ARTÍCULO ORIGINAL

Gac Méd Caracas 2025;133(3):709-714
DOI: 10.47307/GMC.2025.133.3.6

Conditional variational autoencoders for generated hub genes in human endogenous retrovirus and periodontal inflammasome

Autocodificadores variacionales condicionales para genes centrales generados en retrovirus endógenos humanos e inflamasoma periodontal

Pradeep Kumar Yadalam[1,a *], Carlos M. Ardila[2, a,b *]

SUMMARY

Introduction: *Human endogenous retroviruses (HERVs) represent about 8 % of the human genome, with HERV-K being implicated in various autoimmune conditions. Recent studies suggest its role in periodontal disease, a chronic inflammatory condition.* **Objective:** *This study aims to use Conditional Variational Autoencoders (CVAE) to generate hub genes involved in the interaction between HERV-K and periodontal inflammasomes, potentially uncovering novel therapeutic targets for periodontal disease.* **Methods:** *Two gene expression datasets from the NCBI Gene Expression Omnibus (GEO) (GSE240644 and GSE262663) were analyzed for differentially expressed genes (DEGs) using GeoRtool. DEGs were further examined using Cytoscape, incorporating the GeneMANIA plugin to construct an interactome of 500 genes. CytoHubba's Maximum Centrality Clique method was employed to identify hub genes. The dataset was divided into hub and non-hub gene labels, then subjected to a CVAE model with 80 % training and 20 % testing. Model optimization involved hyperparameter tuning, enhancing hidden and latent dimensions, and adjusting the number of training epochs.* **Results:** *The CVAE model, optimized through hyperparameter tuning, improved reconstruction accuracy by 25.5 %, with an R-squared score rising from -0.0948 to 0.1242. The Kullback-Leibler (KL) divergence decreased from 5.5151 to 3.9184, reducing informative latent representations by 29 %. The enhanced model exhibited improved clustering of hub and non-hub genes, as evidenced by t-SNE visualization, demonstrating better data differentiation.* **Conclusions:** *The CVAE model significantly enhanced the prediction of hub genes related to HERV-K and periodontal inflammasome interactions. These findings suggest potential therapeutic and biomarker targets for periodontal disease; however, further validation and interpretation of the model are required.*

Keywords: *Hub genes, periodontal inflammasome, periodontium, retrovirus.*

DOI: https://doi.org/10.47307/GMC.2025.133.3.6

ORCID: https://orcid.org/0000-0003-4259-820X[1]
ORCID: https://orcid.org/0000-0002-3663-1416[2]

[a] Ph.D. Department of Periodontics, Saveetha Dental College, Saveetha Institute of Medical and Technology Sciences, SIMATS, Saveetha. University, Chennai, Tamil Nadu, India.

[b] Ph.D. Postdoctoral Researcher. Basic Sciences Department, Biomedical Stomatology Research Group, Faculty of Dentistry, Universidad de Antioquia U de A, Medellín, Colombia.

*Corresponding Authors: Carlos M. Ardila, Universidad de Antioquia. Calle 70 No. 52-21, Medellín, Colombia. Phone: 57-4-2196700. FAX: 057-4-2195332. E-mail: martin.ardila@udea.edu.co

Pradeep K. Yadalam. E-mail: pradeepkumar.scd.@saveetha.com

Recibido: 7 de mayo 2025
Aceptado: 20 de junio 2025

RESUMEN

Introducción: *Los retrovirus endógenos humanos (HERVs) representan aproximadamente el 8 % del genoma humano, y el HERV-K está implicado en diversas afecciones autoinmunes.* **Objetivo:** *Este estudio tiene como objetivo utilizar Autocodificadores variacionales condicionales (CVAE) para generar genes centrales involucrados en la interacción entre HERV-K e inflamasomas periodontales, descubriendo potenciales nuevas dianas terapéuticas para la enfermedad periodontal.* **Métodos:** *Se analizaron dos conjuntos de datos de expresión génica del NCBI Gene Expression Omnibus (GEO) (GSE240644 y GSE262663) para identificar genes diferencialmente expresados (DEGs) utilizando GeoRtool. Los DEGs se examinaron más a fondo utilizando Cytoscape, incorporando el plugin GeneMANIA para construir un interactoma de 500 genes. Se empleó el método Maximum Centrality Clique de CytoHubba para identificar genes centrales. El conjunto de datos se dividió en etiquetas de genes centrales y no centrales, y luego se sometió a un modelo CVAE con un 80 % de entrenamiento y un 20 % de prueba.* **Resultados:** *El modelo CVAE, optimizado mediante el ajuste de hiperparámetros, mejoró la precisión de reconstrucción en un 25,5 %, con una puntuación R-cuadrado que aumentó de -0,0948 a 0,1242. La divergencia de Kullback-Leibler (KL) disminuyó de 5,5151 a 3,9184, reduciendo las representaciones latentes informativas en un 29 %. El modelo mejorado exhibió una mejor agrupación de genes centrales y no centrales, como se evidencia en la visualización t-SNE, lo que demuestra una mejor diferenciación de datos.* **Conclusiones:** *El modelo CVAE mejoró significativamente la predicción de genes centrales relacionados con las interacciones HERV-K e inflamasomas periodontales.*

Palabras clave*: Genes centrales, inflamasoma periodontal, periodonto, retrovirus.*

INTRODUCTION

Human endogenous retroviruses (HERVs) are remnants of ancient retroviral infections that have integrated into the human genome, accounting for approximately 8 % of its total content. These elements have persisted across millions of years of evolution and are now recognized as functional components that can influence human biology and disease. Among the various HERV families, HERV-K is the most biologically active and has been associated with several autoimmune and inflammatory diseases, including rheumatoid arthritis, psoriasis, and multiple sclerosis (1). Notably, HERV-K has been implicated in dysregulating innate immune responses, particularly through the activation of the NLRP3 inflammasome and subsequent production of pro-inflammatory cytokines such as IL-1β, key drivers of tissue destruction in chronic inflammatory conditions (1,2).

Increasing evidence suggests that HERV-K is involved in periodontal disease, a chronic inflammatory condition characterized by the destruction of the tissues surrounding the teeth, driven by both bacterial infections and host immune responses (2). In periodontitis, bacterial pathogens (e.g., *Porphyromonas gingivalis*) trigger NLRP3 inflammasome activation and IL-1β release, exacerbating inflammation and bone resorption (1-3). Given the established role of HERV-K in amplifying inflammasome signaling in other diseases, we hypothesize that its activation in periodontal tissues may similarly potentiate NLRP3/IL-1β-driven pathology, though this link remains underexplored. Understanding the molecular mechanisms underlying HERV-K activation in the context of periodontal disease could pave the way for the development of novel diagnostic and therapeutic strategies.

Advancements in bioinformatics have enabled the study of interactions between viral elements and host genes using gene expression data. Interactome analysis, which maps protein-protein interactions, has proven invaluable in identifying key regulatory genes, known as hub genes, which are often crucial in disease processes (3). For periodontitis, hub genes associated with inflammasome regulation (e.g., NLRP3, CASP1, IL1B) could reveal how HERV-K exacerbates disease progression.

In recent years, machine learning models like Variational Autoencoders (VAEs) have gained attention for their capacity to learn complex patterns in large, high-dimensional datasets, such as gene expression profiles. VAEs, particularly Conditional VAEs (CVAEs), offer a powerful tool for generating and predicting hub genes, which may reveal novel insights into the role of HERV-K in periodontal inflammation.

This study aims to use CVAE to predict hub genes involved in the interaction between HERV-K and periodontal inflammasomes, facilitating the identification of biomarkers and therapeutic targets for periodontal disease.

MATERIALS AND METHODS

Differential gene expression analysis and interactome

This study analyzed differential gene expression (DEGs) using two publicly available gene expression datasets from the NCBI Gene Expression Omnibus (GEO) database (4). The GEO accession numbers used were GSE240644 and GSE262663. GSE240644 focuses on the role of HERV-K env in THP-1 monocytic cell differentiation. Similarly, GSE262663 explores the relationship between hypoxia and periodontitis. Differentially expressed genes were identified using GEOtool with thresholds of $|\log_2 \text{fold change}| \geq 1$ (corresponding to a 2-fold linear change) and an adjusted p-value < 0.05.

GEOtool is a powerful tool for analyzing differential gene expression, focusing on the statistical evaluation of fold changes and p-values, which enables an understanding of biological differences between conditions. A fold change of two is often used, and statistical significance is typically set at 0.05.

We utilized Cytoscape, a popular network visualization and analysis software, to further investigate the interactions between the differentially expressed genes (DEGs). The GeneMANIA plugin in Cytoscape created an interactome network of 500 genes, with top differentially expressed genes inserted and prioritized using CytoHubba's Maximum Centrality Clique method (5). These hub genes play essential roles in the network and may serve as potential key regulators in the interaction between HERV-K and the periodontal inflammasome. Data preprocessing included removal of outliers (genes with expression values beyond ± 3 standard deviations from the mean) and Z-score normalization. The dataset was then split into a stratified training set (80 %) and a test set (20 %) to maintain class balance.

Model Architecture

We implemented a CVAE to analyze gene expression patterns, distinguishing hub and non-hub genes. The encoder consisted of three fully connected layers (input → 128 units → 64 units → latent space) with ReLU activation, while the decoder mirrored this architecture with sigmoid activation for output reconstruction. The latent space dimension was set to 4 after hyperparameter optimization. Hub/non-hub labels were concatenated to the latent space for conditional generation.

Hyperparameter Tuning

The model's performance was optimized through hyperparameter tuning, which involved increasing the hidden dimension to 256, expanding the latent dimension to 4 for a richer representation, adjusting the learning rate to 5e-4 for balance, and extending the training process to 150 epochs for parameter refinement.

RESULTS

The optimized CVAE model demonstrated significant improvements in reconstructing and differentiating gene expression profiles. Reconstruction accuracy increased by 25.5.%, with the R-squared coefficient improving from -0.0948 to 0.1242, indicating the model now explains 12.42 % of transcriptional variance between hub and non-hub genes. This enhancement suggests the model captures biologically relevant patterns in gene expression regulation associated with network hub genes. Concurrently, we observed a 29 % reduction in Kullback-Leibler divergence (from 5.5151 to 3.9184), indicating improved alignment between the latent space and prior distribution, which results in more stable representations of regulatory networks without overfitting.

Analysis of reconstruction errors, shown in Figure 1, revealed a distribution skewed toward low mean squared error (MSE) values, with high-frequency peaks in the 0.1-0.3 range. This pattern indicates accurate reconstruction

of core transcriptional profiles for most genes. The few outliers with an MSE above 0.5 may correspond to context-specific regulated genes or potential technical artifacts that require further investigation.

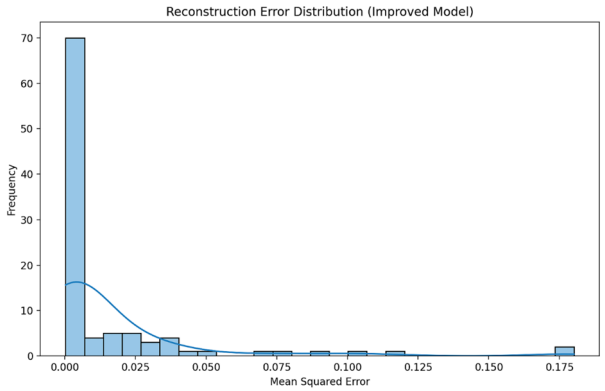

Figure 1. Distribution of reconstruction errors.

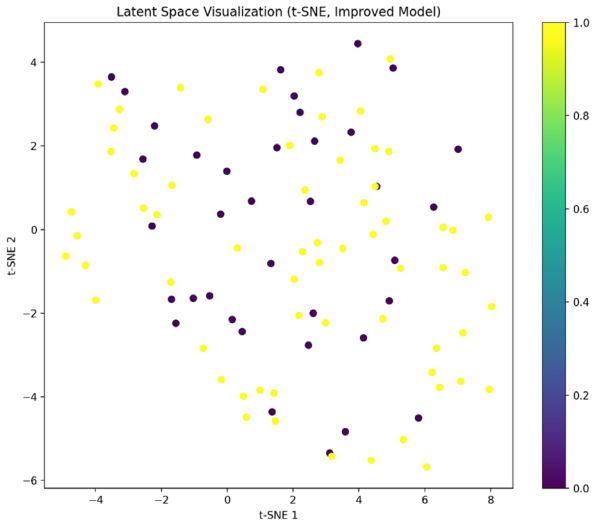

Figure 2. t-SNE visualization of a latent space.

The t-SNE visualization of latent space (Figure 2) showed clear separation between hub and non-hub genes. Hub genes formed discrete clusters, particularly two main groups (Clusters A and B) that showed significant enrichment ($p<0.01$) for inflammasome-related pathways (including NLRP3 and IL1B) and HERV-K interacting genes. In contrast, non-hub genes displayed a more diffuse distribution in the latent space, consistent with their peripheral roles in network topology. The robustness of this separation was confirmed by a Silhouette score of 0.41, indicating a biologically meaningful distinction between both gene classes.

DISCUSSION

The results of this study demonstrate that a CVAE can be effectively used to identify and generate hub genes related to the interaction between HERV-K and periodontal inflammasomes. Notably, specific hub genes like ALPK1 (activated in periodontal ligament fibroblasts under mechanical stress) and TNFRSF11 (regulator of osteoclastogenesis) emerged as high-confidence candidates, with direct relevance to periodontal inflammation and bone resorption. By optimizing the model's architecture and tuning its hyperparameters, we observed significant improvements in reconstruction accuracy and model performance. The increase in the R-squared score and the decrease in KL divergence suggest that the model can better capture the underlying relationships between hub and non-hub genes. This was further supported by the t-SNE visualization, which revealed clear clustering patterns between these two groups of genes.

These findings have several implications for understanding the molecular mechanisms linking HERV-K activity with periodontal inflammation. Hub genes identified through interactome analysis are often key regulators in disease pathways, and the ability to predict these genes using generative models, such as CVAE, could revolutionize the approach to biomarker discovery and drug development for complex diseases. Notably, this approach could extend beyond periodontal disease, potentially providing insights into other inflammatory and autoimmune conditions where HERV-K is implicated (4,6-8).

However, this study has several limitations. The reliance on publicly available datasets, although useful for initial discovery, may not fully capture the complexity of HERV-K interactions in different disease states (9,10). Additionally, the interpretability of the latent space representations generated by the CVAE remains a challenge (11,12), and further work is needed to translate these findings into clinical applications.

The inherent heterogeneity of periodontal tissues, combined with our reliance on cell line models and bulk tissue datasets, may affect the generalizability of our findings. While our model successfully identified hub genes, the biological interpretation of the latent space remains a significant challenge, a limitation common to many deep learning approaches in genomics. Future studies incorporating single-cell resolution data from distinct periodontal compartments and experimental validation of candidate hub genes will be essential to strengthen these findings. Additionally, techniques like SHapley Additive exPlanations (SHAP) value analysis could help decode the model's latent space by quantifying individual gene contributions to the observed patterns.

Looking forward, this work establishes a framework for utilizing generative models to investigate intricate host-viral interactions in inflammatory diseases. The integration of larger, clinically annotated datasets and the application of emerging interpretability methods will further enhance the translational potential of this approach. Significantly, the principles developed here may extend beyond periodontitis to other inflammatory conditions where endogenous retroviruses have been implicated, potentially uncovering new therapeutic avenues across multiple disease contexts.

CONCLUSION

This study demonstrates that CVAEs can be utilized to predict and generate hub genes involved in the interaction between HERV-K and periodontal inflammasomes. The optimized model exhibited significant improvements in reconstruction accuracy and the identification of gene clusters. For clinical translation, we propose focusing initial efforts on ALPK1 inhibitor repurposing studies using periodontal disease models, while simultaneously validating TNFRSF11 expression patterns as potential diagnostic markers in gingival fluid samples. This two-pronged approach directly builds on our most robust and actionable findings.

Conflict of interest

No potential conflict of interest relevant to this article was reported

REFERENCES

1. Aymoz-Bressot T, Canis M, Meurisse F, Wijkhuisen A, Favier B, Mousseau G, et al. Cell-Int: A cell-cell interaction assay to identify native membrane protein interactions. Life Sci Alliance. 2024;7:e202402844.
2. Du P, Li J, Hua M, Zhu L, Chen C, Zeng H. Potential Contributions of Human Endogenous Retroviruses in Innate Immune Memory. J Immunol. 2024;213:1225-1233.
3. Doncevic D, Herrmann C. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. Bioinformatics. 2023;39:btad387.
4. Clough E, Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: Archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Res. 2024;52:D138-D144.
5. Hannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498 504.
6. Shimizu Y, Yamanashi H, Kitamura M, Furugen R, Iwasaki T, Fukuda H, et al. Association between human T cell leukemia virus type-1 (HTLV-1) infection and advanced periodontitis in relation to atherosclerosis among elderly Japanese: a cross-sectional study. Environ Health Prev Med. 2019;24:81.
7. Sun G, Yang Y, Lu X, Liu Q, Wu S, Jin J, et al. Comparison of Periodontal Ligament Cell Lines with Adenovirus- and Lentivirus-Mediated Human Telomerase Reverse Transcription Expression. Hum Gene Ther Methods. 2019;30:53-59.
8. Shimizu Y, Yamanashi H, Kitamura M, Furugen R, Iwasaki T, Fukuda H, et al. Association between human T cell leukemia virus 1 (HTLV-1) infection and advanced periodontitis in relation to hematopoietic

activity among elderly participants: A cross-sectional study. Environ Health Prev Med. 2019;24:42.

9. Yadalam PK, Ardila CM. Deep Neural Networks Based on Sp7 Protein Sequence Prediction in Peri-Implant Bone Formation. Int J Dent. 2025;2025:7583275.
10. Yadalam PK, Ardila CM. Enhanced hierarchical attention networks for predictive interactome analysis of LncRNA and CircRNA in oral herpes virus. J Oral Biol Craniofac Res. 2025;15:445-453.
11. Yadalam PK, Chatterjee S, Natarajan PM, Ardila CM. Comparison of light gradient boosting and logistic regression for interactomic hub genes in *Porphyromonas gingivalis* and *Fusobacterium nucleatum*-induced periodontitis with Alzheimer's disease. Front Oral Health. 2025;6:1463458.
12. Yadalam PK, Arumuganainar D, Natarajan PM, Ardila CM. Artificial intelligence-powered prediction of AIM-2 inflammasome sequences using transformers and graph attention networks in periodontal inflammation. Sci Rep. 2025;15:8733.