



OPEN

Accurate prediction by AlphaFold2 for ligand binding in a reductive dehalogenase and implications for PFAS (per- and polyfluoroalkyl substance) biodegradation

Hao-Bo Guo^{1,2}, Vanessa A. Varaljay¹, Gary Kedziora³, Kimberly Taylor^{1,2}, Sanaz Farajollahi^{1,2}, Nina Lombardo^{1,2}, Eric Harper¹, Chia Hung¹, Marie Gross^{1,4}, Alexander Perminov^{1,5}, Patrick Dennis¹, Nancy Kelley-Loughnane^{1✉} & Rajiv Berry^{1✉}

Despite the success of AlphaFold2 (AF2), it is unclear how AF2 models accommodate for ligand binding. Here, we start with a protein sequence from *Acidimicrobiaceae* TMED77 (T7RdhA) with potential for catalyzing the degradation of per- and polyfluoroalkyl substances (PFASs). AF2 models and experiments identified T7RdhA as a corrinoid iron-sulfur protein (CoFeSP) which uses a nonpseudocobalamin (BVQ) cofactor and two Fe₄S₄ iron-sulfur clusters for catalysis. Docking and molecular dynamics simulations suggest that T7RdhA uses perfluorooctanoic acetate (PFOA) as a substrate, supporting the reported defluorination activity of its homolog, A6RdhA. We showed that AF2 provides processual (dynamic) predictions for the binding pockets of ligands (cofactors and/or substrates). Because the pLDDT scores provided by AF2 reflect the protein native states in complex with ligands as the evolutionary constraints, the Evoformer network of AF2 predicts protein structures and residue flexibility in complex with the ligands, i.e., in their native states. Therefore, an apo-protein predicted by AF2 is actually a holo-protein awaiting ligands.

AlphaFold2¹ (AF2) has achieved near-experimental accuracy for predicting protein structures from the primary sequences. This breakthrough, together with the developments of other tools including RoseTTAFold², allow us to understand the protein structure–function relationships with atomic precision. The performance of AF2, however, was found to produce contradictory results in some assessments. For instance, it is the subject of debate whether AF2 fails to predict the impact of point mutations in protein structure³ and stability⁴; whereas other studies indicated that the structures and phenotypic effects of the point mutations can be correctly predicted⁵ or assisted⁶ by AF2. With these controversies in mind, further modifications to the AF2-predicted structures are required to appropriately understand the protein functions⁷, including the addition of ligands (cofactors and/or substrates). Moreover, as proteins are not static and generally perform functions in the cell but not in crystals, it is important to examine if AF2 can capture protein dynamics in aqueous environments⁸.

Over 60 years before the first glimpse of a protein structure⁹, the “key-lock” model¹⁰ was proposed to describe how proteins perform functions via ligand binding. In this model the protein is described as a lock awaiting the ligand as a key to unleash its function. Unfitted ligands—including water molecules—would fail to unlock the protein. In later studies, especially with the help of protein and protein–ligand complex structures that were becoming increasingly abundant, a refined “induced-fit” model^{11,12} was proposed which accommodates the conformational changes of the protein upon ligand-binding. To illustrate the long-range (allosteric) effect of ligand binding, the conformation selection model was proposed^{13,14}, in which the binding conformations pre-exist in the protein such that the ligand binding could spontaneously occur. Both sequences and structures (shapes, sizes, and locations) of the binding pockets in proteins are thought to evolve to facilitate association with different ligands¹⁵. Although only a small number of representative pockets in proteins have been estimated,

¹Material and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH 45433, USA. ²UES Inc., Dayton, OH 45432, USA. ³GDIT Inc., Wright-Patterson Air Force Base, Dayton, OH 45433, USA. ⁴University of Dayton, Dayton, OH 45469, USA. ⁵Miami University, Oxford, OH 45056, USA. ✉email: nancy.kelley-loughnane.1@us.af.mil; rajiv.berry@us.af.mil

ligand specificity of proteins may emerge in evolution without functional constraints¹⁶. In this regard most of the previously or currently recognized hard-to-degrade chemicals, including polymers¹⁷ and per- and polyfluoroalkyl substances (PFAS)¹⁸, could serve as the “keys” for certain lock proteins.

With AF2 it is likely that every single sequenced protein has its high-resolution 3D structure available in the database^{19,20} or can be accurately predicted. In the Protein Data Bank^{21,22}, however, the majority of the structures derived from experiments (crystallography, cryoEM, NMR, etc.) are complexes, including multimers (often with symmetry)²³, as well as bound with cofactors²⁴ and other ligands²⁵. This information is as important as the protein structures themselves for informing the protein functions and mechanisms, ever since the first solved protein structure of myoglobin⁹. However, the initial structures that we obtain from AF2¹, AlphaFold-multimer²⁶, or AF2Complex²⁷ are apo-proteins, i.e., proteins without ligands. Given the important roles of ligands play in the protein functions, it is crucial to determine whether AF2 is suitable for predicting structure and function of such proteins. In another words, do the apo-proteins predicted by AF2 have the proper binding pockets (cavities) for the ligand binding? Here, we used a multiple-ligand protein to answer this question. As the PFAS substrate (e.g., PFOA) can bind to the protein as a substrate, it has the potential for biodegrading the PFAS contaminants.

A recent work found a bacterium *Acidimicrobiaceae* sp. A6 that, when cultured in the presence of either perfluorooctanoic acid (PFOA) or perfluorooctane sulfonate (PFOS), was able to defluorinate these chemicals with an observed release of fluoride ion, shorter-chain perfluorinated products, and acetate²⁸. The key enzyme for defluorination of PFOA/PFOS was identified as a reductive dehalogenase subunit A (RdhA) in GenBank (id: MK358462.1)²⁸. However, only partial sequence of this enzyme (A6RdhA hereafter) was available with a missing C-terminus of over 100 AAs compared with known reductive dehalogenases including PceA²⁹ and NpRdhA³⁰. Sequence mining starting from the partial A6RdhA sequence revealed a full protein sequence from the bacteria *TMED77* in a metagenomic assembly of the Mediterranean Sea microorganisms³¹, which shares 98% sequence identity with the known part of A6RdhA protein. This protein is referred to as the T7RdhA in present work. It is worth noting that the *TMED77* bacterium belongs to the same *Acidimicrobiaceae* family as *A. sp. A6*.

In the present work, we showed that T7RdhA is a PceA-like protein²⁹ which utilizes two Fe₄S₄ iron sulfur clusters and a norpseudocobalamin (BVQ) cofactor. We constructed AF2 models of T7RdhA, and for the highest-ranked model, both BVQ cofactor and Fe₄S₄ clusters can be put on the binding pockets precisely. Molecular dynamics (MD) simulations were performed on this model with no ligand (apo-form), partially bound by cofactors (either BVQ or Fe₄S₄), or with both cofactors, and with both cofactors and a substrate (PFOA). The results indicate that the AF2 is able to predict the binding pockets for both cofactors and substrates in the protein models, with regard to the binding pockets dynamics³². The model used in the MD simulation was constructed using AF2 V2.0.1 (July 2021 version). A newer version of AF2 V2.2.2 (downloaded in July 2022) was employed to construct additional 90 models and compared with the MD model. High similarity of the new models with the MD model illustrates the reproducibility of AF2. Interestingly, we show that the diversity of AF2 models resemble the MD results. We perform residue-interaction network (RIN) analyses using the MD trajectories of the model with both BVQ, Fe₄S₄, and PFOA. We identified the binding pockets for both cofactors and the substrate in T7RdhA, which will help to search and design proteins for PFAS biosequestration and degradation.

Results

T7RdhA (and potentially A6RdhA) is a CoFeSP. From a sequence similarity network (SSN) constructed using the NCBI nr database³³, T7RdhA and the T7RdhA-like proteins comprise highly conserved residues for the binding of a corrinoid cofactor and two Fe₄S₄ iron-sulfur clusters (see Fig. S1 in the supplementary information, SI). These proteins were termed as corrinoid iron-sulfur proteins (CoFeSPs)³⁴. The corrinoid cofactor or these proteins include the cobalamin (B12) in NpRdhA³⁰ and B12-derivatives such as the norpseudocobalamin (BVQ) in PceA²⁹. NpRdhA uses the B12 cofactor and belongs to an aerobic bacterium *Nitratireductor pacificus*^{30,35}. However, PceA that uses the BVQ cofactor is carried by the anaerobic bacterium *Sulfurospirillum multivorans*^{29,36}. It is likely T7RdhA uses the BVQ cofactor not only the *Acidimicrobiaceae* bacterial family is anaerobic²⁸, but also because T7RdhA belongs to the PceA branch (Fig. S1) in the clustering of the T7RdhA-like proteins from the SSN. A cross-linked binding mode has been found in both PceA- and NpRdhA-like proteins, in which two Fe₄S₄-binding motifs are required for binding of each of the two Fe₄S₄ clusters (Fig. S2). Moreover, we cloned and expressed T7RdhA in *Escherichia coli*, and verified that T7RdhA binds both a corrinoid cofactor and two iron-sulfur clusters (Fig. S3). The network-assisted de novo structured prediction approach and experimental verifications indicate that T7RdhA is a CoFeSP. Besides the cofactors BVQ, Fe₄S₄-A and Fe₄S₄-B, the PFOA substrate is also docked into T7RdhA (Methods), and Fig. 1 shows the binding of all four ligands in T7RdhA.

AlphaFold2 confidence score is consistent with the residue flexibility of T7RdhA complex. We showed previously that the per-residue pLDDT (predicted local distance difference test) scores accompanying the predicted protein models by AF2 also anticipate the residue flexibilities for globular proteins, protein dimers and intrinsically disordered proteins⁸. However, AF2 only provides the apo-forms of the proteins or protein-multimers, and the knowledge of cofactors and/or substrates related to the protein functions can only be acquired from experiments or literature. In the case of T7RdhA, since it is likely a CoFeSP which performs the functions utilizing the corrinoid (BVQ) cofactor and two Fe₄S₄ clusters, and presumably the PFOA substrate can be bound to the active site of the protein for catalysis, we asked if AF2 can predict the binding of these proposed ligands.

To answer this question, we performed MD simulations on five different systems: (1) T7RdhA complexed with BVQ, two Fe₄S₄ clusters (Fe₄S₄-A and Fe₄S₄-B) and the PFOA substrate; (2) apo-T7RdhA with no ligand; (3) T7RdhA complexed with BVQ; (4) T7RdhA complexed with two Fe₄S₄ clusters; and (5) T7RdhA complexed

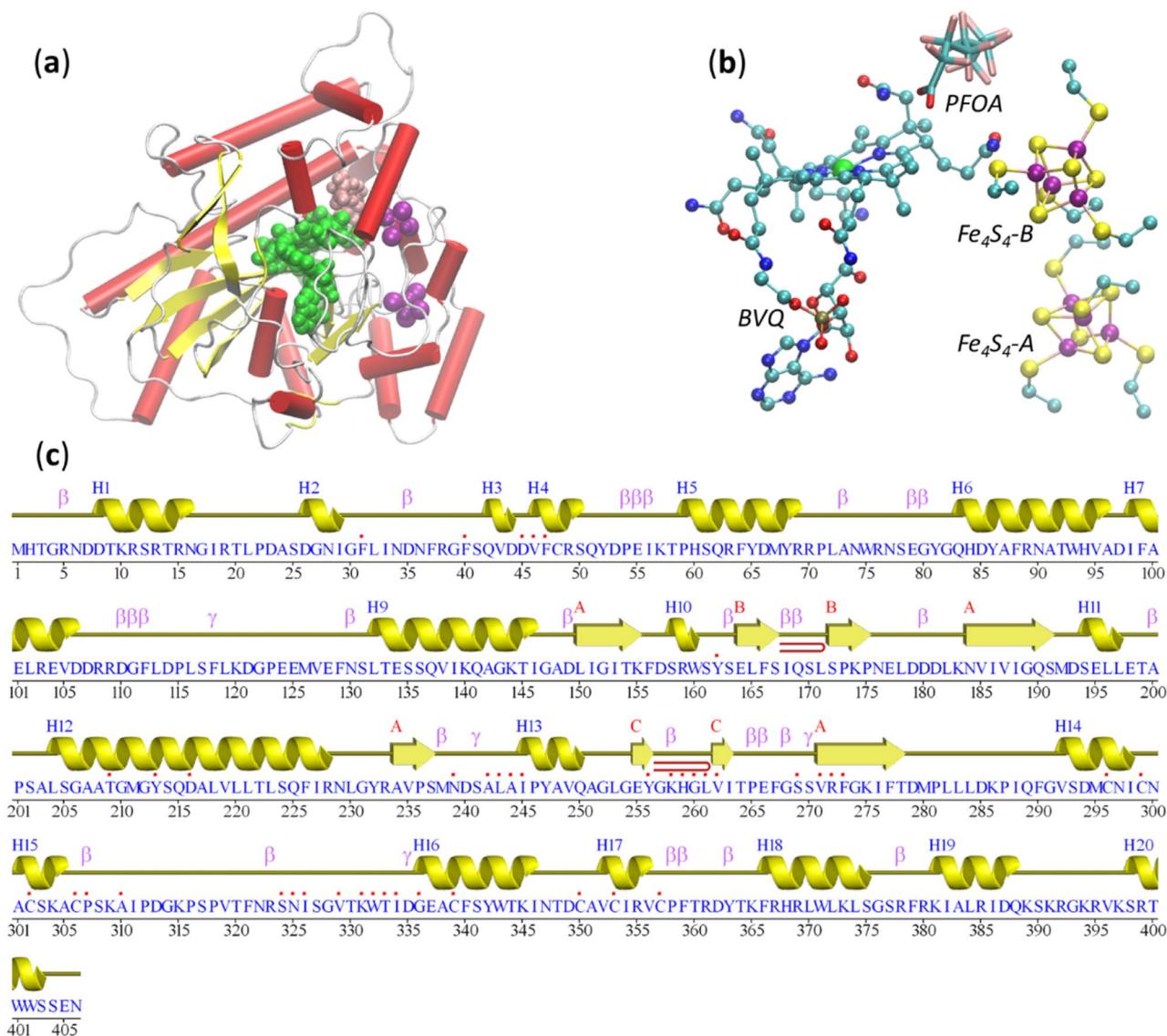


Figure 1. T7RdhA structure. **(a)** The structure of T7RdhA in complex with BVQ (green) cofactor, two Fe₄S₄ (purple) clusters and the PFOA (pink) substrate. In the protein cartoon α -helices are in red, β -sheets are in yellow and coils in white. **(b)** A closer view of the BVQ cofactor, Fe₄S₄ clusters and the binding Cys residues, and the PFOA substrate. Fe in purple, S in yellow, Co in green, F in pink, C in cyan, N in blue and O in red. All structures are plotted using VMD⁶⁶. **(c)** A wire presentation of the secondary structures in T7RdhA plotted by PDBsum³⁷. Note that the red dots on top of the amino acids indicate that the residue is involved in cofactor/substrate binding. The β -strands form three sheets (A, B and C). Positions of β - and γ -turns in the loop regions are labeled.

with the BVQ cofactor and two Fe₄S₄ clusters. T7RdhA is a well-folded globular protein. The binding of cofactors or ligands does not significantly alter the conformation of T7RdhA. However, the residue flexibility vary significantly for all the five systems, as shown in Fig. 2. It has been suggested that the diversity of AF2 models would yield biological insights that might be otherwise ignored from a single snapshot of the protein structure¹. In addition to these five systems, we constructed 320 more AF2 structures and calculated the residue fluctuations among these structures for comparison (line 6 in Fig. 2, see below).

For the well-folded globular proteins, the residue flexibility profiles measured by the root-mean square fluctuation (RMSF) from MD simulations were found to be highly consistent with the AF2-scores from the AF2 predictions, which is a reverse normalization of the pLDDT scores⁸

$$AF2_i = (pLDDT_{\max} - pLDDT_i) / (pLDDT_{\max} - pLDDT_{\min}), \quad (1)$$

where AF2_i is the AF2-score of the i-th residue calculated from the pLDDT-score of the Ca atom of the i-th residue⁸. For systems 1–5, 300 ns MD simulations was performed, and the RMSF profile was calculated using the last 100 ns trajectory. For all five models positive correlation between the AF2-scores and RMSFs were observed (Fig. 2), and the best fit comes from the T7RdhA model in complex with BVQ, two Fe₄S₄ clusters and a PFOA

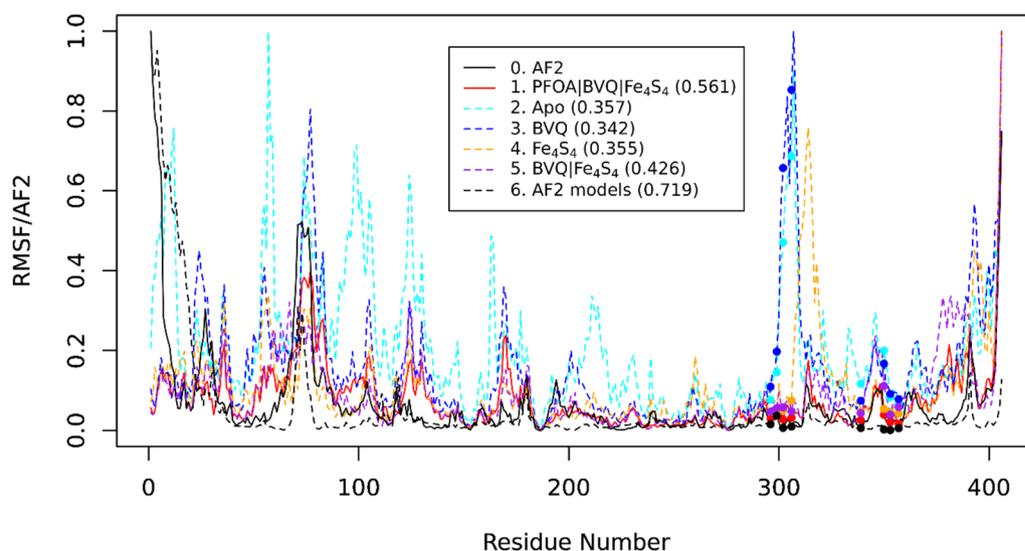


Figure 2. Root-mean square fluctuation profiles (RMSF) of five systems in comparison with the AF2 scores. The dots correspond to the Cys residues that are covalently bonded to the Fe atoms in the Fe_4S_4 clusters. AF2 (0, black solid) is a reverse normalization of the per-residue pLDDT scores of the AF2 protein model. The protein in complex with the PFOA substrate, BVQ cofactor and two Fe_4S_4 clusters (1, red solid) shows the best match with the AF2 scores with Pearson's correlation coefficient (PCC) of 0.561. The other systems include apo-T7RdhA (2, light blue dashed), T7RdhA with BVQ (3, dark blue dashed), T7RdhA with two Fe_4S_4 clusters (4, orange dashed) and T7RdhA with both BVQ and two Fe_4S_4 clusters (5, purple dashed) are also plotted with PCC in parentheses. The RMSF calculated from 320 models (6, black dashed) matches well with the AF2-scores (PCC=0.719).

substrate (system 1)—except for an inconsistency at the C-terminus region for which AF2 anticipates it highly flexible but the RMSF from MD indicates it is instead relatively rigid. Nevertheless, for the important binding regions such as three β -sheets and the helices H12 to H16 (Fig. 2), RMSF of system 1 is highly consistent with the AF2-scores. In contrast, systems 2 and 3 indicate that without the Fe_4S_4 clusters, the regions containing the binding Cys residues (H14 to H16, including the loop region between H15 and H16) are highly flexible, contradicting the AF2 predictions. In system 4, binding of the Fe_4S_4 lead to relatively small flexibility of the binding Cys residues, however, the loop region between H15 and H16 still shows significantly higher flexibility compared with the AF2 prediction. System 5 also shows better consistency between MD and AF2 prediction, slightly lower than system 1. We also measured the root-mean square deviation (RMSD) of all five systems. The mean RMSD of the last 10 ns are calculated as shown in Table 1, and again, system 1 shows the lowest RMSD value. The RMSD profiles from the last 100 ns of all five systems are shown in Fig. S4 in the SI.

Diversity of AF2 models resembles the MD simulation. The original AF2 publication suggested that diversity of AF2 models (i.e., via multiple runs) may yield new biological insights by predicting alternative configurations of the proteins¹. In this work, the T7RdhA model for the MD simulations was constructed using an old AF2 version (V2.0.1, released July 2021). We used the newer AF2 version (V2.2.2, downloaded July 2022) to construct 320 T7RdhA models (64 independent runs, each gives 5 models). These models show relatively low

System	Description	RMSF ¹ (Å)	PCC ²	p value ²	RMSD ³ (Å)
1	T7RdhA + BVQ + Fe_4S_4 -A/ Fe_4S_4 -B + PFOA	0.8 ± 0.6	0.561	5.2 × 10 ⁻³⁵	1.4
2	Apo-T7RdhA	0.9 ± 0.4	0.357	1.2 × 10 ⁻¹³	1.7
3	T7RdhA + BVQ	1.0 ± 0.6	0.342	1.4 × 10 ⁻¹²	2.4
4	T7RdhA + Fe_4S_4 -A/ Fe_4S_4 -B	1.1 ± 0.7	0.355	1.6 × 10 ⁻¹³	2.5
5	T7RdhA + BVQ + Fe_4S_4 -A/ Fe_4S_4 -B	0.9 ± 0.5	0.426	2.7 × 10 ⁻¹⁹	1.5
6	AF2 models ⁴	0.5 ± 0.6	0.719	7.1 × 10 ⁻⁶⁶	0.5

Table 1. Comparing protein dynamics metrics with the AF2 predictions. ¹The root-mean-square fluctuation (RMSF) from a 100 ns MD simulation after 200 ns equilibration. ²Pearson's correlation coefficient (PCC) and p-values between the RMSF and the predicted AF2-score. ³Mean RMSD from the last 10 ns trajectory referenced with the initial structure of the 100 ns MD. ⁴320 AF2 models are used to calculate the RMSF of all residues. The RMSD is averaged from all 320 models referenced to the model used in the MD simulations. The large error (0.6) may originate from a single rare model (Fig. S5).

RMSD to the MD model (Fig. S5 in the SI), indicating the reproducibility of the AF2 algorithm. By combining the configuration of all 105 new T7RdhA models, we also calculated the residue RMSF values and compared with the AF2 scores from the MD model (Fig. 2). Interestingly, this RMSF profile (black dashed line) shows a PCC = 0.719 ($p = 7.1 \times 10^{-66}$) to the AF2-score profile. In this profile not only the BVQ- or Fe_4S_4 -binding regions, but also the dynamic N-terminus is consistent with the AF2-scores. Therefore, from the protein sequence, AF2 not only provided dynamics information of all residues via the (pLDDT or AF2-scores)⁸, it seems that multiple AF2 runs can mimic the MD simulation, i.e., the ensemble of AF2 models were covered by the structural ensemble generated by MD simulation; Fig. S5a shows the structural variations of all AF2 models.

Residue interaction network and the cofactor/ligand binding modes of T7RdhA. The protein residue distance maps usually defined as the distance d_{ij} between the C_β atoms (C_α for Gly) of residues i and j , and there is a contact between these two residues if d_{ij} is shorter than a criterion (e.g., 8 Å)³⁸. A residue interaction network (RIN)³⁹ can be constructed based on the contact map in which all residues are regarded as vertices and the contacts as edges. Here, we used a modified approach to define d_{ij} as the shortest distance between non-hydrogen atoms of two residues, and a cutoff of 3.5 Å is used to identify contacts. This approach would avoid potential false contact assignments, see Methods and Fig. S6 in the SI. The distance map of system 1 obtained from the modified approach is shown in Fig. 3a, which has very similar patterns as the predicted aligned error (PAE) map provided by AF2, as shown in Fig. 3b. The RIN for system 1, in which the cofactors (BVQ, Fe_4S_4 -A and Fe_4S_4 -B) and the substrate (PFOA) were treated as individual vertices, constructed from the final snapshot of the 300 ns MD is shown in Fig. 3c.

To capture the dynamics of the RIN, we constructed the RIN every 1 ns from the last 100 ns MD and monitored the residues that interact with the BVQ cofactor, both Fe_4S_4 clusters, as well as the PFOA substrate. The distributions of these residues are shown in Fig. 4. We observed more residues interact with the BVQ than those that interact with the two Fe_4S_4 clusters and PFOA combined. The motifs involved in BVQ binding include short helical segments H2, H3, H4, and longer helices H12, H13, H16, and H17; two strands from β -sheet A and the β -hairpin C are involved in BVQ binding (see Fig. 1c for all motif names). The β -hairpin C also interacts with Fe_4S_4 -A at the loop region via two positively charged residues H259 and K258, as $\text{Fe}_4\text{S}_4\text{Cys}_4$ carries negative charge (-2 for oxidized and -3 for reduced states, respectively). In this simulation we used a reduced Fe_4S_4 -A (distal), an oxidized Fe_4S_4 -B (proximal) and a reduced BVQ (with Co^{+1}), see methods. H16 is involved in the interactions for BVQ, Fe_4S_4 -B and PFOA. In particular, the aromatic residue W343 that is conserved in other T7RdhA-like proteins (Fig. S2) shows interactions with BVQ (82%), PFOA (80%) and Fe_4S_4 -B (15%). Y213 from H12 has been considered to mediate the reductive dehalogenation in PceA (Y246 of *S. multivorans* PceA)²⁹, and its interactions with BVQ (98%) and PFOA (40%) may be needed for potential defluorination. F47 from a small helix H4 also interacts with BVQ (100%) and PFOA (90%). We noticed that the residues interact with PFOA ($>50\%$) are either aromatic (Y68, Y65, F47, W343, F64, W93, F340) or positively charged (R89), which may be a unique feature of the binding pocket for PFAS substrates.

Discussion

AlphaFold2 correctly predicts cofactor/ligand binding in T7RdhA. AF2 opens an avenue in biology on which the functions and interactions mediated by proteins can be understood with the assistance of highly accurate atomic models. However, the structures predicted by AF2, either single-chain monomers or multi-chain oligomers, are in apo-forms, i.e., unbound form. Even the necessary solvents are missing in the structures predicted by AF2. Cofactors play an important, sometimes essential role in protein folding and functions⁴⁰. Folding and functions of proteins may also be assisted by the substrate that they bind⁴¹. We asked how reliable are the AF2 models in depicting the structures and dynamics of proteins upon cofactor and/or ligand binding? This is a critical question to answer for protein systems with cofactor/substrate, especially for understanding the interactions among them, as well as for protein–protein interactions.

Previous publications discussed the above question on ligand binding⁴², peptide binding⁴³, and protein–protein interactions⁴⁴. In this work, the functional T7RdhA structure incorporates the natural corrinoid norpseudo cobalamin (BVQ)⁴⁵, together with two Fe_4S_4 iron-sulfur clusters (Fe_4S_4 -A and Fe_4S_4 -B), which is known as the “Nature’s modular structures”⁴⁶. We showed that when cofactors (BVQ/ Fe_4S_4) and substrate (PFOA) are present in the correct pockets, the residue flexibility calculated from molecular dynamics simulations can best rationalize the AF2-scores by AlphaFold2, which is an inverse normalization of the pLDDT scores. In the complex model, the residue distance map also mirrors the predicted aligned error map by AlphaFold2. Our results indicate that the AF2 structures already have the pre-built pockets for the correct cofactors and ligands. We also showed that multiple AF2 structures (320 T7RdhA models in the present work) can also capture the protein dynamics. The diversity of protein structures, in our opinion, originates from protein dynamics and can be recaptured by AF2 in the structure modeling.

A processual view of protein structure–function relationships. The protein function is determined by the protein structure. However, a static protein does not perform the function without dynamics and interactions. The processual nature of reality⁴⁷ applies to all biomolecules, including proteins. We collected the structures of different systems (systems 1–5 in Fig. 2) during the MD simulation, together with selected AF2 models (system 6), and compared these snapshots in Fig. 5.

The structures of all models after 300 ns MD are similar to the original AF2 model (Fig. 5b). Without MD trajectories and the residue flexibilities, it would be difficult to tell which system has a ligand or ligands. Our results support that the protein interactions and functions are based on their intrinsic processual nature⁴⁷. For

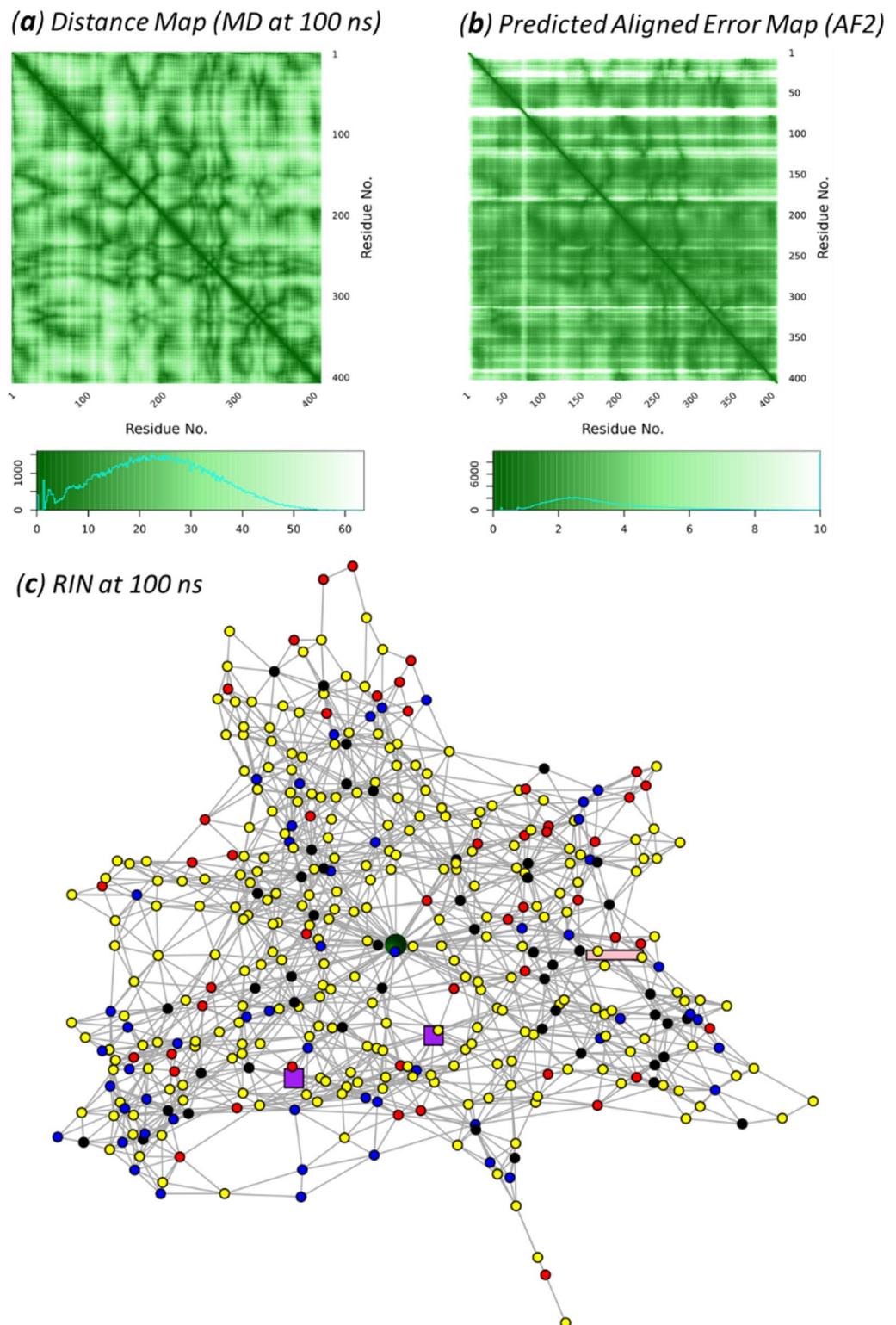


Figure 3. T7RdhA distance map, PAE map and RIN. **(a)** Distance map of T7RdhA at 100 ns compared with **(b)** the PAE map from AF2. **(c)** The RIN at 100 ns constructed from the contact map. The BVQ cofactor in sphere (dark green), the Fe₄S₄ clusters in square (purple) and PFOA in rectangle (pink). Blue nodes are positively charged residues (Lys and Arg) and red are negatively charged residues (Glu and Asp); black nodes are aromatic residues (His, Tyr, Phe and Trp); and all other residues in yellow. The PIN is plotted using the R package igraph⁵³.

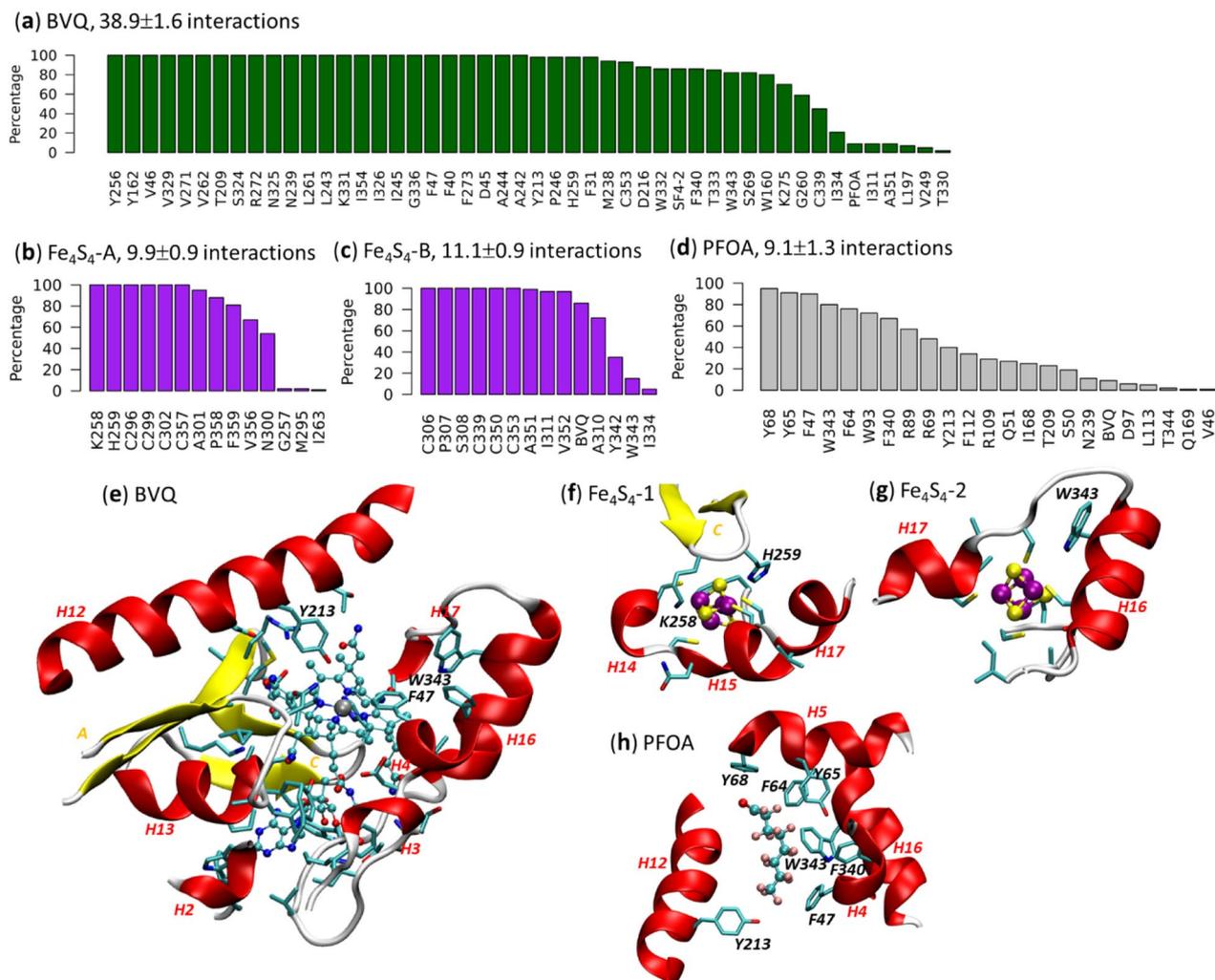


Figure 4. The protein-cofactor/ligand interactions. Distribution of interaction residues from 100 RINs (constructed from snapshots of a 100 ns MD) to (a) BVQ cofactor, (b) Fe₄S₄-A cluster, (c) Fe₄S₄-B cluster and (d) PFOA substrate. Number of interactions (mean ± standard deviation) detected in all RINs are indicated. All residues are ranked by the percentage of interactions observed in all RINs. The y-axis indicates the percentage of the interactions in all RINs. Representative clusters for the interactions centering (e) BVQ, (f) Fe₄S₄-A, (g) Fe₄S₄-B and (h) PFOA. Motifs (sequences in Fig. 1c) that contain interacting residues and some of the important residues are labeled. Cobalt is colored in silver, iron in purple, fluorine in pink, carbon in cyan, nitrogen in blue and oxygen in red. All structures in (e)–(h) are plotted using VMD⁶⁶.

other AF2 models such as those of missense mutations³, it might not be fair to make a judgement based on a static configuration.

Implications for PFAS biodegradation. The persistence and accumulation of per- and polyfluoroalkyl substances, or PFASs, in the environment, and their adverse effects on human health have led to the current global concern⁴⁸. The T7RdhA sequence is highly similar to the partial sequence of A6RdhA from the *Acidimicrobiaceae* sp. A6 which degrade both PFOA and PFOS under anaerobic conditions^{28,49}. Nevertheless, the full A6RdhA sequence and the defluorination mechanisms remain unclear. From the structure modeling and MD simulations, we confirmed the participation of both corrinoic cofactors (BVQ) and iron-sulfur clusters (Fe₄S₄) by experiment. The binding mode of the cofactors and the PFOA ligand have been identified using a dynamic residue interaction network from the MD trajectories. We also showed that AF2 combined with MD simulation can help to identify proteins with targeted functions such as PFAS bioremediation.

Methods and materials

Multiple-sequence alignment & sequence similarity network. A combined A6RdhA/T7RdhA Hidden Markov Model (HMM) was constructed from 529 non-redundant similar sequences identified via blastp from the NCBI and UniProt KB databases. Briefly, these sequences were identified to clade together (with a consensus support value of 100) with A6RdhA/T7RdhA in an amino acid tree using MAFFT v7.453⁵⁰ multiple-sequence alignment, and were then used to construct an HMM profile using the program HMMer

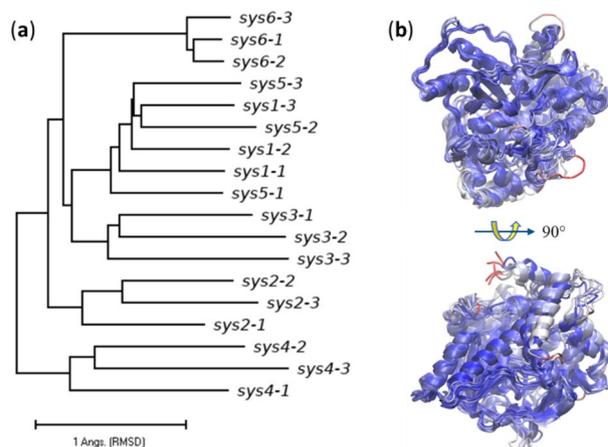


Figure 5. Comparisons of protein structures during MD and AF2 models. **(a)** A structure-based phylogenetic tree using the protein structure snapshots during the MD at different simulation time (100 ns, 200 ns and 300 ns for snapshots 1–3) for the MD systems 1–5 and AF2 structures (system 6), see Fig. 2. The initial MD structure is sys6-1, and sys6-2 and sys6-3 has the RMSD values 0.391 and 0.937 Å to sys6-1, respectively. **(b)** Superimposed structures of systems 1–6 colored by RMSD to the initial MD structure (sys6-1). Systems 1–5 were collected from the MD simulations after 300 ns. A BWR color scheme was applied with blue for low, red for high and white for in-between RMSD values. The overall RMSD for all pairs of structures plotted in the phylogenetic tree is 2.45 ± 0.57 Å. The superimposed structures are plotted using VMD⁶⁶.

(v3.3.2)⁵¹. The first portion of the NCBI non-redundant database, nr00 (8,812,511 sequences), was queried using this HMM profile using the HMMer default threshold values. The resulting 1279 (including T7RdhA) sequences were submitted to the EFI (Enzyme Function Initiative) enzyme similarity tool for generation of the sequence similarity network (SSN) with values $\leq 10^{-5}$ and an alignment cutoff of 20 corresponding to an id% of ~ 30 ⁵². Network clustering and the T7RdhA clique identification was performed using the *igraph* package in R⁵³. The multiple-sequence alignment by MAFFT was visualized using WebLogo (v3.6.0)⁵⁴. TM-align⁵⁵ was used for structure alignment and RMSD calculations. The calculated RMSD matrix was converted to phylogeny using the *ape* package in R⁵⁶, and visualized by Mega-X⁵⁷.

AlphaFold 2 structure predictions. The T7RdhA model used in the MD simulation and other T7RdhA-like proteins models (all 39 models in the SSN shown in Fig. S1) were constructed using AlphaFold2 V2.0.1 (installed in July 2021). 320 more T7RdhA models for the protein-structure-based RMSF profile in Fig. 2 (system 6), were built by a newer version of AlphaFold2 (V2.2.2, installed in July 2022)¹.

Molecular dynamics simulations. The molecular dynamics simulations were performed using NAMD⁵⁸. The CHARMM force field (c36m)^{59,60} was employed for the protein and a modified TIP3P model⁶¹ for the solvent water molecules. The CHARMM-format force field parameters of norpseudo-B12 (BVQ)⁶² and Fe₄S₄ iron sulfur cluster⁶³ under different redox states have been adopted. The force field parameters of the PFOA molecule were derived from the TEAM (Transferable, Extensible, Accurate and Modular) force field in the Direct Force Field (DFE, v7.2)⁶⁴ software, and have been listed in the Appendix-1 in the SI.

The BVQ and Fe₄S₄ cofactors in the crystallographic structures of PceA (e.g., 4UQU²⁹) can be superimposed very well, with the eight Cys residues precisely bound the Fe₄S₄ iron atoms. We used TM-align⁵⁵ to calculate the rotation-translation matrix between the PceA template (pdb 4UQU²⁹) and the AF2 models, then applied this matrix to get the initial coordinates of both BVQ and Fe₄S₄ cofactors. All hydrogen atoms have been added using the HBuild function of CHARMM⁶⁵. The covalent bond between Fe₄S₄ cofactors and their binding Cys residues were generated using the Patch function of CHARMM⁶⁵. The Appendix-2 in the SI shows the details for constructing both the oxidized and reduced Fe₄S₄ clusters in the model for MD. The whole system was put in a solvent box with H₂O molecules added at least 15 Å to the edge of the protein system. The solvation and neutralization (using Na⁺ and/or Cl⁻) were carried out by the Solvate and Autoionization packages of VMD⁶⁶. A reduced BVQ (Co(I)), oxidized Fe₄S₄-B (the proximal, Fe₄S₄(Cys)₄²⁻) and reduced Fe₄S₄-A (the distal, Fe₄S₄(Cys)₄³⁻) were used in the MD simulations.

After solvation and neutralization, the whole system was optimized by 50,000 steps. Then the temperature of the system was “naturally” increased to 300 K with a rate of 0.001 K/timestep. A constant-pressure, constant-temperature (NPT) ensemble was used in the MD simulation with the system pressure of 1 atm and temperature of 300 K maintained by the Langevin piston controls. The *rigidBonds* option was applied to fix the bond lengths involving hydrogen atoms and a timestep of 2 fs was used for the simulations. The van der Waals interaction cutoff switching was set as between 9 and 11 Å. For the long-range interactions, the particle mesh Ewald summation with a grid spacing of 1.35 Å was applied. 310 ns MD simulations were performed for all systems (Fig. 2) and the last 100 ns were taken for further analysis.

Residue interaction network. The residue interaction network (RIN) or contact map of a protein was based on the distance map with a criterion^{38,39}. A common approach, for example, is to measure the C_{α} - C_{β} distances (C_{α} for Gly), and if the measured distance between residues R_i and R_j is shorter than 8 Å, then there is a contact between R_i and R_j . This approach, however, we found may lead to incorrect assignment (Fig. S6). We adopted an alternative approach. Considering the hydrogen bond interaction X-H...Y (where X/Y can be C, N, O, S in proteins), the distance between X and Y for a typical H-bonds are in the range of 2–3 Å, and is ~3.5 Å for a C-H...O hydrogen bond in protein⁶⁷. Here, for residues (vertices) R_i and R_j we define the distance d_{ij} as the shortest distance between all heavy atoms. The distance map under this approach (Fig. 3a) agree well with the PAE map predicted by AF2 (Fig. 3b). The contact map is further defined based on the distance map: if d_{ij} is shorter than 3.5 Å, we define an interaction (edge) between R_i and R_j . We then construct a binary adjacency matrix (1 for interaction and 0 for non-interaction) based this definition. The network analysis was performed using the *igraph*⁵³ package in R. The distance analysis was performed using the *bio3d*⁶⁸ package in R. The BVQ cofactor, the Fe₄S₄-A and Fe₄S₄-B clusters, and the PFOA substrate was treated as a residue (vertex) in the RIN.

Ligand binding. AutoDock Vina (V1.2.0)⁶⁹ was used for ligand docking. Using the T7RdhA-BVQ- Fe₄S₄ system (system 5), after 10 ns MD equilibration, the PFOA ligand was docked into the protein complex (solvent and ions removed), and the top-score ligand was used to construct system 1. The force field parameters of the PFOA ligand can be found in the Appendix in SI.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The code for calculating the rotation-translation matrix using TM-align, the R codes for data analyses and visualizations are available upon request (H.-B.G., haobo.guo.ctr@us.af.mil).

Received: 12 September 2022; Accepted: 21 February 2023

Published online: 11 March 2023

References

- Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
- Buel, G. R. & Walters, K. J. Can alphafold2 predict the impact of missense mutations on structure?. *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).
- Pak, M. A. *et al.* Using AlphaFold to predict the impact of single mutations on protein stability and function. bioRxiv 2021. <https://doi.org/10.1101/2021.09.19.460937>.
- McBride, J. M., Polev, K., Reinharz, V., Grzybowski, B. A. & Tlustý, T. AlphaFold2 can predict structural and phenotypic effects of single mutations. arXiv 2022: 2004.068860[q-bio.BM]. <https://arxiv.org/abs/2204.06860>.
- Schmidt, A. *et al.* Predicting the pathogenicity of missense variants using features derived from AlphaFold2. bioRxiv 2022. <https://doi.org/10.1101/2022.03.05.483091>.
- Bagdonas, H., Fogarty, C. A., Fadda, E. & Agirre, J. The case for post-predictional modifications in the AlphaFold protein structure database. *Nat. Struct. Mol. Biol.* **28**, 869–870 (2021).
- Guo, H.-B. *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci. Rep.* **12**, 10696 (2022).
- Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, 662–666 (1958).
- Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.* **27**, 2985–2993 (1894).
- Koshland, D. E. Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **44**, 98–104 (1958).
- Koshland, D. E. Jr. The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* **33**, 2375–2378 (1994).
- Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
- Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **3**, 19 (2011).
- Najmanovich, R. J. Evolutionary studies of ligand binding sites in proteins. *Curr. Opin. Struct. Biol.* **45**, 85–90 (2017).
- Skolnick, J. & Gao, M. Interplay of physics and evolution in the likely origin of protein biochemical function. *Proc. Natl. Acad. Sci. USA* **110**, 9344–9349 (2013).
- Lu, H. *et al.* Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
- Wackett, L. P. & Robinson, S. L. The ever-expanding limits of enzyme catalysis and biodegradation: Polyaromatic, poly chlorinated, polyfluorinated and polymeric compounds. *Biochem. J.* **477**, 2875–2891 (2020).
- Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
- Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
- Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
- Goodsell, D. S. *et al.* RCSB protein data bank: Enabling biomedical research and drug discovery. *Protein Sci.* **29**, 52–65 (2020).
- Duarte, J. M., Dutta, S., Goodsell, D. S. & Burley, S. K. Exploring protein symmetry at the RCSB protein data bank. *Emerg. Top. Life Sci.* <https://doi.org/10.1042/ETLS20210267> (2022).
- Mukhopadhyay, A. *et al.* Finding enzyme cofactors in protein data bank. *Bioinformatics* **35**, 3510–3511 (2019).
- Stierand, S. & Rarey, M. Drawing the PDB: Protein–ligand complexes in two dimensions. *ACS Med. Chem. Lett.* **1**, 540–545 (2010).
- Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021. <https://doi.org/10.1101/2021.10.04.463034>.
- Gao, M., An, D. N., Parks, J. M. & Skolnick, J. AF2Complex predicts physical interactions in multimeric proteins with deep learning. *Nature Commun.* **13**, 1744 (2022).
- Huang, S. & Jaffé, P. R. Defluorination of Perfluorooctanoic Acid (PFOA) and Perfluorooctane Sulfonate (PFOS) by Acidimicrobium sp. Strain A6. *Environ. Sci. Technol.* **53**, 11410–11419 (2019).
- Bommer, M. *et al.* Structural basis for organohalide respiration. *Science* **346**, 455–458 (2014).
- Payne, K. A. *et al.* Reductive dehalogenase structure suggests a mechanism for B12-dependent dehalogenation. *Nature* **517**, 513–516 (2015).
- Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: A resource for marine microbiology. *PeerJ* **5**, e3558 (2017).
- Stank, A., Kokh, D. B., Fuller, J. C. & Wade, R. C. Protein binding pocket dynamics. *Acc. Chem. Res.* **49**, 809–815 (2016).

33. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
34. Cooper, C. J. *et al.* Structure determination of the HgcAB complex using metagenome sequence data: Insights into microbial mercury methylation. *Commun. Biol.* **3**, 320 (2020).
35. Halliwell, T., Fisher, K., Payne, K. A. P., Rigby, S. E. J. & Leys, D. Catabolic reductive dehalogenase substrate complex structures underpin rational repurposing of substrate scope. *Microorganisms* **8**, 1344 (2020).
36. Kunze, C. *et al.* Cobamide-mediated enzymatic reductive dehalogenation via long-range electron transfer. *Nat. Commun.* **8**, 15858 (2017).
37. Laskowski, R. A. & Thornton, J. M. PDBsum extras: SARS-CoV-2 and AlphaFold models. *Prot. Sci.* **31**, 283–289 (2022).
38. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2022).
39. Clementel, D. *et al.* Ring 30: Fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Res.* **50**, W651–W656 (2022).
40. Wittung-Stafshede, P. Role of cofactors in protein folding. *Acc. Chem. Res.* **35**, 201–208 (2002).
41. Xu, Q., Guo, H.-B., Wlodawer, A. & Guo, H. The importance of dynamics in substrate-assisted catalysis and specificity. *J. Am. Chem. Soc.* **128**, 5994–5995 (2006).
42. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: Enriching the AlphaFold models with ligands and co-factors. bioRxiv 2021. <https://doi.org/10.1101/2021.11.26.470110>.
43. Chang, L. & Perez, A. AlphaFold encodes the principles to identify high affinity peptide binders. bioRxiv 2022. <https://doi.org/10.1101/2022.03.18.484931>.
44. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein–protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
45. Krautler, B. *et al.* The cofactor of tetrachloroethene reductive dehalogenase of *Dehalospirillum multivorans* is norpseudo-B12, a new type of natural corrinoid. *Helvetica* **86**, 3698–3716 (2003).
46. Beinert, H., Holm, R. H. & Münck, E. Iron-sulfur clusters: Nature's modular, multipurpose structures. *Science* **277**, 653–659 (1997).
47. Jaeger, J. & Monk, N. Everything flows: A process perspective of life. *EMBO Rep.* **16**, 1064–1067 (2015).
48. Buck, R. C. *et al.* Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integr. Environ. Assess. Manag.* **7**, 513–541 (2011).
49. Ruiz-Uriguen, M., Shuai, W., Huang, S. & Jaffe, P. R. Biodegradation of PFOA in microbial electrolysis cells by *Acidimicrobiaceae* sp. *Strain A6*. *Chemosphere* **292**, 133506 (2022).
50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
51. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
52. Zallot, R., Oberg, N. & Gerlt, J. A. The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182 (2019).
53. Han, W.-S., Lee, J., Pham, M.-D. & Yu, J. X. iGraph: A framework for comparisons of disk-based graph indexing techniques. *Proc. VLDB Endow.* **3**, 449–459 (2010).
54. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
55. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
56. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
57. Kumar, S., Stecher, G., Li, M., Niyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
58. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
59. Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
60. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
61. Price, D. J. & Brooks, C. L. 3rd. A modified TIP3P water potential for simulation with Ewald summation. *J. Chem. Phys.* **121**, 10096–10103 (2004).
62. Pavlova, A., Parks, J. M. & Gumbart, J. C. Development of CHARMM-compatible force-field parameters for cobalamin and related cofactors from quantum mechanical calculations. *J. Chem. Theory Comput.* **14**, 784–798 (2018).
63. Carvalho, A. T. P. & Swart, M. Electronic structure investigation and parametrization of biologically relevant iron-sulfur clusters. *J. Chem. Inf. Model.* **54**, 613–620 (2014).
64. Direct Force Field 7.0, Aeon Technology, Inc. San Diego (2011).
65. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
66. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
67. Guo, H.-B., Beahm, R. F. & Guo, H. Stabilization and destabilization of the C₈-H...OC hydrogen bonds involving proline residues in helices. *J. Phys. Chem. B* **108**, 18065–18072 (2004).
68. Grant, B. J., Skjaerven, L. & Yao, X.-Q. The Bio3D packages for structural bioinformatics. *Protein Sci.* **30**, 20–30 (2021).
69. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).

Acknowledgements

This work was supported by funding from the OUSD (R&E) ARAP Program. The structural modeling and MD simulations were performed using the DoD HPC. We appreciate Dr. Jerry Parks for the cobalamin force field parameters, and Dr. Marcel Swart for providing the Fe₄S₄ iron sulfur cluster force field parameters. We thank Dr. Peter Jaffe for helpful discussions on the *Acidimicrobiaceae* sp. A6 organism and the mechanism of the A6RdhA enzyme. We appreciate the MAPS TEAM from the Summit Country Day School for the 3D print of the T7RdhA-ligand model (Fig. S7).

Author contributions

Conceptualization, H.-B.G., R.B.; methodology, H.-B.G., V.V., S.F., P.D., R.B.; validation, all authors; writing—original draft preparation, H.-B.G.; writing—review and editing, all authors; project administration. N.K.-L., R.B.; funding acquisition, N.K.-L., R.B. All authors have read the final revision of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-30310-x>.

Correspondence and requests for materials should be addressed to N.K.-L. or R.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023