

Exploring the biological functions of uncharacterized viral proteins

Ana Gagunashvili

*The thesis has been submitted to the Faculty of Natural Sciences and Medicine of
Ilia State University following the requirements for obtaining the degree of Master of Science.*

Program: Applied Genetics

Research supervisor: Vincenzo Lagani, Ph.D.

Ilia State University

Tbilisi, 2026

Statement

To the Dean of Faculty of Natural Sciences and Medicine LEPL Ilia State University

MS. Elena Zhuravliova

Faculty of Natural Sciences and Medicine

Program: Applied Genetics

Master's thesis student: Ana Gagunashvili

As the author of the submitted thesis, I declare that the thesis is my original work and does not contain materials previously published, accepted, or defined by other authors, which are not mentioned or cited according to citation rules.

Ana Gagunashvili 

15.06.2026

.

Abstract

Advances in metagenomic sequencing have revealed an enormous diversity of previously unknown viruses. However, the majority of proteins encoded by these viral genomes remain functionally uncharacterized. Traditional annotation methods based on sequence similarity often fail to assign reliable functions to highly divergent viral proteins, necessitating alternative computational approaches.

This study evaluated the use of Deep Functional Residue Identification (DeepFRI), a deep learning-based protein function prediction framework, for the large-scale annotation of viral proteins. Protein sequences derived from the MetaVR database were analyzed to generate Gene Ontology Molecular Function predictions and associated confidence scores. Acquired Functional predictions were integrated with viral genome metadata, including environmental source, genome type, genome quality, host taxonomy, and genome topology, to investigate factors associated with prediction confidence and functions of viral proteins.

The majority of the predicted functions were assigned low confidence scores, with 79.29% classified as low confidence and only 5.44% as high confidence. This finding is consistent with the extensive presence of “viral dark matter” and highlights the challenges associated with viral protein annotation. High-confidence predictions were dominated by conserved molecular functions, including ATP binding, carbohydrate derivative binding, DNA binding, and structural constituent of the ribosome. These functions appeared repeatedly across multiple viral categories, suggesting that DeepFRI is most effective at identifying evolutionarily conserved biochemical activities.

Significant differences in prediction confidence were observed across environmental sources, genome types, host taxa, genome quality groups, and genome topologies, with viral genome type and host taxonomy exhibiting the most pronounced variation. Nevertheless, similar functional categories were consistently identified across diverse viral populations. Furthermore, genome

length and the number of coding sequences exhibited only weak relationships with prediction confidence, indicating that genome complexity alone does not account for the difficulty of viral protein annotation.

Overall, this study demonstrates the potential of deep learning approaches for large-scale exploration of viral protein function while emphasizing the continuing challenge posed by “viral dark matter”. The findings contribute to ongoing efforts to improve functional annotation of viral proteins and provide insights into the distribution of molecular functions across diverse viral genomes.

Keywords: Viral protein functions, UViG, MetaVR DeepFRI, Deep learning, GO terms, EC Terms,

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Professor Vincenzo Lagani, Ph.D., for his invaluable guidance, support, and encouragement throughout this research project. His expertise, professionalism, and extensive knowledge were instrumental in shaping this work and helping me navigate the challenges encountered during the research process.

I am especially grateful for his patience, willingness to help, and continuous support, particularly when technical and computational difficulties arose. His organized approach to research, attention to detail, and commitment to scientific excellence provided an inspiring example.

Beyond his academic guidance, I greatly appreciate the understanding and encouragement he consistently demonstrated. Working under his supervision has been both a privilege and a valuable learning experience. The qualities he embodies as a scientist, mentor, educator, and human being are qualities that I hope to develop throughout my life.

I would also like to sincerely thank everyone who contributed to my academic journey and supported me during the completion of this thesis.