The Journal of Chemical Physics

Prediction of amyloid aggregation rates by machine learning and feature selection

Cite as: J. Chem. Phys. **151**, 084106 (2019); https://doi.org/10.1063/1.5113848
Submitted: 08 June 2019 . Accepted: 01 August 2019 . Published Online: 28 August 2019

Wuyue Yang, Pengzhen Tan, Xianjun Fu, and Liu Hong

View Online Export Citation CrossMark

Prediction of amyloid aggregation rates by machine learning and feature selection

Cite as: J. Chem. Phys. 151, 084106 (2019); doi: 10.1063/1.5113848
Submitted: 8 June 2019 • Accepted: 1 August 2019 •
Published Online: 28 August 2019

View Online Export Citation CrossMark

Wuyue Yang,[1] Pengzhen Tan,[1] Xianjun Fu,[2] and Liu Hong[1,a)]

AFFILIATIONS

[1] Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing 100084, China
[2] Institute for Literature and Culture of Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan 250355, China

[a)] Author to whom correspondence should be addressed: zcamhl@tsinghua.edu.cn

ABSTRACT

A novel data-based machine learning algorithm for predicting amyloid aggregation rates is reported in this paper. Based on a highly nonlinear projection from 16 intrinsic features of a protein and 4 extrinsic features of the environment to the protein aggregation rate, a feedforward fully connected neural network (FCN) with one hidden layer is trained on a dataset composed of 21 different kinds of amyloid proteins and tested on 4 rest proteins. FCN shows a much better performance than traditional algorithms, such as multivariable linear regression and support vector regression, with an average accuracy higher than 90%. Furthermore, by the correlation analysis and the principal component analysis, seven key features, folding energy, HP patterns for helix, sheet and helices cross membrane, pH, ionic strength, and protein concentration, are shown to constitute a minimum feature set for characterizing the amyloid aggregation kinetics.

I. INTRODUCTION

Many important neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease, and Mad Cow disease, are all closely related to the aggregation of amyloid proteins.[1–3] To make a quantitative description of the process, the aggregation rate plays a central role,[4,5] which also shows a great impact on the progress of amyloidosis, the choice of inhibitors, the strategy of medical treatment, etc.[6,7] Experimentally, the aggregation rate is generally measured through the time trajectories of ThT fluorescence intensity, a reflection of the fiber content. In the current paper, we are going to show that the data based machine learning algorithm provides a new way to make reliable estimations.

In recent years, machine learning, especially deep learning,[8,9] has shown an outstanding performance in applications in many diverse areas, such as image and speech recognition,[10,11] pilotless automobiles,[12] and Go playing.[13,14] More importantly, it provides a promising way to uncover the relations deeply buried inside the data without the help of a human based preknowledge. So, can we take an advantage of machine learning to automatically uncover the rules governing the amyloid aggregation kinetics and to make quantitative reliable predictions on the aggregation rate without referring to either experiments or physical models? This not only is a question of scientific interest but also has practical application values in amyloid-related biochemistry and medicine.

In the past, there were already several early attempts toward this direction. Chiti *et al.*[15] showed that the effect of specific amino acid mutations on the aggregation rates of unfolded polypeptides was correlated to a remarkable extent with changes in simple physicochemical properties, such as hydrophobicity (HB), secondary structure propensity, and charge. This conclusion was further confirmed by the study of DuBay *et al.*,[16] in which they showed that the absolute aggregation rate of unstructured proteins could be determined by linear regression based on a knowledge of the hydrophobicity, hydrophobic patterns, charge, ionic strength (IS), pH, and protein concentration. Various existing prediction algorithms on amyloid aggregation were summarized and tested with *in vivo* data by Belli *et al.*[17] In this study, we have made an alternative preliminary attempt by using the feedforward fully connected neural network (FCN) to see to what extent we can predict the absolute aggregation rate of amyloid proteins based on their given physiochemical properties.

Amyloid aggregation, in nature, is a self-assembling process of amyloid proteins.[1–3] Unlike protein folding and misfolding, which are governed by intramolecular interactions, amyloid aggregation is a consequence of the balance between intermolecular and intramolecular interactions, such as hydrophobic interaction, hydrogen bond, and π-π stacking.[18,19] Since all of them belong to weak interactions, many factors may contribute to the aggregation process to a remarkable extent.

According to their different origins, the physiochemical features of amyloid proteins can be roughly classified into two groups—the intrinsic and the extrinsic.[16] The intrinsic features are those mainly determined by the amyloid protein itself, e.g., the sequence length, molecular weight, amino acid composition, isoelectric point (pI), hydrophobicity, radius of gyration, secondary structure, total charge in the native state, misfolded state, and so on (strictly speaking, some features, such as the total charge, also vary with the temperature or pH value. However, here we only use their values in the native state, such that they can be regarded as intrinsic too.) The intrinsic features are a reflection of the physical, chemical, sequential, and structural properties of amyloid proteins and can be identified even independent of the aggregation process. The extrinsic features are directly related to the setup of experiments, such as the temperature, pH value, salt concentration, protein concentration, buffer condition, stirring rate, metals,[20] crowding agents,[21] lipid bilayers,[22] and so on. They comprise a detailed description of the conditions under which we perform our studies.

Now the major task is to construct the nonlinear projection from those intrinsic and extrinsic features to the amyloid aggregation rate, which is accomplished through a feedforward fully connected neural network. The materials and methods of our study, including the collection of kinetic data and protein features, the preparation of training set and test set, the basic setup of FCN, as well as brief introductions on other classical machine learning algorithms, are listed in Sec. II. Our main result, the predictive ability of FCN on the current problem in comparison with other algorithms, is discussed in detail in Sec. III, with a highlight on a minimal feature set for making reliable predictions achieved by the correlation analysis and the principle component analysis (PCA). Section IV includes a short conclusion and discussions on potential generalizations of FCN.

II. MATERIALS AND METHODS

A. Intrinsic and extrinsic features for amyloid aggregation

As we have claimed, many features may play a nonnegligible role during the aggregation process of amyloid proteins, including both the intrinsic physical, chemical, sequential, and structural information of the protein itself and the extrinsic experimental conditions. In the current study, far from complete, we have collected 16 intrinsic features—the sequence length (N), molecular weight (M_w), isoelectric point (pI), fraction of hydrophobic residues (F_h), hydrophobicity (HB), helix content (F_a), sheet content (F_b), coil content (F_c), fraction of proline residues (F_p), radius of gyration (R_g), number of nonlocal contacts (N_c), folding energy [ΔE, Chemistry at HARvard Macromolecular Mechanics (CHARMM) force field], total charge (Q), HP pattern for α-helix (HP_a), HP pattern for β-sheet (HP_b), and HP pattern for helices cross membrane (HP_m), as well as 4 extrinsic factors—temperature (T), pH value, ionic strength (IS) and protein concentration (C), respectively. Especially, the molecular weight, number of nonlocal contacts, folding energy, total charge, and radius of gyration are all normalized by the sequence length in order to remove this apparent dependence. Without specific mention, all values are calculated with respect to the native protein structure given in the Protein Data Bank. For intrinsic disordered proteins (IDPs), the calculation of their folding energy and HP pattern is a bit tricky since, in principle, the IDPs have no stable tertiary structure under normal physiological conditions. To avoid unnecessary difficulties, here we use their solution NMR structures reported in the Protein Data Bank as a reference. Based on our calculations, the IDPs overall show a much lower secondary structure content than normal globular proteins. Their folding energy is also relatively high. Both are consistent with the characteristics of IDPs.

B. Dataset for machine learning

In our current dataset, there are 25 different amyloid (or amyloidlike) proteins collected from the literature (see Table I). Among them, 21 randomly selected proteins are classified into the training set (140 data points), while the rest 4 proteins are taken as the test (22 data points).

Training set: β-lactoglobulin, M-TTR, β_2-microglobin, actin, stefin B, γC-crystallin, tubulin, WW domain, thaumatin, thaumatin-like, patatin, α-lactalbumin, insulin, polyQ, CsgA, lysozyme, ure2p, CsgB, hIAPP, $A\beta_{40}$, and $A\beta_{4\text{-}40}$.

Test set: calcitonin, apo CII, and NM domain of Sup35 and $A\beta_{42}$.

C. Setup of feedforward fully connected neural network

The FCN is a kind of feed-forward neutral network and has been extensively used on a variety of tasks, including the computer vision, speech recognition, machine translation, and medical diagnosis.[23] Mathematically, FCN defines a nonlinear mapping as

$$\mathrm{f}(x) = \sum_{k=1}^{m} a_k \cdot \sigma(b_k \cdot x + c_k),$$

where m is the number of nodes in the hidden layer. a_k and b_k are the weight terms, and c_k is the bias term. σ is a nonlinear activation function. Our goal is to minimize the gap between the prediction f(x) and the real value y. By making use of the Back Propagation (BP) algorithm,[24] we can get an estimation on a_k, b_k, and c_k.

In the current study, the Python Deep Learning library,Keras,[25] is adopted to build a two-layers neural network framework (Fig. 1). We have 20 inputs, 1 outputs, and one hidden layer with 30 nodes. The batch size (i.e., the size of each batch of the training dataset) is set as 4. We use the adaptive moment estimation optimizer,[26] the Rectified Linear Unit (shortened as ReLU)[27] as the activation function, and the mean absolute error as the loss function. In FCN, there are plenty of unknown parameters. To adjust them, we use the GridSearchCV method in the scikit-learn.[28] Through extensive

TABLE I. A list of amyloid proteins used in FCN.

Protein	PDB ID	Experimental method	References
β-lactoglobulin	3NPO	X-ray diffraction	29
M-TTR	2NBP	Solution NMR	30
Tubulin	1JFF	Electron crystallography	31
Insulin	3I40	X-ray diffraction	32
Calcitonin	1BYV	Solution NMR	33
α-synuclein	1XQ8	Solution NMR	34
Thaumatin	1RQW	X-ray diffraction	35
Patatin	4PK9	X-ray diffraction	36
α-lactalbumin	1A4V	X-ray diffraction	37
ure2p	1HQO	X-ray diffraction	38
apo CII	1I5J	Solution NMR	39
WW domain	1E0L	Solution NMR	40
γC-crystallin	2NBR	Solution NMR	41
Lysozyme	253L	X-ray diffraction	42
$A\beta_{40}$	1BA4	Solution NMR	43
$A\beta_{42}$	1IYT	Solution NMR	44
Actin	3HBT	X-ray diffraction	45
β_2-microglobin	1LDS	X-ray diffraction	46
hIAPP	2L86	Solution NMR	47
Stefin B	1STF	X-ray diffraction	48
Thaumatinlike	2I0W	X-ray diffraction	35
polyQ	3IOR	X-ray diffraction	49
CsgA	2N59	Solution NMR	50
CsgB	2N59	Solution NMR	51
$A\beta_{4-40}$	1BA4	Solution NMR	52
NM domain of Sup35	1R5B	X-ray diffraction	53
LL37	2K6O	Solution NMR	54
Cecropin A2	1F0H	Solution NMR	55

simulations, an optimal combination of parameters is determined. To avoid overfitting, several algorithms, such as early stopping and L1 norm, have been implemented during the training in order to minimize overfitting as much as possible. The code for FCN is available at https://github.com/yangwuyue/Amyloid-Aggregation-Rates-Prediction.

D. Regression evaluation metrics

To evaluate the accuracy of machine learning predictions, following regression evaluation metrics are taken, i.e., the Mean Square Error (MSE) and Pearson's and Spearman's correlation coefficients

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \widehat{y_k}\right)^2,$$

$$\text{Pearson's correlation coefficient } R\_P = \frac{\mathrm{Cov}\left(y_k, \widehat{y_k}\right)}{\sqrt{D\left(y_k\right)D\left(\widehat{y_k}\right)}},$$

$$\text{Spearman's correlation coefficient } R\_S = \frac{\mathrm{Cov}\left(rg_{y_k}, rg_{\widehat{y_k}}\right)}{\sqrt{D\left(rg_{y_k}\right)D\left(rg_{\widehat{y_k}}\right)}},$$

where y_k is the true value, $\widehat{y_k}$ is the predicted value, and n is the number of samples. rg_{y_k} and $rg_{\widehat{y_k}}$ are ranked y_k and $\widehat{y_k}$, respectively, which are either in ascending or descending order.

E. Other classical machine learning algorithms

1. Multivariable linear regression (MLR)

Multivariable linear regression (MLR) is the most widely used approach in statistics for modeling the relationship between a dependent variable Y and one or more explanatory variables denoted by X. In this case, we take $Y = (\ln(t_{1/2}))$ and $X = (N, M_w, pI, \ldots, pH)$ representing all intrinsic and extrinsic factors. The coefficient A in the predictive equation $Y = AX + \varepsilon$ is obtained by the least square estimation.

2. Support vector regression (SVR)

Support Vector Regression (SVR) is a nonlinear regression model for data classification and regression analysis.[56] Given the training data $\{(x_1,y_1),(x_2,y_2), \ldots,(x_n,y_n)\}$, we want to get a regression model $f(x) = w^T x + b$, where w and b are adjustable parameters, to minimize the loss function

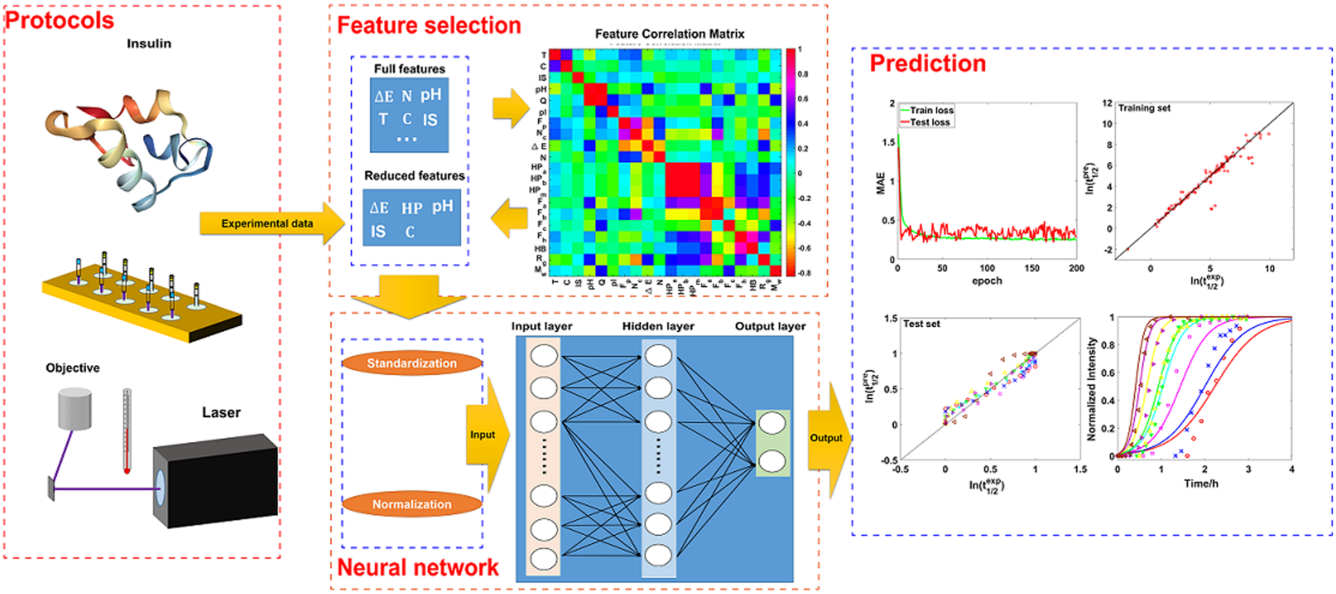

FIG. 1. Flowchart for machine learning and feature selection.

$$\min\left[\frac{1}{2}\|w\|^2 + R\sum_{i=1}^{n}\chi_\epsilon\big(f(x_i) - y_i\big)\right],$$

where R is a regularization constant and

$$\chi_\epsilon(m) = \begin{cases} 0, & |m| \le \epsilon, \\ |m| - \epsilon, & \text{otherwise,} \end{cases}$$

is the ϵ-insensitive loss function. In default, $\epsilon = 0.1$ and $R = 1$.

By introducing relaxation variables, Lagrange multipliers and using the Karush-Kuhn-Tucher condition, we can get the expression for SVR[56]

$$f(x) = \sum_{i=1}^{n}(\hat{a}_i - a_i)\kappa(x, x_i) + b,$$

where α_i and $\hat{\alpha}_i$ are the Lagrange multipliers. $\kappa(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. Here, we have used the linear kernel (LSVR) and polynomial kernel (PSVR), respectively.

3. Lasso

Given the training data $\{(x_1,y_1),(x_2,y_2), \ldots,(x_n,y_n)\}$, Lasso is derived from an extension of the linear model $f(x) = w^T x$. By taking the square loss function $\frac{1}{n}\|y - w^T x\|^2 + \lambda\|w\|^2$, it is equivalent to the following constrained optimization problem:[57]

$$\min \frac{1}{n}\|y - w^T x\|^2,$$

$$\text{s.t.} \|w\|_1 \le C,$$

whose solution gives the selected independent variables.

4. Random Forest (RF)

Random forest is an ensemble learning method,[58] which takes the decision tree as a weak learner and adds randomness to the selection of attributes. As an extension of the bagging algorithm, the random forest has a strong generalization ability to prevent overfitting. In terms of parameter setting, we set the number of trees in the forest as 100.

III. RESULTS AND DISCUSSION

A. Half time plays a key role in characterizing the aggregation kinetics

Through experimental measurements, such as fluorescence spectroscopy and light scattering, we obtain time trajectories reflecting the changes in fiber content. How to quantify these trajectories is a central problem in amyloid kinetics. Recent developments in chemical kinetics analysis[5,59,60] provide us a powerful and systematic way to characterize and understand the amyloid aggregation kinetics at the molecular level.

Among various kinetic parameters, the half-time $t_{1/2}$, defined as the time point when the fiber content reaches one half of its equilibrium value, and the apparent fiber growth rate k_{app}, defined as the normalized fiber growth rate at the half time, play a central role in quantifying the aggregation process [see Fig. 2(a)]. In practice, many amyloid aggregation curves can be fitted by an empirical formula involving $t_{1/2}$ and k_{app} as[60,61]

$$\frac{G(t)}{G(0)} = \frac{1}{1 + \exp\left[-k_{app}\left(t - t_{1/2}\right)\right]}, \qquad (1)$$

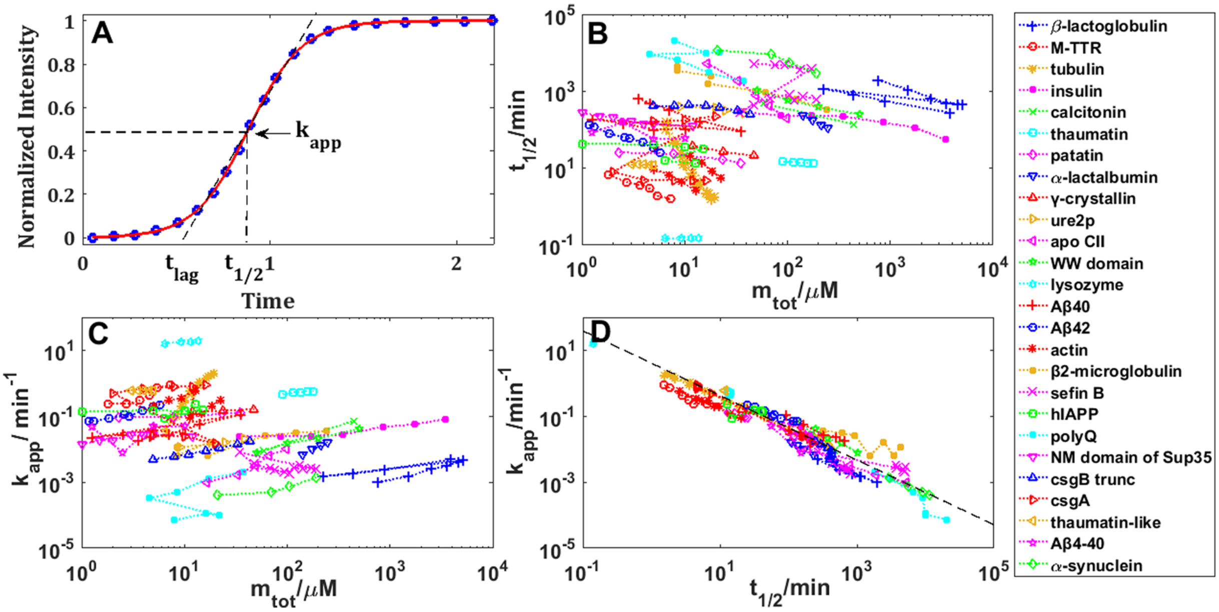

FIG. 2. (a) Determination of the half-time and the apparent fiber growth rate. [(b)–(d)] Scaling laws among the protein concentration, half time, and apparent fiber growth rate for 26 different kinds of amyloid proteins.

where $G(t)$ stands for experimental observations, such as the ThT fluorescence intensity, at time t; and $G(0)$ is the initial value. As a consequence, if a preknowledge of $t_{1/2}$ and k_{app} is obtained, the kinetic curves can be reproduced in a high accuracy.

A well-known result derived from chemical kinetics analysis is that both the half time and the apparent fiber growth rate follow certain scaling relations with respect to the protein concentration [Figs. 2(b) and 2(c)]. The scaling exponent is a critical index to distinguish the detailed amyloid aggregation mechanism at the molecular level.[59–61] In addition, as many previous studies have revealed,[60,61] the half time and the apparent fiber growth rate are inversely related to each other, despite the protein concentration and many other intrinsic and extrinsic features [Fig. 2(d)].

B. FCN provides a reliable way for predicting the half time

Now an interesting question arises naturally: can we make reliable estimations on the half-time $t_{1/2}$ and the apparent fiber growth rate k_{app} based on a preknowledge of intrinsic and extrinsic features? If so, we may achieve the goal of accurately predicting the amyloid aggregation kinetics without doing any experiment or referring to any physical model.

In this study, we adopt the fully connected neural network[23] to predict the half time. A combination of 16 intrinsic features and 4 extrinsic features from 25 different amyloid (or amyloidlike) proteins (162 data points in total; see Table I) is taken as the input (see Sec. II), while the half time $\boldsymbol{t_{1/2}}$ is the output (the apparent fiber growth rate $\boldsymbol{k_{app}}$ can be extracted from the inverse relation).

Most diseases related proteins (no mutants), such as α-synuclein,[34] islet amyloid polypeptide,[62] Aβ,[44] and insulin,[32] are all on the list. Among them, 21 randomly selected proteins are classified into the training set, while the rest 4 proteins are taken as the test. This is currently the largest kinetic dataset for amyloid proteins reported in the literature. In our dataset, the protein concentration is varied for about four orders of magnitude from 1 micromolar to 10 millimolar. The pH condition is from the acid condition with pH = 2 to the alkaline condition with pH = 8. Meanwhile, the half time for protein aggregation is changed for more than six orders of magnitude, from several seconds to hundred hours. According to the sequence alignment (https://www.ebi.ac.uk/Tools/psa/), Aβ_{40}, Aβ_{42}, and A$\beta_{4\text{-}40}$, thaumatin and thaumatinlike, hIAPP, and calcitonin have a sequence identity over 30% and can be considered as homology proteins (see supplementary material). Therefore, our dataset overall shows a relatively low sequence identity (4/25).

Compared to classical machine learning algorithms, such as Multivariable Linear Regression (MLR) and Support Vector Regression (SVR), FCN shows an outstanding performance over the others. For the training set, the mean square errors (MSEs) of $\boldsymbol{t_{1/2}}$ are decreased from 1.33 for MLR and 1.38 for PSVR to 0.42 for FCN (see Table I). Meanwhile, for the test set, the MSE of FCN is about 0.49, significantly lower than 440 for MLR and 3.3 for PSVR. A similar conclusion is reached based on either Pearson's correlation coefficient or a nonparametric measure, such as Spearman's correlation coefficient. Most importantly, a point-to-point comparison in Fig. 3 directly illustrates the excellent correlation between the experimental data and predictions by FCN.

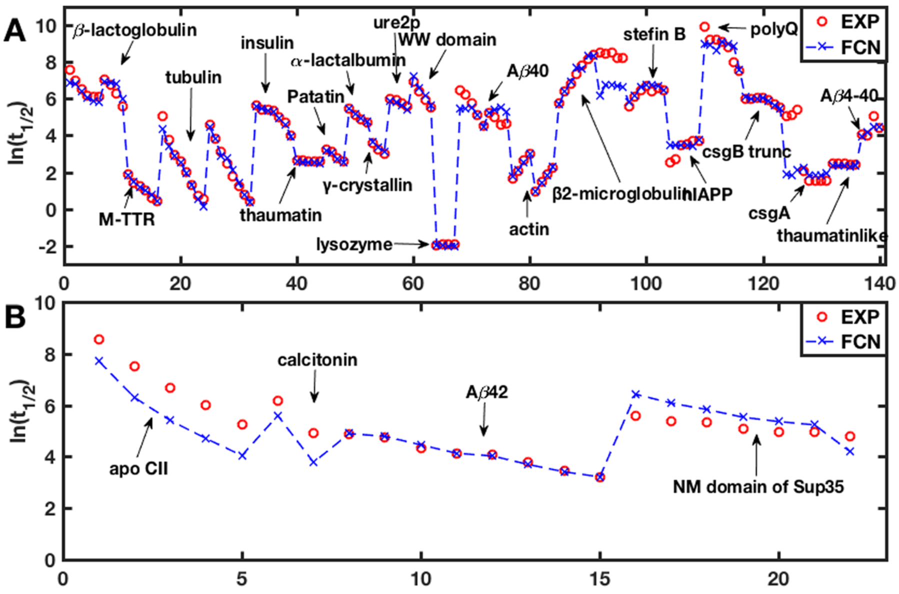

FIG. 3. A point-to-point comparison between the experimental data (red circles) and FCN predictions (blue crosses linked by dashed lines) on (a) the training set (140 points) and (b) the test set (22 points), respectively.

C. Critical tests further validate the predictive ability of FCN

By including data with features relevant to those in the test set into the training set, the accuracy of FCN can be greatly improved. This is a common feature of neural networks. In Fig. 4, we take α-synuclein as a test. Limited information—four data points reflecting different α-synuclein concentrations are added into the training set of FCN, while the pH and temperature dependence of α-synuclein (ten data points) are taken for validation. Comparing the results before and after dataset expansion [see Figs. 4(b) and 4(c)], it is clearly seen that a much better agreement is

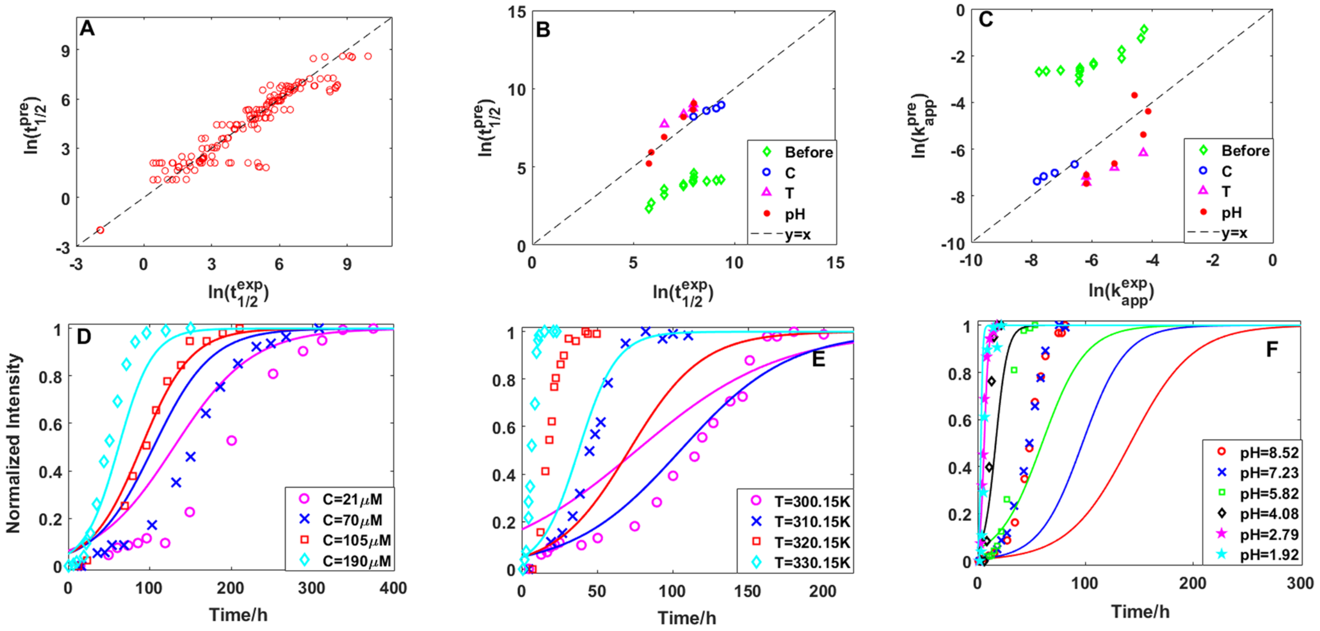

FIG. 4. Dataset expansion helps to improve the FCN accuracy on predictions of α-synuclein aggregation kinetics. [(a)–(c)] Correlations between the experimental data and predictions for α-synuclein aggregation on the training set and test set separately. Note that "before" representatives FCN predictions before dataset expansion. [(d)–(f)] Comparisons between the measured (symbols) and predicted (solid lines) kinetic curves for α-synuclein aggregation under varied concentration, temperature, and pH conditions.

achieved for the FCN predictions after dataset expansion. A similar conclusion also holds for data mixing. Through numerical explorations, we find that FCN shows a better performance on the mixed dataset (140 of 162 data points are randomly classified into the training set, while the rest 22 points belong to the test set). Its predictive accuracy, evaluated by Pearson's correlation coefficient, on the test set is largely improved from 0.84 to 0.96 (averaging over 5 independent samples), meanwhile the MSE is decreased from 0.49 to 0.46 in comparison with the nonmixed dataset. In conclusion, the expansion of training set will significantly reduce the prediction errors and gives us the ability to examine unknown processes without referring to either molecular mechanisms or additional experiments.

To further validate the predictive ability of FCN, a critical test, the aggregation ability of ten nonamyloid proteins, which are randomly picked from the Protein Data Bank, under extreme conditions (high temperature in the current case) is investigated. This study is of particular interest since normal nonamyloid proteins will become aggregated under extreme conditions, such as high temperature, high/low pH, and crowding environments. These conditions facilitate the unfolding of protein's globular structures and make them easy to interact with each other to form aggregates. A direct comparison of the aggregation rates of ten nonamyloid proteins at both high temperatures (100 °C), a condition believed suitable for protein aggregation, and room temperatures (20 °C) shows that an acceleration of 10^3–10^6 times is generally achieved by high heating based on our predictions (see Table S3). It agrees well with our general expectations.

D. Seven key features constitute a minimum set for prediction

A full list of 16 intrinsic and 4 extrinsic features are considered in FCN in Fig. 3; however, not all of them are crucial for a successful prediction. To see this point, the correlation analysis is performed on the normalized data by removing length dependence (see Fig. 1 and the supplementary material), which reveals a clear correlation among eight intrinsic features—the fraction of helix, sheet, and coil residues (F_A, F_B, F_C); HP patterns for helix, sheet, and helices cross membrane (HP_A, HP_B, HP_M); and hydrophobicity (F_H, HB). The same conclusion is applied to the proline content (F_P), nonlocal contact number (N_C), and folding energy (ΔE), or the pH value and total charge (Q) too. Meanwhile, the sequence length (N), molecular weight (M_w), and radius of gyration in folded states (R_g) show a moderate correlation with each other. As expected, no apparent correlation could be identified for the protein concentration (C), temperature (T), ionic strength (IS), and pI value, the former three of which all belong to extrinsic features. As a consequence, 20 intrinsic and extrinsic features are classified into at least seven different groups according to their correlation distance, i.e., {C}, {IS}, {T}, {pI}, {pH, Q}, {{F_P, N_C, ΔE}, N}, {{HP_A, HP_B, HP_M}, {F_A, F_B, F_C}, {F_H, HB}, R_G, M_W}.

Combined with the Principle Component Analysis (PCA), seven representatives, the HP pattern (HP_A, HP_B, HP_M), folding energy (ΔE), pH value, ionic strength (IS), and protein concentration (C) (one representative for each group except temperature T and pI), which make a contribution of over 85% to the whole feature set, constitute a minimum feature set for characterizing the aggregate rate on the current dataset (the contribution of temperature and pI is much smaller according to PCA; see the supplementary material for details). As we have observed, there is little loss in the prediction ability (Pearson's correlation coefficient drops from 0.96 to 0.93) when seven features are used in replace of 20 (see Table II).

From the component analysis and PCA, several interesting results can be obtained. Strong correlations among the HP patterns

TABLE II. Prediction accuracy of different algorithms.

	Training set			Test set		
Method	MSE[a]	Pearson's[b]	Spearman's[c]	MSE	Pearson's	Spearman's
MLR[d]	1.33	0.89	0.87	440	−0.54	−0.41
LSVR[e]	2.68	0.77	0.72	24.5	−0.62	−0.40
PSVR[f]	1.38	0.89	0.86	3.30	−0.36	−0.49
LASSO	3.46	0.71	0.69	2.23	0.44	0.63
RF[g]	0.14	0.99	0.99	2.77	0.32	0.28
FCN_20[h]	0.42	0.97	0.97	0.49	0.84	0.84
FCN_20 (mix)[i]	0.46	0.96	0.96	0.46	0.96	0.93
FCN_7 (mix)[j]	0.40	0.97	0.96	0.69	0.93	0.91

[a] MSE is the abbreviation of the mean square error.

[b] Pearson's is the abbreviation of Pearson's correlation coefficient.

[c] Spearman's is the abbreviation of Spearman's correlation coefficient. See Sec II for details on their definitions and formulas.

[d] MLR is the abbreviation of Multivariable Linear Regression.

[e] LSVR is the abbreviation of Support Vector Regression with linear kernel functions.

[f] PSVR is the abbreviation of Support Vector Regression with polynomial kernel functions.

[g] RF is the abbreviation of the Random Forest.

[h] FCN with 20 input features.

[i] 20 input features and mixed dataset.

[j] 7 reduced features as input and mixed dataset.

and fraction of helix and sheet residues are observed, except for the fraction of coil residues. Furthermore, a high fraction of helical residues means a large radius of gyration and a low nonlocal contact number, while an opposite conclusion holds for the sheet residues. Physically, the content of the secondary structure is determined by the hydrophobicity and the total charge. This fact is confirmed by our component analysis too.

Our findings are consistent with the observations by Chiti *et al.* that specific mutations on the aggregation rates of unfolded polypeptide chains are correlated to a remarkable extent with changes in hydrophobicity, secondary structure propensity, and charge.[15] In addition, the amyloid aggregation rate is shown to be "predictable" by intrinsic properties of the polypeptide sequence, such as the hydrophobicity, hydrophobic–hydrophilic patterning, and charge, and also environmental parameters, such as the pH value, ionic strength, and concentration through a linear regression formula by DuBay *et al.*[16]

IV. CONCLUSION

In summary, in this paper, we have applied the data based machine learning, the fully connected neural network to be exact, to the prediction of amyloid aggregation rates. FCN shows a much better performance on the real dataset than classical methods, such as MLR and SVR. More importantly, our study reveals the promising ability of neural networks in the study of amyloid kinetics, without referring to either experiments or physical models. This fact is also of great interest to many other chemical reaction systems.

Despite of its considerable success in the current application, FCN does not provide a simple explicit formula linking the intrinsic and extrinsic features to the aggregation rate directly, while the linear or nonlinear regression models usually do. The training of FCN also requires far more tagged data than the regression models, which constitutes a major bottleneck in applications. Furthermore, a common drawback of neural network based models is the lack of physics or a human understandable physical interpretation of what we have learned from the neural networks. Recently, there are many studies on the interpretable neural networks, such as the ResNets,[11,63] ODE nets,[64] and PDE nets,[65,66] which try to interpret a neural network flow as numerical calculations of ODEs or PDEs. These efforts for sure will make the machine learning based models more physical and easier understandable.

In the next step, we plan to take data from amino acid mutation experiments,[15,67] metal-ion,[20,68,69] crowding agents,[21] and lipid bilayers[22] mediated aggregation experiments into consideration. To predict how disease-related mutants and other extrinsic factors modulate the aggregation propensities and rates is of the most pressing need in this field. Our current study provides a new framework to examine these effects in quantity. A corresponding major challenge will be how many features have to be included in order to correctly reflect the sequential or environmental specificity. For example, many previous studies[21,70,71] revealed that the excluded volume effect plays a key role in promoting protein aggregation in a macromolecular crowding environment. Besides, several other factors, such as the solution viscosity, the nature of the crowding agent including nonspecific and specific solute–solvent, solvent–solvent, and solute–solute interactions, also make nonnegligible contributions to crowding.

Finally, if the data for protein folding rate are included too, parallel predictions on both protein folding and aggregation kinetics can be performed at the same time based on algorithms of machine learning. A better understanding on the tradeoff between intramolecular and intermolecular interactions inside the native protein structures,[72,73] the free energy landscape for protein folding, misfolding, and aggregation,[74] as well as the "life on the edge" hypothesis for the metastability of proteins *in vivo*[75,76] is expected to be learned from such studies.

SUPPLEMENTARY MATERIAL

See supplementary material for lists of protein features and their values, aggregation abilities of 10 nonamyloid proteins under high heating, results for the protein sequence alignment, protein feature correlation analysis, and principle component analysis.

ACKNOWLEDGMENTS

This work was supported by the 13th 5-Year Basic Research Program of CNPC (2018A-3306), the National Natural Science Foundation of China (Grant Nos. 21877070 and 81473369), the National key research and development program of China (Grant No. 2017YFC1702703), and the Key research and development plan of Shandong province (Grant No. 2016CYJS08A01-1).

The authors declare no competing financial interest.

REFERENCES

[1] A. L. Fink, Folding Des. **3**, R9–R23 (1998).

[2] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333–336 (2006).

[3] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **86**, 27–68 (2017).

[4] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson, Nat. Rev. Mol. Cell Biol. **15**(6), 384 (2014).

[5] L. Hong, C. F. Lee and Y. J. Huang, "Statistical mechanics and kinetics of amyloid fibrillation," in *Biophysics and Biochemistry of Protein Aggregation*, edited by J. M. Yuan and H. X. Zhou (World Scientific Press, 2017).

[6] S. I. A. Cohen, P. Arosio, J. Presto *et al.*, Nat. Struct. Mol. Biol. **22**(3), 207–213 (2015).

[7] J. Habchi, S. Chia, R. Limbocker *et al.*, Proc. Natl. Acad. Sci. U. S. A. **114**, E200–E208 (2017).

[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning (MIT Press, 2016).

[9] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**(7553), 436 (2015).

[10] O. Abdel-Hamid, A. Mohamed, H. Jiang *et al.*, IEEE/ACM Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014).

[11] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, Adv. Neur. Info. Proc. Syst. **27**, 3104–3112 (2014).

[13] D. Silver, A. Huang, C. J. Maddison *et al.*, Nature **529**(7587), 484–489 (2016).

[14] D. Silver, J. Schrittwieser, K. Simonyan *et al.*, Nature **550**(7676), 354 (2017).

[15] F. Chiti, M. Stefani, N. Taddei *et al.*, Nature **424**(6950), 805 (2003).

[16] K. F. DuBay, A. P. Pawar, F. Chiti *et al.*, J. Mol. Biol. **341**(5), 1317–1326 (2004).

[17] M. Belli, M. Ramazzotti, and F. Chiti, EMBO Rep. **12**(7), 657–663 (2011).

[18] C. M. Dobson, Nature **426**, 884–890 (2003).

[19] G. Meric, A. S. Robinson, and C. J. Roberts, Annu. Rev. Chem. Biomol. Eng. **8**, 139–159 (2017).

[20] F. Hane and Z. Leonenko, Biomol. **4**(1), 101–116 (2014).

[21] L. Breydo, K. D. Reddy, A. Piai *et al.*, Biochim. Biophys. Acta, Proteins Proteomics **1844**(2), 346–357 (2014).

[22] R. Friedman, R. Pellarin, and A. Caflisch, J. Mol. Biol. **387**(2), 407–415 (2009).

[23] D. Graupe, *Principles of Artificial Neural Networks*, 3rd ed. (World Scientific Publishers, 2013).

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature **323**(6088), 533–536 (1986).

[25] F. Chollet, Keras. GitHub repository: https://github.com/fchollet/keras/, 2015.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations* (ISCA, 2015).

[27] M. D. Zeiler, M. Ranzato, R. Monga *et al.*, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2013), pp. 3517–3521.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, J. Mach. Learn. Res. **12**(10), 2825–2830 (2011).

[29] S. P. Roefs and K. G. De Kruif, Eur. J. Biochem. **226**(3), 883–889 (1994).

[30] A. R. Hurshman, J. T. White. E. T. Powers *et al.*, Biochemistry **43**(23), 7365 (2004).

[31] W. A. Voter and H. P. Erickson, J. Biol. Chem. **259**(16), 10430–10438 (1984).

[32] L. Nielsen, R. Khurana, A. Coat *et al.*, Biochemistry **40**(20), 6036–6046 (2001).

[33] M. Kamihira, A. Naito, S. Tuzi *et al.*, Protein Sci. **9**(5), 867–877 (2000).

[34] V. N. Uversky, J. Li, and A. L. Fink, J. Biol. Chem. **276**(47), 44284–44296 (2001).

[35] S. D. Stranks, Phys. Rev. E **80**(5), 051907 (2009).

[36] A. M. Pots, H. Gruppen, H. H. J. de Johgh *et al.*, J. Agric. Food Chem. **47**(11), 4593–4599 (1999).

[37] A. Nag and R. S. Berry, J. Chem. Phys. **127**(18), 184503 (2007).

[38] L. Zhu, X. J. Zhang, L. Y. Wang *et al.*, J. Mol. Biol. **328**(1), 235–254 (2003).

[39] K. J. Binger, C. L. L. Pham, L. M. Wilson *et al.*, J. Mol. Biol. **376**(4), 1116–1129 (2008).

[40] N. Ferguson, J. Berriman, M. Petrovich *et al.*, Proc. Natl. Acad. Sci. U. S. A. **100**(17), 9814–9819 (2003).

[41] Y. Wang, S. Petty, A. Trojanowski *et al.* Invest. Ophthalmol. Visual Sci. **51**(2), 672–678 (2010).

[42] A. M. Buswell and A. P. Middelberg, Biotech. Bioeng. **83**(5), 567–577 (2003).

[43] G. Meisl, X. Yang, E. Hellestrand *et al.*, Proc. Natl. Acad. Sci. U. S. A. **111**(26), 9384–9389 (2014).

[44] E. Hellstrand, B. Boland, D. M. Walsh *et al.*, ACS Chem. Neurosci. **1**(1), 13–18 (2009).

[45] L. S. Tobacman and E. D. Korn, J. Biol. Chem. **258**(5), 3207–3214 (1983).

[46] W. F. Xue, S. W. Homans, and S. E. Radford, Proc. Natl. Acad. Sci. U. S. A. **105**(26), 8926–8931 (2008).

[47] C. Cabaleiro-Lago, I. Lynch, K. A. Dawson *et al.*, Langmuir **26**(5), 3453–3461 (2009).

[48] K. Škerget, A. Vilfan, M. Pompe-Novak *et al.*, Proteins: Struct., Funct., Bioinf. **74**(2), 425–436 (2009).

[49] C. C. Lee. R. H. Walters, and R. M. Murphy, Biochemistry **46**(44), 12810–12820 (2007).

[50] X. Wang, N. D. Hammer, and M. R. Chapman, J. Biol. Chem. **283**(31), 21530–21539 (2008).

[51] N. D. Hammer, J. C. Schmidt, and M. R. Chapman, Proc. Natl. Acad. Sci. U. S. A. **104**(30), 12494–12499 (2007).

[52] G. Li, W. Y. Yang, Y. F. Zhao *et al.*, Chem. Eur. J. **24**(51), 13647–13653 (2018).

[53] S. R. Collins, A. Douglass, R. D. Vale *et al.*, PLoS Biol. **2**(10), e321 (2004).

[54] S. Thennarasu, A. Tan, R. Penumatchu *et al.*, Biophys. J. **98**(2), 248–257 (2010).

[55] Z. Zheng, N. Tharmalingam, Q. Liu *et al.*, Antimicrob. Agents Chemother. **61**(7), e00686-17 (2017).

[56] A. J. Smola and B. Schölkopf, Stat. Comput.**14**(3), 199–222 (2004).

[57] R. Tibshirani and J. Roy, J. R. Stat. Soc. B. **73**(3), 273–282 (2011).

[58] L. Breiman, Mach. Learn. **45**, 5–32 (2001).

[59] T. P. J. Knowles, C. A. Waudby, G. L. Devlin *et al.*, Science **326**(5959), 1533–1537 (2009).

[60] S. I. A. Cohen, M. Vendruscolo, C. M. Dobson *et al.*, J. Mol. Biol. **421**(2-3), 160–171 (2012).

[61] L. Hong, X. H. Qi, and Y. Zhang, J. Phys. Chem. B **116**(23), 6611–6617 (2011).

[62] J. R. Peinado, F. Sami, N. Rajpurohit *et al.*, FEBS Lett. **587**(21), 3406–3411 (2013).

[63] K. He, X. Zhang, S. Ren *et al.*, *Computer Vision—ECCV* (IEEE, 2016), pp. 630–645.

[64] R. T. Q. Chen, Y. L. Rubanova, J. Bettencourt, and D. Duvenaud, Adv. Neur. Info. Proc. Sys. **31**, 6571–6583 (2018).

[65] Z. Long, Y. Lu, X. Ma, and B. Dong, "PDE-Net: Learning PDEs from Data," in *35th International Conference on Machine Learning* (ICML, 2018).

[66] Z. Long, Y. Lu, and B. Dong, e-print arXiv:1812.04426 (2018).

[67] A. H. Armstrong, J. Chen, A. F. Mckoy *et al.*, Biochemistry **50**(19), 4058–4067 (2011).

[68] V. Tõugu, A. Tiiman, and P. Palumaa, Metallomics **3**(3), 250–261 (2011).

[69] Y. Yoshiike, K. Tanemura, O. Murayama *et al.*, J. Biol. Chem. **276**(34), 32293 (2001).

[70] L. A. Munishkina, E. M. Cooper, V. N. Uversky *et al.*, J. Mol. Recognit. **17**(5), 456–464 (2004).

[71] A. Magno, A. Caflisch, R. Pellarin. J. Phys. Chem. Lett. **1**(20), 3027–3032 (2010).

[72] K. E. Routledge, G. G. Tartaglia, G. W. Platt *et al.*, J. Mol. Biol. **389**(4), 776–786 (2009).

[73] K. Jong, L. Grisanti, and A. A. Hassanali, J. Chem. Inf. Modell. **57**(7), 1548–1562 (2017).

[74] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**(5204), 1619–1620 (1995).

[75] M. Vendruscolo, T. P. Knowles, and C. M. Dobson, Cold Spring Harbor Perspect. Biol. **3**(12), 1750–1754 (2011).

[76] G. G. Tartaglia, S. Pechmann, C. M. Dobson *et al.*, Trends Biochem. Sci. **32**(5), 204–206 (2007).