A pan-viral map of host dependency factors from multi-omics integration and machine learning across Influenza A, SARS-CoV-2, Zika, and Dengue viruses

Mohadeseh Naseri, André Dietz, Alicia Hiemisch, Marcus Oswald, Rainer Koenig

Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany.
Correspondence: rainer.koenig@uni-jena.de

Identifying Host Dependency Factors

- **What are Host Dependency Factors?**

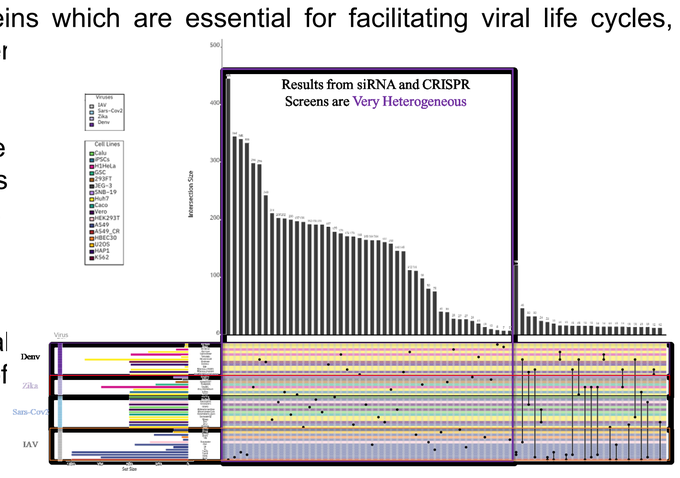
Host Dependency Factors (HDF) are proteins which are essential for facilitating viral life cycles, presenting promising targets for antiviral inter

- **Faced problems to solve:**

However, current methodologies, primarily re on CRISPR/Cas[1-12] or RNA interference-bas gene perturbations[13-15] yield heterogeneous data with limited consensus.

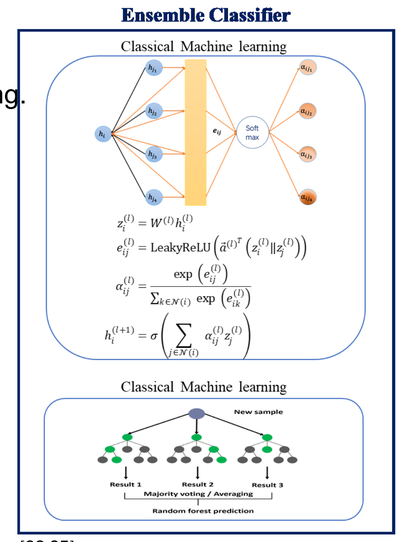
- **Our Aim:**

Our project aims to establish a computational pipeline for the identification and validation of antiviral targets against these viruses.



Model Construction

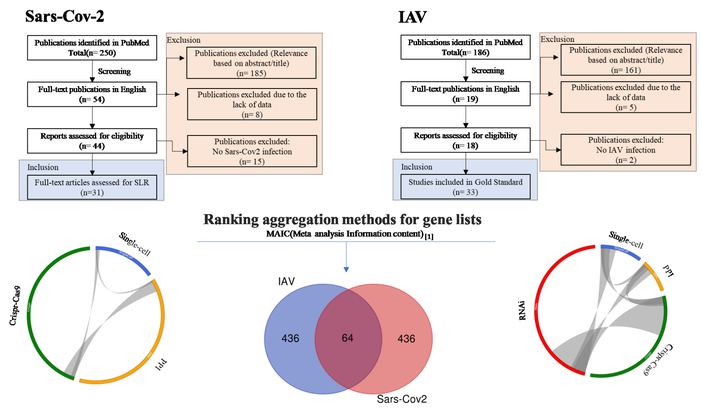
- We propose an the combination of Graph Attention Network (GAT) and Random Forest and for node classification forecasting.
- GAT is a neural network designed for graph data, utilizing attention mechanisms to weight the importance of neighboring nodes during information aggregation.
- Given the essential role of the interactions between viral proteins and human cells, we employed the GAT.
- Compared to various classical machine learning methods (XGB, SVM, and Decision Trees), Random Forest was superior in initial trials, leading us to choose it for combining with GAT.



[23-25]

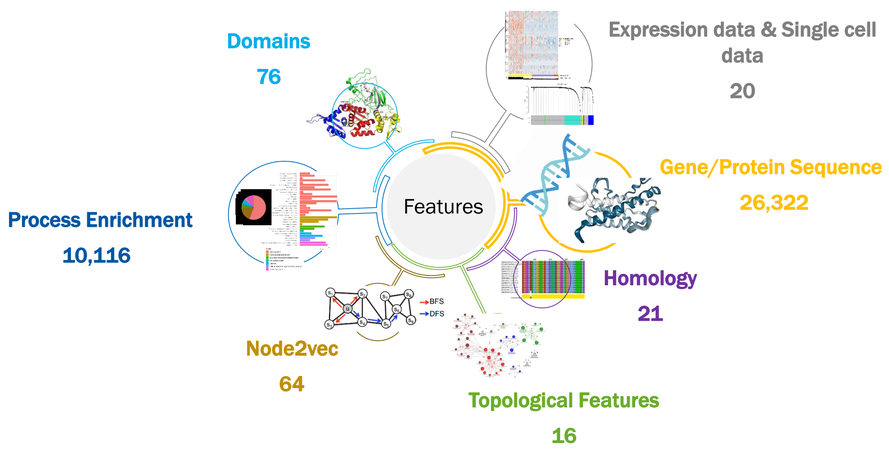
Step I: Setting up the Gold Standard

- Due to the high heterogeneity of th CRISPR data, we incorporated additic sources, such as protein interactions[1 and single-cell RNA-seq [20-21] to enha the gold standard.
- We applied the MAIC[22] (Meta Analysis Information Content) method to aggregate gene lists, selecting the top 500 genes as HDFs and the bottom 1,000 as non-HDFs.



1. Li, Bo, et al. "Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection." Nature communications 11.1 (2020): 164.

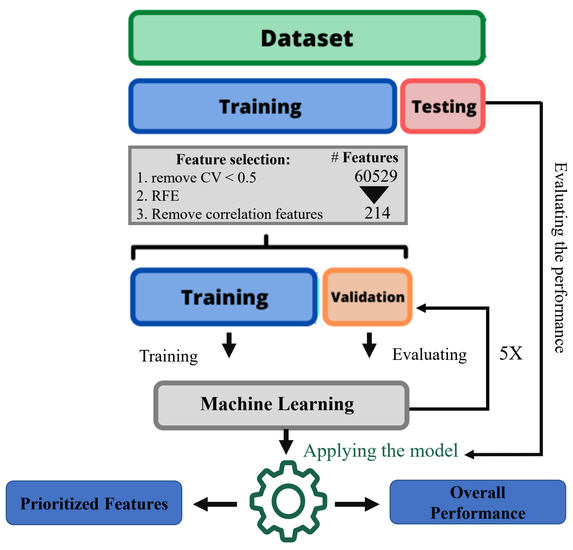
Step II: Assembling the Features



Features for HDF prediction were assembled from various resources. The values indicate the number of features for machine learning.

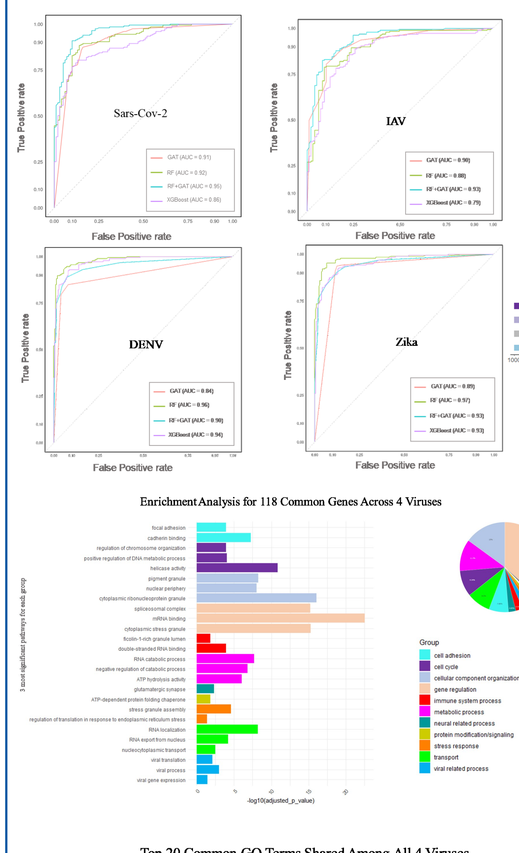
Step III: Machine Learning Pipeline

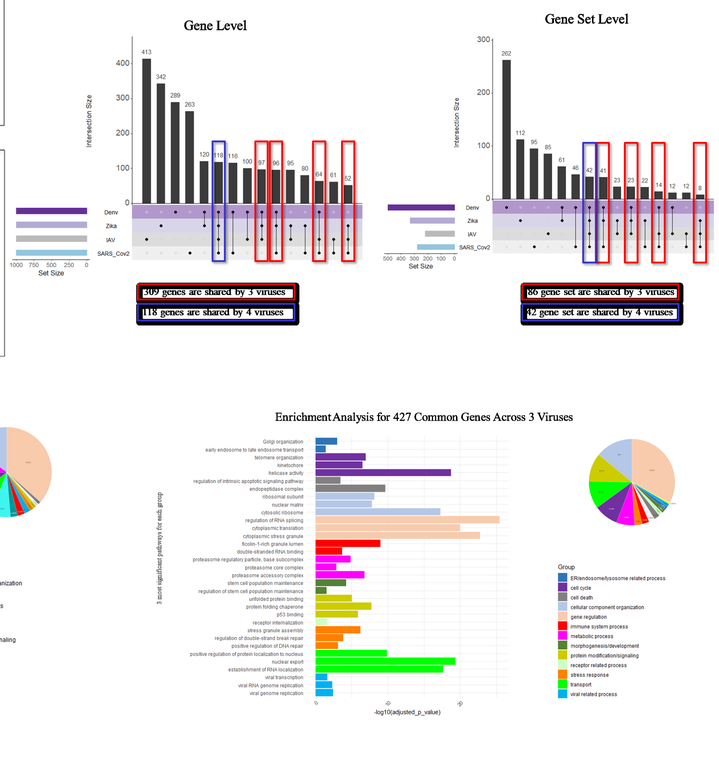
- Utilized 5-fold cross-validation, dividing data into 4/5 training and 1/5 testing sets, ensuring the test set remained to be unseen.
- Repeated the process five times and averaged the results.
- Initiated feature selection with Recursive Feature Elimination, removing highly correlated features.
- Features underwent z-score normalization and modified variance assessment.
- Class imbalances were addressed using the Synthetic Minority Over-sampling Technique.
- Evaluated model performance using the test data.

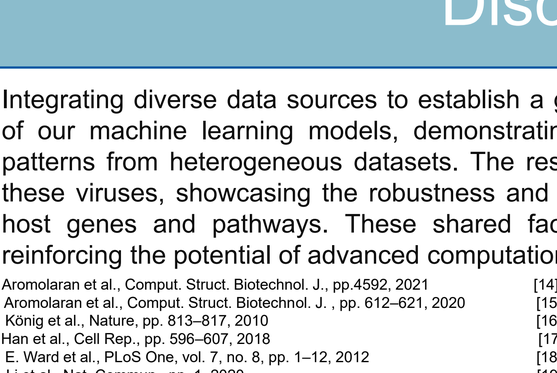


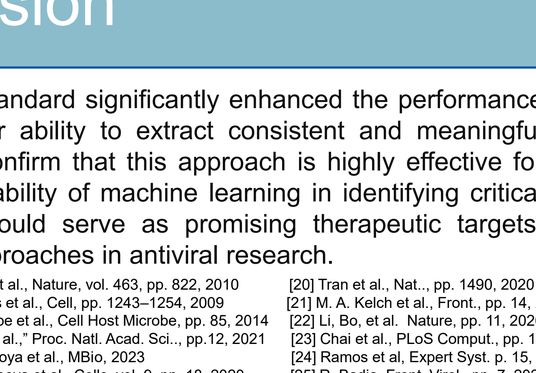
Results

- We defined the top 5% of predictions from the combined GAT and RF model as predicted HDFs, resulting in a total of 1000 predicted HDFs.
- RF+GAT Shows the Best Performance: This model achieves the highest AUC values, indicating improved predictive power.
- The observed increase in shared genes and gene sets highlights the robustness of our approach, confirming that host dependency factors are more conserved than previously identified.
- These analyses reveal that viruses exploit common host pathways, particularly those related to RNA processing, stress response, and viral translation, while also showing virus-specific variations.









Discussion

Integrating diverse data sources to establish a gold standard significantly enhanced the performance of our machine learning models, demonstrating their ability to extract consistent and meaningful patterns from heterogeneous datasets. The results confirm that this approach is highly effective for these viruses, showcasing the robustness and applicability of machine learning in identifying critical host genes and pathways. These shared factors could serve as promising therapeutic targets, reinforcing the potential of advanced computational approaches in antiviral research.

References

[1] M. Grodzki et al.," Genome Med., pp. 1–18, 2022
[2] W. M. Schneider et al., Cell, pp. 120-132.e14, 2021
[3] J. Baggen et al., Nat. Genet., pp. 435–444, 2021
[4] M. Israeli et al., Nat. Commun., pp. 1, 2022
[5] D. E. Gordon et al., Nature, pp. 459–468, 2020
[6] S. Zhu et al., Sci. China Life Sci. , pp. 701–717, 2022
[7] Z. Daniloski et al., Cell, pp. 92-105.e16, 2021
[8] J. Wei et al., Cell, pp. 76-91.e13, 2021
[9] Y. Zhu et al., 2021, pp. 1–11
[10] S. B. Biering et al., Nat. Genet., pp. 1078–1089, 2022
[11] R. Wang et al., Cell, pp. 106-119.e14, 2021
[12] A. Rebendenne et al., Nat. Genet., pp. 1090–1102, 2022
[13] O. Aromolaran et al., Comput. Struct. Biotechnol. J., pp.4592, 2021
[14] O. Aromolaran et al., Comput. Struct. Biotechnol. J. , pp. 612–621, 2020
[15] R. König et al., Nature, pp. 813–817, 2010
[16] J. Han et al., Cell Rep., pp. 596–607, 2018
[17] S. E. Ward et al., PLoS One, vol. 7, no. 8, pp. 1–12, 2012
[18] B. Li et al., Nat. Commun., pp. 1, 2020
[14] A. Karlas et al., Nature, vol. 463, pp. 822, 2010
[15] A. L. Brass et al., Cell, pp. 1243–1254, 2009
[16] T. Watanabe et al., Cell Host Microbe, pp. 85, 2014
[17] Whelan et al.," Proc. Natl. Acad. Sci.., pp.12, 2021
[18] V. R. Montoya et al., MBio, 2023
[19] V. A. Golubeva et al., Cells, vol. 9, pp. 18, 2020
[20] Tran et al., Nat.., pp. 1490, 2020
[21] M. A. Kelch et al., Front., pp. 14, 2023
[22] Li, Bo, et al. Nature, pp. 11, 2020
[23] Chai et al., PLoS Comput., pp. 1,2022
[24] Ramos et al, Expert Syst. p. 15, 2022
[25] R. Badia, Front. Virol., pp. 7, 2022

Funded by the European Union under GA No 101137311.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.