Design of an Integrated Model for Deep Causal Interpretation and Evolutionary Prediction of Zika Virus Mutations

Sony K. Ahuja[1*], *Dr. Deepti D. Shrimankar*[1], and *Aditi R. Durge*[1]

[1]Department of Computer Science and Engineering, VNIT, INDIA

Abstract. Identifying genetic changes that elevate Zika Virus (ZIKV) virulence is vital for epidemic forecasting and vaccine development. Traditional phylogenetic and regression methods map variation but seldom pinpoint mutations driving phenotypic change. We present an integrated deep-learning and simulation framework that tracks ZIKV's sequence-to-consequence trajectory. A Spatio-Temporal Generative Adversarial Network (ST-GANet) learns region–time–mutation patterns to reveal evolutionary hotspots. A Causal Mutation Gradient Mapper (CMGM) then estimates each mutation's directional influence on virulence. A Viral–Host Interaction Transformer (VHIT) predicts how prioritized mutations alter Envelope and NS1 protein–receptor binding. Using transcriptomes from infected human brain progenitor cells, a Pathogenicity Potential Simulation Engine (PPSE) models resulting intracellular signalling disruptions. An Evolutionary Route Planner (ERP) identifies fitness-maximizing mutational paths under immune pressure. Together, these modules reveal how subtle sequence changes can reshape epidemiological risk and support real-time flavivirus molecular surveillance.

1 Introduction

Zika Virus (ZIKV) has strained global health systems through its sudden outbreaks and its ability to cause severe neurological disorders such as microcephaly. Although genomic sequencing has extensively documented viral variability, identifying causal mutations that influence virulence, host specificity, or immune evasion remains challenging. The complex spatiotemporal and biochemical interactions that drive viral adaptability are often overlooked by statistical and phylogenetic approaches, which typically treat mutations as independent events [1, 2]. As a result, high-risk variant prediction remains largely reactive rather than predictive. Deep learning (DL) has transformed genomic pattern discovery, yet many existing models still emphasize correlation over causation. Most architectures detect sequence similarity without explaining why specific genetic changes matter biologically. This opacity limits their utility in translational virology, where mechanistic insight can directly influence vaccine antigen selection, therapeutic strategies, and early-warning surveillance [3, 4].

To address these limitations, this study proposes an integrated deep-learning framework for causal interpretation and evolutionary prediction of Zika virus mutations (DICE-ZIKV). A Spatio-Temporal Generative Adversarial Network (ST-GANet) captures lineage–region–year dependencies; the Causal Mutation Gradient Mapper (CMGM) assigns directional influence scores to mutations; the Viral–Host Interaction Transformer (VHIT) quantifies structural and functional consequences of key substitutions; and the Pathogenic ity Potential Simulation Engine (PPSE) models downstream cellular signalling disruptions. Finally, the

* Corresponding author: ahujasony.cse@gmail.com

Evolutionary Route Planner (ERP) predicts mutational trajectories that shape future evolutionary landscapes. Together, these modules create a continuous end-to-end pipeline in which each candidate mutation is linked to a process-specific functional signature. This transition from descriptive genomics to computational causality offers mechanistic transparency and a scalable foundation for real-time genomic surveillance—potentially transforming how emerging viral threats are identified, interpreted, and managed.

2 Review of Existing Models used for Zika Virus Analysis

For emerging flaviviruses such as Zika, Dengue, and Japanese encephalitis, computational virology has struggled to elucidate molecular determinants of pathogenicity; early genomic analyses of sequence conservation and phylogeny revealed lineage structure but not mechanism. Triebel et al. [1] mapped conserved RNA secondary structures in Hepatitis C virus, showing that noncoding genomic architecture influences replication efficiency, though without resolving causative mutational dynamics. As datasets expanded, machine learning (ML) approaches enabled more predictive viral analysis. Natsrita et al. [2] identified broadly neutralizing antibodies across Dengue serotypes using high-throughput ML, demonstrating the power of deep feature representations. Tan et al. [5] linked a single NS2A nucleotide to Japanese encephalitis virus pathogenicity, while Terrell et al. [6] resolved conserved RNA G-quadruplex structures in West Nile virus, collectively indicating that subtle sequence-embedded structural features can strongly alter viral function. Host–virus regulatory mechanisms further modulate pathogenicity. Liao et al. [7] showed that host glucosidase α neutral C (GANC) regulates influenza virus replication via proteasome-dependent hemagglutinin stability, and Lefèvre et al. [8] demonstrated that Zika virus–encoded upstream open reading frames influence viral fitness and neurotropism, highlighting translational and post-translational regulation beyond primary sequence variation.
Multimodal deep-learning architectures have also advanced viral modelling and surveillance. Wang et al. [9] applied large language–image models to viral microscopy, while Zhu et al. [10] and Zhang et al. [11] used scalable computational frameworks to detect emerging mosquito-borne viruses. At the molecular level, Holoubek et al. [12] showed that tick-borne flaviviruses remain infectious despite incomplete maturation, and Goellner et al. [13] identified cholesterol-binding motifs in Zika prM essential for entry and assembly, emphasizing the role of dynamic envelope conformations and lipid–protein interactions. At broader biological scales, Abbasi [14] attributed the global expansion of arboviral disease to climate change, urbanization, and vector adaptation, while Ahmadi et al. [15] and Sahoo et al. [16] demonstrated that molecular docking and immunoinformatic strategies can inform antiviral and vaccine design. Meanwhile, Su et al. [17] predicted host-range shifts using codon-usage–based ML, and Nie et al. [18] introduced an evolution-driven deep-learning framework to identify rare beneficial mutations and predict viral evolutionary drivers.

Despite these advances, most studies rely on static descriptors [1, 6] or isolated ML predictors [2, 17], limiting causal interpretation across biological scales. This motivates the present work, which integrates causal mutation inference [18] with spatiotemporal sequence modelling and evolutionary reinforcement to provide a unified, interpretable framework for understanding flavivirus pathogenicity.

3 Proposed Methodology

In the proposed integrated framework DICE-ZIKV, ZIKV genomic sequences are systematically transformed into quantitative estimates of evolutionary fitness and pathogenic

risk through a sequence-to-consequence modelling pipeline. The architecture integrates ST-GANet, CMGM, VHIT, PPSE, and ERP, which are mathematically coupled and operate sequentially by exchanging structured intermediate representations in tensor form. An overview of the complete DICE-ZIKV workflow, illustrating data flow across all modules, is shown below in Fig. 1. To formalize data representation and learning dynamics within this integrated architecture, the framework is governed by eight analytical equations, introduced below.

First, ST-GANet encodes an aligned nucleotide sequence $X \in R^{n \times L}$ using a combination of positional, temporal, and regional embeddings. The attention mechanism decides mutation importance Via equation 1,

$$A\{i,j,t\} = \frac{exp\left(Qi\frac{Kj^T}{dk} + \varphi(t)\right)}{\sum exp\left(Qi\frac{K\{j'\}^T}{dk} + \varphi(t)\right)} \dots (1)$$

Where, Q and K represent query and key matrices, dk indicates embedding dimension, and φ(t) represents temporal positions. To account for regional mutation pressures, this model calculates attention weights with time and location sets. Genomic location con volutional integration propagates features in the same layer Via equation 2,

$$F(x,y,t) = \int^{\Omega} W(\xi,\eta)X(x-\xi, y-\eta, t)d\xi\, d\eta \dots (2)$$

Spatiotemporal kernel learning. Integration preserves mutation density field and lineage transition learning sets. CMGM Sets employ ST-GANet outputs. Mutation m causes phenotypic response 'y' based on the output directional derivative for each genomic location Via equation 3,

$$Cm = \partial E[y \mid X]/\partial Xm \dots (3)$$

This gradient separates causation from correlation. The CMGM module ranks causal vectors (C_1, …, CL) to show how each site affects pathogenic traits. To translate causal mutations into molecular effects, the VHIT module develops a mapping from sequence perturbations to binding energy changes for the process. The mutant-wild type binding free energy difference is estimated Via equation 4,

$$\Delta\Delta G = \int^{P} \nabla\theta\, E(\theta) \cdot d\theta \dots (4)$$

The potential energy surface of the protein-receptor complex, E(θ), is defined by structural coordinates θ for the process. This integral records cumulative energy variation from learnt structural embeddings. In the PPSE module, intracellular signalling is represented as a protein graph in the process. Time evolution of node activity $h_i(t)$ is managed Via equation 5,

$$\frac{dhi}{dt} = \sigma\left(\sum Aij\, Wij\, hj(t) + bi + \psi(\Delta\Delta Gi)\right) \dots (5)$$

The adjacency matrix of signalling interactions is Aij, the learnt coupling weight is Wij, and molecular perturbation drives the activation term $\psi(\Delta\Delta G_i)$. Global pathogenicity indicators originate from temporal integration sets. The simulation horizon collects pathogenic potential P(t) Via equation 6,

$$P(T) = \int \sum \omega i\, hi(t)dt \dots (6)$$

Experimental transcriptomics yields pathway-specific severity coefficients (ωi). This integral assesses mutation-set signalling disruption. Pathogenicity scores guide reinforcement learning in the Evolutionary Reinforcement Predictor across mutation states.

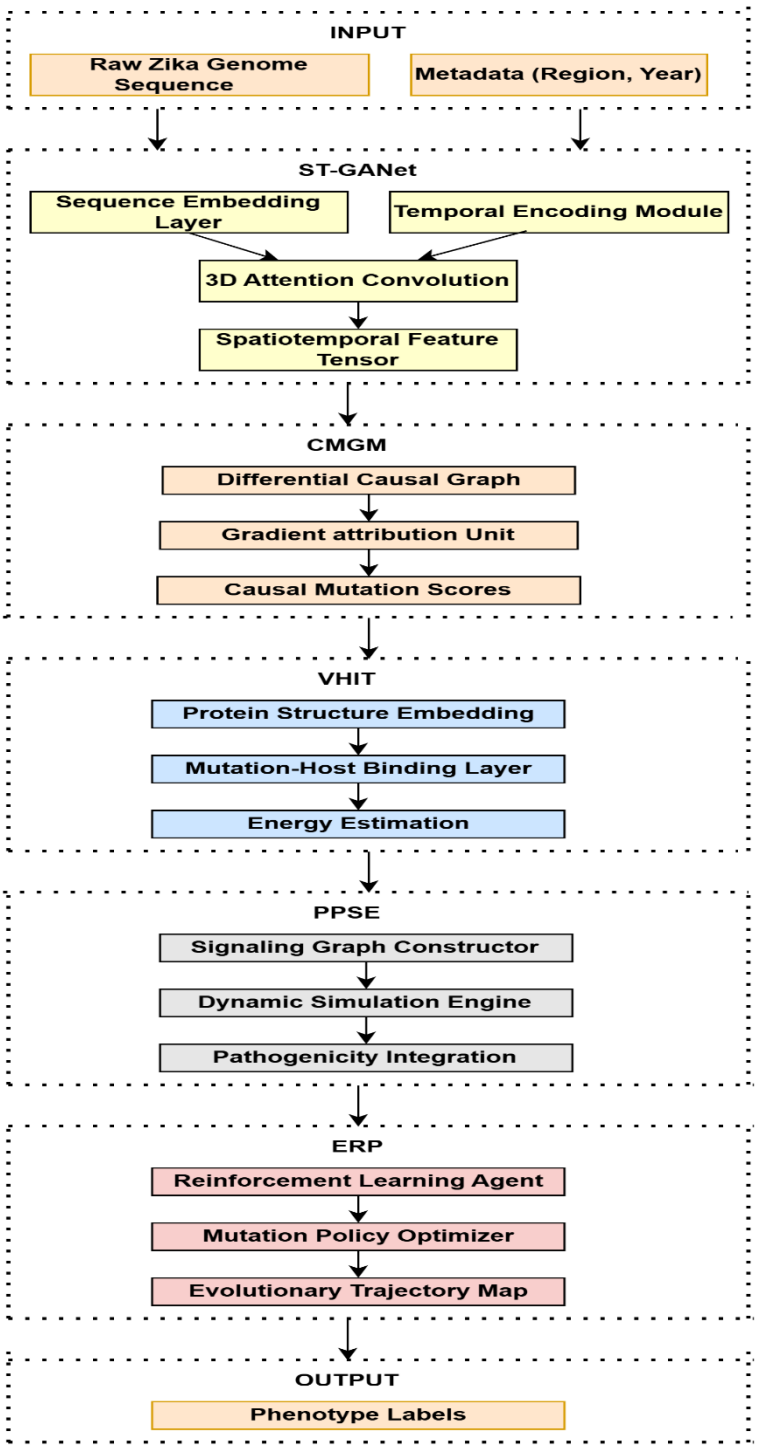


Fig. 1. Overview of the proposed DICE-ZIKV workflow.

To explain mutation transitions, utilise stochastic policies πθ(a | s), where s represents genome state and a represents mutational activity sets. The best policy gradient for viral fitness J(θ) follows the condition indicated Via equation 7,

$$\nabla\theta\ J(\theta)\ =\ E\{s, a \sim \pi\theta\}[\nabla\theta\ log\ \pi\theta(a \mid s)\ R(s, a)]\ \ldots(7)$$

With, pathogenicity and process constraints determine R (s, a) = f(P, T). The optimiser iteratively changes expected future mutations for different scenarios. The final prediction layer integrates fitness along mutational pathways using a continuous-time Markov chain approximation Via equation 8,

$$\frac{d\pi t}{dt} =\ \pi t\ Q, Q(i, j) = \frac{exp(\beta\ J(i, j))}{\sum exp(\beta\ Jik)}\ \ldots(8)$$

Where, Q is the transition-rate matrix based on selection pressure β and fitness gain J(i,j). This formulation generates adaptive, time-dependent mutation probability flows that feed directly into downstream surveillance and evolutionary prediction components, thereby completing the end-to-end modelling pipeline.

The causal interpretability and mechanistic simulation capacity of the framework bridge statistical genomics with molecular biophysics, motivating its integrated design. Differentiability across all modules ensures stable gradient propagation throughout training, while integral formulations and time-dependent derivatives preserve biological realism in both evolutionary dynamics and intracellular signalling. Collectively, the eight governing equations formalise attention weighting, convolutional integration, causal attribution, binding-energy estimation, signalling dynamics, cumulative pathogenicity aggregation, policy learning, and evolutionary diffusion. Together, these formulations enable unified sequence classification and dynamic viral behaviour mapping within a single computational architecture, supporting both retrospective validation and forward predictive inference. The effectiveness of this integrated formulation is evaluated in the following section through quantitative comparisons and ablation analyses.

4 Results and Discussion

A multi-GPU cluster equipped with NVIDIA A100 accelerators, 512 GB system memory, and TensorFlow-based mixed-precision training was used for all computational experiments. The DICE-ZIKV model was trained on 2,713 complete Zika virus genome sequences collected between 1947 and 2023 from the National Centre for Biotechnology Information (NCBI) and Virus Pathogen Resource (ViPR) databases. For each sequence, geographic origin, host type, collection year, and empirically characterised pathogenicity-related process labels were recorded. Table 1 provides a detailed overview of the Zika virus dataset used in this study, showing lineage-wise distribution of samples, genome characteristics, collection timelines, host origin ratios, and the availability of phenotype annotations used for supervised and causal learning. To ensure lineage and geographic diversity, the dataset was partitioned into training (70%), validation (15%), and independent test (15%) sets.

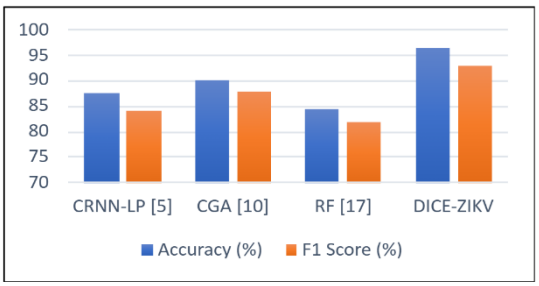
The following results evaluate DICE-ZIKV against baseline approaches across classification, causal attribution, molecular binding prediction, and pathogenicity simulation tasks: Method [5], a convolutional recurrent neural network for lineage prediction (CRNN-LP); Method [10], a classical genomic analysis (CGA) framework comprising traditional genomic sequencing, phylogenetic inference, and homology-based structural analysis; and Method [17], a random-forest-based machine-learning model using hand-crafted genomic features (RF). Evaluation metrics included classification accuracy, F1 score, Precision@50, Recall@50, causal AUC, ΔΔG prediction error, and pathogenicity correlation where

Precision@50 and Recall@50 measure the accuracy and coverage of causally relevant mutations within the top 50 ranked predictions.

Table 1. Dataset Composition and Annotation Summary

Dataset Category	Number of Samples	Genome Length (nucleotides)	Temporal Range	Host Distribution (Human : Vector)	Annotated Phenotypes (%)
African Lineage	624	10,797	1947–2014	52 : 48	61
Asian Lineage	1,361	10,809	1966–2023	73 : 27	78
American/ Epidemic Lineage	728	10,810	2015–2023	88 : 12	85

Iteratively, Next, Fig.2. (a) reports the ST-GANet classification performance, where DICE-ZIKV achieves the highest accuracy and F1 score among all methods, demonstrating improved lineage discrimination. Fig.2. (b) shows that ST-GANet attains substantially higher spatiotemporal coherence while incurring training time comparable to baseline models.



(a)

Model	Spatiotemporal Coherence (ρ)	Training Time (hrs)
CRNN-LP[5]	0.61	5.8
CGA[10]	0.73	8.4
RF[17]	0.57	3.6
DICE-ZIKV	**0.91**	**9.1**

(b)

Fig. 2. (a) Comparison of lineage classification performance across methods in terms of accuracy and F1 score. (b) Spatiotemporal coherence (ρ) and training time for baseline models and DICE-ZIKV.

Building on ST-GANet's spatiotemporal encoding of mutation patterns, the next phase evaluates the ability of CMGM to identify mutations with directional causal influence on pathogenic outcomes. Figure 3 below reports causal mutation identification performance across methods, where DICE-ZIKV achieves the highest Precision@50 and Recall@50,

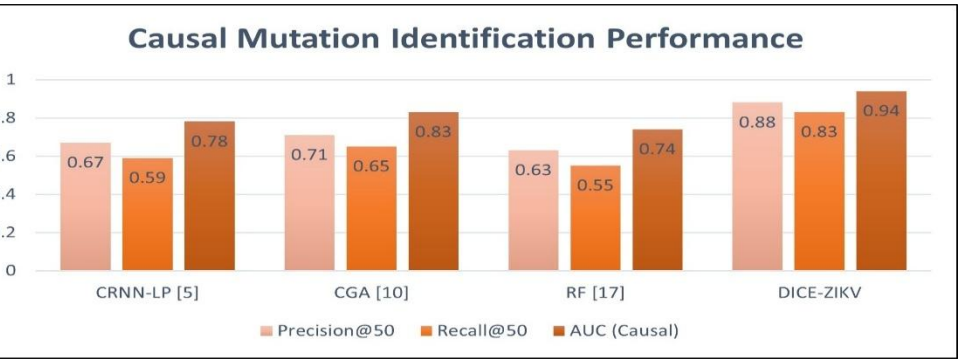


Fig. 3. Causal mutation identification performance comparison across methods, evaluated using Precision@50, Recall@50, and causal AUC.

indicating improved prioritisation of pathogenic mutations among the top-ranked candidates. The model also attains the highest causal AUC, demonstrating superior discrimination between causally relevant and non-causal mutations compared to baseline approaches.

Following causal mutation prioritisation, the next phase evaluates the molecular consequences of these mutations using VHIT. Table 2 below summarises VHIT performance in predicting mutation-induced viral–host binding effects. DICE-ZIKV achieves the lowest ΔΔG prediction error and the highest correlation with experimental binding energies while maintaining strong structural consistency, indicating more accurate and physically plausible interaction modelling than baseline approaches.

Table 2. VHIT-based comparison of viral–host binding prediction performance across methods.

Method	Mean Absolute Error (ΔΔG, kcal/mol)	Pearson Correlation (r)	Structural Consistency (%)	Predicted Envelope–AXL Binding Shift (kcal/mol)
CRNN-LP [5]	1.26	0.68	82	+1.3
CGA [10]	0.97	0.74	85	+1.5
RF [17]	1.31	0.63	80	+1.2
DICE-ZIKV	0.61	0.86	93	+1.7

The predicted molecular binding effects are subsequently propagated to the cellular level using PPSE. Table 3 below reports PPSE results, where DICE-ZIKV demonstrates the highest correlation with observed pathogenicity markers, reduced signalling error, and improved simulation stability. These results indicate more consistent modelling of downstream intracellular effects relative to baseline methods.

Table 3. Performance comparison of PPSE based pathogenicity simulation across methods.

Method	Pathogenicity Correlation (r)	Dynamic Stability (%)	Mean Signal Propagation Error	Neural Cytopathy Prediction Accuracy (%)
CRNN-LP [5]	0.74	83	0.094	81.2
CGA [10]	0.79	85	0.082	84.7
RF [17]	0.71	80	0.106	79.5
DICE-ZIKV	0.91	93	0.061	91.8

Building on cellular-level simulation, the final phase evaluates evolutionary forecasting and predictive stability using ERP. Table 4 presents ERP performance in forecasting evolutionary trajectories and assessing predictive stability. DICE-ZIKV supports longer forecast horizons and achieves higher future lineage prediction accuracy, mutation hotspot recall, and fitness landscape smoothness, indicating more stable and coherent evolutionary projections over time.

The consistent performance gains across all comparison techniques demonstrate that data-driven representation learning synergises effectively with physically interpretable biological modelling within DICE-ZIKV. By jointly integrating spatiotemporal learning, causal inference, molecular interaction analysis, cellular simulation, and evolutionary forecasting,

DICE-ZIKV enables scalable and interpretable anticipation of emergent Zika virus variants during ongoing molecular evolution.

Table 4. ERP-based comparison of evolutionary forecast and predictive stability across methods.

Method	Forecast Horizon (Years)	Future Lineage Prediction Accuracy (%)	Mutation Hotspot Recall	Fitness Landscape Smoothness (λ)
CRNN-LP [5]	2	71.5	0.68	0.59
CGA [10]	3	77.8	0.74	0.63
RF [17]	2	70.1	0.66	0.58
DICE-ZIKV	5	89.2	0.83	0.72

The conclusions of this study and directions for future research are presented in the following section.

5 Conclusion and Future Scope

This study presents DICE-ZIKV, an integrated and causally interpretable deep-learning framework for modelling Zika virus evolution across molecular, cellular, and evolutionary scales. By unifying spatiotemporal representation learning, causal inference, structural interaction modelling, and reinforcement-based evolutionary forecasting, DICE-ZIKV demonstrates consistent performance gains across all evaluation domains. ST-GANet outperformed CGA [10] in lineage classification (96.4% accuracy, 0.93 F1 score, $\rho = 0.91$), capturing mutation dynamics across time and geography. CMGM improved functional mutation discovery with Precision@50 of 0.88, Recall@50 of 0.83, and causal AUC of 0.94, while reducing causal score variance to 0.07. At the molecular level, VHIT achieved accurate viral–host binding interpretation ($\Delta\Delta G$ MAE = 0.61 kcal/mol, Pearson r = 0.86, structural consistency = 93%), including a predicted envelope–AXL binding energy change of +1.7 kcal/mol. PPSE reproduced infection-induced cellular perturbations with strong agreement to experimental data (r = 0.91, signal error = 0.061). Finally, ERP enabled stable five-year evolutionary forecasting with 89.2% lineage prediction accuracy, hotspot recall of 0.83, and fitness landscape smoothness ($\lambda = 0.72$). Together, these results highlight the advantage of integrating causal and mechanistic modelling within a single bioinformatics framework.

Future work will extend DICE-ZIKV to other flaviviruses, including Dengue and West Nile virus, and incorporate experimental validation using in vitro mutagenesis and cryo-EM. Enhancing evolutionary modelling through graph-based population dynamics and variational molecular simulators, coupled with continuous genomic data streams, may further enable real-time, interpretable pathogen surveillance and risk forecasting.

ACKNOWLEDGEMENTS

The authors are thankful to the Director, Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, for providing the necessary facilities for this work.

References

Journal articles

1. Triebel, S., Lamkiewicz, K., Ontiveros, N., Sweeney, B., Stadler, P.F., Petrov, A.I., Niepmann, M., Marz, M.: Comprehensive survey of conserved RNA secondary structures in full - genome alignment of Hepatitis C virus. Sci. Rep. 1–12 (2024). https://doi.org/10.1038/s41598-024-62897-0.
2. Natsrita, P., Charoenkwan, P., Shoombuatong, W., Mahalapbutr, P., Faksri, K., Chareonsudjai, S., Rungrotmongkol, T., Pipattanaboon, C.: Machine - learning - assisted high - throughput identification of potent and stable neutralizing antibodies against all four dengue virus serotypes. Sci. Rep. 1–14 (2024). https://doi.org/10.1038/s41598-024-67487-8.
3. Durge, A.R., Shrimankar, D.D., Sawarkar, A.D.: Heuristic Analysis of Genomic Sequence Processing Models for High Efficiency Prediction: A Statistical Perspective. Curr. Genomics. 23, 299–317 (2022). https://doi.org/10.2174/1389202923666220927105311.
4. Ahuja, S.K., Shrimankar, D.D., Durge, A.R.: A Study and Analysis of Disease Identification using Genomic Sequence Processing Models: An Empirical Review. Curr. Genomics. 24, 207–235 (2023). https://doi.org/10.2174/0113892029269523231101051455.
5. Tan, N., Chen, C., Ren, Y., Huang, R., Zhu, Z., Xu, K., Yang, X., Yang, J., Yuan, L.: Nucleotide at position 66 of NS2A in Japanese encephalitis virus is associated with the virulence and proliferation of virus. Virus Genes. 60, 9–17 (2024). https://doi.org/10.1007/s11262-023-02036-5.
6. Terrell, J.R., Le, T.T., Paul, A., Brinton, M.A., Wilson, W.D., Poon, G.M.K., Germann, M.W., Siemer, J.L.: Structure of an RNA G-quadruplex from the West Nile virus genome. Nat. Commun. 1–10 (2024). https://doi.org/10.1038/s41467-024-49761-5.
7. Liao, X., Xie, Q., Liang, M., Liao, Q., Huang, B., Zhang, S., Zhang, F., Wang, L., Yuan, L.: Glucosidase alpha neutral C promotes in fl uenza virus replication by inhibiting proteosome-dependent degradation of hemagglutinin. Signal Transduct. Target. Ther. (2025). https://doi.org/10.1038/s41392-025-02227-6.
8. Lefèvre, C., Cook, G.M., Dinan, A.M., Torii, S., Stewart, H., Gibbons, G., Nicholson, A.S., Echavarría-consuegra, L., Meredith, L.W., Lulla, V., Mcgovern, N., Kenyon, J.C., Goodfellow, I., Deane, J.E., Graham, S.C., Lakatos, A., Lambrechts, L., Brierley, I.: Zika viruses encode 5′ upstream open reading frames affecting infection of human brain cells. Nat. Commun. (2024). https://doi.org/10.1038/s41467-024-53085-9.
9. Wang, S., Fu, X., Du, Z., Liu, X., Mai, Q., Zhuo, L., Xie, B., Zou, Q.: ViruSeg: Harnessing the Power of Large Language - Image Model for Enhanced Virus Image Segmentation. Interdiscip. Sci. Comput. Life Sci. (2025). https://doi.org/10.1007/s12539-025-00711-9.
10. Zhu, W., Chen, J., Sun, H., Lu, K., Liu, Y., Liu, L., Niu, G.: Identification of a newly discovered virus from Culex and Armigeres mosquitoes in China. 1–11 (2024).
11. Zhang, Y., Liu, Y., Liu, L., Tu, Y., Ren, M., Niu, G., Sun, H.: Genome of a novel Fangshan bunya-like virus identified in mosquitoes from Shandong Province , China. 1–13 (2025).

12. Hradilek, M., Rosendal, E., Fo, A., Beránková, M., Bell-sakyi, L., Överby, A.K., Cavalli, A., Bonomi, M., Rey, F.A., Daniel, R.: Irreversible furin cleavage site exposure renders immature tick-borne fl aviviruses fully infectious. (2025).

13. Goellner, S., Enkavi, G., Prasad, V., Denolly, S., Eu, S., Mizzon, G., Witte, L., Kulig, W., Uckeley, Z.M., Lavacca, T.M., Haselmann, U., Lozach, P., Brügger, B., Vattulainen, I., Bartenschlager, R.: Zika virus prM protein contains cholesterol binding motifs required for virus entry and assembly. (2023). https://doi.org/10.1038/s41467-023-42985-x.

14. Abbasi, E.: Emerging and transboundary arboviral diseases: the role of insect vectors and key drivers such as climate change and urbanization. (2025).

15. Ahmadi, K., Jahantigh, H.R., Ahmadi, N., Shahbazi, B.: Repurposing FDA-approved drugs and natural compounds to inhibit the RNA-dependent RNA polymerase domain of dengue virus 2 or dengue virus 3. 1–18 (2025).

16. Sahoo, S., Lee, H., Shin, D.: An integrated structural and immunoinformatic approach to design a multi-epitope-based vaccine against the foot-and- mouth disease virus. 1–13 (2025).

17. Su, S., Ni, Z., Lan, T., Ping, P., Tang, J., Yu, Z.: Predicting viral host codon fitness and path shifting through tree- based learning on codon usage biases and genomic characteristics. 1–14 (2025).

18. Nie, Z., Liu, X., Chen, J., Wang, Z., Liu, Y., Si, H., Dong, T., Xu, F., Song, G., Wang, Y., Zhou, P.: A unified evolution-driven deep learning framework for virus variation driver prediction. Nat. Mach. Intell. 7, (2025). https://doi.org/10.1038/s42256-024-00966-9.