RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2020, **10**, 16607

Machine learning model for fast prediction of the natural frequencies of protein molecules†

Zhao Qin, ‡[a] Qingyi Yu[bcd] and Markus J. Buehler *[a]

Natural vibrations and resonances are intrinsic features of protein structures and enable differentiation of one structure from another. These nanoscale features are important to help to understand the dynamics of a protein molecule and identify the effects of small sequence or other geometric alterations that may not cause significant visible structural changes, such as point mutations associated with disease or drug design. Although normal mode analysis provides a powerful way to accurately extract the natural frequencies of a protein, it must meet several critical conditions, including availability of high-resolution structures, availability of good chemical force fields and memory-intensive large-scale computing resources. Here, we study the natural frequency of over 100 000 known protein molecular structures from the Protein Data Bank and use this dataset to carefully investigate the correlation between their structural features and these natural frequencies by using a machine learning model composed of a Feedforward Neural Network made of four hidden layers that predicts the natural frequencies in excellent agreement with full-atomistic normal mode calculations, but is significantly more computationally efficient. In addition to the computational advance, we demonstrate that this model can be used to directly obtain the natural frequencies by merely using five structural features of protein molecules as predictor variables, including the largest and smallest diameter, and the ratio of amino acid residues with alpha-helix, beta strand and 3–10 helix domains. These structural features can be either experimentally or computationally obtained, and do not require a full-atomistic model of a protein of interest. This method is helpful in predicting the absorption and resonance functions of an unknown protein molecule without solving its full atomic structure.

Received 3rd June 2019
Accepted 3rd April 2020

DOI: 10.1039/c9ra04186a

rsc.li/rsc-advances

Introduction

Protein molecules constitute the basic biopolymers that are found in many different forms in the body of living creatures. A protein molecule is usually composed of one or several polypeptide chains that wind together to form a certain folded three-dimensional (3D) protein structure.[1–5] The formation of protein structure is typically a spontaneous self-folding process, driven by the interaction inside or between the building blocks of the polypeptide chains, known as amino acids.[6,7] There are strong interactions, such as the covalent bonds between the three atoms (N–C_α–C) of the backbone and its interactions with the functional groups at the side chains. There are also weak interactions, such as charge interactions, van der Waals interactions and the hydrogen bonding that often exist between any atoms, accounting for the attraction and repulsion forces between amino acids and chains. The combination of different interactions, coupled with the complex 3D irregular structures, makes the dynamics of the protein molecule much more complicate than its constituting bonds, angles or dihedral angles.[8–11] However, many evidences resulting from structural biology and computational modeling show that there is no intrinsic difference between the vibration of a protein molecule and the vibration of a large-scale structure such as a building[12] responding to external forces, especially for the low-frequency range, which does not involve the charge distribution in the atomic electronic states for optical excitation.[6,11] It is reasonable to believe that the protein's hierarchical structure and the interaction among amino acids affects its vibration, which in

[a]*Laboratory for Atomistic and Molecular Mechanics (LAMM), Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA. E-mail: mbuehler@mit.edu; Tel: +1 617 452 2750*

[b]*Dereck Bok Center for Teaching and Learning, Harvard University, 125 Mount Auburn Street, Cambridge, MA 02138, USA*

[c]*Department of Educational Psychology, Counseling and Special Education, State University of New York at Oneonta, 108 Ravine Pkwy, Oneonta, NY 13820, USA*

[d]*Barnes Center at The Arch, Syracuse University, 150 Sims Drive, Syracuse, NY 13244, USA*

† Electronic supplementary information (ESI) available: File 1: database used for linear model: *combine_name_freq2param_max_min_ss.dat*; File 2: R code for linear model: *R_code*; File 3: database used for ML model: *combine_freq2param_max_min_ss.dat*; File 4: Python code for predicting *b* value: *predict_f0.py*; File 5: Python code for predicting *m* value: *predict_power.py*; Video 1: the dynamics of protein 1FH1 by using both superposition and MD simulations: *Equipartition theorem.mp4*. See DOI: 10.1039/c9ra04186a

‡ Present address: Department of Civil and Environmental Engineering, Syracuse University, Syracuse, NY 13244, United States of America.

turn can also be useful to determine the structure changes or effects.[11] Indeed, it has been shown in earlier literature that the mutation of a single amino acid residue can bring about a significant change in the frequencies of the vibrational modes, indicating that the nature of the low-frequency motions can be unexpectedly sensitive to the specific local geometry.[6,11]

The recent evolution of experimental techniques makes molecular-scale high-temporal resolution imaging feasible,[13] making it more possible to monitor the dynamical behavior of nanostructures. However, up to date, the most advanced imaging technique provides atomic and microsecond resolutions, which is only able to capture the dynamic behavior of 10^6 Hz, or equivalent as 0.00003 cm^{-1}, which is five orders of magnitude smaller than the frequency of the first vibration mode of most protein structures.[14] One way to obtain the vibrational modes of a protein through computational analysis is normal mode analysis (NMA),[8,14] which is a computationally intensive process and it requires the atomistic structure of the protein and the force field that defines all the interatomic interactions.

The key steps in this process include building the Hessian matrix, a $3N \times 3N$ square matrix composed of second-order partial derivatives of the potential energy function with the coordinates of atom in all the three dimensions, and computing the eigenvector and eigenvalue of this matrix. For large protein structures, the time and memory requirement for NMA can be huge. For example, a membrane protein that has more than 1000 amino acids requires more than 3.6 GB memory for simply loading the Hessian matrix, making computing more difficult for larger protein structures. Moreover, the method can only be applied to proteins with known high-resolution 3D structure (which must be obtained either from experiments, *e.g.* by using Nuclear Magnetic Resonance, X-ray Diffraction or Cryo-electron microscopy or could be obtained from a computational simulation of protein folding[15,16]). However, neither of the two strategies are cost and time efficient, and often require significant computational resources.

Artificial intelligence (AI), enabled by machine learning (ML) techniques, has demonstrated its advantage in solving sophisticated scientific problems that involve multiple physics-based interactions that are hard to directly model or non-polynomial problems that require extremely large computational power that cannot be solved by brute force methods.[17] It now provides a novel feasible way of solving such problems by utilizing efficient algorithms to searching a high-dimensional parameter space for optimal solutions. In several recent materials-focused studies, such a data-driven material modeling for optimized mechanics of mechanical properties of materials, it has shown a unique and powerful role.[18–21] Moreover, it shows a breakthrough capability to optimize the multiscale and multi-paradigm architecture of materials for high sensitivity to environmental factors, as needed in sensors, electronics and for multi-purpose material applications.[22]

Here we use a dataset of the structural features and the first 64 normal modes of more than 100 000 protein molecular structures to train and test a machine-learning model that is capable of predicting the frequency spectrum of any protein. This computational model will be helpful that allows to directly obtain a protein's frequency spectrum without knowing the high-resolution 3D structure nor solving for its eigenvalues.

Results and discussion

We developed a parallel code to extract the vibrational feature of 110 511 protein structures available in the Protein Data Bank by applying NMA on each of them,[8] which is used to compute all the normal modes as the most general motion of a protein structure, as well as the natural frequency that corresponds to each of the modes.[23] The collection of the all the discrete natural frequencies, as shown in Fig. 1A, defines the locations of all the peaks of a mechanical spectrum where the mechanical resonance of the structure gets more significant than other frequencies. The key process and the content of code to compute the molecular-based vibrational data from a full-chemistry model are given in a former paper.[23] Besides the frequency spectrum, any possible vibration of a protein structure, such as a thermal fluctuation in a certain temperature, can be given by a superposition of its normal modes.[24] The modes are normal in the sense that each mode is orthogonal to all the other modes, suggesting that the mode cannot be expressed by other modes, that is to say, that an excitation of one mode will never cause motion of a different mode.

Here we hypothesize that the frequency spectrum, which is defined by the chemistry and the folded structure of a protein, provides a unique fingerprint to identify a protein from others. Such a feature has been applied in infrared and Raman spectroscopy to identify the content of small chemical groups. Combining these experiments with the library of the frequency spectrum of proteins may enable to identify a protein only if it has a high-resolution structure deposited in the Protein Data Bank. However, the Protein Data Bank only includes a tiny portion of proteins, with most of them only known for their sequence and functions (such as the 147 413 762 protein sequences given in https://www.uniprot.org[25]) but not for high-resolution protein structures.

Indeed, it is difficult to identify the complex structure of a protein, which requires advanced tools including Nuclear Magnetic Resonance, X-ray Diffraction or Cryo-electron microscopy, as well as protein crystal samples. Hence, the ML technique provides a great opportunity to obtain the frequency spectrum of a protein without experimentally solving for its folded 3D structure. To achieve that, we develop an ML model and train it based on a randomly selected portion (80%) of the NMA calculation results as the first 64 natural frequencies of all the protein structures in the Protein Data Bank.[23] This training set is integrated with another table that summarizes 10 structural features of each protein molecule, including the largest diameter (D_{max}, measured as the largest distance between the two infinite parallel sheets that clamp the molecule), the smallest diameter (D_{min}, measured as the smallest distance between the two infinite parallel sheets that clamp the molecule) and the ratio of all the secondary structures given by DSSP (p_H: content of alpha helix; p_B: content of residue in isolated

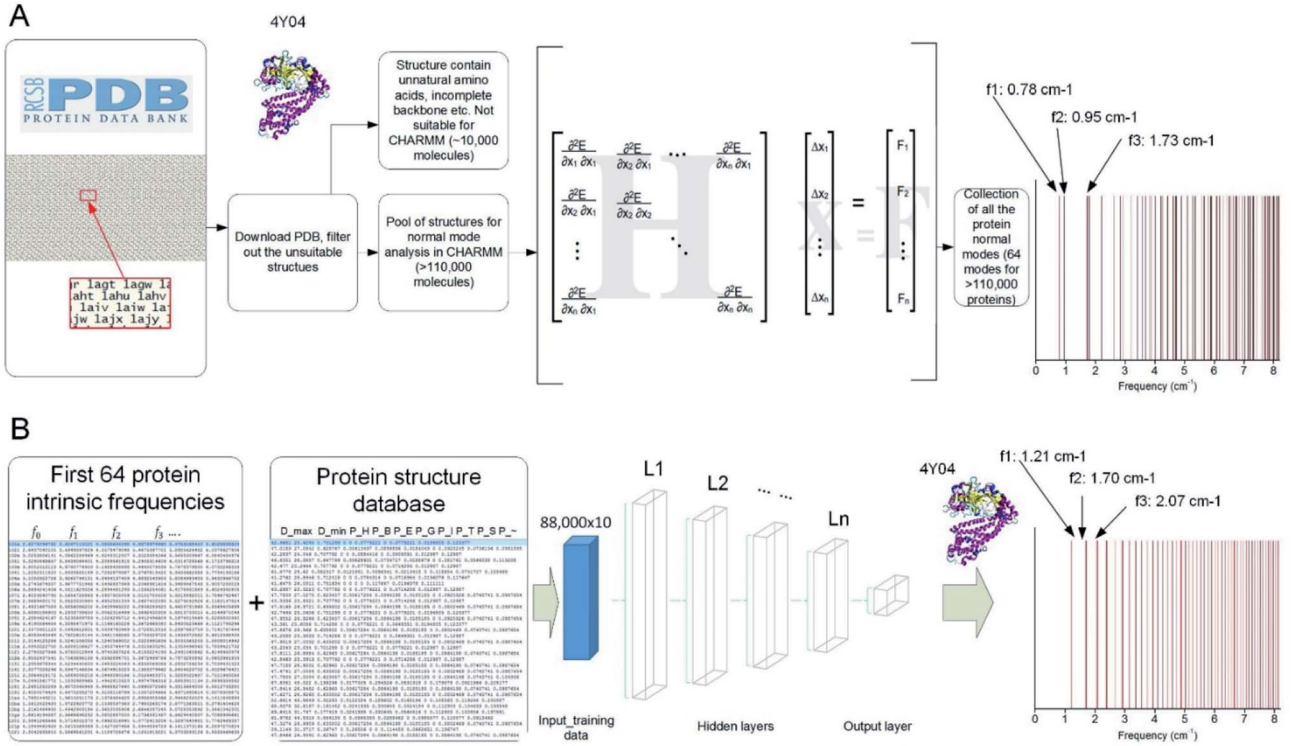

Fig. 1 Overview of the method to run massive in-parallel NMA to obtain the natural frequency spectrum of over 110 000 protein structures present in the Protein Data Bank (PDB) (A) and use the result to train (with 80% of the dataset) and test (with the rest 20% of the dataset) a machine learning model by FNN to predict the frequency spectrum directly by using several structural features (B).

beta-bridge; p_E: content of beta-strand; p_G: content of 3–10 helix; p_I: content of pi-helix; p_T: content of turn; p_S: content of bend and $p_\sim$: content of unstructured parts).

We train the ML model to predict the natural frequencies from the 10 structural features. For the sack of simplicity, we introduce a fitting function for the first 64 natural frequencies f_i ($f0 \ldots 63$) given as

$$f_{i\_\mathrm{fit}} = bi^M \quad (1)$$

Which gives the natural frequency f_{i_fit} for the $(i + 1)$ mode (i starts from 0) where $b = f_0$ for the value of base frequency, M is the power of the function that defines the increment trend of the natural frequencies for larger modes. We also quantitatively compute the standard error of the fitted value for each protein structure by

$$\sigma_{\mathrm{SE}} = \sqrt{\frac{\sum_{i=0}^{63} \left(f_i - f_{i\_\mathrm{fit}}\right)^2}{62}} \quad (2)$$

It has been shown that all the 64 frequencies of each of the protein structures can be fitted by eqn (1) to obtain its $[b, M]$ value for the structure, and the frequencies of over 99% of the protein structure get interpreted with a small standard error $\sigma_{\mathrm{SE}} < 1\ \mathrm{cm}^{-1}$.[23]

By using eqn (1), we train the ML model to predict the value of b and M instead of predicting each of the 64 natural frequencies. The ML is implemented in TensorFlow in Python.[26,27] We start by using the 10 structural features as predictor variables and using the b and M value as the target variables. We use a Feedforward Neuron Network (FNN) with Rectified Linear Unit (ReLU) activation function to read in the training data (80% randomly selected data, ~88 000 records) with a batch size of 10 000 records and develop this data-driven model by minimize the standard error between the predicted and measured $[b, M]$ values by using an Adam optimizer.[28,29] This FNN ML model is realized *via* four hidden layers with 40, 20, 10 and 5 neurons included for each of the layer and a final output layer to evaluate the outcome, as shown in Fig. 1B. We use the remaining 20% data for validation. We run the training for 100 000 epochs, with each of which represents one complete presentation of the training data to be learned by the machine, to ensure there is no further improvement for the optimization. We have tried to increase the number of layers to 10 layers and the width of each layer to 60 neurons and find the increment of depth and width do not improve the standard errors for training and validation, so we keep using this ML architecture for our study.

As a result of the validation, we obtain the predicted $[b, M]$ of the protein structures within the validation dataset. Using the values, we compute each of the natural frequencies as $f_{i\_\mathrm{ML}} = bi^M$ and compare with their measurement result, with the

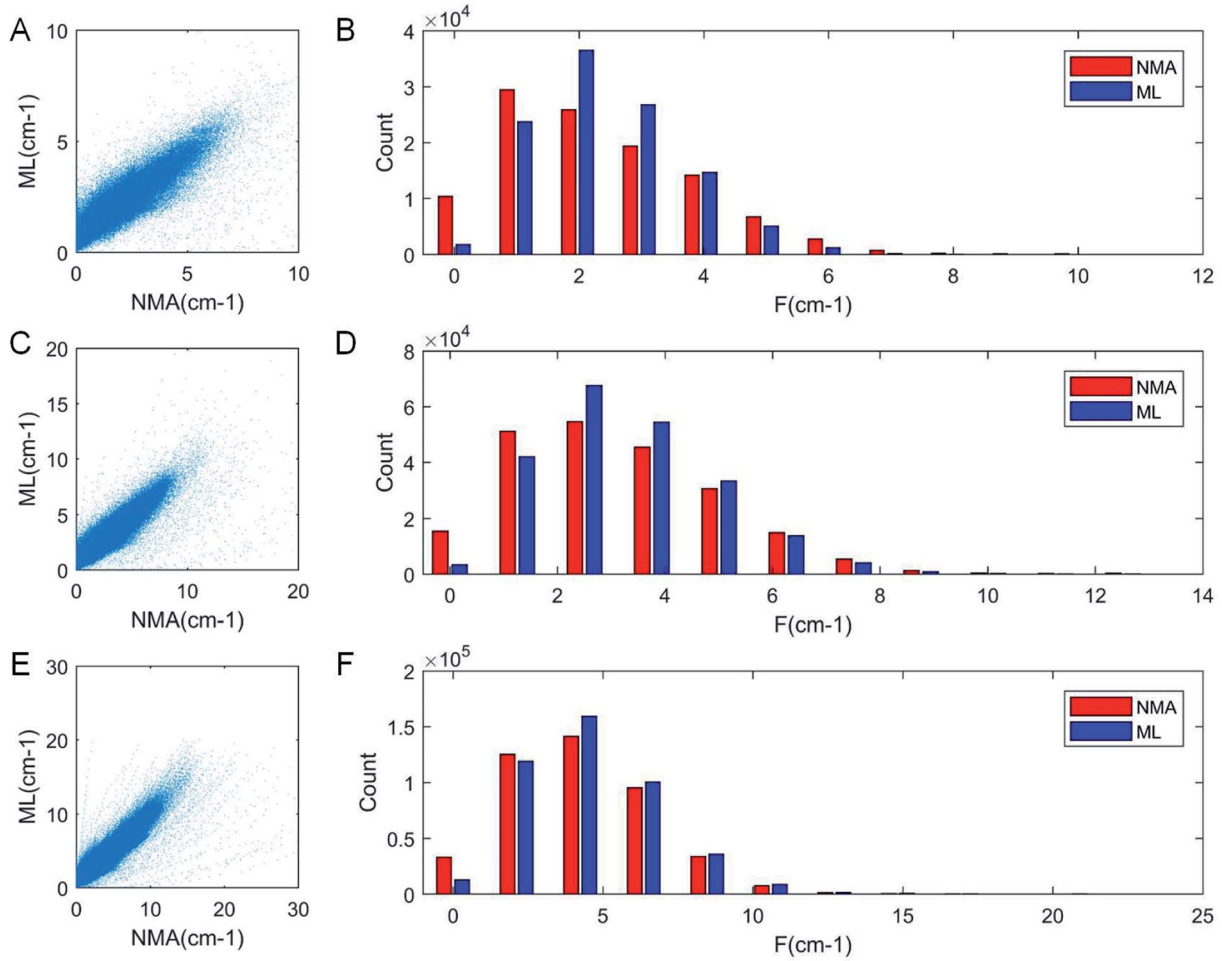

Fig. 2 Comparison of the frequencies predicted by ML and directly obtained from NMA that corresponds to the first 5 (A), 10 (C) and 20 modes (E) of all the protein structures for testing (22 000 structures), with the histogram of the ML and NMA results summarized by bars of different colors in (B), (D) and (F) respectively. It is consistently shown that for low-frequency modes, the distribution has a log-normal shape, with a center peak, that corresponds to most normal modes, concentrates at a low-frequency region.

comparison outcome summarized in Fig. 2. The comparison between the frequencies of the first 5 modes, 10 modes, and 20 modes are given in Fig. 2A, C and E, respectively, with each point $(f_{i\_\mathrm{NMA}}, f_{i\_\mathrm{ML}})$ plotted in the figure. It is shown that the ML prediction and the NMA data have a very strong correlation and most data point concentrated along the line $f_{i\_\mathrm{NMA}} = f_{i\_\mathrm{ML}}$, suggesting the overall good predicting result. Similar to eqn (2), the different between the two dataset is quantified by using the standard error as $\sigma = \sqrt{\frac{\sum_{i=0}^{n} (f_{i\_\mathrm{NMA}} - f_{i\_\mathrm{ML}})^2}{n}}$, where n is the total number of data points in the plot. We find the standard error of computing the first 5, 10 and 20 modes is 0.836, 0.860 and 1.256 cm^{-1}, respectively. The standard error increases for having more modes as the natural frequency increase for higher modes. The prediction error $\sigma = 1.256\ \mathrm{cm}^{-1}$ is still considered relatively low because the mean value and standard deviation of the natural frequency of the 20$^{\mathrm{th}}$ mode of all the protein structures is $6.96 \pm 3.50\ \mathrm{cm}^{-1}$. The histograms of the predicted frequencies, as given in Fig. 2B, D and F for the first 5, 10 and 20 modes, respectively, show a different trend as the more modes yields statistically better agreement between the prediction and measurement, suggesting the larger prediction error comes merely from the nature of the larger data value for higher modes. Moreover, it is interesting to see that the distribution of the frequencies is log-normal all the time, suggesting that the relatively low frequencies (corresponding to the peak of the log-normal distribution) correspond to more normal modes and also account more for mechanical resonance and energy absorption, because of the equipartition principle,[24] which states that each normal mode takes equal kinetic energy for free vibration.

The result is important because it demonstrates that the ML model is able to directly predict the natural frequencies of a protein structure, instead of using the high-resolution 3D structure and NMA calculation or dynamic simulations. The ML model is validated by the testing data and can be applied to any protein molecule with unknown 3D structure but knowing the few structural features. For example, the minimum and

maximum diameters can be measured by dynamic light scattering (DLS), and the secondary structure ratio can be extracted from Fourier Transform Infrared (FTIR) spectroscopy and these experiments are much easier than solving the 3D structure of a protein. It thus provides a convenient way of predicting the spectrum of natural frequencies of a protein. However, many of the eight secondary structures that are defined for FTIR are much less clear than what is defined for the full atomistic protein structures by the Dictionary of Protein Secondary Structure (DSSP),[30] as what is used for the predictor variables. For instance, helix and beta sheet structures are usually better defined than other coiled or random structures for FTIR.[31]

It is useful to understand the contribution of each of the ten predictor variables and find out which are more essential than the others in prediction. Although FNN provides a convenient way of predicting the $[b, M]$ from $[D_{max}, D_{min}, p_H, p_B, p_E, p_T, p_I, p_G, p_S, p_{\sim}]$, the relationship among each target variable and the predictor variables are highly nonlinear and difficult to quantify. We thus take the linear model and compute the pairwise correlation function between the twelve variables by using all the data records, as shown in Fig. 3A. As expected, not all of the predictor variables correlate much to $[b, M]$. b correlates more to $[D_{max}, D_{min}, p_H]$ than others while m correlates more to $[D_{min}, p_E, p_G]$ than others. Using the linear model,[32] we increase the number of predictor variables n_{var} from 1 to 10 and for each n_{var} we test all the possible subsets of predictor variables for ${C_{10}}^{n_{var}}$ times of tests. In total, we perform ${C_{10}}^1 + {C_{10}}^2 + {C_{10}}^3 + \ldots + {C_{10}}^{10} = 2^{10} = 1024$ tests in total. For each n_{var}, we select the best subsets for predicting b (Fig. 3B and C) and M (Fig. 3D and E) values with maximum coefficient of determination (rsq) value, which has value between 0 and 1 and measures the percentage of the response variable variation that is explained by a linear model. It is shown in Fig. 3B and D that the contribution of a unit increment in n_{var} keeps decrease for larger n_{var} value in predicting b and M, respectively. Actually, for $n_{var} > 5$, the further increment of n_{var} has only a small effect. Moreover, by running these 1024 tests, we manage to obtain the best subsets for each n_{var} value, as shown in Fig. 3C and E for b and M, respectively. It is shown that if we are only accessible to one predictor variable, D_{max} and D_{min} will be the most predictive for b, M, respectively. By using five predictor variables, the combination of $[D_{max}, D_{min}, p_H, p_E, p_G]$ will be the most predictive.

We use the knowledge obtained from a linear model to reduce the number of predictor variables of our ML model. We test n_{var} from 1 to 10 and the types of predictor variables are selected according to their importance as given in Fig. 3C. We use the selected predictor variables to train the ML model based on the selected columns of the training data by using the same learning architecture and compute the mean-square error of the testing data during the optimization process, as given in Fig. 4A and B. It is shown that 2000 epochs are enough to yield overall converged mean-square error and $n_{var} > 5$ does not yield a significant difference from $n_{var} = 5$, as what is predicted by the linear model. It is also shown that for $n_{var} = 5$, the ML predicted $[b, M]$ values well agree with the measured $[b, M]$ value of the testing data, as shown in Fig. 4C and D, respectively. The high correlation coefficient (0.88 for b and 0.78 for M) and low standard error (0.35 cm^{-1} for b and 0.084 for M) can be computed from Fig. 4C and D, which support the accuracy of the prediction. In addition to the comparison of the fitting parameters, we can expand the predicted $[b, M]$ values back to f_{i_ML} for each of the protein structure within the testing dataset and compute the standard error σ with f_{i_NMA} as given in Fig. 4E. It is shown that $n_{var} = 5$ has an advantage in reaching small σ

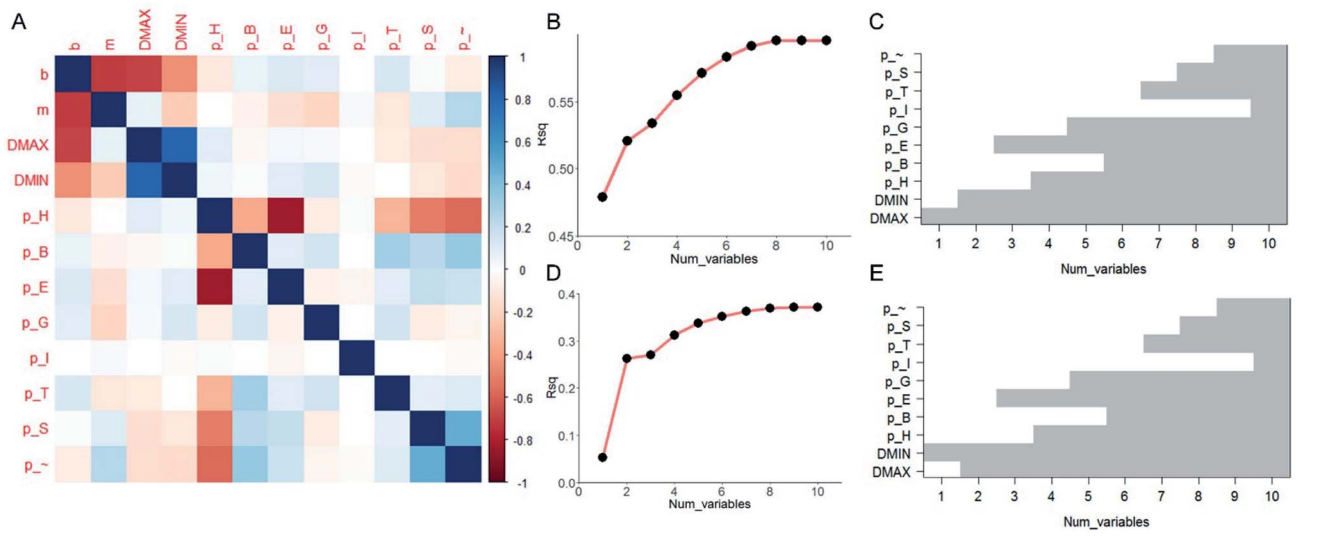

Fig. 3 Results by linear analysis. (A) Visualization of the correlation matrix between pairs out of the twelve variables, including two target variables $[b, M]$ and ten predictor variables $[D_{max}, D_{min}, p_H, p_B, p_E, p_T, p_I, p_G, p_S, p_{\sim}]$, computed by using all the ~110 000 data records. We study the number of predictor variables from 1 to 10 by testing all the possible subsets of predictor variables with 1024 tests in total and present the best subsets that yield the highest coefficient of determination (rsq) value for explaining b (B) and M (D) values. The best subsets are presented in (C) and (E) for b and M values, respectively by the gray column, which shows for a certain number of predictor variables (1 to 10), which variables are included in the best subset that yields the highest rsq value.

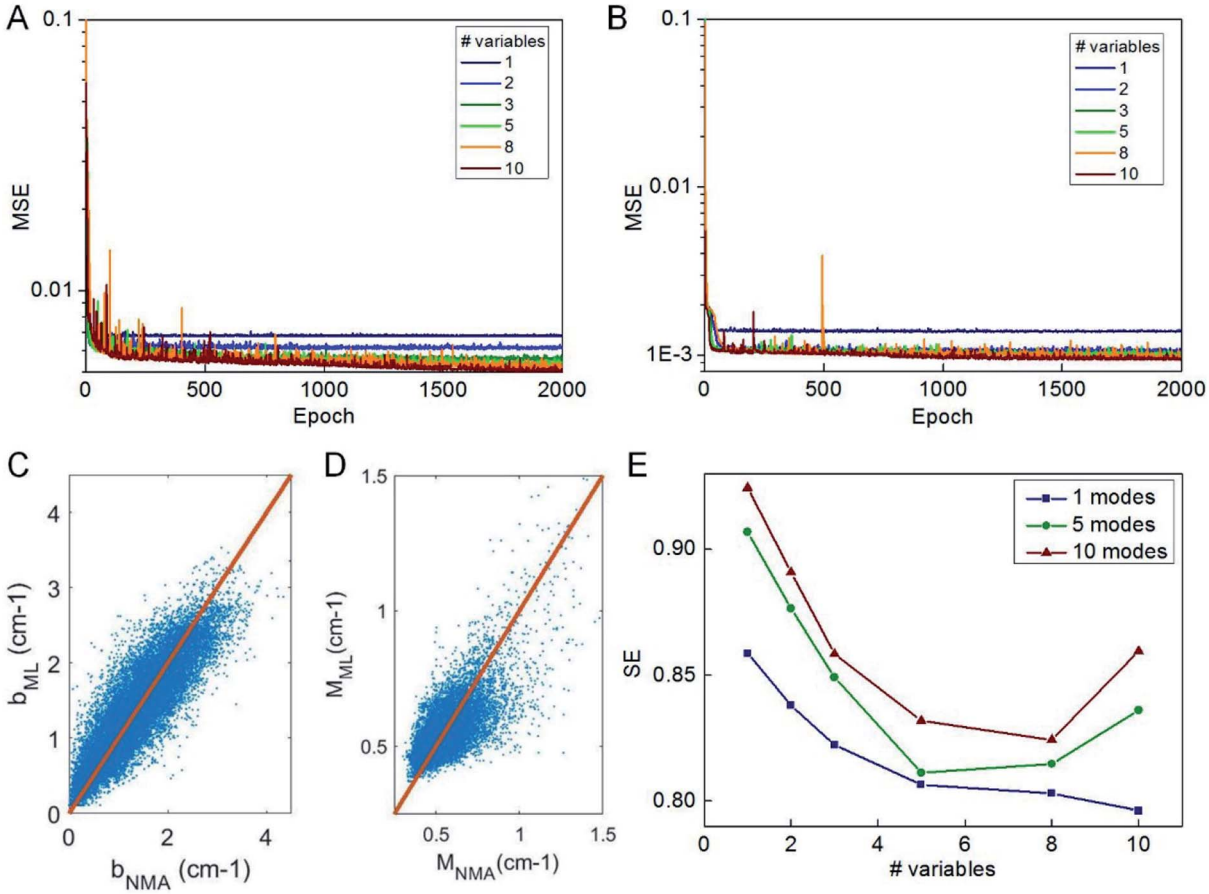

Fig. 4 Performance of ML train by a different number of predictor variables. The MSE computed during the training of predicting *b* (A) and *M* (B) values by using the number of predictor variables from 1 to 10 with the FNN model. The comparison between the ML predicted *b* (C) and *M* value (D) by using the five predictor variables, [D_{max}, D_{min}, p_H, p_E, p_C], as suggested by the linear model, with the NMA result for the 22 000 structures for testing. (E) The standard error (SE) for comparison of the frequencies predicted by ML (trained by the five predictor variables) and directly obtained from NMA that corresponds to the first 1, 5 and 10 modes of all the protein structures for testing.

value for several different comparisons, including comparing of f_1 *per se*, as well as from f_1 to f_5 and from f_1 to f_{10}.

For any protein sequences, recent progress in folding algorithms makes it possible to accurately predict their 3D atomic structures.[33,34] Combining with classical force fields, these structures yields a Hessian matrix that is often too big to solve for their eigenvalues by NMA. The method proposed here provides an efficient way to obtain the eigenvalues without NMA. Thus, it can combined with the recent innovative method of solving eigenvectors from eigenvalues[35] and provide all the normal frequencies and normal modes of a protein. We can use the information to predict the dynamic behavior of a protein. For example, for any protein, we obtain the i^{th} normal modes as a $\{x_i,y_i,z_i\}$ normal frequency f_i. By using the equipartition theorem, we can compute that each mode has the equal kinetic energy parts as $(3/2)Nk_BT/M$ for the first M modes of a protein structure under the temperature of T, where N is the total number of atoms and k_B is the Boltzmann constant. Because each of the normal modes are normal to the other modes, the velocity distribution for atom j in mode i is v_{ij} and should yield

$$\sum_j \frac{1}{2} m_j {v_{ij}}^2 = \frac{3}{2M} N k_B T. \quad (3)$$

Considering the atomic velocity is also given by the magnitude of the particle vibration as $\overrightarrow{v_{ij}} = \overrightarrow{a_{ij}} 2\pi f_i \cos(2\pi t f_i + t_{0i})$, and the magnitude of vibration A_i is given by the $\overrightarrow{a_{ij}} = A_i\{x_j, y_j, z_j\}_i$ as $\{x_j,y_j,z_j\}_i$ the i^{th} normal mode on atom j. Putting these relations to eqn (3), we obtain the magnitude of vibration as

$$A_i = \frac{1}{2\pi f_i} \sqrt{\frac{3Nk_BT}{M\sum_j m_j\left[{x_j}^2 + {y_j}^2 + {z_j}^2\right]}} \quad (4)$$

This equation can be used to calculate the thermal vibration of the protein structure without running molecular dynamics simulations. With the coordinate of each atom given by a function of time t

$$\overrightarrow{x_j} = \sum_{i=1}^{M} A_i\{x_j, y_j, z_j\}_i \sin(2\pi t f_i + t_{0i}) \quad (5)$$

For example, we focus on the protein structure with PDB ID 1FH1 and perform this superposition analysis by using the predicted normal frequencies and obtain its thermal vibration spectrum at 300 K as given in ESI Video 1.†

Discussion

The collection of the natural frequencies of a protein gives the location of all the peaks of the mechanical spectrum where the mechanical resonance of the structure gets more significant than other frequencies, which represents an important fingerprint to identify the protein. Although NMA provides an accurate way to measure all the natural frequencies, it is considered as a very expensive process because of the difficulty in obtaining the high-resolution atomic structure and the time and resource to solve the eigenvalues of the elastic matrix. Here, we demonstrate that ML can be used to directly obtain the natural frequencies without using the atomic structures or solving the eigenvalues, but by merely using several structural features that can be easily obtained. We use a linear model to reduce the number of predictor variables and find that five variables including the largest and smallest diameter, the ratio of amino acid with alpha-helix, beta strand and 3–10 helix domains[36] can be used to predict the natural frequencies with a small standard error.

The selection of the initial structural features, including diameter and secondary structure ratios is more intuitive than fully rational. The reasons are the availability of experimental measurements (DLS, FTIR, *etc.*) and their consistency with *in silico* measurements of atomic protein structures. It forms the fundamental basis of training an ML model with the computational data, which provides the only feasible way of generating massive data records for training and validation. Using linear models, we have successfully reduced the number of predictor variables to five. It is possible that there can be other structure features that can better predict the natural frequencies but the features are not included. Considering the automatic generation of the database and rational analysis based on ML and linear models, as long as these features can be extracted from the high-resolution protein structures, this ML model can be easily updated by considering other geometric features beyond the current ones.

The method proposed here can serve as a useful and computationally efficient approach to predict the absorption and resonance functions of the protein molecule with unknown structure and sequence. It can be useful in categorizing the dynamic function of unknown proteins according to their structural features. This method, combined with the recently evolved techniques that allow prediction of protein structure from its sequence,[15,16] could facilitate predicting the natural frequencies of all the known protein sequences (such as sequences presented in the UniProt,[25] with 99.9% of them not included in the Protein Data Bank), which may be a useful resource to identify protein type according to its frequency spectrum.

Looking ahead, our new method could also be applied to predict the vibrational features of other nanoscale objects such as nanoparticles. It may also be used to extract other material properties besides vibrational spectra, and as such, be broadly applied in a materials-by-design approach where geometric features are identified to achieve certain material properties. Other applications may include material sonification methods, to offer a rapid method to create audible representation of proteins in musical form.[23]

Methods

Normal model analysis

At the foundation of the method, we build a full database of the first 70 normal modes of each of 110 511 natural protein structures composed of only standard amino acids out of the full list of more than 130 000 structures that are currently available in the Protein Data Bank. We developed a bash script that allows integrating multiple open source software with the CHARMM c37b1 program to automatically download, clean and analyze each of the protein molecular structure. The details of the bash script are given in our earlier paper.[23]

We use a Block Normal Mode (BNM) method[8,37] in CHARMM for normal mode analysis on each of the protein structure. BNM projects the full atomic hessian matrix into a subspace spanned by the eigenvectors of blocks, as each block is defined by an amino acid. We save the first 70 modes with lowest frequencies, ordered from lower to higher frequency, of each protein molecule.

The first 6 modes always have zero eigenvalue and zero frequency because they correspond to the rigid-body movement and rotation of the molecule. The higher order modes 7 and onwards, *i.e.* the last 64 normal modes amongst the 70 generated, describe the molecular deformation. We hence only use the frequency value of the last 64 modes out of these 70 ones. These frequency values are denoted in sequence as $f_0, f_1, f_2, \ldots f_{63}$.

Linear model analysis

We use *R* to study the number of predictor variables from 1 to 10 by testing all the possible subsets of predictor variables with 1024 tests in total and present the best subsets that yield the highest coefficient of determination value for explaining *b* and *M* values. We share the links to the database and *R* code of this work.

Feedforward neural network

The machine learning model is implemented in TensorFlow in Python. We use a Feedforward Neuron Network (FNN) with Rectified Linear Unit (ReLU) activation function to read in the training data (80% randomly selected data, ~88 000 records) with a batch size of 10 000 records and develop this data-driven model by minimize the standard error between the predicted and measured [*b*, *M*] values by using an Adam optimizer. This FNN ML model is realized *via* four hidden layers with 40, 20, 10 and 5 neurons included for each of the layer and a final output layer to evaluate the outcome. We use the remaining 20% data for validation. We run the training for 100 000 epochs, with

each of which represents one complete presentation of the training data to be learned by the machine, to ensure there is no further improvement for the optimization. We share the database and python codes of this ML model as ESI.†

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported by ONR (grant #N00014-16-1-2333) and NIH U01 EB014976. Additional support from the Army Research Office (ARO) 73793EG is acknowledged.

References

1 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The Protein Data Bank, *Nucleic Acids Res.*, 2000, **28**(1), 235–242.

2 T. Lazaridis and M. Karplus, "New View" of Protein Folding Reconciled with the Old through Multiple Unfolding Simulations, *Science*, 1997, **278**(5345), 1928–1931, DOI: 10.1126/science.278.5345.1928.

3 E. Paci and M. Karplus, Unfolding Proteins by External Forces and Temperature: The Importance of Topology and Energetics, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**(12), 6521–6526, DOI: 10.1073/pnas.100124597.

4 Z. Qin, L. Kreplak and M. J. Buehler, Hierarchical Structure Controls Nanomechanical Properties of Vimentin Intermediate Filaments, *PLoS One*, 2009, **4**(10), DOI: 10.1371/journal.pone.0007294.

5 M. J. Buehler and S. W. Cranford, *Biomateriomics*, Springer, Netherlands, 2012.

6 C. Rischel, D. Spiedel, J. P. Ridge, M. R. Jones, J. Breton, J. C. Lambry, J. L. Martin and M. H. Vos, Low Frequency Vibrational Modes in Proteins: Changes Induced by Point-Mutations in the Protein-Cofactor Matrix of Bacterial Reaction Centers, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**(21), 12306–12311, DOI: 10.1073/pnas.95.21.12306.

7 T. Ackbarow, X. Chen, S. Keten and M. J. H. Buehler, Multiple Energy Barriers, and Robustness Govern the Fracture Mechanics of -Helical and Beta-Sheet Protein Domains, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(42), 16410–16415, DOI: 10.1073/pnas.0705759104.

8 F. Tama, F. X. Gadea, O. Marques and Y. H. Sanejouand, Building-Block Approach for Determining Low-Frequency Normal Modes of Macromolecules, *Proteins*, 2000, **41**(1), 1–7.

9 Z. P. Xu, R. Paparcone and M. J. Buehler, Alzheimer's A Beta(1-40) Amyloid Fibrils Feature Size-Dependent Mechanical Properties, *Biophys. J.*, 2010, **98**(10), 2053–2062, DOI: 10.1016/j.bpj.2009.12.4317.

10 M. Karplus and J. Kuriyan, Molecular Dynamics and Protein Function, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(19), 6679–6685, DOI: 10.1073/pnas.0408930102.

11 C. H. M. Rodrigues, D. E. V. Pires and D. B. Ascher, DynaMut: Predicting the Impact of Mutations on Protein Conformation, Flexibility and Stability, *Nucleic Acids Res.*, 2018, **46**, W350–W355, DOI: 10.1093/nar/gky300.

12 H. Sun and O. Büyüköztürk, Probabilistic Updating of Building Models Using Incomplete Modal Data, *Mech. Syst. Signal Process.*, 2016, **75**, 27–40, DOI: 10.1016/j.ymssp.2015.12.024.

13 S. S. Wang, Z. Qin, G. S. Jung, F. J. Martin-Martinez, K. Zhang, M. J. Buehler and J. H. Warner, Atomically Sharp Crack Tips in Monolayer MoS_2 and Their Enhanced Toughness by Vacancy Defects, *ACS Nano*, 2016, **10**(11), 9831–9839, DOI: 10.1021/acsnano.6b05435.

14 N. Go, T. Noguti and T. Nishikawa, Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational Modes, *Proc. Natl. Acad. Sci. U. S. A.*, 1983, **80**(12), 3696–3700.

15 S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović and F. Players, Predicting Protein Structures with a Multiplayer Online Game, *Nature*, 2010, **466**, 756–760, DOI: 10.1038/nature09304.

16 S. Conchúir, K. A. Barlow, R. A. Pache, N. Ollikainen, K. Kundert, M. J. O'Meara, C. A. Smith and T. Kortemme, A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design, *PLoS One*, 2015, **10**(9), e0130433, DOI: 10.1371/journal.pone.0130433.

17 D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, Mastering the Game of Go with Deep Neural Networks and Tree Search, *Nature*, 2016, **529**, 484–489, DOI: 10.1038/nature16961.

18 G. X. Gu, C. T. Chen, D. J. Richmond and M. J. Buehler, Bioinspired Hierarchical Composite Design Using Machine Learning: Simulation, Additive Manufacturing, and Experiment, *Mater. Horiz.*, 2018, **5**(5), 939–945, DOI: 10.1039/c8mh00653a.

19 P. Z. Hanakata, E. D. Cubuk, D. K. Campbell and H. S. Park, Accelerated Search and Design of Stretchable Graphene Kirigami Using Machine Learning, *Phys. Rev. Lett.*, 2018, **121**, 255304, DOI: 10.1103/physrevlett.121.255304.

20 G. X. Gu, C. T. Chen and M. J. Buehler, De Novo Composite Design Based on Machine Learning Algorithm, *Extreme Mechanics Letters*, 2018, **18**, 19–28, DOI: 10.1016/j.eml.2017.10.001.

21 C. H. Yu, Z. Qin, F. Martin-Martinez and M. J. Buehler, A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using AI, *ACS Nano*, 2019, **13**, 7471–7482, DOI: 10.1021/acsnano.9b02180.

22 M. Popova, O. Isayev and A. Tropsha, Deep Reinforcement Learning for de Novo Drug Design, *Sci. Adv.*, 2018, **4**(7), eaap7885, DOI: 10.1126/sciadv.aap7885.

23 Z. Qin and M. J. Buehler, Analysis of the Vibrational and Sound Spectrum of over 100,000 Protein Structures and Application in Sonification, *Extreme Mechanics Letters*, 2019, 100460, DOI: 10.1016/j.eml.2019.100460.

24 W. Greiner, D. Rischke, L. Neise and H. Stöcker, *Thermodynamics and Statistical Mechanics; Classical Theoretical Physics*, Springer, New York, 2000.
25 Consortium and T. U. UniProt, A Worldwide Hub of Protein Knowledge, *Nucleic Acids Res.*, 2019, **47**(D1), D506–D515, DOI: 10.1093/nar/gky1049.
26 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin and S. Ghemawat, *et al.*, TensorFlow: A System for Large-Scale Machine Learning, *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, USENIX Association, Savannah, GA, USA, 2016, pp. 265–283.
27 G. Van Rossum and F. L. Drake, *Python Tutorial, Technical Report CS-R9526*, 1995, DOI: 10.1016/j.abb.2004.09.015.
28 J. Schmidhuber, Deep Learning in Neural Networks: An Overview, *Neural Network.*, 2015, **61**, 85–117, DOI: 10.1016/j.neunet.2014.09.003.
29 V. Nair and G. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in *Proceedings of the 27th International Conference on Machine Learning*, 2010.
30 W. Kabsch and C. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features, *Biopolymers*, 1983, **22**(12), 2577–2637, DOI: 10.1002/bip.360221211.
31 H. Yang, S. Yang, J. Kong, A. Dong and S. Yu, Obtaining Information about Protein Secondary Structures in Aqueous Solution Using Fourier Transform IR Spectroscopy, *Nat. Protoc.*, 2015, **10**(3), 382–396, DOI: 10.1038/nprot.2015.024.
32 R Developement Core Team, *R: A Language and Environment for Statistical Computing*, 2008.
33 Z. Qin, L. Wu, H. Sun, S. Huo, T. Ma, E. Lim, P. Y. Chen, B. Marelli and M. J. Buehler, Artificial Intelligence Method to Design and Fold Alpha-Helical Structural Proteins from the Primary Amino Acid Sequence, *Extreme Mechanics Letters*, 2020, **36**, 100652.
34 A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, *et al.*, Improved Protein Structure Prediction Using Potentials from Deep Learning, *Nature*, 2020, **577**(7792), 706–710, DOI: 10.1038/s41586-019-1923-7.
35 P. B. Denton, S. J. Parke, T. Tao and X. Zhang *Eigenvectors from Eigenvalues: A Survey of a Basic Identity in Linear Algebra*, 2019, arXiv e-prints, arXiv:1908.03795.
36 R. A. Silva, S. C. Yasui, J. Kubelka, F. Formaggio, M. Crisma, C. Toniolo and T. A. Keiderling, Discriminating 310- from α-Helices: Vibrational and Electronic CD and IR Absorption Study of Related Aib-Containing Oligopeptides, *Biopolymers*, 2002, **65**(4), 229–243, DOI: 10.1002/bip.10241.
37 L. Ruiz, W. J. Xia, Z. X. Meng and S. Keten, A Coarse-Grained Model for the Mechanical Behavior of Multi-Layer Graphene, *Carbon*, 2015, **82**, 103–115, DOI: 10.1016/j.carbon.2014.10.040.