Supplementary Material

Probing the stochastic dynamics of coronaviruses: Machine learning assisted deep computational insights with exploitable dimensions

T. Mukhopadhyay [a], S. Naskar [b], K. K. Gupta [c], R. Kumar [c], S. Dey [c], S. Adhikari [d]

[a] *Department of Aerospace Engineering, Indian Institute of Technology Kanpur, Kanpur, India*

[b] *Department of Aerospace Engineering, Indian Institute of Technology Bombay, Mumbai, India*

[c] *Department of Mechanical Engineering, National Institute of Technology Silchar, Silchar, India*

[d] *College of Engineering, Swansea University, Swansea, United Kingdom*

In this supplementary document, we have provided the details concerning two critical aspects of the current investigation: finite element modeling of coronavirus structure and support vector regression assisted machine learning model. Basic stages involved in the formation of machine learning model and the statistical approach for assessing the prediction capability of the machine learning model are discussed in detail. Subsequently, this document also presents an overview of the Monte Carlo simulation and sensitivity analysis carried out in the present investigation.

Contents

Figures: S1 to S3

Tables: S1 to S4

SM1. Finite element modeling and validation of coronavirus structure

In this section, we briefly discuss the finite element modeling approach of the equivalent coronavirus structure (refer to figure 1(C) of the main paper) including the geometrical details. The modeled structure of coronavirus consists of a main spherical shell of diameter 120 nm with wall thickness of 8 nm. Multiple spikes are modeled as solid cylindrical beam of length 24 nm and diameter 3 nm. The tip mass of the spikes are modeled as the projected sphere of diameter 6 nm. The most critical aspect of modeling a coronavirus structure is to properly position various numbers of spikes. In the present study we have considered three different virus configurations with different number of spikes (20, 40 and 60 spikes). The positioning of the spikes in coronavirus structure is accomplished based on latitudes and longitudes as depicted figure S1. For example, in case of 40 and 60 spikes attached to the coronavirus structure, four and six rows are modeled respectively with 10 spikes distributed in the latitude direction. The rows in the case of 40 and 60 spikes are separated by 36^{o} and 25.714^{o} angular gap respectively in the longitude direction, while the angular gap between each spike in the latitude direction are maintained as 36^{o} (refer to figure S2).

In the finite element analysis code, meshing of the coronavirus structure is carried out by utilizing 3D hexahedron mesh element with a mesh size of 3.5 nm for the spherical shell and spikes, whereas for the inner soft core 2D quadrilateral mesh with the mesh size of 1.2 nm is used. Following the standard procedure of finite element approach, the element-level mass and stiffness matrices are assembled to obtain the global mass and stiffness matrices of the entire coronavirus structure. The free vibration analysis is performed by solving an eigenvalue problem between the global mass and stiffness matrices, where the natural frequencies and modeshapes are obtained from the eigenvalues and eigenvectors, respectively. Here the block Lanczos algorithm has been utilized to solve the eigenvalue problem concerning vibration analysis of the coronavirus structure.

Figure S1. The division of spherical shell structure into longitude-latitude.

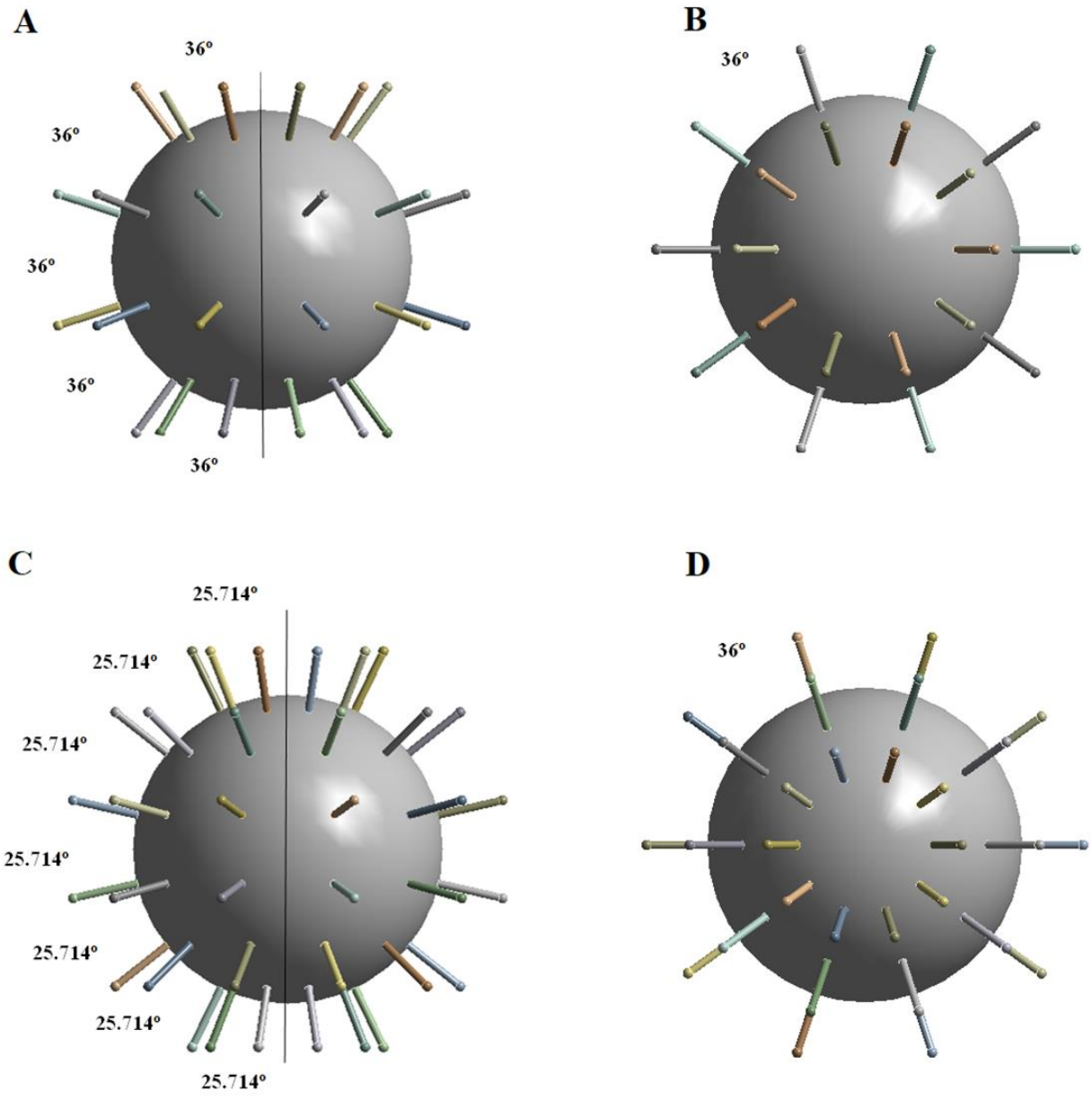

Figure S2. Distribution of different number of spikes in coronavirus structure (front and top views). (A-B) Coronavirus structure with 40 spikes **(C-D)** Coronavirus structure with 60 spikes

Table S1. Material properties of normal and cancerous breast cell [33]

	Elastic modulus (KPa)	Poisson's ratio	Density (Kg/m^3)
MCF- 10A	0.73	0.49	1020
MCF-7	0.425	0.49	994

Table S2. Validation of natural frequencies (normal and cancerous breast cell) with literature [33]

Mode	**Natural frequency**			
	MCF-10A		**MCF-7**	
	Jaganathan et al.	**Present study**	**Jaganathan et al.**	**Present study**
Mode 1	6.2579 ± 2.2	4.9754	4.8369 ± 1.9	7.5147
Mode 2	14.516 ± 1.18	15.105	11.212 ± 1.13	9.4814
Mode 3	17.071 ± 1.29	18.569	13.195 ± 1.34	14.766

Table S3. Fundamental natural frequency of fullerenes C_{60} obtained using finite element modeling [46] and other approaches (Raman spectroscopy and Molecular mechanics)

Natural frequency (THz)	**Free vibration of fullerenes C_{60}**			
	FEM (ANSYS)	**Raman spectroscopy [48]**	**Molecular mechanics [47]**	**Molecular mechanics [49]**
	8.605	8.19	8.69	8.31

Before investigating the dynamic behaviour of coronaviruses, it is necessary to gain adequate confidence on the numerical modeling approach by means of carrying out validation studies. Since suitable literature concerning the vibration analysis of coronaviruses is not available, we resort to other spherical bio/nano structures for the purpose of gaining confidence on the adopted finite element modeling approach. First, we have modeled two different breast cells (normal and cancerous) with exactly same geometric dimensions and material properties (refer to Table S1) as the reference article following the finite element approach and compared the results of computed natural frequencies with the numerical values reported in literature [33]. Both the normal and cancerous cells are modeled as a sphere with an outer diameter 30 micrometers and elastic wall thickness of 200 nanometers. The modeling of cells is carried out as linear elastic material considering four node

shell-63 elements. A good agreement between the results of natural frequencies (refer to Table S2) corroborates the validity of our modelling approach.

The dimension of a breast cell, however, is 100 times bigger than that of the virus under consideration. Thus, to investigate the validity of the current finite element model further, we consider another spherical system (fullerene C_{60}) with dimension 100 times lesser than that of the virus (since the corresponding results of vibration analysis for fullerene C_{60} are available in the literature). We notice from the numerical results presented in Table S3 that a finite element simulation carried out in ANSYS (with the idealization of a spherical structure using brick elements) can produce close results to the fundamental natural frequencies obtained using Raman Spectroscopy and molecular mechanics-based approaches [46-49]. Thus, in the absence of adequate numerical results for validating the natural frequencies, considering two different systems having diameters 100 times more and 100 times less than the diameter of coronaviruses, we show that a finite element approach can produce accurate results for a spherical bio/nano structure. It is expected that the finite element approach would produce accurate results for a spherical structure where the dimension lies in between these two extreme cases. Previous experiences of the authors in this research area also support the observation that the physics behind the low-frequency vibration of a spherical structure is not affected significantly by the change of scale. The two-fold validation presented here provides us adequate confidence to extend the finite element model further for predicting the dynamic behaviour of the present coronavirus structure. The numerical results concerning natural frequencies of coronaviruses can be considered as the first of its kind to be reported in the literature.

SM2. Machine learning modeling and validation

We have explored the dynamic behaviour of coronaviruses following an efficient machine learning assisted framework coupled with finite element approach. The present investigation

involves two different sets of validation. The first validation is for gaining confidence in the adopted finite element approach as described in the preceding section. The second validation concerns the prediction capability of machine learning model, which is used for carrying out simulation-intensive probabilistic quantification and sensitivity analysis for the natural frequencies of coronavirus. The prediction capability of machine learning models is usually assessed by statistical methods (as discussed later in this supplementary document) and scatter plots (refer to figure 4(A-C) of the main paper). We have adopted the support vector regression based machine learning approach, a brief description of which is given in the following paragraphs.

The Support Vector Machines (SVM) is a versatile and robust approach of supervised machine learning. SVM makes use of the kernel tricks to model nonlinear decision boundaries [52 - 53]. SVM was developed as an approach for the classification problems which was further used for the regression problems due to its sparse solution and good generalization. In the present investigation, we have used the nonlinear epsilon-insensitive SVM (ε-SVM) regression. The training dataset consists of predictor variables and observed response values where each sample space is mapped to high dimensional space. The goal of SVR is to reach to a function *f(x)* that deviates from the observed response variables 'y_n' by a value 'ε' and is as flat as possible. For linear function, *f* may be defined as

$$f(x) = wx + b \qquad w \in X, b \in \Re \tag{1}$$

For the minimization of Euclidean norm

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 \\ \text{Subject to} \quad &y_i - wx_i - b \le \varepsilon \\ &wx_i + b - y_i \le \varepsilon \end{aligned} \tag{2}$$

The above problem is applied when function exists and approximates all the elements of sample

space with precise ϵ. SVR can further extend the problem for nonlinear functions, due to the presence of dual formulation. With the implementation of Lagrange multipliers, the standard dualization method is expressed as

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\lambda_i - \lambda_i^*)(\lambda_j - \lambda_j^*)K(x_i, x_j) + \varepsilon\sum_{i=1}^{N}(\lambda_i - \lambda_i^*) - \sum_{i=1}^{N} y_i(\lambda_i - \lambda_i^*)$$

$$\text{Subject to} \quad \sum_{i=1}^{N}(\lambda_n - \lambda_n^*) = 0 \qquad (3)$$

$$\lambda_i, \lambda_i^* \in (0, C)$$

where, C refers to the trade-off between the flatness of the *f(x)* and the threshold deviation larger than ε are tolerated. The new values are predicted using the function

$$f(x) = \sum_{n=1}^{N}(\lambda_n - \lambda_n^*)K(x_n, x) + b \qquad (4)$$

To evaluate the accuracy and performance of SVR based prediction model, the correlation coefficient (R^2 value) is checked. The *R*-square (correlation coefficient) is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5)$$

where, y is the observed value, f is the predicted value, $\bar{y}$ is the mean of observed value, and i denotes the sample index. The value of R^2 provides the statistical signification of closeness of the predicted points to the regression line. It determines goodness of fit for any model, and its value is always in between 0 to 1. As the value of R^2 becomes closer to 1, it is regarded to have a better prediction capability.

Table S4. Accuracy of the SVM models for predicting natural frequencies of coronaviruses

Sample size	1st Natural frequency (R^2)	2nd Natural frequency (R^2)	3rd Natural frequency (R^2)
$N = 16$	0.63	0.72	0.63
$N = 32$	0.96	0.95	0.95
$N = 64$	0.96	0.96	0.94

In the present investigation of coronaviruses, we have used Sobol sequence for generating the set of input parameter, the natural frequencies corresponding to which are obtained using the physics-based finite element model. Three different sample sizes (N) are used to generate the training samples. Based on these training samples, the machine learning models are formed and the prediction capability of the surrogate model is checked using separate samples with respect to the prediction of original simulation model. Table S4 provides the statistical measure of the accuracy of prediction in the form of R^2 for first three natural frequencies. It can be noticed that a sample size of 32 can provide adequate accuracy of prediction. This observation agrees well with the scatter plots presented in figure 4(A-C) of the main paper. Once the machine learning model is formed with adequate accuracy of prediction, it can be regarded as a digital computationally efficient substitute of the original expensive finite element simulation model. In the final stage, the machine learning model could be exploited to predict the responses (i.e. natural frequencies) corresponding to any random set of input parameters, paving the way for carrying out computationally efficient probabilistic quantification and sensitivity analysis.

SM3. Monte Carlo simulation based probabilistic quantification and sensitivity analysis

The probabilistic analysis is carried out in the present investigation using Monte Carlo simulation for obtaining the complete probabilistic descriptions of the natural frequencies. However, since a direct Monte Carlo simulation involving thousands of finite element simulations is computationally expensive and impractical, we have adopted a machine learning based approach as discussed in the preceding section. While preparing the input dataset for carrying out machine learning assisted Monte Carlo simulation, the bound of variation for each of the input material parameters is kept as ±2%. If θ is the considered input parameter then we define the bound as

$\theta_{\min} = \theta(1-\Delta)$ and $\theta_{\max} = \theta(1+\Delta)$, where $\Delta = 0.02$. The i^{th} perturbed sample of Monte Carlo simulation is drawn as $\theta_i = \theta_{\min} + (\theta_{\max} - \theta_{\min})R_i;\ i \in \{1, 2, ...N_{MCS}\}$. Here R_i is a random number generator following a particular probability distribution in the range of 0 to 1. The probability density function plots in figure 4(G-I) are obtained on the basis of the natural frequencies evaluated corresponding to N_{MCS} dataset. Here the machine learning model is used to generate the natural frequencies corresponding to N_{MCS} input dataset instead the original simulation model (i.e. finite element simulation). In the following paragraph, we provide a brief description of the Monte Carlo simulation based stochastic analysis.

Suppose a system is described by a deterministic model providing an input-output relation as follows (in a general framework)

$$w = f(u, v) \tag{6}$$

where *w* is the output vector, *u* is the input vector and *v* is the vector of model parameters. Depending on whether the parameters stochastically vary or not, the model parameter vector can be divided into two groups *x* (deterministic) and *y* (stochastic). Typically the stochastic variable *y* can be described through a probability density function (PDF) denoted as *p*(*y*). The selection of the PDF is based on the information available or should be selected in such a way that could be justified by the maximum information entropy, which suggests a PDF incorporating maximum possible uncertainty into system description. Finally, the probabilistic response of the system such as the expected value of the response can be obtained through a multidimensional integration as follows

$$g = E[f(u, x, y)] = \int f(u, x, y)\, p(y) dy \tag{7}$$

where *g* is the expected value of the system output *w*. The dimension of the integral corresponds to the dimension of *y*. Analytical solution of such integrals is not available (excluding some simple cases) and numerical integration gives inefficient results for the dimensions ≥ 3. Thus, this type of

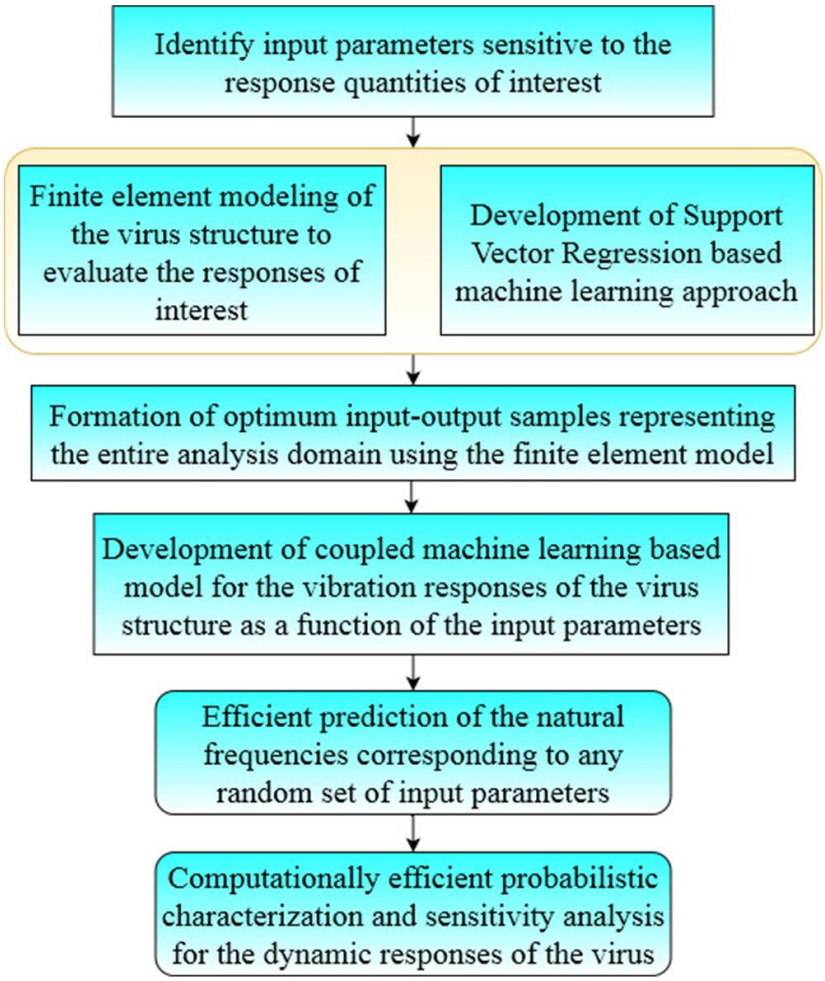

Figure S3. Flowchart of machine learning based prediction algorithm. The plot shows three main stages: generation of optimal input-output dataset for training samples, formation of machine learning model and efficient prediction using the machine learning model for carrying out computationally intensive analyses.

probabilistic integrals are solved using stochastic simulations corresponding to sampling-based methods and such a technique is Monte Carlo simulation (MCS). This method provides a numerical approximation of the integral of the type as in equation (7). The approximate value of stochastic system responses such as expectation g using Central Limit Theorem can be obtained as

$$\hat{g} = \frac{1}{N}\sum_{j=1}^{N} g(u, x, y^j) \tag{8}$$

It is seen that the value of $\hat{g}$ varies as N varies and the value gets closer to theoretical mean as N increases. Now, the question that will arise is how many samples need to be computed or how big N

is. To check if the number of samples is enough for an unbiased value of $\hat{g}$, the coefficient of variation of $\hat{g}$ can be computed using the following expression

$$\delta = \frac{1}{\sqrt{N}} \frac{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(g(u,x,y^j)\right)^2 - \left(\frac{1}{N}\sum_{j=1}^{N} g(u,x,y^j)\right)^2}}{\frac{1}{N}\sum_{j=1}^{N} g(u,x,y^j)} \tag{9}$$

Based on the above expression, often a convergence criterion is adopted to decide the sample size of Monte Carlo simulation (N_{MCS}).

The probabilistic descriptions of four most important individual effects (identified on the basis of sensitivity analysis) and the compound effect of stochasticity are presented in figure 4(G-I) of the main paper. A variance based sensitivity analysis method [56 - 57] is adopted in this study, where the sensitivity indices are calculated as a ratio of the variance (D) of a particular input parameter and the total variance, while the summation of all the sensitivity indices is 1.

$$\begin{aligned} &S_{i_1,\ldots i_s} = \frac{D_{i_1,\ldots i_s}}{D}, 1 \le i_1 \le i_2 \ldots \le i_s \le n \\ &\text{where} \quad \sum S_{i_1} + \sum\left(S_{i_1,i_2} + \ldots + S_{i_1,i_2,\ldots,i_n}\right) = 1 \end{aligned} \tag{10}$$

Here, S_i and D_i represents the sensitivity and variance of i^{th} input parameter. The subscripts with more than one indices show the compound interaction effect of multiple input parameters. The detailed description of forming the machine learning model and the subsequent analyses is depicted in figure S3.