **Small helical bundle monomer** (PDB code 3U3B chain A): The 23 buried residue positions were designed and the remaining residue positions were set to repack only. In all cases, all polar amino acid types were considered at designable residue positions, and the input files and scripts needed to run these benchmarks are provided in the Supporting Information Methods. It should be noted that in actual design scenarios, it can be advantageous to be more restrictive regarding which residue positions are designable and which polar amino acid types are allowed at certain positions.

Additionally, we selected four previously published HBNet designs in order to explore how MC HBNet behaves on design scenarios where HBNet has had proven success: PDB code 5J0K (symmetric homodimer), PDB code 5J10 (symmetric homodimer), PDB code 5J0H (symmetric homotrimer), and PDB code 5IZS (symmetric homotrimer). Both HBNet and MC HBNet were run on these design scenarios with sampling levels Ø, $\chi1$, and $\chi1\chi2$ (scripts included in the Supporting Information). MC HBNet was run with trajectory counts of 10^3, 10^4, 10^5, and 10^6 in order for us to explore the amount of sampling that is required to find high-quality networks for the various cases.

We measured the CPU time, peak memory usage, and the average HBNet statistics for the top 10 networks reported for each run. Benchmarks were measured on the Longleaf cluster at the University of North Carolina at Chapel Hill, using Intel Xeon E5-2680 v4 @ 2.40 GHz CPUs. Due to the ability for most runs to finish within a few hours and HBNet's tendency to take weeks to run if given too large of a design scenario, we declared a 24-h runtime limit. This limit was not applied to the symmetric reruns because they have previously been shown to run in a reasonable amount of time under these conditions. Additionally, this benchmark was run using sampling level Ø on 591 one-sided interface design cases including the three mentioned previously. The only metrics we tracked were the runtimes for HBNet and MC HBNet (using 10^4 trajectories).

RESULTS AND DISCUSSION

Benefits of Extra Chi Sampling. One of the limitations of the previously developed HBNet protocol is that it becomes dramatically slower (see below for case examples) when a larger rotamer library (i.e., more chi torsion angle sampling) is used during the design process. Because of the geometric sensitivity of hydrogen bonding, small changes to chi angles can result in substantial differences in the number of hydrogen bonds that can be made between rotamers, especially for longer side chains, for which lever-arm effects can lead to large changes in polar atom position. Thus, increasing chi sampling is generally expected to lead to more hydrogen bonds from which to sample. In order to quantitatively assess the need to be able to handle larger amounts of extra chi sampling, we measured the effect of extra chi sampling on hydrogen bond network sampling. We collected structures for 2776 native hydrogen bond networks at protein–protein interfaces from the PDB. For each network, we rebuilt the amino acid side chains using Rosetta's pool of rotamers and measured the fraction that could be sampled for each extra chi sampling level defined in Table 1. If every hydrogen bond in the native network could be simultaneously sampled, then the network was deemed "recoverable" for that chi sampling level. We evaluated every combination of rotamers for the residue positions that create the network, so this search was not compromised by stochasticity. As expected, Figure 3 shows the network recovery rate increase as the extra chi sampling level increases. Sampling level $\chi1\chi2$ can sample more than three times the fraction of native networks than can be sampled with no extra chi sampling (Ø).

Similarly, small changes to the backbone can also propagate to substantial changes in side chain hydrogen bonding

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751–2760

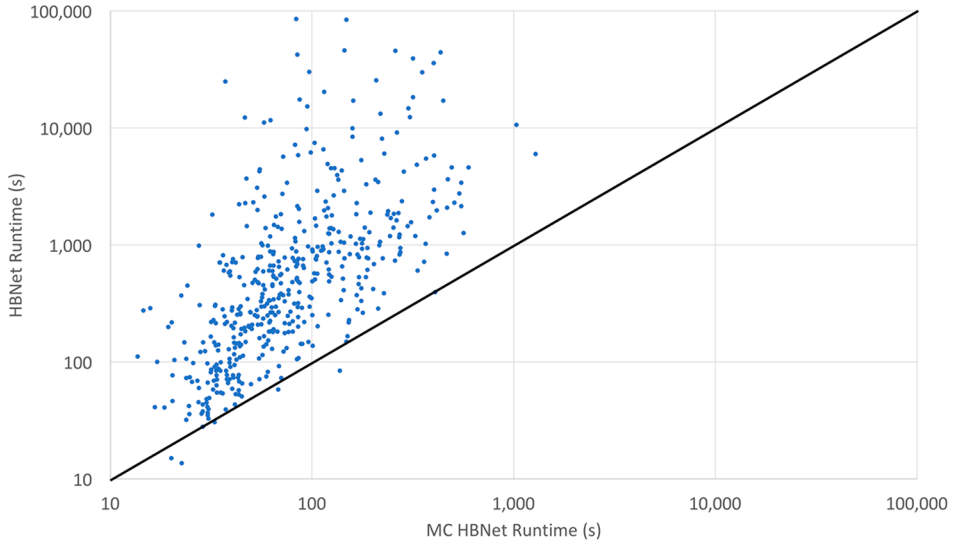


Figure 4. Aggregate data from one-sided interface design benchmarks. The diagonal line represents an equal runtime between the two protocols.

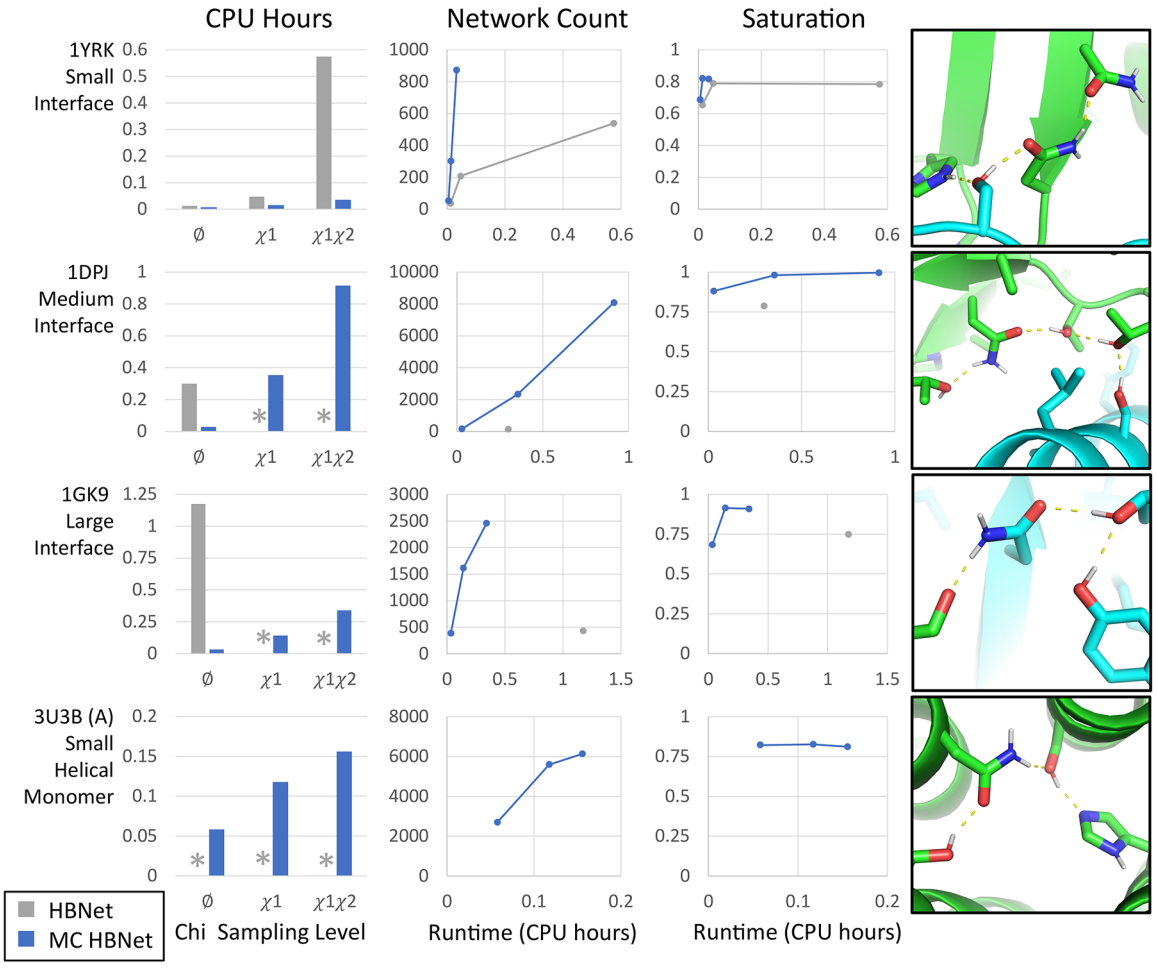


Figure 5. New design problems. Runtime, number of networks found, and saturation for the three interface design benchmarks of various sizes and the small helical monomer. An asterisk in the first column specifies that the protocol did not finish within 24 h. Results of traditional HBNet are shown in gray, and results of the new Monte Carlo protocol are shown in blue. Each case includes a picture of a hand-chosen representative network designed by MC HBNet with $\chi 1\chi 2$.

geometry and the possible networks that can be generated, and increased chi angle sampling can potentially compensate for these changes. Trajectories from Rosetta's Backrub protocol,[36] which incorporates small degrees of flexibility to generate conformational ensembles, illustrates this concept (Figure S1). Running MC HBNet on this ensemble of backbones shows that

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

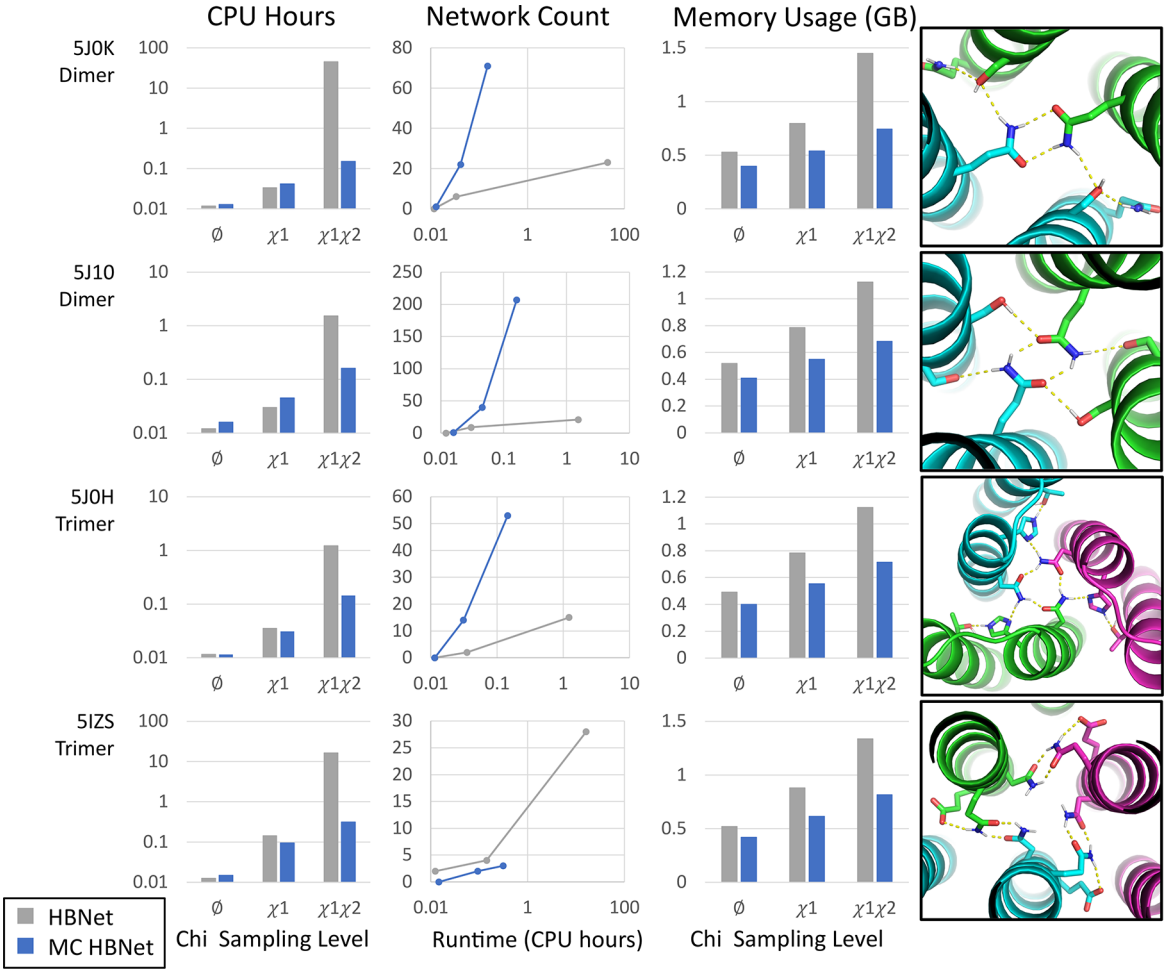


Figure 6. Symmetric interfaces. Runtime, number of networks found, and memory usage for four HBNet designs previously reported.[26] Each case is labeled with its PDB code and includes a picture of a hand-chosen representative network designed by MC HBNet with $\chi1\chi2$. Results of the original HBNet implementation are shown in gray, and results of the new Monte Carlo protocol are shown in blue.

backbone perturbations of as little as ~0.1 RMSD can affect the networks that can be captured (Figure S1, middle), and extra chi angle sampling can recover some of the networks that are missed due to these backbone perturbations (Figure S1, bottom). The ability to identify potential networks, even when the backbone is not in the most favorable conformation, will aid the search for low energy design models when performing design protocols that incorporate backbone sampling along with sequence design.

New Design Cases Enabled by MC HBNet. Many important design problems, for instance designing proteins to bind therapeutic targets, involve large asymmetric interfaces. Our initial attempts to design such cases using the original implementation of HBNet resulted in runtimes that were prohibitively slow. To demonstrate that MC HBNet can address these design cases, we assembled a collection of protein−protein interfaces of various sizes. We first searched for networks using a chi sampling level Ø on 591 protein−protein interfaces and only measured the CPU time consumed by each process (Figure 4). The space above the diagonal line represents results where HBNet takes longer to run than MC HBNet. Not only are most points above the line, but the distance from the line increases as the problem size grows, demonstrating that MC HBNet is faster than HBNet and better equipped to handle large design cases.

We next designed networks at three asymmetric protein−protein interfaces of varying sizes as well as the core of a helical bundle monomer (Figure 5, Table S2). MC HBNet showed a dramatic speed improvement and a better ability to scale to larger levels of extra chi sampling in all four cases. Many HBNet runs were not able to finish within 24 h (denoted by asterisk in Figure 5), while every MC HBNet run took less than 1 CPU hour. Tables S1−S8 also show that MC HBNet is slightly more memory efficient than HBNet for a given extra chi sampling level. All of the top networks reported had 0 unsatisfied polar atoms (which is the first metric used to sort results), meaning that saturation was the primary determinant in the ranking of the networks and assessing network quality. Figure 5 plots the average saturation for the 10 best results reported by Rosetta in the rightmost column. MC HBNet displays the ability to find more networks as a function of time than HBNet (Figure 5, middle column) and higher quality networks as a function of time. These improvements were consistent for both the asymmetric interfaces, as well as the monomeric design case.

HBNet failed to finish exploring the sample space of the small helical monomer design case within our 24-h time limit, even at the smallest extra chi sampling level, but MC HBNet was able to find thousands of hydrogen bond networks within minutes. Designing hydrogen bond networks into large monomeric structures is challenging, particularly if it is not

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

clear which region of the structure to focus on. The sample space of all possible hydrogen bond networks grows dramatically when the requirement of crossing an interface is removed. Our experience in using HBNet for noninterface designs has often resulted in unreasonably long runtimes. This issue can be partially alleviated by manipulating user-defined options, but it is not always obvious to the user how to implement this effectively. This weakness is not present in MC HBNet's algorithm because the runtime of a Monte Carlo trajectory is not dependent on the size of the sample space.

Surprisingly, MC HBNet can find higher quality networks (defined by saturation) than HBNet using the same extra chi sampling level. This result is not expected to be true when comparing any stochastic protocol with its exhaustive counterpart. The difference is that MC HBNet outputs networks that HBNet cannot; HBNet only stores networks that have grown to completion (ignoring the special case for hydroxyls, see Supporting Information). Network quality can be decreased when residues that contain unsatisfiable polar atoms are added to an already satisfied network. The moniker "satisfied subnetworks" is given to these networks that meet all design requirements and still have the ability to grow. MC HBNet can register satisfied subnetworks as results and continue to grow from them, while HBNet does not by default. In short, HBNet is a complete search of an incomplete sample space while MC HBNet is an incomplete search of a more complete sample space, and the latter appears to be a more effective strategy.

Symmetric Homo-Oligomer Benchmarks. We have previously reported success using HBNet to design symmetric homo-oligomers.[26] A handful of these designs were repeated using both HBNet and MC HBNet (Figure 6). Unlike with the Monte-Carlo-motivating benchmarks, we defined stricter filters to the designed networks in order to match the original protocol used to create these designs (details and scripts provided in the Supporting Information). We compared the two protocols by the number of networks that meet these strict design criteria. MC HBNet recapitulates the previously validated networks[26] and is able to find more networks as a function of time, and often as a function of chi sampling level, than HBNet.

MC HBNet still outperforms HBNet for the symmetric homo-oligomers when comparing CPU time for a given chi sampling level; however, the difference is milder than with the asymmetric interfaces. This result is likely due to the small problem size of these design cases. Not only are the proteins relatively small, but the presence of symmetry reduces the design space even further. MC HBNet consistently uses less memory than HBNet, in part due to the ability to use the HBondGraph instead of the traditional Rosetta data structures. The full table of results can be found in Tables S5−S8; MC HBNet benefits noticeably by increasing the number of Monte Carlo trajectories from the default of 10^4 to 10^5 for these symmetric cases.

CONCLUSIONS

MC HBNet is able to sample hydrogen bond networks faster and more effectively than HBNet. Additionally, MC HBNet can better handle large amounts of candidate rotamers per residue position, which increases the number of hydrogen bond networks that can be identified for a given protein backbone or complex. We have implemented MC HBNet within the Rosetta modeling package, but the sampling strategy, data structures, and network selection criteria described here are general and could be straight forwardly implemented within other computational frameworks.

One of our primary motivations for developing MC HBNet was to create a robust protocol that could be used as part of a larger pipeline aimed at *de novo* interface design. When designing new protein−protein interactions, it is generally not clear *a priori* what will be the most favorable way to dock the proteins against each other. For this reason, interface design protocols generally iterate between sampling alternative docked positions and searching for interface sequences that will stabilize the complex. It is important that the sequence search be rapid and reliably produce low energy solutions so that many alternative docked positions can be sampled. MC HBNet is well suited for this task because it can generally finish in the less than a minute for most interface sizes, and it produces multiple solutions that can be independently carried forward for design calculations to optimize the side chains of the neighboring residues; the computational savings afforded by MC HBNet can be reallocated to employ more computationally expensive protocols (e.g., flexible backbone methods) during downstream design to optimize the remaining interface positions that surround the network. In addition to interface design, this type of protocol should prove useful for designing ligand binding sites and catalytic sites that require hydrogen bond networks to stabilize the ligand or transition state.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00033.

Supporting Information Methods: additional details of the original implementation of HBNet relevant to this manuscript; methods for MC HBNet runs on Backrub trajectories to demonstrate the sensitivity to small backbone perturbations. Supporting Information Figures: Figure S1. Small backbone changes affect possible hydrogen bonds and network connectivities. Supporting Information Tables: Table S1, data for "small interface" case, used to generate Figure 5; Table S2, data for "medium interface" case, used to generate Figure 5; Table S3, data for "large interface" case used to generate Figure 5; Table S4, data for "small helical monomer" case used to generate Figure 5; Table S5, data for "symmetric homodimer 5J0K" case, used to generate Figure 6; Table S6, data for "symmetric homodimer 5J10" case, used to generate Figure 6; Table S7, data for "symmetric homodimer 5J0H" case, used to generate Figure 6; Table S8, data for "symmetric homodimer 5IZS" case, used to generate Figure 6 (MC HBNet data from Figure 6 are shown in bold). Scripts used to perform benchmarking and analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: bkuhlman@email.unc.edu.

ORCID

Brian Kuhlman: 0000-0003-4907-9699

Author Contributions

[∇]These authors contributed equally

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

Author Contributions

The manuscript was written and edited by J.B.M., S.E.B., and B.K. Software was written and benchmarked by J.B.M. and S.E.B. Computational strategies for MC HBNet and HBNet were conceived by J.B.M., S.E.B., D.B., and B.K.

Funding

This work was supported by grants from the NIH: GM073960 (B.K.), GM067553 (J.B.M.); a Burroughs Wellcome Fund Career Award at the Scientific Interface (S.E.B.); and the Howard Hughes Medical Institute (D.B).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank many members of the Rosetta Commons community for stimulating discussions, especially Dr. Will Sheffler, Dr. Vikram Mulligan, Dr. Andrew Leaver-Fay, and Zibo Chen.

■ REFERENCES

(1) Judd, E. T.; Stein, N.; Pacheco, A. A.; Elliott, S. J. Hydrogen Bonding Networks Tune Proton-Coupled Redox Steps During the Enzymatic Six-Electron Conversion of Nitrite to Ammonia. *Biochemistry* **2014**, *53* (35), 5638−5646.

(2) Polander, B. C.; Barry, B. A. A Hydrogen-Bonding Network Plays a Catalytic Role in Photosynthetic Oxygen Evolution. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (16), 6112−6117.

(3) Sánchez-Azqueta, A.; Herguedas, B.; Hurtado-Guerrero, R.; Hervás, M.; Navarro, J. A.; Martínez-Júlvez, M.; Medina, M. A Hydrogen Bond Network in the Active Site of Anabaena Ferredoxin-NADP(+) Reductase Modulates Its Catalytic Efficiency. *Biochim. Biophys. Acta, Bioenerg.* **2014**, *1837* (2), 251−263.

(4) Sigala, P. A.; Fafarman, A. T.; Schwans, J. P.; Fried, S. D.; Fenn, T. D.; Caaveiro, J. M. M.; Pybus, B.; Ringe, D.; Petsko, G. A.; Boxer, S. G.; Herschlag, D. Quantitative Dissection of Hydrogen Bond-Mediated Proton Transfer in the Ketosteroid Isomerase Active Site. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (28), E2552−E2561.

(5) Joachimiak, L. A.; Kortemme, T.; Stoddard, B. L.; Baker, D. Computational Design of a New Hydrogen Bond Network and at Least a 300-Fold Specificity Switch at a Protein-Protein Interface. *J. Mol. Biol.* **2006**, *361* (1), 195−208.

(6) Kuroda, D.; Gray, J. J. Shape Complementarity and Hydrogen Bond Preferences in Protein-Protein Interfaces: Implications for Antibody Modeling and Protein-Protein Docking. *Bioinformatics* **2016**, *32* (16), 2451−2456.

(7) Guo, H.; Salahub, D. R. Cooperative Hydrogen Bonding and Enzyme Catalysis. *Angew. Chem., Int. Ed.* **1998**, *37* (21), 2985−2990.

(8) Redzic, J. S.; Bowler, B. E. Role of Hydrogen Bond Networks and Dynamics in Positive and Negative Cooperative Stabilization of a Protein. *Biochemistry* **2005**, *44* (8), 2900−2908.

(9) Livesay, D. R.; Huynh, D. H.; Dallakyan, S.; Jacobs, D. J. Hydrogen Bond Networks Determine Emergent Mechanical and Thermodynamic Properties Across a Protein Family. *Chem. Cent. J.* **2008**, *2* (1), 17.

(10) Tatko, C. D.; Nanda, V.; Lear, J. D.; DeGrado, W. F. Polar Networks Control Oligomeric Assembly in Membranes. *J. Am. Chem. Soc.* **2006**, *128* (13), 4170−4171.

(11) Lombardi, A.; Summa, C. M.; Geremia, S.; Randaccio, L.; Pavone, V.; DeGrado, W. F. Retrostructural Analysis of Metalloproteins: Application to the Design of a Minimal Model for Diiron Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (12), 6298−6305.

(12) Faiella, M.; Andreozzi, C.; de Rosales, R. T. M.; Pavone, V.; Maglio, O.; Nastri, F.; DeGrado, W. F.; Lombardi, A. An Artificial Di-Iron Oxo-Protein with Phenol Oxidase Activity. *Nat. Chem. Biol.* **2009**, *5* (12), 882−884.

(13) Reig, A. J.; Pires, M. M.; Snyder, R. A.; Wu, Y.; Jo, H.; Kulp, D. W.; Calhoun, J. R.; Szyperski, T.; Butch, S. E.; Solomon, E. I.; DeGrado, W. F. Alteration of the Oxygen-Dependent Reactivity of De Novo Due Ferri Proteins. *Nat. Chem.* **2012**, *4* (11), 900−906.

(14) Chino, M.; Maglio, O.; Nastri, F.; Pavone, V.; DeGrado, W. F.; Lombardi, A. Artificial Diiron Enzymes with a De Novo Designed Four-Helix Bundle Structure. *Eur. J. Inorg. Chem.* **2015**, *2015* (21), 3371−3390.

(15) Zhang, S.-Q.; Chino, M.; Liu, L.; Tang, Y.; Hu, X.; DeGrado, W. F.; Lombardi, A. De Novo Design of Tetranuclear Transition Metal Clusters Stabilized by Hydrogen-Bonded Networks in Helical Bundles. *J. Am. Chem. Soc.* **2018**, *140*, 1294.

(16) Guffy, S. L.; Der, B. S.; Kuhlman, B. Probing the Minimal Determinants of Zinc Binding with Computational Protein Design. *Protein Eng., Des. Sel.* **2016**, *29* (8), 327.

(17) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M.; Stein, A.; Houlihan, K.; Dimaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11* (2), 609−622.

(18) Baker, E. N.; Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44* (2), 97−179.

(19) Stranges, P. B.; Kuhlman, B. A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds. *Protein Sci.* **2013**, *22*, 74.

(20) Der, B. S.; Kuhlman, B. Strategies to Control the Binding Mode of De Novo Designed Protein Interactions. *Curr. Opin. Struct. Biol.* **2013**, *23* (4), 639−646.

(21) Fleming, P. J.; Rose, G. D. Do All Backbone Polar Groups in Proteins Form Hydrogen Bonds? *Protein Sci.* **2005**, *14* (7), 1911−1917.

(22) Li, Z.; Yang, Y.; Zhan, J.; Dai, L.; Zhou, Y. Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. *Annu. Rev. Biophys.* **2013**, *42*, 315−335.

(23) Boas, F. E.; Harbury, P. B. Potential Energy Functions for Protein Design. *Curr. Opin. Struct. Biol.* **2007**, *17* (2), 199−204.

(24) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031−3048.

(25) Leaver-Fay, A.; Kuhlman, B.; Snoeyink, J. An Adaptive Dynamic Programming Algorithm for the Side Chain Placement Problem. *Pac Symp. Biocomput* **2005**, 16−27.

(26) Boyken, S. E.; Chen, Z.; Groves, B.; Langan, R. A.; Oberdorfer, G.; Ford, A.; Gilmore, J. M.; Xu, C.; Dimaio, F.; Pereira, J. H.; Sankaran, B.; Seelig, G.; Zwart, P. H.; Baker, D. De Novo Design of Protein Homo-Oligomers with Modular Hydrogen-Bond Network-Mediated Specificity. *Science* **2016**, *352* (6286), 680−687.

(27) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: an Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **2011**, *487*, 545−574.

(28) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived From Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844−858.

(29) Kortemme, T.; Morozov, A. V.; Baker, D. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *J. Mol. Biol.* **2003**, *326* (4), 1239−1259.

(30) Worth, C. L.; Blundell, T. L. Satisfaction of Hydrogen-Bonding Potential Influences the Conservation of Polar Sidechains. *Proteins: Struct., Funct., Genet.* **2009**, *75* (2), 413−429.

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760

(31) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.
(32) Durham, E.; Dorr, B.; Woetzel, N.; Staritzbichler, R.; Meiler, J. Solvent Accessible Surface Area Approximations for Rapid and Accurate Protein Structure Prediction. *J. Mol. Model.* **2009**, *15* (9), 1093−1108.
(33) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, *357* (6347), 168−175.
(34) Wang, G.; Dunbrack, R. L. *PISCES: a Protein Sequence Culling Server*; 2003; Vol. *19*, pp 1589−1591.
(35) Wang, G.; Dunbrack, R. L. PISCES: Recent Improvements to a PDB Sequence Culling Server. *Nucleic Acids Res.* **2005**, *33*, W94−W98.
(36) Smith, C. A.; Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* **2008**, *380* (4), 742−756.

DOI: 10.1021/acs.jctc.8b00033
J. Chem. Theory Comput. 2018, 14, 2751−2760