ACS OMEGA

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

Article

Unsupervised Machine Learning Organization of the Functional Dark Proteome of Gram-Negative "Superbugs": Six Protein Clusters Amenable for Distinct Scientific Applications

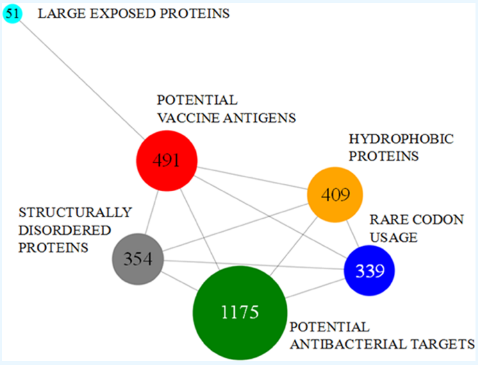
Carlos Sicilia,[⊥] Andrés Corral-Lugo,[⊥] Pawel Smialowski, Michael J. McConnell, and Antonio J. Martín-Galiano*

Cite This: *ACS Omega* 2022, 7, 46131−46145

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Uncharacterized proteins have been underutilized as targets for the development of novel therapeutics for difficult-to-treat bacterial infections. To facilitate the exploration of these proteins, 2819 predicted, uncharacterized proteins (19.1% of the total) from reference strains of multidrug *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* species were organized using an unsupervised *k-means* machine learning algorithm. Classification using normalized values for protein length, pI, hydrophobicity, degree of conservation, structural disorder, and %AT of the coding gene rendered six natural clusters. Cluster proteins showed different trends regarding operon membership, expression, presence of unknown function domains, and interactomic relevance. Clusters 2, 4, and 5 were enriched with highly disordered proteins, nonworkable membrane proteins, and likely spurious proteins, respectively. Clusters 1, 3, and 6 showed closer distances to known antigens, antibiotic targets, and virulence factors. Up to 21.8% of proteins in these clusters were structurally covered by modeling, which allowed assessment of druggability and discontinuous B-cell epitopes. Five proteins (4 in Cluster 1) were potential druggable targets for antibiotherapy. Eighteen proteins (11 in Cluster 6) were strong B-cell and T-cell immunogen candidates for vaccine development. Conclusively, we provide a feature-based schema to fractionate the functional dark proteome of critical pathogens for fundamental and biomedical purposes.


51
LARGE EXPOSED PROTEINS
POTENTIAL VACCINE ANTIGENS
491
HYDROPHOBIC PROTEINS
409
STRUCTURALLY DISORDERED PROTEINS
RARE CODON USAGE
354
339
1175
POTENTIAL ANTIBACTERIAL TARGETS

INTRODUCTION

The lack of functional knowledge of proteins limits their utilization as targets for the development of novel therapeutics for multidrug-resistant pathogens. Uncharacterized and poorly characterized proteins represent a sizable fraction of the bacterial proteome, up to 20−50%, even in strains of model bacterial species.[1,2] The functions of these uncharacterized or "hypothetical" proteins lack experimental characterization, and they cannot be inferred with sufficient certainty with sequence homology-based tools. Thus, they are considered the functional dark proteome.[3,4]

Extensive antibiotic resistance represents a major medical challenge for humanity.[5] The *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* species represent the highest, i.e., a critical concern for the WHO and other public organizations.[6,7] Clinical isolates of these Gram-negative "superbugs" are potentially untreatable with the limited antibiotic panel that is currently available, leading to panresistance and a potential return to the pre-antibiotic era.[8] The existence and dissemination of these pathogens affect multiple aspects of modern medicine such as organ transplantation, anticancer treatment, autoimmune diseases therapy, neonatal and geriatric specialties, ICU stays, and virtually any kind of surgery. Given the slow pace of development of new antibiotics against resistant pathogens,[9] other options, such as the identification of novel antibiotic targets and prophylaxis through vaccination,[10] are currently being explored. In this context, proteins are major players as chemotherapeutic targets, protective antigens, and diagnostic markers that could improve epidemiological tracking. Eventually, any fundamental knowledge regarding persistence and virulence can potentially be utilized for clinical benefit against these problematic species.

Many proteins of these concerning pathogens remain functionally unexplored. Subsets of them have been detected as expressed by proteomics,[11,12] identified as essential during pathogenic processes,[13−15] or deemed conserved enough to be included in the core proteome of the species.[16] However, information regarding the function of protein targets is an important prerequisite for developing biomedical applications.

Received: June 29, 2022
Accepted: October 6, 2022
Published: December 6, 2022

For instance, a description of the mode of action is required for the efficient development of antibiotic compounds.[17] The burden of assigning potential functions to "uncharacterized proteins" often results in their exclusion from studies aiming to identify potential therapeutic targets.

While massive DNA sequencing has become extremely cost-effective over the last 15 years,[18] functional characterization of predicted proteins encoded in bacterial genomes remains a laborious and nonscalable task. Consequently, the relative amount of protein domains with unknown functions (DUFs)[19] in the Pfam database increases as more sequences are available.[20] This increasing gap between sequence availability and functional understanding of proteins is a remaining challenge of the postgenomic era and a huge problem for the biological sciences.[21] To respond to this demand, hypothetical proteins are approached by exclusive resources such as COMBREX-DB[22] and the Dark Proteome Database.[23]

The umbrella term "hypothetical proteins" (HPs) used to define these proteins overlooks their heterogeneity. The reason for the slow pace in the characterization of novel protein functions is manifold. These include, but are not limited to: (a) recalcitrance to experimental work, *e.g.*, low expression using standard recombinant DNA technology or insolubility in common molecular biology buffers; (b) extreme structural disorder resulting in unsatisfactory structural resolution or failed structure-based function prediction; (c) nonessentiality or overlapping roles with other proteins that render inconclusive experiments using knockout mutants; (d) lack of proper microbiological and/or biochemical assays to study nonstandard functions; (e) phylogenetic narrowness (even strain-specific) that do not attract the attention of the scientific community; and (f) prediction of spurious ORFs or relic pseudogenes by gene identification algorithms, typically associated with deviant codon usage and AT content. Altogether, these considerations result in the systematic exclusion of uncharacterized proteins in studies searching for biomedical targets, which prevents the potential discovery of new functions and applications.[24]

Since uncharacterized proteins cannot be studied in full with common homology-based tools, e.g., BLAST or Pfam, alternative approaches for analysis are desirable. Several properties of proteins are linked to their activities irrespective of the strict amino acid sequence. These aspects may be utilized in unsupervised machine learning, which permits the identification of natural specimen clusters that are otherwise incomplete using only human expertise.[25] Unsupervised learning has been applied in protein sciences, for example, as a method for facilitating structural classification.[26,27]

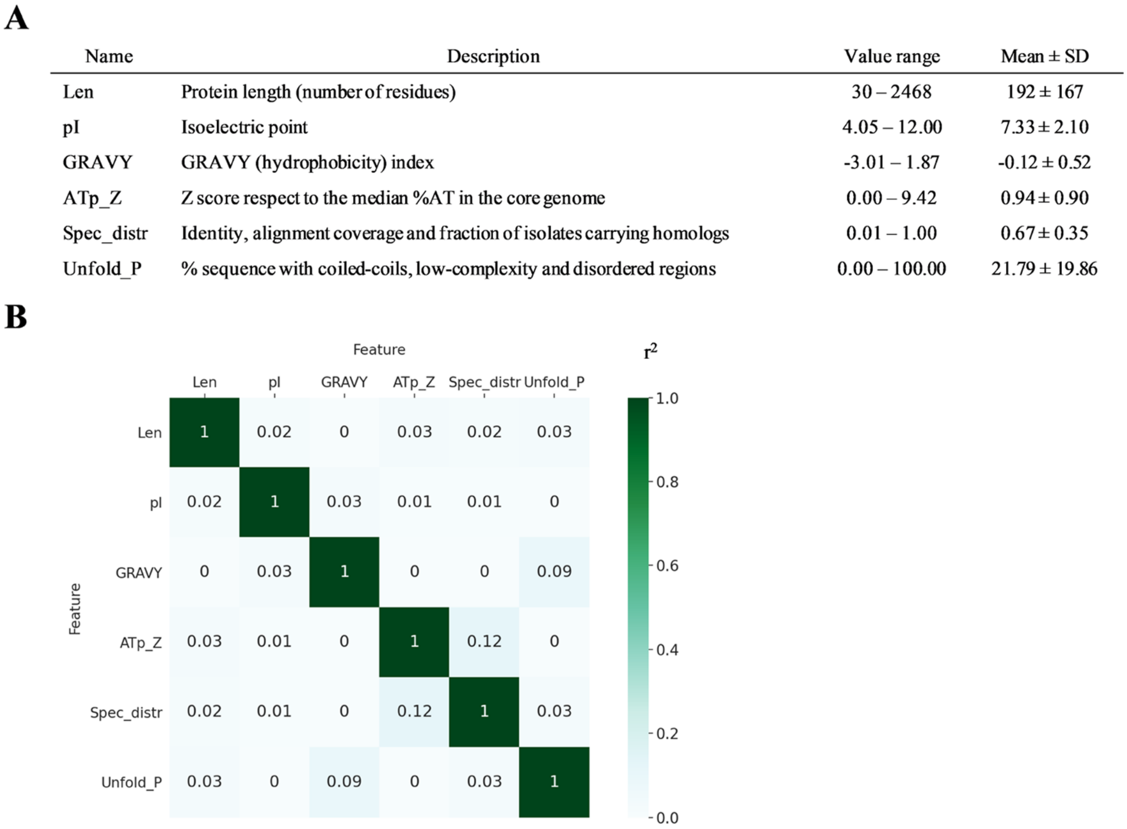
Whether the functional dark proteome from multidrug-resistant species can be utilized in therapeutic objectives remains an open question. In this study, we have taken advantage of the strength resulting from the integration of protein science principles, omic information, and unsupervised classificatory algorithms to identify protein classes within this obscure set of proteins. We achieve the prefractionation of the uncharacterized protein space into manageable groups of proteins. Some of these groups are compatible with biomedical applications, warranting further investigation that explores their use in the control of these challenging infections.

■ EXPERIMENTAL SECTION

Hypothetical Protein Acquisition. Chromosomally-encoded proteomes for the reference strains *A. baumannii* ATCC 19606 (UP000005740), *K. pneumoniae* HS11286 (UP000007841), and *P. aeruginosa* PAO1 (UP000002438) were downloaded (last accession date: 11/May/2021) from the Proteomes RefSeq database section of the UniProt resource.[28] Proteins explicitly described as uncharacterized proteins, those with UPF (unknown protein family), and hypothetical, all without available "biological role" by Gene Ontology[29] were deemed uncharacterized. However, all of those containing known domains and those with "probable", "putative" or uncharacterized general functions were likely remote homologues of known families or superfamilies and were not considered. Fragmented proteins were not considered either. Coding gene sequences and gene feature information were downloaded from the Assembly database.[30] Protein redundancy of selected uncharacterized proteins was assessed by sequence-based clustering by CD-HIT 4.6[31] applying 90% identity and 90% alignment length coverage.

Calculation of Protein Clustering Features. Classificatory properties were calculated as follows. The length, isoelectric point (pI), and GRAVY index were calculated with EXPASY protparam tool.[32] Coiled coils were predicted by ncoils[33] using 0.5 as threshold, low complexity regions by SEG,[34] and disordered sections longer than 10 residues by IUpred.[35] The intra-species conservation index (ranging between 0 and 100) was calculated as the prevalence ratio multiplied by the average identity and alignment coverage of the hits. The prevalence of each protein across the species was calculated as the percentage of isolates that carry protein homologues with identity ≥80% and alignment coverage ≥90%, calculated with BLASTp 2.2.28+.[36] For that, proteomes of a random sample of 200 isolates of the respective species, nonanomalous according to NCBI quality filters from the Assembly database[30] were utilized (Supporting Table S1). The *Z* score for AT content for each gene coding for hypothetical proteins was the number of standard deviation units, higher or lower, with respect to the average AT content of all coding sequences in the genome: 60.3 ± 4.2% for *A. baumannii*, 43.3 ± 6.8% for *K. pneumoniae*, and 33.2 ± 3.9% for *P. aeruginosa*.

Unsupervised Classification. Data profiles containing values for the six classificatory properties were nonparametrically normalized with the *RobustScaler* method available in the *preprocessing* module of the *scikit-learn* machine learning suite.[37] K-means clusters were calculated, applying *k* values between 2 and 20, with the *lloyd* algorithm of the *K-Means* method of the *cluster* scikit-learn module.[37] Clustering was run by applying default hyperparameters: 10 different centroid seeds, up to 300 iterations, and a tolerance of 10^{-4}. The optimal number of clusters, that is, the number of clusters with the best trade-off between maximal inclusivity (highest size) and homogeneity, were estimated by calculating the inter- vs intra-clustering difference. For the latter, the 6-mer Euclidean distances between normalized feature values of all members in the same cluster and between cluster centroids were calculated using the *pairwise* method of the *DistanceMetric* class of *Neighbors* scikit-learn module.[37] Hyperparameter tuning by grid search was carried out by varying the number of seeds (from 5 to 20, step = 5), the number of maximal iterations (from 100 to 500, step = 100), the tolerance to cluster center differences in two consecutive iterations (10^{-3}, 10^{-4} or 10^{-5}), and the algorithm applied (either *lloyd* or *elkan*). Agglomerative hierarchical clustering was carried out with the *Agglomerative Clustering* method of the *cluster* scikit-learn module using the "ward" linkage criterion and applying six



A

Name	Description	Value range	Mean ± SD
Len	Protein length (number of residues)	30 – 2468	192 ± 167
pI	Isoelectric point	4.05 – 12.00	7.33 ± 2.10
GRAVY	GRAVY (hydrophobicity) index	-3.01 – 1.87	-0.12 ± 0.52
ATp_Z	Z score respect to the median %AT in the core genome	0.00 – 9.42	0.94 ± 0.90
Spec_distr	Identity, alignment coverage and fraction of isolates carrying homologs	0.01 – 1.00	0.67 ± 0.35
Unfold_P	% sequence with coiled-coils, low-complexity and disordered regions	0.00 – 100.00	21.79 ± 19.86

B
Feature
Len pI GRAVY ATp_Z Spec_distr Unfold_P
Feature
Len 1 0.02 0 0.03 0.02 0.03
pI 0.02 1 0.03 0.01 0.01 0
GRAVY 0 0.03 1 0 0 0.09
ATp_Z 0.03 0.01 0 1 0.12 0
Spec_distr 0.02 0.01 0 0.12 1 0.03
Unfold_P 0.03 0 0.09 0 0.03 1
r^2
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1. Classificatory protein features utilized in this study. (A) List of the six classificatory features considered in this work. (B). Heatmap showing color-ranked r-squared values for the inter-feature value correlation matrix.

clusters. DBSCAN clustering was carried out with the *DBSCAN* method of the *cluster* scikit-learn module, applying 0.472 as the maximal distance to consider the inter-sample neighborhood and a minimal neighborhood sample number of 10 for core points to render six final clusters.

Calculation of Further Postclassificatory Protein Properties. Correspondence analysis of codon usage of genes coding for hypothetical proteins was carried out with CodonW (http://codonw.sourceforge.net). The percentage of rare codons was defined as the percentage of codons showing relative isocodon prevalence <20% with respect to the most prevalent isocodon for the amino acid in the whole proteome of the species. Operon information was calculated with the Operon-mapper resource.[38] DUF domains were identified by the *hmmscan* program from the HMMER 3.1b1 package[39] against the Pfam v33.1 database,[40] applying their respective gathering thresholds. The number of partners for protein–protein interactions (PPIs) were calculated using STRING v11,[41] applying a combined score threshold of ≥0.7. For reference strains that are not available in STRING, protein equivalences between reference and STRING strains were assigned with BLASTp 2.2.28+, applying *E*-value <0.00001, identity ≥70%, and bidirectional alignment coverage ≥90% thresholds. Transmembrane helices were identified with TMHMM 2.0[42] and considered when they did not overlap with the signal peptide by five or more residues predicted with SignalP 5.0 trained with a Gram-negative data set.[43] Protein solubility was assessed by Proso II.[44] Structural coverage was calculated by applying BLAST with identity ≥25% and bidirectional alignment coverage ≥60% thresholds to protein sequences in the Protein Data Bank (PDB) (last accession: 12/May/2021). Structural homology models required for the prediction of target druggability and discontinuous B-cell epitopes were built with SWISS-MODEL.[45] Essential proteins were downloaded from the DEG 15 database[46] involving conditions such as lung infection,[47] rich medium,[48] and succinate medium.[49] Homologues for essential proteins from strains distinct from reference ones were identified by BLAST, applying 80% identity and 90% mutual alignment length coverage. Betweenness centrality was calculated as the shortest-path betweenness centrality for STRING nodes with the *betweenneess_centrality* method of NetworkX.[50] Pocket detection and their druggability score in HPs was carried out with DogSiteScorer[51] using structures for sequences showing ≥89% identity but PA1233 (37% identity, 53% similarity to PDB id 4f98 sequence). For antibody accessibility to potential antigens, the subcellular location was determined with PSORTb 3.0.[52] Linear B-cell epitopes were predicted with BepiPred 2.0.[53] Discontinuous B-cell epitopes were predicted with Discotope 2.0[54] and ElliPro,[55] applying a score ≥0.5 and a maximal distance <6 Å. HLA class II supermotif cores were predicted with netMHCIIpan 4.0,[56] selecting the five supertype alleles[57] and applying an adjusted rank ≤2.0 and an epitope length of 15 residues.

Control Proteins of Biomedical Interest. Antibacterial agent targets from γ-Proteobacteria, the most specific taxon shared by the three species addressed here, were downloaded from DrugBank.[58] Homologues for antibacterial target sequences were detected in *A. baumannii*, *K. pneumoniae*, and *P. aeruginosa* reference proteomes by BLAST, applying ≥40% identity, ≥80% mutual alignment, and *E*-value <0.001 thresholds. Antigens of these species were selected from a

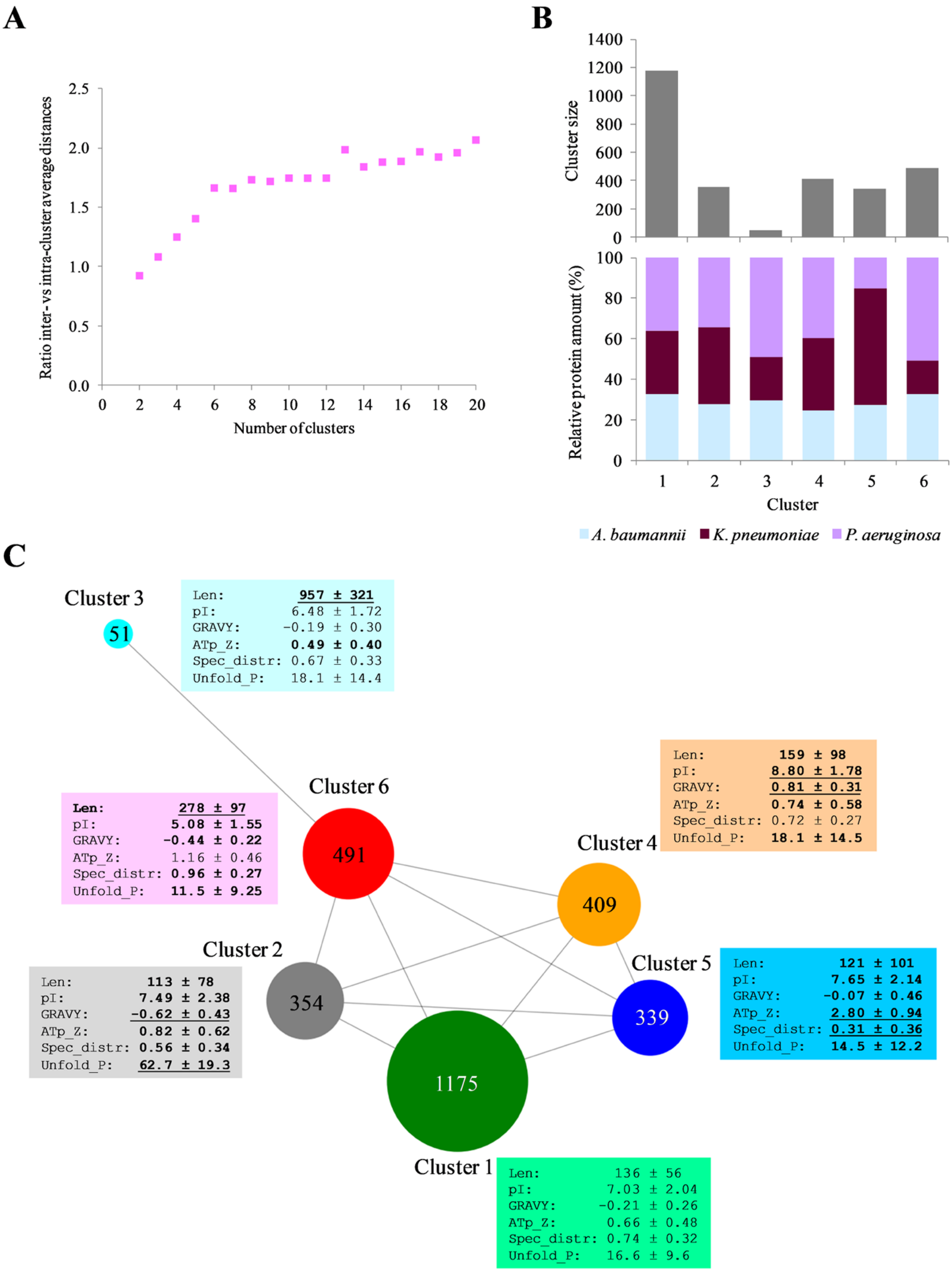


Figure 2. Unsupervised classification of the uncharacterized proteome in Gram-negative superbugs. (A) Ratio between average inter- and intra-cluster Euclidian distances according to the number of clusters. (B) Cluster sizes (upper chart) and relative amount of cluster members by species (lower chart) for a $k = 6$. (C) Undirected weighted graph showing clusters at $k = 6$. Edges correspond to normalized distances between cluster centroids below a cutoff of 4. Average and SD values for non-normalized features for each cluster are indicated. Features showing significant differences to Cluster 1 are highlighted in bold and underlined when dominant (P-value $<10^{-30}$). The cluster size is indicated within the node and made proportional to the sphere diameter. The graph was created using the *NetworkX* and *matplotlib* Python libraries.

comprehensive review.[59] Virulence factor sequences were downloaded from the Virulence Factor Database[60] (last accession: 10/Oct/2021) and identified by BLAST in the proteomes of the reference strains, applying ≥40% identity, ≥80% mutual alignment coverage, and E-value <0.001 thresholds. Values for the six clustering features were calculated as for hypothetical proteins above. Data was nonparametrically normalized with *RobustScaler* by adding

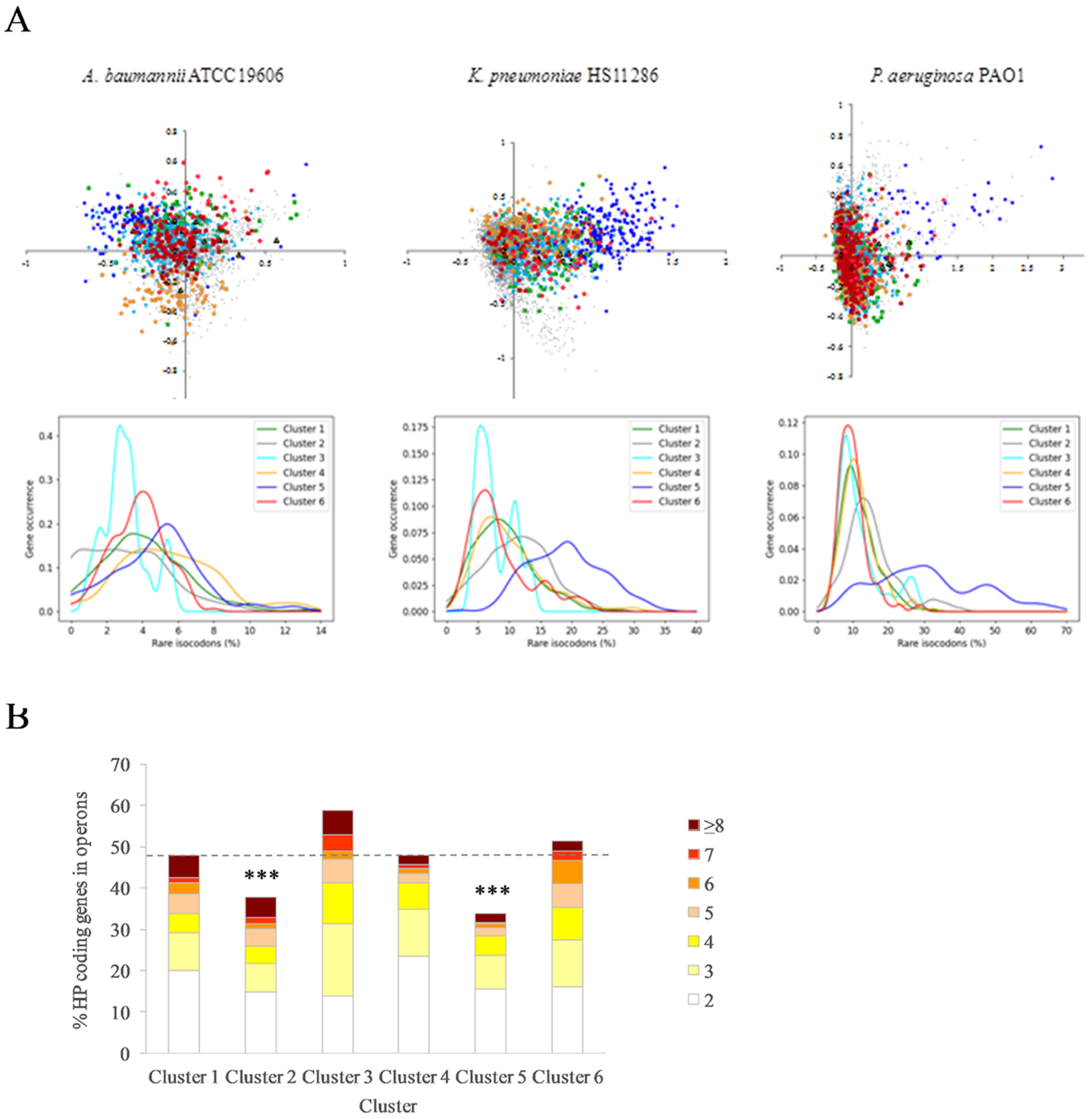


Figure 3. Genetic differences between genes coding for proteins in distinct clusters. (A) Codon usage correspondence analysis (upper panel) and percentage of rare isocodons (lower panel) of uncharacterized protein-encoding genes of the three species. The two principal correspondence analysis axes are shown. Each dot corresponds to gene coding HPs colored according to the cluster. Kernel density estimates for the distribution of rare isocodons was carried out with the *gaussian_kde* method of the *stats* module of *scipy* Python library. (B) Prevalence of operon membership in clusters according to operon size (number of genes). Dashed lines in B correspond to the reference value for Cluster 1. *** $P < 0.001$.

data independently for each antibacterial target, vaccine antigen, or virulence factor to the whole hypothetical protein data set pool. Then, the distance of normalized data of each biomedical control protein to the average of each of the six hypothetical protein clusters was calculated, as described above.

■ RESULTS

Feature-Based Schema to Classify Uncharacterized Proteins. The reference strains of three leading multidrug-resistant species showed comparable numbers of uncharacterized proteins: *A. baumannii* ATCC 19606 (851 HPs, 22.4% of the total proteome), *K. pneumoniae* HS11286 (930 HPs, 17.2%), and *P. aeruginosa* PAO1 (1,038 HPs, 18.6%). The 2,819 protein data set showed very low sequence redundancy (0.3%) at a 90% identity and 90% alignment coverage level.

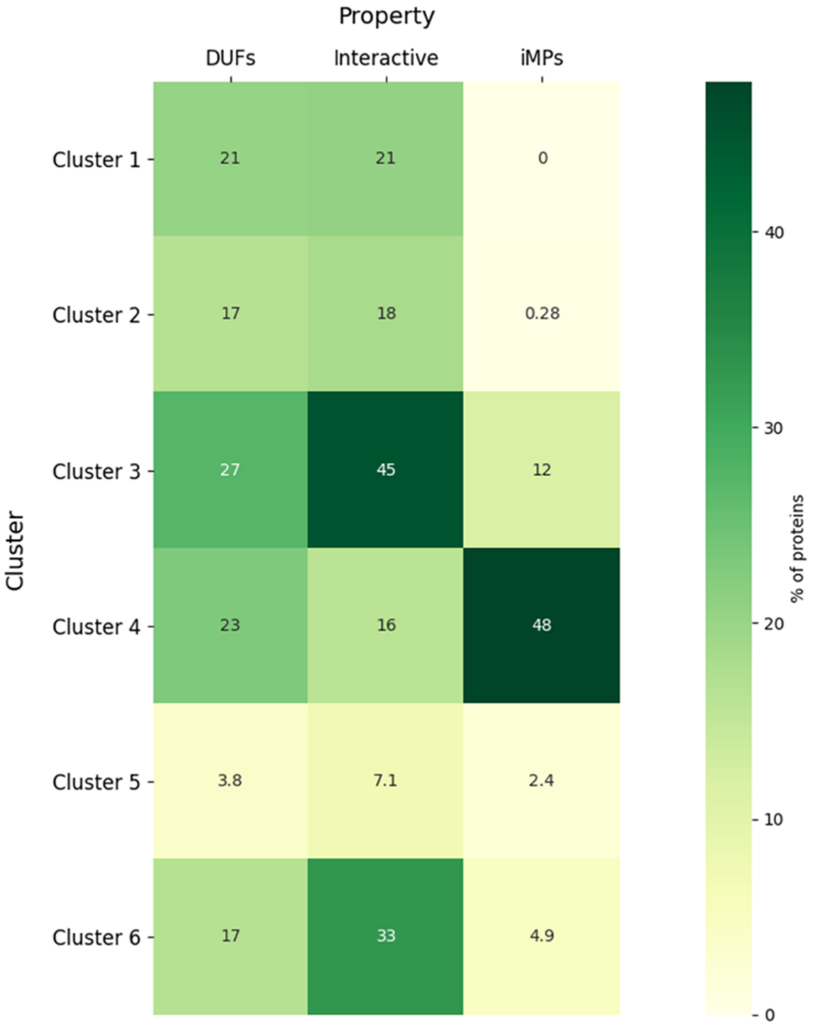


Figure 4. Protein level differences in clusters. Heatmap shows the color-ranked percentage of cluster proteins: containing DUFs, positive for interactome degree (≥5 PPIs) and integral MPs (those containing ≥4 TMHs).

As homology tools have limitations to classify HPs, an alternative feature-based classificatory schema was developed. For that, three global protein properties were utilized: length (Len), pI, and hydrophobicity (GRAVY), besides three additional properties usually related to difficulties in solving protein function: anomalous AT content (ATp_Z), low species isolate occurrence (Spec_distr), and high sequence percentage predicted as structurally disordered (Unfold_P)-(Figure 1A). Notably, these six properties involved values that were continuous, mostly nonzero (nonsparse), and non-correlated ($r^2 \leq 0.12$ in all pair cases) (Figure 1B). Moreover, these are inclusive of other features not explicitly considered for clustering. For instance, length would correlate with molecular weight and the number of domains, pI with an abundance of certain charged residues, and the GRAVY index with the amount of aliphatic and aromatic residues. This 6-mer schema is therefore appropriate for unsupervised machine learning approaches.[61]

Unsupervised Learning Renders Six Optimal Clusters. To minimize the influence of outliers and the different numeric scales of the six properties considered, values were nonparametrically normalized. Next, 6-mer arrays of normalized values for all properties corresponding to the 2,819 HPs were subjected to unsupervised classification by the *k-means* method. Incremental single-linkage clustering was applied for a number (*k* parameter) between 2 and 20 clusters. The difference between the average of inter-cluster with respect to the average of intra-cluster distances increases quickly until a *k* value of six clusters (Figure 2A). After this point, the productivity of every new cluster was minor and clustering likely resulted in over-partitioning of naturally compact clusters. A *k* value of six was therefore considered hereafter due to the production of clusters with optimized inclusivity and homogeneity. Hyperparameter tuning (see Materials and Methods) did not essentially change cluster content, *i.e.*, protein pair coclustering was >99.2%, indicating cluster convergence was reached after a few iterations.

These clusters largely differed in size, where the extremes were Cluster 1 (arbitrary numbering) with 1175 members and Cluster 3 with only 51 members (Figure 2B, upper chart). Clusters showed comparable relative proportions between the three species, although *K. pneumoniae* HPs was relatively enriched in Cluster 5 proteins and *P. aeruginosa* proteins in Cluster 3 and 6 (Figure 2B, lower chart). Clusters showed significantly different values (two-tailed *P*-value by Student's *t* test) for at least two properties with respect to Cluster 1 (selected here as the reference one by its largest size) (Figure 2C). Detailed cluster information containing proteins and

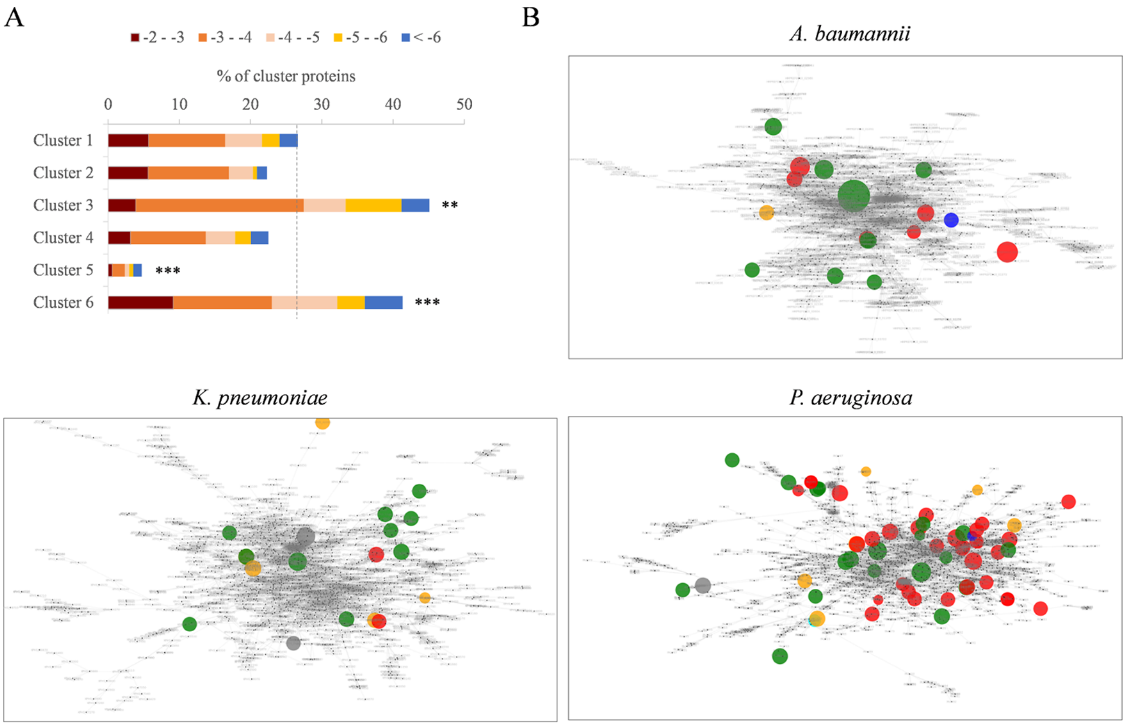


Figure 5. Betweenness centrality in the interactome. (A) Amount of proteins per cluster at distinct BC ranges. (B) Species interactomes. Network graphs of the respective species are shown. Cluster membership is indicated in colors, as shown in Figure 2D. Sphere diameter is proportional to BC.

values for classificatory properties is available in Supporting Table S2.

Besides *k-means*, two other clustering methods are usually utilized in unsupervised learning: hierarchical clustering (HC) and DBSCAN.[61] HC grouping at the level of six clades showed clear cluster correspondences with *k-means*. The exceptions were *k-means* Cluster 1 and 6, which became intermingled after HC (Supporting Figure S1). Overall, both methods showed nearly identical average intra-cluster distances. However, HC showed remarkably lower inter-cluster distances, indicating lower cluster discriminative power. This effect was more exacerbated for DBSCAN, which grouped >99% of the data set into two large heterogeneous clusters (Supporting Figure S1).

Clusters Show Distinct Genetic Trends: Codon Usage and Presence in Operons. Insights into the nature of clusters may be revealed through distinctive values for other properties not considered for clustering due to sparsity, probable correlation within clustering properties, or categorical content character (see Supporting Table S3). The genetic nature of genes encoding proteins of distinct clusters was interrogated first.

Codon selection was analyzed, which is related to gene adaptation to the bacterial tRNA pool involved in the elongation stage of translation in fast-growing bacteria.[62] The two principal axes of the correspondence analysis of codon usage indicate compact occupation for Cluster 6 genes while the disparate location of genes coding for proteins in Cluster 4 and, in particular, Cluster 5 (Figure 3A, upper panel). *A. baumannii* showed the most classical and illustrative view consisting of a central point ball (most housekeeping genes) with two horns (highly expressed genes and anomalous codon content genes, respectively).[63] Most Cluster 5 protein coding genes were located in the disperse horn corresponding to aberrant isocodon utilization. Cluster 5 showed explicit shift distributions toward higher proportions of rare isocodons, defined here as those with <20% occurrence with respect to the preferred isocodon for the amino acid in the whole genome (Supporting Table S4) (Figure 3A, lower panel). This aspect may reveal that several theoretical Cluster 5 HPs are coded by nonexpressed ORFs, as described previously for other genes.[64]

On the other hand, operon membership suggests functional relevance since it indicates pressure for cotranscription into higher-order elements (such as enzymatic pathways or protein complexes) besides favoring useful cohorizontal transfer with other genes with dependent functional reasons.[65] While genes coding for 58.8% and 51.5% of proteins in Cluster 3 and 6, respectively, were part of operons, only 37.8% of genes encoding Cluster 2 ($P < 0.001$, Chi-squared test respect to Cluster 1) and 33.9% of Cluster 5 ($P < 0.001$) proteins were nonmonocistronic (Figure 3B).

Conclusively, 139 HP genes contained >5% rare isocodons, <80 residues, showed presence in <5% isolates in the species, were not in operons, and encoded theoretical polypeptides that did not contain DUFs and did not show any PPI. Among them, around half were *K. pneumoniae* HPs in Cluster 5. These may be spurious genes resulting from false positives by gene detection algorithms or part of mobile genetic elements, pseudogenes, or small untranslated RNAs. Most of these proteins showed only scarce BLAST hits in other species in the same genus (data not shown).

Clusters Show Distinct Protein Trends: Presence of DUFs and Interactomic Relevance. Qualities of the proteins themselves were also evaluated (Figure 4). DUFs were found in 17−27% of proteins in all clusters, except for Cluster 5, where only 3.8% of proteins carried these important PFAM entities ($P < 0.001$, Chi-squared test with respect to Cluster 1). Cluster 3 and 6 proteins had a relatively higher percentage of connected proteins within the interactome ($P <$

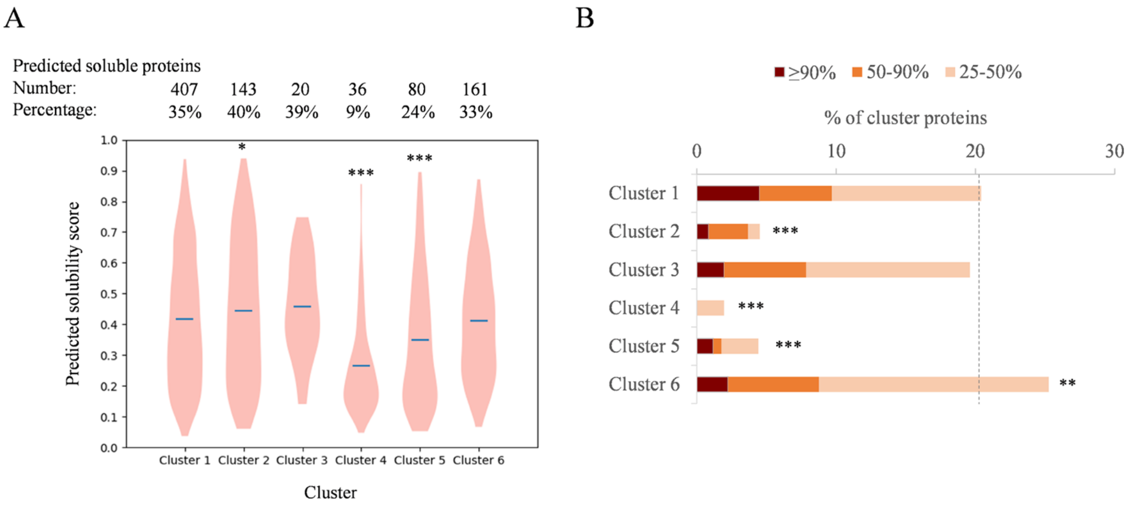


Figure 6. HP workability. (A) predicted solubility. Violin plots showing the distribution of solubility scores calculated with PROSO II are shown. Lines indicate average values. (B) Structural modeling coverage. Identity level bins are color-ranked (see legend). A minimal bidirectional threshold of 60% for alignment length coverage was applied.

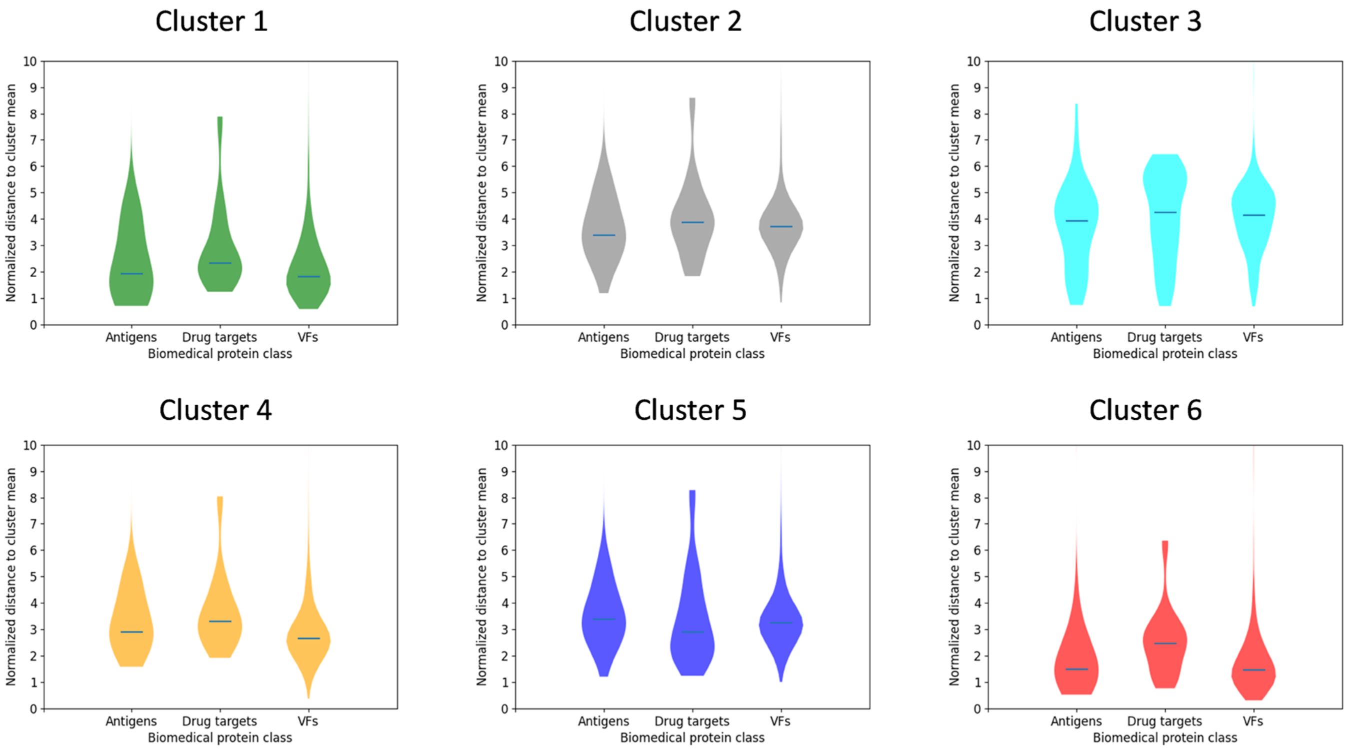


Figure 7. Distance distribution of control biomedical targets to HP cluster centroids. Violin plots of normalized Euclidian distances of validated drug targets, antigens, and VF control data sets with respect to cluster centroids are shown. The lines indicate the medians. Detailed descriptions of biomedical targets and distances to HP clusters are provided in Supporting Table S5.

0.001), whereas only 7.1% of Cluster 5 proteins showed 5 or more partners ($P < 0.001$). Expectedly, the fraction of integral membrane proteins (iMPs, ≥4 TMHs) was much higher for the highly hydrophobic Cluster 4.

Rather than the raw number of partners, the actual relevance of a protein in the interactome is dictated by the degree of irreplaceability of the node within the network context. In this way, important proteins for biomedical applications and bacterial physiology are central.[66] This issue was assessed through the calculation of the betweenness centrality (BC) for all HPs. Cluster 3 and 6 proteins showed higher fractions of central proteins, while most Cluster 5 proteins were peripheral even at the relaxed threshold of BC $<10^{-8}$ (Figure 5A). The actual interactome positioning of the most central HPs is depicted in the network maps for the three species approached (Figure 5B). Cluster 6 and Cluster 1 proteins, the latter favored by its large size, nearly monopolized the most relevant nodes.

Overall, protein-level features support Cluster 1, 3, and 6 content for biomedical relevance while discouraging Cluster 2, 4, and more emphatically, Cluster 5.

Clusters Show Distinct Workability Trends: Predicted Protein Solubility and Structure Availability. A recurrent reason for lack of protein characterization is recalcitrance to

Table 1. Selection Criteria Values for Strong Antibacterial HP Target Candidates

cluster	protein	BC	solubility	spec_distr	PDB25	DUF	essentiality
1	HMMPREF0010_02899	0	0.749	0.949	84.85	DUF493	rich medium
1	KPHS_00170	0.00129	0.693	1.000	100.00	DUF1040	
1	KPHS_38450	0.00150	0.697	0.996	85.86	DUF1131	
1	KPHS_48910	0.00019	0.467	0.993	94.44		lung infection
1	PA1076	0.00232	0.619	0.998	100.00	DUF5064	
1	PA1233	0.00536	0.683	0.985	78.82	DUF2790	
1	PA3931	0.00125	0.719	0.853	93.82		
1	PA4535	0.00198	0.637	0.984	100.00	DUF1780	
2	PA3463	0	0.850	0.980	100.00		succinate medium
3	PA2635	0.00108	0.735	0.929	87.05	DUF839	
6	PA1624	0.00137	0.612	0.815	93.28	DUF4892	
6	PA5545	0.00225	0.751	0.941	91.85		

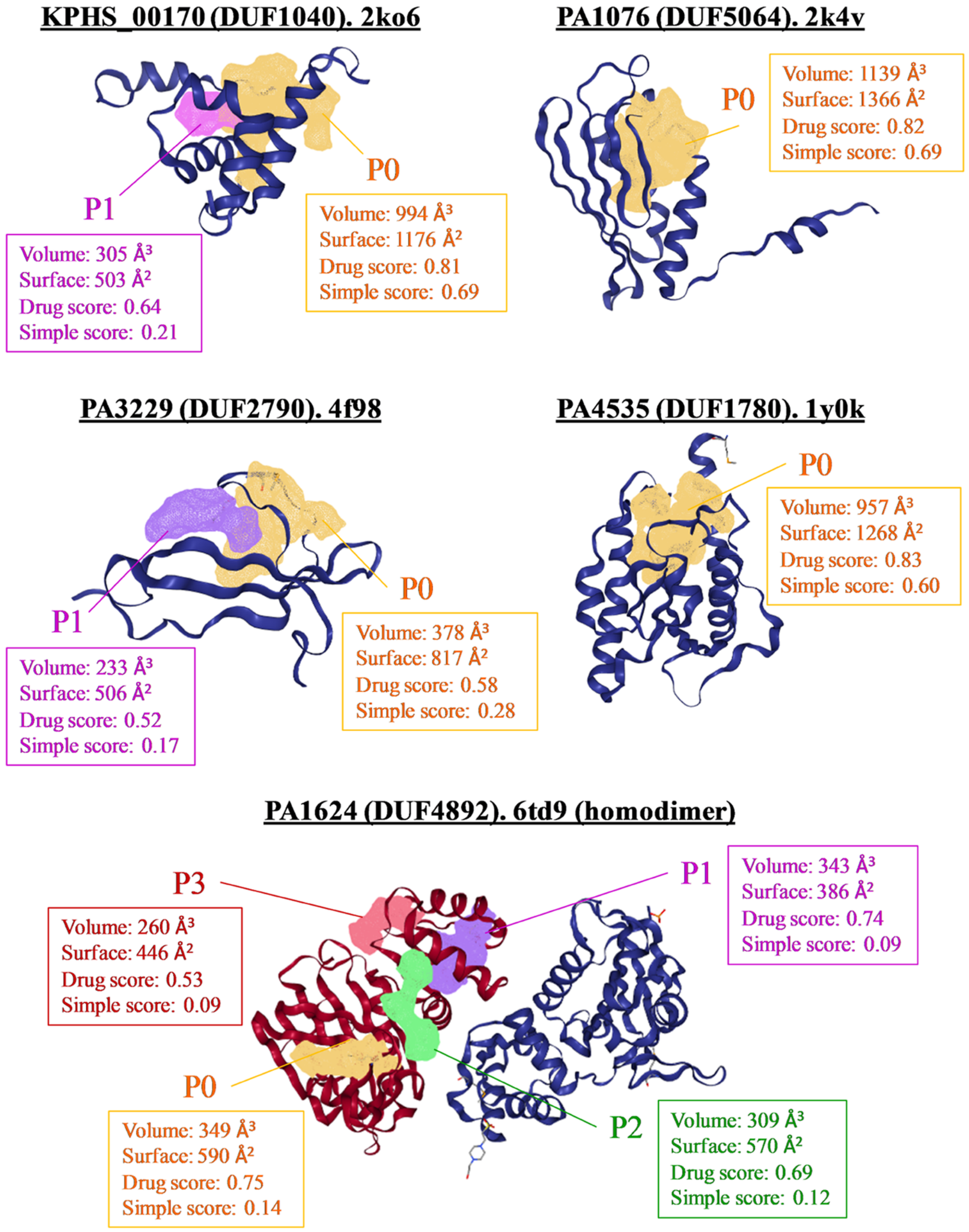
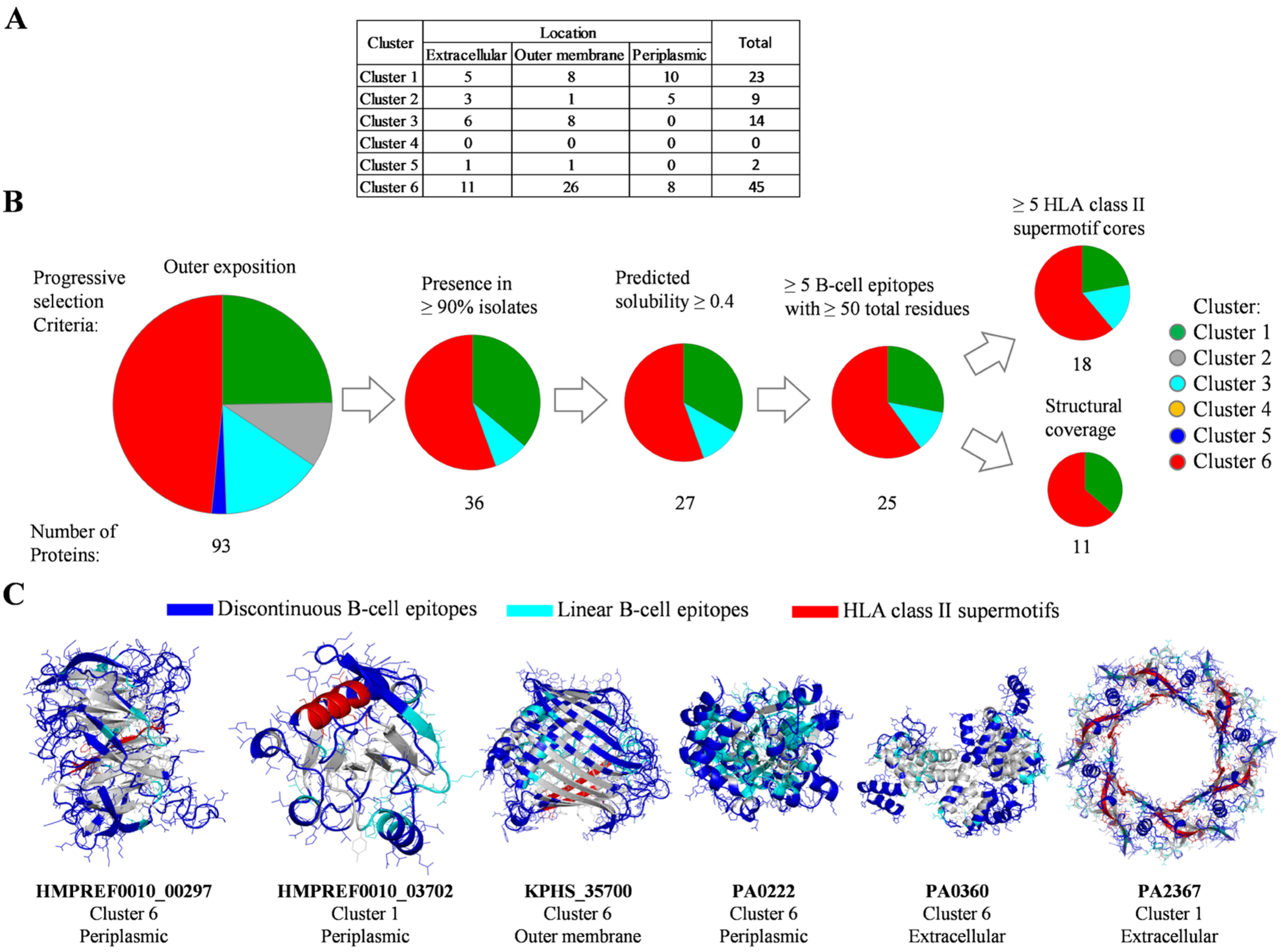


Figure 8. Structural panel of potential targets showing druggable binding pockets. Protein, DUF, and PDB id are indicated in the top. Pockets with druggability scores ≥0.5 are highlighted. Volume, surface, drug, and simple scores of the druggable pockets are listed.

solubility in standard molecular biology buffers. Solubility is a strict prerequisite for activity, ligand binding, and assays aiming to resolve protein structure. Four clusters included at least one-third of solubility-prone proteins according to a supervised



A

Cluster	Location			Total
	Extracellular	Outer membrane	Periplasmic	
Cluster 1	5	8	10	23
Cluster 2	3	1	5	9
Cluster 3	6	8	0	14
Cluster 4	0	0	0	0
Cluster 5	1	1	0	2
Cluster 6	11	26	8	45

Figure 9. Strong vaccine antigen candidates. (A) Number of predicted antibody-exposed proteins per cluster. (B) Antigen selection pipeline. The relative proportion of progressing proteins as antigenic per cluster in the selection pipeline. (C) Structure panel of antigen candidates. Presence of discontinuous and linear B-cell epitopes besides HLA class II supermotifs. Nonepitope residues are shown in gray. Detailed linear epitope data is included in Supporting Table S6.

machine learning prediction (Figure 6A). In contrast, Cluster 5 and Cluster 4 rendered only 24 and 9% predicted soluble proteins (both $P < 0.001$; Student's t test with respect to Cluster 1 solubility score distributions), the latter very likely due to its high hydrophobicity. On the other hand, another cause of undercharacterization is the nonavailability of three-dimensional information of the proteins or their homologues. This prevents the application of highly effective structure-based methods for functional characterization, which outperform sequence-based ones.[67] By far, Cluster 1, 3, and 6 excelled in the three model quality bins with respect to the sequences of PDB content: those considered as resolved (>90% identity), high-quality models (50−90% identity) and even 20% or more HPs in these clusters were at the minimal modeling distance required of ≥25% identity (Figure 6B).

Altogether, Cluster 1, 3, and 6 proteins were predicted to be more amenable to use as biochemical targets through biochemical or structure-based computational methods.

Clusters Show Distinct Distances to Biomedical Control Proteins. A critical question is to what extent proteins of demonstrated biomedical interest are close to HPs based on our 6-mer feature framework. If this is the case, shared feature values may indicate potential application of HP proteins in the clinical biotechnological field. To approach this, validated antibacterial drug targets, vaccine antigens, and VFs in the three species studied here were collected from respective central resources and the literature (see Materials and Methods). On average, the biomedical gold standards showed lower normalized Euclidean distances to the centroids of Cluster 1 and 6 (Figure 7). In particular, antibacterial targets were closer to the Cluster 1 centroid (median distance = 2.33), while antigens and VFs were closer to the Cluster 6 centroid (median distances = 1.48 and 1.50, respectively). The Cluster 4 centroid also showed a short median distance to VFs (2.64), indicating that a substantial number of pathogenicity determinants may be located in the membrane.

Examples of Uncharacterized Antibacterial Target Candidates. Beyond general trends per cluster, we searched for specific strong target candidates for antimicrobial therapy. For that, we applied a criteria selection pipeline for potential uncharacterized antimicrobial targets that rendered 12 proteins (Table 1). These selected proteins showed a predicted solubility >0.45, were present in ≥80% species isolates, showed structural coverage at modeling distance and either high BC values ($\geq 10^{-4}$), and/or were reported as essential in relevant conditions. Among those proteins, we prioritized proteins containing DUFs as these can be independent folding units of biomedical interest.[68] Moreover, monodomain DUF proteins with high sequence conservation and high interactome weight have also been deemed physiologically relevant HPs in *Streptococcus pneumoniae*.[69] Six selected proteins were subjected to a structure-based druggability analysis using the DogSiteScorer method. Five of these proteins contained pockets showing druggability scores >0.5 and could then be considered druggable targets (Figure 8).

Examples of Uncharacterized Vaccine Antigen Candidates. Another potential application of HPs is antigens for vaccine-induced prophylaxis. For example, an unannotated protein has been shown to be immunoprotective against *S. pneumoniae* in a murine model.[70] As immunoprotection against the species approached here is essentially humoral, candidate antigens must be accessible to antibodies. The location of 93 HPs was classified by PSORTb as "extracellular", "outer membrane", or "periplasmic" and then deemed exposed to antibodies (Figure 9A). While Cluster 1 only contained 2.0% exposed proteins, Cluster 3 (27.5%, $P = 4 \times 10^{-4}$, Fisher's exact test) and Cluster 6 (9.2%, $P < 1 \times 10^{-5}$) were proportionally enriched.

These proteins were subjected to a selection pipeline following principles of reverse vaccinology (Figure 9B), *i.e.*, the utilization of multidisciplinary scientific information for rational vaccine design.[71] Among the exposed proteins, 36 HPs showed enough distribution within the species to protect against the most pathogenic lineages. Next, 27 HPs fulfilled at least the relaxed solubility score of ≥0.4 to be considered as workable, as this may be desired for subunit or nucleic acid vaccine types. Twenty-five of them contained at least five linear B-cell epitope regions involving ≥ 50 residues. This indicates immune evasion toward these antigens would require several mutation events in the bacteria, which is unlikely without a considerable fitness penalty. Moreover, structural models for 11 of these B-cell antigen candidates could be built. This permits the assessment of discontinuous regions, that is, distant sections of the protein sequence that colocalize in the final folded protein, which constitute the majority of actual B-cell epitopes.[72] Finally, humoral antigens can also be T-cell immunoantigens since the HLA class II epitope presentation that stimulates T helper CD4+ lymphocytes is also involved in these infections.[73] Although the extent of such protection is essentially determined by the type of human allele, 18 out of the 25 B-cell antigens carried five or more supermotifs, those epitopes recognized by one of the five allelic supertypes that together cover up to 90% of the human population.[57] Six proteins (three periplasmic, two extracellular, one outer membrane), two from Cluster 1 and four from Cluster 6, fulfilled all criteria as attractive immunoantigens (Figure 9C).

DISCUSSION

The lack of functional data available for approximately one-fifth of isolate proteomes diminishes the potential biomedical application of these proteins in different fields. However, there is a pressing need for improved fundamental knowledge and to identify targets to combat infections caused by multidrug-resistant Gram-negative pathogens. This makes the development of alternative schemas to organize the dark proteome necessary. As this cannot be easily achieved, by definition, through classical homology-based tools, here we applied an alternative feature-based schema.

We selected proteins deemed as unannotated according to the central UniProt-Proteome resource from reference strains of three bacterial species classified as critically urgent by the WHO. Taking advantage of unsupervised machine learning techniques, six natural clusters were identified according to general protein properties besides qualities associated with a lack of functional solution. These features were preselected based on principles of protein science and unbiased algorithm performance. Some theoretical analyses of hypothetical proteins of these challenging Gram-negative species are available.[74,75] However, these studies update the annotation by homology search to recently known functions or applying more relaxed searching thresholds rather than approaching strictly unknown polypeptides. By our approach, HPs converged into biologically meaningful groups, i.e., dense zones in the 6th dimensional protein space, in a neutral fashion. Our study supports a high level of heterogeneity between the clusters identified. Our scheme could be refined using different clustering methods, adding further properties or considering the uncharacterized proteomes of other species so that some caution regarding the results obtained should be exerted. However, the observed clusters likely represent the principal trends that explain why a bacterial protein remains unannotated. These ranged from those containing strong candidates with a spectrum of potential biotechnological applications to likely spurious genes.

Cluster 1 represents the largest cluster (41.7% of all HPs) encoded by genes showing the %AT closest to the species average and relatively high isolate distribution. This cluster represents the general proteome by considering its features' values, regardless of annotation. Interestingly, the centroid of Cluster 1 was the closest one to antibacterial target controls. Among all of these cluster members, four proteins were deemed antibacterial target candidates showing 1−2 druggable pockets: KPHS_00170 (DUF1040), PA1076 (DUF5064), PA3229 (DUF2790), and PA4535 (DUF1780). Structures for most of these proteins have been solved by structural genomics projects[76] and are invaluable for further analysis. In addition, the cluster also included two strong vaccine B/T immunoantigen candidates: the periplasmic HMPREF0010_03702 protein and the extracellular PA2367 protein.

Cluster 2 (12.6% of all HPs) rather contains monocistronic, slightly basic proteins with partial species distribution and with remarkably high structural disorder (62.7% on average). Scarce PDB coverage further complicates its biomedical utilization but may serve as material for future protein science approaches involving unstructured regions. These proteins may become organized after protein interactions during important regulation and pathogenic processes.[77,78]

The tiny Cluster 3 (1.8% of all HPs) contains large proteins with over 650 residues. These are very likely multidomain proteins since most protein domains have less than 200 residues.[79] Several Cluster 3 members also tended to be exposed to the cell outside more than any other cluster, which has clinical relevance. Moreover, they can behave as humoral antigens, a fact reinforced by the presence of B-cell epitopes. On the other hand, they allow for direct interaction to host tissues as potential factors for colonization and virulence.

Cluster 4 (14.5% of all HPs) unites most integral membrane, high pI, and high GRAVY index proteins besides moderate isolate distribution within the species. These are recalcitrant to experimental work and have, accordingly, low structural coverage. Whether these are remote undetectable homologues of canonical transporter families recognizing new substrates or they represent novel nontransport functions must be addressed by the membrane protein scientific community.

Cluster 5 (12.0% of all HPs) HPs tend to be small, encoded by genes with salient %AT and codon usage, have low species distribution, and poor Pfam (DUF)/PDB coverage. An open question is whether Cluster 5 proteins are actual proteins. Some of these proteins may have been recently acquired by lateral gene transfer, but a majority of them were not part of operons, which goes against this idea. On the other hand, a large section of these ORFs may be pseudogenes or errors made by gene identification algorithms. This doubt justifies the wide utilization of the "hypothetical protein" concept over uncharacterized proteins. Whether these ORFs actually code for expressible functional protein products should be verified by genomic scientists. Of note, the low predominance of *P. aeruginosa* proteins in Cluster 5 proteins suggests a more strict curation of *P. aeruginosa* PAO1 (for instance, by the well-established Pseudomonas Genome DB resource at https://www.pseudomonas.com/),[80] with respect to reference strains of *A. baumannii* and *K. pneumoniae*, which would benefit from similar resources.

Finally, Cluster 6 (17.4% of all HPs) mainly contains conserved, acidic proteins with relatively larger sizes, low disordered content, and high PDB coverage, and many are coded by genes included in operons. Remarkably, this cluster contains the drug target candidate PA1624, whose inter-homodimer druggable interface has been previously suggested as a potential antimicrobial target.[81] The cluster is enriched in exposed proteins, and its centroid is the closest one to control virulence factors and antigens. Four cluster proteins satisfy the most stringent principles of strong vaccine B/T immunoantigen candidates with high human population coverage for the T helper response (HMPREF0010_00297, KPHS_35700, PA0222, and PA0360).

It should be noted that our approach cannot be compared in goals and complexity to other machine learning efforts. As a remarkable example, AlphaFold utilizes deep neural networks to iteratively predict template-free *ab initio* protein structures with an unprecedented atomic resolution level.[82] This involves massive features, including data from multiple sequence alignments, which were not required for our system. Moreover, the present study focuses on the ∼3000 proteins deemed unresolved by the central resource UniProt for the reference strains, which are those that are utilized in practice by experimentalist groups of the leading multidrug-resistant species. In contrast, AlphaFold has been recently applied to cover the whole protein space known,[83] regardless of whether proteins belong to multidrug-resistant organisms and their functional annotation status.

Conclusively, this study leverages the strength of the unsupervised machine learning tools and multiomic integration to surpass the limits of homology-based tools in exploring the dark proteome with a biomedical perspective. In this way, a functionally agnostic fractionation of the uncharacterized protein space facilitates to fill the ever-increasing gap between sequence availability and fundamental knowledge that can be used for combating difficult-to-treat organisms. The coarse homology-free protein sets we provided with precalculated information (available here as extensive Supporting information files) can then be selected by experimental groups specialized in fields such as chemotherapeutics, unstructured-based regulation, gene prediction, transport, vaccine development, and virulence, which otherwise would remain underutilized. The obtained self-organized map of the functional dark proteome is therefore a resource-saving factor in terms of time, human effort, and monetary cost. This is a prior stage that eases the development of therapeutics for populations affected by these infections.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04076.

Cluster correspondence between *k-means* and other clustering methods Figure S1 (PDF)

List of NCBI-Assembly codes for isolate genomes considered to calculate species distribution of HPs (Tables S1) (XLSX); raw and normalized feature values for cluster proteins (Table S2); values for features not used for clustering (Table S3); list of rare codons (Table S4); data for biomedical control proteins (Table S5); and epitope data for selected HP antigens (Table S6) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Antonio J. Martín-Galiano − *Intrahospital Infections Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III (ISCIII), 28220 Madrid, Spain;* Present Address: Core Scientific and Technical Units, Instituto de Salud Carlos III (ISCIII), Majadahonda, Madrid; orcid.org/0000-0002-6662-329X; Phone: +34 91 822 39 76; Email: mgaliano@isciii.es

Authors

Carlos Sicilia − *Intrahospital Infections Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III (ISCIII), 28220 Madrid, Spain*

Andrés Corral-Lugo − *Intrahospital Infections Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III (ISCIII), 28220 Madrid, Spain*

Pawel Smialowski − *Core Facility Bioinformatics, Biomedical Center Munich, Faculty of Medicine, Ludwig Maximilians Universität München, Munich 80539, Germany; Institute of Stem Cell Research, Helmholtz Center Munich, Planegg-Martinsried 82152, Germany*

Michael J. McConnell − *Intrahospital Infections Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III (ISCIII), 28220 Madrid, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04076

Author Contributions

$^\perp$C.S. and A.C.-L. contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This research was supported by Acción Estratégica en Salud from the ISCIII, Grant MPY 509/19, MPY 380/18, and MPY 516/19. A.C.-L. is the recipient of a Comunidad de Madrid contract by the ISCIII.

Notes

The authors declare the following competing financial interest(s): MM is a founder and shareholder in the biotechnology company Vaxdyn, S.L.

M.J.M. is a founder and shareholder in the biotechnology company Vaxdyn; S.L. Vaxdyn played no role in the present study. No other competing interest is declared for the other co-authors.

■ ACKNOWLEDGMENTS

The authors thank the ISCIII and the Comunidad de Madrid for funding.

■ ABBREVIATIONS USED

DUF, domain with unknown function; HP, hypothetical protein; PDB, protein data bank; pI, isoelectric point; PPI, protein−protein interactions; VFs, virulence factors

■ REFERENCES

(1) School, K.; Marklevitz, J.; K Schram, W.; K Harris, L. Predictive Characterization of Hypothetical Proteins in *Staphylococcus aureus* NCTC 8325. *Bioinformation* **2016**, *12*, 209−220.

(2) Kaur, H.; Singh, V.; Kalia, M.; Mohan, B.; Taneja, N. Identification and Functional Annotation of Hypothetical Proteins of Uropathogenic *Escherichia coli* Strain CFT073 towards Designing Antimicrobial Drug Targets. *J. Biomol. Struct. Dyn.* **2021**, 1−12.

(3) Levitt, M. Nature of the Protein Universe. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11079−11084.

(4) Paik, Y.-K.; Lane, L.; Kawamura, T.; Chen, Y.-J.; Cho, J.-Y.; LaBaer, J.; Yoo, J. S.; Domont, G.; Corrales, F.; Omenn, G. S.; et al. Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* **2018**, *17*, 4042−4050.

(5) Minarini, L. A. D. R.; de Andrade, L. N.; De Gregorio, E.; Grosso, F.; Naas, T.; Zarrilli, R.; Camargo, I. L. B. C. Editorial: Antimicrobial Resistance as a Global Public Health Problem: How Can We Address It? *Front. Public Health* **2020**, *8*, No. 612844.

(6) Tacconelli, E.; Carrara, E.; Savoldi, A.; Harbarth, S.; Mendelson, M.; Monnet, D. L.; Pulcini, C.; Kahlmeter, G.; Kluytmans, J.; Carmeli, Y.; et al. Discovery, Research, and Development of New Antibiotics: The WHO Priority List of Antibiotic-Resistant Bacteria and Tuberculosis. *Lancet Infect. Dis.* **2018**, *18*, 318−327.

(7) Rello, J.; Kalwaje Eshwara, V.; Lagunes, L.; Alves, J.; Wunderink, R. G.; Conway-Morris, A.; Rojas, J. N.; Alp, E.; Zhang, Z. A Global Priority List of the TOp TEn Resistant Microorganisms (TOTEM) Study at Intensive Care: A Prioritization Exercise Based on Multi-Criteria Decision Analysis. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **2019**, *38*, 319−323.

(8) Miriagou, V.; Tzelepi, E.; Daikos, G. L.; Tassios, P. T.; Tzouvelekis, L. S. Panresistance in VIM-1-Producing *Klebsiella pneumoniae*. *J. Antimicrob. Chemother.* **2005**, *55*, 810−811.

(9) Silver, L. L. Challenges of Antibacterial Discovery. *Clin. Microbiol. Rev.* **2011**, *24*, 71−109.

(10) Bloom, D. E.; Black, S.; Salisbury, D.; Rappuoli, R. Antimicrobial Resistance and the Role of Vaccines. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 12868−12871.

(11) Vashist, J.; Tiwari, V.; Kapil, A.; Rajeswari, M. R. Quantitative Profiling and Identification of Outer Membrane Proteins of Beta-Lactam Resistant Strain of *Acinetobacter baumannii*. *J. Proteome Res.* **2010**, *9*, 1121−1128.

(12) Uddin, R.; Jamil, F. Prioritization of Potential Drug Targets against *P. aeruginosa* by Core Proteomic Analysis Using Computational Subtractive Genomics and Protein-Protein Interaction Network. *Comput. Biol. Chem.* **2018**, *74*, 115−122.

(13) Lehoux, D. E.; Sanschagrin, F.; Levesque, R. C. Identification of in Vivo Essential Genes from *Pseudomonas aeruginosa* by PCR-Based Signature-Tagged Mutagenesis. *FEMS Microbiol. Lett.* **2002**, *210*, 73−80.

(14) Maroncle, N.; Balestrino, D.; Rich, C.; Forestier, C. Identification of *Klebsiella pneumoniae* Genes Involved in Intestinal Colonization and Adhesion Using Signature-Tagged Mutagenesis. *Infect. Immun.* **2002**, *70*, 4729−4734.

(15) Struve, C.; Forestier, C.; Krogfelt, K. A. Application of a Novel Multi-Screening Signature-Tagged Mutagenesis Assay for Identification of *Klebsiella pneumoniae* Genes Essential in Colonization and Infection. *Microbiology* **2003**, *149*, 167−176.

(16) Jünemann, S.; Sedlazeck, F. J.; Prior, K.; Albersmeier, A.; John, U.; Kalinowski, J.; Mellmann, A.; Goesmann, A.; von Haeseler, A.; Stoye, J.; Harmsen, D. Updating Benchtop Sequencing Performance Comparison. *Nat. Biotechnol.* **2013**, *31*, 294−296.

(17) Jenssen, M.; Rainsford, P.; Juskewitz, E.; Andersen, J. H.; Hansen, E. H.; Isaksson, J.; Rämä, T.; Hansen, KØ. Lulworthinone, a New Dimeric Naphthopyrone From a Marine Fungus in the Family Lulworthiaceae With Antibacterial Activity Against Clinical Methicillin-Resistant *Staphylococcus aureus* Isolates. *Front. Microbiol.* **2021**, *12*, No. 730740.

(18) Mardis, E. R. The Impact of Next-Generation Sequencing Technology on Genetics. *Trends Genet.* **2008**, *24*, 133−141.

(19) Bateman, A.; Coggill, P.; Finn, R. D. DUFs: Families in Search of Function. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2010**, *66*, 1148−1152.

(20) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; et al. The Pfam Protein Families Database. *Nucleic Acids Res.* **2012**, *40*, D290−301.

(21) Galperin, M. Y.; Koonin, E. V. From Complete Genome Sequence to "complete" Understanding? *Trends Biotechnol.* **2010**, *28*, 398−406.

(22) Chang, Y.-C.; Hu, Z.; Rachlin, J.; Anton, B. P.; Kasif, S.; Roberts, R. J.; Steffen, M. COMBREX-DB: An Experiment Centered Database of Protein Function: Knowledge, Predictions and Knowledge Gaps. *Nucleic Acids Res.* **2016**, *44*, D330−335.

(23) Perdigão, N.; Rosa, A. C.; O'Donoghue, S. I. The Dark Proteome Database. *BioData Min.* **2017**, *10*, 24.

(24) Pawlowski, K. Uncharacterized/Hypothetical Proteins in Biomedical "omics" Experiments: Is Novelty Being Swept under the Carpet? *Briefings Funct. Genomics Proteomics* **2008**, *7*, 283−290.

(25) Hinton, G.; Sejnowski, T. J. *Unsupervised Learning: Foundations of Neural Computation*; The MIT Press, 1999.

(26) Angadi, U. B.; Venkatesulu, M. Structural SCOP Superfamily Level Classification Using Unsupervised Machine Learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2012**, *9*, 601−608.

(27) Alam, F. F.; Shehu, A. Unsupervised Multi-Instance Learning for Protein Structure Determination. *J. Bioinform. Comput. Biol.* **2021**, *19*, 2140002.

(28) UniProt Consortium; Martin, M. J.; Orchard, S.; et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(29) Gene Ontology Consortium; Douglass, E.; Good, B. M.; et al. The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* **2021**, *49*, D325−D334.

(30) Kitts, P. A.; Church, D. M.; Thibaud-Nissen, F.; Choi, J.; Hem, V.; Sapojnikov, V.; Smith, R. G.; Tatusova, T.; Xiang, C.; Zherikov, A.; et al. Assembly: A Resource for Assembled Genomes at NCBI. *Nucleic Acids Res.* **2016**, *44*, D73−80.

(31) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150−3152.

(32) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein Identification and Analysis Tools on Hte ExPASy Server. In *The Proteomics Protocols Handbook*; Humana Press Inc: Totowa, NJ, 2005; pp 571−607.

(33) Lupas, A.; Van Dyke, M.; Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **1991**, *252*, 1162−1164.

(34) Wootton, J. C.; Federhen, S. Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Comput. Chem.* **1993**, *17*, 149−163.

(35) Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347*, 827−839.

(36) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*.

(38) Taboada, B.; Estrada, K.; Ciria, R.; Merino, E. Operon-Mapper: A Web Server for Precise Operon Identification in Bacterial and Archaeal Genomes. *Bioinformatics* **2018**, *34*, 4118−4120.

(39) Mistry, J.; Finn, R. D.; Eddy, S. R.; Bateman, A.; Punta, M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Res.* **2013**, *41*, e121.

(40) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; et al. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412−D419.

(41) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; et al. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47*, D607−D613.

(42) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **2001**, *305*, 567−580.

(43) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37*, 420−423.

(44) Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II–a New Method for Protein Solubility Prediction. *FEBS J.* **2012**, *279*, 2192−2200.

(45) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296−W303.

(46) Luo, H.; Lin, Y.; Liu, T.; Lai, F.-L.; Zhang, C.-T.; Gao, F.; Zhang, R. DEG 15, an Update of the Database of Essential Genes That Includes Built-in Analysis Tools. *Nucleic Acids Res.* **2021**, *49*, D677−D686.

(47) Bachman, M. A.; Breen, P.; Deornellas, V.; Mu, Q.; Zhao, L.; Wu, W.; Cavalcoli, J. D.; Mobley, H. L. T. Genome-Wide Identification of *Klebsiella pneumoniae* Fitness Genes during Lung Infection. *mBio* **2015**, *6*, No. e00775.

(48) Wang, N.; Ozer, E. A.; Mandel, M. J.; Hauser, A. R. Genome-Wide Identification of *Acinetobacter baumannii* Genes Necessary for Persistence in the Lung. *mBio* **2014**, *5*, e01163−01114.

(49) Turner, K. H.; Wessel, A. K.; Palmer, G. C.; Murray, J. L.; Whiteley, M. Essential Genome of *Pseudomonas aeruginosa* in Cystic Fibrosis Sputum. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 4110−4115.

(50) Hagberg, A.; Swart, P.; Chult, D. In *Exploring Network Structure, Dynamics, and Function Using NetworkX*, Proceedings of the 7th Python in Science Conference, 2008.

(51) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360−372.

(52) Yu, N. Y.; Wagner, J. R.; Laird, M. R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S. C.; Ester, M.; Foster, L. J.; Brinkman, F. S. L. PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes. *Bioinformatics* **2010**, *26*, 1608−1615.

(53) Jespersen, M. C.; Peters, B.; Nielsen, M.; Marcatili, P. BepiPred-2.0: Improving Sequence-Based B-Cell Epitope Prediction Using Conformational Epitopes. *Nucleic Acids Res.* **2017**, *45*, W24−W29.

(54) Kringelum, J. V.; Lundegaard, C.; Lund, O.; Nielsen, M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Comput. Biol.* **2012**, *8*, No. e1002829.

(55) Ponomarenko, J.; Bui, H.-H.; Li, W.; Fusseder, N.; Bourne, P. E.; Sette, A.; Peters, B. ElliPro: A New Structure-Based Tool for the Prediction of Antibody Epitopes. *BMC Bioinf.* **2008**, *9*, 514.

(56) Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data. *Nucleic Acids Res.* **2020**, *48*, W449−W454.

(57) Sidney, J.; Steen, A.; Moore, C.; Ngo, S.; Chung, J.; Peters, B.; Sette, A. Five HLA-DP Molecules Frequently Expressed in the Worldwide Human Population Share a Common HLA Supertypic Binding Specificity. *J. Immunol.* **2010**, *184*, 2492−2503.

(58) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.

(59) López-Siles, M.; Corral-Lugo, A.; McConnell, M. J. Vaccines for Multidrug Resistant Gram Negative Bacteria: Lessons from the Past for Guiding Future Success. *FEMS Microbiol. Rev.* **2021**, *45*, No. fuaa054.

(60) Liu, B.; Zheng, D.; Jin, Q.; Chen, L.; Yang, J. VFDB 2019: A Comparative Pathogenomic Platform with an Interactive Web Interface. *Nucleic Acids Res.* **2019**, *47*, D687−D692.

(61) Patel, A. A. *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*; O'Reilly Media, Inc: USA, 2019.

(62) Sharp, P. M.; Bailes, E.; Grocock, R. J.; Peden, J. F.; Sockett, R. E. Variation in the Strength of Selected Codon Usage Bias among Bacteria. *Nucleic Acids Res.* **2005**, *33*, 1141−1153.

(63) Martín-Galiano, A. J.; Wells, J. M.; de la Campa, A. G. Relationship between Codon Biased Genes, Microarray Expression Values and Physiological Characteristics of *Streptococcus pneumoniae*. *Microbiology* **2004**, *150*, 2313−2325.

(64) Kuo, C.-H.; Ochman, H. The Extinction Dynamics of Bacterial Pseudogenes. *PLoS Genet.* **2010**, *6*, e1001050.

(65) Price, M. N.; Huang, K. H.; Arkin, A. P.; Alm, E. J. Operon Formation Is Driven by Co-Regulation and Not by Horizontal Gene Transfer. *Genome Res.* **2005**, *15*, 809−819.

(66) Melak, T.; Gakkhar, S. Comparative Genome and Network Centrality Analysis to Identify Drug Targets of *Mycobacterium tuberculosis* H37Rv. *BioMed Res. Int.* **2015**, *2015*, No. 212061.

(67) Gherardini, P. F.; Helmer-Citterich, M. Structure-Based Function Prediction: Approaches and Applications. *Briefings Funct. Genomics Proteomics* **2008**, *7*, 291−302.

(68) Goodacre, N. F.; Gerloff, D. L.; Uetz, P. Protein Domains of Unknown Function Are Essential in Bacteria. *mBio* **2013**, *5*, e00744−00713.

(69) Martín-Galiano, A. J.; Yuste, J.; Cercenado, M. I.; de la Campa, A. G. Inspecting the Potential Physiological and Biomedical Value of 44 Conserved Uncharacterised Proteins of *Streptococcus pneumoniae*. *BMC Genomics* **2014**, *15*, 652.

(70) Martín-Galiano, A. J.; Escolano-Martínez, M. S.; Corsini, B.; de la Campa, A. G.; Yuste, J. Immunization with SP_1992 (DiiA) Protein of *Streptococcus pneumoniae* Reduces Nasopharyngeal Colonization and Protects against Invasive Disease in Mice. *Vaccines* **2021**, *9*, No. 187.

(71) Rappuoli, R.; Bottomley, M. J.; D'Oro, U.; Finco, O.; De Gregorio, E. Reverse Vaccinology 2.0: Human Immunology Instructs Vaccine Antigen Design. *J. Exp. Med.* **2016**, *213*, 469−481.

(72) Sivalingam, G. N.; Shepherd, A. J. An Analysis of B-Cell Epitope Discontinuity. *Mol. Immunol.* **2012**, *51*, 304−309.

(73) Bayes, H. K.; Bicknell, S.; MacGregor, G.; Evans, T. J. T Helper Cell Subsets Specific for *Pseudomonas aeruginosa* in Healthy Individuals and Patients with Cystic Fibrosis. *PLoS One* **2014**, *9*, e90263.

(74) Pranavathiyani, G.; Prava, J.; Rajeev, A. C.; Pan, A. Novel Target Exploration from Hypothetical Proteins of *Klebsiella pneumoniae* MGH 78578 Reveals a Protein Involved in Host-Pathogen Interaction. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 109.

(75) Sen, T.; Verma, N. K. Functional Annotation and Curation of Hypothetical Proteins Present in A Newly Emerged Serotype 1c of *Shigella flexneri*: Emphasis on Selecting Targets for Virulence and Vaccine Design Studies. *Genes* **2020**, *11*, No. 340.

(76) Michalska, K.; Joachimiak, A. Structural Genomics and the Protein Data Bank. *J. Biol. Chem.* **2021**, *296*, 100747.

(77) Blundell, T. L.; Gupta, M. N.; Hasnain, S. E. Intrinsic Disorder in Proteins: Relevance to Protein Assemblies, Drug Design and Host-Pathogen Interactions. *Prog. Biophys. Mol. Biol.* **2020**, *156*, 34−42.

(78) Seoane, B.; Carbone, A. The Complexity of Protein Interactions Unravelled from Structural Disorder. *PLoS Comput. Biol.* **2021**, *17*, e1008546.

(79) Lin, M. M.; Zewail, A. H. Hydrophobic Forces and the Length Limit of Foldable Protein Domains. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 9851−9856.

(80) Winsor, G. L.; Griffiths, E. J.; Lo, R.; Dhillon, B. K.; Shay, J. A.; Brinkman, F. S. L. Enhanced Annotations and Features for Comparing Thousands of Pseudomonas Genomes in the Pseudomonas Genome Database. *Nucleic Acids Res.* **2016**, *44*, D646−653.

(81) Feiler, C. G.; Weiss, M. S.; Blankenfeldt, W. The Hypothetical Periplasmic Protein PA1624 from *Pseudomonas aeruginosa* Folds into a Unique Two-Domain Structure. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2020**, *76*, 609−615.

(82) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Applying and Improving AlphaFold at CASP14. *Proteins* **2021**, *89*, 1711−1721.

(83) Callaway, E. "The Entire Protein Universe": AI Predicts Shape of Nearly Every Known Protein. *Nature* **2022**, *608*, 15−16.