ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Computationally Decoding NudF Residues To Enhance the Yield of the DXP Pathway

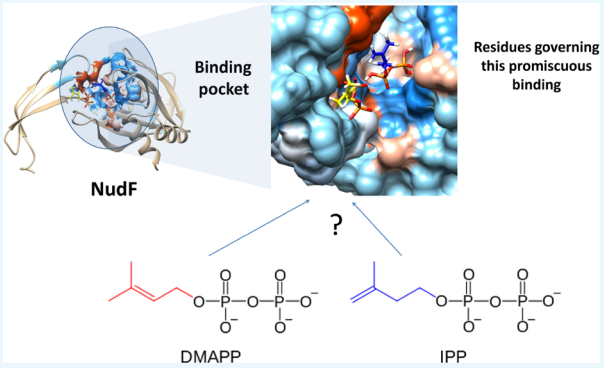
Devi Prasanna and Ashish Runthala*

Cite This: *ACS Omega* 2022, 7, 19898−19912

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Terpenoids form a large pool of highly diverse organic compounds possessing several economically important properties, including nutritional, aromatic, and pharmacological properties. The 1-deoxy-D-xylulose 5-phosphate (DXP) pathway's end enzyme, nuclear distribution protein (NudF), interacting with isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), is critical for the synthesis of isoprenol/prenol/downstream compounds. The enzyme is yet to be thoroughly investigated to increase the overall yield of terpenoids in the *Bacillus subtilis*, which is widely used in industry and is generally regarded as a safe (GRAS) bacterium. The study aims to analyze the evolutionary conservation across the active site for mapping the key residues for mutagenesis studies. The 37-sequence data set, extracted from 103 *Bacillus subtilis* entries, shows a high phylogenetic divergence, and only six one-motif sequences ASB92783.1, ASB69297.1, ASB56714.1, AOR97677.1, AOL97023.1, and OAZ71765.1 show a monophyly relationship, unlike a complete polyphyly relationship between the other 31 three-motif sequences. Furthermore, only 47 of 179 residues of the representative sequence CUB50584.1 are observed to be significantly conserved. Docking analysis suggests a preferential bias of adenosine diphosphate (ADP)-ribose pyrophosphatase toward IPP, and a nearly threefold energetic difference is observed between IPP and DMAPP. The loops are hereby shown to play a regulatory role in guiding the promiscuity of NudF toward a specific ligand. Computational saturation mutagenesis of the seven hotspot residues identifies two key positions LYS78 and PHE116, orderly encoded within loop1 and loop7, majorly interacting with the ligands DMAPP and IPP, and their mutants K78I/K78L and PHE116D/PHE116E are found to stabilize the overall conformation. Molecular dynamics analysis shows that the IPP complex is significantly more stable than the DMAPP complex, and the NudF structure is very unstable. Besides showing a promiscuous binding of NudF with ligands, the analysis suggests its rate-limiting nature. The study would allow us to customize the metabolic load toward the synthesis of any of the downstream molecules. The findings would pave the way for the development of catalytically improved NudF mutants for the large-scale production of specific terpenoids with significant nutraceutical or commercial value.



1. INTRODUCTION

Isoprenoids are the most functionally and structurally varied class of secondary metabolites, with over 55,000 identified molecules,[1,2] and have been used to make aromatic, flavoring, and medicinal molecules.[3−6] As the natural extraction of isoprenoid-based bioactives has led to an overexploitation of plants, the global research interest has now shifted to utilize the generally regarded as safe (GRAS) status microbes like *Bacillus subtilis* as the biocatalytic machinery.[7] A key protein controlling the yield of an end-product molecule is adenosine diphosphate (ADP)-ribose pyrophosphatase or NudF (EC 3.6.1.13) that orderly hydrolyzes dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP) to prenol and isoprenol (Figure 1), collectively termed as isopentenol (C_5 alcohol), a building block of all the higher-order downstream terpenoid molecules.[8] It belongs to the Nudix superfamily (Pfam PF00293; InterPro IPR000086)[9] and is involved in the production of adenosine monophosphate (AMP) and ribose-5-phosphate from ADP-ribose by actively dephosphorylating the phosphate moieties of a variety of substrates. Its close homolog is the NudB protein (EC: 3.6.1.67) from *E. coli* (EcNudB), having a catalytic activity toward geranyl pyrophosphate (GPP) and farnesyl pyrophosphate (FPP).[10] It has been suggested that an additional phosphatase activity (AphA) is needed to hydrolyze IPP to isoprenol because NudB can only catalyze the hydrolysis of IPP to isopentenol.[11] While EcNudB exhibits a strong affinity for DMAPP, the NudF protein of *Bacillus subtilis* (BsNudF) shows no such preference and equally

Received: March 21, 2022
Accepted: May 18, 2022
Published: May 27, 2022

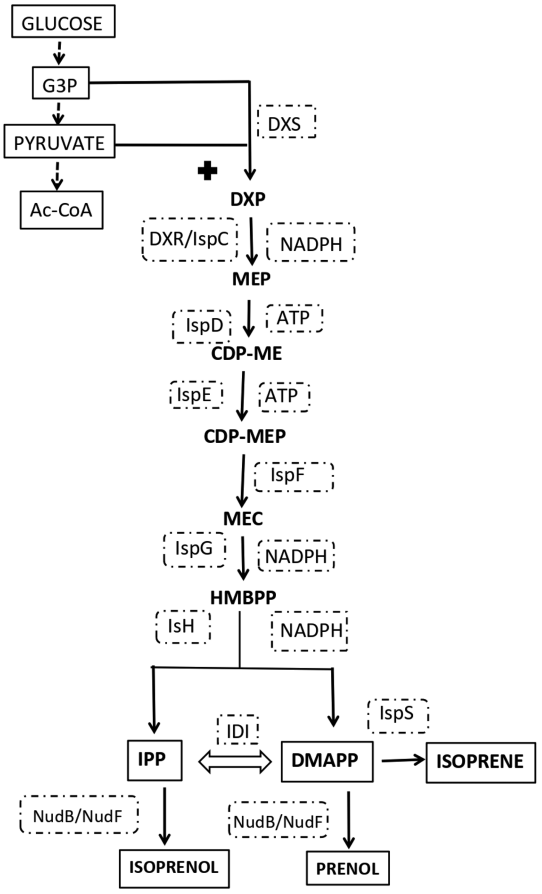


Figure 1. MEP pathway, representing the biocatalysts at all intermediary stages. From the glyceradehyde-3-phosphate (G3P) and pyruvate, it synthesizes the 5-carbon building blocks DMAPP and IPP for producing terpenoids. It includes an ordered set of seven enzymes viz. DXP, DXP reductoisomerase (DXR), 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase (IspD), 4-diphosphocytidyl-2-*C*-methyl-D-erythritol kinase (IspE), 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate synthase (IspF), HMB-PP synthase (IspG), and HMB-PP reductase (IspH), followed by the interplay of IDI and NudF.

interacts with both DMAPP and IPP to orderly produce an equivalent amount of prenol and isoprenol.[12] A varying affinity for various substrates may substantially increase pressure on the stoichiometric flux, and until the connected regulatory network of proteins does not drain out the added molecules at an equivalent rate, it becomes toxic to the cell and inhibits cell growth. Hence, to escape this major bottleneck, a fusion protein of isopentenyl-diphosphate delta isomerase (IDI), also known as isopentenyl pyrophosphate isomerase (IPP isomerase), and EcNudB has been utilized and with an increased expression, the metabolic flux increases the bioproduction rate of only prenol and reduces the isoprenol production. Endogenous overexpression of IspG and 1-deoxy-D-xylulose-5-phosphate synthase (DXP) as well as exogenous expression of YhfR and BsNudF have been shown to increase the production rate of downstream molecules in *E. coli*,[13] and the overexpression of NudF in *B. subtilis* has been shown to improve the isopentenol yield.[12]

Although sufficient IPP and DMAPP concentrations are required for terpenoid synthesis,[8] an excess of these molecules can inhibit cell growth,[12] and thus reduce terpenoid production,[14] and therefore, the NudF enzyme continuously uses these potentially harmful debris molecules. To increase the yield of the DXP pathway, researchers have specifically targeted the first rate-limiting enzyme, DXP[15,16] because its K_{cat}/K_m score is much smaller than that of all the other enzymes.[17] Improving the DXP activity has been considered as the most effective measure to improve the overall yield for several species, including *Streptomyces*,[18] *Lycopersicon esculentum*,[19] and *Synechococcus leopoliensis*.[20,21] Heterologous expression of *Bacillus subtilis* DXP in *E. coli* has also been shown to increase the overall yield.[22] Yet another strategy, using the functionally mutated recombinant poplar DXP,[23] has shown a negligible feedback inhibition for IPP/DMAPP and has been fruitful. To mitigate the toxicity of IPP, the farnesyl diphosphate synthase and IDI synthase have been overexpressed in *E. coli*, and it has led to an 800-fold production of sesquiterpene β-farnesene.[24] Likewise, the balanced metabolic concentration of IPP and DMAPP,[25,26] heterologous expression of *Haematococcus pluvialis* IDI,[27] or overexpression of *Lycium chinense* IDI in *E. coli*[28] have been shown to increase the overall yield. Recently, the modulation of the culture medium has also been demonstrated to increase the isopentenol production by 2.5-folds in *B. subtilis*.[29] Although the majority of these methods have overexpressed the rate-limiting enzymes or have changed the culture conditions, it may lead to a complete metabolic imbalance and significantly minimize the yield of downstream terpenoids. In this regard, directed coevolution of DXP, DXR, and IDI has been shown to increase isoprene production.[28] However, these strategies have also not efficiently focused the major enzyme BsNudF for increasing the overall yield, and directed evolution of NudF's promiscuous active site would therefore be significantly useful to increase the biosynthetic rate of a downstream terpenoid. For understanding the preferential binding of IPP/DMAPP, and catalytic and behavioral switching of BsNudF, this article deciphers the key functional details across the active site for decoding the mutations that could improve its activity. As it comprehensively maps the crucial residues at the functionally important positions, the study will be fruitful in designing a custom set of key interacting residues against a required ligand to attain a theoretically impossible overall yield. Although DXP has been shown to be the rate-limiting enzyme of the nonmevalonate pathway,[17] NudF, as the last biocatalyst to channelize the metabolic flux, should also play a significant role in regulating the overall yield of terpenoids, and its limited molecular engineering study should thus be of prime interest for biological and industrial research to optimally increase the productivity.

2. THEORETICAL CALCULATIONS

2.1. Construction of the Sequence Data set and Functional Affirmation. Screening the NudF protein sequences in the NCBI protein database, a set of 105,910 protein sequences is identified, among which only 103 *Bacillus subtilis* entries are found and are used to build the primary data set. Purging the completely redundant entries with at least 80% alignment coverage through MMSeqs2,[30] a reduced data set of 40 entries is derived, and their length pattern is noted. Three entries WP_139026580.1, WP_139026569.1, and CUB36584.1 orderly encoding only 118, 86, and 55 residues, are found to be the partial sequences and are purged. A final data set of 37 entries, including 1, 6, and 30 sequences orderly encoding 179, 205, and 185 residues, is thus created. As it is necessary to functionally confirm the defined data set for further analysis, the sequences are fed to MEME[31] for screening the conservation of the signature motif and affirming the functions of the sequences to use the functionally correct entries. In the absence of any

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

functionally similar experimentally resolved *B. subtilis* protein, the homologous *Escherichia coli* structure (PDB ID: 5U7E) is used as the representative entry to reliably localize the conserved motifs. Furthermore, to deploy the *Bacillus subtilis* representative sequence for the downstream analysis, the smallest sequence (CUB50584.1) is used from the constructed data set. It is because evolution tends to decrease a protein sequence length to pack the function more optimally.[32]

2.2. Modeling the Representative Sequence. As the *B. subtilis* representative sequence CUB50584.1 is still not experimentally determined, its tertiary structure is modeled through template-based modeling methodology using MODELER, as per the recently published strategy.[33,34] For reliably predicting the model, the wild-type nucleoside diphosphate sugar hydrolase from *Bdellovibrio bacteriovorus* (PDB ID: 5C7Q) is used as the template, sharing a sequence identity of 36%. As the first predicted protein model usually has several nonphysical local clashes, the constructed model is energetically relaxed through the conservative refinement strategy of Galaxyrefine2[35] to maximally retain the topology extracted from the template. The refined model is subsequently evaluated through Swissmodel scoring measures[36] and is also topologically assessed against the one, predicted using the Alphafold algorithm,[37] to assess its credibility.

2.3. Phylogenetic and Conservation Analysis. Statistically significant evolutionary relationships are typically considered for drawing the meaningful phylogenetic connections within the selected set of sequences/species. As a correct sequence alignment is certainly required for extracting the accurate evolutionary relationship, the 37 sequence *B. subtilis* data set is aligned through the hidden Markov model based clustal-omega module of HHPred.[38] The constructed profile is fed to the IQTree server[39] for deriving the evolutionary tree on the basis of the default parameters, using the default ultrafast methodology over 10,000 bootstrap alignments at the minimum correlation coefficient or the convergence threshold of 0.99. The resultant consensus tree is visualized and analyzed through ITOL.[40] Furthermore, the sequence alignment is fed to Consurf[41] for plotting the average sequence conservation scores over the predicted CUB50584.1 model, using the default conditions, and analyzing the sequence variations across the functionally crucial sites. Lastly, on basis of the functionally well-annotated *B. subtilis* (strain 168) homolog P54570, the constructed HMM profile is visualized using Espript[42] to map the encoded domains.

2.4. Active Site Prediction and Docking Analysis with IPP and DMAPP. To reliably channelize the interaction of substrates only at the biologically credible site(s) and to exclude any fake docking solution, the CastP version3 server is used to predict the active site(s).[43] For screening the most promising energetically feasible interaction site(s) of NudF against the substrates IPP and DMAPP, the Dockthor server is used,[44] using the default parameters. Through a set of 24 docking runs, it efficiently docks a ligand using the random seeds for all the rotational, translational, and conformational degrees of freedom of the ligand. Using a multiple solution genetic algorithm and the MMFF94S molecular force field scoring function, it constructs the docking solutions and assesses them on the basis of the docking energy, ligand entropy, and desolvation score to select their top-ranked cluster representatives. It is shown to be more accurate than seven state-of-the-art docking algorithms viz. AutoDock Vina, AutoDock, GOLD, Surflex-Dock, rDock, HPepDock, and Glide.[44]

In the absence of any experimentally resolved IPP-/DMAPP-NudF complex structure, the docking result constructed through this highly accurate Dockthor algorithm would certainly be the reliable data to excavate the functional details. Hence, to attain biologically correct predictions, a complete degree of rotational freedom is orderly allowed for the six and five rotatable bonds of the selected ligands. The resultant solutions are subsequently analyzed using UCSF Chimera to extract the key interacting residues.

2.5. Crucial Residues for Functional Mutagenesis. The accuracy of a rationally evolved enzyme molecule is dependent on the identification of the hotspot residues, whose mutations could enhance its catalytic activity.[45] To analyze the hotspot regions, proximal to the active site and tunnel, and appropriately map the mutational landscape of the predicted protein molecule, the hotspot server 3.1 is used.[46] Although this server is capable of modeling an input protein sequence, the constructed CUB50584.1 model is considered to drive the analysis for the topologically reasonable model. As this server robustly estimates the thermodynamic stability of a mutation through FoldX and Rosetta, it has been demonstrated to successfully exclude disruptive mutations.[47,48] To extract its fullest potential and reliably select the best mutations from the preselected resultant ones, the resulting data are analyzed along with the other residues marked in the preceding steps.

To understand the most promising mutations across the active site, the correlated mutations are mapped for the 179 residue sequence CUB50584.1 sequence through the GREMLIN server.[49] As these coevolution-based contacts are crucial to reliably discriminate the true hotspot residues from the spurious ones, the contact map network is analyzed to mark the true hotspots. Purging the unreliable positions, the true hotspots are analyzed through the Dynamut2 server,[50] trained, and tested on the Protherm database.[51] Its high accuracy is evident from the fact that it outperforms the other measures with a Pearson's correlation of up to 0.72 and 0.64 for the single and multiple point mutations with a root-mean-square error (RMSE) (kcal/mol) of 1.02 and 1.8, respectively, making it a trustworthy strategy for prioritizing the stabilizing/destabilizing mutations.[50]

2.6. Computational Mutagenesis and Flexibility Analysis. On the basis of docking and mutagenesis results, the most favorable binding and substrate preferences are excavated. To robustly quantify the strength of interaction for each of the ligands, the Dynamut2 results are analyzed for the five prioritized residues (ARG18, ALA117, ASP139, GLU140, and ASP141), and their five top-ranking $\Delta\Delta G$ scores are analyzed. To examine these data, NetsurfP2.0,[52] with an 80% correlation with experimentally confirmed data, is preferred to estimate the relative solvent accessible surface area for the five key positions for the native protein. To further examine the topological features of the NudF structure, the flexibility of its Cα-backbone is predicted using CABS-flex 2.0, which computes an average topological fluctuation diversity of 10 medoids of the 10 cluster sets, derived from a 10 ns simulated set of 1000 decoys.[53] To further obtain more insights of the binding affinity of the two ligands, a small 100 ns molecular dynamics (MD) simulation is finally done to analyze the physical movements of ligands and protein atoms. The IPP and DMAPP topologies are constructed through PRODRG2,[54] and the CUB50584.1 model and its two complexes are simulated through the WebGro server.[55]

To perform the simulation, the native protein and its complexes are prepared using the GROMOS96 43a1 force-

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

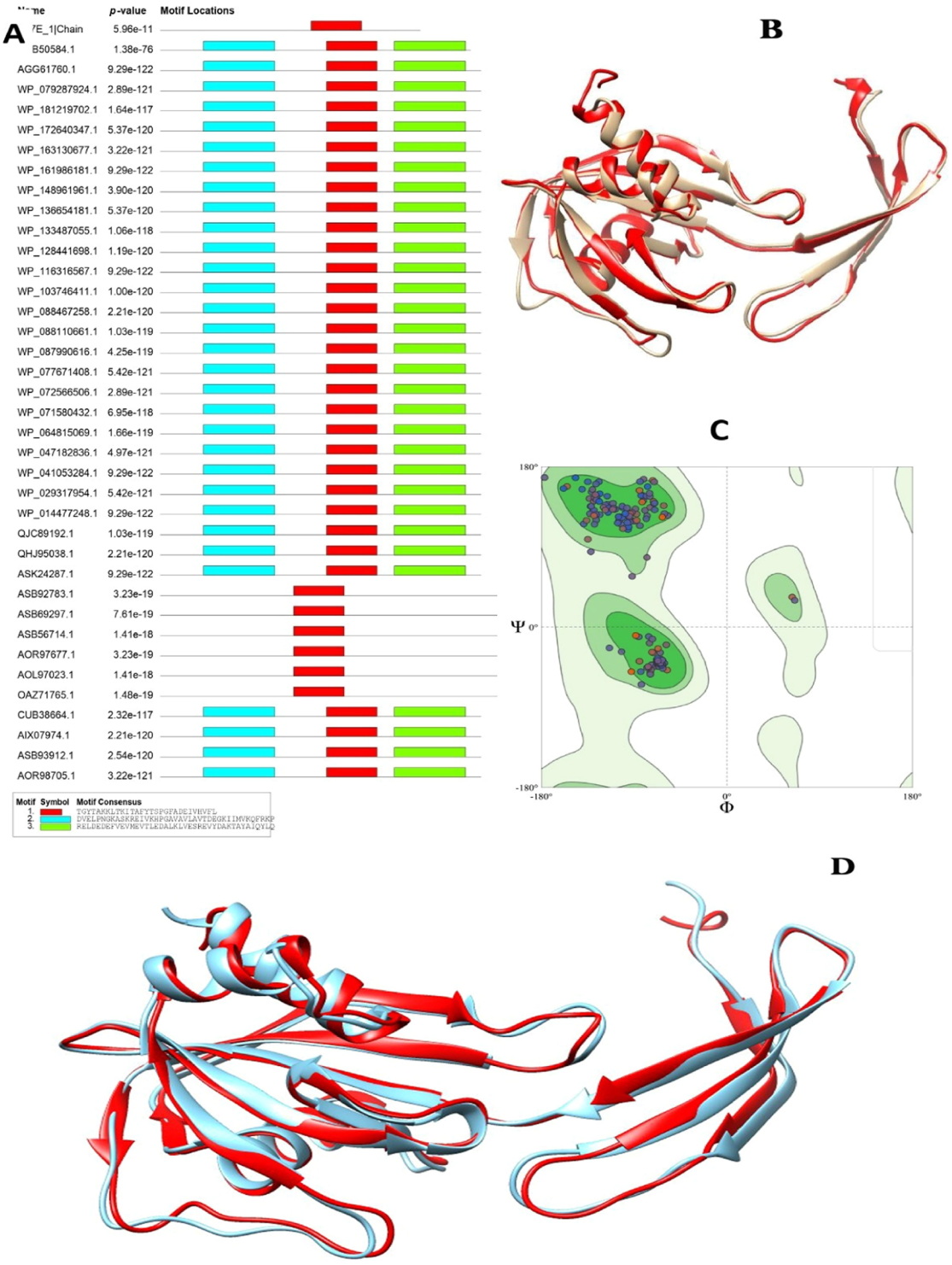


Figure 2. (A) Top-3 motifs, screened against the functionally deciphered structure 5U7E of *E. coli*, (B) structural overlap of the constructed model over the 5U7E, (C) Ramachandran map of the predicted model, and (D) topological superimposition of this model (red) over the Alphafold model (cyan). A complete overlap implies the topological accuracy of the predicted NudF structure.

field.[56] Deploying the SPC water model as a solvent over the triclinic simulation box, the structures are neutralized by adding 0.15 M sodium chloride. To energetically relax the system before MD, the steepest descent algorithm (5000 steps) is applied. Considering 1000 frames per simulation, the structures are simulated for 100 ns using the constant temperature of 300 K and a pressure of 1.0 bar. The trajectory is lastly integrated through the leap-frog algorithm for assessing the resultant trajectory through all of its encoded parameters, viz. time-dependent root-mean-square deviation (RMSD) for the overall structure and RMS fluctuations (RMSF) across its residues, radius of gyration (Rg), number of hydrogen bonds, overall and per-residue solvent accessible surface area (SASA), and ligand RMSD.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

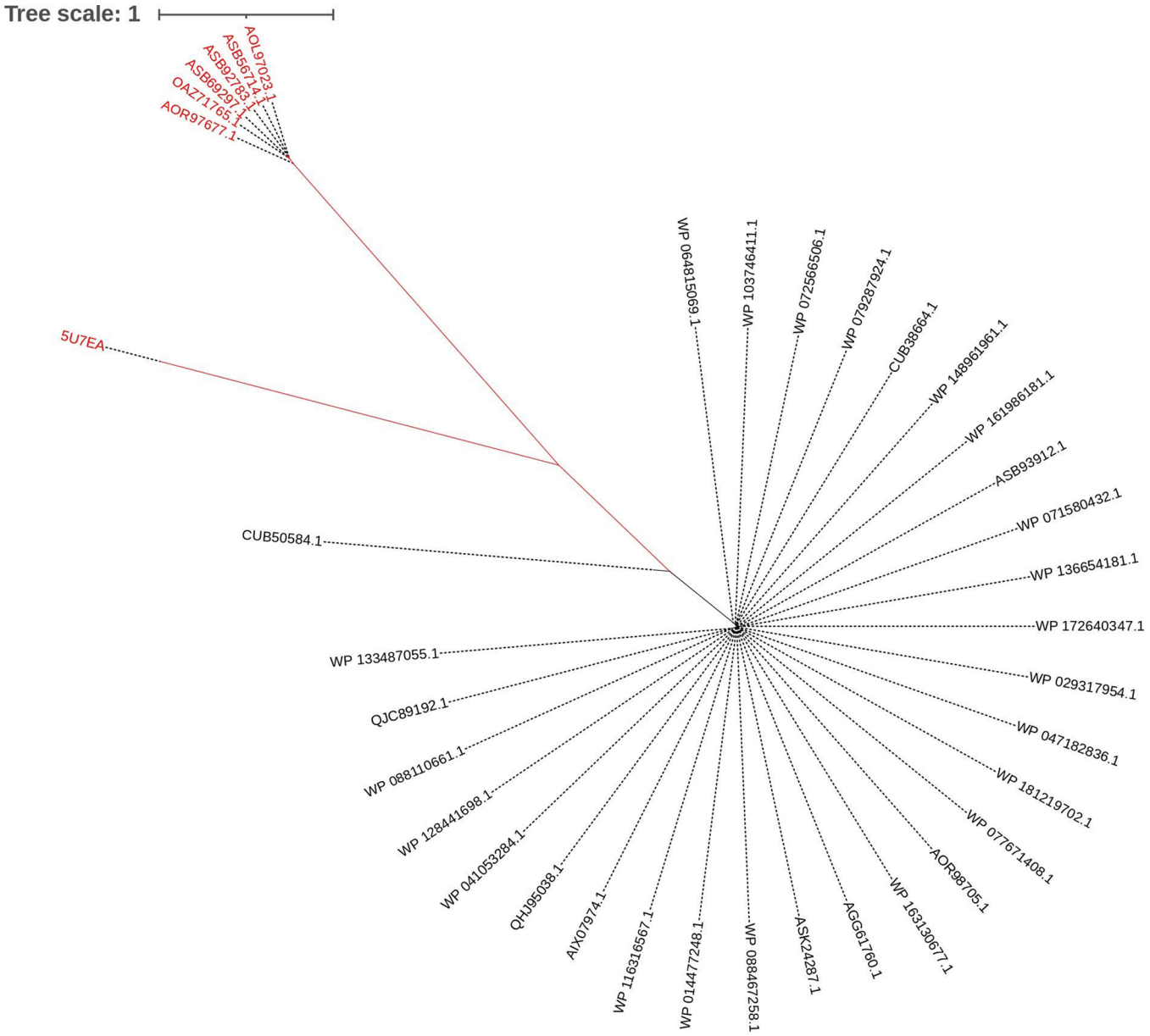


Figure 3. Unrooted circular tree of the 37 sequence NudF data set. The scale bar represents the average number of substitutions per site.

3. RESULTS AND DISCUSSIONS

3.1. Construction of the Sequence Data set and Its Functional Affirmation. To affirm the functional attribute of all the derived 37 sequences, the relative location of their three top-ranked conserved motifs is screened against the functionally deciphered structure (5U7E) of *E. coli* through MEME (Figure 2A). The members of the Nudix superfamily encode the signature sequence $GX_5EX_7REUXEEXG/TU$, where U is either L/V or I,[57] as represented in green in this figure. This conserved sequence is responsible for metal-binding and forms the catalytic site in more than 4000 enzymes in numerous species including eukaryotes, prokaryotes, and viruses.[58] The NudF length range is shown to be 179−205, with 182-residue sequence being the most common. Here, the six 205-residue sequence subset shows a stark feature, that is, ASB92783.1, ASB69297.1, ASB56714.1, AOR97677.1, AOL97023.1, and OAZ71765.1encode only one characteristic motif, as observed for 5U7E, unlike the other *B.subtilis* sequences having all the three conserved motifs.

3.2. Modeling the Representative Sequence. The representative *B. subtilis* sequence CUB50584.1 is modeled using 5C7Q through the template-based modeling methodology using MODELLER9.25. The predicted structure is refined using GalaxyRefine2[35] and is subsequently evaluated through Swissmodel.[36] The model shows a Molprobity score of 1.28, indicating a topological accuracy of a reasonably high accuracy X-ray structure, and its RMSD score is found to be 0.887 against the deployed template (Figure 2B). Furthermore, 98.87% model residues are found localized within the Ramachandran favored regions (Figure 2C). The model also shows a low RMSD of 1.116 against the structure, predicted using Alphafold algorithm,[37] and these two decoys are found to be structurally superimposed, with a marginal deviation of a few loop and terminal residues (Figure 2D). It is thus confirmed that Alphafold specifically works well especially for proteins, sharing a sequence identity lesser than 30% against the known templates, as reported recently.[37] The predicted structure is then reliably deployed for the subsequent analysis.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

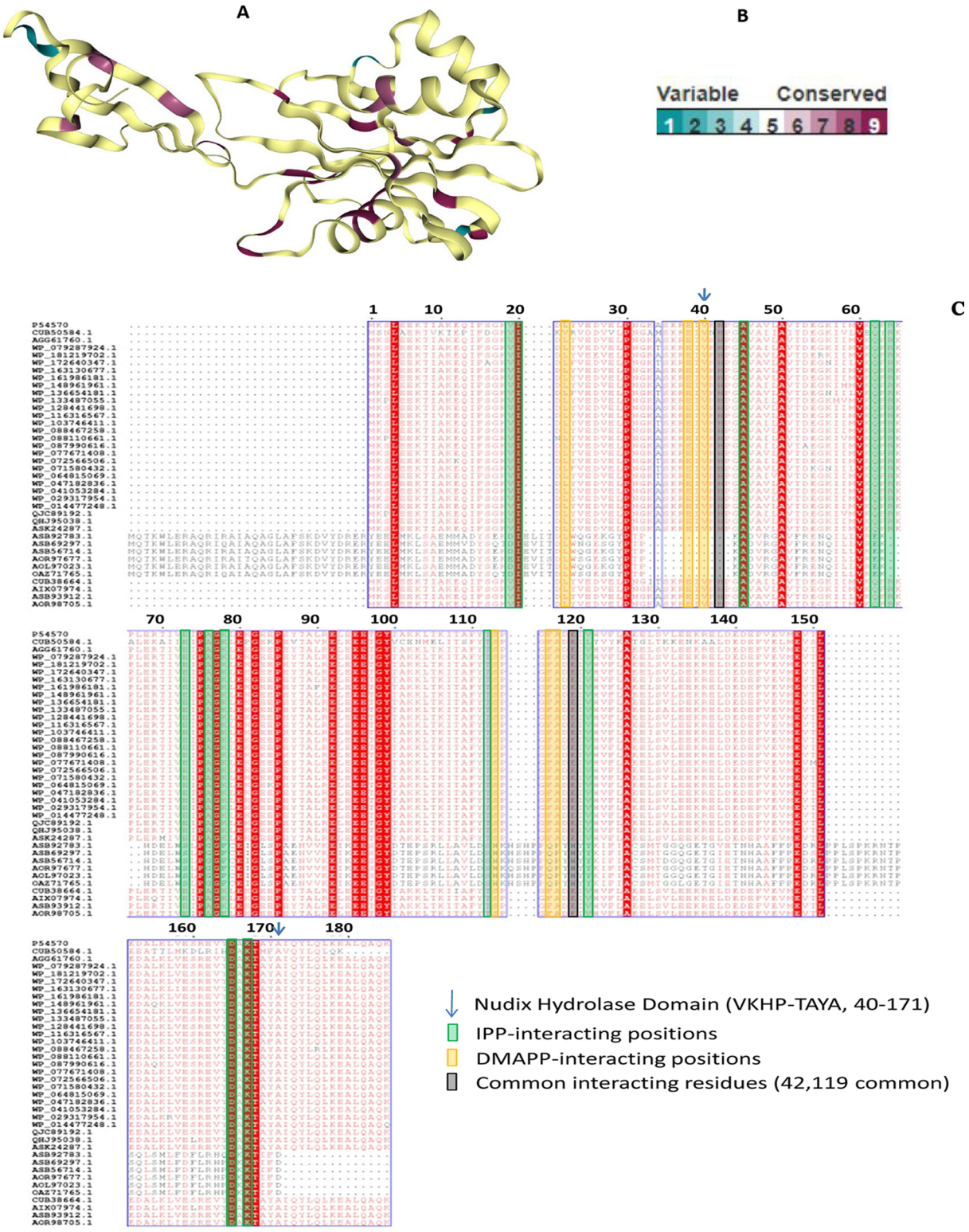


Figure 4. (A) Consurf-derived sequence conservation for the 37 sequence NudF data set, projected onto the representative sequence model CUB50584.1. (B) Color range varying between 1 and 9, maroon referring to the completely conserved positions. (C) Complete overlap of the IPP- and DMAPP-interacting positions and Nudix domains of P54570 (residues 40−171) and CUB50584.1 (residues 42−177).

3.3. Phylogenetic and Conservation Analysis. Feeding the sequence alignment of HHPred-clustal omega to IQTree and building the evolutionary tree using 10,000 bootstrap alignments at the convergence threshold of 0.99, the resultant consensus solution is visualized and analyzed through ITOL (Figure 3). Although the single motif-six sequence subset (AOR97677.1, OAZ71765.1, ASB69297.1, ASB92783.1, ASB56714.1, and AOL97023.1) appears to be clustered with 5U7EA and the representative sequence CUB50584.1, encoding all three motifs, their average sequence identity is 30.3640.339, compared to 14.73 and 21.3630.3304 against 5U7EA and CUB50584.1, respectively. Moreover, all the other sequences

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

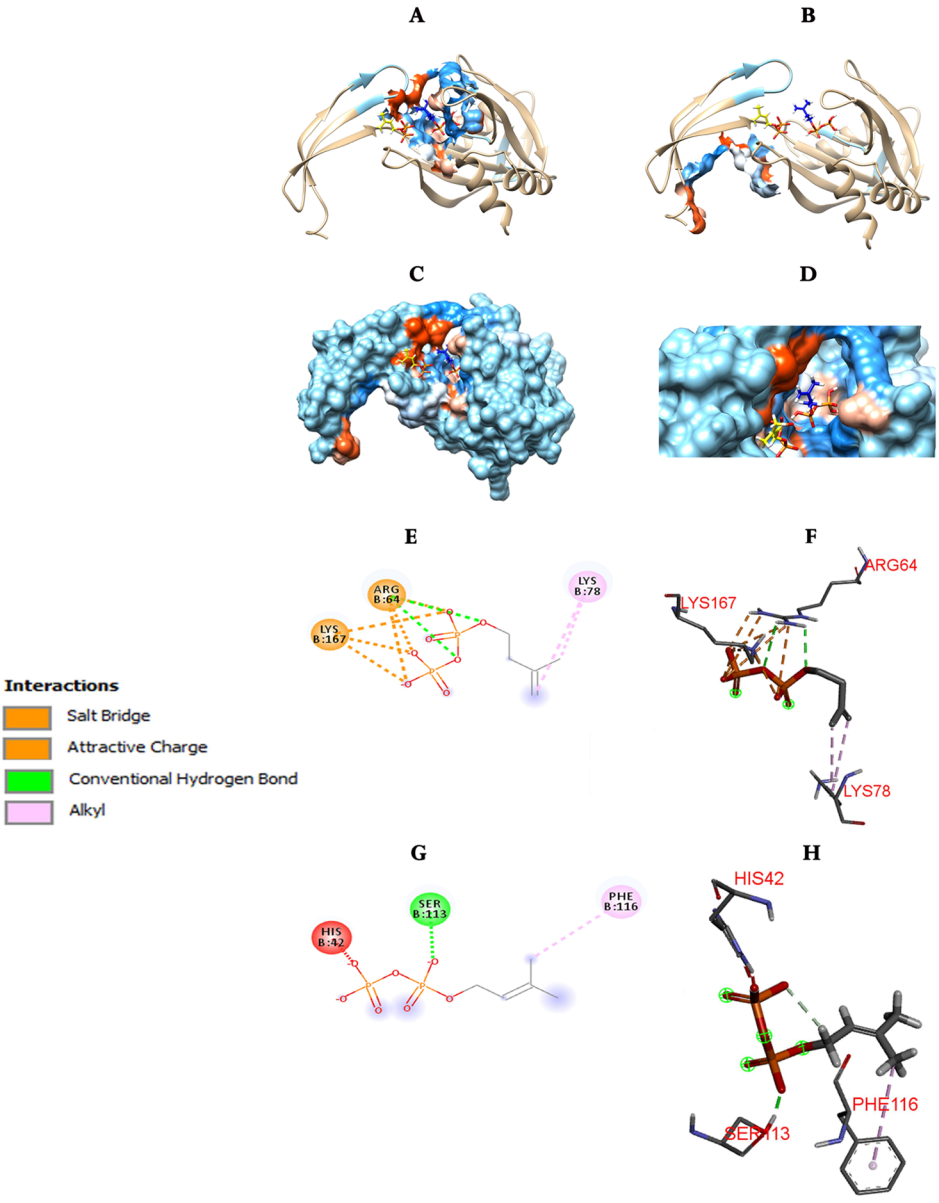


Figure 5. Dockthor results of both IPP and DMAPP. (A) Pocket1 (B) Pocket2 represented in correlation with the two ligands IPP (yellow) and DMAPP (blue), along with the close analytical view of Pocket1, in terms of (C) molecular surface. (D) Active site topology at the vent of the tunnel. (E) Two-dimensional and (F) three-dimensional interaction maps of protein with DMAPP. (G) Two-dimensional and (H) three-dimensional interaction maps of protein with IPP show the key residues interacting within the active site.

are not found to share substantial sequence similarity, as also observed with their scrutiny through HMM-based Clustal Omega alignment.[38] All the other sequences are found uncluttered with any other entry, and as recently shown in the phylogenetic study of 80,000 Nudix homologs, a general monophyly is prominent, besides a few occasional incidences of homoplasy.[59] The constructed profile is further fed to Consurf for mapping the average sequence conservation scores over the predicted structure, using the default conditions,[41] and the sequence variation across the functionally crucial sites is analyzed (Figure 4). The HMM alignment of the 37 sequence data set and the functionally decoded protein P54570 is mapped using Espript[39] to designate the essential sequence features of the CUB50584.1 sequence. It is observed that the Nudix domains of CUB50584.1 and P54570 fully overlap (Figure 4C).

A set of 14 residues, viz. ILE20, ALA45, GLN62, GLY77, GLU80, GLY82, ALA89, GLU92, GLU95, GLU96, THR112, ALA126, THR130, and GLU148 is found to be completely conserved, with a conservation score of 9. Moreover, 33 residues LEU4, THR8, PHE15, PRO30, ASN31, LYS36, ILE39, HIS42, PRO43, ALA50, LYS56, VAL60, LYS65, ILE71, PRO85, THR88, ARG91, LEU93, GLU94, THR97, LEU106, ILE107, GLU119, TYR124, ASP141, VAL144, ALA154, LEU157, ASP166, LYS167, THR168, PHE170, and GLN173 show a conservation score of 7 and 8, indicating significantly higher conservation. Furthermore, 11 residues, viz. PRO13, ARG23, VAL24, ALA33, MET34, LYS69, LYS131, SER150, GLU153, ASP160, and HIS164 are found to be completely variable across the defined data set. It is astonishing that the representative *B. subtilis* sequence encodes mere 11 (6.145%) variant residue loci in contrast to the statistically higher sequence conservation at 47 (26.259%) positions (Figure 4), and still, the enzyme is able to show a highly promiscuous nature, and it indicates that a few key

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

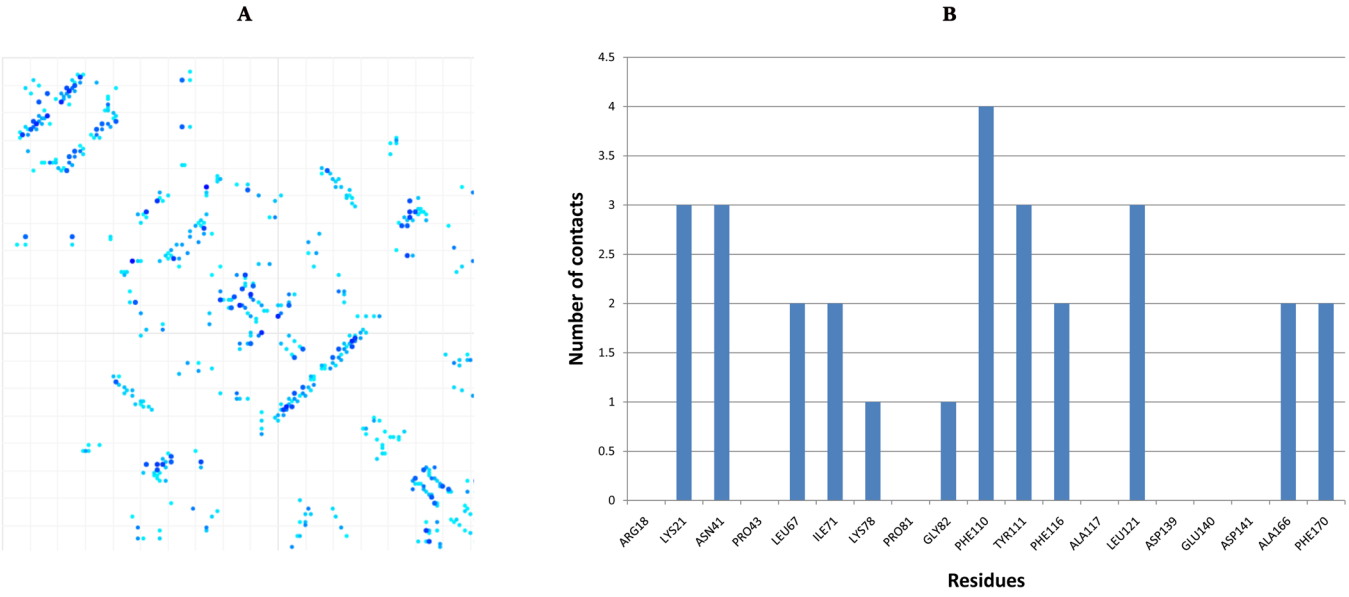


Figure 6. GREMLIN resultant (A) contact map network and (B) number of contacts for the 19 key hotspot residues, placed proximal to the active site.

residues should play a major role within the active site of this enzyme.

3.4. Active Site Prediction and Docking Analysis with IPP and DMAPP. Excluding the shallow openings and using a probe radius of 1.4 Å, CastP[43] delineates the empty concavities on a protein surface to map the volume spectrum of cavities and pockets. It robustly deciphers the surface properties and localizes the functionally important zone and shows only two biologically meaningful active pockets in the predicted NudF model, with a molecular surface area (Å^2) and molecular volume (Å^3) of 501 and 946.6 (Pocket1) and 208.9 and 661.3 (Pocket2), respectively, and this is in accordance with the known structural details of the NudF's *E. coli* homolog, wherein two active sites have been observed.[60] Furthermore, these pockets are orderly found to span a set 12 (MET1, LEU4-A-GLU6, ARG37, ILE39, TYR111, PRO114, GLY115, ALA117, ASP118, and ILE120) and 30 (ARG18-VI-LYS21, VAL40-NH-PRO43, ALA45, VAL60, GLN62-YRK-ALA66, GLU73-IPAG-LYS78, GLU96, THR112, GLU119, LEU121, ASP141, GLU142, ASP165, ALA166, and LYS167) residues. Here, it is worth noting that Dockthor uses only one most-voluminous docking site for both ligands (Pocket1, Figure 5A), and it indicates a preferential binding of ligands over the second superficial site (Pocket2, Figure 5B).

As a modest level of NudF is enough to overcome the IPP/DMAPP toxicity,[12] this enzyme must interact with both of these substrates, and this responsible molecular NudF surface (Figure 5C) closely interacts with the two ligands (Figure 5D). The ligands IPP and DMAPP orderly show an interaction energy (kcal/mol) and affinity score (kcal/mol) of −115.388 and −6.431, and −41.402 and −5.271. Through random forest, Dockthor has tested the predicted binding affinity over the PDBbind v2013 data set, and it has shown the RMS error of 2.256 kcal/mol at a correlation coefficient of 0.705, indicating a reasonable credibility of the predicted docking solution.[61] Examining the docked complexes using the discovery studio, it is observed that DMAPP has two interactions with NudF. While the DMAPP-O atom forms a hydrogen bond with the Ser113-HG atom of NudF with a bond length of 1.355 Å, the DMAPP-C atom shows hydrophobic π-interactions with a DMAPP-C atom with a bond length of 5.085 Å. IPP, on the other hand, it has favorable interactions with NudF, and its oxygen atom forms conventional hydrogen bonds with NudF-ARG64-HH21 and -HH22 atoms, with bond lengths of 2.912 and 2.592, respectively, as well as two salt bridges with ARG64-HH12 and LYS167-HZ2 atoms, with bond lengths of 2.561 and 2.753, respectively. Moreover, the IPP-O atom shows three electrostatic interactions with ARG64-NH1, −NH2, and LYS167-NZ atoms with bond lengths of 5.255, 3.963, and 3.212 Å respectively.

Screening the active site residues within 5 Å of these two ligands, it is observed that 14 residues viz. VAL19, ILE20, HIS42, ALA45, GLN62, ARG64, GLU73, ALA76, LYS78, THR112, GLU119, LEU121, ASP165, and LYS167, and 8 residues viz. VAL22, GLU38, VAL40, HIS42, SER113, PHE116, ALA117, and GLU119 orderly interact with the two ligands IPP (yellow) and DMAPP (blue), as shown in Figure 4C. To analyze it further, the two-dimensional and three-dimensional interaction maps are drawn for DMAPP (Figure 5E,F) and IPP (Figure 5G,H). It indicates that two key hotspot residues LYS78 and PHE116, orderly responsible for interacting with these ligands through one and two residues in the active site, could be the key to specifically alter the active site to stabilize its affinity for the conditionally required ligand.

3.5. Crucial Residues for Functional Mutagenesis. Through GREMLIN,[49] the contact map network of the modeled protein is constructed. For all the predicted contacts, it yields the two-dimensional distance matrix along with their probability scores and is shown to be accurately predicting both direct and indirect residue couplings. The coevolution-based network for CUB50584.1 is analyzed (Figure 6A), and the statistically top-ranked contacts are extracted. A set of 22 functional hotspot residues, viz. ARG18, ILE20, LYS21, ASN41, PRO43, LEU67, ILE71, LYS78, LEU79, PRO81, GLY82, PHE110, TYR111, THR112, PHE116, ALA117, LEU121, ASP139, GLU140, ASP141, ALA166, and PHE170, with a theoretically credible probability score higher than 0.5, are found in NudF. However, LEU79 is found to be completely buried in the structural core of NudF. Moreover, ILE20 and THR112 are found to be completely conserved and are thus excluded. The 19

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

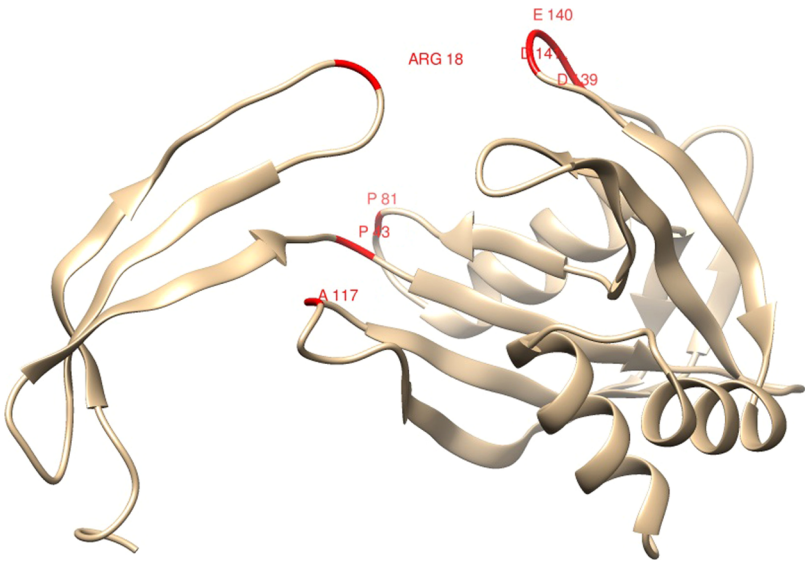


Figure 7. Seven key hotspot residues excavated for designing the functionally improved mutants.

key hotspot residues are found to define a structural network of 0−4 contacts. As previously stated,[62,63] the loop proline residues usually stabilize a structural bend through internal hydrogen bonding, and PRO43 and PRO81 are also not considered for subsequent computational mutagenesis study. Plotting the number of contacts for the hotspot residues, seven positions, viz. ARG18, PRO43, LYS81, ALA117, ASP139, GLU140, and ASP141 are not found to have any contacts (Figure 6B), and as the delicate balance between the flexibility and stiffness is sustained by the key contacts, majorly regulating the functional role(s) of a protein, these positions (Figure 7) are selected for further analysis.

3.6. Computational Mutagenesis and Flexibility Analysis. Excavating the key residues, the Dynamut2 server is used to estimate the stability changes upon a point mutation on the CUB50584.1 model. The method derives the scores through the topological environment property and dynamic behavior of a residue and is recently shown to outperform the prediction measures including FoldX and MAESTRO.[50] Restricting the search to the prioritized residues, the most stabilizing top five mutations are selected. The $\Delta\Delta G$ scores (KJ/mol) of these mutations ranges from −1.67 to 1.06 (Figure 8, Supplementary Table 1). To correctly drive the analysis, the relative surface accessibility (RSA) for the selected residues is computed using NetSurfP2.0.[52] It reveals how easily a residue can be accessed from the protein surface, and as the mutations at this position are less likely to disrupt the overall fold of a protein, the chance of developing a functionally fruitful mutation should be substantially higher. The prioritized residues orderly show an RSA score of 0.393, 0.232, 0.536, 0.696, 0.622, 0.449, and 0.397 (Figure 8A), categorized as exposed or buried at the threshold of 0.25.

To excavate the active site and its key catalytic residues, the average structural fluctuations across NudF are predicted using CABS-flex 2,[53] and a set of 8 loops: loop1−loop8 (Figure 8B), spanning from ILE14-ARG23, ASP26-ARG37, ILE39-VAL46, ALA50-VAL58, GL62-ILE72, GLY77-PRO85, TYR111-LEU121, and THR130-VAL144, marked in red in Figure 8C, are found to be flexible. These loops, encompassing a 48.6% structure, majorly define the activity site topology and are envisaged to substantially regulate the functional nature of NudF. It is interesting to observe that the selected five residues ARG18, ALA117, ASP139, GLU140, and ASP141 are encoded by the loop1, loop7, and loop8, and it could be there that the loop1 acts as a capping loop and loop7 and loop8 make the active site sufficiently voluminous to interact with IPP and DMAPP. It is similar to *E. coli* NudF loop9, which is stabilized by its closed conformation.[60]

To further analyze the dynamic atomic interactions of IPP/DMAPP within the active site, 100 ns NudF and its complexes are simulated using WebGro, as shown recently.[64,65] MD simulation is a robustly accurate strategy to analyze the conformational changes that occur when a ligand is induced to fit.[66] Using the GROMACS-based Webgro[55] server, the protein/complex system is computationally evolved using the classical mechanics algorithm for a short 100 ns timespan, and the conformational stability or binding affinity of a ligand is assessed across the simulation trajectory. To analyze the simulation results, Rg or the average distance between the center of mass and the rotational axis is usually used to estimate the conformational stability of a system against any physicochemical strain,[67] and its lower score implies a higher stability. Here, the apoprotein shows the Rg scoring variations between ∼1.5 and 2.75 nm (Figure 9), and its complex with DMAPP and IPP orderly shows the respective scores of ∼1.8−2.05 and ∼1.875−2.05 nm, indicating that the apoprotein is relatively more unstable than its complex structure, exactly like *E. coli* NudF.[60] Moreover, the IPP complex is more stable than the DMAPP complex because its trajectory mostly crawls at substantially lower scores across the simulation. The RMSD is another helpful measure for estimating the structural stability and the overall deviation from the backbone topology during the complex formation at a specified temperature, as used earlier by GROMACS.[68,69] The backbone-based RMSD trajectory score should be the lowest because it directly gives the deviation of mean atomic coordinates and reflects the macromolecule's conformational stability. The scrutiny of the RMSD across the trajectory shows that the RMSD variations of the apoprotein range between the acceptable range of ∼0.5−2.5 nm in contrast to the respective ranges of 0.3−0.9 and 0.3−0.75 for its DMAPP and IPP complexes (Figure 9). Unlike the DMAPP complex trajectory, showing substantially higher undulations, the IPP complex gets stabilized after ∼60 ns.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

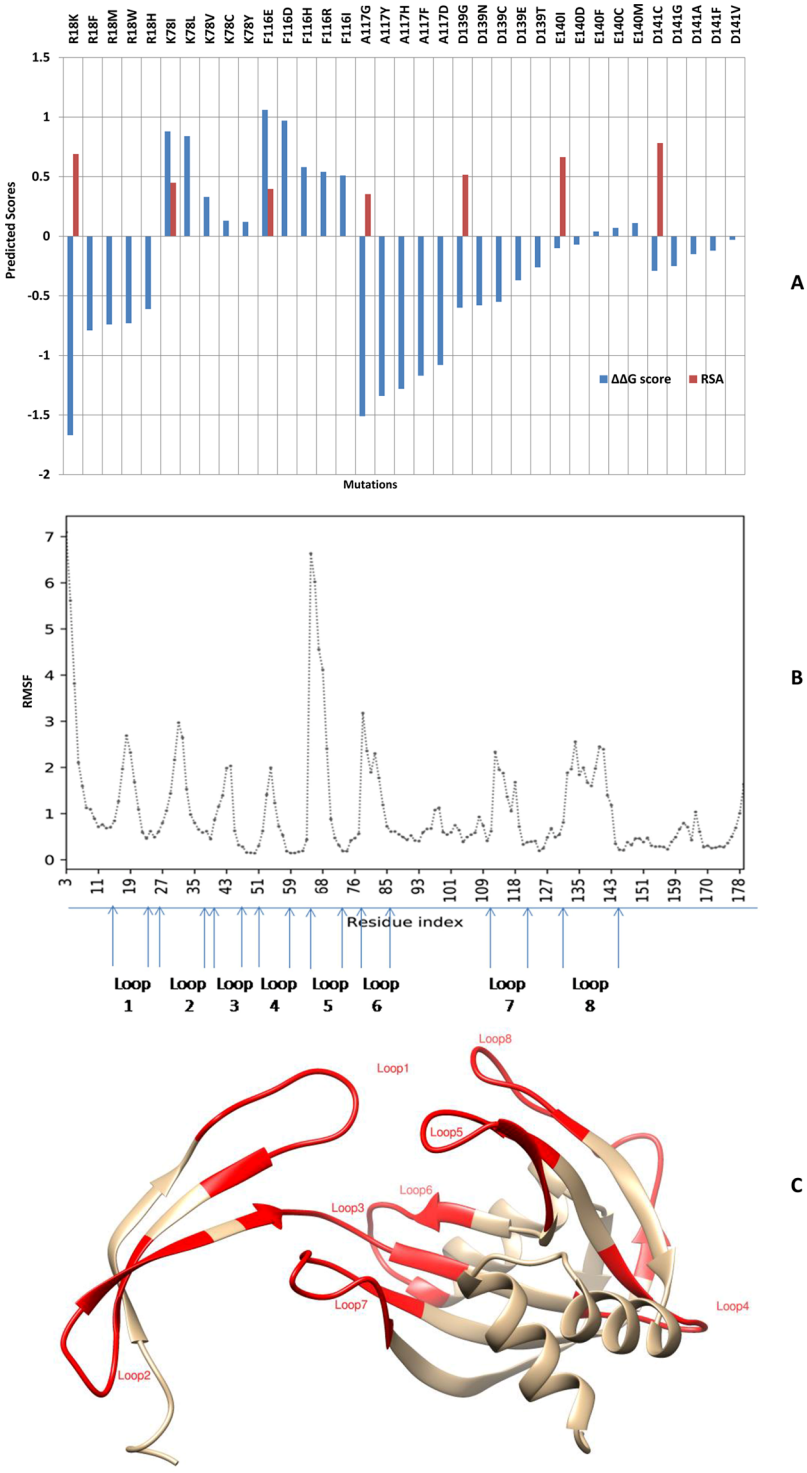


Figure 8. Structural scrutiny. (A) DDG−RSA plot for the five top-ranked mutations of the seven prioritized residues, (B) predicted NudF flexibility showing a high score for the eight loop regions, and (C) structural position of the eight flexible loops. A lower DDG score implies a more energetically stable conformation.

Furthermore, the RMSF, or residue fluctuations across the simulation, reveals a macromolecular system's conformational stability, and its scoring undulations describe structural complexity, with a lower value signifying higher overall stability.[56,63] For NudF, RMSF is found to be within ∼0.5−2 nm (Figure 9), although its three C-term loop residues and two preceding terminal 11-residue α-helices, connected through a five-residue loop, shows a stark increment crossing ∼3 nm. In contrast, the DMAPP and IPP complexes orderly show the RMSF divergence between ∼0.1−0.8 and ∼0.06−0.525 nm, and here, the terminal double-helix does not show any significant displacement. It is interesting to note that the RMSF score across the chain for NudF-IPP is lesser than the NudF-DMAPP scores, indicating that the former is a more stable complex structure.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

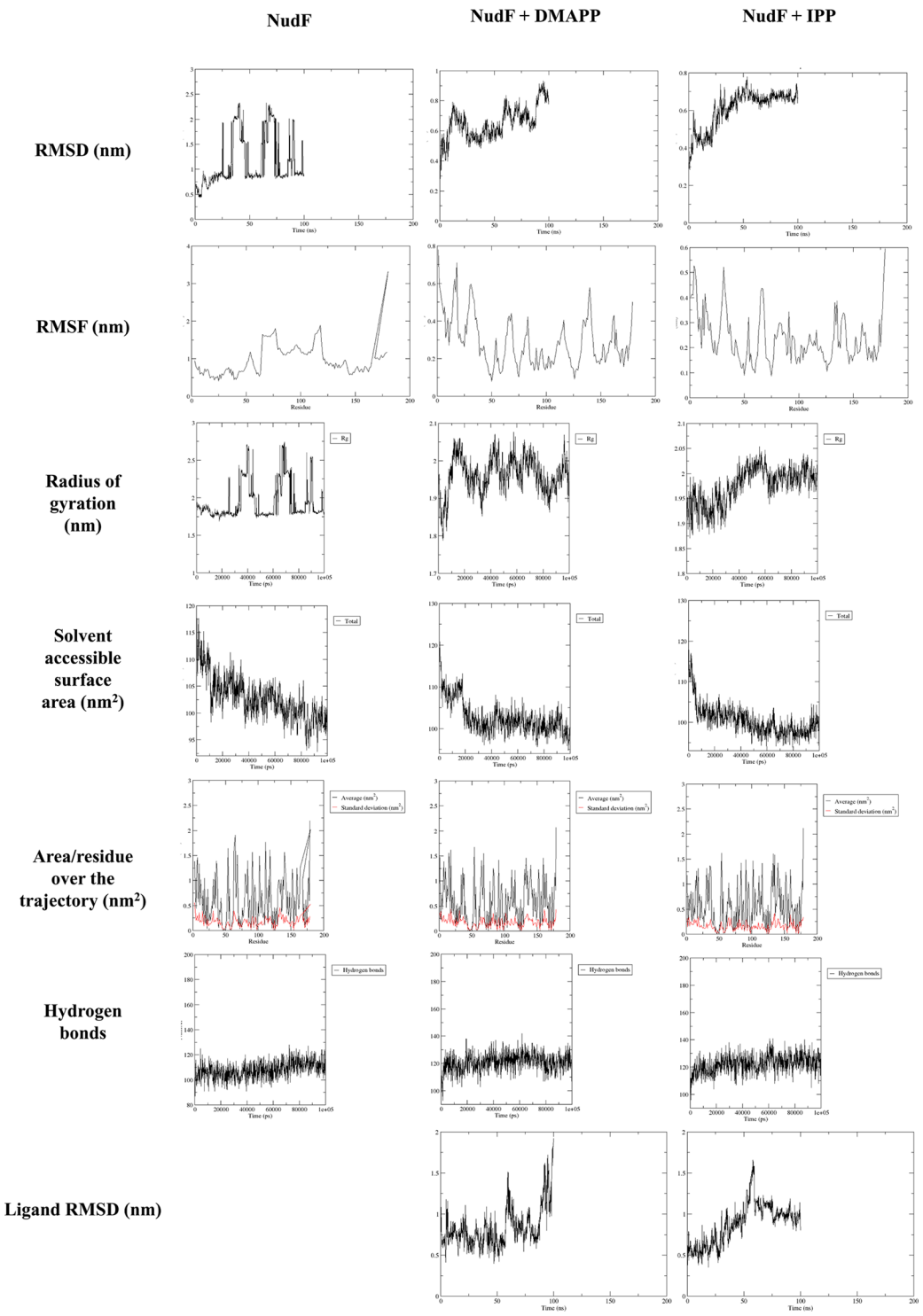


Figure 9. MD analysis for the apoprotein NudF and its DMAPP and IPP complexes for the RMSD, RMSF, radius of gyration, solvent accessible surface area, area/residues, hydrogen bonds, and ligand RMSD scores. It clearly indicates a prioritized binding for IPP within the promiscuous active site of NudF.

As the ligand binding causes conformational changes in the receptor, its respective substructural alterations are likely to affect SASA, and thus, it has been utilized multiple times to assess the level of receptor exposure to the surrounding solvent molecules.[67,68] Although a significant deviation is not observed in the per-residue SASA score, usually showing a comparatively large surface area or a lower compactness of this substructure, the SASA scores are found to be highly variant across the simulation trajectory. The SASA for the NudF structure is observed to consistently drop along the trajectory from 112.5 to 97.5 nm^2 (Figure 9), although the respective score of the DMAPP complex declines from 117.5 to 97.5 nm^2, with a sharp drop at 20 ns. In contrast, the corresponding score for the IPP complex declines from 112.5 to 100 nm^2, and except for a few time intervals, minute variations were observed throughout the simulation period. SASA for the IPP complex drops dramatically

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

from 110 to 103 nm^2 after 5 ns. The equivalent declination for the DMAPP complex, on the other hand, is noticed after 17.5 ns, and its trajectory has significantly more unequal undulations. NudF also exhibits this distinctive declination after 10 ns, and here, it could indicate a significant functional transition, leading to the substantial increase in structural compactness. Moreover, after the initial decline, the SASA graph moves likewise for both DMAPP and IPP complexes.

For estimating the binding affinity of IPP/DMAPP with NudF (Figure 9), the MD trajectories are further investigated to analyze the extent of hydrogen bonds made along the simulation, as usually done.[67,68] Along the trajectory, NudF shows a nearly 100−120 hydrogen bonds. However, the DMAPP and IPP complexes orderly show 100−130 and 110−130 hydrogen bonds, and an almost similar variation throughout the trajectory. It indicates their nearly similar scale of atomic interaction within the active site, as also shown by their substantially similar SASA undulations. Although, observing the earlier results, the affinity of DMAPP should not be comparable to that of IPP, consistency of hydrogen bonds is maintained for both the ligand complexes throughout the trajectory, and it indicates the comparable stability of these complexes. To excavate it further, the protein−ligand hydrogen bonding variations are analyzed through the trajectory, and for the DMAPP complex, the number of hydrogen bonds is found to slowly increase to two with significantly variant undulations. However, for the IPP complex, the number of bonds consistently increases, and after ∼60 ns, nearly one bond is maintained throughout the simulation, showing its more stable interaction. As hydrogen bonds are the major interactions to drive the proper anchoring of ligands within the active site, the higher number of such bonds should be responsible for a stronger interaction.[67,68]

Figure 8 deciphers three key features. First, the ddG score is found to be the highest for PHE116 and LYS78 residues, and it confirms that the two residues certainly hold the key to actively evolve the enzyme against the substrates by increasing the structural stability. Second, ASP139, GLU140, and ASP141 show a remarkably insignificant ddG score, and it implies that these positions are highly crucial for the NudF function and their top-ranked mutations also failed to stabilize the protein, as earlier discussed by Nobel Laureate Frances Arnold.[70] However, these three residues, along with the substructure T130-L138, show a high RSA score, and it indicates that the flexibility of this superficial loop region could impart a significant functional attribute to NudF. Restricting the search to LYS78 and PHE116 indicates that these positions should be significantly crucial for the stability and interaction affinity of the active site. Third, as LYS78 is superficially more exposed in comparison to ARG116 and is only in contact with one other residue, it should be first mutated to study its effect on the overall product yield. It opens venues for their experimental verifications as the top-ranked mutants K78I/K78L and PHE116D/PHE116E could selectively stabilize the conformation and could be responsible for the ligand specificity, urgently needed to design the novel industrially useful NudF enzyme.

The study extends our understanding about ADP-ribose pyrophosphatase and shows that it has a preferential bias for IPP over DMAPP, with −115.388 (kcal/mol) versus −41.402 (kcal/mol), respectively, although if the former is missing, the protein interacts with DMAPP at a much slower rate, and probably, this could be a key signal to IDI to initiate the conversion of DMAPP to IPP. Recently, the promiscuous activity of EcNudB toward geranyl diphosphate and farnesyl diphosphate has been demonstrated to generate several isoprenoid alcohols, including isopentenol, geraniol, and farnesol, as well as their derivatives.[71] Thus, the promiscuous dephosphorylation of NudF should be studied further through various other substrates, and with that, it would open venues to industrially engineer the cells, wherein the downstream reactions could be channeled more actively with minimal cellular regulation.

4. CONCLUSIONS

The research thoroughly examines the *Bacillus subtilis* ADP-ribose pyrophosphatase and its 37 functionally confirmed homologs with the projected tertiary structure of the 179-residue representative sequence CUB50584.1. Besides analyzing the phylogenetic relationship, it maps the highly conserved and variant loci to build the knowledge base for its directed evolution experiments. Although the sequence data set shows a significantly high phylogenetic divergence, 26.259% residues are found to have a statistically higher evolutionary conservation. ADP-ribose pyrophosphatase shows a prioritized interaction with IPP than DMAPP, according to the docking energy data, with −115.388 (kcal/mol) versus −41.402 (kcal/mol). The topological variations are restricted to eight loop regions, maximally encompassing the active site, demonstrating their importance for the ligand binding. Seven residues (ARG18, PRO43, PRO81, ALA117, ASP139, GLU140, and ASP141) are not found to have even one contact in the contact map network of 22 hotspot sites, and mutational analysis for these seven positions shows the highest $\Delta\Delta G$ scores for LYS78 and PHE116, orderly encoded within loop1 and loop7, encapsulating the active site. Quite similar to NudF, the four top-ranked mutants F116E, F116D, K78I, and K78L show the highest $\Delta\Delta G$ scores of 1.06, 0.97, 0.88, and 0.84, and it indicates that these positions should be the key to maximally direct the synthesis of required terpenoids in *Bacillus subtilis*. MD analysis reveals that the NudF structure is unstable, just like *E. coli* NudB, and that its IPP complex is more stable than the DMAPP complex. Thus, the present study must be industrially useful to channelize the entire DXP pathway toward the increased production of prenol or isoprenol or their downstream molecules without generating any metabolic burden.

AUTHORS CONTRIBUTIONS

A.R. conceived the study. D.P. performed the experiments. Both the authors have written the manuscript and have carefully read and approved the manuscript.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01677.

$\Delta\Delta G$ and relative solvent accessible surface area scores of seven key hotspots selected for the computational mutagenesis (PDF)

AUTHOR INFORMATION

Corresponding Author

Ashish Runthala − *Department of Bio-Technology, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP 520002, India;* orcid.org/0000-0002-1835-2755; Email: ashish.runthala@gmail.com

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

Author

Devi Prasanna − *Department of Bio-Technology, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP 520002, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c01677

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors whole-heartedly acknowledge Prof. Balasubramanian Gopal and Prof. Shibasish Chowdhury for driving the career toward directed evolution and metabolic engineering. The authors also thank the University and the department for providing the required resources/support.

■ REFERENCES

(1) Vickers, C. E.; Williams, T. C.; Peng, B.; Cherry, J. Recent advances in synthetic biology for engineering isoprenoid production in yeast. *Curr. Opin. Chem. Biol.* **2017**, *40*, 47−56.

(2) Christianson, D. W. Unearthing the roots of the terpenome. *Curr. Opin. Chem. Biol.* **2008**, *12*, 141−150.

(3) Tippmann, S.; Chen, Y.; Siewers, V.; Nielsen, J. From flavors and pharmaceuticals to advanced biofuels: production of isoprenoids in Saccharomyces cerevisiae. *Biotechnol. J.* **2013**, *8*, 1435−1444.

(4) Phan-Thi, H.; Waché, Y. Behind the myth of the fruit of heaven, a critical review on gac (Momordica cochinchinensis Spreng.) contribution to nutrition. *Curr. Med. Chem.* **2019**, *26*, 4585−4605.

(5) Matulja, D.; Wittine, K.; Malatesti, N.; Laclef, S.; Turks, M.; Markovic, M. K.; Ambrožić, G.; Marković, D. Marine natural products with high anticancer activities. *Curr. Med. Chem.* **2020**, *27*, 1243−1307.

(6) Phulara, S. C.; Pandey, S.; Jha, A.; Chauhan, P. S.; Gupta, P.; Shukla, V. Hemiterpene compound, 3,3-dimethylallyl alcohol promotes longevity and neuroprotection in *Caenorhabditis elegans*. *GeroScience* **2021**, *43*, 791−807.

(7) Zhou, K.; Zou, R.; Zhang, C.; Stephanopoulos, G.; Too, H. P. Optimization of amorphadiene synthesis in bacillus subtilis via transcriptional, translational, and media modulation. *Biotechnol. Bioeng.* **2013**, *110*, 2556−2561.

(8) Formighieri, C.; Melis, A. Carbon partitioning to the terpenoid biosynthetic pathway enables heterologous β-phellandrene production in Escherichia coli cultures. *Arch. Microbiol.* **2014**, *196*, 853−861.

(9) McLennan, A. G. The Nudix hydrolase superfamily. *Cell. Mol. Life Sci.* **2006**, *63*, 123−143.

(10) Mildvan, A. S.; Xia, Z.; Azurmendi, H. F.; Saraswat, V.; Legler, P. M.; Massiah, M. A.; Gabelli, S. B.; Bianchet, M. A.; Kang, L. W.; Amzel, L. M. Structures and mechanisms of Nudix hydrolases. *Arch. Biochem. Biophys.* **2005**, *433*, 129−143.

(11) Kang, A.; George, K. W.; Wang, G.; Baidoo, E.; Keasling, J. D.; Lee, T. S. Isopentenyl diphosphate (IPP)-bypass mevalonate pathways for isopentenol production. *Metab. Eng.* **2016**, *34*, 25−35.

(12) Withers, S. T.; Gottlieb, S. S.; Lieu, B.; Newman, J. D.; Keasling, J. D. Identification of isopentenol biosynthetic genes from Bacillus subtilis by a screening method based on isoprenoid precursor toxicity. *Appl. Environ. Microbiol.* **2007**, *73*, 6277−6283.

(13) Liu, H.; Wang, Y.; Tang, Q.; Kong, W.; Chung, W. J.; Lu, T. MEP pathway-mediated isopentenol production in metabolically engineered Escherichia coli. *Microb. Cell Fact.* **2014**, *13*, 135.

(14) Sivy, T. L.; Fall, R.; Rosenstiel, T. N. Evidence of isoprenoid precursor toxicity in Bacillus subtilis. *Biosci., Biotechnol., Biochem.* **2011**, *75*, 2376−2383.

(15) Lange, B. M.; Wildung, M. R.; Mccaskill, D.; Croteau, R. A family of transketolases that directs isoprenoid biosynthesis via a mevalonate-independent pathway. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 2100−2104.

(16) Estevez, J. M.; Cantero, A.; Reindl, A.; Reichler, S.; Leon, P. 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastIDIc isoprenoid biosynthesis in plants. *J. Biol. Chem.* **2001**, *276*, 22901−22909.

(17) Kuzuyama, T.; Takagi, M.; Takahashi, S.; Seto, H. Cloning and characterization of 1-deoxy-D-xylulose 5- phosphate synthase from Streptomyces sp. Strain CL190, which uses both the mevalonate and nonmevalonate pathways for isopentenyl diphosphate biosynthesis. *J. Bacteriol.* **2000**, *182*, 891−897.

(18) Kuzuyama, T.; Takahashi, S.; Watanabe, H.; Seto, H. Direct formation of 2-C-methyl-D-erythritol 4-phosphate from 1-deoxy-D-xylulose 5-phosphate by 1-deoxyD-xylulose 5-phosphate reductoisomerase, a new enzyme in the non-mevalonate pathway to isopentenyl diphosphate. *Tetrahedron Lett.* **1998**, *39*, 4509−4512.

(19) Rohmer, M. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat. Prod. Rep.* **1999**, *16*, 565−574.

(20) Schwender, J.; Seemann, M.; Lichtenthaler, H. K.; Rohmer, M. Biosynthesis of isoprenoids (carotenoids, sterols, prenyl side-chains of chlorophylls and plastoquinone) via a novel pyruvate/glyceraldehyde 3-phosphate non-mevalonate pathway in the green alga Scenedesmus obliquus. *Biochem. J.* **1996**, *316*, 73−80.

(21) Bach, T. J.; Lichtenthaler, H. K. Inhibition by mevinolin of plant growth, sterol formation and pigment accumulation. *Physiol. Plant.* **2010**, *59*, 50−60.

(22) Leonard, E.; Ajikumar, P. K.; Thayer, K.; Xiao, W. H.; Mo, J. D.; Tidor, B.; Stephanopoulos, G.; Prather, K. L. Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13654−13659.

(23) Banerjee, A.; Preiser, A. L.; Sharkey, T. D. Engineering of recombinant poplar deoxy-D-xylulose-5-phosphate synthase (PtDXS) by site-directed mutagenesis improves its activity. *PLoS One* **2016**, *11*, No. e0161534.

(24) You, S.; Yin, Q.; Zhang, J.; Zhang, C.; Qi, W.; Gao, L.; Tao, Z.; Su, R.; He, Z. Utilization of biodiesel by-product as substrate for high-production of β-farnesene via relatively balanced mevalonate pathway in Escherichia coli. *Bioresour. Technol.* **2017**, *243*, 228−236.

(25) Hahn, F. M.; Hurlburt, A. P.; Poulter, C. D. Escherichia coli open reading frame 696 is IDI, a nonessential gene encoding isopentenyl diphosphate isomerase. *J. Bacteriol.* **1999**, *181*, 4499−4504.

(26) Yoon, S. H.; Kim, J. E.; Lee, S. H.; Park, H. M.; Choi, M. S.; Kim, J. Y.; Lee, S. H.; Shin, Y. C.; Keasling, J. D.; Kim, S. W. Engineering the lycopene synthetic pathway in E. coli by comparison of the carotenoid genes of Pantoea agglomerans and Pantoea ananatis. *Appl. Microbiol. Biotechnol.* **2007**, *74*, 131−139.

(27) Sun, Z.; Cunningham, F. X., Jr.; Gantt, E. Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11482−11488.

(28) Li, Z.; Ji, J.; Wang, G.; Josine, T. L.; Wu, J.; Diao, J.; Wu, W.; Guan, C. Cloning and heterologous expression of isopentenyl diphosphate isomerase gene from Lycium chinense. *J. Plant Biochem. Biotechnol.* **2016**, *25*, 40−48.

(29) Phulara, S. C.; Chaturvedi, P.; Chaurasia, D.; Diwan, B.; Gupta, P. Modulation of culture medium confers high-specificity production of isopentenol in *Bacillus subtilis*. *J. Biosci. Bioeng.* **2019**, *127*, 458−464.

(30) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026−1028.

(31) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202−W208.

(32) Lipman, D. J.; Souvorov, A.; Koonin, E. V.; Panchenko, A. R.; Tatusova, T. A. The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2002**, *2*, 20.

(33) Runthala, A. Probabilistic divergence of a template-based modelling methodology from the ideal protocol. *J. Mol. Model.* **2021**, *27*, 25.
(34) Runthala, A.; Chowdhury, S. Refined template selection and combination algorithm significantly improves template-based modeling accuracy. *J. Bioinf. Comput. Biol.* **2019**, *17*, 1950006.
(35) Lee, G. R.; Won, J.; Heo, L.; Seok, C. GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure. *Nucleic Acids Res.* **2019**, *47*, W451−W455.
(36) Benkert, P.; Biasini, M.; Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **2011**, *27*, 343−350.
(37) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(38) Zimmermann, L.; Stephens, A.; Nam, S. Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A. N.; Alva, V. Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **2018**, *430*, 2237−2243.
(39) Trifinopoulos, J.; Nguyen, L. T.; von Haeseler, A.; Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **2016**, *44*, W232−W235.
(40) Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256−W259.
(41) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **2016**, *44*, W344−W350.
(42) Gouet, P.; Robert, X.; Courcelle, E. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* **2003**, *31*, 3320−3323.
(43) Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* **2018**, *46*, W363−W367.
(44) Santos, K. B.; Guedes, I. A.; Karl, A. L. M.; Dardenne, L. E. Highly Flexible Ligand Docking: Benchmarking of the DockThor Program on the LEADS-PEP Protein-Peptide Data Set. *J. Chem. Inf. Model.* **2020**, *60*, 667−683.
(45) Hu, L. X.; Feng, J. J.; Wu, J.; Li, W.; Gningue, S. M.; Yang, Z. M.; Wang, Z.; Liu, Y.; Xue, Z. L. Identification of six important amino acid residues of MenA from Bacillus subtilis natto for enzyme activity and formation of menaquinone. *Enzyme Microb. Technol.* **2020**, *138*, No. 109583.
(46) Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.* **2018**, *46*, W356−W362.
(47) Pucci, F.; Bernaert, K.; Teheux, F.; Gilis, D.; Rooman, M. Symmetry principles in optimization problems: an application to protein stability prediction. *IFAC-PapersOnLine* **2015**, *48*, 458−463.
(48) Pucci, F.; Bernaerts, K.; Kwasigroch, J. M.; Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **2018**, *34*, 3659−3665.
(49) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **2014**, *3*, No. e02030.
(50) Rodrigues, C. H. M.; Pires, D. E. V.; Ascher, D. B. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* **2021**, *30*, 60−69.
(51) Kumar, M. D.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206.
(52) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; Marcatili, P. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Struct., Funct., Bioinf.* **2019**, *87*, 520−527.
(53) Kuriata, A.; Gierut, A. M.; Oleniecki, T.; Ciemny, M. P.; Kolinski, A.; Kurcinski, M.; Kmiecik, S. CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Res.* **2018**, *46*, W338−W343.
(54) Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr., D: Biol Crystallogr* **2004**, *D60*, 1355−1363.
(55) WebGRO for Macromolecular Simulations. Available from: https://simlab.uams.edu.
(56) Lindahl, E.; Hess, B.; Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306−317.
(57) Bessman, M. J.; Frick, D. N.; O'Handley, S. F. The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J. Biol. Chem.* **1996**, *271*, 25059−25062.
(58) Gabelli, S. B.; Bianchet, M. A.; Xu, W.; Dunn, C. A.; Niu, Z. D.; Amzel, L. M.; Bessman, M. J. Structure and function of the E. coli dihydroneopterin triphosphate pyrophosphatase: a Nudix enzyme involved in folate biosynthesis. *Structure* **2007**, *15*, 1014−1022.
(59) Srouji, J. L.; Xu, A.; Park, A.; Kirsch, J. F.; Brenner, S. E. The evolution of function within the Nudix homology clan. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 775−811.
(60) Gabelli, S. B.; Bianchet, M. A.; Ohnishi, Y.; Ichikawa, Y.; Bessman, M. J.; Amzel, L. M. Mechanism of the Escherichia coli ADP-ribose pyrophosphatase, a Nudix hydrolase. *Biochemistry* **2002**, *41*, 9279−9285.
(61) Guedes, I. A.; Barreto, A. M. S.; Marinho, D.; Krempser, E.; Kuenemann, M. A.; Sperandio, O.; Dardenne, L. E.; Miteva, M. A. New machine learning and physics-based scoring functions for drug discovery. *Sci. Rep.* **2021**, *11*, 3198.
(62) Li, Y.; Reilly, P. J.; Ford, C. Effect of introducing proline residues on the stability of Aspergillus awamori. *Protein Eng., Des. Sel.* **1997**, *10*, 1199−1204.
(63) Trevino, S. R.; Schaefer, S.; Scholtz, J. M.; Pace, C. N. Increasing protein conformational stability by optimizing beta-turn sequence. *J. Mol. Biol.* **2007**, *373*, 211−218.
(64) Vishvakarma, V. K.; Pal, S.; Singh, P.; Bahadur, I. Interactions between main protease of SARS-CoV-2 and testosterone or progesterone using computational approach. *J. Mol. Struct.* **2021**, *1251*, No. 131965.
(65) Tumskiy, R. S.; Tumskaia, A. V. Multistep rational molecular design and combined docking for discovery of novel classes of inhibitors of SARS-CoV-2 main protease 3CLpro. *Chem. Phys. Lett.* **2021**, *780*, No. 138894.
(66) Leach, A. R. Ligand-Based Approaches: Core Molecular Modeling. In *Compr. Med. Chem. II*; Taylor, J. B., Triggle, D. J., Eds.; Elsevier Press: Netherlands, 2007; pp 87−118.
(67) Lobanov, M. Y.; Bogatyreva, N. S.; Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **2008**, *42*, 623−628.
(68) Ghosh, S. K.; Saha, B.; Banerjee, R. Insight into the sequence-structure relationship of TLR cytoplasm's Toll/Interleukin-1 receptor domain towards understanding the conserved functionality of TLR 2 heterodimer in mammals. *J. Biomol. Struct. Dyn.* **2020**, *39*, 5348−5357.
(69) Kumar, B.; Parasuraman, P.; Murthy, T. P. K.; Murahari, M.; Chandramohan, V. In silico screening of therapeutic potentials from Strychnos nux-vomica against the dimeric main protease (Mpro) structure of SARS-CoV-2. *J. Biomol. Struct. Dyn.* **2021**, 1−19.
(70) Bloom, J. D.; Arnold, F. H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 9995−10000.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912

(71) Zada, B.; Wang, C.; Park, J. B.; Jeong, S.; Park, J. E.; Singh, H. B.; Kim, S. W. Metabolic engineering of Escherichia coli for production of mixed isoprenoid alcohols and their derivatives. *Biotechnol. Biofuels* **2018**, *11*, 210.

https://doi.org/10.1021/acsomega.2c01677
ACS Omega 2022, 7, 19898−19912