Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

Macromolecular modeling and design in Rosetta: new methods and frameworks

Koehler Leman, Julia* [1, 2], Weitzner, Brian D [3, 4, 5, 6], Lewis, Steven M [7, 8, 9], RosettaCommons consortium, Bonneau, Richard* [1, 2, 61, 62]

RosettaCommons consortium:
Adolf-Bryfogle, Jared [10], Alam, Nawsad [11], Alford, Rebecca F [3], Aprahamian, Melanie [12], Baker, David [4, 5], Barlow, Kyle A [13], Barth, Patrick [14, 15], Basanta, Benjamin [4, 16], Bender, Brian J [17], Blacklock, Kristin [18], Bonet, Jaume [14, 19], Boyken, Scott [5, 6], Bradley, Phil [20], Bystroff, Chris [21], Conway, Patrick [4], Cooper, Seth [22], Correia, Bruno E [14, 19], Coventry, Brian [4], Das, Rhiju [23], De Jong, René M [24], DiMaio, Frank [4, 5], Dsilva, Lorna [22], Dunbrack, Roland [25], Ford, Alex [4], Frenz, Brandon [9], Fu, Darwin Y [26], Geniesse, Caleb [23], Goldschmidt, Lukasz [4], Gowthaman, Ragul [27, 28], Gray, Jeffrey J [3], Gront, Dominik [29], Guffy, Sharon [7], Horowitz, Scott [30, 31], Huang, Po-Ssu [4], Huber, Thomas [32], Jacobs, Tim M [33], Jeliazkov, Jeliazko R [34], Johnson, David K [35], Kappel, Kalli [36], Karanicolas, John [25], Khakzad, Hamed [19, 37, 38], Khar, Karen R [35], Khare, Sagar D [18, 39, 53, 54, 55], Khatib, Firas [40], Khramushin, Alisa [13], King, Indigo C [4, 9], Kleffner, Robert [22], Koepnick, Brian [4], Kortemme, Tanja [41], Kuenze, Georg [26, 42], Kuhlman, Brian [7], Kuroda, Daisuke [43, 44], Labonte, Jason W [3, 45], Lai, Jason K [15], Lapidoth, Gideon [46], Leaver-Fay, Andrew [7], Lindert, Steffen [12], Linsky, Thomas [4, 5], London, Nir [11], Lubin, Joseph H [3], Lyskov, Sergey [3], Maguire, Jack [33], Malmström, Lars [19, 37, 38, 47], Marcos, Enrique [4, 48], Marcu, Orly [11], Marze, Nicholas A [3], Meiler, Jens [42, 49, 50], Moretti, Rocco [26], Mulligan, Vikram Khipple [1, 4, 5], Nerli, Santrupti [51], Norn, Christoffer [46], Ó'Conchúir, Shane [41], Ollikainen, Noah [41], Ovchinnikov, Sergey [4, 5, 52], Pacella, Michael S [3], Pan, Xingjie [41], Park, Hahnbeom [4], Pavlovicz, Ryan E [4, 5], Pethe, Manasi [53, 54], Pierce, Brian G [27, 28], Pilla, Kala Bharath [32], Raveh, Barak [11], Renfrew, P Douglas [1], Roy Burman, Shourya S [3], Rubenstein, Aliza [18, 55], Sauer, Marion F [56], Scheck, Andreas [14, 19], Schief, William [10], Schueler-Furman, Ora [11], Sedan, Yuval [11], Sevy, Alexander M [56], Sgourakis, Nikolaos G [57], Shi, Lei [4, 5], Siegel, Justin [58, 59, 60], Silva, Daniel-Adriano [4], Smith, Shannon [26], Song, Yifan [4, 5], Stein, Amelie [41], Szegedy, Maria [39], Teets, Frank D [7], Thyme, Summer B [4], Wang, Ray Yu-Ruei [4], Watkins, Andrew [23], Zimmerman, Lior [11],

==========

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 Lyell Immunopharma Inc., Seattle, WA 98109
7 Dept of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
8 Dept of Biochemistry, Duke University, Durham, NC 27710, USA
9 Cyrus Biotechnology, Seattle, WA 98101, USA
10 Dept of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA
11 Dept of Microbiology and Molecular Genetics, IMRIC, Ein Kerem Faculty of Medicine, Hebrew University of Jerusalem, 91120, Jerusalem, Israel
12 Dept of Chemistry and Biochemistry, Ohio State University, Columbus, OH, 43210, USA
13 Graduate Program in Bioinformatics, University of California San Francisco, CA 94158, USA
14 Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
15 Baylor College of Medicine, Department of Pharmacology, Houston, TX 77030, USA
16 Biological Physics Structure and Design PhD Program, University of Washington, Seattle, WA 98195, USA
17 Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, USA
18 Institute of Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
19 Swiss Institute of Bioinformatics, Lausanne, Switzerland

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

20 Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
21 Dept of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA
22 Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
23 Dept of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA
24 DSM Biotechnology Center, 2613 AX Delft, The Netherlands
25 Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia PA 19111, USA
26 Dept of Chemistry, Vanderbilt University, Nashville, TN 37232, USA
27 University of Maryland Institute for Bioscience and Biotechnology Research, Rockville, MD 20850, USA
28 Dept of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA
29 Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw
Żwirki i Wigury 101, 02-089 Warsaw
30 Dept of Chemistry & Biochemistry, University of Denver, Denver, CO 80208, USA
31 The Knoebel Institute for Healthy Aging, University of Denver, Denver, CO 80208, USA
32 Research School of Chemistry, Australian National University, Canberra ACT 2601, Australia
33 Program in Bioinformatics and Computational Biology, Dept of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
34 Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA
35 Center for Computational Biology, University of Kansas, Lawrence, KS 66047, USA
36 Biophysics Program, Stanford University, Stanford CA 94305, USA
37 Institute for Computational Science, University of Zurich, CH-8057 Zurich, Switzerland
38 S3IT, University of Zurich, CH-8057 Zurich, Switzerland
39 Dept of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
40 Dept of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA
41 Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, CA 94158, USA
42 Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA
43 Medical Device Development and Regulation Research Center, School of Engineering, University of Tokyo, Tokyo 113-8656, Japan
44 Dept of Bioengineering, School of Engineering, University of Tokyo, Tokyo 113-8656, Japan
45 Dept of Chemistry, Franklin & Marshall College, Lancaster, PA 17604, USA
46 Dept of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, 76100, Israel
47 Division of Infection Medicine, Dept of Clinical Sciences Lund, Faculty of Medicine, Lund University, SE-22184, Lund, Sweden
48 Institute for Research in Biomedicine Barcelona, The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
49 Depts of Chemistry, Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA
50 Institute for Chemical Biology, Vanderbilt University, Nashville, TN 37232, USA
51 Dept of Computer Science, University of California Santa Cruz, CA, USA
52 Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA
53 Dept of Chemistry and Chemical Biology, The State University of New Jersey, Piscataway, NJ 08854, USA
54 Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
55 Computational Biology and Molecular Biophysics Program, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
56 Chemical and Physical Biology Program, Vanderbilt Vaccine Center, Vanderbilt University, Nashville, TN 37235, USA
57 Dept of Chemistry and Biochemistry, University of California Santa Cruz, CA, USA
58 Dept of Chemistry, University of California, Davis, CA, USA
59 Dept of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA
60 Genome Center, University of California, Davis, CA, USA
61 Dept of Computer Science, New York University, New York, 10003, NY, USA

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

62 Center for Data Science, New York University, New York, 10003, NY, USA

* Corresponding authors: Richard Bonneau (Bonneau@nyu.edu) & Julia Koehler Leman (Julia.koehler.leman@nyu.edu)

Abstract
The Rosetta software suite for macromolecular modeling, docking, and design is widely used in pharmaceutical, industrial, academic, non-profit, and government laboratories. Rosetta's advantage is interoperability between several broad modeling capabilities and it consistently ranks highly when compared to other leading methods created for highly specialized protein modeling and design tasks. Developed for over two decades by a global community of scientists at more than 60 institutions, Rosetta has been refactored and extended continuously and now comprises over three million lines of code. Here we discuss methods and applications developed in the last five years, including 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, and modeling of peptides, antibodies and other proteins in the immune system, nucleic acids, non-standard amino acids, carbohydrates, and membrane proteins. We briefly discuss improvements to the score function, user interfaces, and usability of the software. Rosetta is available at www.rosettacommons.org.

Introduction
It has long been understood that, in biological systems, structure determines function. This relationship has motivated decades of experimental determination of protein structure and function. Many computational packages have been developed to provide valuable guidance to experimental methods, one of which is the Rosetta modeling and design suite. Most computational tools are specialized for a small number of specific purposes; in this regard Rosetta is different, and over two decades has been expanded to include broad capabilities that span many bioinformatics and structural-bioinformatics tasks. Computational structural biology tools and frameworks with similar comprehensive scope are few, but key to progress in biology. Schrodinger[1], the Molecular Operating Environment[2,3], and Discovery Studio[4] are computational chemistry platforms for advanced modeling and design for structural biology, drug discovery and material science, based on molecular mechanics, molecular dynamics and quantum mechanics calculations. The HHSuite[5] includes various tools for bioinformatics, sequence alignments, structure prediction and modeling. The BioChemicalLibrary[6] (BCL) includes tools structure prediction, drug discovery, and several sequence-to-structure methods using machine learning approaches. The Integrative Modeling Platform[7] (IMP) allows modeling of large macromolecular complexes by incorporation of various types of experimental data. OpenBabel[8] is a ChemInformatics toolbox supporting molecular mechanics calculations but is most heavily used for interconversion of file formats.

Molecular dynamics simulation packages like CHARMM[9], AMBER[10], GROMACS[11], OPLS[12], Desmond[13], and FoldX[14] simulate most atoms explicitly with a physics-based energy function that relies on solving Newton's equation of motion. These methods can be used for folding small proteins, model refinement, high-resolution phenomena such as ion flow through membrane channels, and modeling interactions with small molecules and are therefore highly complementary to Rosetta. OpenMM[15] is an API (application programming interface) for setting up molecular simulations and can be used as a library or standalone application.

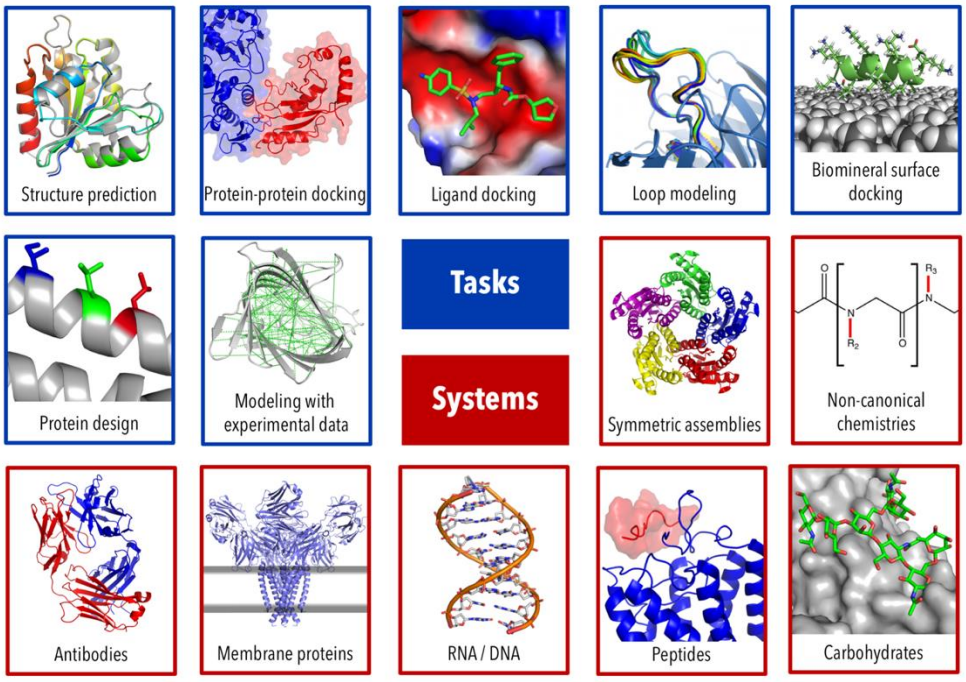
Many other tools are available for more specialized tasks, for instance for *de novo* modeling (AlphaFold[16], QUARK[17], RaptorX[18], MULTICOM[19]), homology modeling (Modeller[20], SwissModel[21]), fold recognition (iTasser[22]), protein-protein docking (HADDOCK[23], Zdock[24], ClusPro[25], Gramm-X[26], PatchDock[27]), ligand docking (AutoDock[28], FlexX[29], Glide[30]) and the numerous other tasks that require molecular modeling. A large body of methods other than Rosetta is listed and cited in the Supplement to this paper, as the focus of this perspective is the description of methods recently developed in Rosetta. Given that it is well beyond the scope of this work to give a fully comprehensive picture of computational chemistry or biomolecular modeling software, we focus this perspective on Rosetta, which a proven state-of-the-art code for a wide variety of bio-macromolecular prediction and design tasks. One of Rosetta's advantages is the inter-operability of its large number of applications. The challenge however is to track the scope of functionality

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

available to scientists who wish to use the software. This perspective is meant to be an illustrated guide to the new, returning, or seasoned user; to help find the right protocol hiding in the Rosetta haystack. We present the list of tools that were developed in the past five years, therefore providing direction as to where to find information for specific modeling problems. For detailed instructions for each of the specific protocols, we refer to the links in the Supplement which direct to the appropriate corners of the expansive Rosetta documentation and codebase.

Development of Rosetta started in the mid-1990s for protein structure prediction and to gain insights into the protein folding problem[31], which remains a grand challenge of 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 to incorporating NMR data, loop modeling, protein design, and interaction with peptides and nucleic acids (Figure 1). Over more than 20 years, the community of developers and scientists, the RosettaCommons, grew from a single academic laboratory to laboratories at over 60 institutions around the globe[32]. The software has undergone several transitions, including in programming language and implementation, with the latest protocols based on Rosetta3, which was first released in 2008[33]. The score function has been continuously improved, detailed descriptions of which can be found in previous articles [34] and [35]. Throughout Rosetta's lifetime, efforts to improve interfaces to the code, and documentation have drastically improved usability and enable modular application to new problems. As part of our sustained focus on accessibility, usability, and scientific reproducibility, we developed several interfaces, (PyRosetta[36], RosettaScripts[37], Foldit[38]) and emphasized publishing protocol captures[39] that accompany manuscripts. As the software's interfaces have grown more versatile, development has accelerated and branched in many directions. However, this extensibility and the very large number of scientists that combine modules in unthought combinations make 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, we have compiled the latest method developments in the Rosetta software suite from the past five years, divided into several modeling categories. The supplement contains a more detailed tour of the protocols discussed with extensive links to documentation, resources on the web, and limitations and competitors to each method.

Figure 1: Capabilities of the Rosetta macromolecular modeling suite
Some popular tasks that can be addressed in Rosetta (blue) and major systems that can be modeled (red). This is an incomplete list of Rosetta's broad modeling capabilities.



Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

1. General overview and challenges

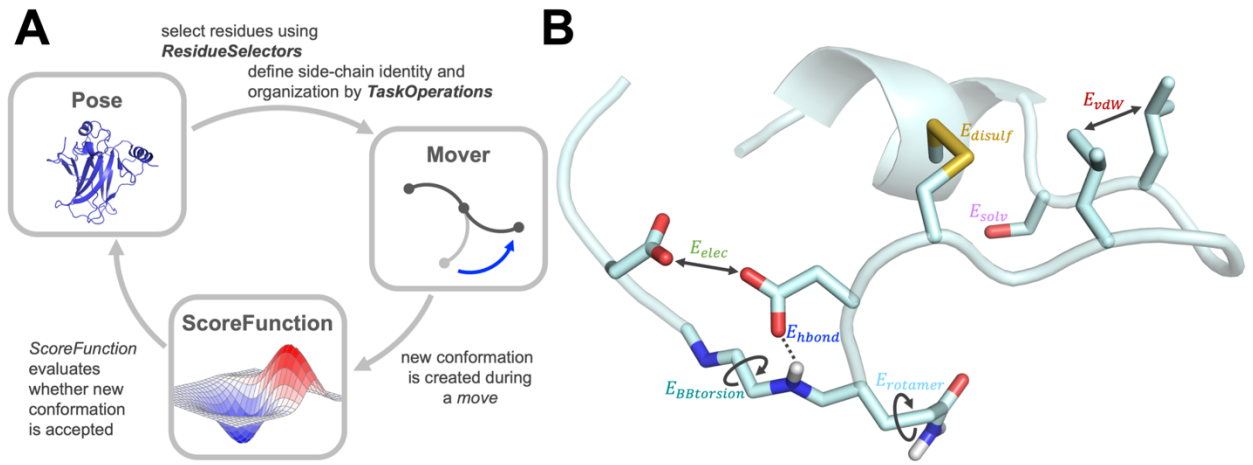
The general outline of a typical Rosetta protocol is depicted in Figure 2A: the conformation of a biomolecule (the *Pose*) is altered, either deterministically or stochastically, via a *Mover* and the resulting conformation is evaluated by a *ScoreFunction*. The *Move* is accepted based on the Metropolis criterion and the energy difference between the original and the new conformation:

$$\begin{aligned} &\text{if } E_{new} < E_{orig} && \text{accept} \\ &\text{if } E_{new} \geq E_{orig} && \text{accept with probability } P = e^{-((E_{new}-E_{orig})/T)} \end{aligned}$$

Many trajectories are generated, and the final models are evaluated based on the scientific objective. This setup highlights common limitations in Rosetta protocols involving sampling, scoring (discussed in the score function section), or technical challenges. Many protocols suffer from under-sampling[40], especially when flexibility is involved. Sampling is a limitation for structure prediction, especially for large structures; protein design; unconstrained global protein-protein docking which leads to success in only about 30% of the cases; local docking is limited by backbone flexibility and deteriorates with larger flexibility in the binding interface; small molecule docking relies on correct identification of the binding interface and is limited by flexibility between unbound and bound states, and loop and antibody modeling suffer from sampling challenges, especially for loops longer than 12 residues. Huge conformational search spaces are also prohibitive for RNA modeling due to the size of their torsion space (see RNA section), membrane proteins due to their size, and carbohydrates because of branching and flexibility.

Figure 2: Main elements of Rosetta are scoring and sampling
(A) 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. (B) The score function consists of a weighted linear combination of various score terms, highlighted in the figure and described above.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2



E_{vdW}	Lennard-Jones for attractive / repulsive interaction	E_{solv}	implicit solvation model penalizes buried polar atoms
E_{hbond}	hydrogen bonding allows buried polar atoms	$E_{BBtorsion}$	backbone torsion preferences from mainchain potential
E_{elec}	electrostatic interaction between charges	$E_{rotamer}$	sidechain torsion angles from rotamer library
E_{disulf}	disulfide bonds between cysteines	E_{ref}	unfolded state reference energy for design

Some Rosetta applications suffer from technical challenges in implementation, for instance a unified framework for various types of experimental data is lacking (see Supplement), code usability revealed by lack of documentation, protocol captures, or support (e.g. DNA modeling, new *de novo* structure prediction protocols), and a need for implementation of more diverse chemistries, for instance for specific carbohydrates, spin labels, non-canonical chemistries, and lipids. Technical challenges are either historical or due to lack of interest in the community to develop and advance methods in these unique areas.

2. Rosetta's brain: its score function

Rosetta's score function has been continuously improved over many years[41] with guiding principles including: improving speed of computation, increasing extensibility, and improving accuracy across multiple tasks. The main score function is a linear combination of weighted score terms that balances physics-based or statistically derived potentials describing *van der Waals* energies, hydrogen bonds, electrostatics, disulfide bonds, residue solvation, backbone torsion angles, sidechain rotamer energies, and an average unfolded state reference energy (Figure 2B):

$$E = E_{vdW} + E_{hbond} + E_{elec} + E_{disulf} + E_{solv} + E_{BBtorsion} + E_{rotamer} + E_{ref}$$

Some energy terms are decomposed into several components to be able to parameterize each of them separately. For instance, the *van der Waals* energy is split into attractive and repulsive terms between different residues, in addition to an intra-residue repulsive term. A complete account of the all-atom score function was published recently[34].

The newest score function REF2015[35] features two main advancements. First, reproducibility of thermodynamic observables (such as liquid-phase properties[12] and liquid-to-vapor transfer free energies[42]) was added to the optimization objectives, in addition to structure[43]-based tests. Second, a new, derivative-free optimization technique was developed, which is suitable for robust optimization of >100 parameters. Further, a new energy term was added that takes into consideration non-ideality of bond lengths and angles in cartesian space[44]. The cartesian term[44] is also the basis for a *cartesian_ddG* method that has been used to calculate ΔΔGs of mutation to probe changes in protein stability. Only the backbones and side chains of residues nearby the mutation site are allowed to move[45]. Due to the local optimization, this protocol is much faster than *ddg_monomer*[46], while retaining the same level of accuracy. The default Rosetta score function is now also compatible with an expanded palette of chemical building-blocks: canonical and non-canonical

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

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[47,48]. As noted above, score functions that enable simultaneous modeling of protein and RNA are being explored[49]. The score function 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[50,51], 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[52].

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[53] has been used to design *de novo* coiled coils with interaction specificity mediated by designed hydrogen bond networks, including homo-oligomers[53], membrane proteins[54], and large sets of orthogonal heterodimers[55]. An improvement to HBNet uses a Monte Carlo search procedure to sample hydrogen bond networks with drastically improved performance[56]. 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[57]. The potential is accessible through the WaterBoxMover (or ExplicitWaterMover) in RosettaScripts.

There are several limitations to the score function: (1) it does not directly model entropy[58], which has been shown to improve sampling efficiency[59]. However, rotamer bond angles, solvation, fragments and pair terms all implicitly model free energy, which at these temperatures and solvation densities account for more than half of the entropy. (2) In most cases, knowledge-based score terms are derived from high-resolution crystal structures, which represent a single state on the energy landscape measured with a specific experimental method, does not represent flexibility well and is in a solid-state environment (compared to solution NMR); (3) knowledge-based terms are less comprehensible than physics-based terms; (4) different score functions are required for different applications, which shouldn't be the case as nature has a single energy function, indicating that the score functions are only approximations of the truth; (5) scoring correlates with the number of score terms and scoring has become slower, yet more accurate, over time; (6) the solvation model is implicit, which is fast, but hinders explicit modeling of ions, water molecules, or lipid environments accurately; (7) score functions for specific applications such as for RNA, membrane proteins, carbohydrates, non-canonicals, or lipids are immature compared to the score functions for 'mainstream' applications in Rosetta.

3. Major applications

Predicting protein structures

Rosetta was originally developed for *de novo* protein structure prediction, which is accomplished by assembling fragments from known protein structures *via* a Monte Carlo procedure and evaluating the models with the score function. While the community's main objective has moved to protein and bio-macromolecular design over the past decade, performance in the CASP blind prediction challenge remains respectable[60], with ranking for refinement and prediction of multimeric complexes among the top three groups. Meanwhile, many groups have developed specialized tools exploiting evolutionary couplings and machine learning methods, for instance Google's DeepMind developed AlphaFold[16] with outstanding performance in the recent CASP12[61]. Other highly ranking methods are iTasser[17] (Yang Zhang), MULTICOM[19] (Jianglin Cheng), and QUARK[17] (Yang Zhang).

Improvements in homology modeling were achieved by multi-template modeling in RosettaCM[62] (now available on the new Robetta[60,63] server), which hybridizes the most homologous portions from multiple templates into a single model while modeling missing residues *de novo*[64]. 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[65]) has led to enormous improvements in model quality. To sample the conformational space further, 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

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

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 score function 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[66] as well as homology models[67]. Model refinement or generating ensembles of structures (useful in particular for design) can be accomplished by several algorithms in Rosetta: *FastRelax*[68], *Backrub*[69], or using vicinity sampling in the KIC/Next-Generation-KIC loop modeling algorithms[70,71].

Loop modeling[72] was implemented early in Rosetta[73,74] to close gaps in models or sample loop conformations, with initial approaches relying on fragments sampling and iterative Cyclic Coordinate Descent (CCD)[75] for chain closure. Subsequent developments introduced inverse kinematic closure (termed “KIC”), relying on polynomial resultants to analytically solve for closed conformations, producing more native-like loops[76,77]. Next-Generation KIC (NGK)[71] made improvements to sampling by employing diversification (i.e. wider range of conformations) and intensification (i.e. focus around previously generated conformations), substantially increasing the fraction of near-native models[71] and allowing modeling of longer loops. GeneralizedKIC[50] (GenKIC) samples loop geometries between fixed endpoints including non-standard peptide chemistries, for instance L- and D-α-amino acids, β-amino acids, peptoids, oligoureas, or side-chain connections, covalently-attached ligands or crosslinkers, or chemistries that conventional loop-modelling algorithms do not typically handle.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

Method	Lab developed
Score function	
REF2015 score function[34,35]	Frank DiMaio, David Baker
cartesian_ddG[35]	Frank DiMaio, Phil Bradley
HBNet[53,56]	David Baker, Brian Kuhlman
HBNetEnergy[53]	Richard Bonneau, David Baker*
AACompositionEnergy	Richard Bonneau, David Baker*
AARepeatEnergy	Richard Bonneau, David Baker*
VoidsPenaltyEnergy	Richard Bonneau, David Baker*
NetChargeEnergy	Richard Bonneau, David Baker*
BuriedUnsatPenalty	Richard Bonneau, David Baker*
Protein structure prediction	
fragment picker[78]	Dominik Gront*,**
RosettaCM[62]	David Baker
iterative hybridize[66,67]	David Baker, Sergey Ovchinnikov*
Loop modeling	
NGK (next-generation KIC) [71]	Tanja Kortemme
GenKIC (generalized KIC) [50]	Richard Bonneau, David Baker*
LoopHashKIC	Tanja Kortemme
Consensus_Loop_Design[79,80]	David Baker
Protein-protein docking	
RosettaDock4.0[81]	Jeffrey Gray
Rosetta SymDock2[82]	(Ingemar André), Jeffrey Gray
Small molecule ligand docking	
RosettaLigand[83–85]	Jens Meiler
RosettaLigandEnsemble[86]	Jens Meiler
pocket optimization[87,88]	John Karanicolas
DARC[89–91]	John Karanicolas
Modeling of antibodies and immune system proteins	
RosettaAntibody[92–95]	Jeffrey Gray
AbPredict[96,97]	Sarel Fleishman
RosettaMHC[98]	Nik Sgourakis
TCRModel[99]	Brian Pierce
SnugDock[100]	Jeffrey Gray
Design of antibodies and immune system proteins	
RAbD[101] (Rosetta AntibodyDesign)	Bill Schief, Roland Dunbrack
Epitope removal[102,103]	David Baker, Cyrus Biotechnology
AbDesign[104,105]	Sarel Fleishman
Protein design	
SEWING[106,107]	Brian Kuhlmann
RosettaRemodel[108]	Possu Huang*,**
LooDo[109]	Sagar Khare
RECON[110]	Jens Meiler
curved β-sheet design[79]	David Baker
biased forward folding[79]	David Baker
fold_from_loops[111]	Bruno Correia*,**
FunFolDes[112]	Bruno Correia
Protein interface design	
FlexDDG[113]	Tanja Kortemme
Coupled Moves[114]	Tanja Kortemme & DSM Biotechnology Center
Parametric design[54,115]	Richard Bonneau*
Peptides and peptidomimetics	
FlexPepDock[116,117]	Ora Schueler-Furman
PIPER-FlexPepDock[118]	Ora Schueler-Furman
PeptiDerive[119]	Ora Schueler-Furman
Method	**Lab developed**
simple_cycpep_predict[50,51,115]	Richard Bonneau, David Baker*
MFPred[120]	Sagar Khare
RosettaSurface[121–123]	Jeffrey Gray
Modeling with experimental data	
cryoEM *de novo*[124]	Frank DiMaio, David Baker
cryoEM: RosettaES[125]	Frank DiMaio
cryoEM: iterative refinement[126,127]	(formerly David Baker) Frank DiMaio
cryoEM: automated refinement[128]	Frank DiMaio
NMR: CS-Rosetta[129]	Nik Sgourakis
NMR: PCS-Rosetta, GPS-Rosetta[130,131]	Thomas Huber
RosettaNMR framework[132]: using RDC/PRE/PCS/NOE/CS for ab initio, protein-protein docking, ligand docking, symmetric assembly	Jens Meiler, Richard Bonneau (Jeffrey Gray)
mass-spec: HRF hydroxyl radical footprinting[133,134]	Steffen Lindert
mass-spec: PyTXMS[135]	Lars Malmstroem
RNA modeling	
SWA (stepwise assembly) [136,137]	Rhiju Das
SWM (stepwise Monte-Carlo) [138]	Rhiju Das
FARFAR (fragment assembly medium resolution structure prediction) [139–141]	Rhiju Das
ERRASER (refinement into EM density maps) [142,143]	Rhiju Das
CS-Rosetta-RNA (modeling with NMR data) [144]	Rhiju Das
RECCES (Reweighting of Energy-function Collection with Conformational Ensemble Sampling)	Rhiju Das
DRRAFTER (*de novo* modeling of protein-RNA complexes into EM densities) [145]	Rhiju Das
Membrane proteins	
RosettaMP framework[146]: mp_ddg, mp_dock, mp_relax, mp_symdock	Jeffrey Gray, Richard Bonneau
RosettaMP toolkit[147]: mp_score, mp_transform, mp_mutate_relax, helix_from_sequence	Jeffrey Gray, Richard Bonneau
mp_lipid_acc[148]	Richard Bonneau
mp_domain_assembly[149]	Richard Bonneau
RosettaCM for membrane proteins[39]	Jens Meiler
Carbohydrates	
RosettaCarbohydrate framework[150,151]	Jeffrey Gray, William Schief
User interfaces	
PyRosetta[36,152]	Jeffrey Gray
RosettaScripts[37,39]	Sarel Fleishman*,**
InteractiveRosetta[153]	Chris Bystroff
Foldit Standalone[38,154–156]	Seth Cooper*,**, Firas Khatib*,**, Justin Siegel, Scott Horowitz, David Baker
ROSIE server[157,158]	Jeffrey Gray
Miscellaneous	
Metalloproteins[47,48]	David Baker, Richard Bonneau*
Waters[57]	Frank DiMaio
SimpleMetrics	William Schief
AmbRose	Sagar Khare
RosettaRC	William Schief

* the main developer(s) in this lab was/were formerly in the lab of David Baker when this application was developed
** the main developer now has his/her own lab

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

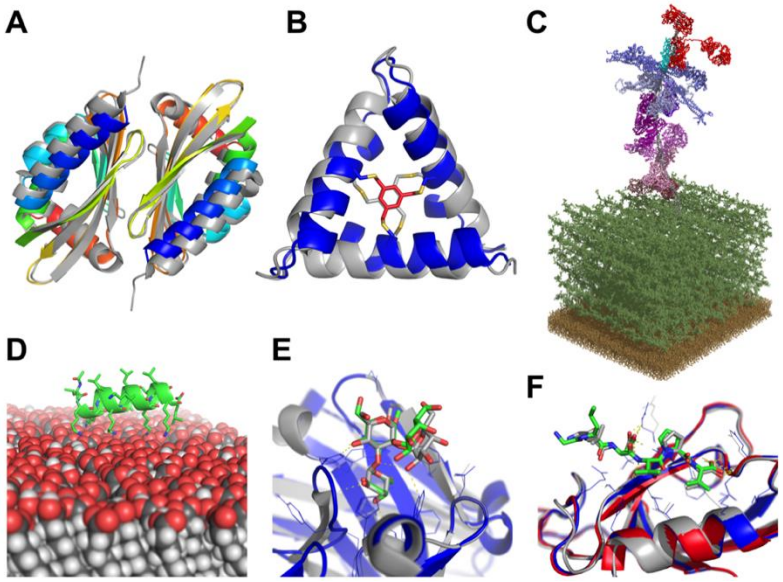
Modeling protein–protein complexes

Another early method was RosettaDock, which predicts the structure of protein-protein complexes from input monomers. The most recent iteration, RosettaDock4.0[81] incorporates protein flexibility from pre-generated protein ensembles, mimicking conformer selection. This has improved sampling efficiency by automatically adjusting the sampling rate based on the diversity of the input ensembles. Scoring has 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 Protein Data Bank (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[82] 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 our symmetry framework[159]. 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 3: Rosetta can successfully address diverse biological questions

(A) Curved β-sheet design: 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 designed *de novo* and features a curved β-sheet, a large pocket, and a homodimer interface[79]. (B) Parametric design: overlay of the *de novo* designed macrocycle 3H1 in blue and the NMR structure in gray (PDBID 5v2g). This “CovCore” (covalent core) miniprotein is held together covalently by a hydrophobic cross-linker at its core (in red for the design and gray for the NMR structure)[115]. (C) PyTXMS: the interactome of M1 protein (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 in a complex mixture of intact bacteria and human plasma by PyTXMS. All models are provided by Rosetta: 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 contains over 200 chemical cross-links[135]. (D) RosettaSurface: 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 in solution and assumes helical structure[122] when on the surface, as experiments show. (E) RosettaCarbohydrate: flexible docking of a carbohydrate antigen to an antibody. The crystal structure is in gray (PDBID 1mfa) and the model in blue, with the carbohydrate in green. Antibody coordinates were taken from the PDB and glycan coordinates started from a randomized backbone conformation and rigid-body orientation[150]. (F) PIPER-FlexPepDock: high-resolution model of a peptide-protein complex generated using PIPER-FlexPepDock (model: blue; solved structure in gray, PDBID 1mfg). The model was generated 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)[118].

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2



Docking of small molecule ligands into proteins

Structure-based drug design has become a common approach for drug optimization due to increasing numbers of deposited structures in the PDB. RosettaLigand[83] has demonstrated success in predicting small molecule-protein interactions.

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[160]. RosettaLigandEnsemble[86] 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's inhibitor-bound conformation often differs from the unbound or protein-protein bound conformation. The pocket optimization approach 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[87,88]. The specific conformations sampled through this approach match "druggable" alternate conformations observed in ligand-bound structures[87,88], implying that these states are 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[161].

Modeling antibodies and other immune system proteins

Due to the therapeutic significance of antibodies, several protocols have been developed 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. RosettaAntibody[92–95] is a protocol for homology modeling of antibodies[95]. It identifies homologous templates, assembles them into a single structure and then models CDR H3 loops *de novo* while simultaneously refining the VH-VL orientation[162]. Recent advances have focused on using multiple templates[162], incorporating a key structural constraints[163,164] into CDR H3 modeling, modeling camelid antibodies[94] and antibodies on the scale of the human repertoire[165,166]. AbPredict[96] predicts antibody structures without homologous templates. Instead, it samples backbone fragments and rigid-body orientations from known antibody structures, without relying on sequence homology, therefore being able to accurately model cases with sequence identity as low as 10%.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

AbPredict2 is available as a webserver[97]. SnugDock[100] is a method for antibody-antigen docking. SnugDock takes as input a plausible 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. Recent advances include a CDR H3 structural constraint[163,164] and docking of camelid antibodies[167]. Limitations in antibody modeling depend on the task: docking is limited by the knowledge of the binding site (global vs. local docking); structure prediction, design and refinement are limited by protein flexibility, and modeling of CDRs or other loops is challenging if they are longer than 12 to 15 residues.

Design of antibodies and immune system proteins

RosettaAntibodyDesign[101] (RAbD) is based on RosettaAntibody[94] (see below) and allows 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. *RAbD* uses North-Dunbrack CDR clustering[168], reducing deleterious sequence mutations, and was benchmarked on 60 diverse antibody-antigen interfaces from both λ and κ antibodies. Experimental benchmarking of two antibody-antigen complexes showed affinity improvements between 10 and 50-fold.

Rosetta has been integrated with experimental immunogenic epitope data, MHC epitope prediction tools, and host genomic data to enable the design of proteins with reduced immunogenicity while retaining function and stability[102]. The approach implements 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 features design by cutting experimentally determined antibody structures along conserved positions to create interchangeable segments and then recombining them to produce a conformationally diverse set of novel antibody models[104,105]. The models are docked to a target of interest, either locally to a specific epitope, or globally, followed by an optimization step comprised of rigorous backbone sampling and sequence design for improving model stability and binding affinity.

Designing new proteins and functions

Protein design[169] (where the objective is identification of a sequence that best represents a given structure) relies on several of the same core functionalities needed for protein prediction, and synergy and interoperability between design and prediction models has always been a core Rosetta design principle. That this circular dependence can be achieved is highlighted by the recently implemented biased forward folding method: During computational *de novo* protein design[170], 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[79] uses three (instead of the typical 200) fragments per residue position. Fragments are chosen based on the RMSD to the native structure in design.

Protein design is somewhat easier when starting from known starting structures and when redesigning for thermostability or features like the protein surface[171]. This is more successful because much information about sequence-structure relationships is readily available in public databases. Most difficult problems are *de novo* design (without a template structure) and design for novel folds or functions. Successes in these cases are sparse and require sampling of enormous conformational spaces, depending on the protein size, several 100s of thousands to millions of models. Another simplification of *de novo* design is thermostabilization of the protein, essentially creating rigid structures that are mostly non-functional, by expanding the energy gap between folded and unfolded designs to facilitate structural characterization. To date, novel functional designs mostly exploit known structures and the next frontier is the design of novel functions onto *de novo* scaffolds. Moreover, nature typically does not design for the global minimum energy conformation (in terms of stability) because proteins require flexibility to carry out their functions.

De novo design and design of novel protein functions towards therapeutic intervention is addressed by various methods in Rosetta: The SEWING protocol creates *de novo* designs by recombining parts of protein

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

structures from randomly-selected helical building blocks[106]. 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[107]. A somewhat similar algorithm has been implemented for antibody design (AbDesign, see below), which was generalized for enzyme design[172]. A more general approach is RosettaRemodel, which performs protein design by rebuilding parts or all of the structure[108] from fragments of known proteins structures. RosettaRemodel 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.

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[173] (MSD) approach evaluates strengths and weaknesses of a single sequence on multiple backbones, for instance binding to one but not another protein partner. REstrained CONvergence[110] (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 achieves on average 70% sequence recovery (a 30% increase compared to MSD!) for large multi-state design problems, such as antibody affinity maturation or prediction of evolutionary sequence profiles of flexible backbones[174,175].

Design of protein function can be accomplished by *motif grafting*, i.e. grafting a known motif or predicted active- or binding-site from a template structure onto a new protein. This approach has been used for antibodies and vaccine design[111] 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 protocol 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[112]. Performance of the folding stage can be improved by selecting fragments according to the target topology via the *StructFragmentMover*.

Designing interfaces between proteins and interaction partners

Problems related to protein design include designing interfaces of proteins with their interaction partners such as proteins or small molecule ligands and predicting ΔΔGs of mutation (e.g. alanine scanning). Predicting ΔΔGs of mutations for protein stability or protein-protein interactions is a difficult problem with low correlation coefficients (0.5-0.7)[176], 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" score function without modeling backbone flexibility[177,178] have typically outperformed methods that allow protein flexibility and use hard-repulsive score functions[179]. FlexDDG[113] was created to improve protein-protein interface ΔΔG predictions and generalize them to residues other than Ala. The protocol creates conformational ensembles using backrub sampling[180], then repacks sidechains, minimizes torsions and computes change in protein-protein interaction ΔΔG by averaging across the ensembles. On 1240 interface mutants, FlexDDG outperforms the earlier *ddg_monomer* application, which was originally created and validated to predict changes in stability upon mutation, not interfaces.

Symmetric protein assemblies can now be modeled using parametric design. Nature created super-helical coiled-coils that are well-described by geometric equations using Crick parameters[181], 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[54,115] 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.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

Modeling peptides and peptidomimetics

The inherent flexibility of peptides imparts a large conformational search space to them, which leads to challenging modeling problems; when peptide modeling is combined with another simulation, e.g. docking, the increase in conformational space makes the modeling task virtually impossible using traditional approaches. PIPER-FlexPepDock[118] is to our knowledge the only global peptide docking protocol. It rigid-body docks these fragments using PIPER FFT-based docking[182], and refines the complex using FlexPepDock[116]. PIPER-FlexPepDock can generate highly accurate peptide-protein complexes from a peptide sequence and a free receptor structure (Figure 3F). Performance decreases in case of receptor flexibility and when fragments are not available in the fragment database.

Conformations of cyclic peptides can be sampled with *simple_cycpep_predict*, which restricts the conformational search space through cyclization[50,51,115] via the Generalized Kinematic Closure (GenKIC) algorithm (see "loop modeling" below). *Simple_cycpep_predict* does not rely on protein fragments and can model non-canonical chemistries (Figure 3B), being a generalization of earlier protocols.

Experimental protein structure determination is challenging for proteins on solid surfaces such as biominerals, self-assembled monolayers, inorganic catalysts, and nanomaterials. RosettaSurface[121] 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[122] to improve scoring, which remains difficult because the score function has been optimized for soluble proteins in aqueous solvent.

Using experimental data to direct modeling

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 accuracy. 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 these types of data have been successfully used for high-resolution structure determination. Since cryoEM density maps are often of low resolution, *de novo* structure determination methods require a combinatorial search procedure to unambiguously assign all densities to residues in the protein. RosettaES[125] is an enumerative sampling approach that does not require initial assignment of densities; it 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 energy and fit to the density.

If assignment is complete but the data are low-resolution, refinement into density maps is necessary. Several methods have been developed for density maps in the 3.0-4.5Å resolution range. More recently, an automated fragment-guided refinement pipeline[128] 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[150] (below) were connected[151] 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 the software's development, creating RosettaNMR. Chemical shifts were used for fragment picking using CS-Rosetta[129], which could be used in conjunction with NOE, RDC[183], PCS[130,131,184] and PRE data. Improvements, for instance through RASREC resampling[185] allowed the use of sparse[186] or unassigned data[187], easier-to-obtain data (backbone-only[188]), modeling larger and more complex proteins[189], membrane proteins[190], symmetric systems[191], and combination with data from SAXS[192], cryoEM[193], distance restraints from homologous proteins[194] and evolutionary couplings[195]. CS-Rosetta also has the AutoNOE[196,197] 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 several types of NMR restraints (CS, RDC, PCS, PRE, NOE) in one consistent framework[132] that could be applied to structure prediction, protein-protein docking, protein-ligand docking, and symmetric assemblies.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

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 foot-printing has been implemented that can improve model quality in structure prediction[133,134]. Finally, data from chemical cross-linking mass spectrometry has been incorporated into an automated workflow to identify protein-protein interactions. The PyTXMS[135] 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 nucleic acids and their interactions with proteins

DNA and RNA modeling face a multitude of challenges due to a lack of structures leading to under-developed score functions, low quality alignments, the sampling torsion space is much larger than for proteins (70 residue RNA comparable to 200 residue protein), and a lack of interest in the scientific community leading to a gap in knowledge. Moreover, in contrast to protein helices where sequence information is displayed on the helix exterior through side-chains, helical RNA sidechains point inwards, therefore hiding sequence information from the environment, making prediction of tertiary or non-local contacts vastly more difficult. Non-local contacts are mediated by loops, further enormously challenging prediction algorithms.

Several advances have been made in the representation of nucleic acids in Rosetta. The *StepWise Monte Carlo* protocol (SWM) has achieved RNA structure predictions reaching atomic accuracy[138]; the approach provides an acceleration over the original enumerative *StepWise Assembly* (SWA) method[136,137]. A version of SWA that rebuilds one nucleotide at a time enables fine-grained correction of errors in RNA coordinates fit into crystallographic or cryo-EM maps by *Enumerative Real-space Refinement ASsisted by Electron density under Rosetta*[142,143] (ERRASER).

The most recent advances in RNA tools expand the fragment assembly protocol to support modeling RNA-protein complexes through simultaneous folding and docking[141]. RNA-protein interactions are handled via additional knowledge-based score terms that supplement the low-resolution RNA score function. Free energy perturbations from RNA or protein mutations can be modeled with the Rosetta-Vienna $\Delta\Delta G$ protocol[49]. 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*)[145].

Redesign and prediction of protein-DNA interfaces is also possible[198,199] and has been accomplished with flexible protein backbones[200], genetic algorithms[198,200,201] and motif-biased rotamer sampling[202,203]. 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 score function that is optimized[203,204] for these highly polar and solvated interfaces. The software further supports prediction of specificity and affinity[205] and the prediction of DNA binding preferences of homologous proteins. Multi-template modeling in RosettaCM[62] was successfully applied to this challenge[206]. 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[207]. However, experimental difficulties have limited our understanding of their structures[208]. Previously, Yarov-Yarovoy[209,210] and Barth[211] implemented tools for low- and high-resolution structure prediction of membrane proteins, termed RosettaMembrane. These tools were recently re-engineered for compatibility with Rosetta3[33] into a platform called RosettaMP[146]. 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[176], PyMOL visualization[212], and assembly of symmetric proteins. In addition, a set of basic modeling tools[147] is 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

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

in the membrane. RosettaMP has further enabled rapid development of new modeling tools including structure-based detection of lipid exposed residues in the membrane[148] and domain assembly of full-length protein models from structures of transmembrane and soluble domains[149]. The RosettaCM protocol for multi-template homology modeling has also been adapted to membrane proteins[39].

Describing membrane protein energetics is challenging since the proteins live in an anisotropic environment and tend to bury polar solvent molecules (e.g. water, ions) that stabilize the structure and participate in important conformational transitions. Implicit membrane models often fail to reliably model membrane protein interiors. A method SPaDES was developed based on a hybrid explicit-implicit solvent model that enhanced the prediction and design of membrane protein structures[213].

Limitations to membrane protein modeling are similar but less severe than for RNA modeling: there are fewer structures in databases, fewer method developers in this field and hence fewer available tools. As a consequence, the score function is much less mature compared to the latest score functions for soluble proteins: the implicit solvent hydrophobic slab model is a very coarse-gained representation of the membrane. Ongoing efforts expand this model by including pores, lipid specificity and different thicknesses[214], yet many effects remain to be acknowledged such as measurement-specific or observed membrane geometries (micelles, bicelles, nanodiscs, vesicles, different pore types, fusion and fission of multiple membranes) and macroscopic physical phenomena like membrane tension and fluidity. Challenges in including these effects are experimental measurements for parameterization of these models and adaptation of a multitude of scoreterms.

Adding carbohydrates to the modeling process

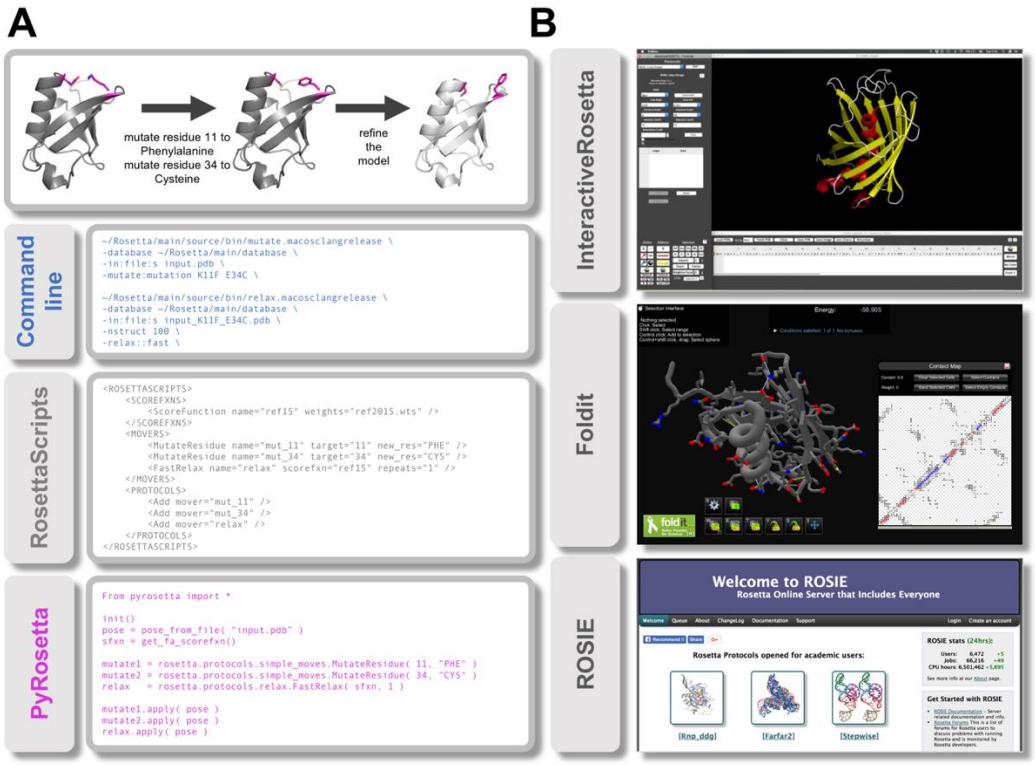
Carbohydrates are fundamental to life[215,216], but because of challenges in experimental characterization and computational sampling and scoring, their structures have been historically under-studied. The RosettaCarbohydrate framework[150] allows modeling of carbohydrate structures and complexes. The framework is integrated into the software 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[217]. 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, 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[151]. Limitations in the carbohydrate framework are the increased sampling space due to carbohydrate flexibility and branching, implementation of different chemistries due to branching and cyclization also requiring adjustments to the score function. Developments in this area have only started in the past years and much work has yet to be done.

4. User interfaces and usability

Advances have also focused on improving usability of Rosetta through developing several user interfaces to suit different use cases and styles (Figure 4). The command line interface was the first and is still the most-often used interface to Rosetta methods. In addition, the software features two major scripting interfaces: RosettaScripts and PyRosetta. RosettaScripts[37] is a popular scripting interface that uses E**x**tensible **M**arkup **L**anguage (XML) to build fairly complex protocols using core machinery[33], without requiring knowledge of the codebase. PyRosetta[36,152] is a collection of Python bindings to the source code, allowing custom protocol development that is flexible and fast, but requires familiarity with the underlying codebase. Other interfaces are InteractiveRosetta[153] and the gaming interface Foldit Standalone[154,156], further described in the Supplement.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

Figure 4: User interfaces to the codebase
(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 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 code objects but requires knowledge of the codebase. (B) Point-and-click interfaces to the codebase. 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 protocol. It currently includes servers for 21 Rosetta methods. [The InteractiveRosetta panel was reproduced with permission from *Bioinformatics*.]

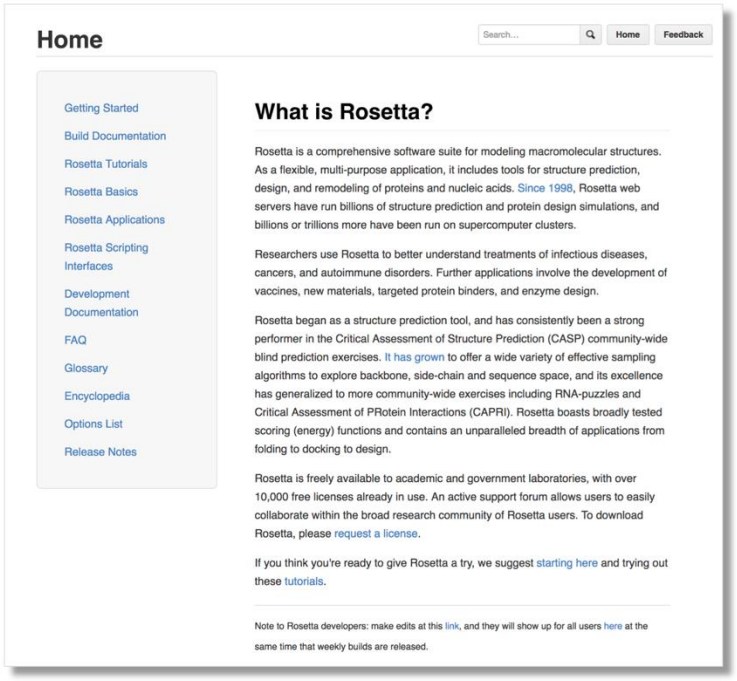


Our community has devoted an enormous effort to enhance the user friendliness of Rosetta by rewriting and adding documentation (Figure 5). We now use a public-facing Gollum wiki (https://www.rosettacommons.org/docs/latest/Home) for various levels of documentation, such as application documentation, tutorials 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 which has drastically improved the quantity and quality of documentation.

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 our community. However, implementing and maintaining such servers comes at a substantial cost. To make it easier to provide protocols webservers, ROSIE (Rosetta Online Server that Includes Everyone)[158,218] (http://rosie.rosettacommons.org/) implements a simple framework for “serverification” of protocols. ROSIE currently contains 21 webservers, with additional protocols continually being added.

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

Figure 5: Main external documentation page
In 2015, our community performed a complete overhaul of our documentation. 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 the software.



A look into the future

Rosetta development is ongoing and will continue to focus on expanding the scope of protein design and modeling by integrating high-throughput experimental data with high-throughput computational methods, which in turn impacts score function development and aids in developing novel therapeutic interventions[219]; restructuring the software for massively parallel computing architectures (e.g. GPUs, TPUs) and quantum computers[220]; greater use of machine-learning (e.g. deep-learning) approaches (e.g. for score function development); modeling more realistic cellular environments; and improving user interfaces to continue making our software accessible to more scientists. The predictive powers implemented in Rosetta that we have reviewed above can be leveraged not only to analyze and verify existing data but to inform the experiments that will galvanize engineering industrial enzymes, enable the creation of novel biomaterials, and accelerate the discovery of new potent therapeutics.

Conclusion

The Rosetta software is developed by a large, global community that aims to solve complex problems through real time collaborative code development. In the last five years, great strides have been made in our software. More protocols are available now that enable modeling a broader range of biological and chemical macromolecular systems. Prediction accuracies have improved through advances in the score function, which is a combination of physics-based and knowledge-based potentials that were fit against known structures and thermodynamic observables. Incorporating experimental data into the modeling process has been facilitated and improved. 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[32], and the complexities of the protocols used to solve real-world problems. The improvements to documentation allow users to quickly start using or developing custom protocols, while facilitating user support for the various interfaces (command line, RosettaScripts, PyRosetta, etc.). Over the years, these applications have

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

moved beyond tackling basic science questions (i.e. the protein folding and design challenges) to more application-based scientific developments. The myriad of advances described above have made integration of Rosetta into existing experimental and computational scientific workflows increasingly useful and standard, as evidenced by the large number of licenses (~30,000 academic and ~70 commercial), 11 spin-off companies that were created from the RosettaCommons[32], and the ever-increasing number of citations from labs beyond those affiliated with RosettaCommons.

Acknowledgements:
RosettaCommons is supported by NIH R01 GM073151 to Kuhlman, NSF, the Packard Foundation, the Beckman Foundation, the Alfred P. Sloan Foundation, and the Simons Foundation. This work was also supported by 100,000,000 CPU-hour donation from Google Inc to Conway; 125,760,000 CPU-hour allocation on the Mira and Theta supercomputers through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program to Baker, DiMaio, Leaver-Fay, and Mulligan. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science user facility supported under Contract DE-AC02-06CH11357. 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; 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; Hertz Foundation Fellowship to Alford; Howard Hughes Medical Institute to Baker; Hyak supercomputer system supported in part by the University of Washington eScience Institute to Baker and DiMaio labs; 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; National Science Centre, Poland, 2018/29/B/ST6/01989 to Gront; 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 Cooper, Khatib; NIH 5F32GM110899-02 to Linsky; NIH F31GM123616 to Jeliazkov; NIH F32CA189246 to Labonte; NIH P01 U19AI117905, R01 AI113867, UM1 AI100663 to Schief; NIH R00 GM120388 to Horowitz; NIH R01 AI143997 to Sgourakis; NIH R01 DK097376 to Meiler; NIH R01 GM073960 to Kuhlman; NIH R01 GM076324 to Siegel; NIH R01 GM078221 to Gray,; NIH R01 GM080403 to Meiler; NIH R01 GM084453 to Dunbrack; NIH R01 GM088277 to Bradley; NIH R01 GM092802 to Baker; NIH R01 GM092802 to Baker; NIH R01 GM098101 to Kortemme; NIH R01 GM099842 to Meiler; NIH R01 GM099959 to Karanicolas; NIH R01 GM110089 to Kortemme; NIH R01 GM117189 to Kortemme; NIH R01 GM117968 to Kuhlman; NIH R01 GM121487 to Bradley; NIH R01 GM123089 to DiMaio; NIH R01 GM126299 to Pierce; NIH R01 GM127578 to Gray; NIH R01 GM099827 to Bystroff; NIH R01 GM092802 to Baker; NIH R01 HL122010 to Meiler; NIH R01067553 to Kuhlman; NIH R01 GM078221 to Gray; NIH R01084433 to Baker; NIH R01088277 to Thyme; NIH R21 AI121799 to Meiler; NIH R21 CA219847 to Das; NIH R21 GM102716 to Das; NIH R35 GM122517 to Dunbrack; NIH R35 GM125034 to Sgourakis; NIH RL1CA133832 to Baker; NIH U19 AI117905 to Meiler; NIH/NCI Cancer Center Support Grant P30 CA006927 to Karanicolas; NSF 1507736 to Gray; NSF 1627539 to Siegel; NSF 1629879 to Cooper; NSF 1805510 to Siegel; NSF 1827246 to Siegel; NSF CHE 1305874 to Meiler; NSF CHE 1750666 to Lindert; NSF CISE 1629811 to Meiler; NSF CNS-1629811 to Meiler; NSF DBI-1262182 to Kortemme; NSF DBI-1564692 to Kortemme; NSF DMR 1507736 to Gray; NSF GRF DGE-1433187 to Rubenstein; NSF Graduate Research Fellowship to Alford, Kappel, Koepnick, Thyme; NSF MCB1330760 to Khare; NSF MCB1716623 to Khare; Open Philanthropy to Coventry; PhRMA Informatics Pre-Doctoral Fellowship U22879-001 to Smith; PhRMA foundation Predoctoral Fellowship to Fu; RosettaCommons to Goldschmidt, Rubenstein, DiMaio, Cooper, Watkins, Szegedy, Geniesse, Blacklock, Das, Khare, 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; XSEDE, which is supported by NSF ACI-1548562; NIH R01 GM097207 to Barth; MCB120101 XSEDE allocation to Barth.

Author contributions:
JKL wrote the manuscript with help from BDW. All authors edited and approved the manuscript and were substantially involved in developing the methods described, either by conception of the ideas or by implementing the methods into Rosetta. The idea for this paper was conceived by RB.

Conflicts of Interest:

Rosetta software has been licensed to numerous non-profit and for-profit organizations. Rosetta Licensing is managed by UW CoMotion, and royalty proceeds are managed by the RosettaCommons. Under institutional participation agreements between the University of Washington, acting on behalf of the RosettaCommons, their respective institutions may be entitled to a portion of revenue received on licensing Rosetta software including programs described here. Baker, Malmström, Gront, Meiler, Schueler-Furman, Gray, Sgourakis, Lindert, Karanicolas, Bonneau, Kortemme, and Bradley are unpaid board members of the RosettaCommons. As members of the Scientific Advisory Board of Cyrus Biotechnology, Baker and Gray are granted stock options. Yifan Song, Indigo C. King, Steven M. Lewis, Brandon Frenz, Karen Khar and Ryan Pavlovicz are currently employed at Cyrus Biotechnology with granted stock options. Cyrus Biotechnology distributes the Rosetta software. Brian D. Weitzner and Scott E. Boyken hold equity in Lyell Immunopharma. Vikram K. Mulligan is a co-founder of and shareholder in Menten Biotechnology Labs, Inc. The content of this manuscript is relevant to work performed at Lyell and Menten. Justin B. Siegel is a co-founder and shareholder of Digestiva, Inc. and PvP Biologics Inc. David Baker is a co-founder, shareholder, or advisor to the following companies: ARZEDA, PvP Biologics, Cyrus Biotechnology, Cue Biopharma, Icosavax, Neoleukin Therapeutics, Lyell Immunotherapeutics, Sana Biotechnology, and A-Alpha Bio.

References:

1. Schrodinger - Biologics Design. at <https://www.schrodinger.com/science-articles/biologics-design>
2. Molecular Operating Environment (MOE) | MOEsaic | PSILO. at <https://www.chemcomp.com/Products.htm>
3. Vilar, S., Cozza, G. & Moro, S. Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Curr. Top. Med. Chem.* **8,** 1555–1572 (2008).
4. Ref. Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017, San Diego: Dassault Systèmes, 2016. at <https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/>
5. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J. & Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20,** 473 (2019).
6. Vu, O., Mendenhall, J., Altarawy, D. & Meiler, J. BCL::Mol2D—a robust atom environment descriptor for QSAR modeling and lead optimization. *J. Comput. Aided. Mol. Des.* **33,** 477–486 (2019).
7. Webb, B., Viswanath, S., Bonomi, M., Pellarin, R., Greenberg, C. H., Saltzberg, D. & Sali, A. Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci.* **27,** 245–258 (2018).
8. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **3,** 33 (2011).
9. Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).

10. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–74 (2004).
11. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26,** 1701–18 (2005).
12. 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>
13. Shaw, D. E. A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. *J. Comput. Chem.* **26,** 1318–28 (2005).
14. Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J. & Rousseau, F. A graphical interface for the FoldX forcefield. *Bioinformatics* **27,** 1711–1712 (2011).
15. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R. & Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13,** e1005659 (2017).
16. Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Jones, D., Silver, D., Kavukcuoglu, K., Hassabis, D. & Senior, A. De novo structure prediction with deep-learning based scoring. *Thirteen. Crit. Assess. Tech. Protein Struct. Predict.* **December,** (2018).
17. Zhang, C., Mortuza, S. M., He, B., Wang, Y. & Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins Struct. Funct. Bioinforma.* **86,** 136–151 (2018).
18. Wang, S., Sun, S. & Xu, J. Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins Struct. Funct. Bioinforma.* **86,** 67–77 (2018).
19. Adhikari, B., Hou, J. & Cheng, J. Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins Struct. Funct. Bioinforma.* **86,** 84–96 (2018).
20. Fiser, A. & Sali, A. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods Enzymol.* **374,** 461–491 (2003).
21. Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.* **45,** D313–D319 (2017).
22. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2015).
23. van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J. & Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428,** 720–725 (2016).
24. Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T. & Weng, Z. ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30,** 1771–1773 (2014).
25. Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., Kholodov, Y., Ritchie, D. W., Vajda, S. & Kozakov, D. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E4286-93 (2016).
26. Tovchigrechko, A. & Vakser, I. a. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* **34,** 310–314 (2006).
27. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **33,** W363-7 (2005).
28. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** NA-NA (2009).
29. FlexX version 4.1; BioSolveIT GmbH, Sankt Augustin, Germany, 2019, www.biosolveit.de/FlexX.
30. Tubert-Brohman, I., Sherman, W., Repasky, M. & Beuming, T. Improved Docking of Polypeptides with Glide. *J. Chem. Inf. Model.* **53,** 1689–1699 (2013).
31. Sorenson, J. M. & Head-Gordon, T. Matching simulation and experiment: a new simplified model for simulating protein folding. *J. Comput. Biol.* **7,** 469–81 (2000).
32. Koehler Leman, J., Weitzner, B. D., Renfrew, P. D., Lewis, S. M., Moretti, R., Watkins, A. M., Mulligan, V. K., Lyskov, S., Adolf-Bryfogle, J., Labonte, J. W., Consortium, R., Bystroff, C., Schief,

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

W., Schueler-Furman, O., Baker, D., Bradley, P., Dunbrack, R., Kortemme, T., Leaver-Fay, A., Strauss, C. E., Meiler, J., Kuhlman, B., Gray, J. J. & Bonneau, R. Better together: Elements of successful scientific software development in distributed collaborative community. *Accept. PlosCompBio* (2019).

33. 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).
34. 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).
35. 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).
36. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26,** 689–691 (2010).
37. 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).
38. 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).
39. 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
40. Simoncini, D., Allouche, D., de Givry, S., Delmas, C., Barbe, S. & Schiex, T. Guaranteed Discrete Energy Optimization on Large Protein Design Problems. *J. Chem. Theory Comput.* **11,** 5980–9 (2015).
41. 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).
42. 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).
43. 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).
44. 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).
45. 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).
46. 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).
47. 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).
48. Mulligan, V. K. Manuscript in preparation. (2019).
49. Kappel, K., Jarmoskaite, I., Vaidyanathan, P. P., Greenleaf, W. J., Herschlag, D. & Das, R. Blind

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

tests of RNA–protein binding affinity prediction. *Proc. Natl. Acad. Sci.* **116,** 8336–8341 (2019).
50. 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).
51. 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).
52. 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).
53. 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).
54. 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).
55. 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).
56. 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).
57. Pavlovicz, R. E., Park, H. & DiMaio, F. Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking. *bioRxiv* 618603 (2019). doi:10.1101/618603
58. Bhowmick, A., Sharma, S. C., Honma, H. & Head-Gordon, T. The role of side chain entropy and mutual information for improving the de novo design of Kemp eliminases KE07 and KE70. *Phys. Chem. Chem. Phys.* **18,** 19386–19396 (2016).
59. König, R. & Dandekar, T. Solvent entropy-driven searching for protein modeling examined and tested in simplified models. *Protein Eng.* **14,** 329–35 (2001).
60. Park, H., Kim, D. E., Ovchinnikov, S., Baker, D. & DiMaio, F. Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. *Proteins Struct. Funct. Bioinforma.* **86,** 283–291 (2018).
61. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins Struct. Funct. Bioinforma.* **86,** 7–15 (2018).
62. Song, Y., Dimaio, F., Wang, R. Y.-R. R., Kim, D. E., Miles, C., Brunette, T., Thompson, J. & Baker, D. High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).
63. New Robetta server - http://new.robetta.org/.
64. 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,**
65. 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).
66. 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).
67. 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).
68. 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).
69. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A simple model of backbone flexibility

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

improves modeling of side-chain conformational variability. *J. Mol. Biol.* **380,** 757–74 (2008).
70. 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).
71. Stein, A. & Kortemme, T. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* **8,** e63090 (2013).
72. Lin, M. S. & Head-Gordon, T. Improved Energy Selection of Nativelike Protein Loops from Loop Decoys. *J. Chem. Theory Comput.* **4,** 515–21 (2008).
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. 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).
79. 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).
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. 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).
82. Roy Burman, S. S., Yovanno, R. A. & Gray, J. J. Flexible Backbone Assembly and Refinement of Symmetrical Homomeric Complexes. *Structure* (2019). doi:10.1016/j.str.2019.03.014
83. Meiler, J. & Baker, D. RosettaLigand: protein-small molecule docking with full side-chain flexibility. *Proteins* **65,** 538–48 (2006).
84. DeLuca, S., Khar, K. & Meiler, J. Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One* **10,** e0132508 (2015).
85. Davis, I. W. & Baker, D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **385,** 381–392 (2009).
86. Fu, D. Y. & Meiler, J. RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach. *ACS Omega* **3,** 3655–3664 (2018).
87. 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).
88. 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).
89. 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).
90. Khar, K. R., Goldschmidt, L. & Karanicolas, J. Fast Docking on Graphics Processing Units via Ray-Casting. *PLoS One* **8,** e70661 (2013).
91. Gowthaman, R., Lyskov, S. & Karanicolas, J. DARC 2.0: Improved Docking and Virtual Screening at Protein Interaction Sites. *PLoS One* **10,** e0131612 (2015).
92. Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* **37,** W474-479 (2009).
93. 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).

94. 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).
95. 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).
96. 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).
97. 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
98. 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).
99. Gowthaman, R. & Pierce, B. G. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res.* **46,** W396–W401 (2018).
100. 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).
101. 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).
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. 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,** 687–90 (2016).
107. 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).
108. 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).
109. 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).
110. 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).
111. 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).
112. 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).

113. 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).
114. 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).
115. 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).
116. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* **78,** 2029–40 (2010).
117. 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).
118. 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).
119. Sedan, Y., Marcu, O., Lyskov, S. & Schueler-Furman, O. Peptiderive server: derive peptide inhibitors from protein-protein interactions. *Nucleic Acids Res.* **44,** W536-41 (2016).
120. 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).
121. 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).
122. 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).
123. Pacella, M. S. & Gray, J. J. A Benchmarking Study of Peptide–Biomineral Interactions. *Cryst. Growth Des.* **18,** 607–616 (2018).
124. 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).
125. 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).
126. 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).
127. 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).
128. 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).
129. Nerli, S. & Sgourakis, N. G. CS-ROSETTA. *Methods Enzymol.* (2018). doi:10.1016/BS.MIE.2018.07.005
130. 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).
131. Schmitz, C., Vernon, R., Otting, G., Baker, D. & Huber, T. Protein structure determination from pseudocontact shifts using ROSETTA. *J. Mol. Biol.* **416,** 668–77 (2012).
132. Kuenze, G., Bonneau, R., Koehler Leman, J. & Meiler, J. Integrative protein modeling in RosettaNMR from sparse paramagnetic restraints. *bioRxiv* 597872 (2019). doi:10.1101/597872
133. 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).
134. 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
135. 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).

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. 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).
139. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7,** 291–294 (2010).
140. Cheng, C. Y., Chou, F.-C. & Das, R. Modeling Complex RNA Tertiary Folds with Rosetta. *Methods Enzymol.* **553,** 35–64 (2015).
141. Kappel, K. & Das, R. Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. *Structure* **27,** 140-151.e5 (2019).
142. 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).
143. 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
144. 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).
145. 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).
146. 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).
147. Koehler Leman, J., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics* **11,** 1–3 (2016).
148. 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).
149. 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
150. 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).
151. 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).
152. 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>
153. Schenkelberg, C. D. & Bystroff, C. InteractiveROSETTA: A graphical user interface for the PyRosetta protein modeling suite. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv492
154. 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).
155. 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).
156. 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
157. Lyskov, S., Chou, F. C., Conch??ir, S. ??, 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,** 5–7 (2013).
158. 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).

159. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS One* **6,** e20450 (2011).
160. 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).
161. 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).
162. 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).
163. 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).
164. 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).
165. 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).
166. 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).
167. 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).
168. North, B., Lehmann, A. & Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. *J. Mol. Biol.* **406,** 228–256 (2011).
169. Vaissier Welborn, V. & Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **119,** 6613–6630 (2019).
170. Marcos, E. & Silva, D.-A. Essentials of *de novo* protein design: Methods and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8,** e1374 (2018).
171. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures. *bioRxiv* 431635 (2018). doi:10.1101/431635
172. 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).
173. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A Generic Program for Multistate Protein Design. *PLoS One* **6,** e20937 (2011).
174. 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).
175. Sauer, M. F., Sevy, A. M., Crowe, J. E. & Meiler, J. Manuscript submitted. (2019).
176. 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).
177. 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).
178. Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci. STKE* **2004,** pl2 (2004).
179. 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).
180. 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).
181. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6,** 685–689 (1953).

182. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* **65,** 392–406 (2006).
183. 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).
184. 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).
185. 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).
186. Bowers, P. M., Strauss, C. E. M. & Baker, D. De novo protein structure determination using sparse NMR data. 311–318 (2000).
187. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15404–9 (2003).
188. 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).
189. 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).
190. 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).
191. 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).
192. 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).
193. 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).
194. 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).
195. Braun, T., Koehler Leman, J. & Lange, O. F. Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction. *PLoS Comput. Biol.* **11,** (2015).
196. 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).
197. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59,** 147–159 (2014).
198. 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).
199. Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J., Stoddard, B. L. & Baker, D. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441,** 656–659 (2006).
200. 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).
201. Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10,** 45–52 (2003).
202. 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

Peer-reviewed version available at *Nature Methods* **2020**; doi:10.1038/s41592-020-0848-2

design and directed evolution. *Nucleic Acids Res.* **42,** 2564–2576 (2014).
203. 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).
204. Yanover, C. & Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* **39,** 4564–76 (2011).
205. Ashworth, J. & Baker, D. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* **37,** e73 (2009).
206. 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).
207. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5,** 993–6 (2006).
208. Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins Struct. Funct. Bioinforma.* **83,** 1–24 (2015).
209. Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins* **62,** 1010–1025 (2006).
210. 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).
211. Barth, P., Schonbrun, J. & Baker, D. Toward high-resolution prediction and design of transmembrane helical protein structures. **2007,** (2007).
212. Baugh, E. H., Lyskov, S., Weitzner, B. D. & Gray, J. J. Real-time PyMOL visualization for Rosetta and PyRosetta. *PLoS One* **6,** e21931 (2011).
213. Lai, J. K., Ambia, J., Wang, Y. & Barth, P. Enhancing Structure Prediction and Design of Soluble and Membrane Proteins with Explicit Solvent-Protein Interactions. *Structure* **25,** 1758-1770.e8 (2017).
214. Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein structure prediction and design in a biologically-realistic implicit membrane. *bioRxiv* 630715 (2019). doi:10.1101/630715
215. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3,** 97–130 (1993).
216. 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).
217. 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).
218. 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).
219. Audacious Project - Institute for Protein Design. (2019). at <https://www.ipd.uw.edu/audacious/>
220. Mulligan, V. K., Melo, H., Merritt, H. I., Slocum, S., Weitzner, B. D., Watkins, A. M., Renfrew, P. D., Pelissier, C., Arora, P. S. & Bonneau, R. Designing Peptides on a Quantum Computer. *bioRxiv* 752485 (2019). doi:10.1101/752485