Cite as: R. Krishna *et al.*, *Science*
10.1126/science.adl2528 (2024).

Generalized biomolecular modeling and design with RoseTTAFold All-Atom

Rohith Krishna[1,2]†, Jue Wang[1,2]†, Woody Ahern[1,2,3]†, Pascal Sturmfels[1,2,3], Preetham Venkatesh[1,2,4]‡, Indrek Kalvet[1,2,5]‡, Gyu Rie Lee[1,2,5]‡, Felix S. Morey-Burrows[6], Ivan Anishchenko[1,2], Ian R. Humphreys[1,2], Ryan McHugh[1,2,4], Dionne Vafeados[1,2], Xinting Li[1,2], George A. Sutherland[6], Andrew Hitchcock[6], C. Neil Hunter[6], Alex Kang[2], Evans Brackenbrough[2], Asim K. Bera[2], Minkyung Baek[7], Frank DiMaio[1,2], David Baker[1,2,5]*

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98105, USA. [3]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA. [4]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA. [5]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA. [6]School of Biosciences, University of Sheffield, Sheffield S10 2TN, UK. [7]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea.

***Corresponding author. Email: dabaker@uw.edu**

†These authors contributed equally to this work.

‡These authors contributed equally to this work.

Deep learning methods have revolutionized protein structure prediction and design but are currently limited to protein-only systems. We describe RoseTTAFold All-Atom (RFAA) which combines a residue-based representation of amino acids and DNA bases with an atomic representation of all other groups to model assemblies containing proteins, nucleic acids, small molecules, metals, and covalent modifications given their sequences and chemical structures. By fine tuning on denoising tasks we obtain RFdiffusionAA, which builds protein structures around small molecules. Starting from random distributions of amino acid residues surrounding target small molecules, we design and experimentally validate, through crystallography and binding measurements, proteins that bind the cardiac disease therapeutic digoxigenin, the enzymatic cofactor heme, and the light harvesting molecule bilin.

The deep neural networks AlphaFold2 (AF2) (*1*) and RoseTTAFold (RF) (*2*) enable high-accuracy prediction of protein structures from amino acid sequence. However, in nature, proteins rarely act alone; they form complexes with other proteins in cell signaling, interact with DNA and RNA during transcription and translation, and interact with small molecules both covalently and noncovalently during metabolism. Modeling such general biomolecular assemblies composed of polypeptide chains, covalently modified amino acids, nucleic acid chains, and arbitrary small molecules remains an outstanding challenge. One approach is to model the protein chains using AF2 or RF, and then successively add in the non-protein components using classical docking methods (*3*–*9*); however, systematically evaluating and optimizing such procedures is not straightforward. RF has been extended to model both protein and nucleic acids by increasing the size of the residue alphabet to 28 (20 amino acids, four DNA bases, and four RNA bases) with RoseTTAFold nucleic acid (RFNA) (*10*), but general biomolecular system modeling is a more challenging problem given the great diversity of possible small molecule components. An approach capable of accurately predicting the three-dimensional structures of biomolecular assemblies starting only from knowledge of the constituent molecules (and not their 3D structures) would have broad impact on structural biology and drug discovery and open the door to deep learning-based design of protein-small molecule assemblies.

We set out to develop a structure prediction method capable of generating 3D coordinates for all atoms of a biological unit, including proteins, nucleic acids, small molecules, metals, and chemical modifications (Fig. 1A). The first obstacle we faced in taking on the broader challenge of generalized biomolecular system modeling was how to represent the components. Existing protein structure prediction networks represent proteins as linear chains of amino acids, and this representation can be readily extended to nucleic acids. However, many of the small molecules that proteins interact with are not polymers, and it is unclear how to model them as a linear sequence. A natural way to represent the bonded structure of small molecules is as graphs whose nodes are atoms and whose edges represent bond connectivity. This graph representation is not suitable for proteins as they contain many thousands of atoms; hence, modeling whole proteins at the atomic level is computationally intractable. To overcome this limitation, we sought to combine a sequence-based description of biopolymers (proteins and nucleic acids) with an atomic graph representation of small molecules and protein covalent modifications.

Generalizing structure prediction to all biomolecules

We modeled the network architecture after the RoseTTAFold2 (RF2) protein structure prediction network, which accepts 1D sequence information, 2D pairwise distance information from homologous templates, and 3D coordinate information and iteratively improves predicted structures through many hidden layers (*11*). We retain the representations of protein and nucleic acid chains from RF2 and represent arbitrary small molecules, covalent modifications, and unnatural amino acids as atom-bond graphs. To the 1D track, we input the chemical element type of each non-polymer atom; to the 2D track, the chemical bonds between atoms; and to the 3D track, information on chirality [whether chiral centers are (*R*) or (*S*)]. For the 1D track, we supplement the 20 residue and eight nucleic acid base representation in RFNA with 46 new element type tokens representing the most common element types found in the Protein Data Bank (PDB) (table S5). For the 2D track atom-bond embedding, we encode pairwise information about whether bonds between pairs of atoms are single, double, triple, or aromatic bonds. These features are linearly embedded and summed with the initial pair features at the beginning of every recycle of the network, allowing the network to learn about bond lengths, angles, and planarity. Since the 1D and 2D representations in the network are invariant to reflections, we encode stereochemistry information in the third track by specifying the sign of angles between the atoms surrounding each chiral center (fig. S1); at each block in the 3D track the gradient of the deviation of the actual angles from the ideal values (with respect to the current coordinates) is computed and provided as an input feature to the subsequent block (Fig. 1B).

Unlike proteins and nucleic acid sequences, molecular graphs are permutation invariant, and hence, the network should make the same prediction irrespective of small molecule element token order. In AF2 and RF2, the sequence order of amino acids and bases is represented by a relative position encoding; for atoms, we omit such an encoding and leverage the permutation invariance of the network attention mechanisms. We also modify the coordinate updates: in AF2 and RF, protein residues are represented by the coordinates of the Cα and the orientation of the N-C-C rigid frame (or the P coordinate and the OP1-P-OP2 frame orientation in RFNA) and along the 3D track the network generates rotational updates to each frame orientation and translational updates to each coordinate. To generalize this representation in RFAA, heavy atom coordinates are added to the 3D track and move independently based only on a predicted translational update to their position. Thus, immediately after input, the full system is represented as a disconnected gas of amino acid residues, nucleic acid bases, and freely moving atoms, which is successively transformed through the many blocks of the network into physically plausible assembly structures. For the loss function to guide parameter optimization, we develop an all-atom version of the Frame Aligned Point Error (FAPE) loss introduced in AF2 by defining coordinate frames for each atom in an arbitrary molecule based on the identities of its bonded neighbors and, as with residue based FAPE, successively aligning each coordinate frame and computing the coordinate error on the surrounding atoms (Fig. 2A; for greater sensitivity to small molecule geometry, we upweight contributions involving atoms; see supplementary methods). In addition to atomic coordinates, the network predicts atom and residue-wise confidence (pLDDT) and pairwise confidence (PAE) metrics to enable users to identify high-quality predictions. A full description of the RFAA architecture is provided in the supplementary methods.

Training RFAA

We curated a protein-biomolecule dataset from the PDB including protein-small molecule, protein-metal, and covalently modified protein complexes, filtering out common solvents and crystallization additives. Following clustering (30% sequence identity) to avoid bias toward overrepresented structures, we obtained 121,800 protein-small molecule structures in 5,662 clusters, 112,546 protein-metal complexes in 5,324 clusters, and 12,689 structures with covalently modified amino acids in 1,099 clusters for training. To help the network learn the general properties of small molecules rather than features specific to the molecules in the PDB, we supplemented the training set with small molecule crystal structures from the Cambridge Structural Database (*12*). Each training example is sampled uniformly from the set of organic non-polymeric molecules, and the network predicts the coordinates for the asymmetric unit given atomic graph information. To further help the network learn about general atomic interactions, we take advantage of the commonalities between atomic interactions within proteins and many of the atomic interactions between proteins and small molecules and augment the training data by inputting portions of proteins as atoms rather than residues (a process we term *atomization*). We atomize randomly selected subsets of three to five contiguous residues by deleting the sequence and template features and providing instead atom, bond, and chirality information for the atoms in those residues (an alanine would be replaced by five atom tokens, one for each heavy atom). Since the atoms are still part of the polypeptide chain, we provide the relative position of the atom tokens with respect to the other residue tokens by adding an extra bond token that corresponds to an "atom-to-residue" bond and develop a positional encoding to account for atom-residue bonds (see supplementary methods). To increase prediction accuracy on biological polymers, we train the network on protein monomer, protein complex, and protein-nucleic acid complex examples as previously described (*10*, *11*). All examples were cropped to have 256 tokens during the initial stages of training and 375 tokens during fine-tuning. The progress of training was monitored using independent validation sets consisting of 10% of the protein sequence clusters (see table S4).

Unlike previous protein-only deep learning architectures (*13*–*15*), RFAA can model full biomolecular systems. In the following sections, we describe the performance of RFAA on different structure modeling tasks. We adopted the philosophy that a single model trained on all available data over all modalities would have the greatest ability to generalize and be more accessible than a series of models specialized for specific problems.

Predicting protein-small molecule complexes

To enable blind testing of RFAA prediction performance, we enrolled an RFAA server in the blind CAMEO ligand docking evaluation, which carries out predictions on all structures submitted to the PDB each week with each enrolled server and evaluates their performance (*16*–*18*). These structures can have multiple protein chains, ligands, and metal ions (for further results on metal ions, see fig. S2). Of the CAMEO targets, 43% are predicted confidently by RFAA (PAE Interaction < 10), and 77% of those high-confidence structures are quite accurate, with < 2 Å ligand RMSD (Fig. 2B). One of the other servers is an implementation of a leading non-deep learning protein small molecule docking method AutoDock Vina by the CAMEO organizers that predicts the protein structure by homology modeling (*19*–*24*), runs AutoDock to dock the small molecules, and ranks the poses using the Vina scoring function (*9*, *19*). RFAA consistently outperformed the other servers in CAMEO on protein-small molecule modeling; for example, on cases modeled by both the RFAA and the AutoDock Vina servers, RFAA models 32% of cases successfully (< 2 Å ligand RMSD) compared to 8% for the Vina server (Fig. 2B; the Vina performance by an expert would likely be considerably improved because of the complexities of fully automatic multiple step modeling pipelines). The most common RFAA failure mode is the placement of small molecules in the correct pockets but not in the correct orientation (fig. S3; for further exploration of failure modes, see supplementary methods).

There were no other deep learning docking methods (*5*, *25*–*29*) enrolled in CAMEO, but we can instead compare performance on a set of PDB structures that were solved after our training set date cutoff (*30*) (most earlier deep learning based docking tools have focused on the "bound" docking problem where the crystal structure of the target (including sidechains) are provided, and hence are less well suited to CAMEO). On this benchmark, RFAA predicts 42% of complexes successfully compared to DiffDock, which predicts 38% of complexes successfully (Fig. 2D; RFAA predicts the protein backbone and side chains in addition to the small molecule dock, whereas DiffDock receives the crystal structure of the protein from the bound complex as input). In cases where both the bound protein structure and the pocket residues are provided, physics-based methods such as AutoDock Vina outperform RFAA (52% vs 42%), which has the much harder task of predicting both the protein backbone and sidechain details and the dock from sequence alone (fig. S4A).

To further benchmark the network, we assembled a dataset of recent PDB entries with small molecules bound that were deposited after the cutoff date for our training set and predicted full structure models for all 5,421 complexes (1,529 protein sequence clusters at 30% sequence identity). The network performs better for clusters with overlap with the training set, but also generates accurate predictions for proteins with low (BLAST e-value > 1) sequence similarity to the training set (35% vs. 24% success rate, respectively; Fig. 2F). We observe a similar pattern for ligand clusters (across 1,310 ligand clusters); whereas the network makes more accurate predictions for ligands seen in training, it also can make accurate predictions on ligands that are not similar to those in training (<0.5 Tanimoto similarity; 19% vs. 14% success rate) (Fig. 2F). In cases where RFAA predicts ligand placement with high confidence and RF2 has high confidence (PAE Interaction <10 and pLDDT >0.8 respectively), RFAA makes higher accuracy protein structure predictions than RF2 (fig. S3A), indicating that training with ligand context can improve overall protein prediction accuracy. Some examples of shifts predicted by RFAA but not by RF2 include domain movements, subtle backbone movements, and flipping of side chain rotamers to accommodate the ligand in the pocket (fig. S3, B and C).

Unlike previous methods, RFAA is able to jointly predict interactions between proteins and multiple non-protein ligands in a single forward pass. Figure 2D shows three examples of recently solved structures with three or more components for which RFAA predictions had <2 Å ligand RMSD (when the proteins are aligned). There are homologous complexes in the training set so these are not de novo predictions, but they demonstrate that RFAA can learn the multicomponent assembly prediction task. The right panel shows a prediction for DNA polymerase (*31*) (PDB ID: 7u7w) with a bound DNA, non-hydrolyzable guanine triphosphate and magnesium ion; the network received no examples of higher order assemblies containing proteins with both small molecules and nucleic acids during training, but is likely synthesizing information from multiple related binary complexes that are in the training set.

To assess whether the network can distinguish compounds known to bind from related compounds, we compared protein-small molecule complex predictions for the PoseBusters dataset for the compound known to bind and decoy molecules including small molecules with the highest Tanimoto similarity in the dataset. In 75.1% of cases the PAE interaction metric of the "decoy" complex was higher (indicating lower confidence) than the native complex (fig. S7). Direct optimization on this discrimination task would likely further improve performance.

To determine the extent to which the network is reasoning over the detailed structure of protein-small molecule interactions, we investigated the correlation between prediction accuracy and the interaction energy computed by a molecular force field. We found that predictions for protein-small molecule complexes in our recent PDB set with lower computed binding energies (by Rosetta ΔG) (*32*, *33*) were more accurate (Fig. 2G; 50%, 25%, and 22% success rates for <-30, -30-0, and >0 Rosetta Energy Units, respectively) suggesting the network considers the detailed interactions between the protein and small molecule (although reasoning over these interactions very differently than human designed force fields).

Predicting structures of covalent modifications to proteins

Many essential protein functions, such as receptor signaling, immune evasion, and enzyme activity, involve covalent modifications of amino acid side chains with sugars, phosphates, lipids, and other molecules (*34*–*37*). RFAA models such modifications by treating the residue and chemical moiety as atoms (with the corresponding covalent bond to the atom token in the residue) and the rest of the protein structure as residues (Fig. 3A). Unnatural amino acids can be modeled in the same way.

We benchmarked the performance of RFAA on covalent modification structure prediction on 931 recent entries in the PDB (post-May, 2020), and found that the network made accurate predictions (Modification RMSD<2.5 Å) in 46% of cases (Modification RMSD is the RMSD of the modified residue and chemical modification when the rest of the protein is aligned). As in the protein-small molecule complex case, confident predictions tend to be more accurate: 60% of structures are predicted with high confidence (PAE Interaction <10), and 63% of those predictions are accurate (<2.5 Å modification RMSD) (Fig. 3B). Although the network makes slightly more accurate predictions on cases with sequence similarity (>25% identity) to proteins in the training set, there are still many cases (27.5%) that do not have sequence overlap to the training set that are predicted with high accuracy (Fig. 3C). RFAA models interactions with covalently bound cofactors and covalently bound drugs with median RMSDs of 0.99 Å and 2.8 Å respectively (Fig. 3, D and E).

Prediction of glycan structure has applications in therapeutics, vaccines, and diagnostics (*38*–*40*). RFAA can accurately model carbohydrate groups introduced by glycosylations with a median RMSD over our test set of 3.2 Å (Fig. 3D). RFAA successfully predicts glycan conformations on the N-acetylglucosamine-1-phosphotransferase (GNPT) gamma subunit (PDB ID: 7s69), and human sperm TMEM ectodomain (PDB ID: 7ux0), which have low sequence homology (<30%) to the RFAA training set (Fig. 3F) and have multiple monosaccharides and different branching patterns (*41*, *42*). RFAA is not simply learning how structure building programs model glycans as the predictions match the experimental density maps (fig. S8C). The network is able to make accurate predictions of glycan interactions even when the sequences were distant from the sequences in the training set, and on glycans with chains up to seven monosaccharides (fig. S8).

It is difficult to compare to other methods because, to our knowledge, previous deep learning-based tools do not model covalent modifications to proteins. Accurate and robust modeling of covalent modifications in predicted structures should contribute to the understanding of biological function and mechanism.

De novo small molecule binder design

Previous work on small molecule binding protein design has involved docking molecules into large sets of native or expert-curated protein scaffold structures (*43*, *44*). Diffusion based methods can generate proteins in the context of a protein target that bind with considerable affinity and specificity (*45*) and can be trained to explicitly condition on structural features (*46*). However, current deep learning based generative approaches do not explicitly model protein-ligand interactions, so they are not directly applicable to the small molecular binder design problem [in RFdiffusion, a heuristic attractive-repulsive potential encouraged the formation of pockets with shape complementarity to a target molecule, but the approach was unable to model the details of protein-small molecule interactions (*45*)]. A general method that can generate protein structures around small molecules and other non-protein targets to maximize favorable interactions could be broadly useful.

We reasoned that RFAA could enable protein design in the context of non-protein biomolecules following fine-tuning on structure denoising. We developed a diffusion model, RFdiffusion All-Atom (RFdiffusionAA), by training a denoising diffusion probabilistic model (DDPM) initialized with the RFAA structure-prediction weights to denoise corrupted protein structures conditioned on the small molecule and other biomolecular context (Fig. 4A). Input structures from the protein-small molecule dataset described above were noised through progressive addition of 3D Gaussian noise to the Cα coordinates and Brownian motion on the manifold of rotations, and the model was trained to remove this noise. In contrast to training for the unconditional generation problem and incorporating conditional information through forms of guidance (*47*, *48*), this training procedure results in an explicitly conditional model that learns the distribution of proteins conditioned on biomolecular substructure. To enable the inclusion of specific protein functional motifs when desired, we also train the network to scaffold a variety of discontiguous protein motifs both in the presence and absence of small molecules. To generate proteins, we initialize a Gaussian distribution of residue frames with randomized rotations around a fixed small molecule motif; at each denoising step t, we predict the fully denoised X_0 state and then update all residue coordinates and orientations by taking a step toward this conformation while adding noise to match the distribution for X_{t-1}. As with RFdiffusion, we investigated the use of auxiliary potentials to influence trajectories to make more contacts between small molecules and binders but found these unnecessary (see fig. S10C).

We evaluated RFdiffusionAA in silico by generating protein structures in the context of four diverse small molecules. Starting from random residue distributions surrounding each of the small molecules, iterative denoising yielded coherent protein backbones with pockets complementary to the small molecule target. Following sequence design using LigandMPNN (*49*, *50*), Rosetta GALigandDock (*32*) energy calculations were used to evaluate the protein-small molecule interface and AF2 predictions to evaluate the extent the sequence encodes the designed structure (*45*, *51*). The computed binding energies of RFdiffusionAA designs are far better ($p<1.56E-12$) than those obtained using a heuristic attractive/repulsive potential with protein-only RFdiffusion (fig. S10C). AF2 structure predictions had backbone RMSD < 2 Å to the RFdiffusionAA design models in all cases (fig. S10C). For each small molecule, RFdiffusionAA generates diverse protein structural solutions to the binding problem that differ from native binders to these ligands (figs. S11 and S12).

Experimental characterization of designed binders

To experimentally evaluate RfdiffusionAA across a range of design scenarios, we designed binders for three diverse small molecules: one with no protein motif included in the design parameters, one with a single residue protein motif, and one with a four-residue protein motif (Fig. 4). The proteins were produced in *E. coli*, and ligand binding was measured experimentally.

Digoxigenin (DIG) is the aglycone of digoxin, a small molecule used to treat heart diseases with a narrow therapeutic window (*52*), and digoxigenin-binding proteins could help reduce toxicity (*53*). Previous attempts to design digoxigenin-binding proteins relied on protein scaffolds with experimentally determined structures and prespecified binding pockets and interacting motifs (*54*). We used RFdiffusionAA to design digoxigenin-binding backbones without any prior assumption about the protein-ligand interface or backbone structure (Fig. 4A). Sequences were obtained using LigandMPNN and Rosetta FastRelax (*55*) and 4,416 designs were selected based on consistency with AF2 predictions and Rosetta metrics (see supplementary methods). Experimental characterization identified several DIG-binding proteins (figs. S29 and S30 and supplementary methods); the highest affinity binder has a 343 nM K_d for free digoxigenin (measured by isothermal titration calorimetry; Fig. 4B) and is stable at temperatures up to 95°C.

Heme is a cofactor for a wide range of oxidation reactions and oxygen transport (cytochrome P450 and hemoglobin are two notable examples), with catalytic function enabled by pentacoordinate iron binding and an open substrate pocket (*56*, *57*). Designed heme-binding proteins with these features have considerable potential as a platform for the development of new enzymes (*58*). We diffused proteins around heme with the central iron coordinated by a cysteine and a placeholder molecule just above the porphyrin ring to keep the axial heme binding site open for potential substrate molecules. Of 168 designs selected based on AF2 predicted confidence (pLDDT), backbone RMSD to design, and RMSD of the predicted cysteine rotamer to the design, 135 were well expressed in *E. coli*, and 90 had UV/Vis spectra consistent with Cys-bound heme (as judged by the Soret maximum wavelength after in vitro heme loading) (*59*). We further purified 40 of the designs and found that 33 were monomeric and retained heme-binding through size exclusion chromatography (SEC). For 26 of the designs, we mutated the putative heme-coordinating cysteine residue to alanine which led to a notable change in the Soret features in all cases (Fig. 4 and figs. S13 to S16). Twenty designs exhibit high thermostability, retaining their heme binding at temperatures above 85°C, and do not unfold at temperatures up to 95°C (Fig. 4C and figs. S13 to S16). We solved the crystal structure of heme-loaded design HEM_3.C9 to 1.8 Å resolution (PDB ID: 8vc8) and found it to closely match the design model (0.86 Å Cα RMSD). The crystal structure verifies that heme is bound through Cys-ligation in a pentacoordinate fashion with an open distal pocket (in agreement with spectroscopic data) and is further held in place with hydrogen bonds to two arginines, as designed (fig. S17).

Bilins are brilliantly colored pigments that play important roles across diverse biological kingdoms. When bilins are constrained by protein scaffolds, such as phycobiliproteins in the megadalton phycobilisome antenna complexes of cyanobacteria and some algae (*60*), their absorption features narrow, their extinction coefficients increase, and their fluorescence is dramatically enhanced. We sampled diffusion trajectories conditioned on the structure of a bilin molecule attached to a four residue peptide corresponding to a motif recognized by the CpcEF bilin lyase (*61, 62*). We evaluated 94 designs with a whole cell screen using phycoerythrobilin (PEB) as the chromophore and identified nine proteins dissimilar to each other and to CpcA (fig. S18A) that bind bilin based on pigmentation or fluorescence (a 9.6% hit rate). We purified three designs - BIL_C11, BIL_H4, and BIL_F9 - with absorption maxima at 573, 605, and 607 nm compared to 557 nm for the CpcA-PEB [Fig. 4D and fig. S8B; the extent of red shifting correlates with computed electrostatic potential around the chromophore (fig. S19)]. Conformationally restricted bilins typically display higher fluorescence yields, absolute fluorescent yields for the BIL_C11, BIL_H4, and BIL_F9 designs are 38%, 11% and 25%, respectively, based on an earlier determination of the absolute fluorescence quantum yield for CpcA-PEB of 67% (*63*) (fig. S18C). These values are much higher than obtained previously with maquette scaffolds (FΦ values of 2-3%), which displayed limited bilin incorporation and less pronounced spectral enhancements (*64*). The strong coloration, absorption and emission for these designs were absent from control *E. coli* strains that synthesize only the PEB bilin and the CpcE/F lyase, or PEB, CpcE/F and maltose binding protein (fig. S20). The 34/30 nm range in absorption/emission covered by just one design round using a single chromophore raises the exciting prospect of tailoring the spectral profiles of designed biliproteins by manipulating the conformational flexibility of the bilin and the protein microenvironment. De novo designed antenna complexes could harvest light over a wider range of the UV-visible spectrum to enhance photosynthetic energy capture and conversion (*65*), and fluorescent reporter probes with tunable excitation/emission maxima would be useful biochemical tools.

The experimental validation of digoxigenin, heme and bilin binding proteins demonstrates that RFdiffusionAA can readily generate novel proteins with custom binding pockets for diverse small molecules. Unlike prior methods that rely on redesigning existing scaffolds, RFdiffusionAA builds proteins from scratch around the target compound, resulting in high shape-complementary in the binding pockets and reducing the need for expert knowledge. The ability of RFdiffusionAA to generalize is highlighted by the sequence and structural dissimilarity between the designs and proteins in the PDB that bind related molecules (Tanimoto similarity > 0.5); the most similar protein in the PDB that binds a related molecule has a TMscore of 0.59 for the highest affinity digoxigenin binder, less than 0.62 for all the characterized heme binders, and less than 0.52 for the bilin binders (fig. S21). In all cases there is no detectable sequence similarity to any known protein.

Discussion

RoseTTAFold All-Atom (RFAA) demonstrates that a single neural network can be trained to accurately model a wide range of general biomolecular assemblies containing a wide diversity of non-protein components. RFAA can make high-accuracy predictions on protein-small molecule complexes, with 32% of CAMEO targets predicted under 2 Å RMSD, and for covalent modifications to proteins, predicting 46% of recently solved covalent modifications under 2.5 Å RMSD, and generate accurate models for complexes of proteins with two or more non-protein molecules (small molecules, metals, nucleic acids, etc.). Training on more extensive datasets and/or architectural improvements will likely be necessary to generate consistently accurate predictions for protein-small molecule complexes on par with the accuracy deep networks can achieve on protein systems alone. The new prediction capabilities do not come at the expense of performance on the classic protein structure prediction problem: RFAA achieves similar protein structure prediction accuracy as AF2 (median GDT of 85 vs. 86) and protein-nucleic acid complex accuracy as RFNA (median allatom-LDDT of 0.74 vs. 0.78) (fig. S22).

Our prediction and design results suggest that RFAA has learned detailed features of protein-small molecule complexes. First, the network is able to make high-accuracy predictions for protein sequences and ligands that differ considerably from those in the training dataset (Figs. 2F and 3C), and prediction accuracy is higher for complexes with more favorable computed interaction energies using the Rosetta physically based model (Fig. 2G). Second, our RFdiffusionAA-generated bilin, heme, and digoxigenin binders have very different structures than proteins that bind these compounds in the PDB. RFAA should be immediately useful for modeling protein-small molecule complexes, in particular multicomponent biomolecular assemblies for which there are few or no alternative methods available, and for designing small molecule binding proteins and sensors.

Methods summary

A detailed description of the dataset curation, modeling of biological inputs, data pipeline, RFAA architecture, training, in silico design methods, and experimental validation can be found in the supplementary materials.

REFERENCES AND NOTES

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline

2. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U.

Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi:10.1126/science.abj8754 Medline

3. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin, Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004). doi:10.1021/jm0306430 Medline
4. M. L. Hekkelman, I. de Vries, R. P. Joosten, A. Perrakis, AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023). doi:10.1038/s41592-022-01685-y Medline
5. G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, DiffDock: Diffusion steps, twists, and turns for molecular docking. arXiv:2210.01776 [q-bio.BM] (2022).
6. R. V. Honorato, J. Roel-Touris, A. M. J. J. Bonvin, MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Front. Mol. Biosci.* **6**, 102 (2019). doi:10.3389/fmolb.2019.00102 Medline
7. M. Holcomb, Y.-T. Chang, D. S. Goodsell, S. Forli, Evaluation of AlphaFold2 structures as docking targets. *Protein Sci.* **32**, e4530 (2023). doi:10.1002/pro.4530 Medline
8. A. M. Díaz-Rovira, H. Martín, T. Beuming, L. Díaz, V. Guallar, S. S. Ray, Are deep learning structural models sufficiently accurate for virtual screening? Application of docking algorithms to AlphaFold2 predicted structures. *J. Chem. Inf. Model.* **63**, 1668–1674 (2023). doi:10.1021/acs.jcim.2c01270 Medline
9. J. Eberhardt, D. Santos-Martins, A. F. Tillack, S. Forli, AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021). doi:10.1021/acs.jcim.1c00203 Medline
10. M. Baek, R. McHugh, I. Anishchenko, D. Baker, F. DiMaio, Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. bioRxiv 2022.09.09.507333 [Preprint] (2022); https://doi.org/10.1101/2022.09.09.507333.
11. M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, F. DiMaio, Efficient and accurate prediction of protein structure using RoseTTAFold2. bioRxiv 2023.05.24.542179 [Preprint] (2023); https://doi.org/10.1101/2023.05.24.542179.
12. C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, The Cambridge Structural Database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016). doi:10.1107/S2052520616003954 Medline
13. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi:10.1126/science.ade2574 Medline
14. R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, J. Peng, High-resolution de novo structure prediction from primary sequence. bioRxiv 2022.07.21.500999 [Preprint] (2022); https://doi.org/10.1101/2022.07.21.500999.
15. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034 [Preprint] (2022); https://doi.org/10.1101/2021.10.04.463034.
16. J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, T. Schwede, Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **86**, 387–398 (2018). doi:10.1002/prot.25431 Medline
17. J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, T. Schwede, The Protein Model Portal—A comprehensive resource for protein structure and model information. *Database* **2013**, bat031 (2013). doi:10.1093/database/bat031 Medline
18. J. Haas, R. Gumienny, A. Barbato, F. Ackermann, G. Tauriello, M. Bertoni, G. Studer, A. Smolinski, T. Schwede, Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins* **87**, 1378–1387 (2019). doi:10.1002/prot.25815 Medline
19. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018). doi:10.1093/nar/gky427 Medline
20. N. Guex, M. C. Peitsch, T. Schwede, Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **30**, S162–S173 (2009). doi:10.1002/elps.200900140 Medline
21. S. Bienert, A. Waterhouse, T. A. P. de Beer, G. Tauriello, G. Studer, L. Bordoli, T. Schwede, The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017). doi:10.1093/nar/gkw1132 Medline
22. G. Studer, C. Rempfer, A. M. Waterhouse, R. Gumienny, J. Haas, T. Schwede, QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics* **36**, 1765–1771 (2020). doi:10.1093/bioinformatics/btz828 Medline
23. M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7**, 10480 (2017). doi:10.1038/s41598-017-09654-8 Medline
24. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022). doi:10.1093/nar/gkab1061 Medline
25. H. Stärk, O.-E. Ganea, L. Pattanaik, R. Barzilay, T. Jaakkola, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato, EquiBind: Geometric deep learning for drug binding structure prediction. arXiv:2202.05146 [q-bio.BM] (2022).
26. W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, S. Zheng, TANKBind: Trigonometry-Aware Neural NetworKs for drug-protein binding structure prediction. bioRxiv 2022.06.06.495043 [Preprint] (2022); https://doi.org/10.1101/2022.06.06.495043.
27. Z. Liao, R. You, X. Huang, X. Yao, "DeepDock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information" in *Proceedings 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, I. Yoo, J. Bi, X. Hu, Eds. (IEEE, 2019), pp. 311–317.
28. Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, A. Anandkumar, State-specific protein-ligand complex structure prediction with a multi-scale deep generative model. arXiv:2209.15171 [q-bio.QM] (2022).
29. G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, G. Ke, Uni-Mol: A universal 3D molecular representation learning framework. ChemRxiv 10.26434/chemrxiv-2022-jjm0j [Preprint] (2022); https://doi.org/10.26434/chemrxiv-2022-jjm0j.
30. M. Buttenschoen, G. M. Morris, C. M. Deane, PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. arXiv:2308.05777 [q-bio.QM] (2023).
31. C. Chang, C. Lee Luo, Y. Gao, In crystallo observation of three metal ion promoted DNA polymerase misincorporation. *Nat. Commun.* **13**, 2346 (2022). doi:10.1038/s41467-022-30005-3 Medline
32. H. Park, G. Zhou, M. Baek, D. Baker, F. DiMaio, Force field optimization guided by small molecule crystal lattice data enables consistent sub-angstrom protein-ligand docking. *J. Chem. Theory Comput.* **17**, 2000–2010 (2021). doi:10.1021/acs.jctc.0c01184 Medline
33. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta All-Atom Energy Function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017). doi:10.1021/acs.jctc.7b00125 Medline
34. H. Bagdonas, C. A. Fogarty, E. Fadda, J. Agirre, The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **28**, 869–870 (2021). doi:10.1038/s41594-021-00680-9 Medline
35. S. Ramazi, J. Zahiri, Posttranslational modifications in proteins: Resources, tools and prediction methods. *Database* **2021**, baab012 (2021). doi:10.1093/database/baab012 Medline

36. C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, Glycosylation in health and disease. *Nat. Rev. Nephrol.* **15**, 346–366 (2019). doi:10.1038/s41581-019-0129-4 Medline

37. J. M. Lee, H. M. Hammarén, M. M. Savitski, S. H. Baek, Control of protein stability by post-translational modifications. *Nat. Commun.* **14**, 201 (2023). doi:10.1038/s41467-023-35795-8 Medline

38. R. J. Woods, Predicting the structures of glycans, glycoproteins, and their complexes. *Chem. Rev.* **118**, 8005–8024 (2018). doi:10.1021/acs.chemrev.8b00032 Medline

39. J. Adolf-Bryfogle, J. W. Labonte, J. C. Kraft, M. Shapavolov, S. Raemisch, T. Lütteke, F. DiMaio, C. D. Bahl, J. Pallesen, N. P. King, J. J. Gray, D. W. Kulp, W. R. Schief, Growing glycans in Rosetta: Accurate de novo glycan modeling, density fitting, and rational sequon design. bioRxiv 2021.09.27.462000 [Preprint] (2021); https://doi.org/10.1101/2021.09.27.462000.

40. S. Jo, H. S. Lee, J. Skolnick, W. Im, Restricted N-glycan conformational space in the PDB and its implication in glycan structure modeling. *PLOS Comput. Biol.* **9**, e1002946 (2013). doi:10.1371/journal.pcbi.1002946 Medline

41. A. Gorelik, K. Illes, K. H. Bui, B. Nagar, Structures of the mannose-6-phosphate pathway enzyme, GlcNAc-1-phosphotransferase. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203518119 (2022). doi:10.1073/pnas.2203518119 Medline

42. S. Tang, Y. Lu, W. M. Skinner, M. Sanyal, P. V. Lishko, M. Ikawa, P. S. Kim, Human sperm TMEM95 binds eggs and facilitates membrane fusion. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2207805119 (2022). doi:10.1073/pnas.2207805119 Medline

43. M. J. Bick, P. J. Greisen, K. J. Morey, M. S. Antunes, D. La, B. Sankaran, L. Reymond, K. Johnsson, J. I. Medford, D. Baker, Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **6**, e28909 (2017). doi:10.7554/eLife.28909 Medline

44. N. F. Polizzi, W. F. DeGrado, A defined structural unit enables de novo design of small-molecule-binding proteins. *Science* **369**, 1227–1233 (2020). doi:10.1126/science.abb8330 Medline

45. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). doi:10.1038/s41586-023-06415-8 Medline

46. B. Ni, D. L. Kaplan, M. J. Buehler, Generative design of de novo proteins based on secondary structure constraints using an attention-based diffusion model. *Chem* **9**, 1828–1849 (2023). doi:10.1016/j.chempr.2023.03.020 Medline

47. L. Wu, B. L. Trippe, C. A. Naesseth, D. M. Blei, J. P. Cunningham, Practical and asymptotically exact conditional sampling in diffusion models. arXiv:2306.17775 [stat.ML] (2023).

48. J. Ingraham, M. Baranov, Z. Costello, V. Frappier, A. Ismail, S. Tie, W. Wang, V. Xue, F. Obermeyer, A. Beam, G. Grigoryan, Illuminating protein space with a programmable generative model. bioRxiv 2022.12.01.518682 [Preprint] (2022); https://doi.org/10.1101/2022.12.01.518682.

49. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022). doi:10.1126/science.add2187 Medline

50. J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, D. Baker, Atomic context-conditioned protein sequence design using LigandMPNN. bioRxiv 2023.12.22.573103 [Preprint] (2023); https://doi.org/10.1101/2023.12.22.573103.

51. B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, T. Jaakkola, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv:2206.04119 [q-bio.BM] (2022).

52. Digitalis Investigation Group, The effect of digoxin on mortality and morbidity in patients with heart failure. *N. Engl. J. Med.* **336**, 525–533 (1997). doi:10.1056/NEJM199702203360801 Medline

53. R. J. Flanagan, A. L. Jones, Fab antibody fragments: Some applications in clinical toxicology. *Drug Saf.* **27**, 1115–1133 (2004). doi:10.2165/00002018-200427140-00004 Medline

54. C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, D. Baker, Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013). doi:10.1038/nature12443 Medline

55. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008). doi:10.1146/annurev.biochem.77.062906.171838 Medline

56. T. L. Poulos, Heme enzyme structure and function. *Chem. Rev.* **114**, 3919–3962 (2014). doi:10.1021/cr400415k Medline

57. X. Huang, J. T. Groves, Oxygen activation and radical transformations in heme proteins and metalloporphyrins. *Chem. Rev.* **118**, 2491–2553 (2018). doi:10.1021/acs.chemrev.7b00373 Medline

58. I. Kalvet, M. Ortmayer, J. Zhao, R. Crawshaw, N. M. Ennist, C. Levy, A. Roy, A. P. Green, D. Baker, Design of heme enzymes with a tunable substrate binding pocket adjacent to an open metal coordination site. *J. Am. Chem. Soc.* **145**, 14307–14315 (2023). doi:10.1021/jacs.3c02742 Medline

59. M. Sono, J. H. Dawson, L. P. Hager, The generation of a hyperporphyrin spectrum upon thiol binding to ferric chloroperoxidase. Further evidence of endogenous thiolate ligation to the ferric enzyme. *J. Biol. Chem.* **259**, 13209–13216 (1984). doi:10.1016/S0021-9258(18)90679-4 Medline

60. N. Adir, S. Bar-Zvi, D. Harris, The amazing phycobilisome. *Biochim. Biophys. Acta Bioenerg.* **1861**, 148047 (2020). doi:10.1016/j.bbabio.2019.07.002 Medline

61. A. Marx, N. Adir, Allophycocyanin and phycocyanin crystal structures reveal facets of phycobilisome assembly. *Biochim. Biophys. Acta* **1827**, 311–318 (2013). doi:10.1016/j.bbabio.2012.11.006 Medline

62. C. Zhao, A. Höppner, Q.-Z. Xu, W. Gärtner, H. Scheer, M. Zhou, K.-H. Zhao, Structures and enzymatic mechanisms of phycobiliprotein lyases CpcE/F and PecE/F. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13170–13175 (2017). doi:10.1073/pnas.1715495114 Medline

63. S. F. H. Barnett, A. Hitchcock, A. K. Mandal, C. Vasilev, J. M. Yuen, J. Morby, A. A. Brindley, D. M. Niedzwiedzki, D. A. Bryant, A. J. Cadby, D. Holten, C. N. Hunter, Repurposing a photosynthetic antenna protein as a super-resolution microscopy label. *Sci. Rep.* **7**, 16807 (2017). doi:10.1038/s41598-017-16834-z Medline

64. J. A. Mancini, M. Sheehan, G. Kodali, B. Y. Chow, D. A. Bryant, P. L. Dutton, C. C. Moser, De novo synthetic biliprotein design, assembly and excitation energy transfer. *J. R. Soc. Interface* **15**, 20180021 (2018). doi:10.1098/rsif.2018.0021 Medline

65. A. Hitchcock, C. N. Hunter, R. Sobotka, J. Komenda, M. Dann, D. Leister, Redesigning the photosynthetic light reactions to enhance photosynthesis - the PhotoRedesign consortium. *Plant J.* **109**, 23–34 (2022). doi:10.1111/tpj.15552 Medline

66. K.-L. Wu, J. A. Moore, M. D. Miller, Y. Chen, C. Lee, W. Xu, Z. Peng, Q. Duan, G. N. Phillips Jr., R. A. Uribe, H. Xiao, Expanding the eukaryotic genetic code with a biosynthesized 21st amino acid. *Protein Sci.* **31**, e4443 (2022). doi:10.1002/pro.4443 Medline

67. L. L. Rade, W. C. Generoso, S. Das, A. S. Souza, R. L. Silveira, M. C. Avila, P. S. Vieira, R. Y. Miyamoto, A. B. B. Lima, J. A. Aricetti, R. R. de Melo, N. Milan, G. F. Persinoti, A. M. F. L. J. Bonomi, M. T. Murakami, T. M. Makris, L. M. Zanphorlin, Dimer-assisted mechanism of (un)saturated fatty acid decarboxylation for alkene production. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2221483120 (2023). doi:10.1073/pnas.2221483120 Medline

68. Y. Yuan, G. Jia, C. Wu, W. Wang, L. Cheng, Q. Li, Z. Li, K. Luo, S. Yang, W. Yan, Z. Su, Z. Shao, Structures of signaling complexes of lipid receptors S1PR1 and S1PR5 reveal mechanisms of activation and drug recognition. *Cell Res.* **31**, 1263–1274 (2021). doi:10.1038/s41422-021-00566-x Medline

69. K. Le, M. J. Soth, J. B. Cross, G. Liu, W. J. Ray, J. Ma, S. G. Goodwani, P. J. Acton, V. Buggia-Prevot, O. Akkermans, J. Barker, M. L. Conner, Y. Jiang, Z. Liu, P. McEwan, J. Warner-Schmidt, A. Xu, M. Zebisch, C. J. Heijnen, B. Abrahams, P. Jones, Discovery of IACS-52825, a potent and selective DLK inhibitor for treatment of chemotherapy-induced peripheral neuropathy. *J. Med. Chem.* **66**, 9954–9971 (2023). doi:10.1021/acs.jmedchem.3c00788 Medline

70. A. Schenkmayerova, M. Toul, D. Pluskal, R. Baatallah, G. Gagnot, G. P. Pinto, V. T. Santana, M. Stuchla, P. Neugebauer, P. Chaiyen, J. Damborsky, D. Bednar, Y. L. Janin, Z. Prokop, M. Marek, Catalytic mechanism for Renilla-type luciferases. *Nat. Catal.* **6**, 23–38 (2023). doi:10.1038/s41929-022-00895-z

71. E. Konia, K. Chatzicharalampous, A. Drakonaki, C. Muenke, U. Ermler, G. Tsiotis, I. V. Pavlidis, Rational engineering of *Luminiphilus syltensis* (*R*)-selective amine

transaminase for the acceptance of bulky substrates. *Chem. Commun.* **57**, 12948–12951 (2021). doi:10.1039/D1CC04664K Medline

72. K. Raja Reddy, M. Totrov, O. Lomovskaya, D. C. Griffith, Z. Tarazi, M. C. Clifton, S. J. Hecker, Broad-spectrum cyclic boronate β-lactamase inhibitors featuring an intramolecular prodrug for oral bioavailability. *Bioorg. Med. Chem.* **62**, 116722 (2022). doi:10.1016/j.bmc.2022.116722 Medline
73. R. Krishna, J. Wang, Woody Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Dryad (2024); https://doi.org/10.5061/dryad.mcvdnck6v.
74. R. Krishna, Generalized biomolecular modeling with RoseTTAFold All-Atom. Zenodo (2024); https://doi.org/10.5281/zenodo.10699231.
75. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011). doi:10.1038/nmeth.1818 Medline
76. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). doi:10.1186/s12859-019-3019-7 Medline
77. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000). doi:10.1093/nar/28.1.235 Medline
78. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, "Learning inverse folding from millions of predicted structures" in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research (PMLR)*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato, Eds. (PMLR, 2022), pp. 8946–8970.
79. R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004). doi:10.1021/jm030580l Medline
80. J. Yang, A. Roy, Y. Zhang, BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2013). doi:10.1093/nar/gks966 Medline
81. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017). doi:10.1038/nbt.3988 Medline
82. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011). doi:10.1186/1758-2946-3-33 Medline
83. R. M. Roshan, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, "MSA transformer" in *Proceedings of the 38th International Conference on Machine Learning*, vol. 138 of *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang, Eds. (PMLR, 2021), pp. 8844–8856.
84. N. Bhattacharya, N. Thomas, R. Rao, J. Daupras, P. K. Koo, D. Baker, Y. S Song, S. Ovchinnikov, "Single layers of attention suffice to predict protein contacts," Paper presented at the ICLR 2021 Workshop EBM, 7 May 2021.
85. F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, SE (3)-transformers: 3D roto-translation equivariant attention networks. arXiv:2006.10503 [cs.LG] (2020).
86. N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, P. Riley, Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. arXiv:1802.08219 [cs.LG] (2018).
87. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library. arXiv:1912.01703 [cs.LG] (2019).
88. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in vol. 1 of *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2019), pp. 4171–4186.
89. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020). doi:10.1073/pnas.1914677117 Medline
90. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, W. M. Skiff, UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992). doi:10.1021/ja00051a040
91. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004). doi:10.1002/prot.20264 Medline
92. I. Anishchenko, M. Baek, H. Park, N. Hiranuma, D. E. Kim, J. Dauparas, S. Mansoor, I. R. Humphreys, D. Baker, Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins* **89**, 1722–1733 (2021). doi:10.1002/prot.26194 Medline
93. A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, H. Bye-A-Jee, A. Cukura, P. Denny, T. Dogan, T. G. Ebenezer, J. Fan, P. Garmiri, L. J. da Costa Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasaamy, A. Lock, A. Luciani, M. Lugaric, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, A. Mishra, K. Moulang, A. Nightingale, S. Pundir, G. Qi, S. Raj, P. Raposo, D. L. Rice, R. Saidi, R. Santos, E. Speretta, J. Stephenson, P. Totoo, E. Turner, N. Tyagi, P. Vasudev, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. J. Bridge, L. Aimo, G. Argoud-Puy, A. H. Auchincloss, K. B. Axelsen, P. Bansal, D. Baratin, T. M. Batista Neto, M.-C. Blatter, J. T. Bolleman, E. Boutet, L. Breuza, B. C. Gil, C. Casals-Casas, K. C. Echioukh, E. Coudert, B. Cuche, E. de Castro, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, V. Muthukrishnan, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M. Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C. J. A. Sigrist, K. Sonesson, S. Sundaram, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, J. Zhang; UniProt Consortium, UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023). doi:10.1093/nar/gkac1052 Medline
94. I. A. Chen, V. M. Markowitz, K. Chu, K. Palaniappan, E. Szeto, M. Pillay, A. Ratner, J. Huang, E. Andersen, M. Huntemann, N. Varghese, M. Hadjithomas, K. Tennessen, T. Nielsen, N. N. Ivanova, N. C. Kyrpides, IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017). doi:10.1093/nar/gkw929 Medline
95. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011). doi:10.1371/journal.pcbi.1002195 Medline
96. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi:10.1186/1471-2105-10-421 Medline
97. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
98. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022). doi:10.1038/s41586-022-04654-9 Medline
99. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian~16

Revision C.01 (Gaussian Inc., 2016).

100. S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011). doi:10.1002/jcc.21759 Medline

101. C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11**, e1493 (2021). doi:10.1002/wcms.1493

102. S. Vázquez Torres, P. J. Y. Leung, I. D. Lutz, P. Venkatesh, J. L. Watson, F. Hink, H.-H. Huynh, A. H.-W. Yeh, D. Juergens, N. R. Bennett, A. N. Hoofnagle, E. Huang, M. J. MacCoss, M. Expòsit, G. R. Lee, P. M. Levine, X. Li, M. Lamb, E. N. Korkmaz, J. Nivala, L. Stewart, J. M. Rogers, D. Baker, De novo design of high-affinity protein binders to bioactive helical peptides. bioRxiv 10.1101/2022.12.10.519862 [Preprint] (2022); https://doi.org/10.1101/2022.12.10.519862.

103. J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J.-H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, D. Baker, Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022). doi:10.1126/science.abn2100 Medline

104. N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. De Munck, S. N. Savvides, D. Baker, Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023). doi:10.1038/s41467-023-38328-5 Medline

105. G. R. Lee, S. J. Pellock, C. Norn, D. Tischer, J. Dauparas, I. Anischenko, J. A. M. Mercer, A. Kang, A. Bera, H. Nguyen, I. Goreshnik, D. Vafeados, N. Roullier, H. L. Han, B. Coventry, H. K. Haddox, D. R. Liu, A. H.-W. Yeh, D. Baker, Small-molecule binding and sensing with a designed protein family. bioRxiv 2023.11.01.565201 [Preprint] (2023); https://doi.org/10.1101/2023.11.01.565201.

106. R. M. Alvey, A. Biswas, W. M. Schluchter, D. A. Bryant, Attachment of noncognate chromophores to CpcA of *Synechocystis* sp. PCC 6803 and *Synechococcus* sp. PCC 7002 by heterologous expression in *Escherichia coli*. *Biochemistry* **50**, 4890–4902 (2011). doi:10.1021/bi200307s Medline

107. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022). doi:10.1126/science.add1964 Medline

108. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019). doi:10.1038/s41592-019-0357-3 Medline

109. E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, "Protein identification and analysis tools on the ExPASy server" in *The Proteomics Protocols Handbook*, Springer Protocols Handbooks, J. M. Walker, Ed. (Humana Press, 2005), pp. 571–607

110. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010). doi:10.1107/S0907444909047337 Medline

111. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011). doi:10.1107/S0907444910045749 Medline

112. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007). doi:10.1107/S0021889807021206 Medline

113. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010). doi:10.1107/S0907444909052925 Medline

114. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004). doi:10.1107/S0907444904019158 Medline

115. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall III, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018). doi:10.1002/pro.3330 Medline

116. J. D. Hunter, Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007). doi:10.1109/MCSE.2007.55

117. M. L. Waskom, seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021). doi:10.21105/joss.03021

118. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas; SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020). doi:10.1038/s41592-019-0686-2 Medline

119. A. Grosdidier, V. Zoete, O. Michielin, SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **39**, W270–W277 (2011). doi:10.1093/nar/gkr366 Medline

120. O. Trott, A. J. Olson, 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). doi:10.1002/jcc.21334 Medline

ACKNOWLEDGMENTS

We thank Luki Goldschmidt and Kandise VanWormer for maintaining the computational and wet lab resources at the Institute for Protein Design. We thank Joe Watson and David Juergens for helpful conversations during the development of the method and providing the figure color scheme used in this work, Justas Dauparas for helpful conversations about model improvements and developing LigandMPNN, Guangfeng Zhou for help with GALigandDock and Madison Kennedy for reading earlier drafts of the manuscript. We thank Lukas Milles for providing custom plasmids. We thank Tina Nguyen for assistance with the SEC-MALS experiment. Crystallographic data were collected on AMX of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). We also thank Xavier Robin for helping set up our CAMEO server and providing the Vina and AD4 servers for us to benchmark our results and Martin Buttenschoen for providing metadata from their PoseBusters benchmark results. **Funding:** We thank Microsoft for generous donation of Azure Compute Credits, and Perlmutter grant NERSC award BER-ERCAP0022018 for access to the Perlmutter high performance computing resources. This work was supported by gifts from Microsoft (R.K., P.S., D.B.), the Howard Hughes Medical Institute (D.B, G.R.L.),, the New Faculty Startup Fund from Seoul National University (M.B.), the Schmidt Futures program (J.W., R.M., F.D.), the Open Philanthropy Project Improving Protein Design Fund (J.W., I.K, G.R.L.), grant no. INV-010680 from the Bill and Melinda Gates Foundation (W.A.), the Audacious Project (P.V.), the Washington State General Operating Fund supporting the Institute for Protein Design (P.V.), Defense Threat Reduction Agency (DTRA) (G.R.L.), the Washington Research Foundation, Innovation Fellows Program (G.R.L.), Faculty of Science PhD studentship from the University of Sheffield (F.S.M-B), Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.: 75N93022C00036 (I.A.), Amgen (I.R.H.), Juvenile Diabetes Research Foundation International (JDRF) grant # 2-SRA-2018-605-Q-R (X.L.), Helmsley Charitable Trust Type 1 Diabetes (T1D) Program Grant # 2019PG-T1D026 (X.L.), Defense Threat Reduction Agency Grant HDTRA1-19-1-0003 (X.L.), Bill and Melinda Gates Foundation #OPP1156262 (X.L.), Synergy award 854126 from the European Research Council (G.A.S., C.N.H), Royal Society University Research Fellowship (award no. URF\R1\191548) (A.H.), Human frontiers science program grant RGP0061/2019 (F.D.). **Author contributions:** Designed the research: R.K., I.A., M.B., F.D. and D.B. Developed RFAA architecture and training regimen: R.K., J.W. and F.D. Evaluated RFAA on different structure prediction tasks: P.S., R.K., I.R.H. and R.M. Developed RFdiffusionAA: W.A. Generated designs for digoxigenin binders: P.V and G.R.L. Performed experiments for digoxigenin binders: P.V, G.R.L, D.V. and

X.L. Generated designs and performed experiments for heme binders: I.K. Generated designs for bilin binders: W.A. Performed experiments for bilin binders: F.S.M-B. Contributed code and ideas: I.A., G.A.S., M.B. and F.D. Performed the crystallography experiments: A.K, E.B, A.B. Offered supervision throughout the project: D.B., A.H. and C.N.H. Wrote the manuscript: R.K., J.W., W.A and D.B. All authors read and contributed to the manuscript. **Competing interests:** R.K., J.W., W.A., A.I., F.D. and D.B. have filed for a provisional patent covering the work presented. The other authors declare no competing interests. **Data and materials availability:** Additional data files containing successful design models and possible training set sequences can be found at Dryad (*73*). Code and neural network weights are available at https://github.com/baker-laboratory/RoseTTAFold-All-Atom and https://github.com/baker-laboratory/rf_diffusion_aa_public/ for RF-allatom and RFdiffusion All-Atom, respectively. These repositories are archived at Zenodo (*74*). All other data are available in main text or supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adl2528
Materials and Methods
Figs. S1 to S34
Tables S1 to S16
References (*75*–*120*)
MDAR Reproducibility Checklist

Submitted 9 October 2023; accepted 27 February 2024
Published online 7 March 2024
10.1126/science.adl2528

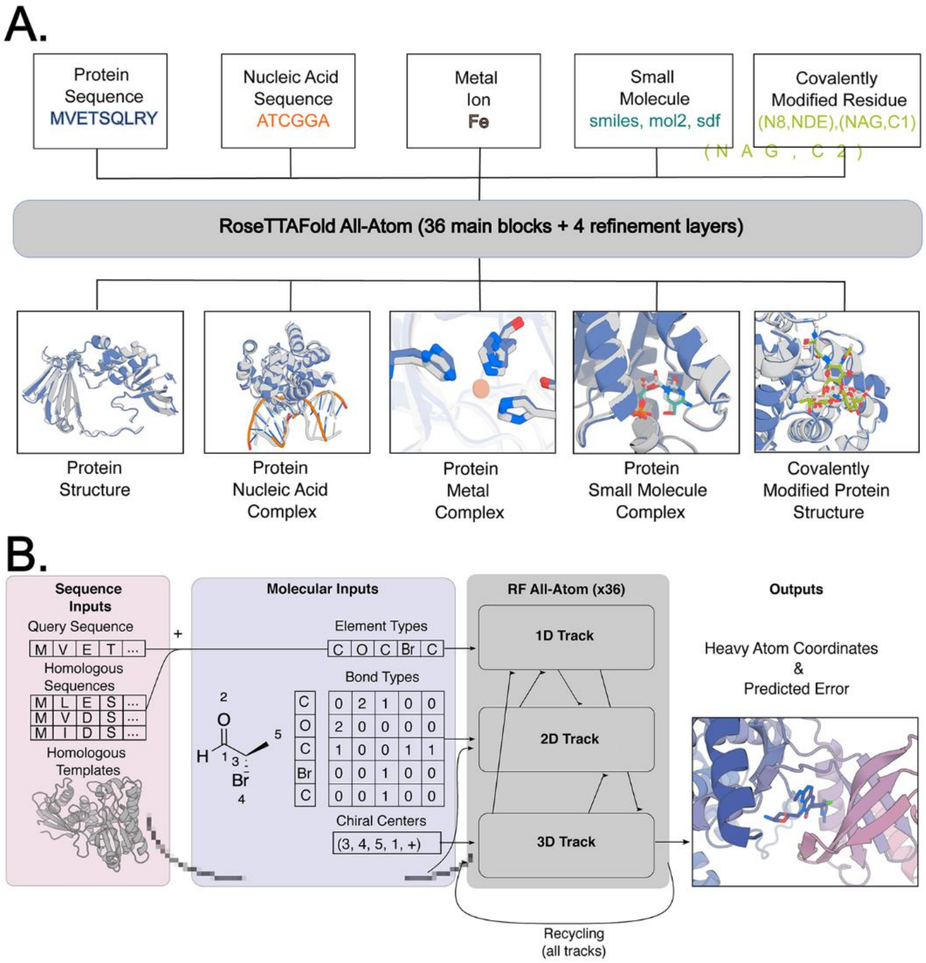

Fig. 1. General biomolecular modeling with RoseTTAFold All-Atom. (**A**) RFAA takes input information about the molecular composition of the biomolecular assembly to be modeled, including protein amino acid and nucleic acid base sequences, metal ions, small molecule bonded structure, and covalent bonds between small molecules and proteins. (**B**) Processing of molecular input information. Small molecule information is parsed into element types (46 possible types), bond types, and chiral centers. Covalent bonds between proteins and small molecules are provided as a separate token in the bond adjacency matrix. The three-track architecture mixes 1D, 2D, and 3D information and predicts all-atom coordinates and model confidence.

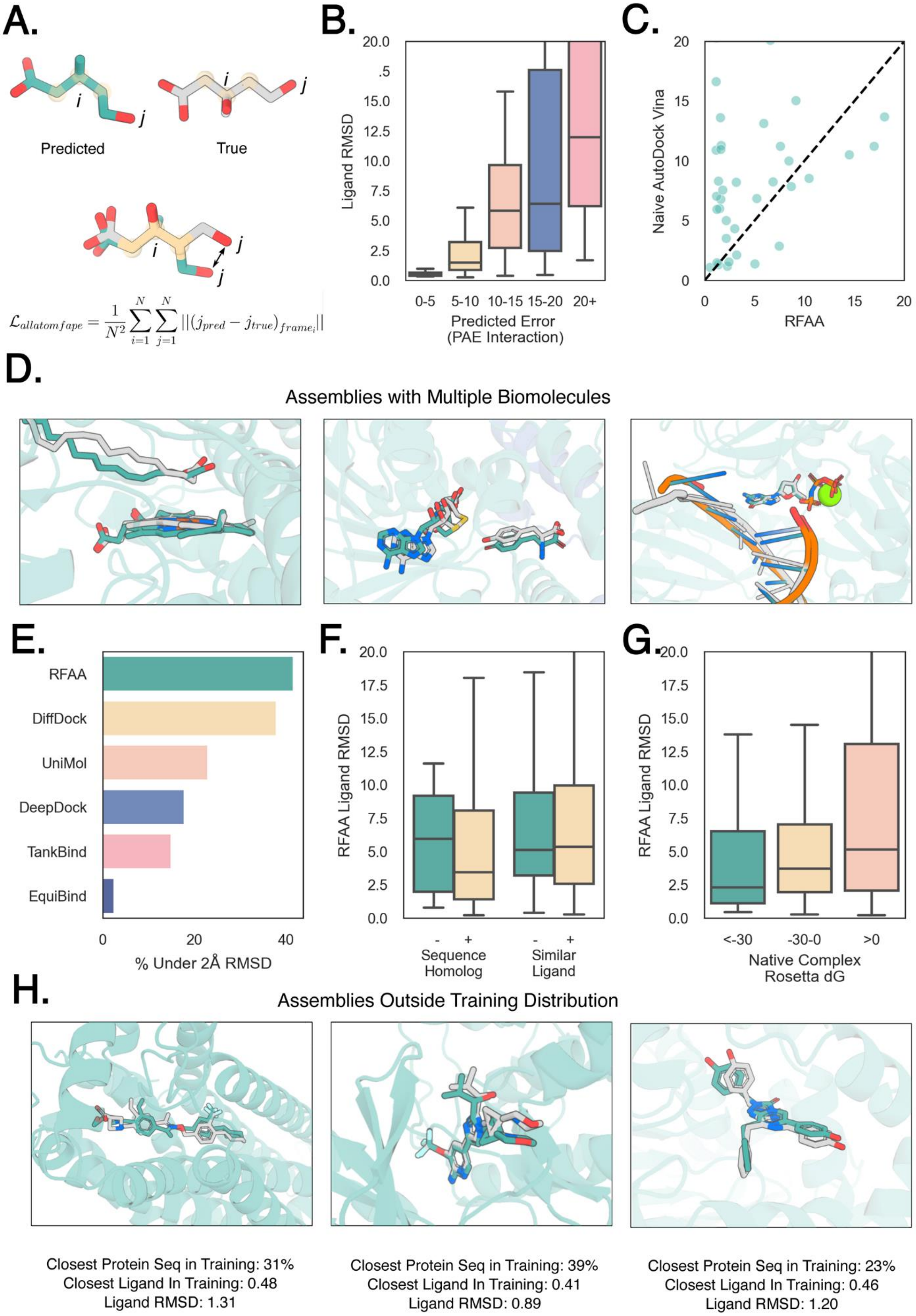
A.
i
j
Predicted
i
j
True
i
j
j
$\mathcal{L}_{allatomfape} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}||(j_{pred} - j_{true})_{frame_i}||$
B.
Ligand RMSD
20.0
.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0-5
5-10
10-15
15-20
20+
Predicted Error
(PAE Interaction)
C.
Naive AutoDock Vina
20
15
10
5
0
0
5
10
15
20
RFAA
D.
Assemblies with Multiple Biomolecules
E.
RFAA
DiffDock
UniMol
DeepDock
TankBind
EquiBind
0
20
40
% Under 2Å RMSD
F.
RFAA Ligand RMSD
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
-
+
Sequence
Homolog
-
+
Similar
Ligand
G.
RFAA Ligand RMSD
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
<-30
-30-0
>0
Native Complex
Rosetta dG
H.
Assemblies Outside Training Distribution
Closest Protein Seq in Training: 31%
Closest Ligand In Training: 0.48
Ligand RMSD: 1.31
Closest Protein Seq in Training: 39%
Closest Ligand In Training: 0.41
Ligand RMSD: 0.89
Closest Protein Seq in Training: 23%
Closest Ligand In Training: 0.46
Ligand RMSD: 1.20

Fig. 2. RoseTTAFold All-Atom can accurately predict protein-small molecule complex structures. All panels: Predicted protein structure (aligned to native): transparent teal, predicted ligand conformation: teal, native ligand conformation: gray. All boxplots cut off at 20 Å for clarity. (**A**) Every "atom" node is assigned a local coordinate frame based on the identities of its neighbors. To compute the main loss in the network, we align each atom's coordinate frame in the predicted and true structures and measure the error over all the other atoms. (**B**) Model accuracy correlates with error predictions. Computed for CAMEO targets (05/20/23-7/29/23; 261 protein-small molecule interfaces). Ligand RMSD was computed by CAMEO organizers. (**C**) RFAA outperforms AutoDock Vina on CAMEO targets (Week 8/12/23-09/02/23; 149 protein-small molecule interfaces). Both servers have to model the protein, find pockets for all ligands present in the solved structure, and the correct docks for all ligands. Ligand RMSD for both servers was computed by CAMEO organizers, AutoDock Vina server set up by CAMEO organizers. (**D**) Three examples of successful predictions with multiple biomolecules. From left to right: fatty acid decarboxylase (PDB ID: 8d8p; Seq ID: 34%; from CAMEO) with a heme cofactor and a lipid substrate, a dimeric tyrosine methyltransferase (PDB ID: 7ux8; Seq ID: 28%; CASP15 Target: T1124) with an *S*-adenosyl homocysteine and tyrosine interaction and a DNA polymerase (PDB ID: 7u7w; Seq ID: 100%) bound to DNA, a nucleotide and a metal ion (*31*, *66*, *67*). (**E**) Comparison to other deep learning-based docking methods. In this case, each method was applied in their respective training regime. For RFAA this means only having sequence and minimal atomic graph inputs, whereas for other methods this involves providing the bound crystal structure. Ligand RMSD was computed using PoseBusters suite, and a single example present in our training set was removed for all methods in comparison. (**F**) Comparison of RFAA predictions on recently solved PDBs that are novel compared to the training set (Homolog <1 BLAST e-value, Similar Ligand >0.5 Tanimoto Similarity). Each set is clustered based on sequence/ligand similarity, and a random cluster representative is chosen for each. (**G**) Comparison of RFAA prediction accuracy to Rosetta ΔG energy estimates for the native complex (over 940 cases that were successfully processed by Rosetta). RFAA makes more accurate predictions for native complexes with low Rosetta energy. (**H**) Three examples of successful predictions with low similarity to the training set. From left to right: G protein-coupled S1P receptor (PDB ID: 7ew1; Seq ID: 31%), complex of DLK bound to an inhibitor (PDB ID: 8ous; Seq ID: 39%), a *Renilla* luciferase bound to an azacoelenterazine (non-native substrate; PDB ID: 7qxr; Seq ID: 23%) (*68*–*70*).

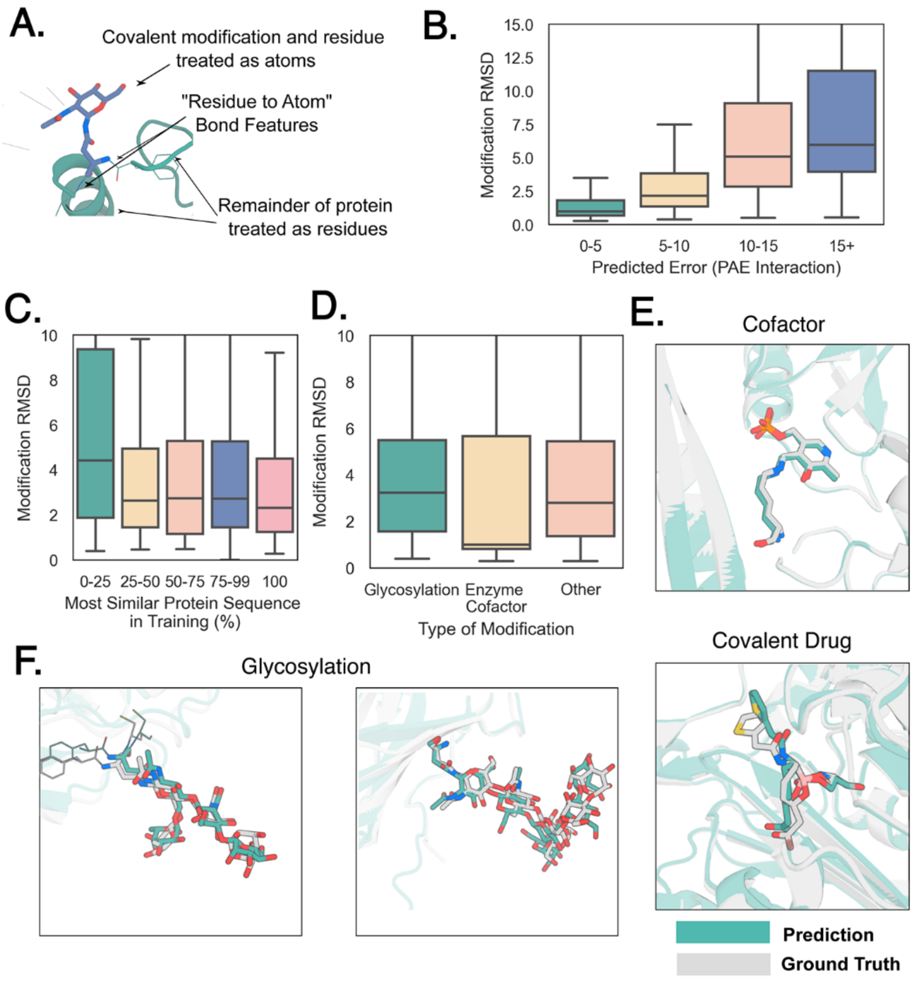

Fig. 3. **Accurate prediction of protein covalent modifications.** All panels: transparent teal: predicted protein structure, transparent gray: native structure, teal: predicted covalent modification, gray: native covalent modification. (**A**) Schematic describing how RFAA models covalent modifications to proteins. The chemical moiety that modifies the residue and the residue are modeled as atom nodes, and the rest of the protein is modeled as residues (with MSA and template inputs). (**B**) Model accuracy correlates with predicted error on a set of 938 recently solved structures with covalent modifications. Modification RMSD is computed by aligning the protein structure within 10 Å and computing RMSD over the modified residue and chemical modification. Boxplot cut off at 15 Å for clarity. (**C**) Comparison of sequence identity to training set and model accuracy. Models are generally accurate even with low sequence homology to the training set. (**D**) Comparison of model accuracy for different types of covalent modifications. (**E**) Top: Example of successfully predicted covalently linked enzyme cofactor (PDB ID: 7p3t; Seq ID: 28%), which is a structure of a (R)-selective amine transaminase. Bottom: example of a covalently bound drug candidate (PDB ID: 7ti1, Seq ID: 27%), which is a β-lactamase enzyme bound to cyclic boronic acid inhibitor (*71*, *72*). (**F**) Accurate predictions of glycans on the N-acetylglucosamine-1-phosphotransferase (GNPT) gamma subunit (PDB ID: 7s69; No BLAST hits), human sperm TMEM ectodomain (PDB ID: 7ux0; Seq ID: 26%) (*41*, *42*).

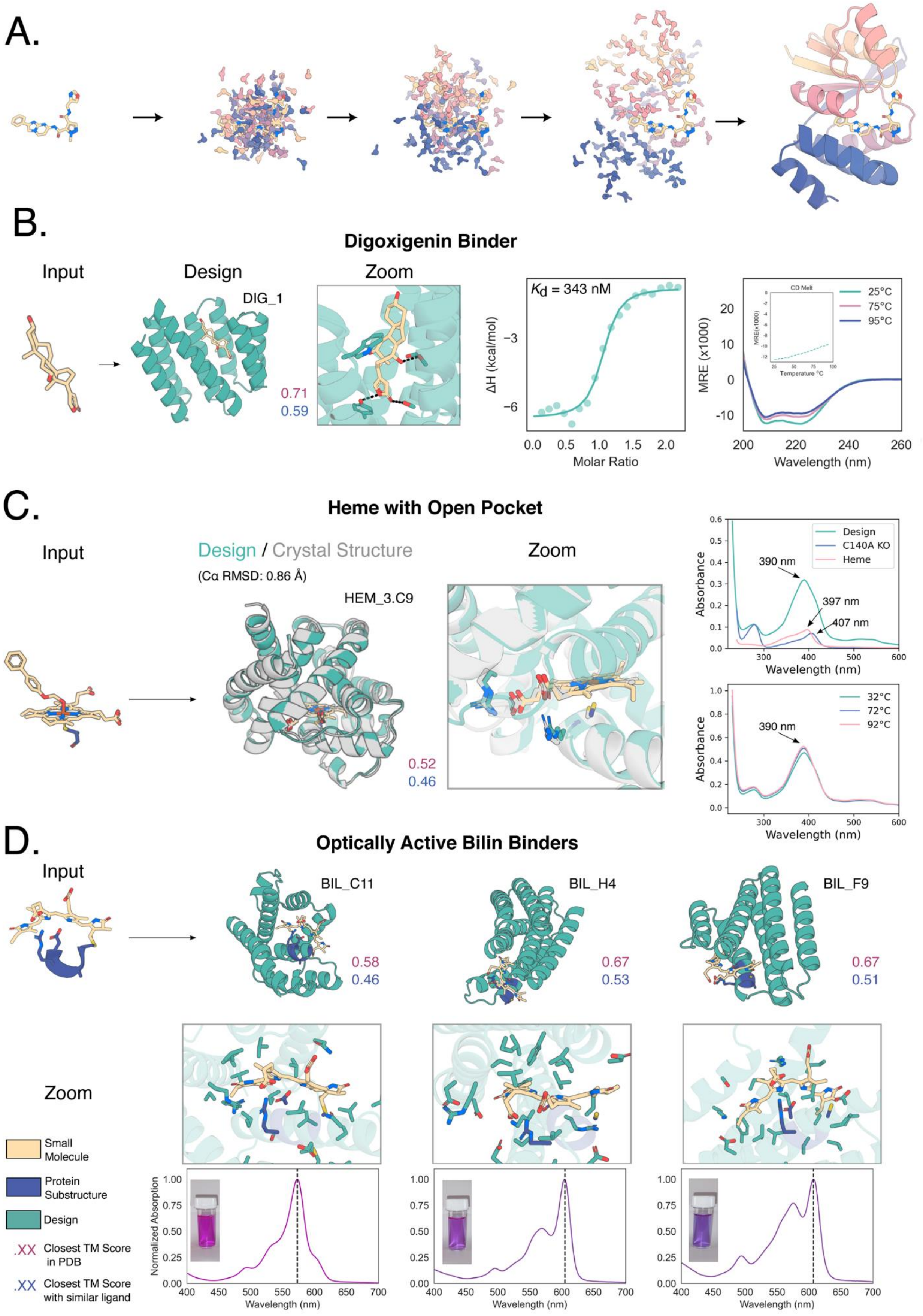
A.
B.
Digoxigenin Binder
Input
Design
Zoom
DIG_1
0.71
0.59
K_d = 343 nM
ΔH (kcal/mol)
−3
−6
0.0 0.5 1.0 1.5 2.0
Molar Ratio
CD Melt
MRE(x1000)
Temperature °C
25 50 75 100
MRE (x1000)
20
10
0
-10
200 220 240 260
Wavelength (nm)
25°C
75°C
95°C
C.
Heme with Open Pocket
Input
Design / Crystal Structure
(Cα RMSD: 0.86 Å)
HEM_3.C9
Zoom
0.52
0.46
Absorbance
Design
C140A KO
Heme
390 nm
397 nm
407 nm
0.0 0.1 0.2 0.3 0.4 0.5 0.6
300 400 500 600
Wavelength (nm)
32°C
72°C
92°C
390 nm
0.0 0.2 0.4 0.6 0.8 1.0
300 400 500 600
Wavelength (nm)
D.
Optically Active Bilin Binders
Input
BIL_C11
0.58
0.46
BIL_H4
0.67
0.53
BIL_F9
0.67
0.51
Zoom
Small Molecule
Protein Substructure
Design
.XX Closest TM Score in PDB
.XX Closest TM Score with similar ligand
Normalized Absorption
0.00 0.25 0.50 0.75 1.00
400 450 500 550 600 650 700
Wavelength (nm)

Fig. 4. Experimental characterization of RFdiffusionAA designed binders. All panels: input ligand shown in yellow, input protein motif shown in blue, and diffused protein shown in teal. Purple text: Closest TM Score to any protein in the training set, Blue text: Closest TM Score to any protein with a similar ligand bound in the training set (Tanimoto >0.5). (**A**) Schematic depicting the random initialization of residues surrounding a small molecule and progressive denoising by RfdiffusionAA. (**B**) Characterization of dioxigenin binder design. (From left to right) Input motif to RFdiffusionAA, designed protein, zoom in view of binding site sidechains. Isothermal Calorimetry (FP) measuring binding affinity (K_d = 343 nM), CD trace (26 μM protein concentration; inlay CD Melt showing intensity at 220 nm across a broad range of temperatures). (**C**) Characterization of heme binding designs. (From left to right) Input motif to RFdiffusionAA, designed protein aligned to its crystal structure (PDB ID: 8vc8); zoom in view of binding site; (top) UV-Vis spectra of designed protein matches expected spectra for penta-coordinated heme and mutating cysteine to alanine abolishes binding; (bottom) designed protein retains heme binding at temperatures up to 90°C. (**D**) Characterization of bilin binding designs. (Row 1, left to right) Input motif to RFdiffusionAA, three designs with different predicted structural topologies. (Row 2, left to right) Zoom in view of binding sites for each design. (Row 3, left to right) Normalized absorption spectra for the three designs shown. Designs have a range of maximum absorption wavelengths and hence different colors in solution (inset).