nature methods

PERSPECTIVE

https://doi.org/10.1038/s41592-020-0848-2

Macromolecular modeling and design in Rosetta: recent methods and frameworks

The Rosetta software for macromolecular modeling, docking and design is extensively used in laboratories worldwide. During two decades of development by a community of laboratories at more than 60 institutions, Rosetta has been continuously refactored and extended. Its advantages are its performance and interoperability between broad modeling capabilities. Here we review tools developed in the last 5 years, including over 80 methods. We discuss improvements to the score function, user interfaces and usability. Rosetta is available at http://www.rosettacommons.org.

The understanding that molecular structure determines biological function has motivated decades of experimental determination of protein structure and function. Many computational packages have been developed to guide experimental methods and elucidate macromolecular structure, including Rosetta. Rosetta offers capabilities spanning many bioinformatics and structural-bioinformatics tasks. Computational structural biology frameworks with similarly comprehensive scope are few, but key to progress in biology. Schrödinger[1], the Molecular Operating Environment[2] and Discovery Studio[3] 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[4] includes tools for bioinformatics, sequence alignments, structure prediction and modeling. The BioChemicalLibrary[5] (BCL) includes tools for structure prediction and drug discovery, and several sequence-to-structure methods using machine learning approaches. The Integrative Modeling Platform[6] (IMP) models large macromolecular complexes by incorporating various types of experimental data. OpenBabel[7] is a ChemInformatics toolbox supporting molecular mechanics calculations, being most heavily used for interconversion of file formats.

Molecular dynamics packages like CHARMM[8], AMBER[9], GROMACS[10] and others 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, modeling phenomena such as ion flow through membrane channels, and modeling interactions with small molecules and are therefore highly complementary to Rosetta. OpenMM[11] 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[12,13], QUARK[14], RaptorX[15]), homology modeling (Modeller[16], SwissModel[17]), fold recognition (iTasser[18]), protein–protein docking (HADDOCK[19], Zdock[20], ClusPro[21]), ligand docking (AutoDock[22], FlexX[23], Glide[24]) and many other tasks requiring molecular modeling. As the focus here is on Rosetta developments, a comprehensive list of related methods is listed in the Supplementary Note.

Development of Rosetta started in the mid-1990s; it was initially aimed at protein structure prediction and protein folding[25]. Over time, the number of applications grew to address diverse modeling tasks, from protein–protein or protein–small molecule docking to incorporating nuclear magnetic resonance (NMR) data, loop modeling, protein design, and interaction with peptides and nucleic acids (Fig. 1 and Tables 1 and 2). 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 wordwide[26]. The software has undergone several transitions, including in programming language and implementation, with the latest protocols based on Rosetta3, first released in 2008[27]. The score function has been continuously improved and was described in refs. [28,29]. As part of our sustained focus on accessibility, usability and scientific reproducibility, we developed several interfaces (PyRosetta[30], RosettaScripts[31], Foldit[32]) and emphasized publishing protocol captures[33] to accompany manuscripts. As those interfaces have grown more versatile and modular, development has accelerated and branched in many directions. However, the interoperability, extensibility and modularity enable scientists to combine modules in a wide variety of combinations, making it difficult to keep up with all the developments within the software and the scientific community. Here we have compiled the latest method developments in Rosetta from the past 5 years, divided into several categories; we provide direction on where to find further information for specific modeling problems. The Supplementary Note contains more details on the protocols, with extensive links to documentation, resources on the web, limitations, and competitors.

General overview and challenges

A typical Rosetta protocol is outlined in Fig. 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:

$$\text{if } E_{\text{new}} < E_{\text{orig}} \text{ accept}$$

$$\text{if } E_{\text{new}} \geq E_{\text{orig}} \text{ accept with probability } P = e^{-\left(\left(E_{\text{new}} - E_{\text{orig}}\right)/T\right)}$$

Many independent 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 "Rosetta's score function" below), or technical challenges. Many protocols suffer from undersampling[34], especially when flexibility is involved. Sampling is a limitation for structure prediction (especially for large structures), protein design and unconstrained global protein–protein docking.

A full list of authors and affiliations appear at the end of the paper.

For example, even with local docking we are limited by backbone flexibility and performance deteriorates with larger flexibility in the binding interface. Small-molecule docking similarly relies on correct identification of the binding interface and is limited by flexibility between unbound and bound states. Enormous conformational search spaces are also prohibitive for RNA modeling because of the size and combinatorics of the torsion space (see "Modeling nucleic acids and their interactions with proteins" below), membrane proteins because of their size, and carbohydrates because of branching and flexibility.

Some Rosetta applications suffer from technical challenges in implementation; a lack of documentation, protocol captures or support; and a need for more diverse chemistries for biomolecules. Technical challenges are either historical or due to lack of interest in the community to develop and advance methods in these unique areas.

Rosetta's score function

Rosetta's score function has been continuously improved over many years[35] 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 and statistically derived potentials describing respectively van der Waals energies, hydrogen bonds, electrostatics, disulfide bonds, residue solvation, backbone torsion angles, sidechain rotamer energies, and an average unfolded state reference energy (Fig. 2b):

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

Some energy terms are decomposed into several components 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 detailed account of the all-atom score function was published recently[28].

The newest score function[29], REF2015, reproduces thermodynamic observables (such as liquid-phase properties[36] and liquid-to-vapor transfer free energies[37]) in addition to structure-based[38] tests. It also utilizes a new, derivative-free optimization technique, 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[39]. The cartesian term[39] is also the basis for a cartesian_ddG method, which has been used to calculate $\Delta\Delta G$ values of mutations (where ΔG is the free energy of folding) to assess changes in protein stability. Only the backbones and side chains of residues near the mutation site are allowed to move[40]. Due to the local optimization, this protocol is much faster than the previous gold-standard ddg_monomer[41] while retaining the same level of accuracy. REF2015 is now compatible with an expanded palette of chemical building blocks—canonical and non-canonical L-α-amino acids and their D-amino acid counterparts, exotic achiral amino acids, peptoids and oligoureas—and can model metalloproteins[42]. Score functions that enable simultaneous modeling of protein and RNA are being explored[43]. REF2015 is now thread-safe and fully mirror symmetric; that is, enantiomers in mirror conformations score identically. Guidance energy terms for design have been added to encourage certain features, such as specific amino acid compositions[44,45], 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[46].

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[47] has been used to design de novo coiled coils with interaction specificity mediated by designed hydrogen bond networks, including homo-oligomers[47], membrane proteins[48] and large sets of orthogonal heterodimers[49]. An improvement to HBNet uses a Monte Carlo search to sample hydrogen bond networks with notably improved performance[50]. We further developed a statistical potential to place highly coordinated water molecules on the surface of biomolecules. On a dataset 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[51]. The potential is accessible through the ExplicitWaterMover (formerly WaterBoxMover) in RosettaScripts.

There are still several limitations to the score function. (1) It does not directly estimate entropy[52], which has been shown to improve sampling efficiency[53]. However, rotamer bond angles, solvation, fragments and pair terms all implicitly model this component of the 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, representing a single state on the energy landscape, and do not represent flexibility well (in comparison to solution NMR). (3) Knowledge-based terms are less interpretable and transferable than physics-based terms. (4) Scoring performance scales with the number of score terms and has become slower, although more accurate, over time. (5) The solvation model is implicit, and hence fast, but hinders explicit modeling of ions, water molecules or lipid environments. (6) Several score functions for specific applications (RNA, membrane proteins, carbohydrates, non-canonical amino acids) are still developing.

Major applications

Predicting protein structures. Rosetta was originally developed for de novo protein structure prediction, assembling fragments from known protein structures via a Monte Carlo procedure and evaluating the models with the score function. While the community's main goals have moved to macromolecular design over the past decade, performance in the CASP13 blind prediction challenge remains respectable[54], with ranking for refinement and prediction of multimeric complexes among the top three groups. Meanwhile, other groups have refined their tools exploiting evolutionary couplings and machine learning: for instance Google's DeepMind developed AlphaFold[12,13], which uses Rosetta for refinement, with outstanding performance in the recent CASP13[54]. Another highly ranking method is the Zhang server, built on iTasser[14] and QUARK[14].

Homology modeling was improved by using multiple templates in RosettaCM[55] (now available on the new Robetta[56,57] server), which hybridizes the most homologous portions from multiple templates into a single model while modeling missing residues de novo[55]. Without a template, predicting protein structures de novo remains one of the most challenging tasks in structural biology, even though the incorporation of evolutionary coupling constraints (for instance, from GREMLIN[58]) has led to enormous improvements in model quality. An iterative hybridization approach improves sampling and uses a genetic algorithm that recombines models from an input pool to create models that have features from their parents but are also distinct. Creating several child models in each iteration, updating the input pool, and performing 30–50 iterations lead to improved model accuracy because features that are scored favorably are repeatedly used in the recombination, such that the models in the pool converge over time. Iterative hybridization has been used to improve model quality of de novo predicted models[59] as well as homology models[60]. Model refinement or generating ensembles of structures (useful for design) can be accomplished by several algorithms in Rosetta: FastRelax[61], Backrub[62] or vicinity sampling[63] using kinematic closure (KIC) or next-generation-KIC (NGK) loop

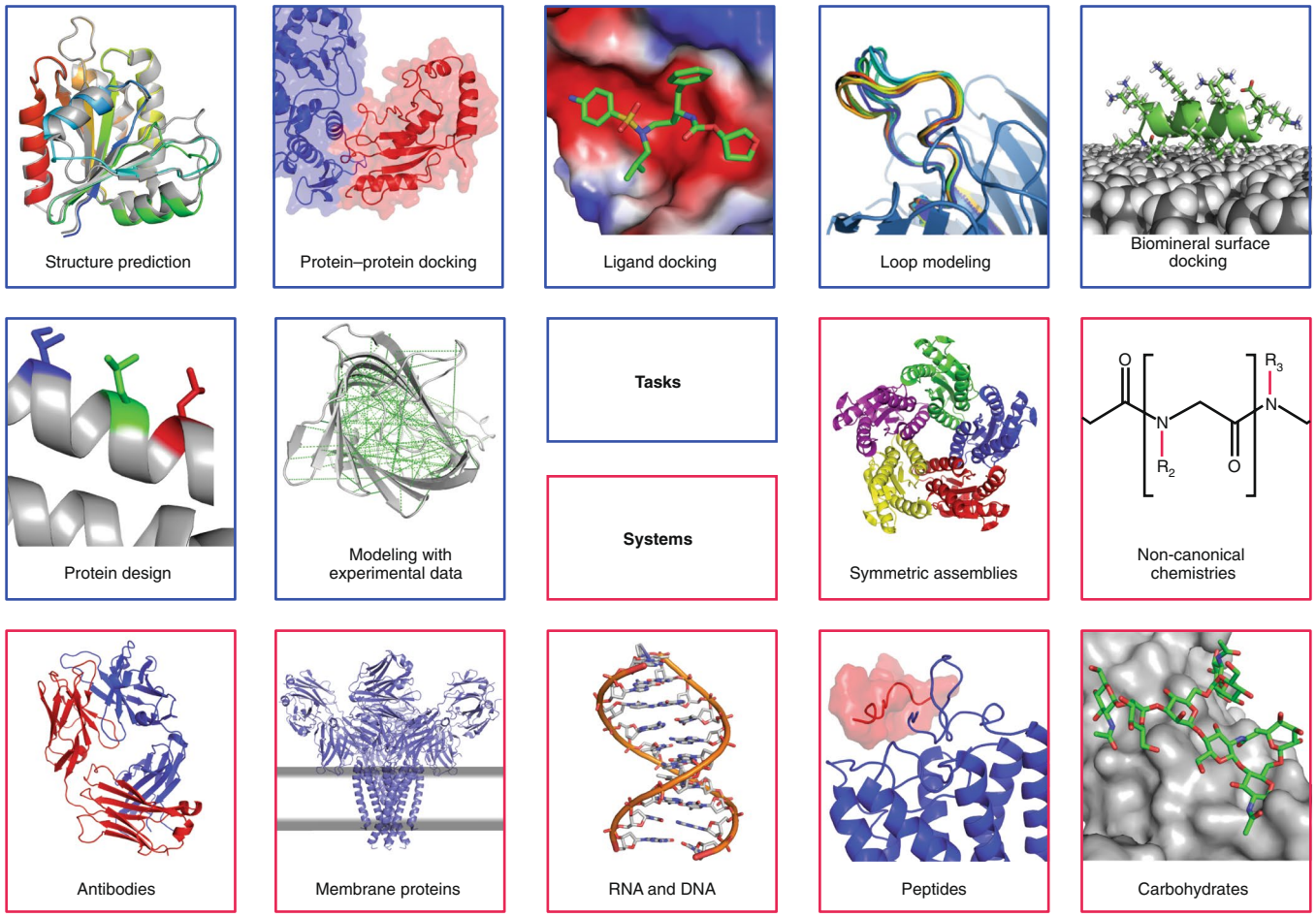

Fig. 1 | Capabilities of the Rosetta macromolecular modeling suite. Some popular tasks that can be addressed in Rosetta (blue) and major systems that can be modeled (red). Note that this is an incomplete list of Rosetta's broad modeling capabilities.

modeling[64]. Loop modeling[65] was implemented early in Rosetta[66,67], with initial approaches relying on fragment sampling and iterative cyclic coordinate descent (CCD)[68] for chain closure. Later, a KIC approach relied on polynomial resultants to analytically solve for closed conformations, producing more native-like loops[69,70]. Next-generation KIC[64] is an innovation that improves sampling by employing diversification (that is, wider range of conformations) and intensification (that is, focus around previously generated conformations), substantially increasing the fraction of near-native models[64] and modeling longer loops. A related method, GeneralizedKIC[44] (GenKIC), samples loop geometries between fixed endpoints, including non-standard peptide chemistries or chemistries that conventional loop-modeling algorithms do not typically handle.

Modeling protein–protein complexes. Another early expansion of Rosetta's functionality was RosettaDock, a method for predicting the structure of protein–protein complexes. The latest version, RosettaDock4.0[71], incorporates protein flexibility from pre-generated protein ensembles, mimicking conformer selection. This has improved sampling efficiency by automatically adjusting the sampling rate on the basis of the diversity of the input ensembles. Scoring has been improved by a six-dimensional coarse-grained scoring scheme called motif_dock_score, employing score grids generated from known complexes in the Protein Data Bank (PDB). In local docking benchmarks with backbone deviations of up to 2.2 Å, RosettaDock4.0 successfully docked ~50% of complexes[71]. For symmetric homomers, Rosetta SymDock2[72] uses the same six-dimensional scoring scheme as RosettaDock. Symmetry information can be extracted from a homologous complex, or from a global docking search for a given point symmetry using our symmetry framework[73]. An induced-fit-based all-atom refinement step 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[72]. These accuracies can be markedly improved when adding restraints.

Docking small-molecule ligands into proteins. Structure-based drug design has become a key drug optimization tool and leverages the vast array of knowledge contained in the increasing numbers of deposited structures in the PDB. RosettaLigand[74] has demonstrated success in predicting small molecule–protein interactions. Later in the drug development process, medicinal chemists optimize ligands on the basis of structure–activity relationships by synthesizing different ligands that share a core chemical scaffold and are assumed to bind to their target in a similar fashion[75]. RosettaLigandEnsemble[76] improves sampling during ligand docking by taking advantage of ligand similarities and docking a congeneric series of ligands simultaneously, allowing a placement that works for all considered ligands while optimizing the binding interface for each ligand independently. Experimental structure–activity relationships can help identify preferred binding modes. Small-molecule ligands can also

Table 1 | Overview of recent methods developed in Rosetta

Method	Developing laboratory
Score function	
REF2015 score function[28,29]	Frank DiMaio, David Baker
cartesian_ddG[29]	Frank DiMaio, Phil Bradley
HBNet[47,50]	David Baker, Brian Kuhlman
HBNetEnergy[47]	Richard Bonneau, David Baker[a]
AACompositionEnergy	Richard Bonneau, David Baker[a]
AARepeatEnergy	Richard Bonneau, David Baker[a]
VoidsPenaltyEnergy	Richard Bonneau, David Baker[a]
NetChargeEnergy	Richard Bonneau, David Baker[a]
BuriedUnsatPenalty	Richard Bonneau, David Baker[a]
Protein structure prediction	
fragment picker[190]	Dominik Gront[a,b]
RosettaCM[55]	David Baker
iterative hybridize[59,60]	David Baker, Sergey Ovchinnikov[a]
Loop modeling	
NGK (next-generation KIC)[64]	Tanja Kortemme
GenKIC (generalized KIC)[44]	Richard Bonneau, David Baker[a]
LoopHashKIC	Tanja Kortemme
Consensus_Loop_Design[101,191]	David Baker
Protein–protein docking	
RosettaDock4.0[71]	Jeffrey Gray
Rosetta SymDock2[72]	(Ingemar André)[c], Jeffrey Gray
Small molecule ligand docking	
RosettaLigand[74,192,193]	Jens Meiler
RosettaLigandEnsemble[76]	Jens Meiler
pocket optimization[77,78]	John Karanicolas
DARC[194–196]	John Karanicolas
Modeling of antibodies and immune system proteins	
RosettaAntibody[80–83]	Jeffrey Gray
AbPredict[89,90]	Sarel Fleishman
RosettaMHC[197]	Nik Sgourakis
TCRModel[198]	Brian Pierce
SnugDock[91]	Jeffrey Gray
Design of antibodies and immune system proteins	
RAbD[93] (RosettaAntibodyDesign)	Bill Schief, Roland Dunbrack
Epitope removal[95,96]	David Baker, Cyrus Biotechnology
AbDesign[97,98]	Sarel Fleishman
Protein design	
SEWING[103,104]	Brian Kuhlmann
RosettaRemodel[106]	Possu Huang[a,b]
LooDo[199]	Sagar Khare
RECON[108]	Jens Meiler
curved β-sheet design[101]	David Baker
biased forward folding[101]	David Baker
fold_from_loops[111]	Bruno Correia[a,b]
FunFolDes[112]	Bruno Correia
Protein interface design	
FlexDDG[117]	Tanja Kortemme
Coupled Moves[200]	Tanja Kortemme, DSM Biotechnology Center
Parametric design[48,120]	Richard Bonneau[a]

[a]The main developer(s) in this lab were formerly in the lab of David Baker when this application was developed. [b]The main developer(s) now have their own labs. [c]Names in parentheses were either initial developers or previously involved in development but are no longer involved in development and maintenance of this part of the code.

be used as competitive inhibitors of protein–protein interactions. However, a protein's inhibitor-bound conformation often differs from the unbound or protein–protein bound conformation; thus Rosetta's ability to model protein conformational flexibility is key. Rosetta's 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 bias sampling such that low-energy pocket-containing states are preferentially explored[77,78]. The sampled conformations match 'druggable' alternative conformations observed in ligand-bound structures[77,78], making these states excellent starting points for virtual screening. Pockets sampled on a protein surface can then be matched to complementary ligands by using the pocket as the starting point for pharmacophore-based screening[79].

Modeling and designing antibodies and immune system proteins. Due to the therapeutic significance of antibodies, several antibody-specific and immune-specific protocols have been developed for structure prediction, docking and design, with specific protocols targeting immunoglobulin G, T-cell receptors, displayed antigens of the major histocompatibility complex (MHC), and other soluble antigens and immunogens. RosettaAntibody[80–83] is a protocol for modeling of antibodies. It identifies homologous templates, assembles them into a single structure and then models complementarity-determining region (CDR) H3 loops de novo while refining the orientation of the variable domain of the heavy and light chains[84]. Recent advances use multiple templates[84], incorporate key structural constraints[85,86] into CDR H3 modeling, and model camelid antibodies[82] and antibodies on the scale of the human repertoire[87,88]. AbPredict[89] 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 accurately modeling cases with sequence identity as low as 10%. AbPredict2 is available as a webserver[90]. SnugDock[91] is a related method for antibody–antigen docking, taking as input a plausible starting conformation and optionally an ensemble of antibodies and antigens. SnugDock then runs local docking to refine both the antibody–antigen interface and the heavy chain–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[85,86] and docking camelid antibodies[92]. Limitations in antibody modeling depend on the task: docking is limited by 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.

RosettaAntibodyDesign[93] (RAbD) is based on RosettaAntibody[82] 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[94], reducing deleterious sequence mutations, and was benchmarked on 60 diverse antibody–antigen interfaces from complexes including both λ and κ light chains. 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 design proteins with reduced immunogenicity while retaining function and stability[95]. The approach implements machine-learning-based epitope prediction for 28 different alleles, restricts design to select 15-mer epitope regions, and uses greedy stepwise protein design[96] to eliminate the most immunogenic epitopes with the least mutations, avoiding disruptive core mutations likely to destabilize the protein. Another method, AbDesign, splits experimentally determined antibody structures along conserved positions to create interchangeable segments and then recombines them to produce a diverse set

Table 2 | Overview of additional recent methods developed in Rosetta

Method	Developing laboratory
Peptides and peptidomimetics	
FlexPepDock[123,201]	Ora Schueler-Furman
PIPER-FlexPepDock[121]	Ora Schueler-Furman
PeptiDerive[202]	Ora Schueler-Furman
simple_cycpep_predict[44,45,120]	Richard Bonneau, David Baker[a]
MFPred[203]	Sagar Khare
RosettaSurface[124,125,204]	Jeffrey Gray
Modeling with experimental data	
cryo-EM de novo[205]	Frank DiMaio, David Baker
cryo-EM: RosettaES[126]	Frank DiMaio
cryo-EM: iterative refinement[206,207]	Frank DiMaio[a,b]
cryo-EM: automated refinement[127]	Frank DiMaio
NMR: CS-Rosetta[130]	Nik Sgourakis
NMR: PCS-Rosetta, GPS-Rosetta[132,133]	Thomas Huber
RosettaNMR framework[148]: using RDC/PRE/PCS/NOE/CS for ab initio protein–protein docking, ligand docking, symmetric assembly	Jens Meiler, Richard Bonneau, (Jeffrey Gray)[c]
mass-spec: HRF hydroxyl radical footprinting[149,150]	Steffen Lindert
mass-spec: PyTXMS[151]	Lars Malmström
RNA modeling	
SWA (stepwise assembly)[153,154]	Rhiju Das
SWM (stepwise Monte-Carlo)[152]	Rhiju Das
FARFAR (fragment assembly medium resolution structure prediction)[157,208,209]	Rhiju Das
ERRASER (refinement into EM density maps)[155,156]	Rhiju Das
CS-Rosetta-RNA (modeling with NMR data)[210]	Rhiju Das
RECCES (Reweighting of Energy-function Collection with Conformational Ensemble Sampling)[211]	Rhiju Das
DRRAFTER (de novo modeling of protein-RNA complexes into EM densities)[158]	Rhiju Das
Membrane proteins	
RosettaMP framework[172]: mp_ddg, mp_dock, mp_relax, mp_symdock	Jeffrey Gray, Richard Bonneau
RosettaMP toolkit[174]: mp_score, mp_transform, mp_mutate_relax, helix_from_sequence	Jeffrey Gray, Richard Bonneau
mp_lipid_acc[175]	Richard Bonneau
mp_domain_assembly[176]	Richard Bonneau
RosettaCM for membrane proteins[33]	Jens Meiler
Carbohydrates	
RosettaCarbohydrate framework[128,129]	Jeffrey Gray, William Schief
User interfaces	
PyRosetta[30,182,212]	Jeffrey Gray
RosettaScripts[31,33]	Sarel Fleishman[a,b]
InteractiveRosetta[183]	Chris Bystroff
Foldit Standalone[32,184,185,213]	Seth Cooper[a,b], Firas Khatib[a,b], Justin Siegel, Scott Horowitz, David Baker
ROSIE server[186,187]	Jeffrey Gray
Miscellaneous	
Metalloproteins[42]	David Baker, Richard Bonneau[a]
Waters[51]	Frank DiMaio
SimpleMetrics	William Schief
AmbRose	Sagar Khare
RosettaRC	William Schief

[a]The main developer(s) in this lab were formerly in the lab of David Baker when this application was developed. [b]The main developer(s) now have their own labs. [c]Names in parentheses were either initial developers or previously involved in development but are no longer involved in development and maintenance of this part of the code.

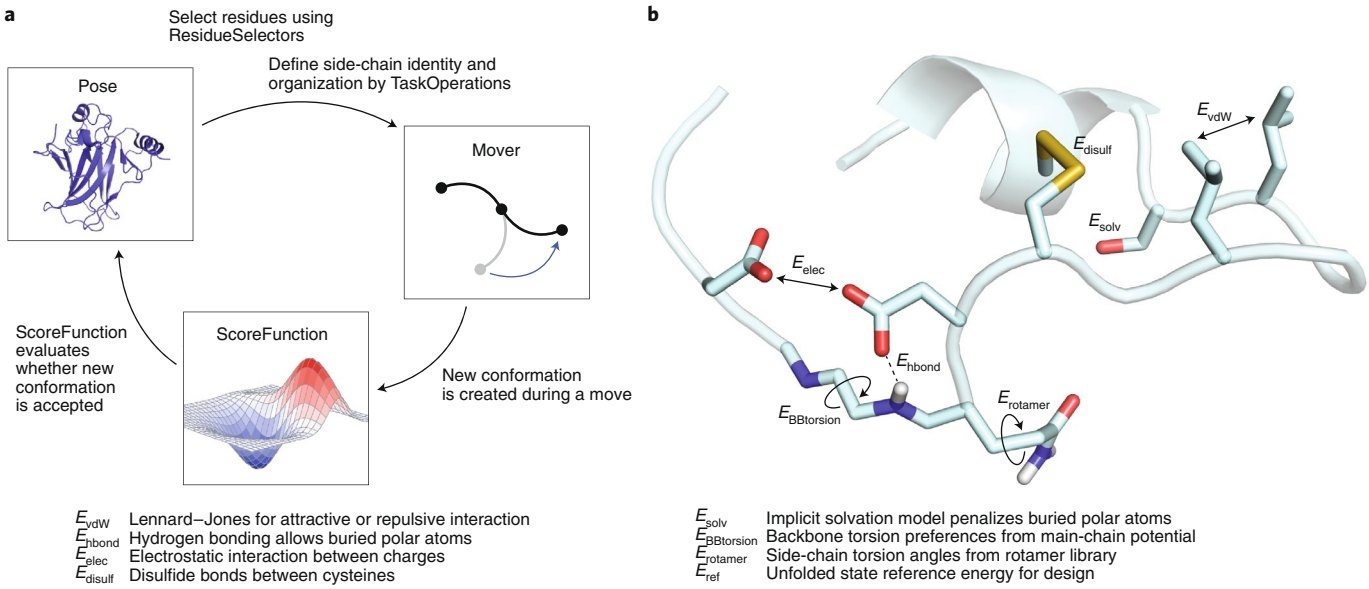

Fig. 2 | Main elements of Rosetta are scoring and sampling. a, Three main elements are required in a Rosetta protocol. The Pose is the biomolecule, such as a protein, RNA, DNA, small molecule, or glycan, in a specific conformation. Residues in the Pose can be selected via ResidueSelectors and the behavior 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 during sampling. Many independent sampling trajectories are generated, and the final models are evaluated according to 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 in the text.

of novel antibody models[97,98]. The models are docked to a target of interest, either locally to a specific epitope or globally, followed by an optimization step comprising rigorous backbone sampling and sequence design for improving model stability and binding affinity.

Designing new proteins and functions. Protein design[99] relies on several of the same core functionalities needed for structure prediction, and synergy and interoperability between design and prediction models has always been a core Rosetta principle. For example, this synergy is well illustrated by the biased forward folding method: during de novo protein design[100], a 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 limit that space and test more designs, biased forward folding[101] uses 3 (instead of 200) fragments per residue position, with fragments being chosen on the basis of the r.m.s. deviation to the native structure used to instantiate the design process. Protein design is easier when starting from known structures and when redesigning for well understood objectives such as thermostability[102]. More difficult design objectives include de novo design (without a template structure) and design for novel folds or functions. Successes in these cases require sampling of enormous conformational spaces, depending on the protein size. 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.

Design of novel protein structures and functions toward therapeutic intervention is addressed by various methods in Rosetta. SEWING creates de novo designs by recombining parts of protein structures from randomly selected helical building blocks[103]. SEWING's requirement-driven approach allows users to specify features that should be incorporated into their designs during backbone generation without 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 ligand binding sites for ligand-binding protein design[104]. A similar algorithm was implemented for antibody design (AbDesign, see above), which was generalized for enzyme design[105]. A more general approach is RosettaRemodel, performing protein design by rebuilding parts or all of the structure[106] from fragments of known protein structures. RosettaRemodel uses a blueprint file in which the user defines secondary and supersecondary structure of the desired fold. Remodel interfaces with various Rosetta protocols and allows de novo modeling; fixed-backbone sequence design; refinement; loop insertion, deletion and remodeling; disulfide engineering; domain assembly; and motif grafting.

A common task is not only design toward a certain goal (positive design), but design away from undesired features (negative design). Such a multi-state design[107] approach evaluates strengths and weaknesses of a single sequence on multiple backbones — for instance, binding to one but not another protein partner. REstrained CONvergence[108] (RECON) allows 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 multi-state design) for large multi-state design problems, such as antibody affinity maturation or the prediction of evolutionary sequence profiles of flexible backbones[109,110].

Protein function can be designed by motif grafting—that is, 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 that includes multi-segment motif grafting, different residue length motif insertion, the 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. Protein design problems include interface design between proteins and proteins or small-molecule ligands and prediction of $\Delta\Delta G$ values of mutation (for example, alanine scanning). Predicting $\Delta\Delta G$ values of mutations for protein stability or protein–protein interactions is difficult with low correlation coefficients (0.5–0.7)[113] 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 alanine scanning (mutating residues into alanine), methods that use a 'soft-repulsive' score function without modeling backbone flexibility[114,115] typically outperform methods that allow protein flexibility and use hard-repulsive score functions[116]. FlexDDG[117] improves protein–protein interface $\Delta\Delta G$ predictions and generalizes them to residues other than alanine. The protocol creates conformational ensembles using Backrub sampling[118], then repacks sidechains, minimizes torsions and computes the change in protein–protein interaction $\Delta\Delta G$ by averaging across the ensembles. On 1,240 interface mutants, FlexDDG outperformed the earlier ddg_monomer application, which was created to predict changes in stability upon mutation, not interfaces.

Symmetric protein assemblies modeled using parametric design. Nature created superhelical coiled coils that are well described by geometric equations using Crick parameters[119], including variables for the radius of the bundle, major helical twist and minor helix rotation about the primary axis. Several Movers, such as MakeBundle, PerturbBundle and BundleGridSampler, allow one to design helical bundles[48,120] and β-barrels on the basis of predefined or sampled parameters. These parametric methods do not rely on fragments libraries and can be applied to non-canonical coiled-coil heteropolymers.

Modeling peptides and peptidomimetics. The inherent flexibility of peptides imparts a large conformational search space to them, leading to challenging modeling problems; when peptide modeling is combined with another simulation—for example, docking—the increase in conformational space makes the modeling task challenging by any method. PIPER-FlexPepDock[121] is Rosetta's global peptide docking protocol. It rigid-body docks fragments using PIPER FFT-based docking[122] and refines the complex using FlexPepDock[123]. PIPER-FlexPepDock can generate peptide–protein complexes from a peptide sequence and a free receptor structure (Fig. 3f). Performance decreases in cases of receptor flexibility.

Cyclic peptide conformations can be sampled with simple_cycpep_predict, restricting the conformational search space through cyclization[44,45,120] via the GenKIC algorithm (see "Predicting protein structures" above). Simple_cycpep_predict does not rely on protein fragments and can model non-canonical chemistries (Fig. 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[124] samples protein conformations ab initio in both the solution and adsorbed states (Fig. 3d) to account for adsorption-induced conformational changes. Experimental data can be incorporated[125] to improve scoring.

Using experimental data to direct modeling. Using experimental data in modeling can vastly restrict the conformational space, allowing the modeling of larger, more complex biomolecules to greater accuracy. Electron density maps generated by cryo-electron microscopy (cryo-EM) or X-ray crystallography have improved in quality and become substantially more available in the past decade, and methods to incorporate them can produce high-resolution structures. To deal with variations in the resolution of these methods, RosettaES[126] samples enumeratively, not requiring initial assignment of densities; it gradually extends the model one residue at a time until all residues are 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 of low resolution, refinement into density maps is necessary. Several methods have been developed for density maps in the 3.0–4.5Å resolution range. More recently, an automated fragment-guided refinement pipeline[127] 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 on the basis of their fit to the validation map, model geometry from MolProbity and fit to the full map; and then optimizes against the full map. Further, the frameworks for electron density maps and carbohydrate modeling[128] (below) were connected[129], allowing refinement of carbohydrates into low-resolution density maps.

NMR data were incorporated into de novo structure prediction early on, embodied in RosettaNMR. Chemical shifts (CS) were used for fragment picking using CS-Rosetta[130], which could be used with nuclear Overhauser enhancements (NOEs), residual dipolar couplings (RDCs)[131], pseudo-contact shifts (PCSs)[132–134] and paramagnetic relaxation enhancement (PRE) data. Improvements—for instance through RASREC resampling[135]—allowed the use of sparse[136] or unassigned data[137]; the use of easier-to-obtain data (backbone-only[138]); the modeling of larger and more complex proteins[139], membrane proteins[140] and symmetric systems[141]; and combination with data from small-angle X-ray scattering (SAXS)[142], cryo-EM[143], distance restraints from homologous proteins[144] and evolutionary couplings[145]. CS-Rosetta also has the AutoNOE[146,147] module for automated assignment of NOE data for use in structure calculations. RosettaNMR was recently overhauled and reconciled with CS-Rosetta and PCS-Rosetta to seamlessly integrate several types of NMR restraints (CS, RDC, PCS, PRE and NOE) in one consistent framework[148] for structure prediction, protein–protein docking, protein–ligand docking and symmetric assemblies.

Covalent-labeling mass spectrometry data provide information on relative solvent exposure of residues, yielding information on protein tertiary structure. A low-resolution score term that allows use of hydroxyl radical footprinting has been implemented that can improve model quality in structure prediction[149,150]. Moreover, data from chemical cross-linking mass spectrometry has been incorporated into an automated workflow to identify protein–protein interactions. The PyTXMS[151] protocol combines the sensitivity of mass spectrometry for analyzing complex samples with the power of Rosetta structural modeling and protein–protein docking to efficiently sample the vast conformational space and identify interactions (Fig. 3c). A machine-learning algorithm based on high-resolution first-stage mass spectrometry (MS1) data guides the potential binding interface selection, being validated and adjusted by a repository of structural models and second-stage mass spectrometry (MS2, data-dependent acquisition) samples.

Modeling nucleic acids and their interactions with proteins. DNA and RNA modeling requires addressing a multitude of challenges due to a lack of structures leading to underdeveloped score functions, low quality alignments, and a much larger sampling

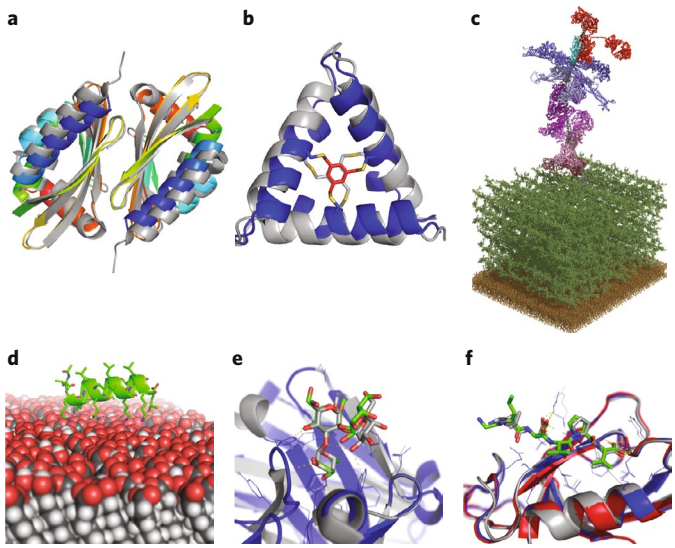

Fig. 3 | Rosetta can successfully address diverse biological questions. **a**, Curved β-sheet design: overlay of the designed homo-dimeric curved β-sheet (dcs-E_4_dim_cav3) in rainbow and the crystal structure in gray (PDB 5U35). The protein is designed de novo and features a curved β-sheet, a large pocket and a homodimer interface[101]. **b**, Parametric design: overlay of the de novo designed macrocycle 3H1 in blue and the NMR structure in gray (PDB 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)[120]. **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; membrane, brown). This 1.8-MDa structure contains over 200 chemical cross-links[151] and is measured in a complex mixture of intact bacteria and human plasma. All models are provided by Rosetta: M1 protein (gray), immunoglobulin G (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 α-1-antitrypsin (SerpinA1; plum). **d**, RosettaSurface: model of an LK-α peptide (LKKLLKLLKKLLKL, with a periodicity of 3.5 under the assumption of a helical conformation) on a hydrophilic self-assembled monolayer surface. The peptide is unstructured in solution and assumes helical structure[125] when on the surface, as experiments show. **e**, RosettaCarbohydrate: flexible docking of a carbohydrate antigen to an antibody. The crystal structure is in gray (PDB 1MFA) and the model in blue, with the carbohydrate in green. Antibody coordinates were taken from PDB and glycan coordinates started from a randomized backbone conformation and rigid-body orientation[128]. **f**, PIPER-FlexPepDock: high-resolution model of a peptide–protein complex (model, blue; solved structure, gray; PDB 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, PDB 2H3L; red)[121].

torsion space than for proteins (that of a 70-residue RNA being comparable to that of a 200-residue protein). In contrast to protein helices, where side chains display sequence information on the helix exterior, helical RNA side chains point inwards, therefore hiding sequence information from the environment, making prediction of tertiary or non-local contacts more difficult. Non-local contacts are mediated by loops, 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 prediction reaching atomic accuracy[152]; the approach provides an acceleration over the original enumerative StepWise Assembly (SWA) method[153,154]. 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 ERRASER (Enumerative Real-space Refinement Assisted by Electron Density under Rosetta)[155,156].

The most recent advances in RNA tools expand the fragment assembly protocol to support modeling RNA–protein complexes through simultaneous folding and docking[157]. RNA–protein interactions are handled via 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[43]. 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)[158]. Redesign and prediction of protein–DNA interfaces[159,160] have been accomplished with flexible protein backbones[161], genetic algorithms[159,161,162] and motif-biased rotamer sampling[163,164]. A potential limitation is the reliance on fixed DNA backbone conformations, as DNA backbone conformations can be flexible. Key to successful protein–DNA design is a score function optimized[164,165] for these highly charged and solvated interfaces. Rosetta supports prediction of specificity and affinity[166], the prediction of DNA binding preferences of homologous proteins, and multi-template modeling in RosettaCM[55,167].

Modeling membrane proteins. Membrane proteins constitute about 30% of all proteins and are targets for over 60% of pharmaceuticals on the market[168]. However, experimental difficulties have limited our understanding of their structures[169]. Previously, Yarov-Yarovoy[170] and Barth[171] implemented tools for low- and high-resolution structure prediction of membrane proteins, termed RosettaMembrane. These tools were re-engineered for compatibility with Rosetta3[27] into a platform called RosettaMP[172]. 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[113], PyMOL visualization[173] and assembly of symmetric proteins. Additionally, a set of basic modeling tools[174] allows scoring, transformation of a membrane protein into the membrane coordinate frame, modeling of single-transmembrane-span helices de novo, introduction of mutations, and visualization in the membrane. RosettaMP has enabled rapid development of new tools, including those for structure-based detection of lipid-exposed residues in the membrane[175] and domain assembly of full-length protein models from structures of transmembrane and soluble domains[176]. The RosettaCM protocol for multi-template homology modeling has also been adapted to membrane proteins[33].

Describing membrane protein energetics is challenging as these proteins reside in an anisotropic environment and bury polar solvent molecules (for example, water and ions) that stabilize the structure and participate in important conformational transitions. Implicit membrane models often fail to reliably model membrane protein interiors. The method SPaDES is based on a hybrid explicit–implicit solvent model that enhances the prediction and design of membrane protein structures[177]. 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. Consequently, the score function is less mature than the latest score functions for soluble proteins: the implicit solvent hydrophobic slab model is a coarse-gained representation of the membrane. Ongoing efforts expand this model by including pores, lipid specificity and different thicknesses[178], yet many effects remain to be acknowledged, such as measurement-specific or observed membrane geometries (micelles, bicelles, nanodiscs, vesicles, different pore types, and fusion and fission of multiple membranes) and macroscopic physical phenomena such as membrane tension and fluidity. Challenges in including these effects are experimental

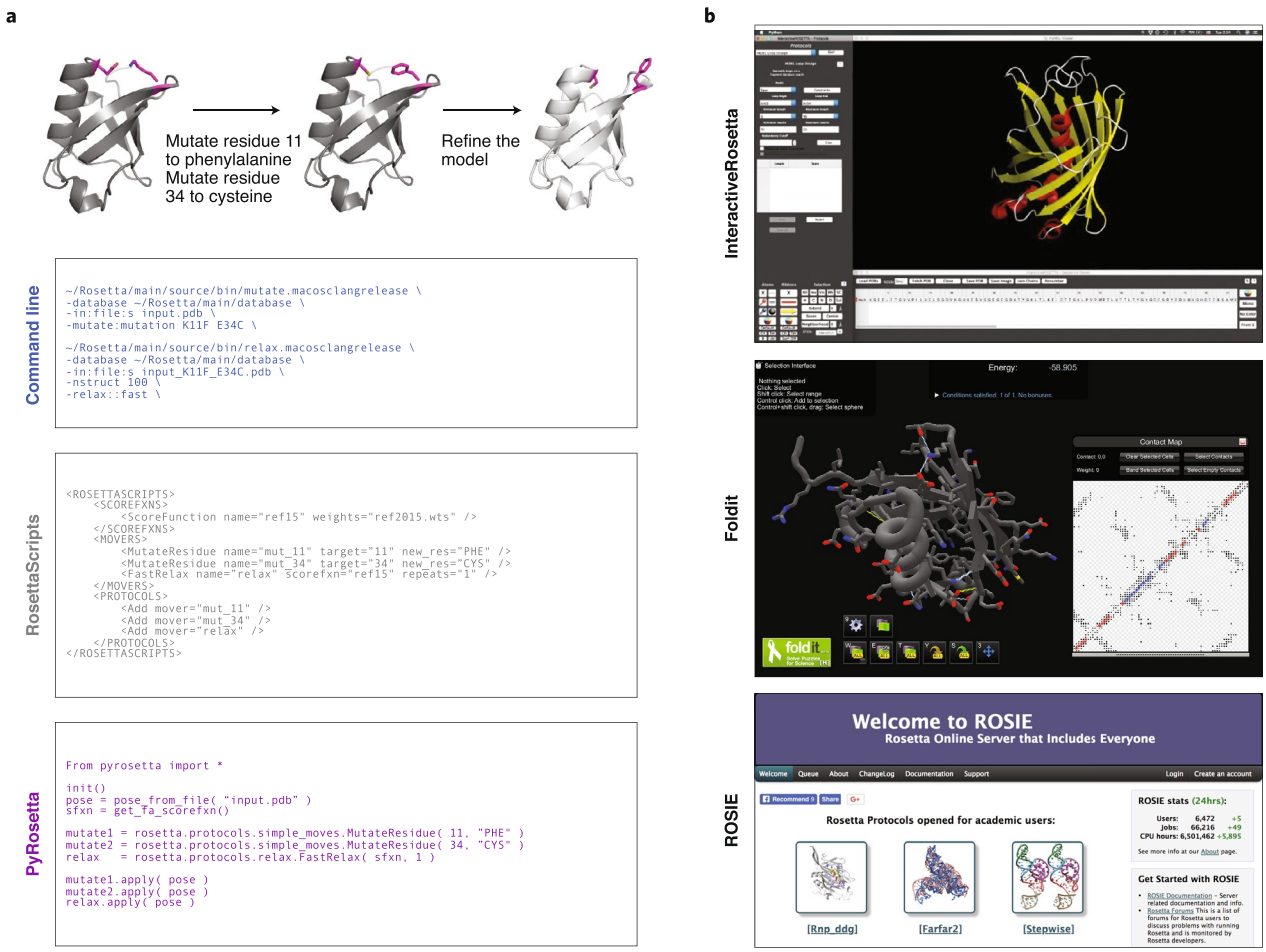

Fig. 4 | User interfaces to the codebase. a, Rosetta can be run from a terminal and offers three interfaces to the codebase. The top panel outlines the task to be accomplished: making two mutations in a protein and then refining the structure. The panels underneath show how this task can be accomplished 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 video game primarily used to crowd-source real-world scientific puzzles and is also usable on custom proteins of interest. It can run some popular applications via a game interface. ROSIE hosts a multitude of servers, each executing a particular protocol. It currently includes servers for 25 Rosetta methods. (InteractiveRosetta and Foldit panels reprinted from refs. [184,214] under Creative Commons licenses.).

measurements for parameterization of these models and adaptation of a multitude of score terms.

Adding carbohydrates to the modeling process. Carbohydrates are fundamental to life[179,180], but because of challenges in experimental characterization and computational sampling and scoring, their structures have been historically under-studied. The RosettaCarbohydrate framework[128] models carbohydrate structures and complexes such as glycosylated proteins or protein–sugar complexes (Fig. 3f) with the same algorithms one would use for proteins. RosettaCarbohydrate can handle commonly studied and uncommon 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[181]. Because saccharide residues are stored as distinct data structures, we can integrate bioinformatic and statistical data into these algorithms, opening the door for glycoengineering and design applications. RosettaCarbohydrate has been integrated with 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 can be refined into electron density maps[129]. Limitations in the carbohydrate framework include the increased sampling space due to carbohydrate flexibility and branching, and the need to model many different chemistries with possible branching and cyclization. Developments in this area have only recently started, and much work remains.

User interfaces and usability

Advances have also focused on improving usability of Rosetta through several user interfaces to suit different use cases and workflow styles (Fig. 4). The command line was the first and is still the most often used interface to Rosetta methods. Additionally, Rosetta features two popular scripting interfaces: RosettaScripts and PyRosetta. RosettaScripts[31] uses Extensible Markup Language (XML) to build complex protocols using core machinery[27], without requiring knowledge of the codebase. PyRosetta[30,182] is a collection of Python bindings to the source code, allowing flexible and

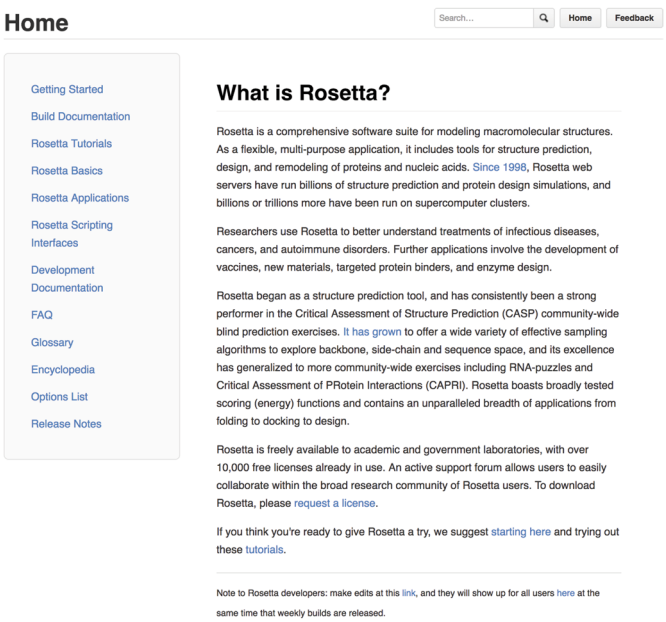

Fig. 5 | Main external documentation page. In 2015, our community performed a complete overhaul of our documentation. Documentation is now hosted on a Gollum wiki, which is version controlled and easily editable by members of our community. Accessibility and ability to edit the documentation has improved the user experience of the software.

fast custom protocol development, but requires familiarity with the underlying codebase. Other interfaces are InteractiveRosetta[183] and the gaming interface Foldit Standalone[184,185] (Supplementary Note).

We devoted an enormous effort to rewriting and adding documentation (Fig. 5). A public-facing Gollum wiki (https://www.rosettacommons.org/docs/latest/Home) houses various levels of documentation, such as application documentation, tutorials for beginning users, and static protocol captures that accompany manuscripts for scientific reproducibility (see Supplementary Note 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. Web servers 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 protocol web servers, ROSIE (Rosetta Online Server that Includes Everyone)[186,187] (http://rosie.rosettacommons.org/) implements a simple framework for 'serverification' of protocols. ROSIE currently contains 25 webservers, with additional protocols continually being added.

Conclusion

The Rosetta software is developed by a large, global community aiming to solve complex problems through real-time collaborative code development. In the last 5 years, great strides have been made in our software. More protocols 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 modeling has been facilitated and improved. Further, our community now develops 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[26] 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 and facilitate user support for the various interfaces (command line, RosettaScripts, PyRosetta, and so forth). Over the years, these applications have moved beyond tackling basic science questions (that is, the protein folding and design challenges) to more application-based scientific developments. The myriad 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, including most of the largest pharmaceutical companies), the 11 spin-off companies that were created from RosettaCommons[26], and the ever-increasing adoption by labs beyond those affiliated with RosettaCommons.

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 computation, influencing score function development and aiding in the development of therapeutic interventions[188]; restructuring the software for massively parallel computing architectures (for example, GPUs and TPUs) and quantum computers[189]; greater use of machine-learning (for example, deep-learning) approaches (for example, for score function development); modeling more realistic cellular environments; and improving user interfaces to make Rosetta accessible to more scientists. The predictive powers that we have reviewed above can be leveraged not only to analyze and verify existing data but also to inform experiments that will galvanize the engineering of industrial enzymes, enable the creation of novel biomaterials, and accelerate the discovery of potent new therapeutics.

Code availability.

Rosetta is licensed and distributed through https://www.rosettacommons.org. Licenses for academic, non-profit and government laboratories are free of charge; there is a license fee for industry users.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0848-2.

Received: 29 April 2019; Accepted: 22 April 2020;
Published online: 1 June 2020

References

1. Schrödinger. Biologics design. https://www.schrodinger.com/science-articles/biologics-design (2020).
2. Chemical Computing Group. Molecular Operating Environment (MOE) | MOEsaic | PSILO. https://www.chemcomp.com/Products.htm (2020).
3. Dassault Systèmes. BIOVIA, Discovery Studio Modeling Environment, release 2017. https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/ (2016).
4. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
5. 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).
6. Webb, B. et al. Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci.* **27**, 245–258 (2018).
7. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).

8. Brooks, B. R. et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
9. 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**, 1157–1174 (2004).
10. Van Der Spoel, D. et al. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
11. Eastman, P. et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
12. Senior, A. W. et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).
13. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
14. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).
15. Xu, J. & Wang, S. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins* **87**, 1069–1081 (2019).
16. Fiser, A. & Sali, A. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374**, 461–491 (2003).
17. Bienert, S. et al. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.* **45 D1**, D313–D319 (2017).
18. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
19. van Zundert, G. C. P. et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
20. Pierce, B. G. et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
21. Padhorny, D. et al. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl Acad. Sci. USA* **113**, E4286–E4293 (2016).
22. 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**, 455–461 (2010).
23. BioSolveIT GmbH. FlexX version 4.1. http://www.biosolveit.de/FlexX (2019).
24. Tubert-Brohman, I., Sherman, W., Repasky, M. & Beuming, T. Improved docking of polypeptides with Glide. *J. Chem. Inf. Model.* **53**, 1689–1699 (2013).
25. Sorenson, J. M. & Head-Gordon, T. Matching simulation and experiment: a new simplified model for simulating protein folding. *J. Comput. Biol.* **7**, 469–481 (2000).
26. Koehler Leman, J. et al. Better together: Elements of successful scientific software development in a distributed collaborative community. *PLoS Comput. Biol.* **16**, e1007507 (2020).
27. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
28. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
29. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
30. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
31. Fleishman, S. J. et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161 (2011).
32. Cooper, S. et al. Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
33. Bender, B. J. et al. Protocols for molecular modeling with Rosetta3 and RosettaScripts. *Biochemistry* https://doi.org/10.1021/acs.biochem.6b00444 (2016).
34. Simoncini, D. et al. Guaranteed discrete energy optimization on large protein design problems. *J. Chem. Theory Comput.* **11**, 5980–5989 (2015).
35. Leaver-Fay, A. et al. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523**, 109–143 (2013).
36. 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.* **118**, 11225–11236 (1996).
37. 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).
38. O'Meara, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
39. 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).
40. 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).
41. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011).
42. Mills, J. H. et al. Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy. *J. Am. Chem. Soc.* **135**, 13393–13399 (2013).
43. Kappel, K. et al. Blind tests of RNA-protein binding affinity prediction. *Proc. Natl Acad. Sci. USA* **116**, 8336–8341 (2019).
44. Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
45. Hosseinzadeh, P. et al. Comprehensive computational design of ordered peptide macrocycles. *Science* **358**, 1461–1466 (2017).
46. 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**, 1336–1341 (2007).
47. Boyken, S. E. et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).
48. Lu, P. et al. Accurate computational design of multipass transmembrane proteins. *Science* **359**, 1042–1046 (2018).
49. Chen, Z. et al. Programmable design of orthogonal protein heterodimers. *Nature* **565**, 106–111 (2019).
50. 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).
51. Pavlovicz, R.E., Park, H. & DiMaio, F. Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking. Preprint at *bioRxiv* https://doi.org/10.1101/618603 (2019).
52. 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).
53. König, R. & Dandekar, T. Solvent entropy-driven searching for protein modeling examined and tested in simplified models. *Protein Eng.* **14**, 329–335 (2001).
54. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—round XIII. *Proteins* https://doi.org/10.1002/prot.25823 (2019).
55. Song, Y. et al. High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
56. Robetta. http://new.robetta.org/ (2020).
57. Park, H., Kim, D. E., Ovchinnikov, S., Baker, D. & DiMaio, F. Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. *Proteins* **86**(Suppl. 1), 283–291 (2018).
58. 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. USA* **110**, 15674–15679 (2013).
59. Ovchinnikov, S. et al. Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).
60. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl Acad. Sci. USA* **115**, 3054–3059 (2018).
61. Tyka, M. D. et al. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).
62. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.* **380**, 757–774 (2008).
63. Kapp, G. T. et al. Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proc. Natl Acad. Sci. USA* **109**, 5277–5282 (2012).
64. Stein, A. & Kortemme, T. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* **8**, e63090 (2013).
65. Lin, M. S. & Head-Gordon, T. Improved energy selection of nativelike protein loops from loop decoys. *J. Chem. Theory Comput.* **4**, 515–521 (2008).
66. Rohl, C. A., Strauss, C. E. M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656–677 (2004).
67. Wang, C., Bradley, P. & Baker, D. Protein-protein docking with backbone flexibility. *J. Mol. Biol.* **373**, 503–519 (2007).
68. Canutescu, A. A. & Dunbrack, R. L. Jr. Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963–972 (2003).
69. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).

70. Mandell, D. J. & Kortemme, T. Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420–428 (2009).
71. 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).
72. Roy Burman, S. S., Yovanno, R. A. & Gray, J. J. Flexible backbone assembly and refinement of symmetrical homomeric complexes. *Structure* **27**, 1041–1051.e8 (2019).
73. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. *PLoS One* **6**, e20450 (2011).
74. Meiler, J. & Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **65**, 538–548 (2006).
75. 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).
76. Fu, D. Y. & Meiler, J. RosettaLigandEnsemble: a small-molecule ensemble-driven docking approach. *ACS Omega* **3**, 3655–3664 (2018).
77. 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).
78. 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).
79. 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).
80. Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* **37**, W474–W479 (2009).
81. 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* **82**, 1611–1623 (2014).
82. Weitzner, B. D. et al. Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* **12**, 401–416 (2017).
83. 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* **74**, 497–514 (2009).
84. 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).
85. Finn, J. A. et al. Improving loop modeling of the antibody complementarity-determining region 3 using knowledge-based restraints. *PLoS One* **11**, e0154811 (2016).
86. 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).
87. DeKosky, B. J. et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl Acad. Sci. USA* **113**, E2636–E2645 (2016).
88. Jeliazkov, J. R. et al. Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. *Front. Immunol.* **9**, 413 (2018).
89. 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).
90. Lapidoth, G., Parker, J., Prilusky, J. & Fleishman, S. J. AbPredict 2: a server for accurate and unstrained structure prediction of antibody variable domains. *Bioinformatics* **35**, 1591–1593 (2019).
91. 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).
92. 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**, 6357–6367 (2011).
93. Adolf-Bryfogle, J. et al. RosettaAntibodyDesign (RAbD): a general framework for computational antibody design. *PLoS Comput. Biol.* **14**, e1006112 (2018).
94. North, B., Lehmann, A. & Dunbrack, R. L. Jr. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406**, 228–256 (2011).
95. King, C. et al. Removing T-cell epitopes with computational protein design. *Proc. Natl Acad. Sci. USA* **111**, 8577–8582 (2014).
96. Nivón, L. G., Bjelic, S., King, C. & Baker, D. Automating human intuition for protein design. *Proteins* **82**, 858–866 (2014).
97. Lapidoth, G. D. et al. AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins* **83**, 1385–1406 (2015).
98. Baran, D. et al. Principles for computational design of binding antibodies. *Proc. Natl Acad. Sci. USA* **114**, 10900–10905 (2017).
99. Vaissier Welborn, V. & Head-Gordon, T. Computational design of synthetic enzymes. *Chem. Rev.* **119**, 6613–6630 (2019).
100. Marcos, E. & Silva, D.-A. Essentials of de novo protein design: methods and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1374 (2018).
101. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
102. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures. *Proc. Natl Acad. Sci. USA* **117**, 1059–1068 (2020).
103. Jacobs, T. M. et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
104. 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).
105. Lapidoth, G. et al. Highly active enzymes by automated combinatorial backbone assembly and sequence design. *Nat. Commun.* **9**, 2780 (2018).
106. Huang, P.-S. et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* **6**, e24109 (2011).
107. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A generic program for multistate protein design. *PLoS One* **6**, e20937 (2011).
108. Sevy, A. M., Jacobs, T. M., Crowe, J. E. Jr. & 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).
109. Sevy, A. M. et al. Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses. *Proc. Natl Acad. Sci. USA* **116**, 1597–1602 (2019).
110. Sauer, M. F., Sevy, A. M., Crowe, J. E. & Meiler, J. Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLoS Comput. Biol.* **16**, e1007339 (2020).
111. Correia, B. E. et al. Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206 (2014).
112. Bonet, J. et al. Rosetta FunFolDes — a general framework for the computational design of functional proteins. *PLoS Comput. Biol.* **14**, e1006623 (2018).
113. Kroncke, B. M. et al. Documentation of an imperative to improve methods for predicting membrane protein stability. *Biochemistry* **55**, 5002–5009 (2016).
114. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl Acad. Sci. USA* **99**, 14116–14121 (2002).
115. Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci. STKE* **2004**, pl2 (2004).
116. Conchúir, Ó. et al. Web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. *PLoS One* **10**, e0130433 (2015).
117. Barlow, K. A. et al. Flex ddG: Rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).
118. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **380**, 742–756 (2008).
119. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6**, 685–689 (1953).
120. Dang, B. et al. De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures. *Proc. Natl Acad. Sci. USA* **114**, 10852–10857 (2017).
121. Alam, N. et al. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS Comput. Biol.* **13**, e1005905 (2017).
122. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* **65**, 392–406 (2006).
123. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* **78**, 2029–2040 (2010).
124. Pacella, M. S., Koo, 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).
125. 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).
126. 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).
127. Wang, R. Y.-R. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016).
128. 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).
129. Frenz, B. et al. Automatically fixing errors in glycoprotein structures with Rosetta. *Structure* **27**, 134–139.e3 (2019).
130. Nerli, S. & Sgourakis, N. G. CS-ROSETTA. *Methods Enzymol.* **614**, 321–362 (2019).

131. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* **124**, 2723–2729 (2002).
132. Yagi, H. et al. Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. *Structure* **21**, 883–890 (2013).
133. Schmitz, C., Vernon, R., Otting, G., Baker, D. & Huber, T. Protein structure determination from pseudocontact shifts using ROSETTA. *J. Mol. Biol.* **416**, 668–677 (2012).
134. 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). 2 Pt B.
135. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* **80**, 884–895 (2012).
136. Bowers, P. M., Strauss, C. E. M. & Baker, D. De novo protein structure determination using sparse NMR data. *J. Biomol. NMR* **18**, 311–318 (2000).
137. Meiler, J. & Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl Acad. Sci. USA* **100**, 15404–15409 (2003).
138. Raman, S. et al. NMR structure determination for larger proteins using backbone-only data. *Science* **327**, 1014–1018 (2010).
139. Lange, O. F. et al. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl Acad. Sci. USA* **109**, 10873–10878 (2012).
140. Reichel, K. et al. Systematic evaluation of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins* **85**, 812–826 (2017).
141. Sgourakis, N. G. et al. Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc.* **133**, 6288–6298 (2011).
142. Rossi, P. et al. A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins* **83**, 309–317 (2015).
143. Demers, J.-P. et al. High-resolution structure of the *Shigella* type-III secretion needle by solid-state NMR and cryo-electron microscopy. *Nat. Commun.* **5**, 4976 (2014).
144. Thompson, J. M. et al. Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc. Natl Acad. Sci. USA* **109**, 9875–9880 (2012).
145. 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**, e1004661 (2015).
146. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
147. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59**, 147–159 (2014).
148. Kuenze, G., Bonneau, R., Koehler Leman, J. & Meiler, J. Integrative protein modeling in RosettaNMR from sparse paramagnetic restraints. *Structure* **27**, 1721–1734.e5 (2019).
149. 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).
150. Aprahamian, M. L. & Lindert, S. Utility of covalent labeling mass spectrometry data in protein structure prediction with Rosetta. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.9b00101 (2019).
151. Hauri, S. et al. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **10**, 192 (2019).
152. Watkins, A. M. et al. Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci. Adv.* **4**, eaar5316 (2018).
153. Sripakdeevong, P., Kladwang, W. & Das, R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl Acad. Sci. USA* **108**, 20573–20578 (2011).
154. Das, R. Atomic-accuracy prediction of protein loop structures through an RNA-inspired Ansatz. *PLoS One* **8**, e74830 (2013).
155. 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).
156. Chou, F.-C., Echols, N., Terwilliger, T. C. & Das, R. RNA structure refinement using the ERRASER-Phenix pipeline. in *Nucleic Acid Crystallography* 269–282 (Springer, 2016); https://doi.org/10.1007/978-1-4939-2763-0_17
157. Kappel, K. & Das, R. Sampling native-like structures of RNA-protein complexes through Rosetta folding and docking. *Structure* **27**, 140–151.e5 (2019).
158. Kappel, K. et al. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nat. Methods* **15**, 947–954 (2018).
159. Thyme, S. B. et al. Exploitation of binding energy for catalysis and design. *Nature* **461**, 1300–1304 (2009).
160. Ashworth, J. et al. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656–659 (2006).
161. Ashworth, J. et al. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38**, 5601–5608 (2010).
162. Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45–52 (2003).
163. Thyme, S. B. et al. Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Res.* **42**, 2564–2576 (2014).
164. 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).
165. Yanover, C. & Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* **39**, 4564–4576 (2011).
166. Ashworth, J. & Baker, D. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* **37**, e73 (2009).
167. Thyme, S. B. et al. Massively parallel determination and modeling of endonuclease substrate specificity. *Nucleic Acids Res.* **42**, 13839–13852 (2014).
168. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
169. Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins* **83**, 1–24 (2015).
170. Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins* **62**, 1010–1025 (2006).
171. Barth, P., Schonbrun, J. & Baker, D. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl Acad. Sci. USA* **104**, 15682–15687 (2007).
172. Alford, R. F. et al. An integrated framework advancing membrane protein modeling and design. *PLoS Comput. Biol.* **11**, e1004398 (2015).
173. Baugh, E. H., Lyskov, S., Weitzner, B. D. & Gray, J. J. Real-time PyMOL visualization for Rosetta and PyRosetta. *PLoS One* **6**, e21931 (2011).
174. Koehler Leman, J., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics* **33**, 754–756 (2017).
175. 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).
176. Koehler Leman, J. & Bonneau, R. A novel domain assembly routine for creating full-length models of membrane proteins from known domain structures. *Biochemistry* https://doi.org/10.1021/acs.biochem.7b00995 (2017).
177. 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).
178. Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein structure prediction and design in a biologically realistic implicit membrane. *Biophys. J.* **118**, 2042–2055 (2020).
179. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3**, 97–130 (1993).
180. Varki, A. et al. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2009).
181. 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).
182. Gray, J. J., Chaudhury, S., Lyskov, S. & Labonte, J. W. The PyRosetta interactive platform for protein structure prediction and design: a set of educational modules. (CreateSpace, 2014).
183. Schenkelberg, C. D. & Bystroff, C. InteractiveROSETTA: a graphical user interface for the PyRosetta protein modeling suite. *Bioinformatics* **31**, 4023–4025 (2015).
184. Kleffner, R. et al. Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. *Bioinformatics* **33**, 2765–2767 (2017).
185. 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. Foundations of Digital Games – FDG '18* https://doi.org/10.1145/3235765.3235770 (ACM Press, 2018).
186. Lyskov, S. et al. Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PLoS One* **8**, e63906 (2013).
187. 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).
188. Institute for Protein Design. Audacious Project. https://www.ipd.uw.edu/audacious/ (2019).
189. Mulligan, V.K. et al. Designing peptides on a quantum computer. Preprint at *bioRxiv* https://doi.org/10.1101/752485 (2019).
190. 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).
191. Marcos, E. et al. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).

192. DeLuca, S., Khar, K. & Meiler, J. Fully flexible docking of medium sized ligand libraries with RosettaLigand. *PLoS One* **10**, e0132508 (2015).
193. Davis, I. W. & Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* **385**, 381–392 (2009).
194. Gowthaman, R. et al. DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sites. *J. Med. Chem.* **59**, 4152–4170 (2016).
195. Khar, K. R., Goldschmidt, L. & Karanicolas, J. Fast docking on graphics processing units via Ray-Casting. *PLoS One* **8**, e70661 (2013).
196. Gowthaman, R., Lyskov, S. & Karanicolas, J. DARC 2.0: improved docking and virtual screening at protein interaction sites. *PLoS One* **10**, e0131612 (2015).
197. Toor, J. S. et al. A recurrent mutation in anaplastic lymphoma kinase with distinct neoepitope conformations. *Front. Immunol.* **9**, 99 (2018).
198. Gowthaman, R. & Pierce, B. G. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res.* **46 W1**, W396–W401 (2018).
199. 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* **86**, 354–369 (2018).
200. 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).
201. 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).
202. Sedan, Y., Marcu, O., Lyskov, S. & Schueler-Furman, O. Peptiderive server: derive peptide inhibitors from protein-protein interactions. *Nucleic Acids Res.* **44 W1**, W536–W541 (2016).
203. 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).
204. Pacella, M. S. & Gray, J. J. A benchmarking study of peptide–biomineral interactions. *Cryst. Growth Des.* **18**, 607–616 (2018).
205. Wang, R. Y.-R. et al. De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods* **12**, 335–338 (2015).
206. DiMaio, F. et al. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10**, 1102–1104 (2013).
207. DiMaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12**, 361–365 (2015).
208. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7**, 291–294 (2010).
209. Cheng, C. Y., Chou, F.-C. & Das, R. Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol.* **553**, 35–64 (2015).
210. Sripakdeevong, P. et al. Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. *Nat. Methods* **11**, 413–416 (2014).
211. Chou, F. C., Kladwang, W., Kappel, K. & Das, R. Blind tests of RNA nearest-neighbor energy prediction. *Proc. Natl Acad. Sci. USA* **113**, 8430–8435 (2016).
212. Ford, A. S., Weitzner, B. D. & Bahl, C. D. Integration of the Rosetta suite with the python software stack via reproducible packaging and core programming interfaces for distributed simulation. *Protein Sci.* **29**, 43–51 (2020).
213. Khatib, F. et al. Algorithm discovery by protein folding game players. *Proc. Natl Acad. Sci. USA* **108**, 18949–18953 (2011).
214. Hooper, W. F., Walcott, B. D., Wang, X. & Bystroff, C. Fast design of arbitrary length loops in proteins using InteractiveRosetta. *BMC Bioinformatics* **19**, 337 (2018).

Acknowledgements

RosettaCommons is supported by NIH R01 GM073151 to B. Kuhlman, NSF, the Packard Foundation, the Beckman Foundation, the Alfred P. Sloan Foundation and the Simons Foundation. This work was also supported by a 100,000,000 CPU-hour donation from Google Inc to P.C. and a 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 D.B., F.D., A.L.-F. and V.K.M. 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. Supported by AHA 18POST34080422 to G.K., AMED J-PRIDE JP18fm0208022h to D.K., the Biltema Foundation to B.E.C. and Boehringer Ingelheim Fonds to C.N.; 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 United States to P.C.; DFG KU 3510/1-1 to G.K.; DP120100561 to T.H.; DP150100383 to T.H. and K.B.P.; EMBO long-term fellowship ALTF 698-2011 to A. Stein; EPFL-Fellows H2020 Marie Sklodowska-Curie to J.B.; European Research Council Grant 310873 to O.S.-F. and N.A.; European Research Council Grant 310873 to Y. Sedan and O.M.; European Research Council Starting grant 716058 to B.E.C. and A. Scheck; FT0991709 to T.H.; Foundation of Knut and Alice Wallenberg 20160023 to L.M.; a Hertz Foundation Fellowship to R.F.A.; the Howard Hughes Medical Institute to D.B.; Hyak supercomputer system supported in part by the University of Washington eScience Institute to the D.B. and F.D. labs; Israel Science Foundation 2017717 to O.S.-F. and N.A.; Japan Society for the Promotion of Science JP17K18113 to D.K.; MCB1330760 to S.D.K.; Marie Curie International Outgoing Fellowship FP7-PEOPLE-2011-IOF 298976 to E.M.; National Science Centre, Poland, 2018/29/B/ST6/01989 to D.G.; NIAID T32AI007244 to J.A.-B.; NIAID U19 AI117905 to A.M.S.; NIEHS P42ES004699 to J.B.S.; NIGMS Ruth L Kirschstein National Research Service Award T32GM008268 to P.C.; NIGMS T32 GM007628 to B.J.B.; NIH 1R35 GM122579 to R. Das; NSF DMREF award 1728858 and DMR-0820341 to R.B.; NIH 1UH2CA203780 to S.C. and F.K.; NIH 5F32GM110899-02 to T.L.; NIH F31GM123616 to J.R.J.; NIH F32CA189246 to J.W.L.; NIH P01 U19AI117905, R01 AI113867 and UM1 Al100663 to W.S.; NIH R00 GM120388 to S.H.; NIH R01 AI143997 to N.G.S.; NIH R01 DK097376, R01 GM080403, R01 HL122010 and R01 GM099842 to J. Meiler; NIH R01 GM073960, R01 GM117968 and R01067553 to B. Kuhlman; NIH R01 GM076324 to J.B.S.; NIH R01 GM127578 and R01 GM078221 to J.J.G.; NIH R01 GM084453 to R. Dunbrack; NIH R01 GM088277 and R01 GM121487 to P. Bradley; NIH R01 GM092802, R01 GM092802, R01084433 and GM092802 to D.B.; NIH R01 GM098101, R01 GM110089 and R01 GM117189 to T.K.; NIH R01 GM099959 to J.K.; NIH R01 GM123089 to F.D.; NIH R01 GM126299 to B.G.P.; NIH R01 GM099827 to C.B.; NIH R01088277 to S.B.T.; NIH R21 AI121799 to J. Meiler; NIH R21 CA219847 and R21 GM102716 to R. Das; NIH R35 GM122517 to R. Dunbrack; NIH R35 GM125034 to N.G.S.; NIH RL1CA133832 to D.B.; NIH U19 AI117905 to J. Meiler; NIH/NCI Cancer Center support grant P30 CA006927 to J.K.; NSF 1507736 and NSF DMR 1507736 to J.J.G.; NSF 1627539, 1805510 and 1827246 to J.B.S.; NSF 1629879 to S.C.; NSF CHE 1305874, CISE 1629811 and CNS-1629811 to J. Meiler; NSF CHE 1750666 to S. Lindert; NSF DBI-1262182 and DBI-1564692 to T.K.; NSF GRF DGE-1433187 to A.R.; NSF Graduate Research Fellowships to R.F.A., K.K., B. Koepnick and S.B.T.; NSF MCB1330760 and MCB1716623 to S.D.K.; Open Philanthropy to B.C.; PhRMA Informatics Pre-Doctoral Fellowship U22879-001 to S.S.; a PhRMA Foundation Predoctoral Fellowship to D.Y.F.; RosettaCommons to L.G., A.R., F.D., S.C., A.W., M.S., C.G., K.B., R. Das, S.D.K., J. Koehler Leman and K.K.; Career Award at the Scientific Interface from Burroughs Wellcome Fund to S.E.B.; Simons Foundation to V.K.M., R.B., P.D.R. and J. Koehler Leman; a Stanford Graduate Fellowship to K.K.; a Starter Grant from the European Research Council to G.L.; Swiss National Science Foundation – NCCR Molecular Systems Engineering 51NF40-141825 to B.E.C.; Swiss National Science Foundation 310030_163139 to B.E.C.; Swiss National Science Foundation SNF 200021 160188 to L.M. and H.K.; UCSF/UCB Graduate Program in Bioengineering to X.P.; USA-Israel Binational Science Foundation 2009418 to B.R., L.Z. and N.L.; USA-Israel Binational Science Foundation 2009418 and 2015207 to O.S.-F. and N.A.; USA-Israel Binational Science Foundation 2015207 to A.K.; Washington Research Foundation Innovation Postdoctoral Fellowship to B.D.W.; XSEDE, which is supported by NSF ACI-1548562; NIH R01 GM097207 to P. Barth; and the MCB120101 XSEDE allocation to P. Barth. The authors would like to thank Jason C. Klima for his work on PyRosetta.

Author contributions

J.K.L. wrote the manuscript with help from B.D.W. 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 R.B.

Competing interests

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. D.B., L.M., D.G., J.M., O.S.-F., J.J.G., N.G.S., S.L., J.K., R.B., T.K. and P.B. are unpaid board members of the RosettaCommons. As members of the Scientific Advisory Board of Cyrus Biotechnology, D.B. and J.J.G. are granted stock options. Y.S., I.C.K., S.M.L., B.F., K.R.K. and R.E.P. are employed at Cyrus Biotechnology with granted stock options. Cyrus Biotechnology distributes the Rosetta software. B.D.W. and S.E.B. hold equity in Lyell Immunopharma. V.K.M. is a cofounder of and shareholder in Menten Biotechnology Labs, Inc. The content of this manuscript is relevant to work performed at Lyell and Menten. J.B.S. is a cofounder and shareholder of Digestiva, Inc. and PvP Biologics Inc. D.B. is a cofounder of, shareholder in, or advisor to the following companies: ARZEDA, PvP Biologics, Cyrus Biotechnology, Cue Biopharma, Icosavax, Neoleukin Therapeutics, Lyell Immunotherapeutics, Sana Biotechnology and A-Alpha Bio.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0848-2.

Correspondence should be addressed to J.K.L. or R.B.

Peer review information Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Julia Koehler Leman[1,2,63] ✉, Brian D. Weitzner[3,4,5,6,63], Steven M. Lewis[7,8,9,63], Jared Adolf-Bryfogle[10], Nawsad Alam[11], Rebecca F. Alford[3], Melanie Aprahamian[12], David Baker[4,5], Kyle A. Barlow[13], Patrick Barth[14,15], Benjamin Basanta[4,16], Brian J. Bender[17], Kristin Blacklock[18], Jaume Bonet[14,19], Scott E. Boyken[5,6], Phil Bradley[20], Chris Bystroff[21], Patrick Conway[4], Seth Cooper[22], Bruno E. Correia[14,19], Brian Coventry[4], Rhiju Das[23], René M. De Jong[24], Frank DiMaio[4,5], Lorna Dsilva[22], Roland Dunbrack[25], Alexander S. Ford[4], Brandon Frenz[5,9], Darwin Y. Fu[26], Caleb Geniesse[23], Lukasz Goldschmidt[4], Ragul Gowthaman[27,28], Jeffrey J. Gray[3,29], Dominik Gront[30], Sharon Guffy[7], Scott Horowitz[31,32], Po-Ssu Huang[4], Thomas Huber[33], Tim M. Jacobs[34], Jeliazko R. Jeliazkov[29], David K. Johnson[35], Kalli Kappel[36], John Karanicolas[25], Hamed Khakzad[19,37,38], Karen R. Khar[9,35], Sagar D. Khare[18,39,40,41,42], Firas Khatib[43], Alisa Khramushin[11], Indigo C. King[4,9], Robert Kleffner[22], Brian Koepnick[4], Tanja Kortemme[44], Georg Kuenze[26,45], Brian Kuhlman[7], Daisuke Kuroda[46,47], Jason W. Labonte[3,48], Jason K. Lai[15], Gideon Lapidoth[49], Andrew Leaver-Fay[7], Steffen Lindert[12], Thomas Linsky[4,5], Nir London[11], Joseph H. Lubin[3], Sergey Lyskov[3], Jack Maguire[34], Lars Malmström[19,37,38,50], Enrique Marcos[4,51], Orly Marcu[11], Nicholas A. Marze[3], Jens Meiler[45,52,53], Rocco Moretti[26], Vikram Khipple Mulligan[1,4,5], Santrupti Nerli[54], Christoffer Norn[49], Shane Ó'Conchúir[44], Noah Ollikainen[44], Sergey Ovchinnikov[4,5,55], Michael S. Pacella[3], Xingjie Pan[44], Hahnbeom Park[4], Ryan E. Pavlovicz[4,5,9], Manasi Pethe[40,41], Brian G. Pierce[27,28], Kala Bharath Pilla[33], Barak Raveh[11], P. Douglas Renfrew[1], Shourya S. Roy Burman[3], Aliza Rubenstein[18,42], Marion F. Sauer[56], Andreas Scheck[14,19], William Schief[10], Ora Schueler-Furman[11], Yuval Sedan[11], Alexander M. Sevy[56], Nikolaos G. Sgourakis[57], Lei Shi[4,5], Justin B. Siegel[58,59,60], Daniel-Adriano Silva[4], Shannon Smith[26], Yifan Song[4,5,9], Amelie Stein[44], Maria Szegedy[39], Frank D. Teets[7], Summer B. Thyme[4], Ray Yu-Ruei Wang[4], Andrew Watkins[23], Lior Zimmerman[11] and Richard Bonneau[1,2,61,62] ✉

[1]Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, USA. [2]Department of Biology, New York University, New York, New York, USA. [3]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Department of Biochemistry, University of Washington, Seattle, WA, USA. [5]Institute for Protein Design, University of Washington, Seattle, WA, USA. [6]Lyell Immunopharma Inc., Seattle, WA, USA. [7]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [8]Department of Biochemistry, Duke University, Durham, NC, USA. [9]Cyrus Biotechnology, Seattle, WA, USA. [10]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA. [11]Department of Microbiology and Molecular Genetics, IMRIC, Ein Kerem Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel. [12]Department of Chemistry and Biochemistry, Ohio State University, Columbus, OH, USA. [13]Graduate Program in Bioinformatics, University of California San Francisco, San Francisco, CA, USA. [14]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [15]Baylor College of Medicine, Department of Pharmacology, Houston, TX, USA. [16]Biological Physics Structure and Design PhD Program, University of Washington, Seattle, WA, USA. [17]Department of Pharmacology, Vanderbilt University, Nashville, TN, USA. [18]Institute of Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [19]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [20]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [21]Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA. [22]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [23]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [24]DSM Biotechnology Center, Delft, the Netherlands. [25]Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, PA, USA. [26]Department of Chemistry, Vanderbilt University, Nashville, TN, USA. [27]University of Maryland Institute for Bioscience and Biotechnology Research, Rockville, MD, USA. [28]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA. [29]Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD, USA. [30]Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland. [31]Department of Chemistry & Biochemistry, University of Denver, Denver, CO, USA. [32]The Knoebel Institute for Healthy Aging, University of Denver, Denver, CO, USA. [33]Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia. [34]Program in Bioinformatics and Computational Biology, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [35]Center for Computational Biology, University of Kansas, Lawrence, KS, USA. [36]Biophysics Program, Stanford University, Stanford, CA, USA. [37]Institute for Computational Science, University of Zurich, Zurich, Switzerland. [38]S3IT, University of Zurich, Zurich, Switzerland. [39]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [40]Department of Chemistry and Chemical Biology, The State University of New Jersey, Piscataway, NJ, USA. [41]Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [42]Computational Biology and Molecular Biophysics Program, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [43]Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA, USA. [44]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [45]Center for Structural Biology, Vanderbilt University, Nashville, TN, USA. [46]Medical Device Development and Regulation Research Center, School of Engineering, University of Tokyo, Tokyo, Japan. [47]Department of Bioengineering,

School of Engineering, University of Tokyo, Tokyo, Japan. [48]Department of Chemistry, Franklin & Marshall College, Lancaster, PA, USA. [49]Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel. [50]Division of Infection Medicine, Department of Clinical Sciences Lund, Faculty of Medicine, Lund University, Lund, Sweden. [51]Institute for Research in Biomedicine Barcelona, The Barcelona Institute of Science and Technology, Barcelona, Spain. [52]Departments of Chemistry, Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [53]Institute for Chemical Biology, Vanderbilt University, Nashville, TN, USA. [54]Department of Computer Science, University of California Santa Cruz, Santa Cruz, CA, USA. [55]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA. [56]Chemical and Physical Biology Program, Vanderbilt Vaccine Center, Vanderbilt University, Nashville, TN, USA. [57]Department of Chemistry and Biochemistry, University of California Santa Cruz, Santa Cruz, CA, USA. [58]Department of Chemistry, University of California, Davis, Davis, CA, USA. [59]Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, California, USA. [60]Genome Center, University of California, Davis, Davis, CA, USA. [61]Department of Computer Science, New York University, New York, NY, USA. [62]Center for Data Science, New York University, New York, NY, USA. [63]These authors contributed equally: Julia Koehler Leman, Brian D. Weitzner, Steven M. Lewis. ✉e-mail: Julia.koehler.leman@nyu.edu; Bonneau@nyu.edu