Article:

Krishna, R. orcid.org/0009-0003-6387-9622, Wang, J. orcid.org/0000-0002-1139-6640, Ahern, W. orcid.org/0009-0006-1247-8847 et al. (19 more authors) (2024) Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science, 384 (6693). ISSN 0036-8075

https://doi.org/10.1126/science.adl2528

Title: Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom

Authors: 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[1,2,7°], Gyu Rie Lee[1,2,7°], Felix S Morey-Burrows[5], 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[5], Andrew Hitchcock[5], C Neil Hunter[5], Alex Kang[2], Evans Brackenbrough[2], Asim K Bera[2], Minkyung Baek[6], Frank DiMaio[1,2], David Baker[1,2,7*]

‡Equal contribution
°Equal contribution
*To whom correspondence should be addressed: dabaker@uw.edu

Affiliations:

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. School of Biosciences, University of Sheffield, Sheffield, S10 2TN, UK
6. School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea
7. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

One sentence summary: The authors develop a deep learning method for prediction and design of complexes of proteins, small molecules, and nucleic acids.

Abstract:

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 finetuning 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.

Main Text:

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 (Figure 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 (Figure 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 (Figure 2A; for greater sensitivity to small molecule geometry, we upweight contributions involving atoms; see Supplemental 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 Supplemental 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 towards 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 (Supplemental 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 Figure 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 (Figure 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 (Figure 2C; 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 (Figure S3; for further exploration of failure modes, see Supplemental 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 (Figure 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 (Figure 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; Figure 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) (Figure 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 S4A), 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 (Figure S4B-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 do 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 (Figure 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 (Figure 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 (Figure 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 (where Modification RMSD is defined as 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) (Figure 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 (Figure 3C). RFAA models interactions with covalently bound cofactors and covalently bound drugs with median RMSDs of 0.99 Å and 2.8 Å respectively (Figure 3D-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 (Figure 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 (Figure 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 (Figure 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 small molecules and other biomolecular context (Figure 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 predict the denoised structures. In contrast to training for the unconditional generation problem and incorporating conditional information through forms of guidance(*47*),(*48*), we train 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 towards 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 Figure 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 (Figure S10C). AF2 structure predictions had backbone RMSD < 2 Å to the RFdiffusionAA design models in all cases (Figure S10C). For each small molecule, RFdiffusionAA generates diverse protein structural solutions to the binding problem that differ from native binders to these ligands (Figure S11, Figure 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 (Figure 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 (Supplemental Methods). Experimental characterization identified several DIG-binding proteins (Figures S29-30, Supplemental Methods); the highest affinity binder has a 343 nM K_d for free digoxigenin (measured by isothermal titration calorimetry, Figure 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 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 (Figure 4; Figure S13-16). Twenty designs exhibit high thermostability, retaining their heme binding at temperatures above 85 °C, and do not unfold at temperatures up to 95 °C (Figure 4C and Figure S13-16). 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 (Figure 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 (Figure 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 (Figure 5C, S8B; the extent of red shifting correlates with computed electrostatic potential around the chromophore (Figure S19)). Conformationally restricted bilins typically display higher fluorescence yields, absolute fluorescence 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*) (Figure 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 (Figure 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 (related meaning 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 (Figure 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 will likely be necessary to generate consistently accurate predictions for new protein-small molecule complexes on par with the accuracy deep networks can achieve on protein systems alone. These 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) (Figure 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 (Figure 2F, 3C), and prediction accuracy is higher for complexes with more favorable computed interaction energies using the Rosetta physically based model (Figure 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.

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).

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).

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).

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).

5. G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv [q-bio.BM]* (2022), (available at http://arxiv.org/abs/2210.01776).

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).

7. M. Holcomb, Y.-T. Chang, D. S. Goodsell, S. Forli, Evaluation of AlphaFold2 structures as docking targets. *Protein Sci.* **32**, e4530 (2023).

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).

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).

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), p. 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), p. 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).

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).

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), p. 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* (2022), p. 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 Suppl 1**, 387–398 (2018).

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).

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).

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).

20. N. Guex, M. C. Peitsch, T. Schwede, Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis*. **30 Suppl 1**, S162–73 (2009).

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).

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).

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).

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).

25. H. Stärk, O.-E. Ganea, L. Pattanaik, R. Barzilay, T. Jaakkola, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato, Eds., EquiBind: Geometric deep learning for drug binding structure prediction. *arXiv [q-bio.BM]* (17--23 Jul 2022), pp. 20503–20521.

26. W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, S. Zheng, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds., TANKBind: Trigonometry-Aware Neural NetworKs for drug-protein binding

structure prediction. *bioRxiv* (2022), pp. 7236–7249.

27. Z. Liao, R. You, X. Huang, X. Yao, DeepDock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information. *on Bioinformatics and …* (2019) (available at https://ieeexplore.ieee.org/abstract/document/8983365/).

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 [q-bio.QM]* (2022), (available at http://arxiv.org/abs/2209.15171).

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* (2022), doi: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 [q-bio.QM]* (2023), (available at http://arxiv.org/abs/2308.05777).

31. C. Chang, C. Lee Luo, Y. Gao, In crystallo observation of three metal ion promoted DNA polymerase misincorporation. *Nat. Commun.* **13**, 2346 (2022).

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).

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).

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).

35. S. Ramazi, J. Zahiri, Posttranslational modifications in proteins: resources, tools and prediction methods. *Database* . **2021** (2021), doi:10.1093/database/baab012.

36. C. Reily, T. J. Stewart, M. B. Renfrow, J. Novak, Glycosylation in health and disease. *Nat. Rev. Nephrol.* **15**, 346–366 (2019).

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).

38. R. J. Woods, Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* **118**, 8005–8024 (2018).

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), p. 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).

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).

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).

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** (2017), doi:10.7554/eLife.28909.

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).

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).

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).

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 [stat.ML]* (2023), (available at http://arxiv.org/abs/2306.17775).

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), p. 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).

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), p. 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 [q-bio.BM]* (2022), (available at http://arxiv.org/abs/2206.04119).

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).

53. R. J. Flanagan, A. L. Jones, Fab antibody fragments: some applications in clinical toxicology. *Drug Saf.* **27**, 1115–1133 (2004).

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).

55. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).

56. T. L. Poulos, Heme enzyme structure and function. *Chem. Rev.* **114**, 3919–3962 (2014).

57. X. Huang, J. T. Groves, Oxygen Activation and Radical Transformations in Heme Proteins and Metalloporphyrins. *Chem. Rev.* **118**, 2491–2553 (2018).

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).

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).

60. N. Adir, S. Bar-Zvi, D. Harris, The amazing phycobilisome. *Biochim. Biophys. Acta Bioenerg.* **1861**, 148047 (2020).

61. A. Marx, N. Adir, Allophycocyanin and phycocyanin crystal structures reveal facets of phycobilisome assembly. *Biochim. Biophys. Acta*. **1827**, 311–318 (2013).

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).

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).

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** (2018), doi:10.1098/rsif.2018.0021.

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).

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).

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).

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).

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).

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. *Nature Catalysis*. **6**, 23–38 (2023).

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).

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).

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 (67). Code and neural network weights are available at https://github.com/baker-laboratory/RoseTTAFold-All-Atom and https://github.com/baker-laboratory/rf_diffusion_all_atom for RF-allatom and RFdiffusion All-Atom, respectively. These repositories are archived at Zenodo (68). All other data is available in main text or Supplemental Materials.

Supplementary Materials
Materials and Methods
Figs. S1 to S34
Tables S1 to S16

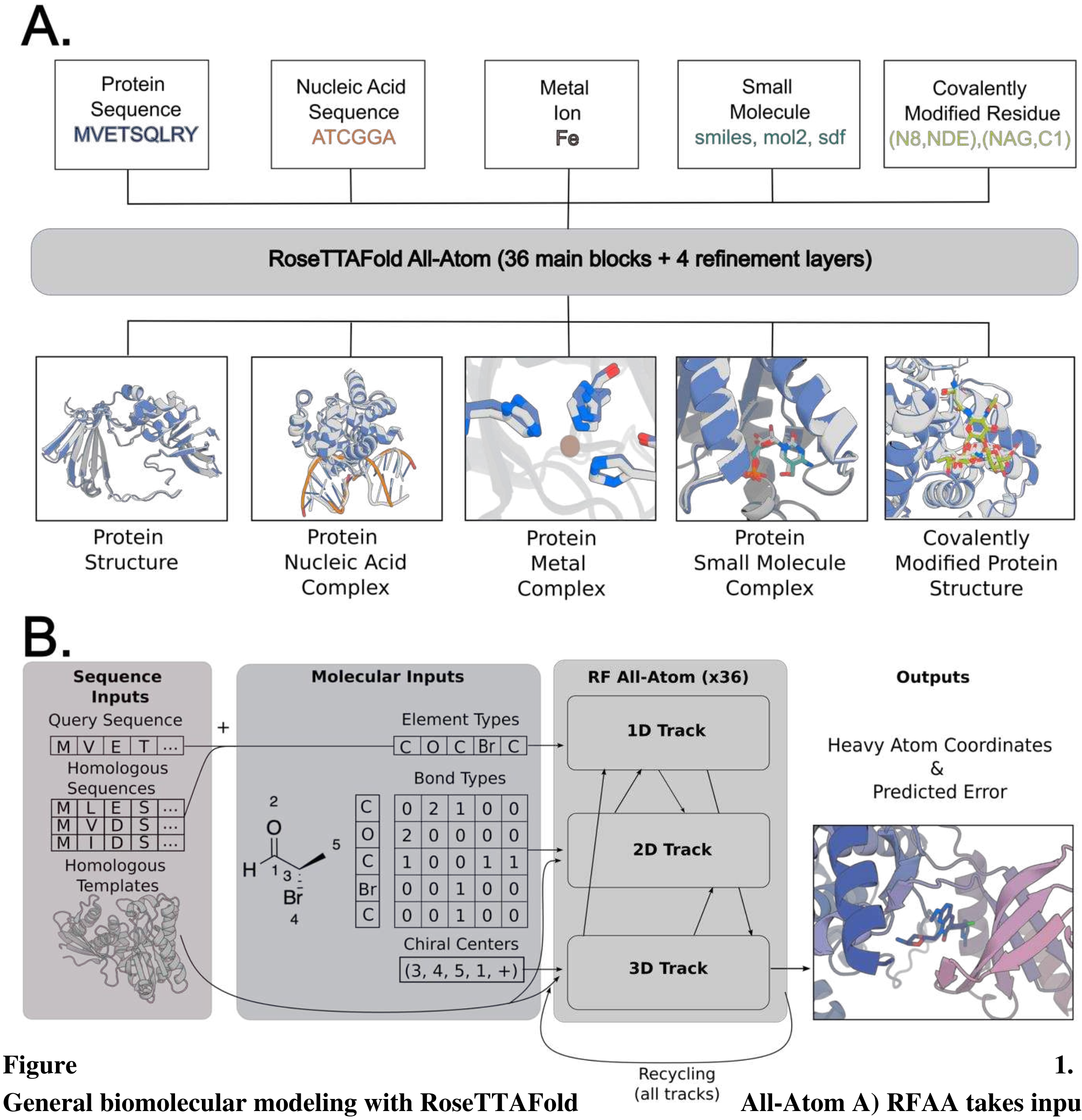

Figure 1. General biomolecular modeling with RoseTTAFold All-Atom A) RFAA takes input information about the molecular composition of the biomolecular assembly to be modeled, including protein amino acid and nucleic acid base sequences, metal ions, small molecule bonded structure, and covalent bonds between 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.

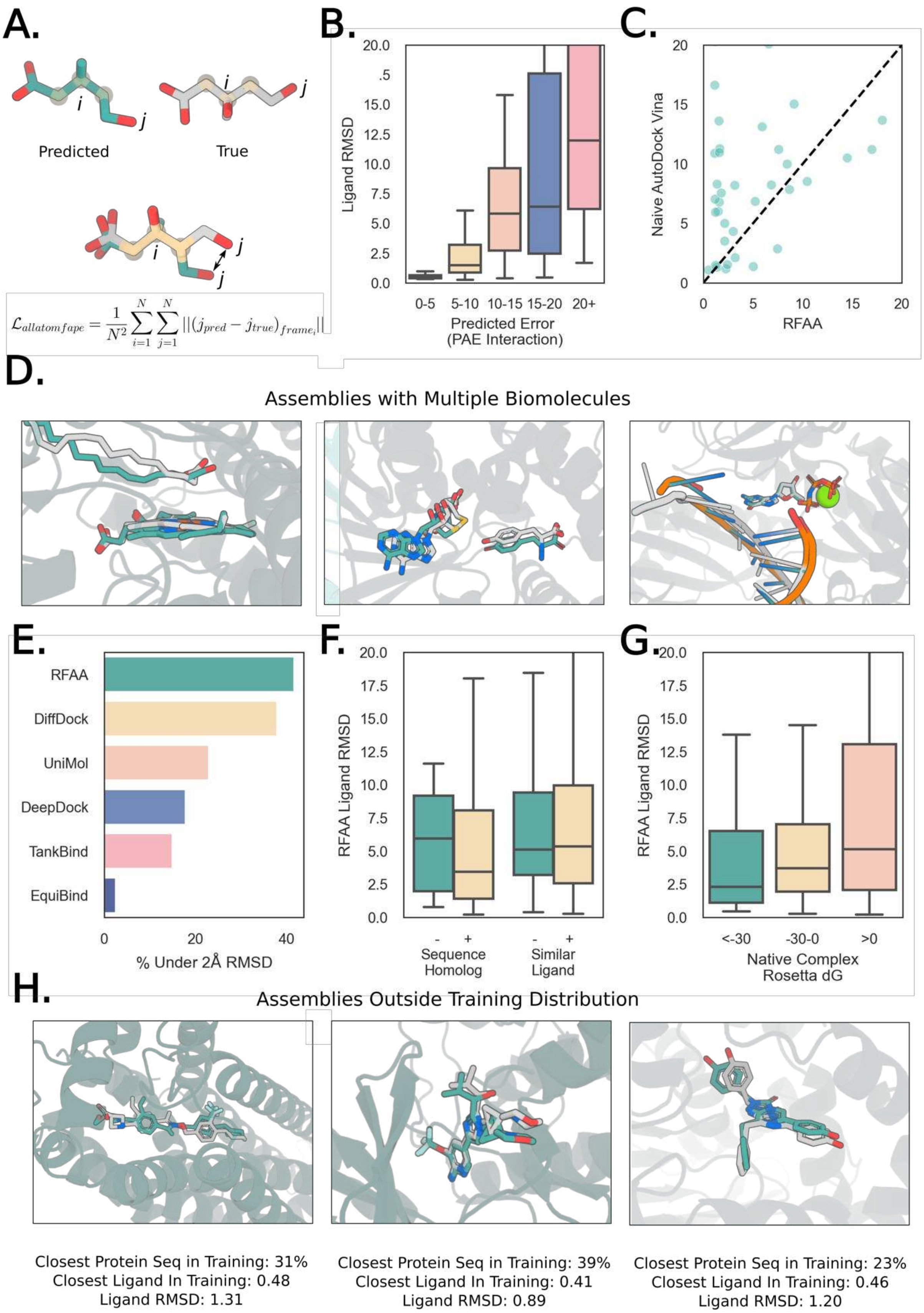
A.
Predicted
True
i
j
$\mathcal{L}_{allatomfape} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}||(j_{pred} - j_{true})_{frame_i}||$
B.
Ligand RMSD
20.0
.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0-5
5-10
10-15
15-20
20+
Predicted Error
(PAE Interaction)
C.
Naive AutoDock Vina
RFAA
0
5
10
15
20
D.
Assemblies with Multiple Biomolecules
E.
RFAA
DiffDock
UniMol
DeepDock
TankBind
EquiBind
% Under 2Å RMSD
0
20
40
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
<-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

Figure 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*)

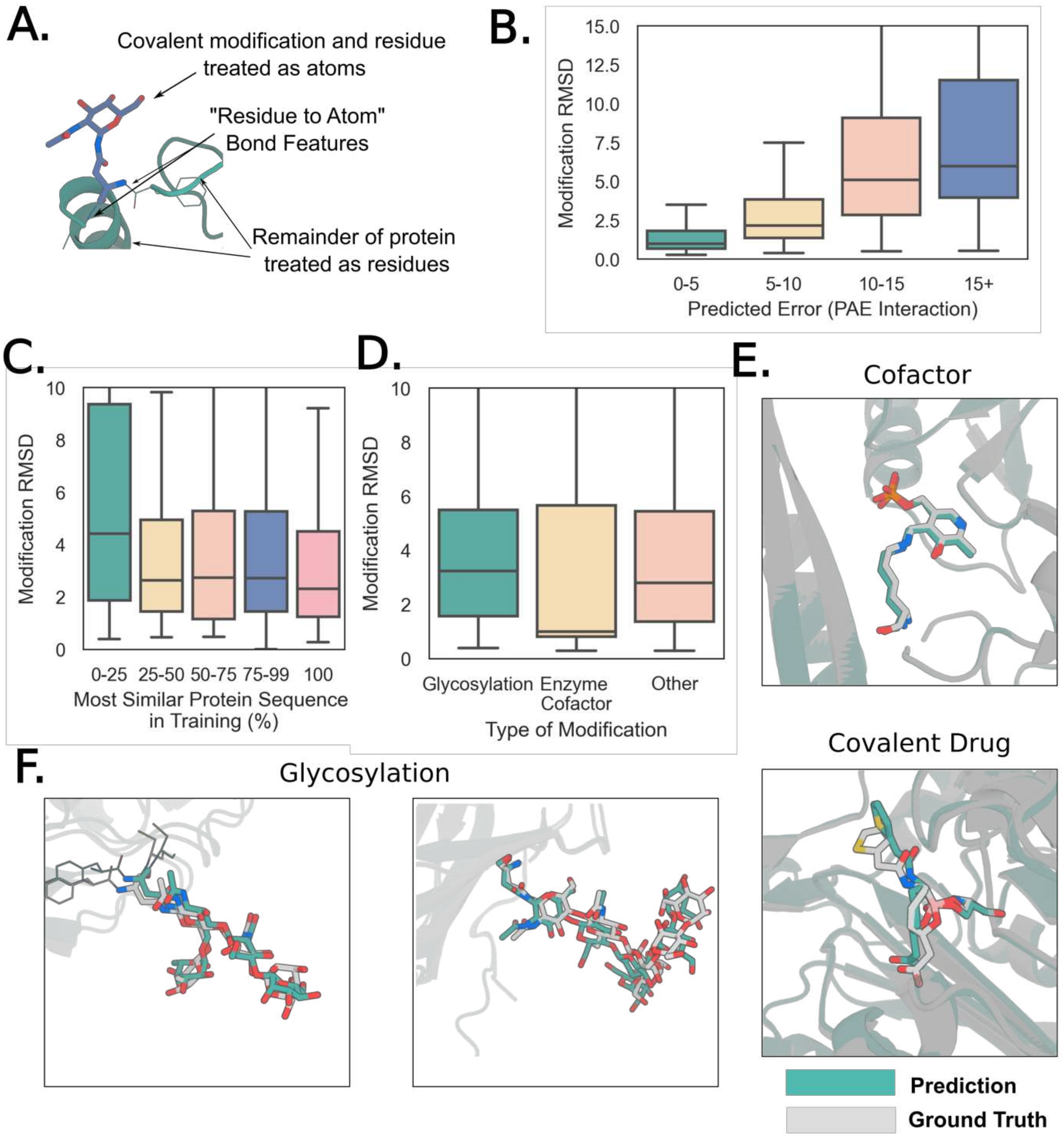

Figure 3. Accurate prediction of protein covalent modifications. All panels: transparent teal: predicted protein structure, transparent gray: native structure, teal: predicted covalent modification, gray: native covalent modification. A) Schematic describing how RFAA models covalent modifications to 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*).

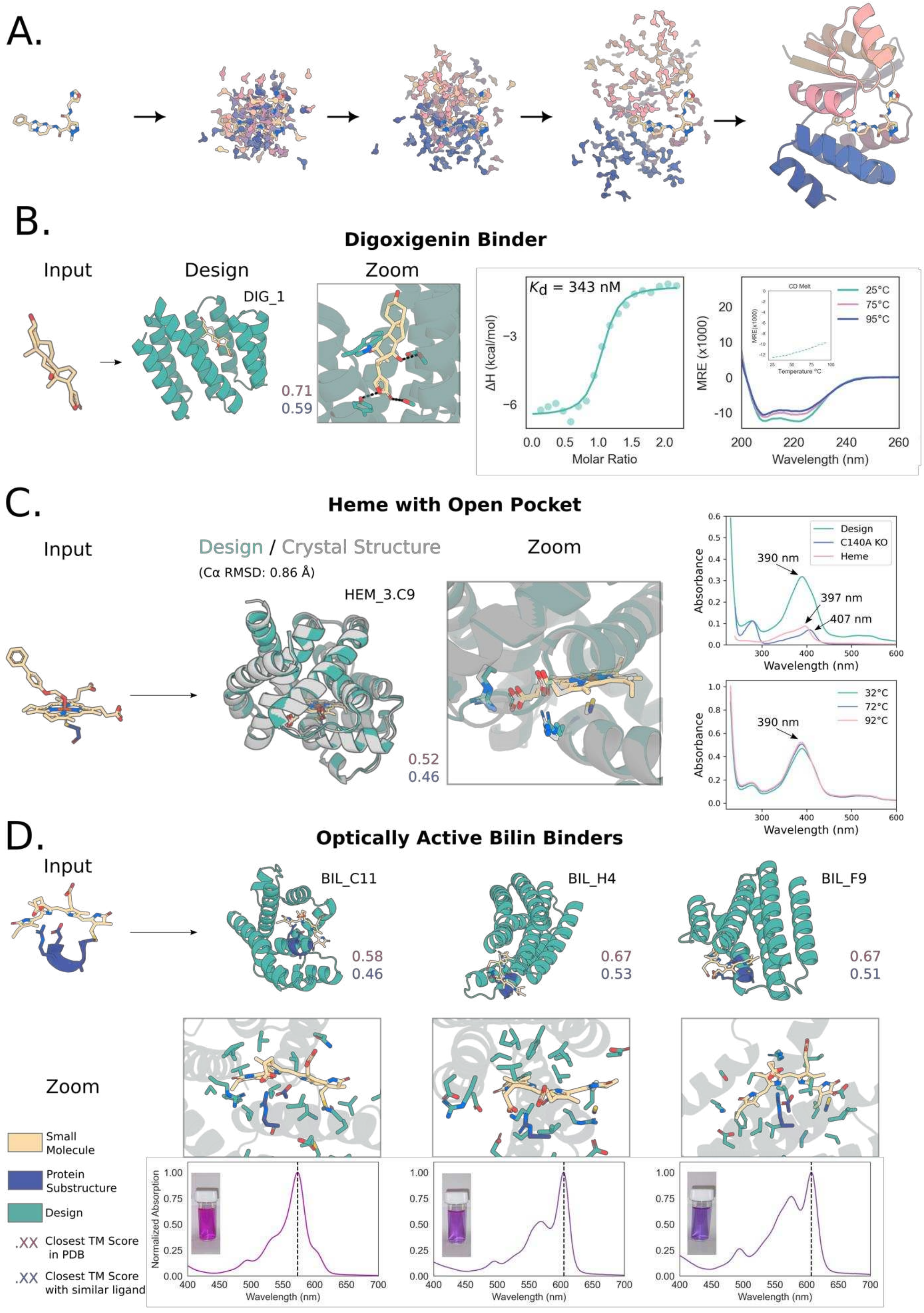
A.
B.
Digoxigenin Binder
Input
Design
Zoom
DIG_1
0.71
0.59
K_d = 343 nM
ΔH (kcal/mol)
-3
-6
0.0
0.5
1.0
1.5
2.0
Molar Ratio
CD Melt
MRE(x1000)
Temperature °C
25°C
75°C
95°C
MRE (x1000)
20
10
0
-10
200
220
240
260
Wavelength (nm)
C.
Heme with Open Pocket
Input
Design / Crystal Structure
(Cα RMSD: 0.86 Å)
HEM_3.C9
Zoom
0.52
0.46
Design
C140A KO
Heme
390 nm
397 nm
407 nm
Absorbance
Wavelength (nm)
32°C
72°C
92°C
390 nm
Absorbance
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
1.00
0.75
0.50
0.25
0.00
400
450
500
550
600
650
700
Wavelength (nm)

Figure 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 (ITC) 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).