Biochemistry

pubs.acs.org/biochemistry Article

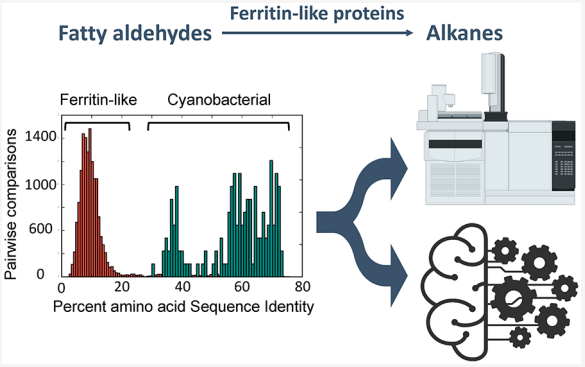
Discovery, Design, and Structural Characterization of Alkane-Producing Enzymes across the Ferritin-like Superfamily

Wai Shun Mak, XiaoKang Wang, Rigoberto Arenas, Youtian Cui, Steve Bertolani, Wen Qiao Deng, Ilias Tagkopoulos, David K. Wilson, and Justin B. Siegel*

Cite This: *Biochemistry* 2020, 59, 3834–3843 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: To complement established rational and evolutionary protein design approaches, significant efforts are being made to utilize computational modeling and the diversity of naturally occurring protein sequences. Here, we combine structural biology, genomic mining, and computational modeling to identify structural features critical to aldehyde deformylating oxygenases (ADOs), an enzyme family that has significant implications in synthetic biology and chemoenzymatic synthesis. Through these efforts, we discovered latent ADO-like function across the ferritin-like superfamily in various species of Bacteria and Archaea. We created a machine learning model that uses protein structural features to discriminate ADO-like activity. Computational enzyme design tools were then utilized to introduce ADO-like activity into the small subunit of *Escherichia coli* class I ribonucleotide reductase. The integrated approach of genomic mining, structural biology, molecular modeling, and machine learning has the potential to be utilized for rapid discovery and modulation of functions across enzyme families.

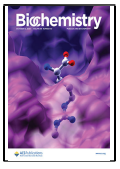


There has been considerable progress made in the ability to engineer protein function over the past two decades through the use of directed evolution and rational design techniques.[1−6] Current methods often focus on using a small set of well-established proteins into which a highly targeted or evolutionarily selected set of mutations are introduced.[7−9] While powerful, current engineering practices leave the vast majority of protein sequence and functional space unexplored. Through recent advances in DNA sequencing and computational modeling, new paradigms for protein engineering are being made possible.[10−14] Here we explore how to integrate established rational and computational protein engineering approaches with genomic mining- and machine learning-based methodologies to identify and engineer functionality within the ferritin-like superfamily.

The proteins in the ferritin-like superfamily are defined by the signature four-helix bundle fold with a carboxylate-bridged, non-heme dimetal center coupled to a structurally divergent collection of active sites.[15] Proteins within this superfamily, while sharing a geometrically similar dimetal center (Figure 1B), fulfill many different functions, including fatty acid and DNA biosynthesis, iron storage, and the relief of oxidative stress.[15] These enzymes are also frequently found to possess promiscuous activities beyond their natural functions. For example, methane monooxygenase is found to catalyze epoxidation, N-oxygenation, dehalogenation, and desaturation of benzylic compounds, while its natural physiological function is hydroxylation of methane.[16] Similarly, steroyl-acyl-carrier protein desaturase can catalyze hydroxylation and sulfoxidation reactions.[16] A particularly interesting member of the ferritin-like superfamily is the aldehyde deformylating oxygenase (ADO) from cyanobacteria.[17] This family of enzymes has garnered significant attention since its discovery due to its notable implications in synthetic biology and chemoenzymatic synthesis applications. This enzyme catalyzes the conversion of fatty aldehydes to n-1 alkanes, enabling organisms to produce drop-in fuel molecules[18] (Figure 1A). In addition, biophysical and mechanistic studies have revealed a unique reaction mechanism that has not been observed in Nature.[19−29] In light of the catalytic promiscuity observed for other members of the ferritin-like superfamily, we hypothesized that ADO-like activity (alkane production activity from fatty aldehyde) may exist in other non-heme di-iron oxygenase scaffolds if the relevant structural features are present.

Through genomic mining efforts, we discovered enzymes with ADO-like activity occur widely among Bacteria and

Received: August 10, 2020
Revised: September 16, 2020
Published: September 16, 2020



https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834–3843

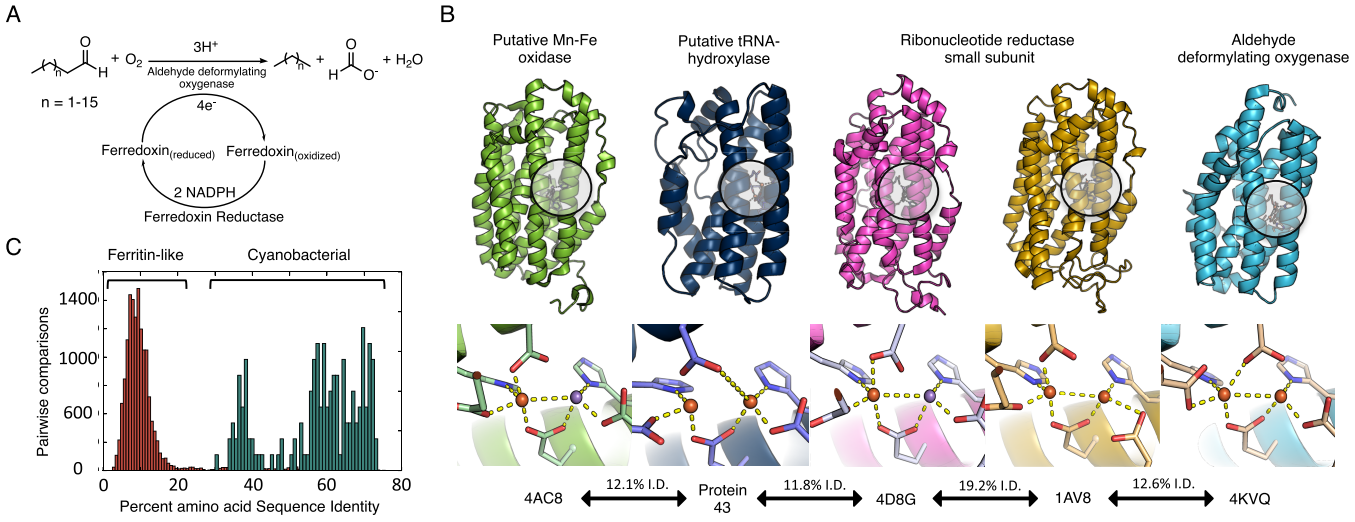


Figure 1. Aldehyde deformylating oxygenase (ADO) reaction and its relationship to the broader ferritin-like superfamily. (A) ADO catalyzes the conversion of fatty aldehydes to the corresponding n-1 alkanes. (B) Despite being substantially divergent in amino acid sequence, dimetal, carboxylate-bridged proteins within the ferritin superfamily possess a highly conserved metal coordination geometry. The different metal colors represent iron (orange) and manganese (purple) found in the crystal structures. Protein Data Bank codes of each protein are listed at the bottom of each structure with their pairwise sequence identity shown between them. Protein 43 is a new crystal structure from *Synechococcus* sp. determined in this work. (C) All-to-all pairwise amino acid sequence identity of the 114 ferritin-like genes evaluated in this work. The average identity of this protein set covers a highly diverse sequence space with an average pairwise amino acid sequence identity of 12.6%, compared to that of 60% in the native cyanobacterial family.

Archaea. Combining structural biology, molecular modeling, and machine learning, we have identified the minimal requirements for ADO-like activity to occur. Our discovery suggests that ADO-like activity can be achieved by non-cyanobacterial proteins very distal in sequence space and may be buried in many other proteins within the ferritin-like superfamily. This finding also opens doors to an entirely new suite of diverse proteins for further biophysical studies of di-iron protein chemistry that will serve as the critical stepping stone to the future successful use of these enzymes within the bioeconomy.[30]

RESULTS

Genomic Mining and Discovery of New ADO-like Activity. The ferritin-like superfamily is one of the largest and oldest protein families.[31] Our search started with the well-studied ADO sequence from *Prochlorococcus marinus* (PmADO) [Protein Data Bank (PDB) entry 4KVQ] as a starting point. Jackhmmer[32] was used to find orthologues, and roughly 40000 unique sequences were identified. A series of filters (detailed in Genes selection in Methods in the Supporting Information) maximizing the sequence space coverage were used to identify a subset of 114 sequences with an average pairwise sequence identity of 12.6% (Figure 1C, Data S1 and S2, and Table S1). Their synthetic genes were synthesized and expressed in *Escherichia coli*, which resulted in 75 of them being solubly expressed at levels where they could be readily purified and functionally characterized (Figure 2, Figure S1, and Table S2). Because a di-iron center is critical for activity in established ADOs, the metal contents of each enzyme (Fe, Mn, Co, Ni, and Zn) were quantitated using inductively coupled plasma mass spectrometry (ICP-MS) (Figure 2 and Table S2). All expressed proteins were screened for *in vitro* alkane production using octanal (C8) and octadecanal (C18) as representative substrates of long and medium chain aldehydes. A total of 32 of 75 expressed proteins were found to be active on at least one of the tested substrates (Figure 2). Among these active proteins, 14 of them have not been previously reported to have ADO-like activity and belong to families that are distal (<20% identical in amino acid sequence) to the characterized ADO family in cyanobacteria that has an average sequence identity of 60% (Figure 1C). These proteins include members of the tRNA-(MS[2]IO[6]-A)-hydroxylase (MiaE) family, the ribonucleotide reductase (RNR) family, and multiple families with unknown function (Figure 2). This screening shows that their abilities to produce alkanes are on average 1 order of magnitude weaker than those from the cyanobacterial family and are likely present in the form of latent activities (Table S1). Among the 43 expressed proteins with no activity, ICP-MS analysis showed that 11 of these proteins had only <0.1 iron-protein equivalent in the sample. Further analysis of the data shows that no significant correlation was observed between the level of iron and the level of alkane in the reaction (Figure S2). This indicates that some critical features beyond a di-iron center that are needed for this ADO-like function to occur do exist.

Protein Structures. Prior to this work, no structures of noncyanobacterial proteins with ADO-like activity had been determined. Therefore, to help elucidate the structural determinants of ADO function, we obtained seven crystal structures of six proteins [proteins 4, 12, 19, 43, 60, and 84 (Figure 2); crystallographic procedures can be found in the Methods in the Supporting Information]. These proteins were chosen to diversify the structures available that can perform the alkane-producing function both within and outside the established ADO family. All structures were of high quality as reflected in the refinement statistics (Table S3), and the general four-helix bundle architecture was conserved throughout. On the contrary, significant diversity was observed in the active site, which could have potential effects on catalytic activity. Despite the nearly identical purifications used throughout this work, the iron occupancy estimated from the

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

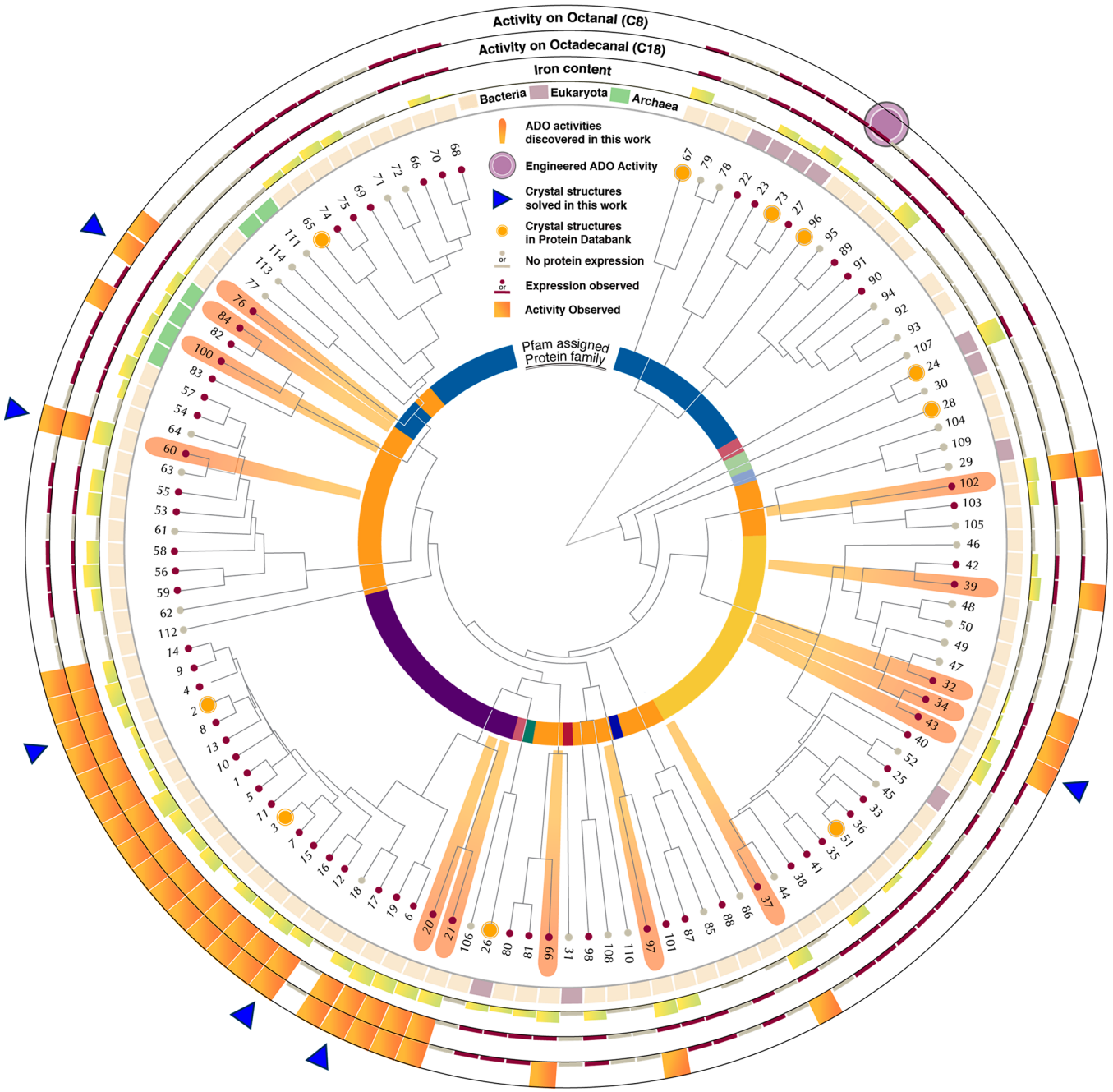


Figure 2. Functional characterization of alkane-producing activity. A phylogenetic tree based on the sequence alignment for the 114 proteins was constructed using the Geneious program.[33] The PFAM annotation for each sequence is indicated through the color of the inner circle: light green for fatty acid desaturase, light blue for phenol hydroxylase, yellow for MiaE, royal blue for *P*-aminobenzoate *N*-oxygenase, blue for ribonucleotide reductase, purple for cyanobaterical aldehyde deformylating oxygenase, rose for ferritin, dark green for rubrerythrin, red for the PF04305 family, and orange for multiple other families with no pfam assigned. The domain of life [Bacteria (pink), Arachaea (light green), or Eukaryota (violet)] of each protein sequence is indicated in the first outer circle. Proteins that were found to be expressed in the soluble fraction [74 of 114 (gray dots under the protein numbering ring)] were screened for *in vitro* alkane production and metal content quantification. Each purified protein solution was subjected to ICP-MS to quantify the metal content of the solution; a full bar on the iron chart denotes 2 equiv of iron atoms per protein molecule. The same solution was evaluated for the production of alkanes upon incubation with either octanal or octadecanal as a substrate under standardized reaction conditions, the product of which was screened by GC-FID and validated via GC-MS (Table S2 and Data S3; assay and detection procedures can be found in the Methods in the Supporting Information). Enzymes with detectable levels of alkane observed are labeled with an orange box on the two outer rings labeled octanal and octadecanal. The 14 newly discovered genes with alkane observed are highlighted by rose spokes in the phylogenetic tree. All proteins with a previously known crystal structure are identified with orange circles on the node of the tree, and proteins whose structures were determined in this work are indicated with blue triangles on the outer rim of the tree. The purple circle indicates the protein where ADO activity is introduced. A full table of all of the data illustrated in this figure and detailed experimental methods are available in the Methods in the Supporting Information and Table S2.

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

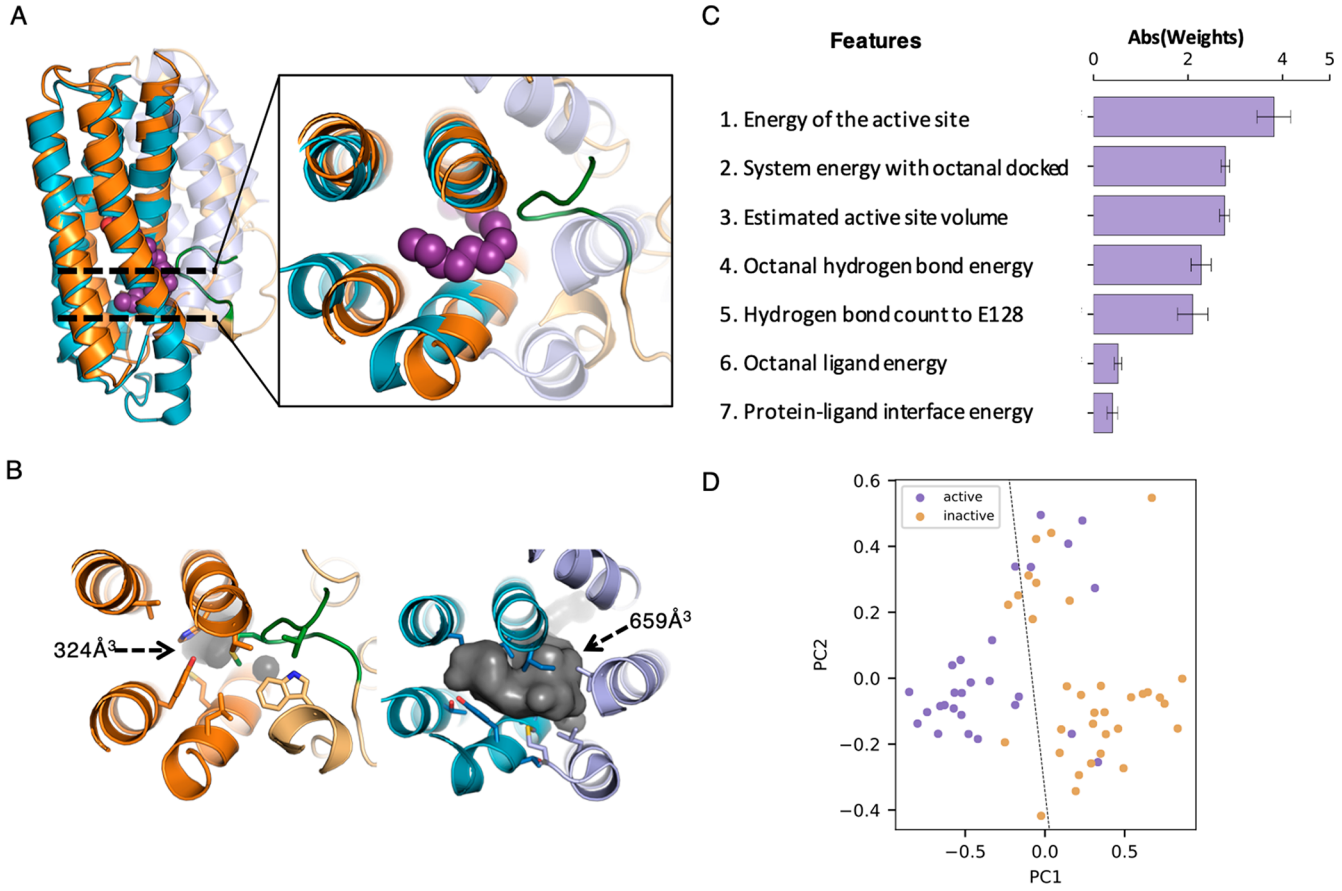


Figure 3. Elucidation of function-conferring structural features of ADO-like proteins. (A) Overlay of proteins 43 (orange) and 3 (blue, the previously established ADO from *P. marinus*) showing that the core four-helix bundles (helix highlighted) of these proteins are structurally conserved despite sharing a low pairwise sequence identity of 10%. The active site view of protein 43 reveals a structured loop (green) spanning residues 212−217, which obstructs part of the pocket and is predicted to clash with the fatty aldehyde binding pocket observed in protein 3 (purple sphere). (B) Active site pocket volume (gray) of protein 43 that is roughly 2-fold smaller than that in protein 3 as calculated by Caver.[34] (C) Weight (shown in absolute value) for each feature generated by Lasso models (the error bar represents one standard deviation). All features were generated on the basis of our current mechanistic understanding and interactions observed in the well-studied *P. marinus* ADO. Details of the selection of features can be found in the Methods in the Supporting Information and Discussion S1). (D) The 60 data points were visualized in the space of the first principal component (PC). The proteins fall into two distinguishable clusters with the first PC being the major explanatory variable. The first PC is a linear combination of the raw features, among which the feature named overall system energy with octanal docked contributed most to the first PC. The dashed line represents a separating line generated by a support vector machine classifier with the circled points being the support vectors. The blue circles represent active proteins, whereas the red circles represent inactive proteins.

structures varied from zero to full, suggesting that iron depletion may contribute to the low catalytic efficiencies observed in this class of enzymes. The reasons for this are not obvious in the cases of NpADO/protein 4 and GvADO/protein 12; however, the StADO/protein 84 sequence is missing one of the chelating glutamate residues, and two independent structures showed that there are either 0 (monoclinic form) or 1 equiv (orthorhombic form) of iron per protein as a result of the disrupted iron binding site. Examination of active sites also revealed a variety of ligands bound to the presumed substrate binding site, including proprionic acid soaked into SsADO2/protein 19 and three dodecyldimethylamine oxide (DDAO) molecules (NpADO/protein 4), which was essential for crystallization in the latter case. The GvADO/protein 12 structure revealed an endogenously bound stearate in an "L"-shaped conformation that was similar to the stearate found in the ADO from *Synechococcus elongatus* (PDB entry 4RC8). On the basis of these observations, it is possible that fatty acids and other lipids may inhibit a substantial fraction of ADO activity *in vivo*.

A notable relationship between active site volume and function was gained from the crystal structures determined in this work. SsADO/protein 43, for which the structure was determined in this work, is annotated to be a tRNA-hydroxylase-like protein from *Synechococcus* sp. (strain CC9605) and is roughly 10% identical in sequence to pmADO, the first reported ADO with a structure [protein 3 (Figure 2)]. *In vitro* alkane screening revealed that protein 43 is active only on the octanal substrate under the tested conditions. On the basis of a structural alignment of the distinctive four-helix bundle domains between proteins 43 and 3, it reveals that despite having a low degree of sequence homology, there is significant secondary and tertiary homology with a root-mean-square deviation (RMSD) of 2.06 Å over the entire metal-coordinating four-helix bundle region (Figure 3A). The active site of this putative hydroxylase, as suggested by the pocket found next to the di-iron center, has a volume of 324 Å^3, which is 2-fold smaller than that found in protein 3 (659 Å^3)[34] (Figure 3B). This decrease in the volume of protein 43 is in part ascribed to a loop formed by residues 212−217

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

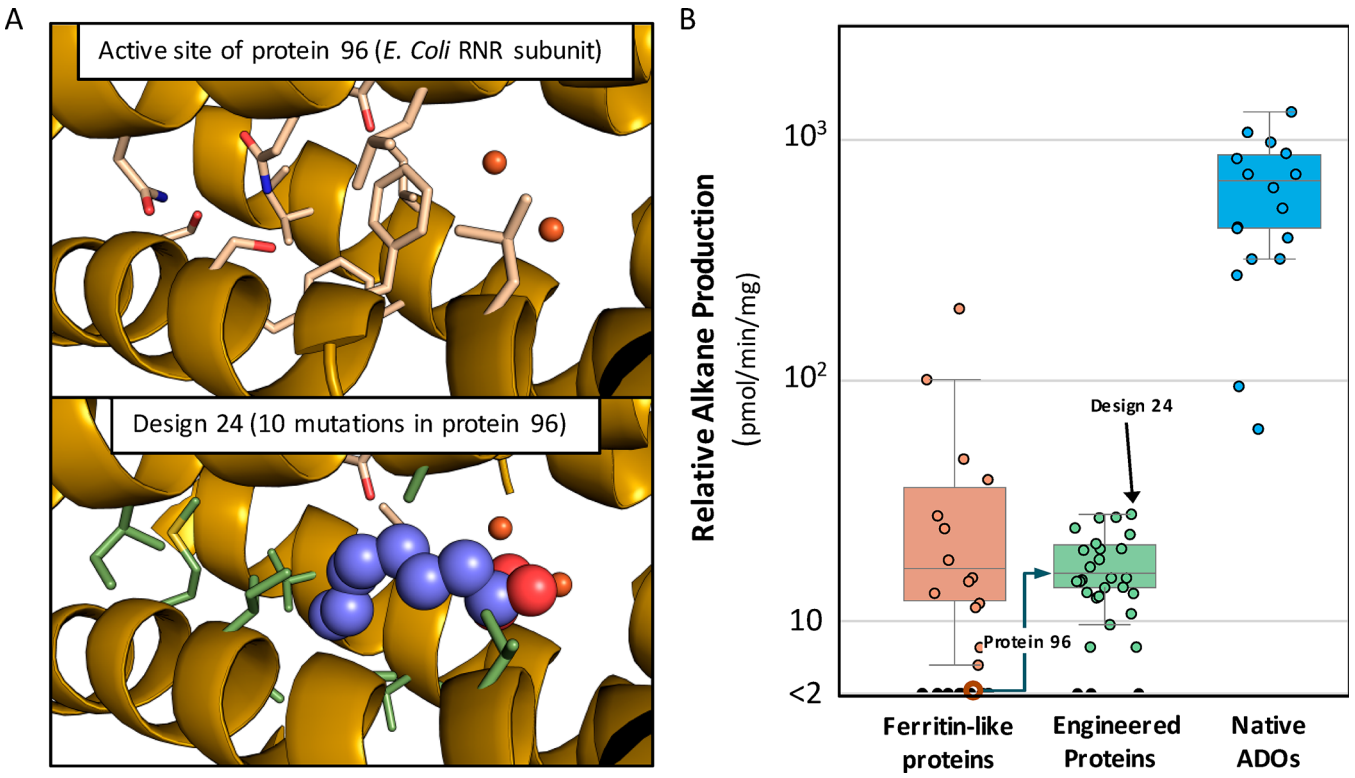


Figure 4. Engineering ADO-like activity in the small subunit of *E. coli* RNR. (A) The top panel illustrates the active site of protein 96 (the *E. coli* RNR small subunit), which has no measurable open volume present next to the iron center (rust-red spheres). Computational protein design tools were used to redesign protein 96 to introduce ADO-like. The bottom panel illustrates a model of the most active mutant where the 10 mutated amino acids are highlighted as green sticks in the active site. The octanal substrate in the model is depicted as spheres with purple carbons, and the di-iron in small rust-red spheres. (B) Box plot representation of the levels of alkane observed in ferritin-like, engineered, and native ADO proteins. Design 24 (10 mutations, N76I, Q80A, L83A, L203V, F208A, S211I, F212A, S215M, I231V, and A235V) represents the most active mutant obtained in this work.

inserting into the active site (Figure 3A,B). These structural features show that the active site of protein 43 is not large enough to accommodate the C18 substrate and thus exhibits no activity. Further examination of other structures determined in this work shows that this volume−specificity relationship is commonly observed. The structure of protein 84 (an uncharacterized archaeal protein) reveals that it has a cavity next to the di-iron center with the size of 359 Å^3, and it is inactive on octadecanal under the tested conditions. Protein 60, annotated as a VlmB homologue and active on both octanal and octadecanal, has a cavity size of 1200 Å^3 based on our structure. The structures of two other cyanobacterial proteins (proteins 12 and 19) also concur with this notion, having volumes of 993 and 747 Å^3, respectively. These data validate prior observation that the function of these proteins can be modulated by controlling active site volumes.[22] Of course, the size of the active site is not the only essential feature for the existence of ADO-like activity. Eight other noncyanobacterial ferritin-like proteins with structures have been tested in this work, and none of them are active under the tested conditions. Their active site volume ranges from no detectable volume to 864 Å^3 (Figure S3). Therefore, it is evident that other structural determinants are at play for governing the chemistry of this protein function.

Construction of a Predictor for ADO-like Function. To efficiently navigate through the complex sequence and structural landscape of ADO-like activity, we examined multiple structural features such as di-iron coordination geometry, hydrophobic interactions with the alkyl chain, and the shape complementarity between the protein pocket and the substrate. To evaluate structural features for as many proteins as possible in our data set, RosettaCM[35] was used to generate molecular models for the proteins with no crystal structure available. Using the models, an octanal ligand was docked into the active site of each expressed protein using RosettaLigand[36] with functional constraints to ensure the protein and ligand were modeled in a catalytically relevant orientation (Data S4). Specifically, the ligand was modeled after the $Fe_2^{III/III}$-peroxychemiacetal species as identified by previous mechanistic studies of ADO.[24] Structural features were then calculated on the basis of force field energies and other metrics that describe chemical interactions (Figure 3C,D, Data S5 and S6, and Discussion S1). To rank the importance of these features, a machine learning model was employed to generate a model capable of discriminating structural factors essential to activity based on our data.

L1-regularized logistic regression models[37] were built to evaluate the relative importance of features in discriminating ADO-like protein activity. In the task of activity classification to low and high, generalized linear models had an area under the ROC curve (AUCROC) of 0.85 ± 0.02 and an AUC for the Precision-Recall curve (AUCPR) of 0.67 ± 0.1 (Figure S4C−F). The relative importance of each feature indicated by the weight generated by LASSO was close to the result by principal component analysis (PCA) with a difference in two ranks of 3 (Figure 3C,D and Table S4). The two rank lists are

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

Table 1. Kinetic Characterization of Selected Enzymes[a]

protein	organism	k_{cat} (s^{-1})	K_M (M)	k_{cat}/K_M (M^{-1} s^{-1})	putative function	PID to Np ADO
4	*N. punctiforme*	$6.5 \times 10^{-3} \pm 2.8 \times 10^{-4}$	$4.1 \times 10^{-5} \pm 5.7 \times 10^{-6}$	160 ± 23	aldehyde deformylating oxygenase	100
19	*Synechococcus* sp. RS9917	$1.8 \times 10^{-3} \pm 1.2 \times 10^{-4}$	$5.4 \times 10^{-4} \pm 9.7 \times 10^{-5}$	3.3 ± 0.65	aldehyde deformylating oxygenase	37.1
43	*Synechococcus* sp. CC9605	$2.2 \times 10^{-3} \pm 9.2 \times 10^{-5}$	$1.3 \times 10^{-3} \pm 1.3 \times 10^{-4}$	1.7 ± 0.18	tRNA-(MS(2)IO(6)A)-hydroxylase-like	6.6
84	*Sulfolobus tokodaii*	not determined	$>1.5 \times 10^{-3}$	0.50 ± 0.014	uncharacterized protein	10.4
96*	engineered protein 96	not determined	$>1.5 \times 10^{-3}$	1.5 ± 0.15	computationally designed ADO from protein 96	18.1
102	*K. albida*	$4.2 \times 10^{-3} \pm 1.1 \times 10^{-4}$	$3.2 \times 10^{-4} \pm 2.7 \times 10^{-5}$	14 ± 1.2	uncharacterized protein	8.8

[a]The protein number corresponds to the number shown in Figure 2. The last column, percent sequence identity (PID) to NpADO, is the pairwise percent amino acid sequence identity of the protein to Np ADO. Values and the reported standard error are calculated using nonlinear or linear regression, as appropriate, on the basis of the mean activity observed across at least three independent measurements at five to seven different substrate concentrations of octanal. Saturation was not observed for protein 84 or 96* (design 24 of protein 96), and therefore, the turnover rate could not be determined and only a limit for K_M is reported in addition to the measured catalytic efficiency. More details are provided in the Methods in the Supporting Information, and fitted curves are shown in Figure S5.

correlated by a p value of 0.65 (τ-test). The results from this analysis suggest that (1) the Rosetta energy of the active site, (2) the total system energy, and (3) the overall active site volume are predictive of ADO-like activity. The successful establishment of correlation between function and selected structural features not only corroborates the apparent requirements of a di-iron center with appropriate active site volume and shape but also evokes a bold hypothesis that incorporating these basic features would allow a noncyanobacterial protein to produce alkane using fatty aldehyde.

Introduction of ADO Activity into Ribonucleotide Reductase (RNR). To conduct a rigorous evaluation of the ability of the identified structural features to confer function, we aimed to introduce ADO-like catalytic activity into the class I RNR family, for which no proteins with ADO activity had been previously observed. On the basis of the availability of structural data, di-iron metal binding data, and protein expression level with our system, we selected the RNR small subunit from *E. coli* [PDB entry 1AV8 (protein 96 in Figure 2)] as the specific chassis for design.

The crystal structure of this protein reveals that of the features identified, protein 96 has insufficient steric volume to accommodate a substrate adjacent to the di-iron center; however, proper coordination and hydrogen bonding of the di-iron center were present (Figure 4A). Therefore, a subset of features that target this specific deficiency, including the active site volume and pocket−substrate shape complementarity, were chosen to aid in our design endeavor for activity introduction (Figure 3C). Computational protein design was used to create a cavity with a volume on par with that of octanal-specific protein 43. Amino acid changes predicted to optimize packing around the targeted octanal substrate while maintaining the structural integrity of the proteins were introduced. Because most of the native amino acids in the active site of protein 96 are either polar or bulky hydrophobic residues, the entire binding pocket has been completely redesigned with small hydrophobic residues to improve its shape complementarity toward octanal. A total of 25 proteins designed over three rounds were screened for ADO-like activity on the octanal substrate (Table S5). Of these proteins, 16 were solubly expressed and 13 had activity on the octanal substrate (Table S5). The most active design (design 24) contains a total of 10 mutations, including N76I, Q80A, L83A, L203V, F208A, S211I, F212A, S215M, I231V, and A235V (Figure 4A). The active site volume as measured from the mutant model is estimated to be 415 $Å^3$, which is 20% larger than that of protein 43. These results demonstrate that the molecular features confirmed as being critical by machine learning adequately provide the minimal requirements for this ADO-like activity to occur.

Evaluating the Utility of the Machine Learning Model as a Postdesign Classifier. With the ability to introduce ADO-like activity into protein 96, the opportunity to evaluate the machine learning model as a postdesign classifier using a variety of established design methodologies could be conducted. Therefore, a new set of 21 mutant proteins were constructed using the protein design packages PROSS[38] and Funclib[39] (using default settings), PROSS guided by an alignment generated from ADOs from cyanobacteria, Foldit to revert specific mutations back to native from the previously identified best design from protein 96, and Foldit to design mutations to disrupt hydrogen bonds around the di-iron center (Table S6). This test set aims to challenge the model's ability to classify mutants that were either designed to impart ADO-like activity (PROSS and Funclib) or mutated to remove function (Foldit point mutations). Of the 21 proteins built, 19 were solubly expressed and characterized using our standard protocol. To maximize our chance of identifying all true positives using our classifier, the average prediction cutoff that identified all true positive genes in the training set was used to predict active genes in the set of the 19 tested proteins. Using this criterion, 14 of the 19 expressed genes in the test set are predicted to be active, and 11 of them are true positives as confirmed by experimental characterization. This translates to a prediction accuracy of 78.5%. The model can be found in the Methods in the Supporting Information under the statistical test and machine learning protocol. These results demonstrate that the machine learning classification model could effectively serve as a postdesign classifier to help select active mutants for experimental characterizations.

Kinetic Characterization of Selected Enzymes on Octanal. To further understand the structure−function relationships of these proteins with latent ADO-like activity, kinetic characterization was carried out on a subset of six enzymes against octanal. This panel of proteins includes the most active design from protein 96 (design 24), native ADO

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

from *Nostoc punctiforme* (NpADO), multiple other proteins whose crystal structures have been determined in this work [proteins 4, 19, 43, and 84 (Table 1)], and a noncyanobacterial protein with the highest level of alkane production observed from the screening (protein 102). Results show that the catalytic efficiency of NpADO was 9.55 mM^{-1} min^{-1}, the highest among those of the six characterized proteins. Its turnover number and Michaelis constant of 0.39 min^{-1} and 0.041 mM, respectively, are within 2-fold of those of the ADO from *P. marinus* (*Pm*) measured previously[40] (Table 1). Another cyanobacterial ADO from *Synechococcus* sp. RS9917 (protein 19), whose crystal structure has been determined in this work, has a k_{cat}/K_M of 0.20 mM^{-1} min^{-1}, which is 48-fold lower than that of NpADO. Structural analysis of this protein overlaid with *Pm* ADO shows that their active sites are highly conserved. Using all residues within 5 Å of the dodecyldimethylamine oxide bound in NpADO for comparison, the only difference between these two proteins is residue V108 in protein 19, which is an isoleucine on NpADO. The engineered protein from protein 96 of design 24 has a k_{cat}/K_M of 0.088 mM^{-1} min^{-1}, which is 44% of that observed with protein 19. Further mechanistic studies will be necessary to understand the remaining features required to improve the catalytic efficiency of protein 19 and the engineered protein to the level of NpADO, and potentially beyond. Another notable observation is the catalytic efficiency of protein 102 from *Kutzneria albida*. This uncharacterized protein with no Pfam assigned shares a pairwise sequence identity of only 8.8% with NpADO. Kinetic characterization reveals that it has a k_{cat}/K_M of 0.80 mM^{-1} min^{-1}, which is 4-fold higher than that of protein 19, and a k_{cat} value of 0.25 min^{-1}, which is 64% of NpADO. This result shows that the level of ADO activity possessed by the cyanobacterial protein can be achieved by proteins vastly different in sequence space.

DISCUSSION

In this work, a combination of genomic mining, structural biology, molecular modeling, and machine learning algorithms were used to study the structure−function relationship of the ferritin-like superfamily to ADO-like activity. Our genomic mining efforts have also revealed that ADO-like activity is widely observed outside of the narrow sequence range of the cyanobacterial scaffolds that previously defined this family of enzymes.[41] The successful introduction of this alkane-producing enzymatic activity into a natively inactive RNR protein demonstrates that the minimal requirement for the alkane-producing activity to occur is a properly coordinated di-iron center next to a hydrophobic pocket.

The crystal structures determined in this work have also revealed significant insights into the structure−function relationship of this family of proteins. While our kinetic results show that the activities observed in the newly identified enzymes are not simply a function of iron loading, the varying equivalents of iron detected in the structures of the expressed proteins may explain the turnover rate discrepancies described in other studies.[20,22,40] In addition, the diverse types of lipids bound to the di-iron center as revealed from our structures (Table S3) suggest that different small molecules (particularly endogenously bound fatty acids) may inhibit the activities of this enzyme family. This further complicates comparisons of catalytic efficiencies between studies if growth and purification methods are not consistent. Future studies that will examine the ability of various reagents and metabolites to inhibit this enzyme chemistry could provide valuable insights for engineering efforts focused on producing enzymes that are active in a cellular context. Another notable finding is the discovery of alkane forming activity from a scaffold (protein 84) that misses one of the Fe-coordinating glutamate residues. The ability to produce alkanes from aldehyde with an active site that is distinct from that of cyanobacterial ADOs suggests that this protein could be performing this reaction via a different mechanism. Future biophysical and spectroscopic studies will help shed light on the exact mechanism that these proteins employ to catalyze their alkane-forming activities.

The successful introduction of ADO-like activity into the small subunit of protein 96 demonstrates the functional modularity of the ferritin-like protein scaffolds as well as the minimal requirements for this ADO-like activity to occur. While our data indicate that a properly coordinated di-iron center with a hydrophobic pocket may be enough to support this enzyme activity, the low catalytic efficiencies from our designed synthetic proteins indicate that other important structural factors are necessary for this enzyme activity to approach the level observed in cyanobacterial ADOs.[40] Previous works have proposed that other elements that may be critical to ADO activity include a well-positioned hydrogen bond to anchor the aldehyde oxygen for tighter substrate binding and a channel of water molecules leading to the di-iron center from the protein surface to facilitate proton transfer.[26] These features have not been explicitly designed in protein 96 in this work as the backbone geometry was not predicted to be compatible with the desired hydrogen bond. Therefore, engineering efforts to redesign the structure of protein 96 or mine for enzymes with compatible backbones for the introduction of this could be an effective way to discover ADO-like enzymes with higher levels of activity.

Despite the limited training data sets, the machine learning methods that we used were adequate to capture information from relevant features of ADO-like activity and create a classifier with a high level of accuracy. There is much room for improvement, something that is also evident by strong classification performance, but poor regression performance between the training and testing sets (Figure S4). Increasing both the size and the diversity of the data set will be a boon to future analyses, as they will increase the generalization accuracy and robustness of the machine learning predictors, and it will allow the application of more sophisticated methods that would otherwise lead to overfitting with the current data set. However, we expect the current classifier developed here to be of significant value in future mining and design efforts for enzymes with ADO-like activity.

In conclusion, these latent ADO-like activities open numerous protein targets for biophysical studies on di-iron protein chemistry, which are invaluable to the search of relevant enzymes for future synthetic biology applications.[42−44] More broadly, the integration of genomic mining, structural biology, machine learning, and computational protein design as presented here holds significant promise for the discovery and design of structural features responsible for enzyme function.

METHODS

All methods are provided in the Supporting Information. No unexpected or unusually severe safety hazards were encountered.

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00665.

Sodium dodecyl sulfate−polyacrylamide gel electrophoresis analysis, protein amino acid and DNA sequences, data used to generate Figure 2 (Table S2), Michaelis−Menten plots, GC-MS traces, and Rosetta modeling files (PDF)

List of features for Lasso (XLS)

List of features for PCA (XLSX)

Crystallographic data (ZIP)

Accession Codes

Certain sequences have become obsolete (highlighted with asterisks below) over the course of this study, and therefore, a UniProt ID is no longer available for some of the sequences in this study. The UniProt IDs of the closest homologues identified using Hmmer are used as representative accession IDs for these sequences. See the Fasta files provided in the Supporting Information for the exact sequences used. All crystal structures have been deposited in the Protein Data Bank under the accession codes listed in Table S3: protein 1, B8HSZ3_CYAP4; protein 2, ALDEC_SYNE7; protein 3, ALDEC_PROMM; protein 4, ALDEC_NOSP7; protein 5, K9RTZ6_SYNP3; protein 6, K9P6M5_CYAGP; protein 7, A3YYU8_9SYNE; protein 8, B4WJ48_9SYNE; protein 9, K9SZC3_9SYNE; protein 10, L8LLH3_9CHRO; protein 11, A0A166UJA4_9CYAN; protein 12, Q7NGM3_GLOVI; protein 13, K9Z5G1_CYAAP; protein 14, ALDEC_SYNY3; protein 15, A5GUD0_SYNR3; protein 16, Q2JX62_SYNJA; protein 17, B8HLW2_CYAP4; protein 18, K9EQR0_9CYAN; protein 19, A3Z6M0_9SYNE; protein 20, A0A406MN71_9ZZZZ; protein 21, A0A0F5ZY09_9GAMM; protein 22, A0A0D9QH06_PLAFR; protein 23, F9W6C1_TRYCI; protein 24, A0A061EJ81_THECC*; protein 25, A0A0A2WK64_BEABA; protein 26, YCX8_CYAPA; protein 27, H3FJK2_PRIPA; protein 28, Q84AQ2_PSEST; protein 29, A0A0A2 V841_BEABA; protein 30, A0A061EJ81_THECC; protein 31, A0BNK4_PARTE; protein 32, A5GR51_SYNR3; protein 33, M4XFB3_PSEDE; protein 34, Q7V8X6_PROMM; protein 35, C7R6F3_KANKD; protein 36, F6ABL4_PSEF1; protein 37, A6VX42_MARMS; protein 38, Q2SK44_HAHCH; protein 39, Q7TTZ9_RHOBA; protein 40, D5VCV9_MORCB; protein 41, E1VK89_9GAMM; protein 42, D2R2X6_PIRSD; protein 43, Q3AMA5_SYNSC; protein 44, B3PHJ7_CELJU; protein 45, Q9I2V2_PSEAE; protein 46, Q2IMP0_ANADE; protein 47, K9P9K3_CYAGP; protein 48, Q7VDJ1_PROMA; protein 49, A0A356NDR7_9SYNE*; protein 50, A0A0A2ALS1_PROMR*; protein 51, Q88KV1_PSEPK; protein 52, MIAE_SALTY; protein 53, A0A542XQM6_9ACTN*; protein 54, A0A1S1QEC2_9ACTN*; protein 55, D9W5G4_9ACTN; protein 56, A0A022M8K7_9ACTN; protein 57, A0A0M8SS09_9ACTN*; protein 58, H2K8B9_STRHJ; protein 59, C6WBU0_ACTMD; protein 60, E4N6B3_KITSK; protein 61, A8L8J1_FRASN; protein 62, Q7MZA5_PHOLL; protein 63, Q84F34_STRVF*; protein 64, R4L8R2_9ACTN; protein 65, RIR2H_MYCTA; protein 66, E6SGH3_THEM7; protein 67, A0A0E9CGZ9_CHLTH; protein 68, C8WQL9_ALIAD; protein 69, C0Z9B9_BREBN; protein 70, D5WWU8_KYRT2; protein 71, I0I9S1_CALAS; protein 72, A9WJQ1_CHLAA; protein 73, H2QWI9_PANTR; protein 74, D2PRL0_KRIFD; protein 75, RIR2H_SACEN; protein 76, Q5QL41_GEOKA; protein 77, D3FEI3_CONWI; protein 78, Q6ADM4_LEIXX*; protein 79, A0A0R2PVF5_9MICO*; protein 80, T0XWY9_9BACT; protein 81, I0IP74_LEPFC; protein 82, Q970H5_SULTO*; protein 83, F4G2E9_METCR; protein 84, Q970H5_SULTO; protein 85, D2PLC0_KRIFD*; protein 86, A0A0B8ZZA5_BRELN*; protein 87, A0A0T0MIS6_9CELL*; protein 88, V7KWS3_MYCPC; protein 89, E7FXJ4_ERYRH; protein 90, C7NA19_LEPBD; protein 91, J5BG39_ENTFL; protein 92, C6D5U5_PAESJ; protein 93, F6FFK4_MYCHI; protein 94, BNRDF_BPSPB; protein 95, Q6YRG0_ONYPE; protein 96, A0A037YRC8_ECOLX; protein 97, C5CMN9_VARPS; protein 98, A9F8U5_SORC5; protein 99, U4QRD8_9BACT; protein 100, H2C789_9CREN; protein 101, C7PXC6_CATAD; protein 102, W5WGM4_9PSEU; protein 103, E3J563_FRASU; protein 104, A0A1C9VGC7_9BURK; protein 105, Q2J5C1_FRASC; protein 106, S9QQR0_9DELT; protein 107, X1FPJ6_9ZZZZ; protein 108, E3HSR5_ACHXA; protein 109, E6UYQ7_VARPE; protein 110, H8MF88_CORCM; protein 111, A0A1U7EYI2_NATPD; protein 112, D0LYY3_HALO1; protein 113, D3FED6_CONWI; protein 114, D3ST95_NATMM.

■ AUTHOR INFORMATION

Corresponding Author

Justin B. Siegel − *Department of Chemistry, Department of Biochemistry and Molecular Medicine, and Genome Center, University of California, Davis, Davis, California 95616, United States*; Email: jbsiegel@ucdavis.edu

Authors

Wai Shun Mak − *Department of Chemistry, University of California, Davis, Davis, California 95616, United States*; orcid.org/0000-0003-2996-0139

XiaoKang Wang − *Department of Biomedical Engineering, University of California, Davis, Davis, California 95616, United States*

Rigoberto Arenas − *Department of Chemistry and Chemistry Graduate Group, University of California, Davis, Davis, California 95616, United States*; orcid.org/0000-0002-6151-7535

Youtian Cui − *Department of Chemistry, University of California, Davis, Davis, California 95616, United States*

Steve Bertolani − *Department of Chemistry, University of California, Davis, Davis, California 95616, United States*

Wen Qiao Deng − *California College of Arts, San Francisco, California 94107, United States*

Ilias Tagkopoulos − *Department of Biomedical Engineering, Genome Center, and Department of Computer Science, University of California, Davis, Davis, California 95616, United States*

David K. Wilson − *Chemistry Graduate Group and Department of Molecular and Cellular Biology, University of California, Davis, Davis, California 95616, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00665

Author Contributions

J.B.S. conceived the genomic mining strategy. J.B.S., W.S.M., and S.B. designed and performed the *in vitro* and modeling experiments. X.W. and I.T. performed the machine learning experiments. R.A. and D.K.W. performed protein crystallog-

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

raphy experiments. W.Q.D. provided support in figures and data visualization. All authors contributed to the writing of the manuscript and discussions of data, results, and implications of the work.

Funding

Research reported in this publication was supported by the Joint Genome Institute Community Synthesis Project Grant 1109, the work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DEAC02- 05CH11231, the University of California, Davis, National Science Foundation Grants 1827246, 1805510, and 1627539, the National Institute of Environmental Health Sciences of the National Institutes of Health (NIH) under Grant P42ES004699, Sloan Foundation Grant BR2014-012, NIH Grant R01 GM 076324-11, and the Rosetta Commons.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, under Contract DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research and the National Institute of General Medical Sciences, National Institutes of Health (including Grant P41GM103393).

■ REFERENCES

(1) Chen, K., and Arnold, F. H. (2020) Engineering new catalytic activities in enzymes. *Nat. Catal. 3*, 203−213.
(2) Wolf, C., Siegel, J. B., Tinberg, C., Camarca, A., Gianfrani, C., Paski, S., Guan, R., Montelione, G., Baker, D., and Pultz, I. S. (2015) Engineering of Kuma030: a gliadin peptidase that rapidly degrades immunogenic gliadin peptides in gastric conditions. *J. Am. Chem. Soc. 137*, 13106−13113.
(3) Chowdhury, R., and Maranas, C. D. (2020) From directed evolution to computational enzyme engineering—A review. *AIChE J. 66*, No. e16847.
(4) Lutz, S., and Iamurri, S. M. (2018) Protein engineering: past, present, and future. In *Protein Engineering*, pp 1−12, Springer.
(5) Markel, U., Essani, K. D., Besirlioglu, V., Schiffels, J., Streit, W. R., and Schwaneberg, U. (2020) Advances in ultrahigh-throughput screening for directed enzyme evolution. *Chem. Soc. Rev. 49*, 233−262.
(6) Goldsmith, M., and Tawfik, D. S. (2017) Enzyme engineering: reaching the maximal catalytic efficiency peak. *Curr. Opin. Struct. Biol. 47*, 140−150.
(7) Faber, M. S., and Whitehead, T. A. (2019) Data-driven engineering of protein therapeutics. *Curr. Opin. Biotechnol. 60*, 104−110.
(8) Trudeau, D. L., and Tawfik, D. S. (2019) Protein engineers turned evolutionists—the quest for the optimal starting point. *Curr. Opin. Biotechnol. 60*, 46−52.
(9) Zeymer, C., and Hilvert, D. (2018) Directed evolution of protein catalysts. *Annu. Rev. Biochem. 87*, 131−157.
(10) Hou, J., Wu, T., Cao, R., and Cheng, J. (2019) Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins: Struct., Funct., Genet. 87*, 1165−1178.
(11) Yang, K. K., Wu, Z., and Arnold, F. H. (2019) Machine-learning-guided directed evolution for protein engineering. *Nat. Methods 16*, 687−694.
(12) Wu, Z., Kan, S. J., Lewis, R. D., Wittmann, B. J., and Arnold, F. H. (2019) Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A. 116*, 8852−8858.
(13) Hon, J., Borko, S., Stourac, J., Prokop, Z., Zendulka, J., Bednar, D., Martinek, T., and Damborsky, J. (2020) EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res. 48*, W104−W109.
(14) Wrenbeck, E. E., Faber, M. S., and Whitehead, T. A. (2017) Deep sequencing methods for protein engineering and design. *Curr. Opin. Struct. Biol. 45*, 36−44.
(15) Krebs, C., Bollinger, J. M., and Booker, S. J. (2011) Cyanobacterial alkane biosynthesis further expands the catalytic repertoire of the ferritin-like 'di-iron-carboxylate'proteins. *Curr. Opin. Chem. Biol. 15*, 291−303.
(16) Bornscheuer, U. T., and Kazlauskas, R. J. (2004) Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angew. Chem., Int. Ed. 43*, 6032−6040.
(17) Schirmer, A., Rude, M. A., Li, X., Popova, E., and Del Cardayre, S. B. (2010) Microbial biosynthesis of alkanes. *Science 329*, 559−562.
(18) Li, N., Nørgaard, H., Warui, D. M., Booker, S. J., Krebs, C., and Bollinger, J. M., Jr (2011) Conversion of fatty aldehydes to alka (e) nes and formate by a cyanobacterial aldehyde decarbonylase: cryptic redox by an unusual dimetal oxygenase. *J. Am. Chem. Soc. 133*, 6158−6161.
(19) Zhang, J., Lu, X., and Li, J.-J. (2013) Conversion of fatty aldehydes into alk (a/e) nes by in vitro reconstituted cyanobacterial aldehyde-deformylating oxygenase with the cognate electron transfer system. *Biotechnol. Biofuels 6*, 86.
(20) Li, N., Chang, W.-C., Warui, D. M., Booker, S. J., Krebs, C., and Bollinger, J. M., Jr (2012) Evidence for only oxygenative cleavage of aldehydes to alk (a/e) nes and formate by cyanobacterial aldehyde decarbonylases. *Biochemistry 51*, 7908−7916.
(21) Das, D., Ellington, B., Paul, B., and Marsh, E. N. G. (2014) Mechanistic insights from reaction of α-oxiranyl-aldehydes with cyanobacterial aldehyde deformylating oxygenase. *ACS Chem. Biol. 9*, 570−577.
(22) Khara, B., Menon, N., Levy, C., Mansell, D., Das, D., Marsh, E. N. G., Leys, D., and Scrutton, N. S. (2013) Production of Propane and Other Short-Chain Alkanes by Structure-Based Engineering of Ligand Specificity in Aldehyde-Deformylating Oxygenase. *ChemBioChem 14*, 1204−1208.
(23) Waugh, M. W., and Marsh, E. N. G. (2014) Solvent isotope effects on alkane formation by cyanobacterial aldehyde deformylating oxygenase and their mechanistic implications. *Biochemistry 53*, 5537−5543.
(24) Pandelia, M. E., Li, N., Nørgaard, H., Warui, D. M., Rajakovich, L. J., Chang, W.-C., Booker, S. J., Krebs, C., and Bollinger, J. M., Jr (2013) Substrate-triggered addition of dioxygen to the diferrous cofactor of aldehyde-deformylating oxygenase to form a diferric-peroxide intermediate. *J. Am. Chem. Soc. 135*, 15801−15812.
(25) Rajakovich, L. J., Nørgaard, H., Warui, D. M., Chang, W.-C., Li, N., Booker, S. J., Krebs, C., Bollinger, J. M., Jr, and Pandelia, M.-E. (2015) Rapid reduction of the diferric-peroxyhemiacetal intermediate in aldehyde-deformylating oxygenase by a cyanobacterial ferredoxin: Evidence for a free-radical mechanism. *J. Am. Chem. Soc. 137*, 11695−11709.
(26) Buer, B. C., Paul, B., Das, D., Stuckey, J. A., and Marsh, E. N. G. (2014) Insights into substrate and metal binding from the crystal structure of cyanobacterial aldehyde deformylating oxygenase with substrate bound. *ACS Chem. Biol. 9*, 2584−2593.
(27) Shokri, A., and Que, L., Jr (2015) Conversion of Aldehyde to Alkane by a Peroxoiron (III) Complex: A Functional Model for the Cyanobacterial Aldehyde-Deformylating Oxygenase. *J. Am. Chem. Soc. 137*, 7686−7691.
(28) Wang, C., Zhao, C., Hu, L., and Chen, H. (2016) Calculated Mechanism of Cyanobacterial Aldehyde-Deformylating Oxygenase: Asymmetric Aldehyde Activation by a Symmetric Diiron Cofactor. *J. Phys. Chem. Lett. 7*, 4427−4432.

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843

(29) Warui, D. M., Li, N., Nørgaard, H., Krebs, C., Bollinger, J. M., Jr, and Booker, S. J. (2011) Detection of formate, rather than carbon monoxide, as the stoichiometric coproduct in conversion of fatty aldehydes to alkanes by a cyanobacterial aldehyde decarbonylase. *J. Am. Chem. Soc. 133*, 3316−3319.
(30) House, T. W. (2012) National bioeconomy blueprint, April 2012. *Ind. Biotechnol. 8*, 97−102.
(31) Andrews, S. C. (2010) The Ferritin-like superfamily: Evolution of the biological iron storeman from a rubrerythrin-like ancestor. *Biochim. Biophys. Acta, Gen. Subj. 1800*, 691−705.
(32) Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., and Eddy, S. R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res. 43*, No. W30.
(33) Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics 28*, 1647−1649.
(34) Pavelka, A., Sebestova, E., Kozlikova, B., Brezovsky, J., Sochor, J., and Damborsky, J. (2016) CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules. *IEEE/ACM Trans. Comput. Biol. Bioinf. 13*, 505−517.
(35) Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013) High-resolution comparative modeling with RosettaCM. *Structure 21*, 1735−1742.
(36) Lemmon, G., and Meiler, J. (2012) Rosetta Ligand docking with flexible XML protocols. In *Computational Drug Discovery and Design*, pp 143−155, Springer.
(37) Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267−288.
(38) Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., et al. (2016) Automated structure-and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell 63*, 337−346.
(39) Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D. L., Prilusky, J., et al. (2018) Automated design of efficient and functionally diverse enzyme repertoires. *Mol. Cell 72*, 178−186.e5.
(40) Andre, C., Kim, S. W., Yu, X.-H., and Shanklin, J. (2013) Fusing catalase to an alkane-producing enzyme maintains enzymatic activity by converting the inhibitory byproduct H2O2 to the cosubstrate O2. *Proc. Natl. Acad. Sci. U. S. A. 110*, 3191−3196.
(41) Klähn, S., Baumgartner, D., Pfreundt, U., Voigt, K., Schön, V., Steglich, C., and Hess, W. R. (2014) Alkane biosynthesis genes in cyanobacteria and their transcriptional organization. *Front. Bioeng. Biotechnol. 2*, 24.
(42) Kallio, P., Pásztor, A., Thiel, K., Akhtar, M. K., and Jones, P. R. (2014) An engineered pathway for the biosynthesis of renewable propane. *Nat. Commun. 5*, 4731.
(43) Sheppard, M. J., Kunjapur, A. M., and Prather, K. L. (2016) Modular and selective biosynthesis of gasoline-range alkanes. *Metab. Eng. 33*, 28−40.
(44) Rodriguez, G. M., and Atsumi, S. (2014) Toward aldehyde and alkane production by removing aldehyde reductase activity in Escherichia coli. *Metab. Eng. 25*, 227−237.

https://dx.doi.org/10.1021/acs.biochem.0c00665
Biochemistry 2020, 59, 3834−3843