Leveraging Large Language Models for Predicting Microbial Virulence from Protein Structure and Sequence

Felix Quintana
Computer Science Department
Rice University
Houston, Texas, USA

Todd J. Treangen
Computer Science Department
Rice University
Houston, Texas, USA

Lydia E. Kavraki*
kavraki@rice.edu
Computer Science Department
Rice University
Houston, Texas, USA

ABSTRACT

In the aftermath of COVID-19, screening for pathogens has never been a more relevant problem. However, computational screening for pathogens is challenging due to a variety of factors, including (i) the complexity and role of the host, (ii) virulence factor divergence and dynamics, and (iii) population and community-level dynamics. Considering a potential pathogen's molecular interactions, specifically individual proteins and protein interactions can help pinpoint a potential protein of a given microbe to cause disease. However, existing tools for pathogen screening rely on existing annotations (KEGG, GO, etc), making the assessment of novel and unannotated proteins more challenging. Here, we present an LLM-inspired approach that considers protein sequence and structure to predict protein virulence. We present a two-stage model incorporating evolutionary features captured from the DistilProtBert language model and protein structure in a graph convolutional network. Our model performs better than sequence alone for virulence function when high-quality structures are present, thus representing a path forward for virulence prediction of novel and unannotated proteins.

CCS CONCEPTS

• **Applied computing** → **Structural Bioinformatics**.

KEYWORDS

Protein Function, Virulence Prediction, Graph-based models, Large Language Models

ACM Reference Format:
Felix Quintana, Todd J. Treangen, and Lydia E. Kavraki. 2023. Leveraging Large Language Models for Predicting Microbial Virulence from Protein Structure and Sequence. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3584371.3612953

*Corresponding author.

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3612953

1 INTRODUCTION

Recent global biological threats like COVID-19 have shown the urgency to detect pathogens swiftly and promptly. However, detecting pathogens requires improved modeling of what causes disease or virulence, or simply, "what makes a bad bug bad?". There are several challenges to generally modeling and predicting virulence, including the host immune system, pathogen-host interactions, microbe-microbe interactions, as well as the complex interplay of genes responsible for virulence[10]. Despite much progress in this field [3, 8, 26], much remains to be elucidated for viral, bacterial, and eukaryotic virulence across a wide range of hosts. These "confounders" for accurate pathogen-host virulence prediction result from protein interactions occurring at the molecular level. Analyzing a pathogen's proteins and associated functions with detailed characterization can provide insights into their general virulence and distinguish host or environmental-specific virulence and any interplay with different genes to cause virulence.

One approach for describing a protein's function is by its gene ontology (GO) terms. These terms classify a protein's function hierarchically related by three classes: Molecular Function, Biological Process, and Cellular Component[2]. GO annotations describe a protein's function at the most granular level, where a set of GO terms assigned to a protein can provide a well-rounded understanding of the protein's function. However, these terms are not associated with virulence directly. Therefore, virulence-specific protein databases are needed to link virulence factors to specific proteins. A few relevant databases are the Virulence Factor Database (VFDB)[8], VEuPathDB[1], the SeqScreen Functions of Sequences of Concern (FunSoC) database[5], and a recently created resource is the Pathogenesis Gene Ontology database, or PathGO[14].

Methods that rely on these virulence factor databases often use local sequence alignment and sequence similarity to determine the presence or absence of a specific virulence factor. Still, as the protein sequence divergence increases, or for proteins not in the database, the performance of these methods rapidly decreases. Furthermore, while a widely used resource, VFDB has limitations, including poor sequence annotation, discrepancies of pathogenic and non-pathogenic features, and poor GO term diversity[10]. A recent approach that takes advantage of both GO terms and the richer virulence annotation FunSoCs is called Seqscreen[5]. At its core, SeqScreen is a human-in-the-loop machine learning platform for assigning virulence functions (functions of sequences of concern, or FunSoCs) to sequence fragments, from individual sequencing reads to gene fragments, to assembled contigs and ORFs. Although SeqScreen only labels a small fraction of the proteins in UniRef100 with FunSoCs, this is on par with expectations as these are thought

to be low prevalence in microbes. However, if a novel protein arises that is not contained in the SeqScreen FunSoC database or is highly divergent from those contained within, it will struggle to characterize it.

Alternatively, several deep learning methods have recently been deployed such as DeepGO[18], and DeepFRI[12]. The former is a protein sequence tool composed of a deep neural network with convolutional layers, while the latter is a protein structure-based graph convolutional network (GCN). However, the performance of these tools is limited as predicting GO terms contains a similar fundamental problem to predicting virulence, where exotic proteins are poorly understood.

In this work, we propose to exclude similarity scoring and consider the structural features that promote the function or virulence annotation for automatic function and virulence annotation assignment. Proteins determine a wide set of functions from their 3-dimensional conformations, from binding specificity and conferring mechanical stability to catalysis of biochemical reactions[17]. However, the virulence of a protein can come from any or a combination of these functions. A basic understanding of where these interactions can occur is unknown and could provide a better fundamental understanding of what constitutes virulence allowing better characterization of novel proteins.

As deep learning models have become popular in recent years, their application to biological problems has shown substantial progress. Availability of protein structure has recently exploded from deep models like AlphaFold2[16]. Concerning protein structure, deep large language models (LLM), have recently been shown to capture evolutionary features along sequences as demonstrated in ESMFold[19], replacing the information captured from multiple sequence alignments (MSAs) that other state-of-the-art models use. Similarly, its been shown that features extracted from pre-trained, task-agnostic, LLM can significantly increase classification performance in many biological problems[20]. LLM have been deployed for protein function prediction. SPROF-GO [27] surpassed state-of-the-art sequence-based and network approaches by coupling a LLM and stable diffusion. Furthermore, advances in geometric deep learning have provided new modules built for protein structure-based tasks. In particular, geometric vector perceptions (GVP) are a novel neural module transforming euclidean positions to their equivariant form which is beneficial in protein structure-related tasks such as docking[15].

This work takes advantage of LLMs and protein structure to predict the protein virulent function expressed by FunSoCs annotations with unlabeled proteins. We include a baseline method considering only sequence to see what sequence can understand about virulence. The LLM finds local sequential features to make its decision on virulence. The LLM we used is a flavor of ProtBert [6], a masked language model with BERT transformer architecture. ProtBert was trained on UniRef100, composed of over 217 million protein sequences. It has shown to be good at predicting GO terms which provide proximity to function [22]. The baseline model uses a fine-tuned and distilled version of ProtBert called DistilProtBert [11] to classify FunSoCs. Our proposed method is a graph-based approach exploiting GVPs in GNNs coupled with evolutionary information from DistilProtBert. This method allows us to exploit the explicit geometry of the protein and evolutionary information from pairing nearby amino acids.

2 METHODS

2.1 Dataset

The protein dataset was curated from UniProtKB [9] and Swiss-Prot [4] from domain experts manually annotating FunSoCs. We considered all FunSoCs categories available amounting to 32 different FunSoCs non-mutually exclusive. The definition of FunSoCs is a broader set of annotations allowing for diversity amongst each FunSoCs annotation such as cytotoxicity covering small peptides to structures containing sub-units amounting over 4000 amino acids. A full table of each FunSoC and associated number of structures used for training is available in Table S1.

Protein structures were first searched from AlphaFold database [24] updated on November 1st, 2022. If the protein did not exist in their database, it was computed using ColabFold v1.5.[21]. Structures with an average predicted local distance difference test (PLDDT) score of 75 or lower were excluded to limit low-quality structures for a total of 8,000 structures. Experimental structures were not used to limit potential noise from missing domains to unwanted chains. Furthermore, excluding experimental structures had a small impact on the dataset size since most UniProt IDs lacked any experimental structures. Protein structures were also limited to no larger than 2500 amino acids due to memory constraints. Lastly, structures shorter than 50 amino acids were also excluded to remove any sequences lacking tertiary structure. Protein structures were split into training, validation, and test sets with an 80/10/10% ratio split. The splits were enforced on a per-FunSoC basis meaning each FunSoC contained at least 10% held out for testing. Weighted sampling during training was enforced to help elevate the class imbalance between FunSoCs.

2.2 Graph Construction

Graphs were constructed at the residue level where each node is represented the backbone carbon alpha of each amino acid (see Figure 1). Edges were calculated from $C_\alpha - C_\alpha$ contact maps with an 10Å cutoff. A k-nearest neighbor of graph nodes representing C_αs was also tried with a grid search of [5,10,15,20,30] finding similar results to contact maps. A node embedding from DistilProtBert [11] of vector size of 1024, was assigned per node from respective amino acid. Node features are composed of scalar and vector features. The scalar features include the sines and cosines of the dihedral angles ϕ, ψ, ω, and vector features consist of the forward and reverse unit vectors in the directions of adjacent C-alpha atoms from two neighboring amino acids and the unit vector in the imputed direction of C-alpha and C-beta atoms. More explicitly, the unit vector direction is calculated considering perpendicular bisection of the nitrogen, and carbon atom from carboxyl group.

Edge features are composed of scalar and vector features. Scalar features include the encoding of C-alpha distance in terms of 16 Gaussian radial basis functions with centers evenly spaced between 0 and 20 angstroms, and a positional encoding of j-i[25], representing the AA distance alone the 1D protein sequence. Lastly, the vector feature is the unit vector in the direction of connecting carbon alpha atoms.

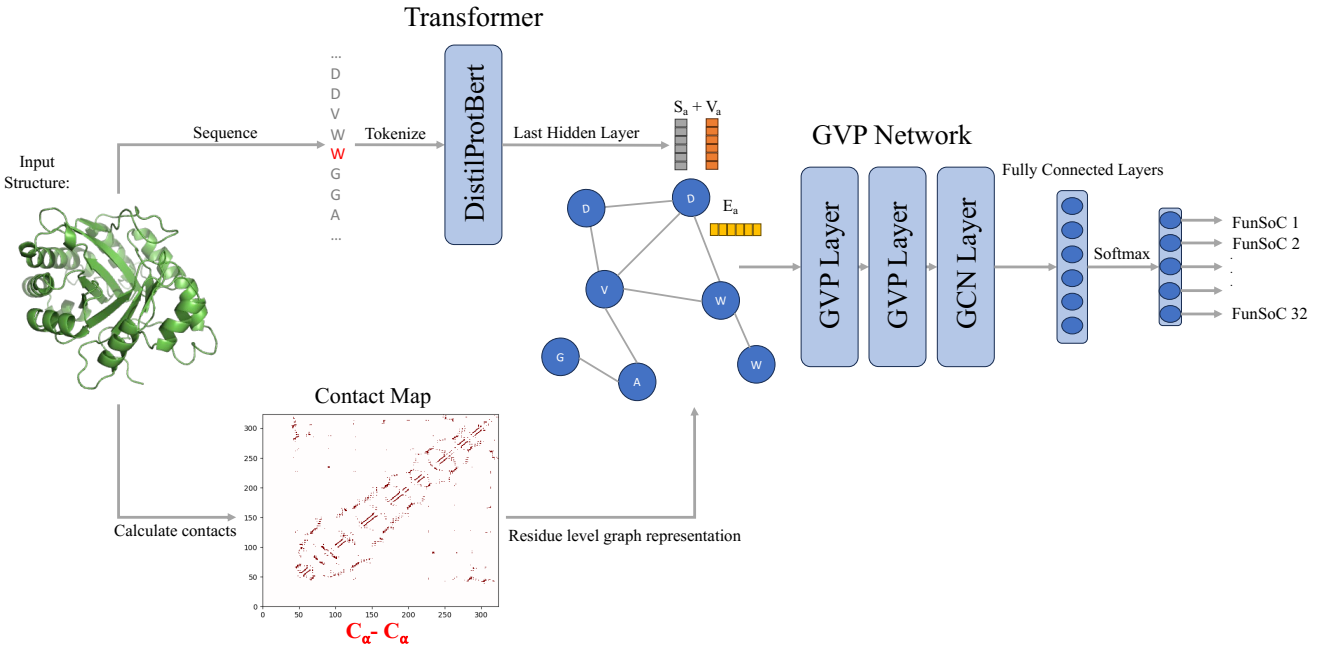

Figure 1: Schematic Overview of Method: Overview of GVP Network+LLM classifier, input protein structure contact map of $C_\alpha - C_\alpha$ contacts less than 10Å is calculated then represented in a graph network. Amino acid embedding from LLM is assigned to each node S_a concatenated with additional directional and positional features V_a. Edge positional and directional features are also assigned E_a. Graph is then passed through GVP Covolutional Network for FunSoCs prediction.

2.3 Model Training and Hyper-parameter tuning

To account for the imbalanced dataset, both models were trained to minimize weighted binary cross entropy giving weight to FunSoCs with fewer examples. The learning rate for both models was 0.0001. Random sampling was also weighted for training and validation splits. No filtering was performed pre-split besides filtering for structures larger than 2500 residues due to computational constraints for training. During training, DistilProtBert's layers were gradually unfrozen to fine-tune to classification task [13]. A token was generated for each amino acid in each protein structure with padding and truncation for a length of 2500. This is the same cut off of largest protein size ensuring each protein gets a token for encoding. A batch size of four was used due to memory constraints from DistilProtBert. To minimize batch bias, accumulative gradients were used for every 8 batches. This was shown to provide a small performance increase. The training was performed on 96GBs of system ram, Nvidia 3070 8GB GPU, and Nvidia 3090 24GB GPU for a duration of eight hours for each model.

3 RESULTS

As the first investigation into virulence without annotation to simulate novel proteins, there is no outright comparison of other approaches. We consider precision and recall as two suitable metrics of performance because high fidelity and sensitivity is required for good pathogen detection. Since there are 32 different FunSoCs labels to consider, we record precision and recall for each FunSoC category and calculate the corresponding F1 score as this summarizes both precision and recall in one value. As shown in Figure 2, the individual performance of each FunSoC category has high variability for both the sequence-only baseline and the proposed model. However, there are several specific categorical results to take note of.

In the cases where the proposed method under-performed to sequence, including development in host and non-viral invasion, there is a large discrepancy in the ratio of high-quality predicted structure quality in the FunSoC category. Although we selected a 75 average PLDDT score cut-off for included structures, many poor-quality structures persisted by containing a high-scoring core and large loops of disordered regions wrapping around the protein. An example is the AlphaFold model of Q5A0E5, a transcriptional regulator NRG1 from Candida albicans. The predicted structure consists of largely disordered loops with small alpha helices boosting the artificial score of the model just enough so to meet the 75 PLLDT threshold. In contrast, the sequence doesn't contain the same challenge as the structure, allowing for a less noisy sample for better prediction.

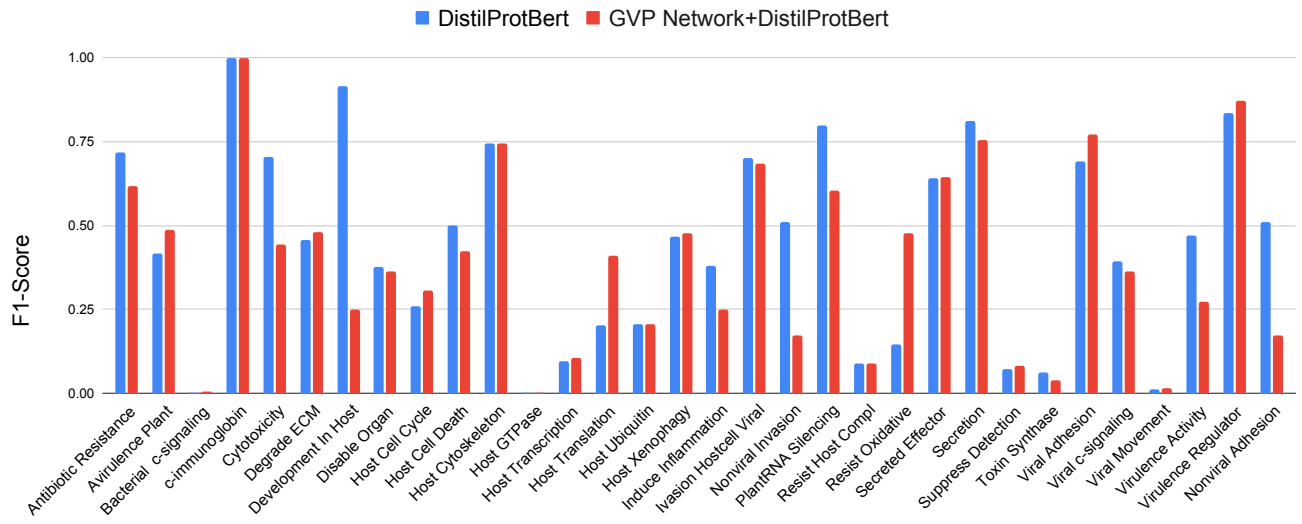

Figure 2: Performance of the two methods in the virulence prediction task. F1 score of the DistProtBert LLM model (blue) and the GVP model (red) across different FunSoC groups.

FunSoCs which performed marginally better on the GVP model (structure-based), correlated to a large ratio of high-quality structures in the dataset. These FunSoCs include viral adhesion and virulence regulator containing an average PLDDT of each structure much higher than the selected cut-off with an average PLDDT of all structures in the FunSoC of 92. Previous work has demonstrated that predicted structures from AlphaFold with an average PLDDT scores of 90 or above display interactions and mutation behavior close to their experimental counterparts[7].

FunSoCs that perform similarly on both the baseline and the proposed method showed limitations for both approaches. In particular, degrade ECM is a FunSoC containing many well-known proteins, including metalloproteases and disintegrin proteins or a protein containing a metalloprotease sub-unit in a chain. Metalloprotease binding and function are well known, relying on a zinc metal motif to cleave proteins. The motif is conserved amongst the metalloprotease family with a similar story for disintegrins, but both methods performed poorly on Degrade ECM. This highlights that both methods may need more refined features and structure for classification. The structure as a whole is not enough to classify the proteins well. The poor performance is also demonstrated by virtually zero scores in bacterial counter signaling and Host GTPase for both sequence and protein structure.

Lastly, the high performance of both the baseline (highlighted in blue) and our structural method (highlighted in red) for the Counter Immunoglobin is misleading. This FunSoCs training data is small, incorporating a total of seven structures, making classification much simpler at the cost of real-world performance. The lack of structures is primarily due to the limited amount of annotated proteins with the designated FunSoC. Counter immunoglobin structures also lack structural and sequence similarity consisting of a few long alpha-helix secondary structures connected by disordered domains. When clustering these seven proteins around 90% sequence similarity the proteins clustered into two groups.

4 DISCUSSION

Here we present a large language model (LLM) based approach for predicting microbial virulence from protein structure. As virulence amongst the different categories varied, so did the protein structure quality and available data. Using protein structure showed a performance uplift in FunSoC categories that contained high-quality structures such as virulence regulator. Conversely, those with low-quality structures performed worse than sequence. The performance discrepancies of structure quality provide evidence of the potential superiority of structure-based models when high-quality data is available. However, both models contained FunSoCs categories that performed poorly. These poor performing FunSoCs, it will be difficult to determine function and virulence from protein structures. However, protein substructures such as binding sites are known to orchestrate protein function [23]. Exploration is still possible to isolate particular substructures associated with protein function. Therefore, substructures for poor-performing FunSoCs could show a performance uplift by their targeted nature.

Furthermore, structure quality is an important consideration. Poor quality structures result in high proportions of disordered loops wrapping around small secondary structures or, in worst cases, no secondary structures failing to form a comprehensive tertiary structure.

Considering the performed structure filtering, more comprehensive filtering is needed to ensure quality structures. One potential approach to consider is Alphafold's predicted aligned error matrix measuring the alignment error for an amino acid in an all vs. all fashion. If presented with high-quality structures, the square matrix would contain low values indicating each amino acid's position has

a high probability of being placed appropriately for the other amino acids. In contrast, structures of low quality will contain regions in this matrix with high values indicating there are regions that Alphafold did not know how to place respectively to the rest of the structure. These high-value regions typically encode for a disordered region. However, if the area of high error is large enough, it indicates Alphafold doesn't know how to place amino acids from a tertiary structure. The predicted aligned error matrix can pinpoint both low scoring and high scoring PLDDT regions that are inconsistent with the overall tertiary structure of the protein

5 CONCLUSION

The described methods represent a first step towards leveraging protein structure and sequence alone to describe and predict microbial virulence broadly. We showed that, when high-quality protein structures are available, a structure-based approach performs better than a state-of-the-art deep learning sequence based model. However, some FunSoCs need further investigation to increase performance. Future work will be conducted to gather higher-quality structures. understand what substructures correlate to promoters and deterrents to associate protein function and virulence.

6 ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Krista Ternus' and Dr. Gene Godbold's numerous contributions and insights into functions of sequences of concern (FunSoCs) and microbial virulence factors. We would also like to thank Dr. Advait Balaji for discussions related to SeqScreen. T.J.T. was supported in part by a NSF CAREER award (IIS-2239114) and a NIH/NIAID grant (P01-AI152999). LEK is supported in part by NIH U01CA258512.

REFERENCES

[1] B. Amos, C. Aurrecoechea, M. Barba, A. Barreto, E. Y. Basenko, W. Bażant, R. Belnap, A. S. Blevins, U. Böhme, J. Brestelli, B. P. Brunk, M. Caddick, D. Callan, L. Campbell, M. B. Christensen, G. K. Christophides, K. Crouch, K. Davis, J. DeBarry, R. Doherty, Y. Duan, M. Dunn, D. Falke, S. Fisher, P. Flicek, B. Fox, B. Gajria, G. I. Giraldo-Calderón, O. S. Harb, E. Harper, C. Hertz-Fowler, M. J. Hickman, C. Howington, S. Hu, J. Humphrey, J. Iodice, A. Jones, J. Judkins, S. A. Kelly, J. C. Kissinger, D. K. Kwon, K. Lamoureux, D. Lawson, W. Li, K. Lies, D. Lodha, J. Long, R. M. MacCallum, G. Maslen, M. A. McDowell, J. Nabrzyski, D. S. Roos, S. S. C. Rund, S. W. Schulman, A. Shanmugasundram, V. Sitnik, D. Spruill, D. Starns, C. J. Stoeckert, S. S. Tomko, H. Wang, S. Warrenfeltz, R. Wieck, P. A. Wilkinson, L. Xu, and J. Zheng. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research*, 50(D1):D898–D911, Oct. 2021.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[3] C. Aurrecoechea, A. Barreto, E. Y. Basenko, J. Brestelli, B. P. Brunk, S. Cade, K. Crouch, R. Doherty, D. Falke, S. Fischer, B. Gajria, O. S. Harb, M. Heiges, C. Hertz-Fowler, S. Hu, J. Iodice, J. C. Kissinger, C. Lawrence, W. Li, D. F. Pinney, J. A. Pulman, D. S. Roos, A. Shanmugasundram, F. Silva-Franco, S. Steinbiss, C. J. Stoeckert, D. Spruill, H. Wang, S. Warrenfeltz, and J. Zheng. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Research*, 45(D1):D581–D591, Nov. 2016.

[4] A. Bairoch. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, Jan. 2000.

[5] A. Balaji, B. Kille, A. D. Kappell, G. D. Godbold, M. Diep, R. A. L. Elworth, Z. Qian, D. Albin, D. J. Nasko, N. Shah, M. Pop, S. Segarra, K. L. Ternus, and T. J. Treangen. SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. *Genome Biology*, 23(1), June 2022.

[6] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, Feb. 2022.

[7] D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, and A. Elofsson. Towards a structurally resolved human protein interaction network. *Nature Structural &amp Molecular Biology*, 30(2):216–225, Jan. 2023.

[8] L. Chen. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(Database issue):D325–D328, Dec. 2004.

[9] T. U. Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020.

[10] R. A. L. Elworth, C. Diaz, J. Yang, P. de Figueiredo, K. Ternus, and T. Treangen. Synthetic DNA and biosecurity: Nuances of predicting pathogenicity and the impetus for novel computational approaches for screening oligonucleotides. *PLOS Pathogens*, 16(8):e1008649, Aug. 2020.

[11] Y. Geffen, Y. Ofran, and R. Unger. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, 38(Supplement_2):ii95–ii98, Sept. 2022.

[12] V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho, and R. Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1), May 2021.

[13] J. Howard and S. Ruder. Universal language model fine-tuning for text classification, 2018.

[14] R. Jacak, J. Proescher, G. Godbold, A. Ernlund, and T. Zudock. PathGO: The Pathogenesis Gene Ontology.

[15] B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

[16] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021.

[17] A. Kessel and N. Ben-Tal. *: Structure, Function, and Motion, Second Edition*. Chapman and Hall/CRC, New York, 2 edition, Mar. 2018.

[18] M. Kulmanov, M. A. Khan, and R. Hoehndorf. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, Oct. 2017.

[19] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar. 2023.

[20] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, Jan. 2023.

[21] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, May 2022.

[22] G. B. Oliveira, H. Pedrini, and Z. Dias. TEMPROT: protein function annotation using transformers embeddings and homology search. *BMC Bioinformatics*, 24(1), June 2023.

[23] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. Recognition of functional sites in protein structures. *Journal of Molecular Biology*, 339(3):607–633, June 2004.

[24] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, Nov. 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] A. R. Wattam, J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, S. Gerdes, C. S. Henry, R. W. Kenyon, D. Machi, C. Mao, E. K. Nordberg, G. J. Olsen, D. E. Murphy-Olson, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, V. Vonstein, A. Warren, F. Xia, H. Yoo, and R. L. Stevens. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research*, 45(D1):D535–D542, Nov. 2016.

[27] Q. Yuan, J. Xie, J. Xie, H. Zhao, and Y. Yang. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings in Bioinformatics*, 24(3), Mar. 2023.

A SUPPLEMENTARY MATERIAL

Table S1: Functional Sequence of Concern Categorical Structures

FunSoC Name	Number of Positive Samples
Antibiotic Resistance	1744
Avirulence Plant	48
Bacterial Counter Signaling	31
Counter Immunoglobin	7
Cytotoxicity	511
Degrade ECM	391
Development In Host	73
Disable Organ	2294
Host Cell Cycle	162
Host Cell Death	307
Host Cytoskeleton	17
Host GTPase	36
Host Transcription	176
Host Translation	70
Host Ubiquitin	20
Host Xenophagy	44
Induce Inflammation	50
Invasion Hostcell Viral	543
Nonviral invasion	58
Plant RNA Silencing Viral	58
Resist Host Complement	32
Resist Oxidative	46
Secreted Effector	67
Secretion	494
Suppress Detection	61
Toxin Synthase	439
Viral Adhesion	459
Viral Counter Signaling	545
Viral Movement	68
Virulence Activity	92
Virulence Regulator	300
Nonviral Adhesion	58