forum acusticum 2026
8-12 September · Graz, Austria

# HERMES: A PHYSICS-INFORMED HIERARCHICAL MESH TRANSFORMER FOR REAL-TIME ROOM EIGENFREQUENCY PREDICTION

Bence Bakos[1,2], Csaba Huszty[2], Gergely Firtha[3], Ferenc Izsák[4], András Lukács[1]

[1] *AI Research Group, Institute of Mathematics, Eötvös Loránd University, Hungary*
[2] *ENTEL Engineering Research & Consulting Ltd., Szépvölgyi út 32, Budapest, Hungary*
[3] *Department of Networked Systems and Services, Faculty of Electric Engineering, Budapest University of Technology and Economics, Hungary*
[4] *Department of Applied Analysis and Comput. Mathematics & HUN-REN–ELTE NumNet Research Group, Eötvös Loránd University, Hungary*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

## Abstract

The accurate prediction of low-frequency room acoustics relies on modal frequencies, traditionally computed using expensive numerical methods such as the Finite Element Method (FEM). This paper presents HERMES (Hierarchical Estimator for Room Modal Eigenvalue Synthesis), a physics-informed hierarchical mesh transformer that predicts the first 200 modal frequencies directly from 3D surface meshes of rooms with diverse geometries. The architecture employs a two-phase transformer backbone that leverages object-level structure for efficient processing of complex meshes, combined with an LSTM-based prediction head grounded in Weyl's law to incorporate physical priors. Trained and evaluated on a synthetic dataset of 7,500 rooms with various shapes and sizes, HERMES achieves a mean absolute error of 0.59 Hz across all room types, consistently outperforming an analytical scaled box baseline, while reducing computational cost by approximately three orders of magnitude compared to conventional eigenvalue solvers. Additionally, we present a sensitivity analysis which quantifies how modal frequency prediction uncertainties propagate into derived acoustic pressure fields. By demonstrating that deviations remain within a stable, linear regime for the error range achieved by HERMES, we confirm our neural approach provides both the efficiency and precision required for reliable acoustic engineering.

**Keywords:** *room acoustics, modal frequencies, deep learning, mesh transformer, sensitivity analysis*

***Corresponding author**: bakosbence@student.elte.hu.*




## 1. Introduction

Accurate modeling of room acoustics at low frequencies is fundamentally governed by wave phenomena, where room modes dictate the steady-state and transient responses. While geometrical acoustics excel at higher frequencies, wave-based numerical methods such as the Finite Element Method (FEM) or Boundary Element Method (BEM) are required to extract modal frequencies. However, these methods are notoriously computationally expensive, prohibiting their use in real-time applications or iterative architectural design.

Recently, deep learning has shown promise as a surrogate for acoustic simulations [1]. One line of work targets end-to-end predictions such as impulse responses [2] or acoustic descriptors [3], [4], while a complementary direction focuses specifically on the modal eigenvalue problem [5]–[7]. Many of these methods utilize physics-informed neural network techniques [6]–[8]. However, existing methods in both categories rely on simplified input representations—such as voxelized volumes or 2D projections—which limits their applicability to complex real-world geometries.

In this work, as part of the SOUNDY project [9], we address these limitations with three contributions: (1) a hierarchical mesh transformer that operates directly on raw 3D surface meshes, avoiding lossy intermediate representations; (2) a physics-informed prediction head grounded in Weyl's law that combines acoustic priors with learned residual corrections; and (3) a sensitivity analysis that quantifies how modal frequency errors propagate into derived acoustic pressure fields. We demonstrate that for prediction errors within the range achieved by HERMES, the resulting deviations in the frequency response remain within a stable, linear regime. This confirms that our neural approach is not

only computationally efficient but also provides the necessary precision for reliable acoustic engineering and design.

## 2. Synthetic Room Dataset

To train a highly generalizable mesh transformer, a vast and diverse dataset of room geometries is required. Using our proprietary framework, specifically developed for this purpose, we generated 7500 synthetic surface room meshes and simulated modal frequencies and functions for each of them. The dataset includes room geometries with different floor shapes (from rectangles to general polygons) with varying volumes ranging from 8 to 3000 $m^3$. These geometries also include artifacts, like boxes and columns, to make the spaces even more diverse. The basic properties of the resulting dataset are summarized in Table 1.

**Table 1:** Summary statistics of the room properties. Gross volume represents the total spatial volume of the room without accounting for any internal artifacts, whereas net volume denotes the remaining space after subtracting the volume occupied by those artifacts.

| Property | Min | Max | Mean |
|---|---|---|---|
| Height (m) | 2.40 | 6.00 | 4.18 |
| BBox Width (m) | 1.33 | 20.00 | 10.89 |
| BBox Depth (m) | 1.68 | 34.95 | 13.35 |
| Floor Area ($m^2$) | 2.28 | 551.91 | 113.22 |
| Gross Volume ($m^3$) | 8.87 | 2,993.57 | 473.27 |
| Net Volume ($m^3$) | 7.52 | 2,967.08 | 455.82 |
| Walls per room | 4 | 8 | 6.25 |

For each generated mesh, the first 200 ground-truth modal frequencies were computed using a custom-developed high-order Finite Element Method (FEM) solver. The solver performs a modal analysis by solving the generalized eigenvalue problem, utilizing quadratic ($2^{nd}$ order) tetrahedral elements for the volumetric discretization. The computed modal frequencies typically span the range from the fundamental room mode (often below 20 Hz) up to approximately 150–400 Hz, depending on the room volume.
The boundary conditions were modeled as frequency-independent surface admittance, where the absorption coefficients of the materials ($\alpha$) were mapped to the real part of the acoustic admittance. This allows the solver to construct a full modal damping matrix, capturing the energy dissipation at the boundaries. The resulting dataset thus provides a diverse representation of both the geometric resonances and the associated modal damping characteristics of complex interior spaces.

## 3. Mesh-Processing Transformer Architecture

Directly applying standard deep learning to 3D meshes is challenging due to their irregular topology and non-Euclidean geometry. Traditional Convolutional Neural Networks (CNNs) require uniform grids, making them poorly suited for unstructured 3D data. Consequently, the field often converts meshes into intermediate representations—such as voxels or point clouds—which can sacrifice fine geometric detail or inflate computational overhead. To process raw geometry directly and overcome these limitations, we leverage a mesh-processing transformer. Through self-attention, this approach effectively captures both local surface features and global topological dependencies without the constraints of a rigid spatial grid.

### 3.1. Input Representation

Let the 3D surface mesh of the room be defined as a set $\mathcal{M}$ consisting of $N$ triangular faces:

$$\mathcal{M} = \{f_i\}_{i=1}^{N}. \tag{1}$$

To capture the necessary acoustic and geometric characteristics, each face $f_i$ is encoded with a feature vector comprising its spatial vertex coordinates, associated material properties (e.g., absorption and scattering parameters), and a discrete object assignment index. Rather than applying explicit feature normalization, the scale differences between the diverse input types are handled implicitly by an initial learnable dense layer, which projects the raw features into a unified latent space before the transformer blocks. This composite representation provides the foundational physical and structural information required to solve for the impulse response and modal frequencies. Furthermore, the object indices—typically available as a byproduct of the room design process—serve as critical structural priors that facilitate the hierarchical processing detailed in the subsequent section.

### 3.2. Hierarchical Transformer Backbone

Before predicting the modal frequencies, the HERMES model extracts a global representation vector for the room using a two-phased, hierarchical transformer architecture. Standard 3D mesh transformers often struggle with the large number of input faces inherent to complex geometries, typically relying on spatial patching heuristics, such as $k$-nearest neighbor grouping [10] or graph partitioning [11] to aggregate local groups of faces. To mitigate this computational bottleneck without losing structural coherence, our backbone leverages the object-level assignments provided in the input representation. The architecture processes the mesh in two distinct phases:

- **Object-Level Aggregation:** An initial transformer module processes the groups of face tokens belonging to the same object, outputting an intermediate representation vector for each distinct object in the room.
- **Global Room Representation:** A subsequent transformer module takes these aggregated object vectors as input tokens to produce a final, unified feature vector for the entire room.

This approach allows the model to efficiently handle complex rooms with high face counts while maintain-

ing the permutation invariance necessary for unorganized mesh data.

### 3.3. Recurrent Head Using Weyl's Law

From Weyl's law, we can approximate the $k$-th eigenfrequency, $f_k$, as follows:

$$f_k \approx a \cdot k^{1/3}, \quad a = \left(\frac{3c^3}{4\pi V}\right)^{1/3}, \tag{2}$$

where $c$ is the speed of sound at room temperature and $V$ is the room volume.

We employ this physical principle as a baseline estimator within the prediction head of our network. Rather than predicting absolute frequencies directly, we use a Long Short-Term Memory (LSTM [12]) network to sequentially predict residual corrections to the Weyl baseline. The final predicted $k$-th frequency is:

$$p_k = \text{Weyl}(k) + p'_k, \tag{3}$$

where $p'_k$ is the learned residual at step $k$. This design grounds the network in acoustic theory while giving it the flexibility to correct for complex, non-rectangular geometries where the idealized Weyl approximation is insufficient.

## 4. Sensitivity Analysis of Acoustic Parameters

While the neural network yields an estimated frequency $\hat{f}_k = f_k + \Delta f_k$, it is crucial to understand how this prediction error ($\Delta f_k$) impacts downstream acoustic design.

### 4.1. Error Propagation

The room transfer function (RTF) or frequency response $H(\omega)$ at low frequencies is modeled as a sum of modal resonances:

$$H(\omega) \approx \sum_k \frac{c_k}{\omega^2 - \omega_k^2 + j2\delta_k\omega_k}, \tag{4}$$

where $\omega_k = 2\pi f_k$ is the angular modal frequency, $\delta_k$ is the damping factor, and $c_k$ represents modal coupling coefficients. We analyze the sensitivity of this acoustic response function to perturbations in $f_k$.

Let $\{f_k\}_{k=1}^K$ be the set of ground-truth eigenfrequencies obtained via FEM. We introduce a stochastic perturbation to these frequencies:

$$\hat{f}_k = f_k(1 + \varepsilon_k), \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where $\sigma$ represents the relative error magnitude.

The sensitivity is assessed by comparing the reference complex pressure field $p(\omega)$ and the perturbed field $\hat{p}(\omega)$ at $N_r$ randomly distributed receiver positions. For each configuration, we compute the mean relative error ($E$) over the frequency range of interest:

$$E = \frac{1}{N_r}\sum_{i=1}^{N_r}\int \left|\frac{p(f,\mathbf{r}_i) - \hat{p}(f,\mathbf{r}_i)}{p(f,\mathbf{r}_i)}\right| \mathrm{d}\omega. \tag{6}$$

This process is repeated across a gradient of average absorption coefficients $\bar{\alpha}$ to investigate the coupling between modal damping and frequency precision. In our numerical framework, we utilize a full modal damping matrix derived from the boundary conditions to ensure that the interaction between shifted resonances and their respective bandwidths is accurately captured. This methodology allows us to define the "acoustic stability" of the prediction: identifying the maximum allowable $\varepsilon_n$ for which the resulting error $E$ remains below a perceptually or engineering-relevant threshold.

### 4.2. Perturbation Results

The numerical results of the perturbation analysis are summarized in Fig. 1. The simulation confirms that the mean relative pressure error $E$ scales linearly with the modal frequency perturbation magnitude $\varepsilon_n$ in the decibel domain. This linear relationship is maintained consistently across various room configurations and absorption profiles, indicating that the global acoustic response's sensitivity to frequency errors is a fundamental statistical property of the modal superposition.

Specifically, the error propagation remains predictable and linear for perturbations up to $\varepsilon_n \approx 0.1$ (10% relative frequency error). Above this threshold, the modal shifts become large enough to cause significant overlap and phase decorrelation, leading to a saturation of the error metric. Given that the HERMES model achieves a mean absolute error of approximately 0.59 Hz—which corresponds to $\varepsilon_n \ll 0.01$ in the considered frequency range—the predicted modal frequencies fall well within the stable linear region, ensuring high fidelity in derived acoustic parameters.

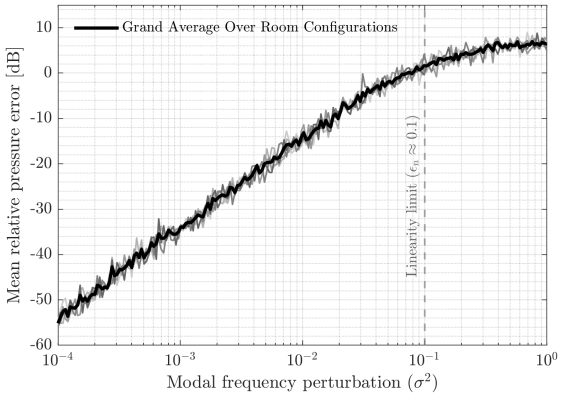


**Figure 1:** Mean relative pressure error as a function of the modal frequency perturbation $\varepsilon_n$. The grey lines represent individual absorption configurations ($\bar{\alpha}$ ranging from 0.0 to 0.5), while the black line shows the grand average. The error scales linearly with the perturbation up to $\varepsilon_n \approx 0.1$.

## 5. Experiments and Results

### 5.1. Implementation Details

The models are trained using the $L^2$ loss on the sequence of predicted eigenfrequencies. Teacher forcing is used for the LSTM sequence generation during training by supplying the ground-truth eigenfrequencies from previous steps as inputs rather than the model's own predictions. To prevent overfitting and enhance generalization capabilities, we apply random rotations to the input rooms along the vertical axis. The dataset was split into 80% training and 20% test sets. The Adam optimizer is used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a base learning rate of $10^{-4}$. Exponential learning rate decay was applied with a total rate of 0.1. The models were trained for 200 epochs with a batch size of 32, and the training process took around 16 hours using a NVIDIA A100 GPU.

### 5.2. Baseline Scaled Box Model

To establish a point of reference for our neural network's performance, we compare it against a naive analytical calculation method. We assign a rectangular bounding box to each room, which is then proportionally scaled down so its volume matches the true volume of the room. Then, we calculate the modal frequencies of this scaled rectangular room using the formula:

$$f_{p,q,r} = \frac{c}{2}\sqrt{\left(\frac{p}{L}\right)^2 + \left(\frac{q}{W}\right)^2 + \left(\frac{r}{H}\right)^2}, \qquad (7)$$

where:

- $c$ is the speed of sound in air (typically around 343 m/s at 20°C),
- $L$, $W$, and $H$ are the length, width, and height of the scaled rectangular box in meters,
- $p$, $q$, and $r$ are non-negative integers.

Note that this analytical formulation naturally breaks down for complex, non-cuboid geometries where Lord Rayleigh's assumptions no longer hold. Additionally, while it provides a computationally inexpensive approximation of the modal frequencies, it heavily relies on the precise total volume of the room—a metric that is difficult to extract from raw, unorganized meshes of complex geometries. Nevertheless, we use it as a baseline because it provides a straightforward order-of-magnitude reference to quantify the error introduced by idealized rectangular estimates, thereby highlighting the necessity of our data-driven approach for realistic room shapes.

### 5.3. Modal Frequency Prediction Performance

Table 2 and Table 3 display the performance of our HERMES model. The tables compare our model to the baseline scaled box model in empty rooms and all rooms, respectively. We evaluate the predictive performance using Mean Absolute Error (MAE) and Mean Relative Error (MRE).

**Table 2:** Comparison of HERMES to the Baseline Box Model for Empty Rooms.

| | HERMES | | Baseline (Box) | |
|---|---|---|---|---|
| | MRE | MAE | MRE | MAE |
| Rectangle | 0.0070 | 0.6236 | 0.0025 | 0.2339 |
| L-Shape | 0.0064 | 0.4186 | 0.0186 | 1.4101 |
| T-Shape | 0.0070 | 0.5984 | 0.0221 | 2.0495 |
| Polygon | 0.0056 | 0.6094 | 0.0152 | 1.8488 |
| All Shapes | 0.0064 | 0.5515 | 0.0162 | 1.5373 |

**Table 3:** Comparison of HERMES to the Baseline Box Model for All Rooms.

| | HERMES | | Baseline (Box) | |
|---|---|---|---|---|
| | MRE | MAE | MRE | MAE |
| Rectangle | 0.0067 | 0.5942 | 0.0244 | 2.7949 |
| L-Shape | 0.0064 | 0.4131 | 0.0277 | 2.2051 |
| T-Shape | 0.0073 | 0.6215 | 0.0442 | 4.5272 |
| Polygon | 0.0066 | 0.7412 | 0.0524 | 7.4371 |
| All Shapes | 0.0068 | 0.5860 | 0.0386 | 4.3557 |

For empty rectangular rooms—the trivial case where the analytical formula is near-exact—the baseline expectedly outperforms HERMES. Our model, on the other hand, performs consistently across different room shapes and significantly outperforms the baseline model in all other cases. HERMES achieves an MSRE (mean squared relative error) of 0.00018 over all rooms which, based on Fig. 1, corresponds to an estimated mean relative pressure error between -40 and -50 dB.

Fig. 2 and Fig. 3 show the performance of our model across different frequency numbers and values, respectively. Two general trends can be observed: later modes are generally easier to predict, however, for outstandingly high modal frequencies – for which large number of training examples are lacking – greater error rates appear. Targeted expansion of the training dataset could help reduce these error rates in the future.

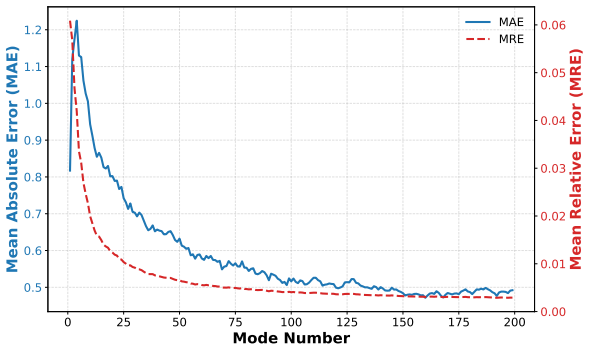


**Figure 2:** Absolute and relative errors on the test set as a function of mode number.

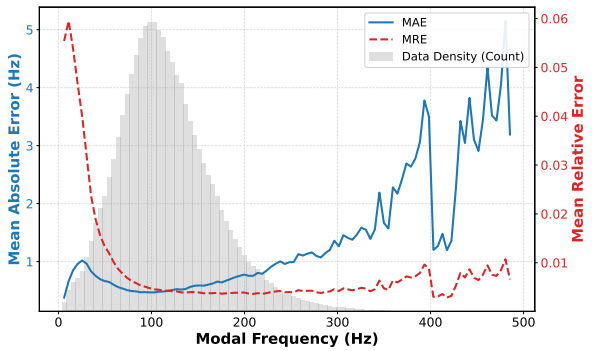


**Figure 3:** Absolute and relative errors on the test set as a function of frequency.

### 5.4. Computation Cost

The primary advantage of our model over traditional methods is a significant reduction in the computational cost required to calculate modal frequencies. Conventional eigenvalue solvers rely on 3D volumetric meshing and large matrix operations, which lead to high memory consumption and long execution times. In contrast, our lightweight neural network achieves orders-of-magnitude improvements in both speed and memory efficiency, as shown in Fig. 4.

In standard numerical methods, increasing a room's dimensions causes the number of tetrahedrons to grow cubically, tying the computational load directly to the room's total volume. Our approach bypasses this bottleneck by relying strictly on the room's surface mesh. While our model's computational complexity still depends on the detail of this input surface (i.e., the triangle count), it remains completely independent of the room's empty volume.

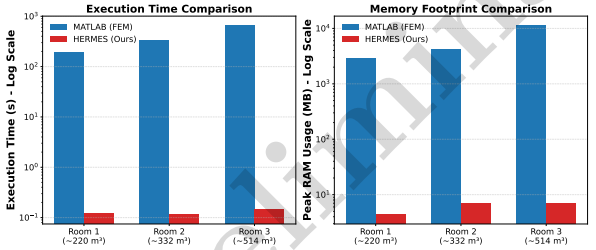


**Figure 4:** Runtime and memory comparison

## 6. Conclusion

This paper presented HERMES, a physics-informed hierarchical mesh transformer for predicting room eigenfrequencies directly from 3D surface meshes. By combining a two-phase transformer backbone with an LSTM-based prediction head grounded in Weyl's law, HERMES achieves a mean absolute error of 0.59 Hz for the first 200 modal frequencies, consistently outperforming the analytical scaled box baseline while operating at orders-of-magnitude lower computational cost than conventional eigenvalue solvers.

The accompanying sensitivity analysis helped make the numerical results more interpretable by demonstrating a clear, linear relationship between perturbations in the modal frequency vectors and the resulting deviations in the acoustic pressure field. This predictable behavior confirms that the model's prediction accuracy translates directly into reliable downstream acoustic performance.

Future work will focus on improving generalization capabilities and reducing errors by extending the training dataset and also expanding the prediction range to higher modal frequencies. Additionally, we aim to investigate the estimation of modal shapes using neural networks, which would enable a complete neural surrogate for the acoustic eigenvalue problem.

## 7. Acknowledgments

The research was supported by grant DIMOP_PLUSZ-1.1.2/A-24, as part of the Soundy project, and by the Hungarian National Research, Development and Innovation Office within the framework of the Thematic Excellence Program 2021 – National Research Sub programme: "Artificial intelligence, large networks, data security: mathematical foundation and applications". The project has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2023 funding scheme.

## References

[1] T. van Waterschoot, "Deep, data-driven modeling of room acoustics: Literature review and research perspectives," *arXiv preprint arXiv:2504.16289*, 2025.

[2] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 924–933.

[3] C. Huszty, B. Bakos, B. Csanády, G. Hidy, and A. Lukács, "Prediction of room acoustic parameters in rectangular rooms using recurrent neural networks," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, vol. 268, 2023, pp. 3058–3069.

[4] B. Bakos, G. Hidy, B. Csanády, C. Huszty, and A. Lukács, "Estimating room acoustic descriptors from bag-of-vectors representation with transformers," *Engineering Applications of Artificial Intelligence*, vol. 145, p. 110 183, 2025.

[5] O. Lundin, E. Zea, J. Cuenca, and U. P. Svensson, "Prediction of eigenfrequencies in non-rectangular rooms with machine learning," in *International Congress on Acoustics, ICA 2022, Gyeongju, Korea, 24-28 Oct, 2022*, 2022, pp. 1–5.

[6] W.-W. Lo, Y.-S. Chen, and M. R. Bai, "Physics-informed neural network for room acoustic transfer function and modal analysis," *Journal of Mechanics*, vol. 41, pp. 510–518, 2025.

[7] S. Schoder, "Physics-informed neural networks for modal wave field predictions in 3d room acoustics," *Applied Sciences*, vol. 15, no. 2, p. 939, 2025.

[8] N. Borrel-Jensen, S. Goswami, A. P. Engsig-Karup, G. E. Karniadakis, and C.-H. Jeong, "Sound propagation in realistic interactive 3d scenes with parameterized sources using deep neural operators," *Proceedings of the National Academy of Sciences*, vol. 121, no. 2, e2312159120, 2024.

[9] Soundy, Entel Engineering Research & Consulting Ltd. Available at `https://soundy.ai`, 2024.

[10] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.

[11] Y. Liang, S. Zhao, B. Yu, J. Zhang, and F. He, "Meshmae: Masked autoencoders for 3d mesh data analysis," in *European conference on computer vision*, Springer, 2022, pp. 37–54.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.