doi:10.20944/preprints201904.0263.v1

Macromolecular modeling and design in Rosetta: new methods and frameworks

Koehler Leman, Julia [1, 2]*, Weitzner, Brian D [3, 4, 5], Lewis, Steven M [6, 7, 8], RosettaCommons consortium, Bonneau, Richard [1, 2, 58, 59]

RosettaCommons consortium:
Adolf-Bryfogle, Jared [9], Alam, Nawsad [10], Alford, Rebecca F [3], Aprahamian, Melanie [11], Baker, David [4, 5], Barlow, Kyle A [12], Basanta, Benjamin [4, 13], Bender, Brian J [14], Blacklock, Kristin [15], Bonet, Jaume [16, 17], Boyken, Scott [5], Bradley, Phil [18], Bystroff, Chris [19], Conway, Patrick [4], Cooper, Seth [20], Correia, Bruno E [16, 17], Coventry, Brian [4], Das, Rhiju [21], De Jong, Rene M [22], DiMaio, Frank [4], Dsilva, Lorna J L [20], Dunbrack, Roland [23], Ford, Alex [4], Frenz, Brandon [8], Fu, Darwin Y [24], Geniesse, Caleb [21], Goldschmidt, Lukasz [4], Gowthaman, Ragul [25, 26], Gray, Jeffrey J [3], Guffy, Sharon [6], Horowitz, Scott [27, 28], Huang, Po-Ssu [4], Huber, Thomas [29], Jacobs, Tim M [30], Jeliazkov, Jeliazko R [31], Johnson, David K [32], Kappel, Kalli [33], Karanicolas, John [23], Khakzad, Hamed [17, 34, 35], Khar, Karen R [32], Khare, Sagar [36], Khatib, Firas [37], Khramushin, Alisa [10], King, Indigo C [4, 8], Kleffner, Robert [20], Koepnick, Brian [4], Kortemme, Tanja [38], Kuenze, Georg [24, 39], Kuhlman, Brian [6], Kuroda, Daisuke [40, 41], Labonte, Jason W [3, 42], Lapidoth, Gideon [43], Leaver-Fay, Andrew [6], Lindert, Steffen [11], Linsky, Thomas [4, 5], London, Nir [10], Lubin, Joseph H [3], Lyskov, Sergey [3], Maguire, Jack [30], Malmström, Lars [17, 34, 35, 44], Marcos, Enrique [4, 45], Marcu, Orly [10], Marze, Nicholas A [3], Meiler, Jens [39, 46, 47], Moretti, Rocco [24], Mulligan, Vikram Khipple [1, 4, 5], Nerli, Santrupti [48], Norn, Christoffer [43], Ó Conchúir, Shane [38], Ollikainen, Noah [38], Ovchinnikov, Sergey [4, 5, 49], Pacella, Michael S [3], Pan, Xingjie [38], Park, Hahnbeom [4], Pavlovicz, Ryan E [4, 5], Pethe, Manasi [50, 51], Pierce, Brian G [25, 26], Pilla, Kala Bharath [29], Raveh, Barak [10], Renfrew, P Douglas [1], Roy Burman, Shourya [3], Rubenstein, Aliza [15, 52], Sauer, Marion F [53], Scheck, Andreas [16, 17], Schief, William [9], Schueler-Furman, Ora [10], Sedan, Yuval [10], Sevy, Alexander M [53], Sgourakis, Nikolaos G [54], Shi, Lei [4, 5], Siegel, Justin [55, 56, 57], Silva, Daniel-Adriano [4], Smith, Shannon [24], Song, Yifan [4, 5], Stein, Amelie [38], Szegedy, Michael [36], Teets, Frank D [6], Thyme, Summer B [4], Wang, Ray Yu-Ruei [4], Watkins, Andrew [21], Zimmerman, Lior [10]

================

1 Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY 10010, USA
2 Dept of Biology, New York University, New York, 10003, NY, USA
3 Dept of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
4 Dept of Biochemistry, University of Washington, Seattle, WA 98195, USA
5 Institute for Protein Design, University of Washington, Seattle, WA 98195, USA
6 Dept of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
7 Dept of Biochemistry, Duke University, Durham, NC 27710, USA
8 Cyrus Biotechnology, Seattle, WA 98101, USA
9 Dept of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA
10 Dept of Microbiology and Molecular Genetics, IMRIC, Ein Kerem Faculty of Medicine, Hebrew University of Jerusalem, 91120, Jerusalem, Israel
11 Dept of Chemistry and Biochemistry, Ohio State University, Columbus, OH, 43210, USA
12 Graduate Program in Bioinformatics, University of California San Francisco, CA 94158, USA
13 Biological Physics Structure and Design PhD Program, University of Washington, Seattle, WA 98195, USA
14 Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, USA
15 Institute of Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
16 Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
17 Swiss Institute of Bioinformatics, Lausanne, Switzerland
18 Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
19 Dept of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA
20 Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
21 Dept of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA

22 DSM Biotechnology Center, 2613 AX Delft, The Netherlands
23 Fox Chase Cancer Center, Philadelphia PA 19111, USA
24 Dept of Chemistry, Vanderbilt University, Nashville, TN 37232, USA
25 University of Maryland Institute for Bioscience and Biotechnology Research, Rockville, MD 20850, USA
26 Dept of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA
27 Dept of Chemistry & Biochemistry, University of Denver, Denver, CO 80208, USA
28 The Knoebel Institute for Healthy Aging, University of Denver, Denver, CO 80208, USA
29 Research School of Chemistry, Australian National University, Canberra ACT 2601, Australia
30 Program in Bioinformatics and Computational Biology, Dept of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
31 Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA
32 Center for Computational Biology, University of Kansas, Lawrence, KS 66047, USA
33 Biophysics Program, Stanford University, Stanford CA 94305, USA
34 Institute for Computational Science, University of Zurich, CH-8057 Zurich, Switzerland
35 S3IT, University of Zurich, CH-8057 Zurich, Switzerland
36 Dept of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
37 Dept of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA
38 Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, CA 94158, USA
39 Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA
40 Medical Device Development and Regulation Research Center, School of Engineering, University of Tokyo, Tokyo 113-8656, Japan
41 Dept of Bioengineering, School of Engineering, University of Tokyo, Tokyo 113-8656, Japan
42 Dept of Chemistry, Franklin & Marshall College, Lancaster, PA 17604, USA
43 Dept of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, 76100, Israel
44 Division of Infection Medicine, Dept of Clinical Sciences Lund, Faculty of Medicine, Lund University, SE-22184, Lund, Sweden
45 Institute for Research in Biomedicine Barcelona, The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
46 Depts of Chemistry, Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA
47 Institute for Chemical Biology, Vanderbilt University, Nashville, TN 37232, USA
48 Dept of Computer Science, University of California Santa Cruz, CA, USA
49 Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA
50 Dept of Chemistry and Chemical Biology, Institute of Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
51 Center for Integrative Proteomics Research, Institute of Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
52 Computational Biology and Molecular Biophysics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
53 Chemical and Physical Biology Program, Vanderbilt Vaccine Center, Vanderbilt University, Nashville, TN 37235, USA
54 Dept of Chemistry and Biochemistry, University of California Santa Cruz, CA, USA
55 Dept of Chemistry, University of California, Davis, CA, USA
56 Dept of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA
57 Genome Center, University of California, Davis, CA, USA
58 Dept of Computer Science, New York University, New York, 10003, NY, USA
59 Center for Data Science, New York University, New York, 10003, NY, USA

* corresponding author: Julia.koehler.leman@nyu.edu

Keywords: structure prediction, Rosetta, computational modeling, protein design

Abstract

The Rosetta software suite for macromolecular modeling, docking, and design is widely used in pharmaceutical, industrial, academic, non-profit, and government laboratories. Despite its broad modeling capabilities, Rosetta remains consistently among leading software suites when compared to other methods created for highly specialized protein modeling and design tasks. Developed for over two decades by a global community of over 60 laboratories, Rosetta has undergone multiple refactorings, and now comprises over three million lines of code. Here we discuss methods developed in the last five years in Rosetta, involving the latest protocols for structure prediction; protein–protein and protein–small molecule docking; protein structure and interface design; loop modeling; the incorporation of various types of experimental data; modeling of peptides, antibodies and proteins in the immune system, nucleic acids, non-standard chemistries, carbohydrates, and membrane proteins. We briefly discuss improvements to the energy function, user interfaces, and usability of the software. Rosetta is available at www.rosettacommons.org.

Introduction

Development of Rosetta started in the mid-1990s for protein structure prediction and to gain insights into the protein folding problem, which to date remains a challenging problem in structural biology. Over time, the number of applications grew to address a wider array of modeling tasks, ranging from protein–protein or –small molecule docking, and incorporation of NMR data to loop modeling, protein design, and interaction with peptides and DNA or RNA (Figure 1). Over the 20 plus years since those early beginnings, the community of developers and scientists, the RosettaCommons, grew from a single academic laboratory to over 60 laboratories around the globe[1]. Rosetta has undergone several transitions in both programming language and implementation, and latest protocols are based on Rosetta3, which was first released in 2008[2]. Rosetta's energy function has been continuously improved over its 20 plus year timeframe, detailed descriptions of which can be found in references [3] and [4]. As the Rosetta community grew, efforts to improve usability, interfaces to the code, and documentation have drastically improved usability and modular application to new problems. As part of our sustained focus on reproducibility and usability, several interfaces were developed, (PyRosetta[5], RosettaScripts[6], Foldit[7]) and emphasis was placed on publishing protocol captures[8] that accompany manuscripts to improve user friendliness and scientific reproducibility. As the software's interfaces have grown more versatile, development has accelerated and branched in many directions. However, this makes it difficult to keep up with all the developments that are happening within the software and the scientific community. To address this growth in functionality, here we have compiled the latest method developments in the Rosetta software suite from the past five years, divided into several modeling categories as outlined below. This report is intended to serve as a guide for Rosetta users and developers — whether new, returning, or seasoned — who want to be updated on newest developments. The supplement contains a tour of the protocols with extensive links to documentation and resources on the web.

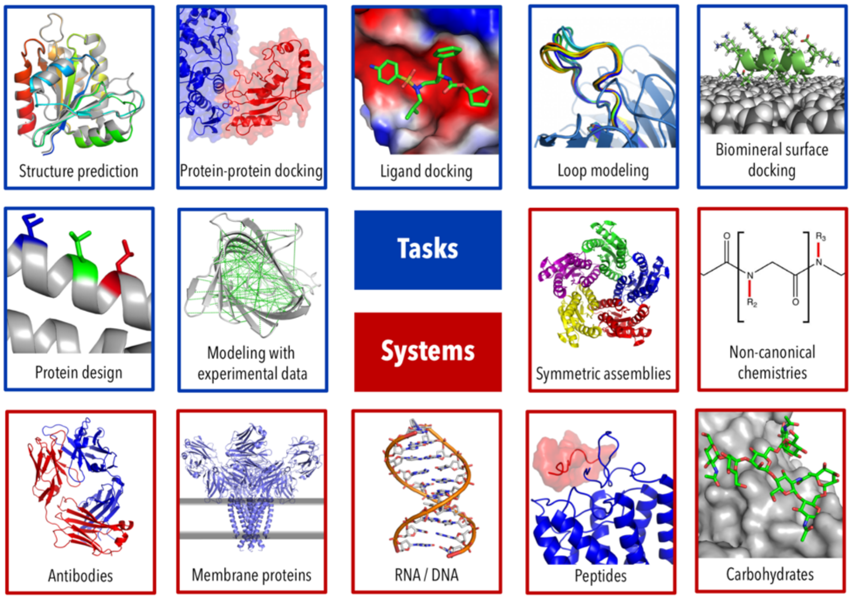
1. Major applications

The general outline of a Rosetta protocol is depicted in Figure 2: the conformation of a biomolecule (the *Pose*) is randomly altered via a *Mover* and the resulting conformation is evaluated by a *ScoreFunction*. The *move* is accepted based on the Metropolis criterion and the energy difference between the original and the new conformation. Many trajectories are generated, and the final models are evaluated based on the scientific objective.

Predicting protein structures

Rosetta was originally developed for protein structure prediction, which is accomplished by assembling fragments from known protein structures *via* a Monte-Carlo procedure and scoring the models with an advanced scorefunction. Since optimizing the fragments for structure prediction can improve model quality, the original fragment picker application was re-implemented as an object-oriented framework that is vastly more flexible and allows incorporation of various types of restraints from secondary structure prediction or experimental data, for instance from NMR chemical shifts[9]. Improvements in homology modeling were achieved by multi-template modeling in RosettaCM, which combines the most homologous portions from multiple templates into a single model while modeling missing residues *de novo*[10]. If a template is absent, protein structures can be predicted *de novo*, which remains one of the most challenging tasks in structural biology, even though the incorporation of evolutionary coupling constraints (for instance from GREMLIN[11]) has led to enormous improvements in model quality. To thoroughly search the conformational search space, an iterative hybridize approach was implemented. It uses a genetic algorithm that recombines models from an input pool to create models that have features from their parents but are also distinctly different. Creating several child models in each iteration, updating the input pool, and performing 30-50 iterations lead to improved model accuracy because features that are scored favorably by the scorefunction are repeatedly used in the recombination, such that the models in the pool converge over time. This approach has been used to improve model quality of *de novo* predicted models[12] as well as homology models[13]. Model refinement or generating ensembles of structures can be accomplished by several algorithms in Rosetta: *FastRelax*[14], Backrub[15], or using vicinity sampling in the KIC/Next-Generation-KIC loop modeling algorithms[16,17] (see loop modeling section).

Figure 1: Capabilities of the Rosetta macromolecular modeling suite
Popular tasks that can be addressed in Rosetta (blue) and major systems that can be modeled (red).

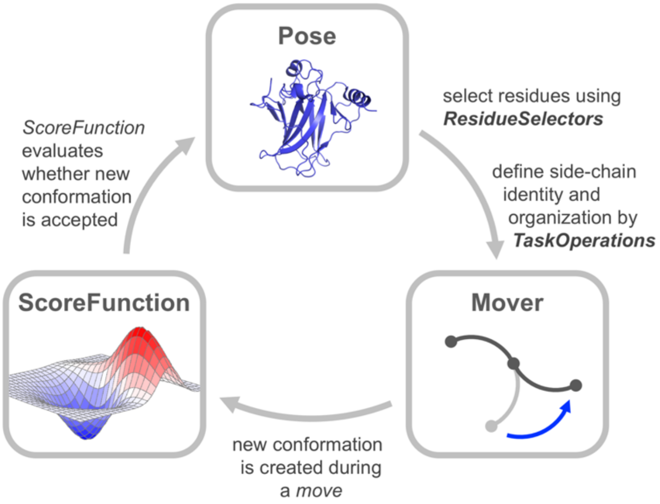


Experimental protein structure determination is challenging for proteins on solid surfaces such as biominerals, self-assembled monolayers, inorganic catalysts, and nanomaterials. RosettaSurface[18] samples protein conformations *ab initio* in both the solution and adsorbed states (Figure 3D) in order to account for adsorption-induced conformational changes. Experimental data can be incorporated into the simulation[19] to improve scoring, which remains difficult because the Rosetta scorefunction has been optimized for proteins in aqueous solution.

Modeling protein-protein complexes

Another early Rosetta method was RosettaDock, which predicts the structure of protein-protein complexes from input monomers. The most recent iteration, RosettaDock4.0[20] incorporates protein flexibility from pre-generated protein ensembles, mimicking conformational selection. The new protocol has improved sampling efficiency by automatically adjusting the sampling rate based on the diversity of the input ensembles. Scoring has also been improved by using a novel, six-dimensional coarse-grained scoring scheme called *motif_dock_score*, which employs score grids generated from known complexes in the PDB. In local docking benchmarks and backbone deviations of up to 2.2 Å, RosettaDock4.0 was able to successfully dock ~50% of complexes. For symmetric homomers, Rosetta SymDock2[21] can be used, which uses the same six-dimensional scoring scheme as in RosettaDock. Symmetry information can be extracted from a homologous complex, or a global docking search can be performed for a given point symmetry using Rosetta's symmetry framework[22]. An induced-fit based all-atom refinement relieves clashes in tightly-packed complexes to give physically realistic models. On a benchmark set of 43 complexes with different cyclic and dihedral symmetries, global docking on homology models had accuracies of 61% and 42% for cyclic and dihedral symmetries, respectively. These accuracies are substantially higher than for other symmetric docking tools and can be dramatically improved when adding restraints.

Figure 2: Main elements of a Rosetta protocol
Three main elements are required in a Rosetta protocol. The *Pose* is the biomolecule, such as a protein, RNA, DNA, small molecule, or glycan, in a specific conformation. Residues in the *Pose* can be selected via *ResidueSelectors* and the behavior for side-chain optimization or mutation can be defined by *TaskOperations*. Specific *Movers* then control how the conformation of the *Pose* is changed, and the new conformation is subsequently evaluated by a *ScoreFunction*. The Metropolis criterion decides whether the new conformation is accepted in the sampling trajectory. Many independent sampling trajectories are generated, and the final models are evaluated based on the purpose of the protocol.



Elucidating interfaces between proteins and small molecule ligands

Structure-based drug design has become a common approach for drug optimization due to increasing numbers of deposited structures in the Protein Data Bank. RosettaLigand[23] has demonstrated success in predicting small molecule-protein interactions. Recent improvements to the algorithm rely on a low-resolution sampling step utilizing the *TransformMover*, that combines translational and rotational perturbations in a single step, and using scoring grids for energy evaluation[24]. Further, the algorithm allows backbone flexibility, mimicking an induced fit algorithm[25]. On a benchmark of 43 complexes, this new algorithm demonstrated an enhanced docking success by 10-15% with an effective 30-fold speedup over the original RosettaLigand performancez, allowing virtual high-throughput screening (vHTS) of medium-sized ligand libraries. Later in the drug development process, when medicinal chemists optimize ligands based on structure-activity relationships (SAR), they synthesize ligands that typically share a core chemical scaffold and are assumed to bind to their target in a similar fashion[26]. RosettaLigandEnsemble[27] improves sampling during ligand docking by taking advantage of ligand similarities and docking a congeneric series of ligands at the same time, allowing for a placement that works for all considered ligands while optimizing the binding interface for each ligand independently. Experimental SARs can be included by promoting certain binding modes.

Another approach for therapeutic intervention is to use small molecule ligands as competitive inhibitors of protein-protein interactions. A common challenge, however, is that the protein conformation bound to the inhibitor is often distinctly different from the unbound or protein-protein bound conformation. The pocket optimization approach in Rosetta identifies protein surface pockets and uses their volume as an additional scoring term: this allows the user to start from an unbound protein structure and carry out biased sampling of a protein such that low-energy pocket-containing states are preferentially explored[28,29]. The specific conformations sampled through this approach match “druggable” alternate conformations observed in ligand-bound structures[28,29], implying that these states can serve as be excellent starting points for virtual screening. The pockets sampled on the protein surface can then be matched to complementary ligands directly, by using the pocket itself as the starting point for pharmacophore-based screening[30]. Alternatively, these pockets can also be used for Rosetta's *Docking Approach using Ray Casting* (DARC[31]) method. DARC uses ray-casting to rapidly position a ligand in the protein surface pocket[31]; by iterating over many candidates, DARC provides a means for very rapid virtual screening. DARC has also been adapted for GPUs[32], and the newer implementation[33] includes features that provide improved performance in virtual screening benchmarks.

Designing new proteins and functions

The inverse problem of protein structure prediction is protein design to identify a sequence that best represents a given structure. In particular, *de novo* design and design of novel protein functions towards therapeutic intervention remains one of the grand challenges in structural biology. This problem has been approached by various methods in Rosetta. The SEWING protocol creates *de novo* designs by recombining parts of protein structures from randomly-selected helical building blocks[34]. SEWING's requirement-driven approach allows users to specify features or properties that should be incorporated into their designs during backbone generation without necessarily requiring a certain size or three-dimensional fold. New features include incorporation of functional motifs such as protein-binding peptides for protein interface design and partial or complete ligand binding sites for ligand-binding protein design[35]. A somewhat similar algorithm has been implemented for antibody design (AbDesign, see below), which was generalized for enzyme design[36]. A more general approach is RosettaRemodel, which performs protein design by rebuilding parts or all of the structure[37] from fragments of known proteins structures. It relies on a blueprint file in which the user defines secondary and supersecondary structure of the fold to be built. Remodel interfaces with a number of Rosetta protocols and can be used for various applications such as *de novo* modeling, fixed-backbone sequence design, refinement, loop insertion, deletion, and remodeling, as well as disulfide engineering, domain assembly, and motif grafting.

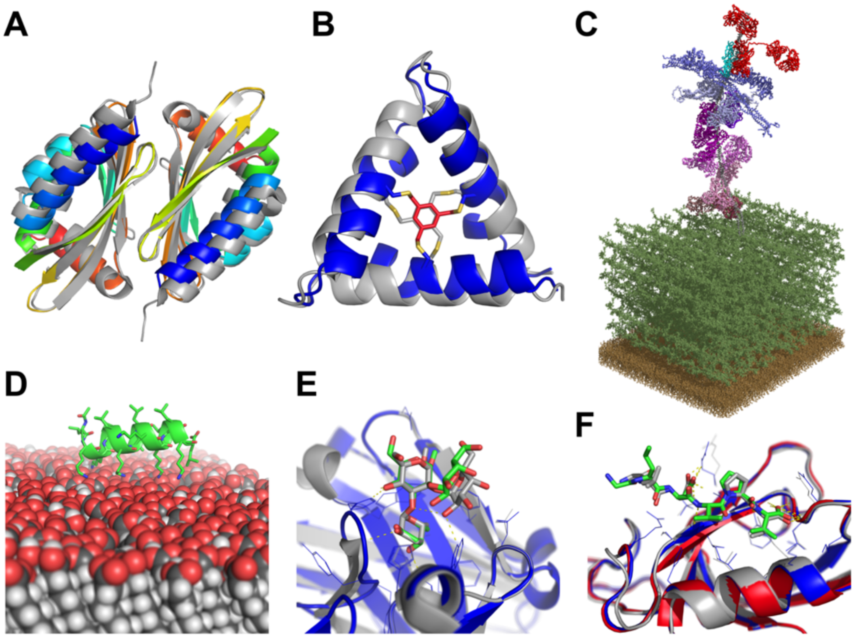
For the design of multifunctional proteins such as biosensors, bioswitches and tunable affinity clamps, the domain insertion protocol LooDo (Loop-directed domain insertion) has been developed. With LooDo, proteins are designed by inserting a domain into another by two flanking linker regions[38]. The linker regions are sampled *via* fragment insertion to determine relative positioning of the domains, followed by generalized kinematic loop closure[39] (GenKIC, see below) and enzyme design to optimize the interface.

In protein design, a common task is not only design towards a certain goal (positive design), but additionally, design away from undesired features (negative design). Such a Multi-State Design[40] (MSD) approach evaluates strengths and weaknesses of a single sequence on multiple backbones to account for positive and negative design, for instance binding to one but not another protein partner. REstrained CONvergence[41] (RECON) takes this idea one step further by allowing each state to sample multiple sequences during the design process, which is iteratively applied by increasing the restraint weight to encourage sequence convergence. RECON is effective for large multi-state design problems, such as antibody affinity maturation or prediction of evolutionary sequence profiles of flexible backbones[42,43].

De novo protein design is somewhat easier for structures involving highly regular helices and sheets, as the principles dictating their conformations are well known. However, designing curved and twisted β-sheets requires a deeper understanding of the structural irregularities that enable them. These principles were implemented in the curved β-sheet design method to design a variety of protein folds with curved sheets (Figure 3A), creating pockets suitable for tailoring ligand-binding and enzymatic active sites[44].

During computational *de novo* protein design[45], a stringent test for the consistency of the designed sequence is whether *ab initio* structure prediction will yield the same structure that was used as a starting point for the design. However, computationally testing a large number of designs is prohibited by the vast conformational search space for *ab initio* structure prediction. To drastically limit that search space and test many more designs, the biased forward folding method[44] uses three (instead of the typical 200) fragments per residue position. Fragments are chosen based on the RMSD to the native structure in design. The designs achieving near-native sampling are more likely to have funnel-shaped energy landscapes and therefore worth assessing with *ab initio* structure prediction.

Figure 3: Rosetta can successfully address diverse biological questions
(A) Overlay of the designed homo-dimeric curved β-sheet (dcs-E_4_dim_cav3) in rainbow and the crystal structure in gray (PDBID 5u35). The protein is completely designed *de novo* and features a curved β-sheet, a large pocket, and a homodimer interface[44]. (B) Overlay of the *de novo* designed macrocycle 3H1 in blue and the NMR structure in gray (PDBID 5v2g). This is an example of a "CovCore" (covalent core) protein that is held together covalently by hydrophobic cross-linkers at its core (in red for the design and gray for the NMR structure)[59]. (C) The interactome of M1 protein (the most important virulence factor of Group A *streptococcus*) and 15 human plasma proteins on the surface of bacteria (peptidoglycan layer (dark green), and the membrane (brown)). This 1.8MDa structure is measured directly in a complex mixture of intact bacteria and human plasma by PyTXMS. All structural models are provided by Rosetta where it consists of M1 protein (gray), IgG (red), four fibrinogens (dark to light blue), six albumins (dark to light pink), coagulation factor XIII A [F13A] (purple), C4bPa (cyan), haptoglobin [HP] (brown), and alpha-1-antitrypsin [SerpinA1] (plum). This complex is supported by more than 200 chemical cross-links[134]. (D) Model of an LK-α peptide (LKKLLKLLKKLLKL with a periodicity of 3.5 assuming a helical conformation) on a hydrophilic self-assembled monolayer surface. In solution, the peptide is unstructured, whereas when on the surface, experiments show that it assumes helical structure[19]. (E) Flexible docking of a carbohydrate antigen to an antibody. The crystal structure is in gray (PDBID 1mfa) and the Rosetta model in blue, with the carbohydrate in green. The coordinates for the antibody were taken from the PDB and the glycan coordinates started from a randomize backbone conformation and rigid-body orientation[111]. (F) High-resolution model of a peptide-protein complex generated using PIPER-FlexPepDock (model: blue; solved structure in gray, PDBID 1mfg). The predicted model was generated starting from a peptide sequence (LDVPV, derived from the C-terminal tail of ErbB2R) and the unbound structure of the receptor (Erbin PDZ domain, PDBID 2h3l, colored in red)[63].



Design of protein function has been accomplished by grafting a known motif from a template protein onto a new protein (*motif grafting*). This approach has been used for antibodies and for vaccine design[46] using the *fold_from_loops* application, where the functional motif is used as a starting point of an extended structure that is folded following the constraints of a target topology. Iterative refinement is carried out via sequence design and structural relaxation before filtering and human-guided optimization. This application has been extended into the *Functional Folding and Design* (FunFolDes) protocol, which includes multi-segment motif grafting, different residue length motif insertion, incorporation of restraints, and folding in the presence of a binding target[47]. Fragments selected according to the structure of the target topology improve the performance of the folding stage via the *StructFragmentMover*.

doi:10.20944/preprints201904.0263.v1

Designing interfaces between proteins and interaction partners

Related problems to protein design are the design of interfaces of proteins with their interaction partners such as proteins or small molecule ligands, and the prediction of ΔΔGs of mutation.

Prediction of ΔΔGs of mutations for protein stability or protein-protein interactions is a difficult problem with low correlation coefficients (0.5-0.7)[48], because the effect of the mutation is small compared to the total energy in the system, and because protein flexibility adds noise to the energies that can mask the effect of mutations. In the simplest case of alanine scanning (mutating into Ala), methods that use a “soft-repulsive” energy function without modeling backbone flexibility[49,50] have typically outperformed methods that allow protein flexibility and use hard-repulsive energy functions[51]. FlexDDG[52] was created to improve protein-protein interface ΔΔG predictions and generalize them to residues other than Ala. The protocol creates conformational ensembles using Rosetta backrub sampling[53], then repacks sidechains, minimizes torsions and computes change in protein-protein interaction ΔΔG by averaging across the ensembles. On 1240 interface mutants, FlexDDG outperforms Rosetta’s *ddg_monomer* application, which was originally created and validated to predict changes in stability upon mutation, not interfaces.

Designing ligand-binding interfaces in proteins is challenging due to inaccuracies in the energy function, the flexibility of ligands, and the sensitivity of protein-ligand interactions to even subtle conformational changes[54]. Flexible backbone design methods that use pre-generated ensembles as a starting point for design[55,56] perform poorly in benchmarks. The *CoupledMoves* protocol couples backbone flexibility with changes in sidechain rotamers or ligand orientation or conformers, and leads to substantial improvements in various benchmarks[57].

Symmetric protein assemblies can now be modeled using parametric design. Nature created super-helical coiled-coils that can be described by geometric equations using Crick parameters[58], which include variables for the radius of the bundle, major helical twist, minor helix rotation about the primary axis, etc. Several Movers such as *MakeBundle*, *PerturbBundle*, and *BundleGridSampler* allow designing helical bundles[59,60] and β-barrels based on pre-defined or sampled parameters. Since parametric methods do not rely on fragments libraries, these modules can be applied to non-canonical coiled-coil heteropolymers.

Modeling peptides and peptidomimetics

When the protein interaction partners are peptides, the inherent flexibility of the peptide and thus the large conformational search space lead to challenging modeling problems. FlexPepDock allows targeted sampling of the peptide flexibility during its docking into a given binding site, either by refining an approximate peptide conformation (FlexPepDock refinement[61]), or by full *ab initio* sampling of the peptide conformation (FlexPepDock *ab initio*[62]). Peptide docking is especially challenging when the binding site on the receptor is unknown. However, it can be simplified based on the observation that the bound peptide conformation is often included in the fragments generated by the *FragmentPicker*. The PIPER-FlexPepDock[63] protocol rigid-body docks these fragments using PIPER FFT-based docking[64], and refines the complex using FlexPepDock[61]. PIPER-FlexPepDock can generate highly accurate peptide-protein complexes from a peptide sequence and a free receptor structure (Figure 3F).

Many protein-protein interactions (PPI) are mediated by often disordered peptide segments that are responsible for most of the binding energy[65–68]. PeptiDerive[69] detects such segments in a PPI complex through a sliding window approach. PeptiDerive was extended to cyclized peptides and is available on the ROSIE[70] server.

Conformations of cyclic peptides can be sampled with *simple_cycpep_predict*, which restricts the conformational search space through cyclization[39,59,71] via the Generalized Kinematic Closure (GenKIC) algorithm (see below). *Simple_cycpep_predict* does not rely on protein fragments and can model non-canonical chemistries (Figure 3B), being a generalization of earlier protocols.

Multi-specificity is common at protein-peptide interfaces, meaning that the protein can interact with multiple substrates at the same interaction site. This can be exploited for the identification and design of novel substrates. Multi-specificity can be modeled with MFPred[72], which is a rapid, flexible-backbone self-consistent mean field theory-based technique. MFPred can predict experimentally determined peptide specificity profiles for a range of receptors, at equivalent or better prediction accuracy and a 10- to 1000-fold lower computational cost when compared to other methods.

Loop modeling for structure prediction and design

Loop modeling was implemented early in Rosetta[73,74] to generate structures for loops or gaps in proteins, with initial approaches relying on fragments using the Cyclic Coordinate Descent (CCD) algorithm[75]. Subsequent developments introduced kinematic closure methods ("termed KIC") into Rosetta which produced more native-like loop conformations[76]. KIC was used for modeling protein surface and interface loops, and to design and refine active site loops or regions binding small molecules[77]. Next-Generation KIC (NGK)[17] made improvements to sampling loop conformations by employing diversification (i.e. sampling a wider range of possible conformations) and intensification (i.e. to focus sampling on previously generated conformations) to identify lowest-energy conformations. NGK substantially increased the fraction of near-native models[17] and allowed modeling of longer loops. GeneralizedKIC[39] (GenKIC) samples or perturbs loop geometries between fixed endpoints for non-standard peptide chemistries, for instance for non-canonical backbones. GenKIC can sample backbone conformations containing L- and D-α-amino acids, β-amino acids, peptoids, oligoureas, or more exotic chemical building-blocks for which template structures do not exist in structural databases. Additionally, GenKIC can sample conformations involving side-chain connections (*e.g.* disulfide bonds, side-chain isopeptide bonds, *etc.*), covalently-attached ligands or crosslinkers, or chemistries that conventional loop-modelling algorithms do not typically handle.

Most Rosetta loop modeling algorithms were primarily developed for structure prediction. However, design poses the opposite problem, finding low-energy sequence–structure combinations that satisfies certain design goals. LoopHashKIC[78] addresses this problem and uses the Rosetta LoopHash algorithm[79] to efficiently query a database of loop conformations based on rigid-body transforms between the first and last loop residues. LoopHashKIC uses LoopHash to identify a suitable peptide fragment, and then uses KIC to find an exact solution to close the backbone. To improve the local sequence-structure compatibility in *de novo* designed loops, the *ConsensusLoopDesign* task operation implemented in Rosetta Scripts allows to restrict the amino acid identities of loops based on sequence profiles of naturally occurring loops with the same ABEGO backbone torsion bins[44,80].

Modeling antibodies and proteins in the immune system

Due to the therapeutic significance of antibodies, a number of protocols have been developed in our community for structure prediction, docking and design that involve antibodies and other proteins in the immune system, such as T-cell receptors (TCR), displayed antigens of the Major Histocompatibility Complex (MHC) and other soluble antigens and immunogens. An overview of the various applications can be found in Table 1.

protocol	**target**	**task**	**comments**
RosettaAntibody[81–84] Gray lab	antibodies	homology modeling / docking	Identifies templates, assembles them into a structure, and models loops *de novo* while refining VH-VL orientation[85]; allows multiple templates; uses a key constraint[86,87] for CDR H3 modeling; good for modeling camelid antibodies[88] and antibodies on the scale of the human repertoire[89,90]
AbPredict[91] Fleishman lab	antibodies	structure prediction	Does not rely on templates, samples backbone fragments and rigid-body orientations from known structures, without considering sequence homology, hence being able to model antibodies accurately with sequence identity as low as 10%; AbPredict2 available as webserver[92]
RosettaMHC[93] Sgourakis lab	antigen / MHC-I / (chaperone or T-cell receptor)	modeling / docking	Predicts peptides antigens that bind to all known MHC-I alleles models peptide/MHC-I structures[94]; code implementation in PyRosetta
RosettaTCR[95] Pierce lab	T-cell receptors	structure prediction	Models TCRs from sequence, via template identification, grafting of loop templates onto framework regions, minimization and loop refinement; to gain structural insights into TCRs, e.g. those targeting cancer neoepitopes[96], or to identify features of sets of TCRs from high throughput

			sequencing; can be combined with docking to generate models of TCR-peptide-MHC complexes[97] or TCRs in complex with non-peptide antigens bound to MHC-like proteins[98].
SnugDock[99] Gray lab	antibody-antigen	docking	Input is starting conformation and optionally an ensemble of antibodies/antigens, then runs local docking to refine both the antibody–antigen interface and the heavy–light chain interface (within the antibody) and re-models the CDR H2/H3 loops at the interface; also includes a CDR H3 structural constraint[86,87] and docking of camelid antibodies[88]
RosettaAntibodyDesign[100] (RAbD) Dunbrack lab	antibody-antigen	design	Based on RosettaAntibody[83]; design of specific CDRs of different clusters and lengths, sequence design using cluster-based CDR profiles or conservative mutations, or *de novo* design of whole antibodies; based on the North-Dunbrack CDR clustering[101], reducing deleterious sequence mutations; benchmarked on a diverse set of 60 interfaces from both lambda and kappa antibodies; experimental benchmarking of two complexes showed affinity improvements between 10 and 50-fold
Epitope removal[102,103] Baker lab	MHC epitopes	design	Includes experimental immunogenic epitope data, MHC epitope prediction tools, and host genomic data to enable protein design with reduced immunogenicity while retaining function and stability[102]; uses a machine learning-based epitope prediction for 28 different alleles, restricts design to select 15mer epitope regions, and uses a greedy stepwise protein design[103] to eliminate the most immunogenic epitopes with as few mutations as possible, avoiding disruptive core mutations likely to destabilize the protein
AbDesign[104,105] Fleishman lab	antibodies	design	Cuts experimentally determined antibody structures along conserved positions to create interchangeable segments, recombines them to produce novel antibody models[104,105]; models are docked to a target of interest, either locally to a specific epitope, or globally, followed by optimization comprised of backbone sampling and sequence design for improving stability and binding affinity

Including experimental data into the modeling process

The use of experimental data in modeling can vastly restrict the conformational search space, therefore allowing the modeling of larger, more complex biomolecules to greater detail. Electron density maps from cryo-electron microscopy (cryoEM) or X-ray crystallography have become more readily available in the past decade and methods to incorporate this type of data have been successfully used for high-resolution structure determination. Since cryoEM density maps are often of too low resolution to unambiguously assign all densities to residues in the protein, *de novo* structure determination methods require a combinatorial search procedure to accomplish that. A *de novo* method described by Wang et al. applies a model building approach[106] for density maps between 3-5Å that fits fragments into densities and scores their match based on secondary structure, fit with density, loop closure, clashes and consistency between overlapping fragments to assign sequence into densities. While this method requires >70% of the map to be assigned initially, the improvement of this method, the RosettaES[107] enumerative sampling approach forgoes this requirement. RosettaES gradually extends the model one residue at a time until all residues have been assigned. At each iteration short fragments are used to sample the nearby conformational space of the growing model, while undergoing a series of clustering and filtering steps based on the Rosetta energy and fit to the density.

If assignment is complete but the data are of low resolution, refinement into density maps is necessary. Several methods have been developed for density maps in the 3.0-4.5Å resolution range. One method[108] iterates between refinement with Phenix in reciprocal space to physically plausible conformations, and Rosetta in real space, because Rosetta's all-atom scorefunction compensates for the lack of high-resolution data, while the density map restraints backbone and sidechain sampling in real space. Refinement can also be accomplished starting from homology models, then density-guided rebuilding and refinement of coordinates and B-factors[109]. More recently, an automated fragment-guided refinement pipeline[110] splits the density map into independent training and validation maps. It finds regions with poor density fit, iteratively rebuilds them with fragments using the training map, filters the models based on their fit to the validation map, model geometry from MolProbity and fit to the full map, and then optimizes against the full map. The frameworks for electron density maps and carbohydrate modeling[111] in Rosetta (below) were connected[112] for refinement of carbohydrates into low-resolution electron density maps from cryoEM or crystallography.

NMR data were incorporated into *de novo* structure prediction early in Rosetta development, creating RosettaNMR. Chemical shifts were used for fragment picking using CS-Rosetta[113], which could be used in conjunction with NOE, RDC[114], PCS[115–117] and PRE data. Improvements, for instance through RASREC resampling[118] allowed the use of sparse[119] or unassigned data[120], easier obtainable data (backbone-only[121]), modeling larger and more complex proteins[122], membrane proteins[123], symmetric systems[124], and combination with data from SAXS[125], cryoEM[126], distance restraints from homologous proteins[127] and evolutionary couplings[128]. CS-Rosetta also has the AutoNOE[129,130] module for automatic assignment of NOESY data for use in structure calculations. RosettaNMR was recently overhauled and reconciled with CS-Rosetta and PCS-Rosetta to allow seamless integration of the different types of NMR restraints (CS, RDC, PCS, PRE, NOE) in one consistent framework[131] that could be applied to structure prediction, protein-protein docking, protein-ligand docking, and symmetric assemblies.

Covalent labeling mass spectrometry data provides information on relative solvent exposure of residues, therefore yielding information on protein tertiary structure. A low-resolution score term from hydroxyl radical footprinting has been implemented in Rosetta that can improve model quality in structure prediction[132,133]. Finally, data from chemical cross-linking mass spectrometry has been incorporated into an automated workflow to identify protein-protein interactions. The PyTXMS[134] method combines the sensitivity of mass spectrometry to analyze complex samples with the power of Rosetta structural modeling and protein-protein docking to efficiently sample the vast conformational space and identify interactions (Figure 3C). A machine learning algorithm based on high resolution MS1 data guide the potential binding interface selection which is then validated and adjusted by a repository of structural models and MS2 (DDA) samples.

Modeling DNA and RNA and their interactions with proteins

Considerable advances have been made in the representation of nucleic acids in Rosetta. The *StepWise Monte-carlo* protocol (SWM) for RNA structure prediction[135] provides a substantial acceleration over the original enumerative *StepWise Assembly* (SWA) method[136,137]. SWM builds loop residues one-by-one, has been generalized for proteins, and allows it to approach new substrates: it can directly model chemically modified nucleotides, L-RNA, and it can dock ligands or add them along with nucleotides in idealized "submotif" positions.

Our *Fragment Assembly of RNA with Full-Atom Refinement* (FARFAR) structure prediction protocol[138,139] also permits working with chemically modified nucleotides, picking fragments for the most chemically similar base available. Homologous fragments can automatically be eliminated from fragment sets to give pseudo-blind prediction results. Structures may also be refined against electron density maps by *Enumerative Real-space Refinement ASsisted by Electron density under Rosetta*[140,141] (ERRASER). Alternatively, ^{1}H NMR data can be used for RNA modeling via the CS-Rosetta-RNA protocol[142].

The fragment assembly protocol also supports modeling RNA-protein complexes through simultaneous folding and docking[143]. RNA-protein interactions are handled via additional knowledge-based score terms that supplement the low-resolution RNA scorefunction. Structure coordinates can further be built into cryo-EM density maps for large RNA-protein complexes with DRRAFTER (*De novo Ribonucleoprotein modeling in Real space through Assembly of Fragments Together with Experimental density in Rosetta*)[144].

doi:10.20944/preprints201904.0263.v1

The scorefunction can be augmented to reward the discovery of particular non-canonical features such as kink turns and tetraloop-receptor interactions, as well as with a six-dimensional statistical potential for loop closure. SWM can be adapted to perform RNA design, and a reference energy term was implemented to calculate the RNA secondary structure of the current sequence. Further scorefunction developments include rewarding the formation of highly stereotyped non-canonical structural features like kink-turns and A-minor interactions.

Redesign and prediction of protein-DNA interfaces is also possible in Rosetta[145,146] and has been accomplished with flexible protein backbones[147] and motif-biased rotamer sampling[148]. However, the biggest limitation of these approaches is that they rely on fixed DNA backbone conformations, which in nature can be highly flexible. Key to successful protein-DNA design is an energy function that is optimized[149] for these highly polar and solvated interfaces. Rosetta further supports prediction of specificity and affinity[150] and the prediction of DNA binding preferences of homologous proteins. Multi-template modeling in RosettaCM[151] was successfully applied to this challenge[152]. To accomplish this, protein homology modeling was followed by docking of multiple competing DNA sequences threaded onto the original crystal structure backbone and comparing the energies of the resulting protein-DNA complexes.

Modeling membrane proteins

Membrane proteins constitute about 30% of all proteins and are targets for over 60% of pharmaceuticals on the market[153]. However, experimental difficulties have limited our understanding of their structures[154]. Previously, Yarov-Yarovoy[155,156] and Barth[157] implemented tools for low- and high-resolution structure prediction of membrane proteins, termed RosettaMembrane. These tools were recently re-engineered for compatibility with Rosetta3[2] into a platform called RosettaMP[158]. RosettaMP implements core modules for representing, sampling, and scoring proteins in the context of an implicit membrane. RosettaMP is compatible with key modeling protocols including docking, design, $\Delta\Delta G$ prediction[48], PyMOL visualization[159], and assembly of symmetric proteins. In addition, a set of basic modeling tools[160] has been implemented, for instance for scoring, transforming a membrane protein into the membrane coordinate frame, *de novo* modeling for single transmembrane span helices, introducing mutations, and visualization in the membrane. RosettaMP has further enabled rapid development of new modeling tools including structure-based detection of lipid exposed residues in the membrane[161] environment and domain assembly of full-length protein models from structures of transmembrane and soluble domains[162]. The RosettaCM protocol for multi-template homology modeling has also been adopted for membrane proteins[8].

Adding carbohydrates to the modeling process

Carbohydrates are fundamental to life[163,164], but because of challenges in experimental characterization and computational sampling and scoring, they have been historically under-studied. The RosettaCarbohydrate framework[111] allows modeling of carbohydrate structures and complexes. The framework is integrated into Rosetta such that it is possible to model glycosylated proteins or protein–sugar complexes (Figure 3F) with the same algorithms one would use for proteins. RosettaCarbohydrate is not limited to commonly studied sugars but can handle the full gamut of carbohydrate structures, including linear, cyclic, and branched structures, sugar modifications, and conjugations. Methods exist for sampling ring conformations, packing substituents, refining glycosidic linkages, sampling from linkage “fragments”, and extending glycan chains. Scoring of saccharide-containing sugars includes a quantum-mechanically derived intrinsic backbone term[165]. Because saccharide residues are stored as distinct data structures, we can integrate bioinformatic and statistical data into our algorithms, which opens the doors for glycoengineering and design applications. RosettaCarbohydrate has been integrated with various other frameworks in Rosetta, such as loop modeling (GenKIC and Stepwise Assembly), refinement (*GlycanTreeModeler*), symmetry, and RosettaScripts-accessible classes such as *MoveMaps* and *ResidueSelectors*. Linkages are automatically determined during PDB read-in. Carbohydrates work with Cartesian minimization, and they can be refined into electron density maps[112].

2. The brain of Rosetta: its scorefunction

Rosetta's energy function has been continuously improved over many years[166] and a full review of the modern all-atom energy function has been published recently[3]. Briefly, the newest energy function REF2015[4] (REF stands for Rosetta Energy Function) features two main advancements. First, reproducibility of thermodynamic observables (such as liquid-phase properties[167] and liquid-to-vapor transfer free energies[168]) has been added to the optimization objectives, in addition to structure[169]-based tests. Second, a new, derivative-free optimization technique has been developed, which is suitable for robust optimization of >100 parameters. Further, a new energy term has been added that takes into consideration non-ideality of bond lengths and angles in cartesian space[170]. The cartesian term[170] is also the basis for a *cartesian_ddG* method that has been used to calculate ΔΔGs of mutation to probe protein stability, in which the backbones and sidechains of residues nearby the mutation site are allowed to move[171]. Due to the local optimization, this protocol is much faster than *ddg_monomer*[172], while retaining the same level of accuracy. The default Rosetta energy function is now also compatible with an expanded palette of chemical building-blocks: canonical and non-canonical L-α-amino acids and their D-amino acid counterparts, exotic achiral amino acids like 2-aminoisobutyric acid (AIB), peptoids, and oligoureas. The ability to model metalloproteins has also been added to Rosetta[173,174]. The scorefunction is now thread-safe and fully mirror symmetric, i.e. enantiomers in mirror conformations score identically. Guidance energy terms for design have been added to encourage certain features, such as specific amino acid compositions[39,71], hydrogen bonding networks, or global or local net charges, and discourage others, such as repeat sequences that hinder NMR assignments, buried unsatisfied hydrogen bond donors and acceptors, or voids within the protein[175]. Further, a consensus scoring method, which utilizes the semi-orthogonal nature of the Rosetta and Amber energy functions, was developed for model ranking to identify most near-native models[176] from the pool of generated decoys. This approach led to the development of a python-based tool (AmbRose) for interconversion between Rosetta and Amber models to facilitate consensus scoring.

Hydrogen bond networks are important for biomolecular structure and catalysis but have been challenging to design because of pairwise interactions that have multi-body, cooperative properties. The HBNet protocol[177] has been used to design *de novo* coiled coils with interaction specificity mediated by designed hydrogen bond networks, including homo-oligomers[177], membrane proteins[60], and large sets of orthogonal heterodimers[178]. An improvement to HBNet uses a Monte Carlo search procedure to sample hydrogen bond networks with drastically improved performance[179]. We further developed a statistical potential to place highly-coordinated water molecules on the surface of biomolecules. On a data set of 153 high-resolution protein-protein interfaces, the method predicts 17% of native interface waters with 20% precision within 0.5 Å of the crystallographic water positions [REF Ryan???]. The potential is accessible through the WaterBoxMover in RosettaScripts.

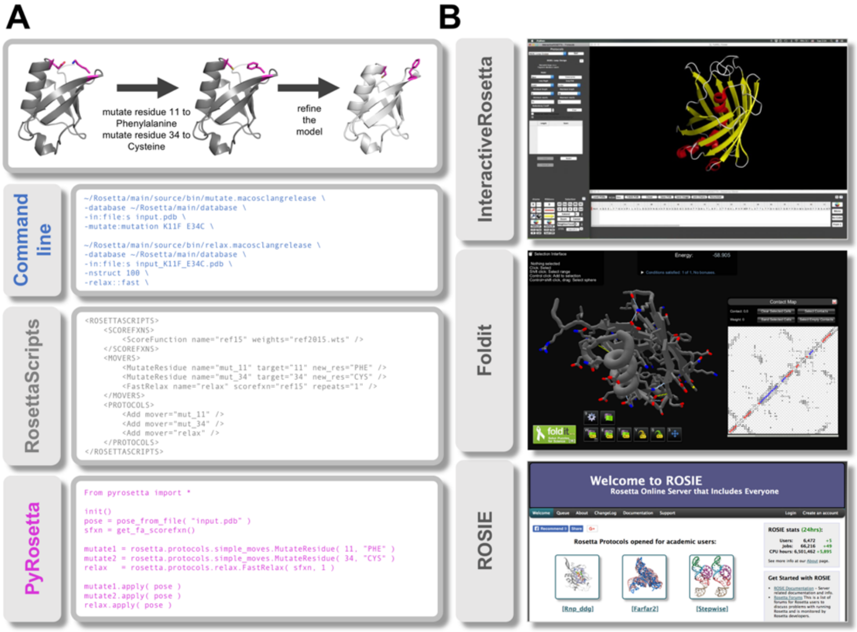
3. User interfaces and usability

Advances in Rosetta were also geared towards different user interfaces and improving usability of the software (Figure 4). The command line interface was the first and is still the most often use interface to the Rosetta codebase. Structure input and output was enhanced by the ability to read mmCIF files using the same mechanisms as PDB files, which permits representation of large complexes that are ill-suited for the PDB format. This comes with the ability to read the Protein Databank's Chemical Component Dictionary, the description of the chemical composition of residues in the officially released PDB structures. Multithreading support has been implemented into Rosetta, for which it has undergone a major refactor of its core architecture for thread-safety, allowing shared-memory parallelism. Multithreading is currently available for specific protocols with planned expansion to other applications.

In addition to the command line interface, Rosetta features two major scripting interfaces. PyRosetta[5,180] is a collection of Python bindings to the ~3 million lines of Rosetta C++ code. The PyRosetta scripts interface with precompiled code objects allows custom protocol development that is flexible and fast, yet requires familiarity with the codebase.

RosettaScripts[6] is a popular scripting interface that uses E**x**tensible **M**arkup **L**anguage (XML) to build fairly complex protocols using core Rosetta machinery[2]. Comprehensive knowledge of the codebase is unnecessary since most of the underlying modules[8] have been thoroughly documented – documentation is now also generated using XML schema, which validates the RosettaScripts XML files on-the-fly. RosettaScripts was further extended and generalized to enhance consistency: *ResidueSelectors* enable selection of residues based on specific properties such as chain, amino acid, secondary structure, index, burial, and others, and can be used in conjunction with *MoveMapFactories*, which control a structure's flexibility during energy minimization. *ResidueSelectors* are also accepted by *TaskOperations* which control side-chain identity and optimization. A more general analysis tool, *SimpleMetrics*, allows custom analyses of models and writes the output into the scorefile. The *SimpleMetrics* system is more general than previous tools, such as the *interface_analyzer* or the *FeaturesReporter*. In order to take full advantage of those protocols, RosettaScripts is now also fully callable from PyRosetta.

Figure 4: User interfaces to Rosetta
(A) Rosetta can be run from a terminal and offers three different interfaces to the codebase. The top panel outlines the task to be accomplished: making two mutations in a protein and then refining the structure. The panels underneath show how this task can be accomplished in the different interfaces. The command line panel shows the executable, input files and options to run two specific Rosetta applications. RosettaScripts is an XML-based scripting language that offers more flexibility by combining *Movers* and *ScoreFunctions* into a custom *Protocol*. PyRosetta offers direct access to the underlying Rosetta code objects but requires knowledge of the Rosetta codebase. (B) Point-and-click interfaces to Rosetta. InteractiveRosetta is a graphical user-interface (GUI) to PyRosetta. It offers controls to the most popular protocols, file formats and options. Foldit is a videogame primarily used to crowd-source real-world scientific puzzles but can also be used on custom proteins of interest. It allows access to some popular applications via a game interface. ROSIE is a super-server hosting a multitude of servers each executing a particular Rosetta protocol. It currently includes servers for 21 Rosetta methods.


A
B
mutate residue 11 to Phenylalanine mutate residue 34 to Cysteine
refine the model
Command line
RosettaScripts
PyRosetta
InteractiveRosetta
Foldit
ROSIE
Energy: -98.903
Contact Map
Welcome to ROSIE
Rosetta Online Server that Includes Everyone
Welcome
Queue
About
ChangeLog
Documentation
Support
Login
Create an account
Recommend 0
Share
Rosetta Protocols opened for academic users:
[Rnp_ddg]
[Farfar2]
[Stepwise]
ROSIE stats (24hrs):
Users: 6,472 +5
Jobs: 66,216 +49
CPU hours: 6,501,462 +5,895
See more info at our About page.
Get Started with ROSIE
ROSIE Documentation - Server related documentation and info.
Rosetta Forums This is a list of forums for Rosetta users to discuss problems with running Rosetta and is monitored by Rosetta developers.

```
~/Rosetta/main/source/bin/mutate.macosclangrelease \
-database ~/Rosetta/main/database \
-in:file:s input.pdb \
-mutate:mutation K11F E34C \

~/Rosetta/main/source/bin/relax.macosclangrelease \
-database ~/Rosetta/main/database \
-in:file:s input_K11F_E34C.pdb \
-nstruct 100 \
-relax::fast \
```

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="ref15" weights="ref2015.wts" />
    </SCOREFXNS>
    <MOVERS>
        <MutateResidue name="mut_11" target="11" new_res="PHE" />
        <MutateResidue name="mut_34" target="34" new_res="CYS" />
        <FastRelax name="relax" scorefxn="ref15" repeats="1" />
    </MOVERS>
    <PROTOCOLS>
        <Add mover="mut_11" />
        <Add mover="mut_34" />
        <Add mover="relax" />
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

```
From pyrosetta import *

init()
pose = pose_from_file( "input.pdb" )
sfxn = get_fa_scorefxn()

mutate1 = rosetta.protocols.simple_moves.MutateResidue( 11, "PHE" )
mutate2 = rosetta.protocols.simple_moves.MutateResidue( 34, "CYS" )
relax   = rosetta.protocols.relax.FastRelax( sfxn, 1 )

mutate1.apply( pose )
mutate2.apply( pose )
relax.apply( pose )
```

InteractiveROSETTA[181] is a graphical interface for the PyRosetta framework that presents easy-to-use controls for several of the most widely-used Rosetta protocols alongside a selection system utilizing PyMOL as a visualizer. InteractiveROSETTA is capable of interacting with remote servers running a standalone Rosetta install, rendering it easy to incorporate more sophisticated protocols that are not accessible in PyRosetta and/or require significant computational resources.

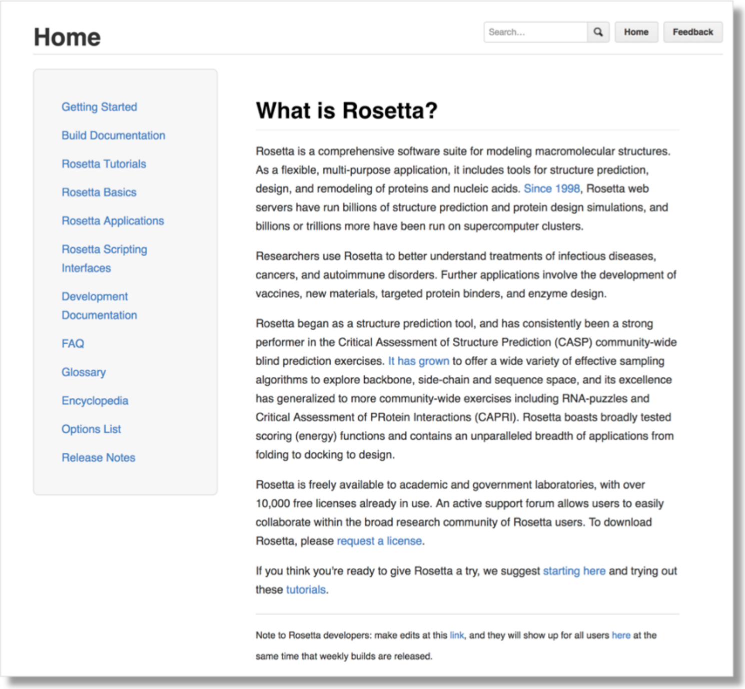
Foldit Standalone[182,183] is a graphical interface to Rosetta based on the Foldit video game[7,184]. Foldit Standalone provides several interactive structure manipulations, including pulling directly on the structure, rigid body docking, and residue mutation, insertion and deletion. Users can apply hard and soft constraints that guide automated moves such as packing and minimization, and provides real-time

doi:10.20944/preprints201904.0263.v1

scoring updates as the structure changes. Additional features include multiple sequence alignments for template-based modeling, along with electron density, Ramachandran map, and contact map visualizations. Further, scientists and educators can now run their own custom Foldit puzzles for a group of their choosing, a new feature called “Custom Contests”[185].

The Rosetta community has further devoted an enormous effort to enhance the user friendliness of Rosetta by rewriting and adding documentation (Figure 5). We now use a Gollum wiki (https://www.rosettacommons.org/docs/latest/Home) for various levels of documentation, such as application documentation, tutorials for beginning users, and static protocol captures that accompany manuscripts for scientific reproducibility (see supplement for links). The Gollum wiki is easily editable by members of the RosettaCommons. The `–info` option can be passed on the command line to see available options in RosettaScripts, and debugging command lines is facilitated by improved error messages. Default options can now be specified that are used for all Rosetta runs (for instance for database paths), in a *rosetta.rc* file. Lastly, code development in C++ is now easier with the help of available code templates.

Figure 5: Main external documentation page for Rosetta
In 2015, our community performed a complete overhaul of the documentation for Rosetta. Documentation is now hosted on a Gollum wiki, which is version controlled and easily editable for members of our community. Accessibility and ability to edit the documentation has drastically improved the user-experience of Rosetta.



What is Rosetta?

Rosetta is a comprehensive software suite for modeling macromolecular structures. As a flexible, multi-purpose application, it includes tools for structure prediction, design, and remodeling of proteins and nucleic acids. Since 1998, Rosetta web servers have run billions of structure prediction and protein design simulations, and billions or trillions more have been run on supercomputer clusters.

Researchers use Rosetta to better understand treatments of infectious diseases, cancers, and autoimmune disorders. Further applications involve the development of vaccines, new materials, targeted protein binders, and enzyme design.

Rosetta began as a structure prediction tool, and has consistently been a strong performer in the Critical Assessment of Structure Prediction (CASP) community-wide blind prediction exercises. It has grown to offer a wide variety of effective sampling algorithms to explore backbone, side-chain and sequence space, and its excellence has generalized to more community-wide exercises including RNA-puzzles and Critical Assessment of PRotein Interactions (CAPRI). Rosetta boasts broadly tested scoring (energy) functions and contains an unparalleled breadth of applications from folding to docking to design.

Rosetta is freely available to academic and government laboratories, with over 10,000 free licenses already in use. An active support forum allows users to easily collaborate within the broad research community of Rosetta users. To download Rosetta, please request a license.

If you think you're ready to give Rosetta a try, we suggest starting here and trying out these tutorials.

Note to Rosetta developers: make edits at this link, and they will show up for all users here at the same time that weekly builds are released.

A limitation of Rosetta is the need for a local installation and compilation in a Unix-like environment. Webservers provide a user-friendly alternative and a number of independent servers have emerged in the Rosetta community. However, implementing and maintaining such servers comes at a substantial cost. To make it easier to provide Rosetta protocols as a webserver, ROSIE (Rosetta Online Server that Includes Everyone)[70,186] (http://rosie.rosettacommons.org/) provides a simple framework for “serverification” of protocols. ROSIE currently contains 21 webservers, with additional protocols continually being added.

Conclusion

The Rosetta software is developed by a large, global community that aims to solve complex problems through collaborative solutions and code implementations. In the last five years, great strides have been accomplished in Rosetta. More protocols are available now that allow us to model a broader range of biologically- and chemically-realistic systems, larger macromolecular complexes and in general more sophisticated systems. Prediction accuracies have improved through advances in the scorefunction, which is a combination of physics-based and knowledge-based potentials that were also fit against thermodynamic observables. Incorporating experimental data into the modeling process has also been facilitated. Further, our community saw the need to develop more general, reusable, user-friendly, and scientifically reproducible protocols. This was motivated by the growth of the software and the developer community, the various user interfaces, the diversity of the community[1], and the complexities of the protocols in this sophisticated piece of software. Over the years, these applications have moved away from simply tackling basic science questions to more application-based scientific developments that will lead to therapeutic outcomes. Documentation has been drastically improved to allow users to quickly start using or developing custom protocols, while facilitating user support for the various Rosetta interfaces (command line, RosettaScripts, PyRosetta). The myriad of advances described above have made integration of Rosetta into existing experimental and computational scientific workflows increasingly useful and expected. Rosetta's predictive powers are highly advantageous, because they cannot only be used to analyze and verify existing data but to inform experiments and therefore enormously accelerate the engineering of useful enzymes, enable the creation of new biomaterials, and permit the discovery of potent new therapeutics.

Acknowledgements:

100MCPU-hour donation from Google Inc to Conway; AHA 18POST34080422 to Kuenze; AMED J-PRIDE JP18fm0208022h to Kuroda; Biltema Foundation to Correia; Boehringer Ingelheim Fonds to Norn; Computing was performed using resources of the Argonne Leadership Computing Facility at Argonne National Laboratory which is supported by the Office of Science of the US to Conway; DFG KU 3510/1-1 to Kuenze; DOE DE-AC02-06CH11357 to Conway; DP120100561 to Huber; DP150100383 to Huber, Pilla; Defense Threat Reduction Agency to King; EMBO long-term fellowship ALTF 698-2011 to Stein; EPFL-Fellows - H2020 Marie Sklodowska-Curie to Bonet; European Research Council Grant 310873 to Schueler-Furman, Alam; European Research Council Grant 310873 to Sedan, Marcu; European Research Council Starting grant - 716058 to Correia, Scheck; FT0991709 to Huber; Foundation of Knut and Alice Wallenberg 20160023 to Malmström; Google Exacycle for Visiting Faculty program to Conway; Hertz Foundation Fellowship to Alford; Howard Hughes Medical Institute to Mulligan, Basanta, Shi; Hyak supercomputer system supported in part by the University of Washington eScience Institute to Conway; Israel Science Foundation 2017717 to Schueler-Furman, Alam; Japan Society for the Promotion of Science JP17K18113 to Kuroda; MCB1330760 to Khare; Marie Curie International Outgoing Fellowship FP7-PEOPLE-2011-IOF 298976 to Marcos; NIAID T32AI007244 to Adolf-Bryfogle; NIAID U19 AI117905 to Sevy; NIEHS P42ES004699 to Siegel; NIGMS Ruth L Kirschstein National Research Service Award T32GM008268 to Conway; NIGMS T32 GM007628 to Bender; NIH 1R35 GM122579 to Das; NIH 1UH2CA203780 to Koepnick, Dsilva, Cooper, Kleffner, Khatib; NIH 5F32GM110899-02 to Linsky; NIH F31GM123616 to Jeliazkov; NIH F32CA189246 to Labonte; NIH P01 AI094419 to Schief; NIH R00 GM12038 to Horowitz; NIH R01 AI143997 to Nerli, Sgourakis; NIH R01 DK097376 to Bender, Meiler; NIH R01 GM073151 to Leaver-Fay, Lyskov, Gray, Meiler, Labonte, Fu; NIH R01 GM073960 to Jacobs; NIH R01 GM076324-11 to Siegel; NIH R01 GM078221 to Marze, Gray, Kuroda, Labonte, Roy Burman, Weitzner; NIH R01 GM080403 to Kuenze, Bender, Meiler; NIH R01 GM084453 to Adolf-Bryfogle, Dunbrack, Weitzner; NIH R01 GM088277 to Bradley; NIH R01 GM092802 to Ovchinnikov; NIH R01 GM092802-02 to Wang, Song, Park; NIH R01 GM098101 to Stein, Kortemme; NIH R01 GM099842 to Moretti, Meiler, Fu; NIH R01 GM099959 to Johnson; NIH R01 GM110089 to Ollikainen, Barlow, Kortemme, Ó Conchúir; NIH R01 GM117189 to Barlow, Kortemme, Ó Conchúir; NIH R01 GM117968 to Jacobs, Teets, Kuhlman, Guffy; NIH R01 GM121487 to Bradley; NIH R01 GM123089 to Pavlovicz, DiMaio; NIH R01 GM126299 to Gowthaman, Pierce; NIH R01 GM127578 to Gray, Labonte; NIH R01 GM99827 to Bystroff; NIH R01 GMN092802-04 to Baker, Conway; NIH R01 HL122010 to Bender, Meiler; NIH R01067553 to Maguire; NIH R01072281 to Alford; NIH R01084433 to Baker, Thyme; NIH R01088277 to Thyme; NIH R21 AI121799 to Meiler; NIH R21 CA219847 to Das; NIH R21 GM102716 to Das; NIH R35 GM122517 to Adolf-Bryfogle, Dunbrack; NIH R35 GM125034 to Nerli, Sgourakis; NIH

RL1CA133832 to Baker, Thyme; NIH U19 AI117905 to Meiler, Sauer; NIH/NCI Cancer Center Support Grant P30 CA006927 to Gowthaman, Karanicolas, Khar; NSF 1507736 to Roy Burman; NSF 1627539 to Siegel; NSF 1629879 to Koepnick, Dsilva, Cooper, Kleffner, Khatib; NSF 1805510 to Siegel; NSF 1827246 to Siegel; NSF CHE 1305874 to Moretti, Bender, Meiler, Fu; NSF CHE 1750666 to Aprahamian, Lindert; NSF CISE 1629811 to Meiler; NSF CNS-1629811 to Smith; NSF DBI-1262182 to Ollikainen, Stein, Kortemme; NSF DBI-1564692 to Pan; NSF DMR 1507736 to Lubin, Gray, Pacella; NSF GRF DGE-1433187 to Rubenstein; NSF Graduate Research Fellowship to Alford, Kappel; NSF MCB1330760 to Pethe; NSF MCB1716623 to Khare; NSF graduate research fellowship to Thyme; Open Philanthropy to Coventry; PhRMA Informatics Pre-Doctoral Fellowship U22879-001 to Smith; PhRMA foundation Predoctoral Fellowship to Fu; R21 CA219847 to Watkins; R21 GM102716 to Geniesse; RosettaCommons to Koepnick, Goldschmidt, Rubenstein, Dsilva, DiMaio, Cooper, Watkins, Kleffner, Szegedy, Geniesse, Blacklock, Das, Khare, Khatib, Koehler Leman, Kappel; SEB is funded by a Career Award at the Scientific Interface from Burroughs Wellcome Fund to Boyken; Simons Foundation to Mulligan, Bonneau, Renfrew, Koehler Leman; Stanford Graduate Fellowship to Kappel; Starter Grant from the European Research Council to Lapidoth; Swiss National Science Foundation - NCCR Molecular Systems Engineering 51NF40-141825 to Correia; Swiss National Science Foundation 310030_163139 to Correia; Swiss National Science Foundation SNF 200021 160188 to Malmström, Khakzad; UCSF/UCB Graduate Program in Bioengineering to Pan; USA-Israel Binational Science Foundation 2009418 to Raveh, Zimmerman, London; USA-Israel Binational Science Foundation 2009418, 2015207 to Schueler-Furman, Alam; USA-Israel Binational Science Foundation 2015207 to Khramushin; Washington Research Foundation Innovation Postdoctoral Fellowship to Weitzner;

Conflicts of Interest:

Yifan Song, Indigo C. King, Steven M. Lewis, Brandon Frenz, and Ryan Pavlovicz are currently employed at Cyrus Biotechnology with granted stock options. Cyrus Biotechnology distributes the Rosetta software, which may include methods described in this paper.

JJG is an unpaid board member of the Rosetta Commons. Under institutional participation agreements between the University of Washington, acting on behalf of the Rosetta Commons, Johns Hopkins University may be entitled to a portion of revenue received on licensing Rosetta software including programs described here.

As a member of the Scientific Advisory Board of Cyrus Biotechnology, JJG is granted stock options. Cyrus Biotechnology distributes the Rosetta software, which may include methods described in this paper.

doi:10.20944/preprints201904.0263.v1

References:

1. Koehler Leman, J., Weitzner, B. D., ... & Bonneau, R. Better together: 20+ years of scientific software development in the Rosetta macromolecular modeling suite. *Prep.* (2019).
2. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J. M., 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., Berrondo, M., Mentzer, S., Popovic, 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.* **487,** 545–74 (2011).
3. 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.* **13,** 1–35 (2017).
4. Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D. & DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12,** 6201–6212 (2016).
5. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26,** 689–691 (2010).
6. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J. & Baker, D. RosettaScripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6,** 1–10 (2011).
7. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. & Players, F. Predicting protein structures with a multiplayer online game. *Nature* **466,** 756–760 (2010).
8. Bender, B. J., Cisneros, A., Duran, A. M., Finn, J. A., Fu, D., Lokits, A. D., Mueller, B. K., Sangha, A. K., Sauer, M. F., Sevy, A. M., Sliwoski, G., Sheehan, J. H., Dimaio, F., Meiler, J. & Moretti, R. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* acs.biochem.6b00444 (2016). doi:10.1021/acs.biochem.6b00444
9. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M. & Baker, D. Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS One* **6,** e23294 (2011).
10. Song, Y., Dimaio, F., Wang, R. Y., Kim, D. E., Miles, C., Brunette, T. J., Thompson, J. & Baker, D. Supplemental Information High-Resolution Comparative Modeling with RosettaCM. **21,**
11. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15674–9 (2013).
12. Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C. & Baker, D. Protein structure determination using metagenome sequence data. *Science (80-.).* **355,** 294–298 (2017).
13. Park, H., Ovchinnikov, S., Kim, D. E., Dimaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 3054–3059 (2018).
14. Tyka, M. D., Keedy, D. A., André, I., Dimaio, F., Song, Y., Richardson, D. C., Richardson, J. S. & Baker, D. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–18 (2011).
15. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.* **380,** 757–74 (2008).
16. Kapp, G. T., Liu, S., Stein, A., Wong, D. T., Remenyi, A., Yeh, B. J., Fraser, J. S., Taunton, J., Lim, W. A. & Kortemme, T. Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proc. Natl. Acad. Sci.* **109,** 5277–5282 (2012).
17. Stein, A. & Kortemme, T. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* **8,** e63090 (2013).
18. Pacella, M. S., Koo, D. C. E., Thottungal, R. A. & Gray, J. J. Using the RosettaSurface Algorithm to Predict Protein Structure at Mineral Surfaces. *Methods Enzymol.* **532,** 343–366 (2013).
19. Lubin, J. H., Pacella, M. S. & Gray, J. J. A Parametric Rosetta Energy Function Analysis with LK Peptides on SAM Surfaces. *Langmuir* **34,** 5279–5289 (2018).
20. Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein–

protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469 (2018).
21. Roy Burman, S. S., Yovanno, R. A. & Gray, J. J. Flexible backbone assembly and refinement of symmetrical homomeric complexes. *bioRxiv* 409730 (2018). doi:10.1101/409730
22. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS One* **6,** e20450 (2011).
23. Meiler, J. & Baker, D. RosettaLigand: protein-small molecule docking with full side-chain flexibility. *Proteins* **65,** 538–48 (2006).
24. DeLuca, S., Khar, K. & Meiler, J. Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One* **10,** e0132508 (2015).
25. Davis, I. W. & Baker, D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **385,** 381–392 (2009).
26. Fu, D. Y. & Meiler, J. Predictive Power of Different Types of Experimental Restraints in Small Molecule Docking: A Review. *J. Chem. Inf. Model.* **58,** 225–233 (2018).
27. Fu, D. Y. & Meiler, J. RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach. *ACS Omega* **3,** 3655–3664 (2018).
28. Johnson, D. K. & Karanicolas, J. Selectivity by Small-Molecule Inhibitors of Protein Interactions Can Be Driven by Protein Surface Fluctuations. *PLOS Comput. Biol.* **11,** e1004081 (2015).
29. Johnson, D. K. & Karanicolas, J. Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface. *PLoS Comput. Biol.* **9,** e1002951 (2013).
30. Johnson, D. K. & Karanicolas, J. Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein–Protein Interactions. *J. Chem. Inf. Model.* **56,** 399–411 (2016).
31. Gowthaman, R., Miller, S. A., Rogers, S., Khowsathit, J., Lan, L., Bai, N., Johnson, D. K., Liu, C., Xu, L., Anbanandam, A., Aubé, J., Roy, A. & Karanicolas, J. DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites. *J. Med. Chem.* **59,** 4152–4170 (2016).
32. Khar, K. R., Goldschmidt, L. & Karanicolas, J. Fast Docking on Graphics Processing Units via Ray-Casting. *PLoS One* **8,** e70661 (2013).
33. Gowthaman, R., Lyskov, S. & Karanicolas, J. DARC 2.0: Improved Docking and Virtual Screening at Protein Interaction Sites. *PLoS One* **10,** e0131612 (2015).
34. Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T. & Kuhlman, B. Design of structurally distinct proteins using strategies inspired by evolution. **352,** (2016).
35. Guffy, S. L., Teets, F. D., Langlois, M. I. & Kuhlman, B. Protocols for Requirement-Driven Protein Design in the Rosetta Modeling Program. *J. Chem. Inf. Model.* **58,** 895–901 (2018).
36. Lapidoth, G., Khersonsky, O., Lipsh, R., Dym, O., Albeck, S., Rogotner, S. & Fleishman, S. J. Highly active enzymes by automated combinatorial backbone assembly and sequence design. *Nat. Commun.* **9,** 2780 (2018).
37. Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R. & Baker, D. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **6,** e24109 (2011).
38. Blacklock, K. M., Yang, L., Mulligan, V. K. & Khare, S. D. A computational method for the design of nested proteins by loop-directed domain insertion. *Proteins Struct. Funct. Bioinforma.* **86,** 354–369 (2018).
39. Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., Buchko, G. W., Pulavarti, S. V. S. R. K., Kaas, Q., Eletsky, A., Huang, P.-S., Johnsen, W. A., Greisen, P. J., Rocklin, G. J., Song, Y., Linsky, T. W., Watkins, A., Rettie, S. A., Xu, X., Carter, L. P., Bonneau, R., Olson, J. M., Coutsias, E., Correnti, C. E., Szyperski, T., Craik, D. J. & Baker, D. Accurate de novo design of hyperstable constrained peptides. *Nature* **538,** 329–335 (2016).
40. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A Generic Program for Multistate Protein Design. *PLoS One* **6,** e20937 (2011).
41. Sevy, A. M., Jacobs, T. M., Crowe, J. E. & Meiler, J. Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Comput. Biol.* **11,** e1004300 (2015).
42. Sevy, A. M., Wu, N. C., Gilchuk, I. M., Parrish, E. H., Burger, S., Yousif, D., Nagel, M. B. M.,

Schey, K. L., Wilson, I. A., Crowe, J. E. & Meiler, J. Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 1597–1602 (2019).

43. Sauer, M. F., Sevy, A. M., Crowe, J. E. & Meiler, J. Manuscript submitted. (2019).

44. Marcos, E., Basanta, B., Chidyausiku, T. M., Tang, Y., Oberdorfer, G., Liu, G., Swapna, G. V. T., Guan, R., Silva, D.-A., Dou, J., Pereira, J. H., Xiao, R., Sankaran, B., Zwart, P. H., Montelione, G. T. & Baker, D. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355,** 201–206 (2017).

45. Marcos, E. & Silva, D.-A. Essentials of *de novo* protein design: Methods and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8,** e1374 (2018).

46. Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., MacPherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., Klevit, R. E., Graham, B. S., Wyatt, R. T., Baker, D., Strong, R. K., Crowe, J. E., Johnson, P. R. & Schief, W. R. Proof of principle for epitope-focused vaccine design. *Nature* **507,** 201–206 (2014).

47. Bonet, J., Wehrle, S., Schriever, K., Yang, C., Billet, A., Sesterhenn, F., Scheck, A., Sverrisson, F., Veselkova, B., Vollers, S., Lourman, R., Villard, M., Rosset, S., Krey, T. & Correia, B. E. Rosetta FunFolDes - A general framework for the computational design of functional proteins. *PLoS Comput. Biol.* **14,** e1006623 (2018).

48. Kroncke, B. M., Duran, A. M., Mendenhall, J. L., Meiler, J., Blume, J. D. & Sanders, C. R. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. *Biochemistry* **55,** 5002–5009 (2016).

49. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 14116–21 (2002).

50. Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci. STKE* **2004,** pl2 (2004).

51. Ó Conchúir, S., Barlow, K. A., Pache, R. A., Ollikainen, N., Kundert, K., O'Meara, M. J., Smith, C. A. & Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* **10,** e0130433 (2015).

52. Barlow, K. A., Ó Conchúir, S., Thompson, S., Suresh, P., Lucas, J. E., Heinonen, M. & Kortemme, T. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **122,** 5389–5399 (2018).

53. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **380,** 742–56 (2008).

54. Dou, J., Doyle, L., Jr Greisen, P., Schena, A., Park, H., Johnsson, K., Stoddard, B. L. & Baker, D. Sampling and energy evaluation challenges in ligand binding protein design. *Protein Sci.* **26,** 2426–2437 (2017).

55. Smith, C. A. & Kortemme, T. Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.* **402,** 460–74 (2010).

56. Smith, C. A. & Kortemme, T. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One* **6,** e20451 (2011).

57. Ollikainen, N., de Jong, R. M. & Kortemme, T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity. *PLoS Comput. Biol.* **11,** e1004335 (2015).

58. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6,** 685–689 (1953).

59. Dang, B., Wu, H., Mulligan, V. K., Mravic, M., Wu, Y., Lemmin, T., Ford, A., Silva, D.-A., Baker, D. & DeGrado, W. F. De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 10852–10857 (2017).

60. Lu, P., Min, D., DiMaio, F., Wei, K. Y., Vahey, M. D., Boyken, S. E., Chen, Z., Fallas, J. A., Ueda, G., Sheffler, W., Mulligan, V. K., Xu, W., Bowie, J. U. & Baker, D. Accurate computational design of multipass transmembrane proteins. *Science (80-.).* **359,** 1042–1046 (2018).

61. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* **78,** 2029–40 (2010).

62. Raveh, B., London, N., Zimmerman, L. & Schueler-Furman, O. Rosetta FlexPepDock ab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors. *PLoS One* **6,**

e18934 (2011).

63. Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D. & Schueler-Furman, O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS Comput. Biol.* **13,** e1005905 (2017).
64. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* **65,** 392–406 (2006).
65. London, N., Raveh, B., Movshovitz-Attias, D. & Schueler-Furman, O. Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? *Proteins* **78,** 3140–9 (2010).
66. Watkins, A. M. & Arora, P. S. Anatomy of β-strands at protein-protein interfaces. *ACS Chem. Biol.* **9,** 1747–54 (2014).
67. Jochim, A. L. & Arora, P. S. Systematic analysis of helical protein interfaces reveals targets for synthetic inhibitors. *ACS Chem. Biol.* **5,** 919–23 (2010).
68. Gavenonis, J., Sheneman, B. A., Siegert, T. R., Eshelman, M. R. & Kritzer, J. A. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.* **10,** 716–22 (2014).
69. Sedan, Y., Marcu, O., Lyskov, S. & Schueler-Furman, O. Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Res.* gkw385 (2016). doi:10.1093/nar/gkw385
70. Moretti, R., Lyskov, S., Das, R., Meiler, J. & Gray, J. J. Web-accessible molecular modeling with Rosetta: The Rosetta Online Server that Includes Everyone (ROSIE). *Protein Sci.* **27,** 259–268 (2018).
71. Hosseinzadeh, P., Bhardwaj, G., Mulligan, V. K., Shortridge, M. D., Craven, T. W., Pardo-Avila, F., Rettie, S. A., Kim, D. E., Silva, D.-A., Ibrahim, Y. M., Webb, I. K., Cort, J. R., Adkins, J. N., Varani, G. & Baker, D. Comprehensive computational design of ordered peptide macrocycles. *Science (80-.).* **358,** 1461–1466 (2017).
72. Rubenstein, A. B., Pethe, M. A. & Khare, S. D. MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory. *PLOS Comput. Biol.* **13,** e1005614 (2017).
73. Rohl, C. A., Strauss, C. E. M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55,** 656–77 (2004).
74. Wang, C., Bradley, P. & Baker, D. Protein-Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373,** 503–519 (2007).
75. Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12,** 963–72 (2003).
76. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–2 (2009).
77. Mandell, D. J. & Kortemme, T. Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20,** 420–8 (2009).
78. Pan, X. & Kortemme, T. Manuscript in preparation. (2019).
79. Tyka, M. D., Jung, K. & Baker, D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *J. Comput. Chem.* **33,** 2483–2491 (2012).
80. Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivón, L. G., Davis, A., Oberdorfer, G., Tripsianes, K., Sgourakis, N. G. & Baker, D. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25,** 1028–1034 (2018).
81. Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* **37,** W474-479 (2009).
82. Weitzner, B. D., Kuroda, D., Marze, N., Xu, J. & Gray, J. J. Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins Struct. Funct. Bioinforma.* **82,** 1611–1623 (2014).
83. Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., Adolf-Bryfogle, J., Biswas, N., Dunbrack, R. L. & Gray, J. J. Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* **12,** 401–416 (2017).
84. Sivasubramanian, A., Sircar, A., Chaudhury, S. & Gray, J. J. Toward high-resolution homology modeling of antibody F $_{v}$ regions and application to antibody-antigen docking. *Proteins Struct. Funct. Bioinforma.* **74,** 497–514 (2009).

85. Marze, N. A., Lyskov, S. & Gray, J. J. Improved prediction of antibody V $_L$ –V $_H$ orientation. *Protein Eng. Des. Sel.* **29,** 409–418 (2016).
86. Finn, J. A., Koehler Leman, J., Willis, J. R., Cisneros, A., Crowe, J. E. & Meiler, J. Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints. *PLoS One* **11,** e0154811 (2016).
87. Weitzner, B. D. & Gray, J. J. Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. *J. Immunol.* **198,** 505–515 (2017).
88. Sircar, A., Sanni, K. A., Shi, J. & Gray, J. J. Analysis and modeling of the variable region of camelid single-domain antibodies. *J. Immunol.* **186,** 6357–67 (2011).
89. DeKosky, B. J., Lungu, O. I., Park, D., Johnson, E. L., Charab, W., Chrysostomou, C., Kuroda, D., Ellington, A. D., Ippolito, G. C., Gray, J. J. & Georgiou, G. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl. Acad. Sci.* **113,** E2636–E2645 (2016).
90. Jeliazkov, J. R., Sljoka, A., Kuroda, D., Tsuchimura, N., Katoh, N., Tsumoto, K. & Gray, J. J. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. *Front. Immunol.* **9,** 413 (2018).
91. Norn, C. H., Lapidoth, G. & Fleishman, S. J. High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments. *Proteins* **85,** 30–38 (2017).
92. Lapidoth, G., Parker, J., Prilusky, J. & Fleishman, S. J. AbPredict 2: a server for accurate and unstrained structure prediction of antibody variable domains. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty822
93. Nerli, S. & Sgourakis, N. G. Manuscript in preparation. (2019).
94. Toor, J. S., Rao, A. A., McShan, A. C., Yarmarkovich, M., Nerli, S., Yamaguchi, K., Madejska, A. A., Nguyen, S., Tripathi, S., Maris, J. M., Salama, S. R., Haussler, D. & Sgourakis, N. G. A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. *Front. Immunol.* **9,** 99 (2018).
95. Gowthaman, R. & Pierce, B. G. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res.* **46,** W396–W401 (2018).
96. Tran, E., Robbins, P. F., Lu, Y.-C., Prickett, T. D., Gartner, J. J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E. M., Kriley, I. R. & Rosenberg, S. A. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N. Engl. J. Med.* **375,** 2255–2262 (2016).
97. Pierce, B. G. & Weng, Z. A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Sci.* **22,** 35–46 (2013).
98. Pierce, B. G., Vreven, T. & Weng, Z. Modeling T cell receptor recognition of CD1-lipid and MR1-metabolite complexes. *BMC Bioinformatics* **15,** 319 (2014).
99. Sircar, A. & Gray, J. J. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.* **6,** e1000644 (2010).
100. Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B. D., Hu, X., Adachi, Y., Schief, W. R. & Dunbrack, R. L. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLOS Comput. Biol.* **14,** e1006112 (2018).
101. North, B., Lehmann, A. & Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. *J. Mol. Biol.* **406,** 228–256 (2011).
102. King, C., Garza, E. N., Mazor, R., Linehan, J. L., Pastan, I., Pepper, M. & Baker, D. Removing T-cell epitopes with computational protein design. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 8577–82 (2014).
103. Nivón, L. G., Bjelic, S., King, C. & Baker, D. Automating human intuition for protein design. *Proteins* **82,** 858–66 (2014).
104. Lapidoth, G. D., Baran, D., Pszolla, G. M., Norn, C., Alon, A., Tyka, M. D. & Fleishman, S. J. AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins* **83,** 1385–406 (2015).
105. Baran, D., Pszolla, M. G., Lapidoth, G. D., Norn, C., Dym, O., Unger, T., Albeck, S., Tyka, M. D. & Fleishman, S. J. Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 10900–10905 (2017).
106. Wang, R. Y.-R., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., Baker, D. & DiMaio, F. De novo protein structure determination from near-atomic-resolution cryo-EM maps.

Nat. Methods **12,** 335–8 (2015).

107. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* **14,** 797–800 (2017).
108. DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D. & Baker, D. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10,** 1102–4 (2013).
109. DiMaio, F., Song, Y., Li, X., Brunner, M. J., Xu, C., Conticello, V., Egelman, E., Marlovits, T. C., Cheng, Y. & Baker, D. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12,** 361–5 (2015).
110. Wang, R. Y.-R., Song, Y., Barad, B. A., Cheng, Y., Fraser, J. S. & DiMaio, F. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *Elife* **5,** (2016).
111. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *J. Comput. Chem.* **38,** 276–287 (2017).
112. Frenz, B., Rämisch, S., Borst, A. J., Walls, A. C., Adolf-Bryfogle, J., Schief, W. R., Veesler, D. & DiMaio, F. Automatically Fixing Errors in Glycoprotein Structures with Rosetta. *Structure* **0,** (2018).
113. Nerli, S. & Sgourakis, N. G. CS-ROSETTA. *Methods Enzymol.* (2018). doi:10.1016/BS.MIE.2018.07.005
114. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* **124,** 2723–9 (2002).
115. Schmitz, C., Vernon, R., Otting, G., Baker, D. & Huber, T. Protein Structure Determination from Pseudocontact Shifts Using ROSETTA. *J. Mol. Biol.* **416,** 668–677 (2012).
116. Yagi, H., Pilla, K. B., Maleckis, A., Graham, B., Huber, T. & Otting, G. Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. *Structure* **21,** 883–890 (2013).
117. Pilla, K. B., Otting, G. & Huber, T. Pseudocontact Shift-Driven Iterative Resampling for 3D Structure Determinations of Large Proteins. *J. Mol. Biol.* **428,** 522–532 (2016).
118. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins Struct. Funct. Bioinforma.* **80,** 884–895 (2012).
119. Bowers, P. M., Strauss, C. E. M. & Baker, D. De novo protein structure determination using sparse NMR data. 311–318 (2000).
120. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15404–9 (2003).
121. Raman, S., Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T. A., Eletsky, A., Szyperski, T., Kennedy, M. A., Prestegard, J., Montelione, G. T. & Baker, D. NMR Structure Determination for Larger Proteins Using Backbone-Only Data. **1014,** (2010).
122. Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H.-W., Aramini, J. M., Ertekin, a., Xiao, R., Acton, T. B., Montelione, G. T. & Baker, D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci.* **109,** 10873–10878 (2012).
123. Reichel, K., Fisette, O., Braun, T., Lange, O. F., Hummer, G. & Schäfer, L. V. Systematic evaluation of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins* **85,** 812–826 (2017).
124. Sgourakis, N. G., Lange, O. F., DiMaio, F., André, I., Fitzkee, N. C., Rossi, P., Montelione, G. T., Bax, A. & Baker, D. Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc.* **133,** 6288–98 (2011).
125. Rossi, P., Shi, L., Liu, G., Barbieri, C. M., Lee, H. W., Grant, T. D., Luft, J. R., Xiao, R., Acton, T. B., Snell, E. H., Montelione, G. T., Baker, D., Lange, O. F. & Sgourakis, N. G. A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins Struct. Funct. Bioinforma.* **83,** 309–317 (2015).
126. Demers, J.-P., Habenstein, B., Loquet, A., Kumar Vasa, S., Giller, K., Becker, S., Baker, D., Lange, A. & Sgourakis, N. G. High-resolution structure of the Shigella type-III secretion needle by solid-state NMR and cryo-electron microscopy. *Nat. Commun.* **5,** 4976 (2014).
127. Thompson, J. M., Sgourakis, N. G., Liu, G., Rossi, P., Tang, Y., Mills, J. L., Szyperski, T., Montelione, G. T. & Baker, D. Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 9875–9880 (2012).
128. Braun, T., Koehler Leman, J. & Lange, O. F. Combining Evolutionary Information and an Iterative

doi:10.20944/preprints201904.0263.v1

Sampling Strategy for Accurate Protein Structure Prediction. *PLoS Comput. Biol.* **11,** (2015).

129. Evangelidis, T., Nerli, S., Nováček, J., Brereton, A. E., Karplus, P. A., Dotas, R. R., Venditti, V., Sgourakis, N. G. & Tripsianes, K. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9,** 384 (2018).
130. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59,** 147–159 (2014).
131. Kuenze, G., Bonneau, R., Leman, J. K. & Meiler, J. Integrative protein modeling in RosettaNMR from sparse paramagnetic restraints. *bioRxiv* 597872 (2019). doi:10.1101/597872
132. Aprahamian, M. L., Chea, E. E., Jones, L. M. & Lindert, S. Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data. *Anal. Chem.* **90,** 7721–7729 (2018).
133. Aprahamian, M. L. & Lindert, S. Utility of Covalent Labeling Mass Spectrometry Data in Protein Structure Prediction with Rosetta. *J. Chem. Theory Comput.* acs.jctc.9b00101 (2019). doi:10.1021/acs.jctc.9b00101
134. Hauri, S., Khakzad, H., Happonen, L., Teleman, J., Malmström, J. & Malmström, L. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **10,** 192 (2019).
135. Watkins, A. M., Geniesse, C., Kladwang, W., Zakrevsky, P., Jaeger, L. & Das, R. Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci. Adv.* **4,** eaar5316 (2018).
136. Sripakdeevong, P., Kladwang, W. & Das, R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl. Acad. Sci.* **108,** 20573–20578 (2011).
137. Das, R. Atomic-Accuracy Prediction of Protein Loop Structures through an RNA-Inspired Ansatz. *PLoS One* **8,** e74830 (2013).
138. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7,** 291–294 (2010).
139. Cheng, C. Y., Chou, F.-C. & Das, R. Modeling Complex RNA Tertiary Folds with Rosetta. *Methods Enzymol.* **553,** 35–64 (2015).
140. Chou, F.-C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods* **10,** 74–76 (2013).
141. Chou, F.-C., Echols, N., Terwilliger, T. C. & Das, R. in 269–282 (Humana Press, New York, NY, 2016). doi:10.1007/978-1-4939-2763-0_17
142. Sripakdeevong, P., Cevec, M., Chang, A. T., Erat, M. C., Ziegeler, M., Zhao, Q., Fox, G. E., Gao, X., Kennedy, S. D., Kierzek, R., Nikonowicz, E. P., Schwalbe, H., Sigel, R. K. O., Turner, D. H. & Das, R. Structure determination of noncanonical RNA motifs guided by 1H NMR chemical shifts. *Nat. Methods* **11,** 413–416 (2014).
143. Kappel, K. & Das, R. Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. *Structure* **27,** 140-151.e5 (2019).
144. Kappel, K., Liu, S., Larsen, K. P., Skiniotis, G., Puglisi, E. V., Puglisi, J. D., Zhou, Z. H., Zhao, R. & Das, R. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nat. Methods* **15,** 947–954 (2018).
145. Thyme, S. B., Jarjour, J., Takeuchi, R., Havranek, J. J., Ashworth, J., Scharenberg, A. M., Stoddard, B. L. & Baker, D. Exploitation of binding energy for catalysis and design. *Nature* **461,** 1300–1304 (2009).
146. Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10,** 45–52 (2003).
147. Ashworth, J., Taylor, G. K., Havranek, J. J., Quadri, S. A., Stoddard, B. L. & Baker, D. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38,** 5601–5608 (2010).
148. Thyme, S. B., Boissel, S. J. S., Arshiya Quadri, S., Nolan, T., Baker, D. A., Park, R. U., Kusak, L., Ashworth, J. & Baker, D. Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Res.* **42,** 2564–2576 (2014).
149. Thyme, S. B., Baker, D. & Bradley, P. Improved Modeling of Side-Chain–Base Interactions and Plasticity in Protein–DNA Interface Design. *J. Mol. Biol.* **419,** 255–274 (2012).
150. Ashworth, J. & Baker, D. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* **37,** e73 (2009).
151. Song, Y., Dimaio, F., Wang, R. Y.-R. R., Kim, D. E., Miles, C., Brunette, T., Thompson, J. & Baker,

doi:10.20944/preprints201904.0263.v1

D. High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).

152. Thyme, S. B., Song, Y., Brunette, T. J., Szeto, M. D., Kusak, L., Bradley, P. & Baker, D. Massively parallel determination and modeling of endonuclease substrate specificity. *Nucleic Acids Res.* **42,** 13839–13852 (2014).

153. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5,** 993–6 (2006).

154. Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins Struct. Funct. Bioinforma.* **83,** 1–24 (2015).

155. Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins* **62,** 1010–1025 (2006).

156. Yarov-Yarovoy, V., Decaen, P. G., Westenbroek, R. E., Pan, C.-Y. Y. C.-Y., Scheuer, T., Baker, D. & Catterall, W. a. Structural basis for gating charge movement in the voltage sensor of a sodium channel. *Proc. Natl. Acad. Sci.* **109,** E93–E102 (2012).

157. Barth, P., Schonbrun, J. & Baker, D. Toward high-resolution prediction and design of transmembrane helical protein structures. **2007,** (2007).

158. Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A. & Gray, J. J. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.* **11,** e1004398 (2015).

159. Baugh, E. H., Lyskov, S., Weitzner, B. D. & Gray, J. J. Real-time PyMOL visualization for Rosetta and PyRosetta. *PLoS One* **6,** e21931 (2011).

160. Koehler Leman, J., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics* **11,** 1–3 (2016).

161. Koehler Leman, J., Lyskov, S. & Bonneau, R. Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP. *BMC Bioinformatics* **18,** 115 (2017).

162. Koehler Leman, J. & Bonneau, R. A novel domain assembly routine for creating full-length models of membrane proteins from known domain structures. *Biochemistry* acs.biochem.7b00995 (2017). doi:10.1021/acs.biochem.7b00995

163. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3,** 97–130 (1993).

164. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W. & Etzler, M. E. *Essentials of Glycobiology*. *Essentials Glycobiol.* (Cold Spring Harbor Laboratory Press, 2009).

165. Nivedha, A. K., Thieker, D. F., Makeneni, S., Hu, H. & Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **12,** 892–901 (2016).

166. Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D. & Kuhlman, B. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523,** 109–43 (2013).

167. Jorgensen, W. L., Jorgensen, W. L., Maxwell, D. S. & Tirado-rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. AM. CHEM. SOC* 11225--11236 (1996). at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.334.2959>

168. Radzicka, A. & Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27,** 1664–1670 (1988).

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

170. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23,** 47–55 (2014).

171. Park, H., Lee, H. & Seok, C. High-resolution protein-protein docking by global optimization: recent advances and future challenges. *Curr. Opin. Struct. Biol.* **35,** 24–31 (2015).

172. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.* **79,** 830–838 (2011).

173. Mills, J. H., Khare, S. D., Bolduc, J. M., Forouhar, F., Mulligan, V. K., Lew, S., Seetharaman, J., Tong, L., Stoddard, B. L. & Baker, D. Computational Design of an Unnatural Amino Acid Dependent Metalloprotein with Atomic Level Accuracy. *J. Am. Chem. Soc.* **135,** 13393–13399 (2013).
174. Mulligan, V. K. Manuscript in preparation. (2019).
175. Leaver-Fay, A., Butterfoss, G. L., Snoeyink, J. & Kuhlman, B. Maintaining solvent accessible surface area under rotamer substitution for protein design. *J. Comput. Chem.* **28,** 1336–41 (2007).
176. Rubenstein, A. B., Blacklock, K., Nguyen, H., Case, D. A. & Khare, S. D. Systematic Comparison of Amber and Rosetta Energy Functions for Protein Structure Evaluation. *J. Chem. Theory Comput.* **14,** 6015–6025 (2018).
177. 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* **352,** 680–7 (2016).
178. Chen, Z., Boyken, S. E., Jia, M., Busch, F., Flores-Solis, D., Bick, M. J., Lu, P., VanAernum, Z. L., Sahasrabuddhe, A., Langan, R. A., Bermeo, S., Brunette, T. J., Mulligan, V. K., Carter, L. P., DiMaio, F., Sgourakis, N. G., Wysocki, V. H. & Baker, D. Programmable design of orthogonal protein heterodimers. *Nature* **565,** 106–111 (2019).
179. Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J. Chem. Theory Comput.* **14,** 2751–2760 (2018).
180. Gray, J. J., Chaudhury, S., Lyskov, S., and Labonte, J. W. The PyRosetta Interactive Platform for Protein Structure Prediction and Design: A Set of Educational Modules. (2014). at <http://www.amazon.com/PyRosetta-Interactive-Platform-Structure-Prediction/dp/1500968277>
181. Schenkelberg, C. D. & Bystroff, C. InteractiveROSETTA: A graphical user interface for the PyRosetta protein modeling suite. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv492
182. Kleffner, R., Flatten, J., Leaver-Fay, A., Baker, D., Siegel, J. B., Khatib, F. & Cooper, S. Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics* **33,** 2765–2767 (2017).
183. Cooper, S., Sterling, A. L. R., Kleffner, R., Silversmith, W. M. & Siegel, J. B. Repurposing citizen science games as software tools for professional scientists. in *Proc. 13th Int. Conf. Found. Digit. Games - FDG '18* 1–6 (ACM Press, 2018). doi:10.1145/3235765.3235770
184. Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., Baker, D. & Players, F. Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 18949–53 (2011).
185. Dsilva, L., Mittal, S., Koepnick, B., Flatten, J., Cooper, S. & Horowitz, S. Creating custom Foldit puzzles for teaching biochemistry. *Biochem. Mol. Biol. Educ.* **47,** 133–139 (2019).
186. Lyskov, S., Chou, F.-C., Conchúir, S. Ó., Der, B. S., Drew, K., Kuroda, D., Xu, J., Weitzner, B. D., Renfrew, P. D., Sripakdeevong, P., Borgo, B., Havranek, J. J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J. J. & Das, R. Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PLoS One* **8,** e63906 (2013).