Machine learning prediction of inter-fragment interaction energies between ligand and amino-acid residues on the fragment molecular orbital calculations for Janus kinase - inhibitor…

Tokutomi, Shusuke
Shimamura, Kohei
Fukuzawa, Kaori
Tanaka, Shigenori

(Citation)
Chemical Physics Letters,757:137883

(Issue Date)
2020-10-16

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90007515

Machine learning prediction of inter-fragment interaction energies between ligand and amino-acid residues on the fragment molecular orbital calculations for Janus kinase - inhibitor complex

Shusuke Tokutomi

Graduate School of System Informatics, Department of Computational Science, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Kohei Shimamura

Department of Physics, Kumamoto University, Kumamoto 860-8555, Japan

Kaori Fukuzawa

Department of Physical Chemistry, School of Pharmacy and Pharmaceutical Sciences, Hoshi University, 2-4-41 Ebara, Shinagawa, Tokyo 142-8501, Japan

Shigenori Tanaka

Graduate School of System Informatics, Department of Computational Science, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Abstract

Inter-Fragment Interaction Energies (IFIEs) obtained by Fragment Molecular Orbital (FMO) method can quantitatively measure the effective interactions between ligand and residues in protein, which are therefore useful for drug discovery. However, it has not been clarified whether the IFIEs can be reproduced using only geometrical (*e.g.*, interatomic distances) information of biomolecular complex without resort to explicit FMO calculations. In this

Email address: tanaka2@kobe-u.ac.jp; Fax:+81-78-803-6620 (Shigenori Tanaka)

study, through machine learning technique, we propose a highly accurate reproduction or prediction scheme for ligand-protein IFIEs using only the distance information as descriptors, thereby drastically saving the computational cost in FMO analysis for a variety of conformations.

Keywords: fragment molecular orbital method (FMO); inter-fragment interaction energy (IFIE); machine learning; ligand-protein complex; Janus kinase (JAK)

---

1. Introduction

Fragment molecular orbital (FMO) method [1–6] is a computational method that divides large molecules such as proteins into relatively small units called "fragments", and then calculates the energy of the whole molecule and the electron density quantum chemically by molecular orbital (MO) calculations for fragment monomers and dimers in environmental potentials. By using this method, we can apply an *ab initio* MO method that has been shown to succeed for small compounds to macromolecules such as proteins without significant loss in accuracy. With the use of the FMO method, we can obtain the inter-fragment interaction energies (IFIEs) [6, 7] with greatly reduced computation time. These IFIEs can be used for drug discovery because the interaction energy between ligand and each amino acid can be quantitatively evaluated. So far, FMO calculations have been performed on various ligand-protein complexes, and have been shown to be useful for virtual screening and other pharmaceutical applications [5, 6, 8–14], whereas their computational costs are still high.

In this paper, we consider kinases as protein systems. Kinases are a group of enzymes that transfer the γ-phosphate group of adenosine triphosphate (ATP) to the hydroxyl group of another protein working as a substrate. Until now, a number of drug compounds targeting kinases that cause various diseases have been identified. Janus Kinase (JAK) [15–17] is a cytokine involved in immunity and a phosphorylating enzyme responsible for intracellular signal transduction. By inhibiting this function, symptoms such as rheumatoid arthritis can be suppressed. In fact, a compound targeting JAK, Tofacitinib [15], is widely used as a therapeutic drug for rheumatoid arthritis

and so on, and can be regarded as the first molecular-target, low-molecular-weight drug in the field of immunosuppression.

In drug screening, extremely high-speed calculations are required to deal with huge numbers of compounds. In addition, incorporation of the dynamical structural changes of ligand-protein systems is often essential. In order to obtain the ligand-protein IFIEs with low cost and high accuracy, we may use the machine learning techniques to avoid actually performing the *ab initio* FMO calculations as much as possible. In earlier research [18], IFIEs of polypeptides were reproduced by a machine learning method using descriptors with inclusion of the electric charge information of molecules known in advance, and a promising result for the IFIE prediction has been obtained. However, it has not been investigated so far whether the IFIEs can be accurately predicted using only the geometrical (*e.g.*, interatomic distances) information of the biomolecular complex instead of performing the actual FMO calculations for respective structures. If machine learning techniques can be utilized for the high-speed prediction of ligand-residue IFIEs, it would be very helpful for drug discovery, incorporating the effects of structural change in the system.

In this study, we propose a ligand-protein IFIE prediction method using only the interatomic distance information as descriptors in the Janus Kinase-Tofacitinib complex system. In the following, after introducing the computational methods, we illustrate the accuracy of the proposed machine-learning model, its difficulties, superiorities over other models, and required computational cost. Finally, future developments based on the present method will be discussed.

2. Methods

2.1. Preparation of datasets

2.1.1. Molecular dynamics simulation

Classical molecular dynamics (MD) simulations were performed to obtain 150 structures as an input dataset for FMO calculations. The integrated computational science system, Molecular Operating Environment (MOE) [19], was used for the initial structure preparation, in which AMBER10:EHT force field [20] was employed. The MD simulations with the AMBER package [21] were then carried out, in which ff14SB force field [22] was employed. The force fields of the ligand and phosphorylated tyrosine specific to Janus kinase were separately prepared: The general AMBER force field (GAFF) with restrained electrostatic potential (RESP) atomic charges [23] was used for the former and TYR-PO3 force field [24] in AMBER parameter database for the latter. The detailed protocols are illustrated below:

1. Structure Preparation

   Crystal Strucuture of JAK1-Tofacitinib complex (PDB entry: 3EYG, Resolution: 1.9 Å, Sequence length: 290 [15]; see Fig. 1) was retrieved from Protein Data Bank (PDB). Missing atoms in the structure were complemented by the "structure preparation" function of MOE. Then, Molecular Mechanics (MM) calculation was performed to eliminate the anomalous approach between atoms by moving only hydrogen atoms using the AMBER10:EHT force field [20]. Simulation box in periodic boundary condition was created with the TIP3PBOX model, in which hydration was performed by adding solvated water molecules with the

thickness of at least 14 Å. In addition, three Na^+ counter ions were arranged to attain the charge neutralization.

2. Production Run

   Molecular dynamics (MD) simulation was performed starting with the initial structure prepared in 1. The simulation conditions were based on previous studies [17] that performed MD calculations on JAK. Structural relaxation was performed by gradually loosening the restraint of heavy atoms with four stages of 25 ps each. Finally, production run of 100 ns was performed at 1 bar and 300 K under NPT conditions.

3. Obtaining Snapshots

   After the MD calculation, Root Mean Square deviation (RMSd) of trajectory was calculated based on the initial structure. The result is shown in Fig. 2. From this result, 150 structures were obtained at every 0.1 ns between 85 and 100 ns, where the molecular structures are relatively stable. Finally, these structures were optimized by MM calculations under the condition that only hydrogen atoms were moved in vacuum after removing the ions and solvent.

2.1.2. Fragment molecular orbital calculation

Fragment Molecular Orbital (FMO) calculations [1–4, 6] are performed to obtain the interaction energies between the ligand and the amino-acid residues, which are the predicition target in this study. The objective variables are Inter-Fragment Interaction Energies (IFIEs), which are described below and predicted using the distance information in ligand-protein system.

In the FMO method, when a molecule is divided into N_f fragments, the electron energy of the entire system is approximately calculated in terms of

the fragment energies as follows, where the index I refers to the monomer and IJ the dimer:

$$E \simeq \sum_{I>J} E_{IJ} - (N_f - 2) \sum_{I} E_I. \tag{1}$$

This can be rewritten as follows:

$$E \simeq \sum_{I>J} (E'_{IJ} - E'_I - E'_J) + \sum_{I>J} \mathrm{Tr}(\Delta \mathbf{P^{IJ}} \mathbf{V^{IJ}}) + \sum_{I} E'_I \tag{2}$$

with $E'_\lambda = E_\lambda - V_\lambda$ and $V_\lambda = \mathrm{Tr}(\mathbf{P}^\lambda \mathbf{V}^\lambda)$. Here, V^λ is the electrostatic environment potential from fragments other than λ; P^λ and ΔP^λ are the density matrix and its difference between dimer and monomers, respectively. The sum of the first and second summand terms on the right side of Eq. (2), $\Delta E_{IJ} = (E'_{IJ} - E'_I - E'_J) + \mathrm{Tr}(\Delta \mathbf{P^{IJ}} \mathbf{V^{IJ}})$, represents an effective interaction between fragments I and J. This ΔE_{IJ} is defined as IFIE.

FMO calculations were performed using software ABINIT-MP [6] for 150 structures prepared as in Sec. 2.1.1 to obtain the IFIEs between ligand and amino-acid residues as objective variables. For the FMO calculation method, the MP2 approximation considering up to the second-order perturbation over the Hartree-Fock (HF) approximation was employed, and for the basis function, 6-31G* was used.

2.2. IFIE prediction

2.2.1. Selection of target residues

In this research, IFIEs were predicted from the distance information between ligand and neighboring residues around it. We have supposed that the IFIEs between the ligand and its distant residues would be difficult to

predict accurately only by the distance information. The present study is thus regarded as a first step toward this direction of research.

Therefore, first, residues to be predicted were selected. The selection criterion was that residues whose center of gravity distance from ligand was within 7 Å were adopted, leading to totally 50 residues (Fig. 1).

For each of the selected 50 residues, the procedures as mentioned in the following subsections were performed. The IFIE prediction model between the ligand and each residue was thus built and evaluated.

2.2.2. Standardization of IFIEs

Before performing the prediction, the IFIEs between ligand and target residues were standardized. Standardized IFIE as $IFIE_{std}$ was calculated as follows:

$$IFIE_{std} = \frac{IFIE - IFIE_{ave}}{IFIE_{var}}, \tag{3}$$

where $IFIE$ is the IFIE value between ligand and target residue for each structure (snapshot), $IFIE_{ave}$ is the average of the IFIEs for calculated 150 structures, and $IFIE_{var}$ is the standard deviation of IFIEs for the 150 structures.

2.2.3. Preparation of descriptors for machine learning

In the following, we illustrate the preparation of descriptors to be used for the prediction model. The preparation consists of three steps: selection of descriptors, their standardization, and dimension reduction for descriptors using sparse modeling.

Selecting Distance Information

Interatomic distance information used for predicting IFIE between ligand and amino-acid residue was selected as follows:

- Information on all interatomic distances between the ligand and the predicted (target) residue
- Information on all interatomic distances between the ligand and the residues on both neighboring sides of the predicted residue (*e.g.*, when predicting ASP880, information on all the interatomic distances between the ligand and ARG879, and between the ligand and LEU881)
- Information on five distances in descending order of the coefficients of variation (see below) between the ligand and other 47 residues (where "other" residues are those not corresponding to the above three residues, *i.e.*, focused and neighboring residues, out of the 50 selected residues to take account of the contributions from surroundings)

Here, the coefficient of variation was calculated as

$$CV = \frac{\sigma}{\mu}, \tag{4}$$

where σ is the standard deviation of the distance between a ligand atom and a residue atom for 150 structures, and μ is the average of the distance between the ligand atom and the residue atom for 150 structures. That is, CV refers to the degree of fluctuations of interatomic distance.

With the above selection, on average, 2297 distance information for each of 50 residues was selected as descriptors.

Standardization of descriptors

After the selection above, these descriptors were standardized as follows:

$$\hat{x} = \frac{x - \mu}{\sigma}, \tag{5}$$

where $\hat{x}$ is the standardized descriptor, x is the interatomic distance selected above, μ is the average of distances over 150 structures, and σ is the standard deviation of distances. From $\hat{x}$ for different interatomic pairs and structures, we then obtain the matrix of standardized descriptors $\mathbf{X_{std}}$.

Preprocessing by sparse modeling

As described above, there are about 2300 descriptors on average for respective ligand-residue (target) pair. Then, to make descriptors sparse, the linear regression with l_1 regularization was performed in advance using $IFIE_{std}$ as the target variable through

$$arg\ min_{\boldsymbol{\omega}}\{||\mathbf{IFIE_{std}} - \mathbf{X_{std}}\boldsymbol{\omega}||_2^2 + \alpha||\boldsymbol{\omega}||_1\}, \tag{6}$$

where $\mathbf{IFIE_{std}}$ is the reference values (vector as data points), $\mathbf{X_{std}}$ is the matrix of descriptors, $\boldsymbol{\omega}$ is the weight, $\mathbf{X_{std}}\boldsymbol{\omega}$ is the prediction values, α is set to 0.01, and $||\boldsymbol{\lambda}||_p$ refers to l_p norm of $\boldsymbol{\lambda}$. Here, when $\boldsymbol{\lambda} \in R^n$, l_p norm is calculated as

$$||\boldsymbol{\lambda}||_p = (\Sigma_{j=1}^{n}|\lambda_j|^p)^{1/p}. \tag{7}$$

Finally, the sparse descriptors $\mathbf{X_{spa}}$ are given by $\mathbf{X_{std}}$ and $\boldsymbol{\omega}$ above:

$$\mathbf{X_{spa}} = \mathbf{X_{std}}\boldsymbol{\omega}. \tag{8}$$

$\mathbf{X_{spa}}$ was used for input to the model in Sec. 2.2.4. Of the about 2300 descriptors selected above, 81 descriptors on average were retained for each of the predicted 50 residues. In the processing above, Scikit-Learn's library [25] was used for the actual calculations.

2.2.4. Construction of prediction model

In the following, Keras [26], a deep learning library of Python, was used for building and training of the neural network model.

Building the Model

Neural Network model to predict IFIE was built, where the input values were descriptors $\mathbf{X_{spa}}$ addressed in Sec. 2.2.3, and the output value was $IFIE_{std}$. The model was composed of two hidden layers and 20 nodes each. Furthermore, Rectified Linear Unit (ReLU) [27] was used as the activation function, and the weights were initialized according to a truncated normal distribution.

Training the Model

The data set obtained in Sec. 2.1 was divided into 105 (training) structures from 85 to 95.5 ns and 45 (test) structures from 95.5 to 100 ns so that the ratio of $training : test = 7 : 3$. The training was repeated 500 times using the optimization method *Adam* [28] so that the mean square error as the loss function was minimized. The batch size was 32.

Evaluating the Model

The weight determined above and new descriptors were input to the trained model, and $IFIE_{std}$ was predicted. The predicted result was inversely transformed into $IFIE$ by Eq. (3), and the coefficient of determination R^2 was calculated from the following equation:

$$R^2 = \frac{Cov(\xi, \eta)^2}{\sigma_\xi^2 \sigma_\eta^2}, \tag{9}$$

where ξ is the predicted IFIEs, η is the reference IFIEs, $Cov(\xi, \eta)$ is the covariance of the predicted and reference IFIEs, and σ_ξ and σ_η are the standard deviations of the predicted and reference IFIEs, respectively.

3. Results and discussion

3.1. Prediction results

3.1.1. Prediction results for 50 residues

Figure 3 shows the results of the prediction performed for each residue. The average of R^2_{test} of 50 residues is 0.85, which means that the IFIEs for most residues could be predicted accurately. As an example, we show the IFIE prediction result between the ligand and LYS911 with the best accuracy (Fig. 4) and the IFIE prediction result between the ligand and LYS908 with the second highest accuracy (Fig. S1 in Supplementary data). From these figures, it can be seen that the accuracy of IFIE prediction between ligand and residue obtained with particularly good precision is close to unity for R^2_{test} (Figs. 4(a) and S1(a)), and there is almost no difference between the predicted and correct IFIE values with wide ranges during the MD trajectory (Figs. 4(b) and S1(b)). On the other hand, the corresponding results for the worst R^2_{test} case are shown in Fig. 4 (c)-(d) for ALA906. We observe in this case that the relatively poor value of $R^2_{test} = 0.607$ is substantially due to the smallness of IFIE magnitude.

In this study, as described in Sec. 2.1.2, IFIE was calculated using the MP2 method in addition to the HF method. That is, IFIE can be devided into

$IFIE_{HF}$ and $IFIE_{CORR}$ as in the following expression Eq. (10). $IFIE_{HF}$ is a component that mainly describes the electrostatic interaction energy, and $IFIE_{CORR}$ is a component that expresses the dispersion interaction energy:

$$IFIE_{MP2} = IFIE_{HF} + IFIE_{CORR}. \tag{10}$$

It is noted that some residues have $IFIE_{CORR} = 0$ due to the Dimer ES approximation in the FMO calculation [6]. Then, in order to assess the difference in the prediction result for each component, the prediction was also performed using $IFIE_{HF}$ and $IFIE_{CORR}$ separately as the objective variables for 33 residues with available $IFIE_{CORR}$, where just the objective variable was changed with the other conditions being the same as in the proposed model above.

The worst cases for the IFIE prediction are shown in Table 1 for ten residues with R^2_{test} being less than 0.78. For all the 33 residues with available $IFIE_{CORR}$, $IFIE_{HF}$ has the values of R^2_{test} lower than those of $IFIE_{CORR}$ by about 0.1 on average. Table 1 shows that the R^2_{test} result of $IFIE_{HF}$ is significantly lower than that of $IFIE_{CORR}$ in each residue, while four residues do not have $IFIE_{HF}$ and $IFIE_{CORR}$ separately due to the Dimer ES approximation [6]. It is thus suggested that the accuracy of IFIE predictions is significantly affected by the $IFIE_{HF}$ component, which primarily represents the electrostatic contribution. There are a number of non-polar residues in Table 1 with R^2_{test} being less than 0.78. Here, we also observe in the case of non-polar residues that the dispersion interactions can be well described in terms of the information on interatomic distances, whereas the electrostatic interactions represented by $IFIE_{HF}$ may not be appropriately

described only in terms of distance information. It is then suggested that the inclusion of other descriptors than the interatomic distances, such as electric charges, may be effective for further improvement on prediction accuracy.

Table 1: R^2_{test} results for each component (HF versus $CORR$) by the proposed model (Model I). "AVERAGE" means the average of the results for 33 residues where $IFIE_{CORR}$ was available. The listed 10 residues were taken by the ten lowest R^2_{test} values ($R^2_{test} < 0.78$). $IFIE_{HF}$ and $IFIE_{CORR}$ were not separately obtained for the residues with * mark due to the Dimer ES approximation [6].

Residue	$IFIE$	$IFIE_{HF}$	$IFIE_{CORR}$
AVERAGE	0.859	0.816	0.913
ALA906	0.607	0.821	0.974
LEU964	0.636	0.727	0.941
GLY962	0.641	0.510	0.939
LEU891*	0.654	-	-
LEU1010	0.748	0.721	0.942
LEU929*	0.765	-	-
ASP1021	0.775	0.774	0.873
LEU959	0.777	0.573	0.975
LYS939*	0.778	-	-
LYS1018*	0.779	-	-

3.1.2. Prediction results for Total-IFIE

Figure 5 shows the results of Total-IFIE which is the sum of all the IFIEs between ligand and 50 residues obtained above. The reproduction

(training) and prediction (test) accuracies were $R^2_{train} = 0.976$ and $R^2_{test} = 0.685$, respectively. Considering the analysis below, we suppose that, while the prediction was only for the 50 residues around the ligand in this study, it could be obtained with similar accuracy even if it is extended to all 290 residues of JAK1.

As seen for the results of the test data set in Fig. 5(b), the difference between the reference value and the predicted value was particularly large for the structures at 99.1 and 100.0 ns, where the difference was about 10 kcal/mol. Considering the results for each residue in these structures, it is observed that the cause for the difference is due to particularly poor prediction results for ASP1021 (which is shown in Table 1 with $R^2_{test} = 0.775$), as seen in Fig. 5(c). There was a difference of about 5 kcal/mol between the correct and predicted values only for ASP1021 with large magnitude of IFIE mainly due to the electrostatic interaction. This fact seems to reflect the sensitivity of electrostatic interaction to structural change, since the charged ASP1021 is a part of the DFG loop that is the activation loop in JAK [15–17]. For this reason, in Fig. 5(b), it is considered that there was a particularly large difference in the Total-IFIE for the structures at 99.1 and 100.0 ns. However, the structural fluctuations of ASP1021 in this study are not so large compared to those of the other selected residues. The average of root mean square fluctuation (RMSF) of the 50 residues obtained by 100 ns MD simulation is 0.574 Å, while that of ASP1021 is 0.450 Å. Therefore, in order to clarify the relationship between structural fluctuations and prediction accuracy for IFIE, more detailed investigations on molecular interactions are needed, including those residues with larger structural fluctuations in the DFG loop

and other parts of JAK.

3.1.3. Prediction results for all IFIEs

All the results for 150 structures of 50 residues (totally 7500 points) are shown in Fig. 6. We here find that R^2_{test} is 0.995, indicating that the prediction is highly accurate in total. Further, in the figure, it can be seen that the deviation and variation are relatively large in the region where the IFIE values are smaller than about -20 kcal/mol. This is considered to be due to the fact that, as shown in Sec. 3.1.2, the prediction results for some particular structures of some residues become worse.

3.2. Comparison with other models

In order to show the superiority of the proposed IFIE prediction model, three other models were built and the R^2_{test} values were compared. The proposed model above is referred to as Model I, and other models II-IV are described below:

- Model II

  In the neural network part of the proposed model (Model I), the output is each IFIE for a single residue. In contrast, this model (Model II) predicts IFIEs of all 50 residues at once by a single training. Accordingly, the 50 descriptors with large coefficient value of variation (CV) were equally taken from the interatomic distance information between the ligand and the respective 50 residues (2500 in total). Here, the batch size was set to 105 and the number of epochs was set to 20,000. Other conditions are the same as in Model I.

- Model III

  This is a model that directly predicts Total-IFIE alone. The 50 descriptors with large CV were equally taken from the information on the interatomic distances between the ligand and the respective 50 residues (2500 in total). Again, the batch size was set to 105 and the number of epochs was set to 20,000. Other conditions are the same as in Model I.

- Model IV

  This model performs the LASSO regression [29] without using the neural network model. The major difference from other models is that the LASSO regression is a linear model, whereas the neural networks in Model I-III are nonlinear. The selected descriptors were the same as in Model I, but the input descriptors were not made sparse.

Table 2 shows a summary for the prediction results by the four models, and Figs. S2-S4 in Supplementary data illustrate the details. On the whole, the results of R^2_{test} for 50 residues by Model I were shown to be superior to those by other models.

On the other hand, in the prediction of Total-IFIE, it was found that the model that directly predicted Total-IFIE (Model III) had better accuracy in R^2_{test}, as seen in Fig. S3. Therefore, when we need to predict IFIE-sum (the sum of IFIEs between ligand and all residues) used as the binding energy between the ligand and the protein, it is considered to be better to use the Total-IFIE prediction model (Model III). However, it should be noted that this model cannot describe the IFIE per residue. It was also found that the

results plotted for all the IFIEs were similarly accurate in all the models owing to the enlarged ranges of IFIE values employed in the plots.

From the results above, it can be concluded that Model I is superior to other models because of the ability to predict IFIE for each residue with high accuracy.

Table 2: Results of R^2_{test} for each model. Model I is the model proposed and recommended in this study. Model II gives the outputs of IFIEs for 50 residues by a single neural network. Model III is a model that directly predicts Total-IFIE (therefore the results for each 50 residue are not obtained). Model IV is a linear regression model.

Models	Average of 50 residues	Total-IFIE	All IFIEs
Model I(Recommended Model)	**0.854**	**0.685**	**0.995**
Model II	0.702	0.391	0.987
Model III	-	0.810	-
Model IV	0.679	0.625	0.986

3.3. Computational cost

Here, the execution time by the proposed model (Model I) is illustrated. The average time required to calculate the IFIEs for 50 residues by the proposed model was 410.3 seconds on a personal computer with 2 CPU cores of Intel Core i5 and 8GB memory. This includes the model training time, and if excluding it, the IFIE can be obtained in much shorter time.

On the other hand, the average FMO execution time per structure was 36528.2 seconds (the number of used nodes was 420 for the 20 structures and

36 for the 130 structures, respectively) on the K computer in Kobe, where 1 node has 8 CPU cores of SPARC64 VIIIfx with 16 GB memory.

From these facts, we see that the computational cost required to obtain IFIEs is drastically reduced if the proposed model is used.

4. Conclusions

We have proposed an IFIE prediction model and others for the JAK1-Tofacitinib complex on the basis of machine learning techniques applied to the FMO calculation data. The proposed model (Model I) has those important characteristics such as the generation of each trained model for every 50 residue surrounding the ligand, the standardization of IFIE as the objective variable, and the method of selecting descriptors with sparse modeling. In this way, it was shown that the proposed model can accurately reproduce the IFIEs between ligand and residues for each 50 pair and the Total-IFIE through machine learning based on neural network model. Furthermore, the prediction of IFIE divided into the HF and correlation energies showed the relatively low accuracy for $IFIE_{HF}$, thus indicating the deficiencies associated with the description of electrostatic interactions. This is a tendency found for residues with low IFIE prediction accuracy. Therefore, it is concluded that the prediction accuracy of IFIE is significantly affected by $IFIE_{HF}$. Then, it was found that the worse prediction impaired at specific residues resulted in the low accuracy in Total-IFIE for some structures. For example, ASP1021 shown here is a part of the DFG loop, which is the activation loop in JAK, and there is room for improvement on the effect of structural change on prediction accuracy. As a whole, we showed the supe-

riority of the proposed model (Model I) over other models (Models II-IV). It may be concluded that the proposed model can predict the IFIE for every residue in ligand-protein complex, and the high prediction accuracy thereof is a great advantage. Finally, the computational cost of the proposed model was assessed. It was found that IFIEs can be obtained much faster using the proposed model than actually performing FMO calculation without significant loss of accuracy.

From the analyses in the present study, we conclude that it is possible through machine learning for limited numbers of FMO calculations to predict IFIEs for JAK residues near Tofacitinib with high accuracy using only the geometrical (distance) information. Therefore, the IFIE predictions for all residues in the JAK1-Tofacitinib complex based on this model and the extension to prediction models for other ligand-kinase complexes are feasible. The transferability of the present method has been demonstrated through involvement of a wide variety of amino-acid residues interacting with ligand molecule. As future developments, the predictions for each energy component [30] to form the IFIEs would also be possible in addition to the application to the solvated biomolecular systems [31]. We may expect that the FMO-IFIE information for relatively small numbers of structures can be used for the prediction of IFIEs for large variety of structures. Besides, the findings obtained in this study will give some insights into the construction of forthcoming FMO-based force fields. Taken together, there is much room for further investigations concerning the combination of *ab initio* FMO method and machine learning techniques [32–35] toward efficient drug discovery.

Acknowledgements

We would like to acknowledge the Grants-in-Aid for Scientific Research (Nos. 17H06353 and 18K03825) from the Ministry of Education, Cultute, Sports, Science and Technology (MEXT), Japan. This research was performed in the activities of the FMO drug design consortium (FMODD). The results of FMO calculations were obtained using the K computer (project ID: hp180147 and hp190119).

References

[1] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, M. Uebayasi, Chem. Phys. Lett. 313 (1999) 701.

[2] T. Nakano, T. Kaminuma, T. Sato, Y. Akiyama, M. Uebayashi, K. Kitaura, Chem. Phys. Lett. 318 (2000) 614.

[3] D.G. Fedorov, K. Kitaura, in Modern Methods for Theoretical Physical Chemistry of Biopolymers, ed. E.B. Starikov, J.P. Lewis, S. Tanaka, Elsevier, Amsterdam, 2006, p. 3-38.

[4] D.G. Fedorov, K. Kitaura, The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems, CRC Press, Boca Raton, FL, 2009.

[5] D.G. Fedorov, T. Nagata, K. Kitaura, Phys. Chem. Chem. Phys. 14 (2012) 7562.

[6] S. Tanaka, Y. Mochizuki, Y. Komeji, Y. Okiyama, K. Fukuzawa, Phys. Chem. Chem. Phys. 16 (2014) 10310.

[7] I. Kurisaki, K. Fukuzawa, Y. Komeiji, Y. Mochizuki, T. Nakano, J. Imada, A. Chmielewski, S.M. Rothstein, H. Watanabe, S. Tanaka, Biophys. Chem. 130 (2007) 1.

[8] K. Fukuzawa, Y. Mochizuki, S. Tanaka, K. Kitaura, T. Nakano, J. Phys. Chem. B 110 (2006) 16102.

[9] O. Ichihara, J. Barker, R.J. Law, M. Whittaker, Mol. Inf. 30 (2011) 298.

[10] R. Hatada, K. Okuwaki, Y. Mochizuki, Y. Handa, K. Fukuzawa, Y. Komeiji, Y. Okiyama, S. Tanaka, J. Chem. Inf. Model. 60 (2020) 3593.

[11] *Quantum Mechanics in Drug Discovery*, edited by A. Heifetz, Methods in Molecular Biology vol. 2114 (Humana Press, New York, NY, 2020).

[12] A. Heifetz, I. Morao, M.M. Babu, T. James, M.W.Y. Southey, D.G. Fedorov, M. Aldeghi, M.J. Bodkin, A. Townsend-Nicholson, J. Chem. Theory Comput. 16 (2020) 2814.

[13] H. Lim, J. Chun, X. Jin, J. Kim, J.H. Yoon, K.T. No, Sci. Rep. 9 (2019) 16727.

[14] S. Tanaka, C. Watanabe, T. Honma, K. Fukuzawa, K. Ohishi, T. Maruyama, J. Mol. Graph. Model. 100 (2020) 107650.

[15] N.K. Williams, R.S. Bamert, O. Patel, C. Wang, P.M. Walden, A.F. Wilks, E. Fantino, J. Rossjohn, I.S. Lucet, J. Mol. Biol. 387 (2009) 219.

[16] Y. Higashi, Folia Pharmacol. Jpn. 144 (2014) 160.

[17] S. Wan, P.V. Coveney, J. Chem. Inf. Model. 52 (2012) 2992.

[18] Y. Mochizuki, K. Sakakura, Y. Akinaga, K. Kato, H. Watanabe, Y. Okiyama, T. Nakano, Y. Komeiji, A. Okusawa, K. Fukuzawa, S. Tanaka, J. Comput. Chem. Jpn. 16 (2017) 119.

[19] Molecular Operating Environment (MOE). Chemical Computing Group (CCG) Inc., Montreal, QC, Canada, 2016.

[20] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, J. Comput. Chem. 25 (2004) 1154.

[21] D.A. Case et al., AMBER 16. University of California, San Francisco, 2016.

[22] J.A. Maier. C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, J. Chem. Theory Comput. 11 (2015) 3696.

[23] C.I. Bayly, P. Cieplak, W. Cornell, P.A. Kollman, J. Phys. Chem. 97 (1993) 10269.

[24] N. Homeyer, A.H.C. Horn, H. Lanig, H. Sticht, J. Mol. Model. 12 (2006) 281.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Machine Learn. Res. 12 (2011) 2825.

[26] Keras Documentation, https://keras.io/ja/

[27] Y. Lecun, Y. Bengio, G. Hinton, Nature 521 (2015) 436.

[28] D.P. Kingma, J. Ba, arXiv:1412.6980 (2014).

[29] R. Tibshirani, J. R. Statist. Soc. B 73 (2011) 273.

[30] D.G. Fedorov, K. Kitaura, J. Comput. Chem. 28 (2007) 222.

[31] Y. Okiyama, K. Fukuzawa, Y. Komeiji, S. Tanaka, Methods Mol. Biol. 2114 (2020) 105.

[32] T. Yoshida, Y. Munei, S. Hitaoka, H. Chuman, J. Chem. Inf. Model. 50 (2010) 850.

[33] R. Kurauchi, C. Watanabe, K. Fukuzawa, S. Tanaka, Comput. Theor. Chem. 1061 (2015) 12.

[34] K. Maruyama, Y. Sheng, H. Watanabe, K. Fukuzawa, S. Tanaka, Comput. Theor. Chem. 1132 (2018) 23.

[35] K. Kato, T. Masuda, C. Watanabe, N. Miyagawa, H. Mizouchi, S. Nagase, K. Kamisaka, K. Oshima, S. Ono, H. Ueda, A. Tokuhisa, R. Kanada, M. Ohta, M. Ikeguchi, Y. Okuno, K. Fukuzawa, T. Honma, J. Chem. Inf. Model. 60 (2020) 3361.

Figure captions

Figure 1: (a) Crystal structure of JAK1-Tofacitinib complex (PDB entry: 3EYG). The 50 residues whose IFIEs with the ligand are predicted in this study are depicted in pink surrounding the ligand, where the distances between the centers of gravity of the ligand and the residues are within 7 Å. (b) Molecular structure of ligand compound, Tofacitinib.

Figure 2: RMSd results for 100 ns MD simulation of JAK1-Tofacitinib complex. 150 structures were obtained at every 0.1 ns between 85 and 100 ns surrounded by a red frame.

Figure 3: IFIE prediction results for 50 residues by the proposed model (Model I). The ordinate refers to the value of R^2, and the abscissa the selected 50 residues. The blue bar represents the result of R^2 for the test data set, and the orange bar the result of R^2 for the training data set. R^2 was calculated from Eq. (9) in Sec. 2.2.4.

Figure 4: (a) IFIE prediction result between ligand and LYS911 by the proposed model (Model I). The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (b) The results for time-series 150 structures are plotted with IFIE on the ordinate and time on the abscissa in the case of LYS911. The orange/blue lines indicate the temporal change of the predicted/correct values according to the change of structure. The left side of the red verti-

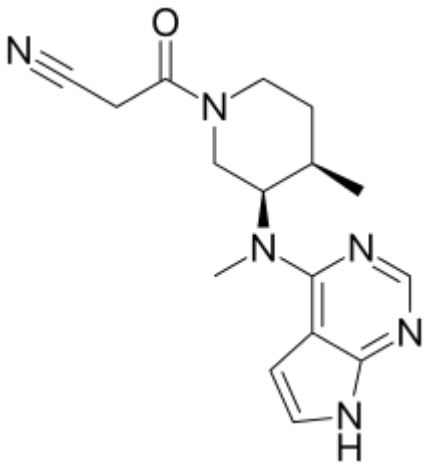
cal line refers to the results for the training data set, and the right side the results for the test data set. (c) IFIE prediction result between ligand and ALA906 by the proposed model (Model I). The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (d) The results for time-series 150 structures are plotted with IFIE on the ordinate and time on the abscissa in the case of ALA906. The orange/blue lines indicate the temporal change of the predicted/correct values according to the change of structure. The left side of the red vertical line refers to the results for the training data set, and the right side the results for the test data set.

Figure 5: (a) Total-IFIE prediction result by the proposed model (Model I). The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (b) The results for time-series 150 structures are plotted with the Total-IFIE on the ordinate and time on the abscissa. The orange/blue lines indicate the temporal change of the predicted/correct values according to the change of structure. The left side of the red vertical line refers to the results for the training data set, and the right side the results for the test data set. Numerical values of Total-IFIE are shown at 99.1 and 100.0 ns, indicating that the deviation between the predicted value (pred) and the correct value (ref) is particularly large there. (c) IFIE prediction result between ligand and

ASP1021 by the proposed model (Model I). The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). The points surrounded by red circles correspond to the structures at 99.1 and 100.0 ns, and it is observed that they are greatly deviated from the line of $y = x$.

Figure 6: All IFIE prediction result by the proposed model (Model I). The results for all 150 structures of 50 residues are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$), which is virtually hidden behind the points.

(a)



(b)

Figure 1

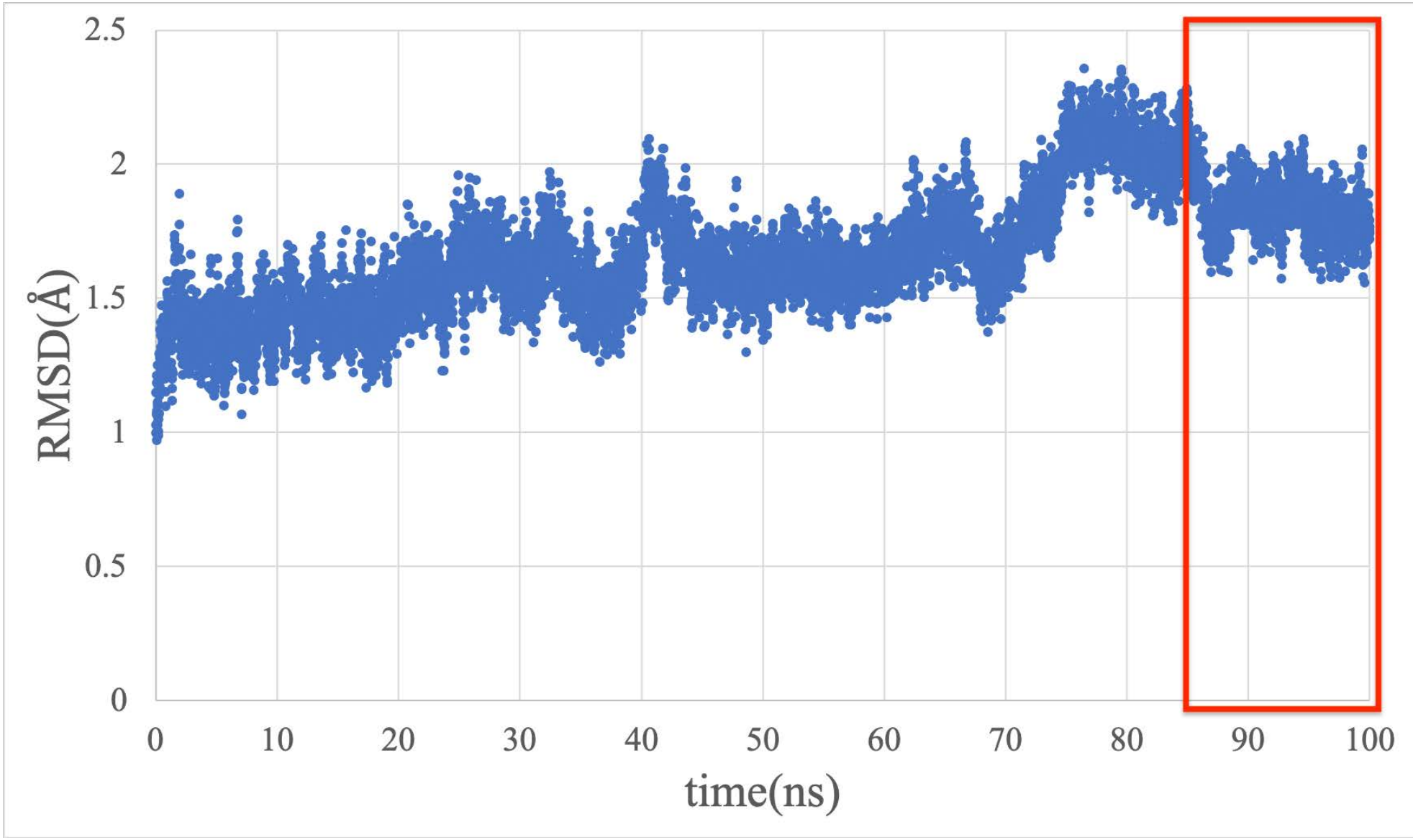


Figure 2

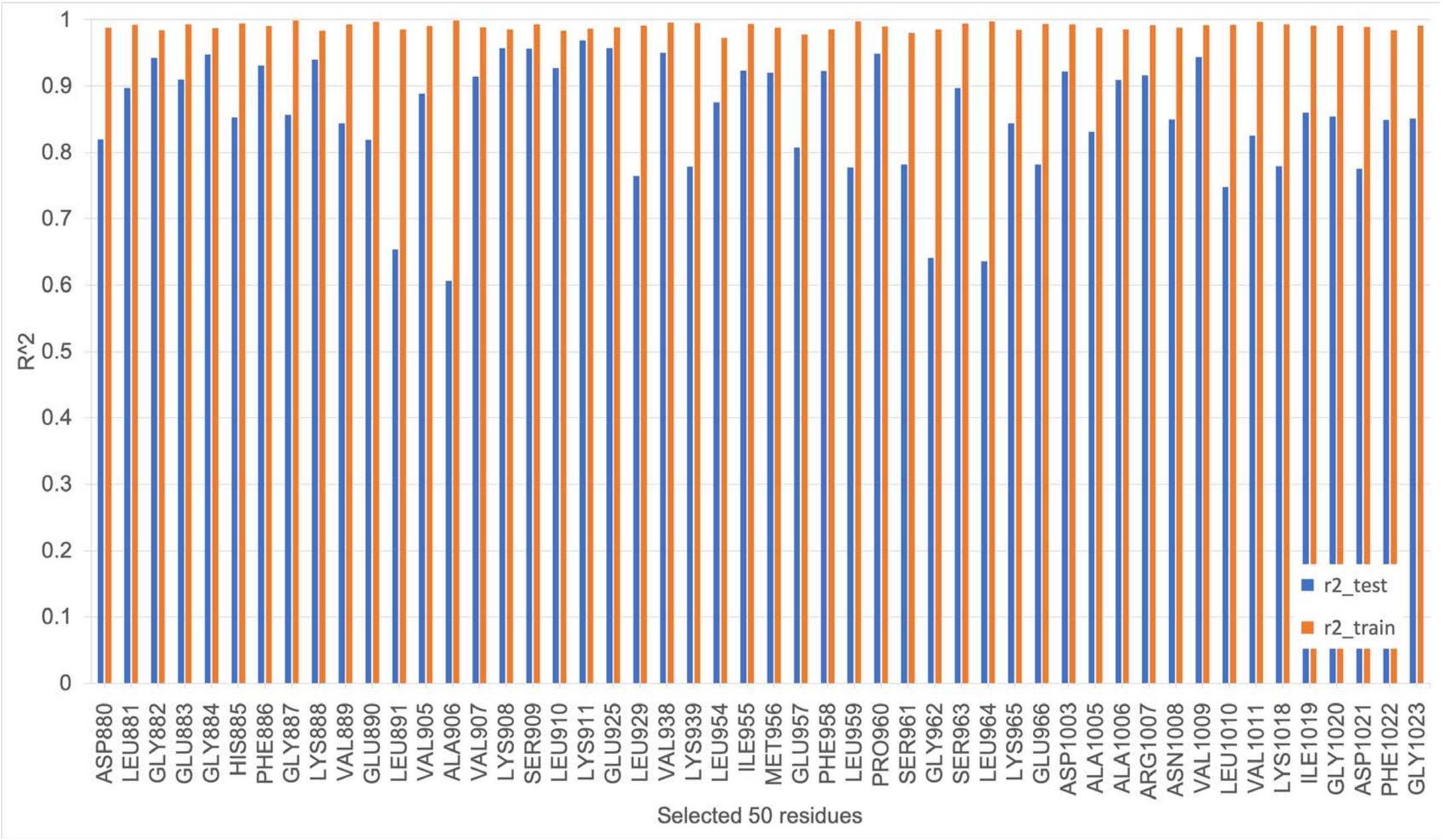

R^2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
r2_test
r2_train
ASP880
LEU881
GLY882
GLU883
GLY884
HIS885
PHE886
GLY887
LYS888
VAL889
GLU890
LEU891
VAL905
ALA906
VAL907
LYS908
SER909
LEU910
LYS911
GLU925
LEU929
VAL938
LYS939
LEU954
ILE955
MET956
GLU957
PHE958
LEU959
PRO960
SER961
GLY962
SER963
LEU964
LYS965
GLU966
ASP1003
ALA1005
ALA1006
ARG1007
ASN1008
VAL1009
LEU1010
VAL1011
LYS1018
ILE1019
GLY1020
ASP1021
PHE1022
GLY1023
Selected 50 residues

Figure 3

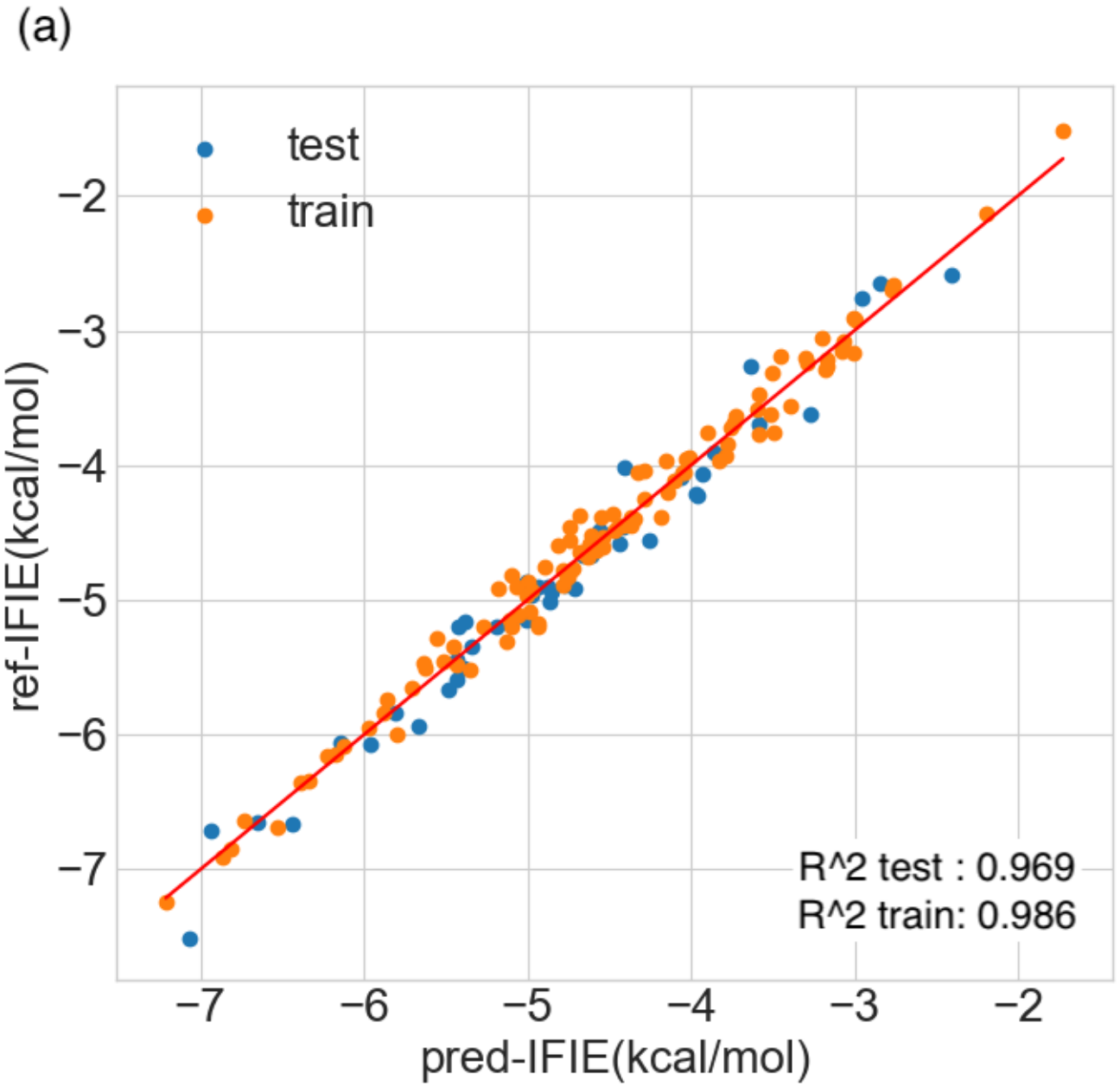

(a)
test
train
ref-IFIE(kcal/mol)
pred-IFIE(kcal/mol)
−2
−3
−4
−5
−6
−7
R^2 test : 0.969
R^2 train: 0.986

Figure 4 (a)

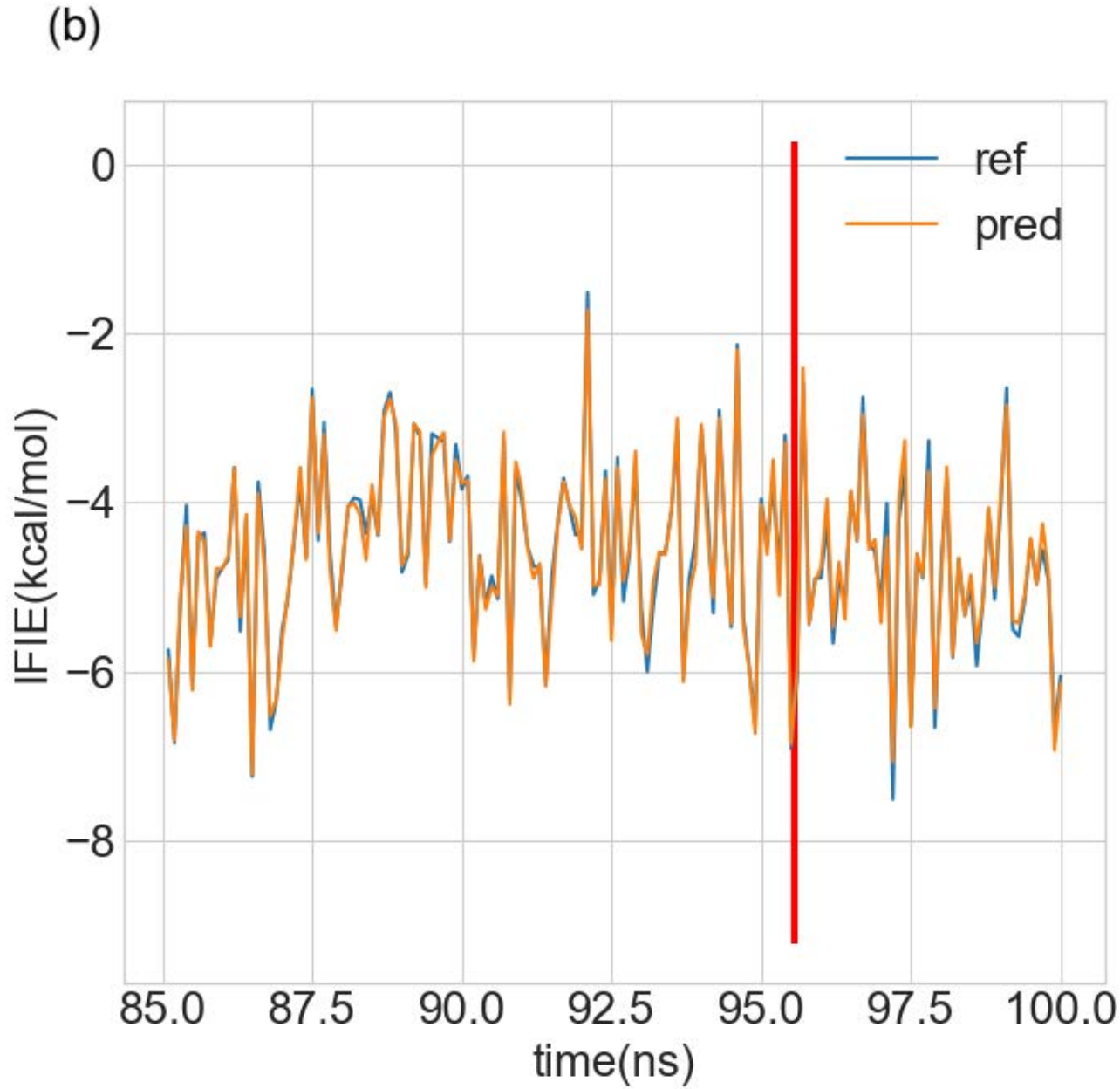

(b)
ref
pred
IFIE(kcal/mol)
0
−2
−4
−6
−8
85.0
87.5
90.0
92.5
95.0
97.5
100.0
time(ns)

Figure 4 (b)

(c)

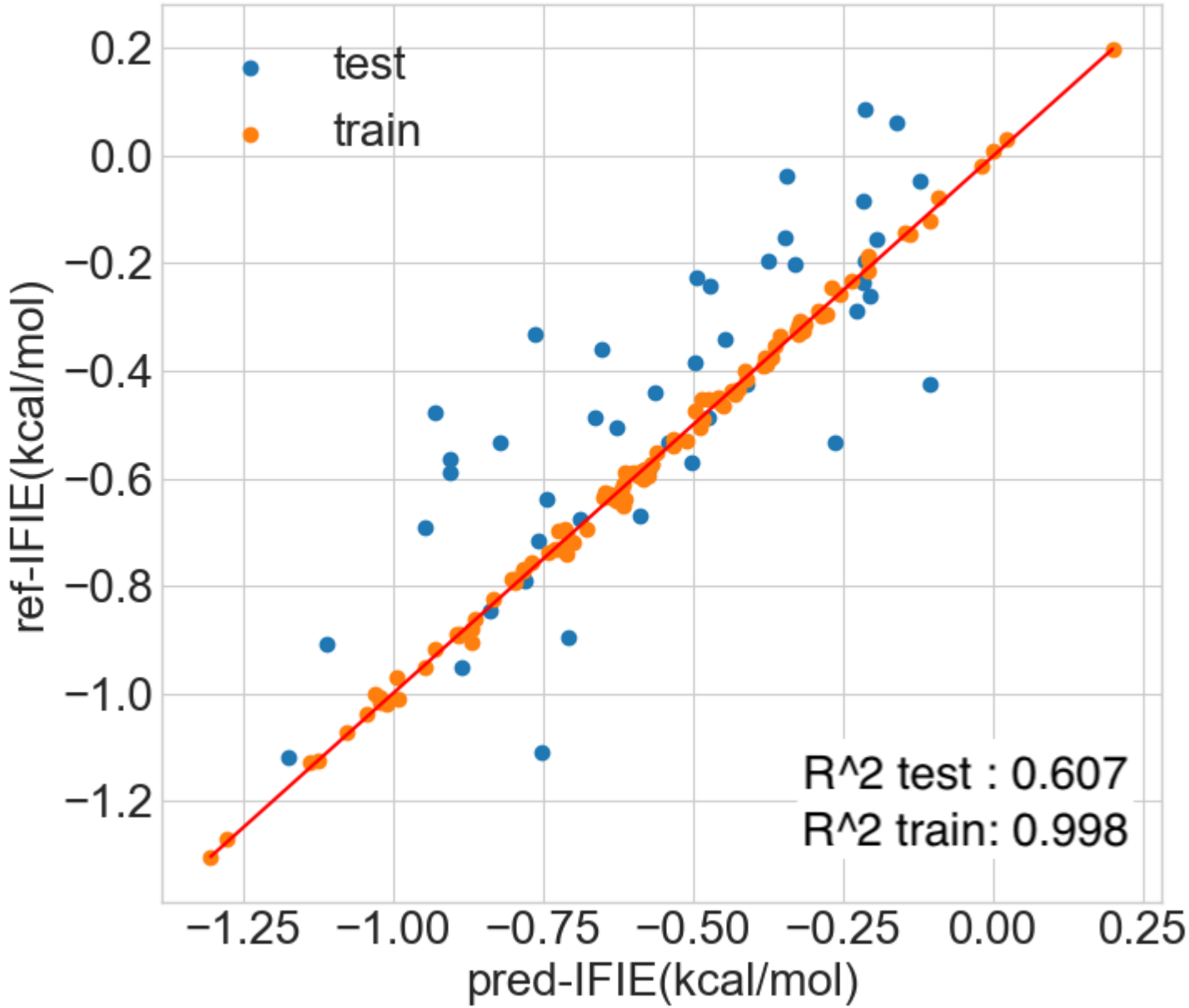

test
train
ref-IFIE(kcal/mol)
pred-IFIE(kcal/mol)
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.25
−1.00
−0.75
−0.50
−0.25
0.00
0.25
R^2 test : 0.607
R^2 train: 0.998

Figure 4 (c)

(d)

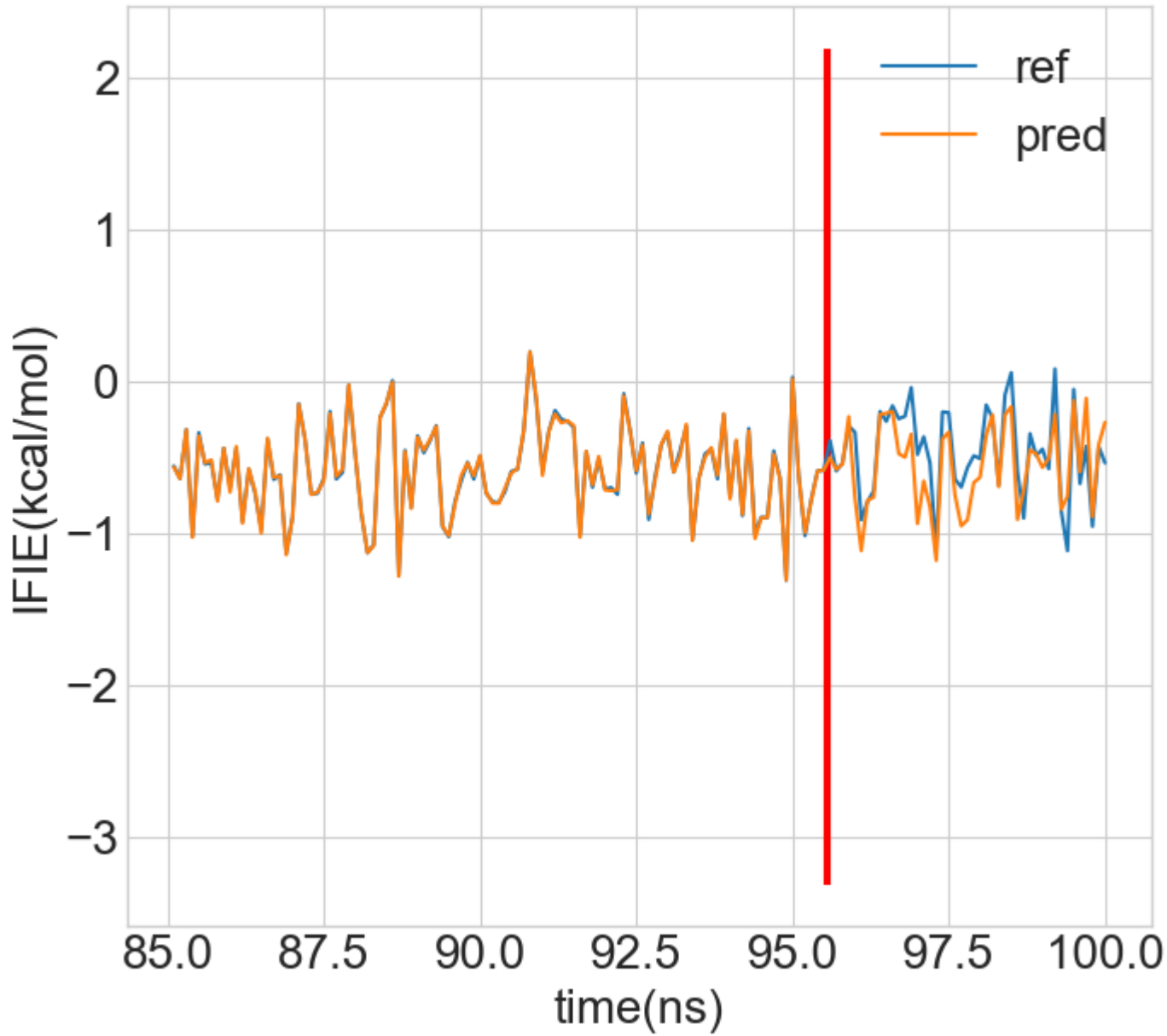

ref
pred
IFIE(kcal/mol)
2
1
0
−1
−2
−3
85.0
87.5
90.0
92.5
95.0
97.5
100.0
time(ns)

Figure 4 (d)

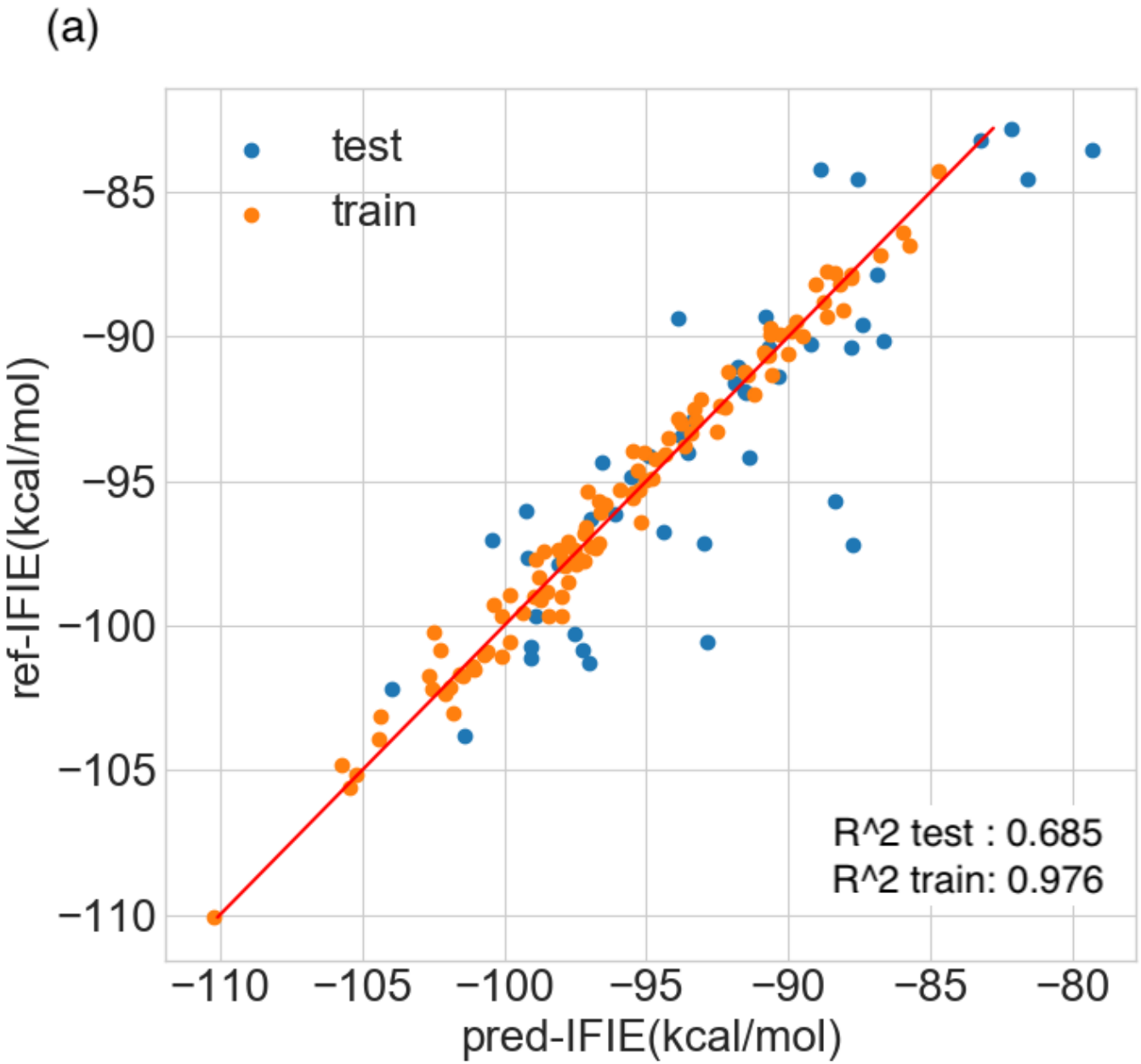

(a)
test
train
−85
−90
−95
−100
−105
−110
ref-IFIE(kcal/mol)
−110
−105
−100
−95
−90
−85
−80
pred-IFIE(kcal/mol)
R^2 test : 0.685
R^2 train: 0.976

Figure 5 (a)

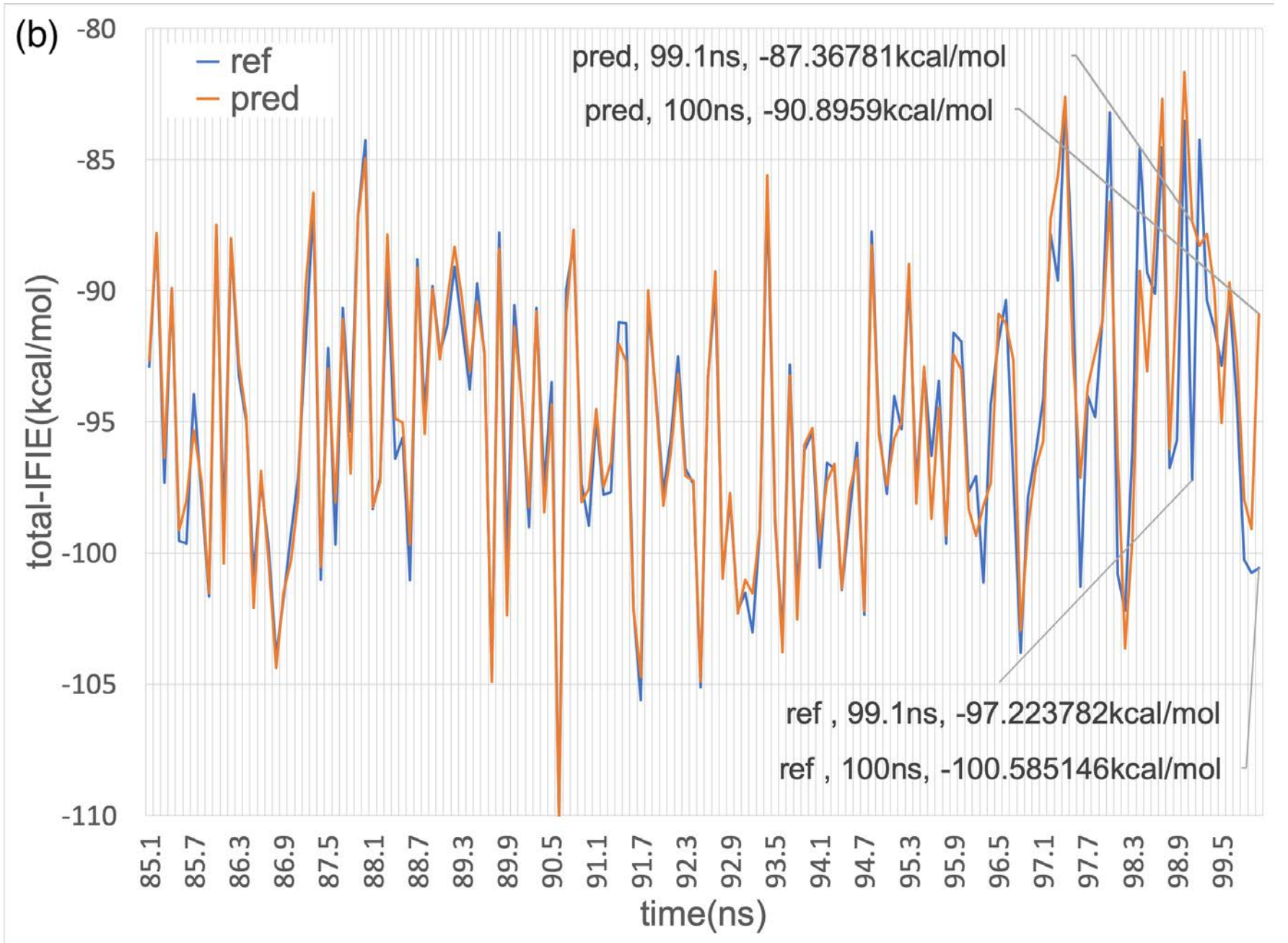

(b)
ref
pred
pred, 99.1ns, -87.36781kcal/mol
pred, 100ns, -90.8959kcal/mol
ref , 99.1ns, -97.223782kcal/mol
ref , 100ns, -100.585146kcal/mol
total-IFIE(kcal/mol)
-80
-85
-90
-95
-100
-105
-110
85.1
85.7
86.3
86.9
87.5
88.1
88.7
89.3
89.9
90.5
91.1
91.7
92.3
92.9
93.5
94.1
94.7
95.3
95.9
96.5
97.1
97.7
98.3
98.9
99.5
time(ns)

Figure 5 (b)

(c)

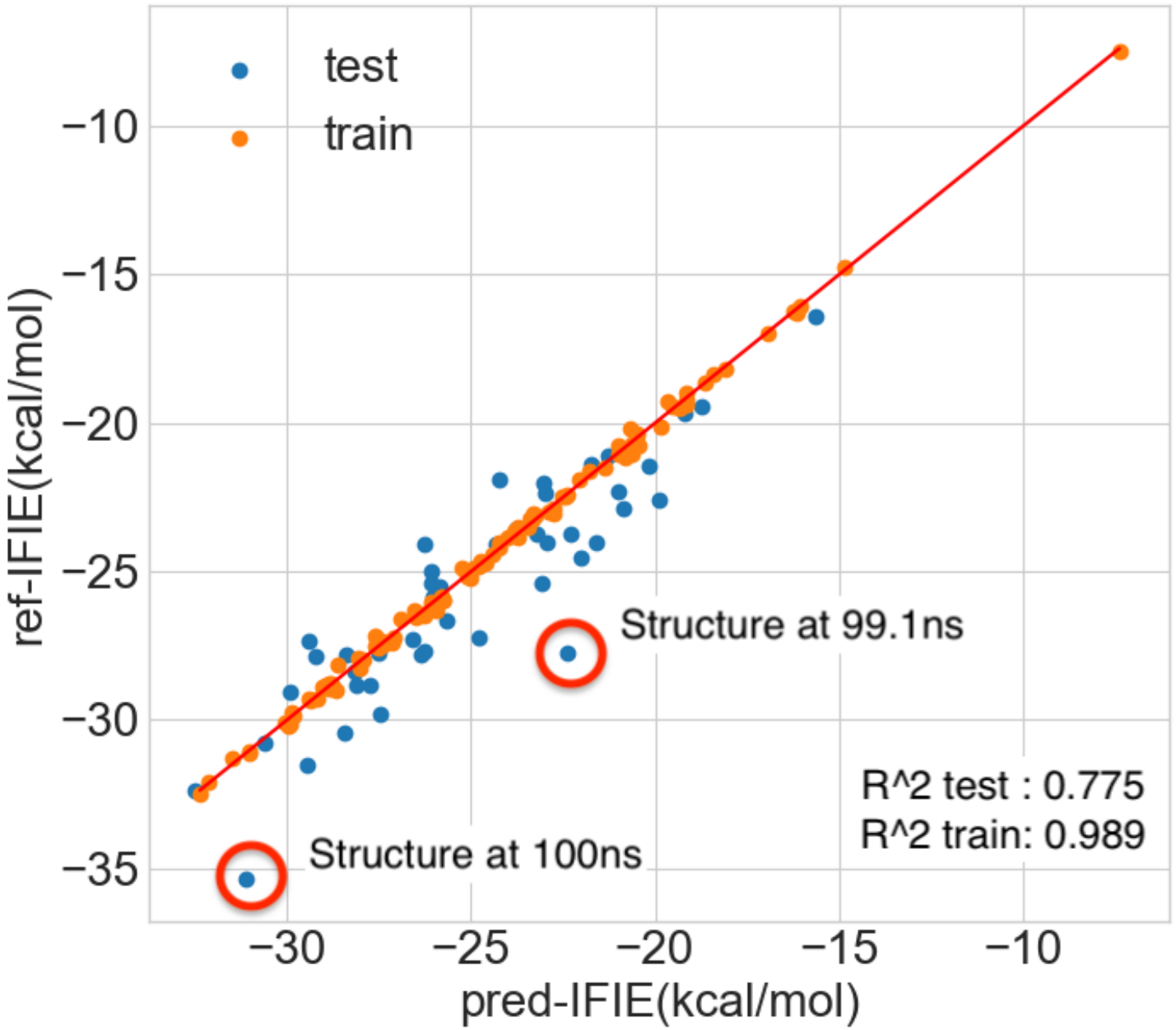

test
train
−10
−15
−20
−25
−30
−35
ref-IFIE(kcal/mol)
Structure at 99.1ns
Structure at 100ns
R^2 test : 0.775
R^2 train: 0.989
−30
−25
−20
−15
−10
pred-IFIE(kcal/mol)

Figure 5 (c)

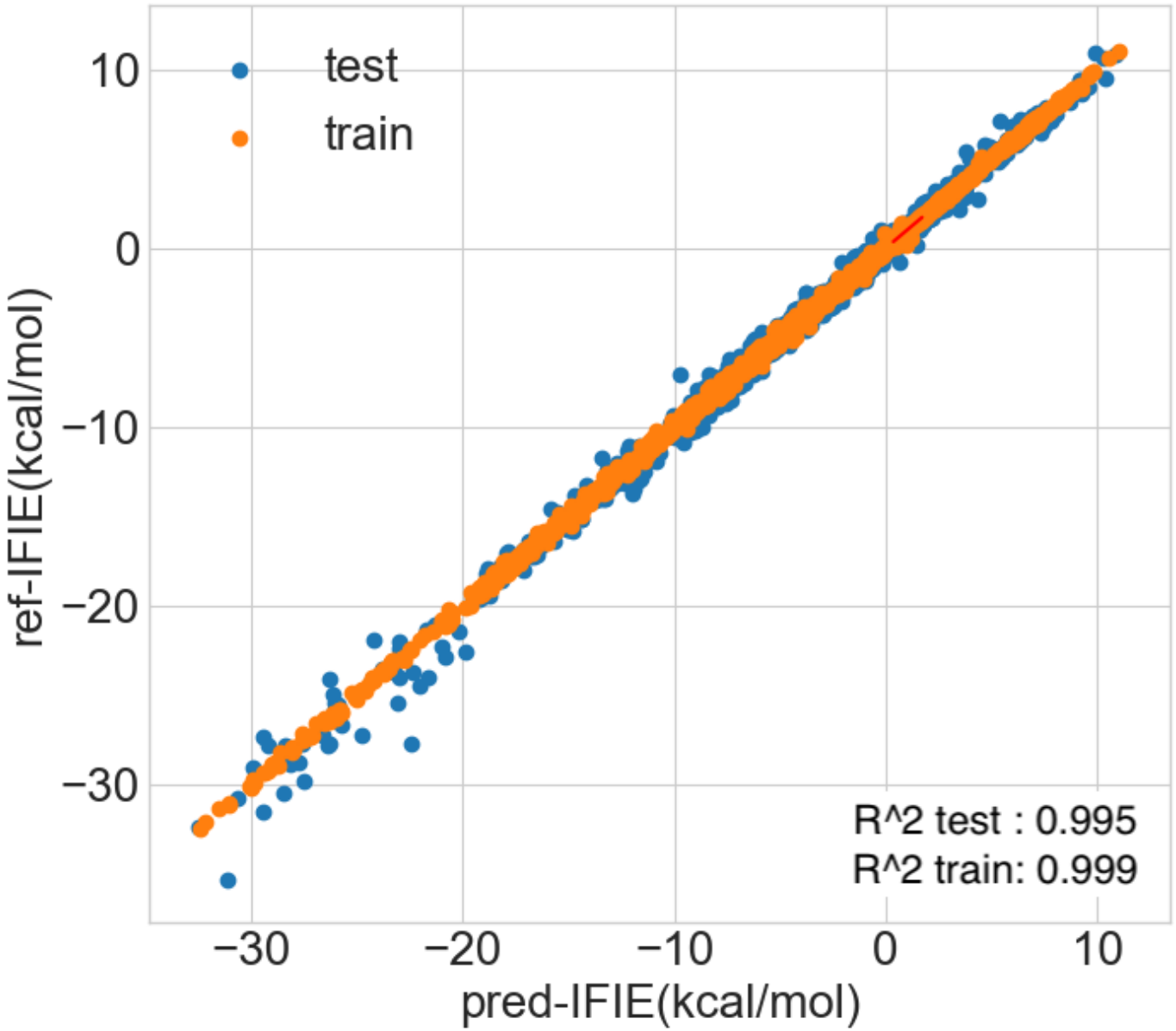

test
train
10
0
−10
−20
−30
ref-IFIE(kcal/mol)
−30
−20
−10
0
10
pred-IFIE(kcal/mol)
R^2 test : 0.995
R^2 train: 0.999

Figure 6

Supplementary data for "Machine learning prediction of inter-fragment interaction energies between ligand and amino-acid residues on the fragment molecular orbital calculations for Janus kinase - inhibitor complex"

Shusuke Tokutomi

Graduate School of System Informatics, Department of Computational Science, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Kohei Shimamura

Department of Physics, Kumamoto University, Kumamoto 860-8555, Japan

Kaori Fukuzawa

Department of Physical Chemistry, School of Pharmacy and Pharmaceutical Sciences, Hoshi University, 2-4-41 Ebara, Shinagawa, Tokyo 142-8501, Japan

Shigenori Tanaka

Graduate School of System Informatics, Department of Computational Science, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

---

Supplementary figures

In this document, supplementary figures S1-S4 are given with their legends.

Email address: tanaka2@kobe-u.ac.jp; Fax:+81-78-803-6620 (Shigenori Tanaka)

Figure captions

Figure S1: (a) IFIE prediction result between ligand and LYS908 by the proposed model (Model I). The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (b) The results for time-series 150 structures are plotted with IFIE on the ordinate and time on the abscissa. The orange/blue lines indicate the temporal change of the predicted/correct values according to the change of structure. The left side of the red vertical line refers to the results for the training data set, and the right side the results for the test data set.

Figure S2: (a) IFIE prediction results for 50 residues by Model II. The ordinate refers to the value of R^2, and the abscissa the selected 50 residues. The blue bar represents the result of R^2 for the test data set, and the orange bar the result of R^2 for the training data set. R^2 was calculated from Eq. (9) in Sec. 2.2.4 in the main text. (b) Total-IFIE prediction result by Model II. The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (c) All IFIE prediction result by Model II. The results for all 150 structures of 50 residues are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set,

and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$), which is virtually hidden behind the points.

Figure S3: Total-IFIE prediction result by Model III. The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$).

Figure S4: (a) IFIE prediction results for 50 residues by Model IV. The ordinate refers to the value of R^2, and the abscissa the selected 50 residues. The blue bar represents the result of R^2 for the test data set, and the orange bar the result of R^2 for the training data set. R^2 was calculated from Eq. (9) in Sec. 2.2.4 in the main text. (b) Total-IFIE prediction result by Model IV. The results for 150 structures are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$). (c) All IFIE prediction result by Model IV. The results for all 150 structures of 50 residues are plotted with the correct value on the ordinate and the predicted value on the abscissa. The blue points refer to the results for the test data set, and the orange points the results for the training data set. If $R^2 = 1$, all points are plotted on the red line ($y = x$), which is virtually hidden behind the points.

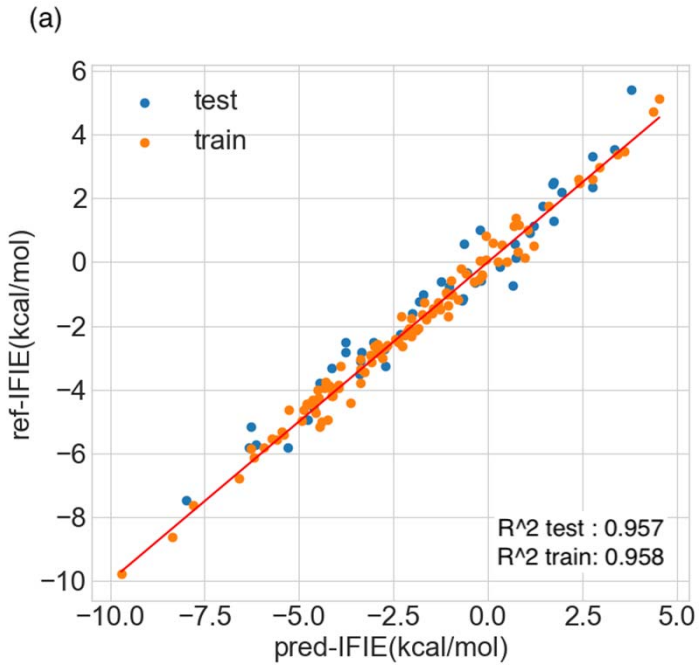


Figure S1 (a)

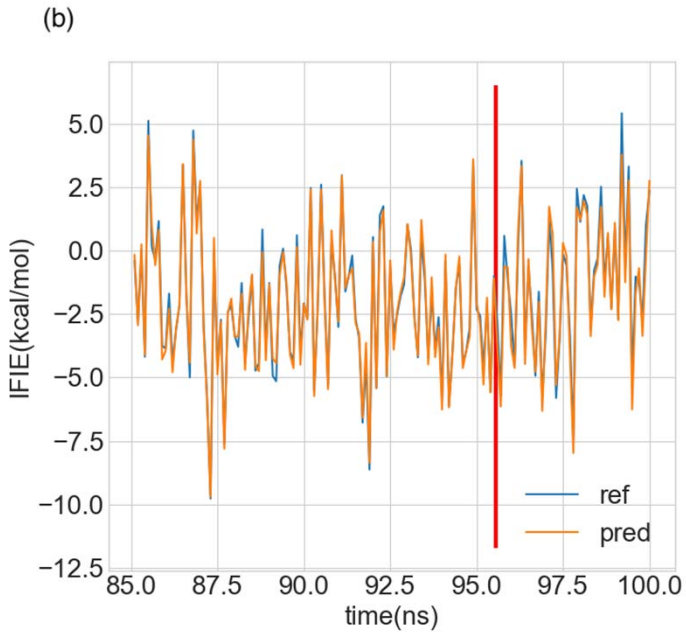


Figure S1 (b)

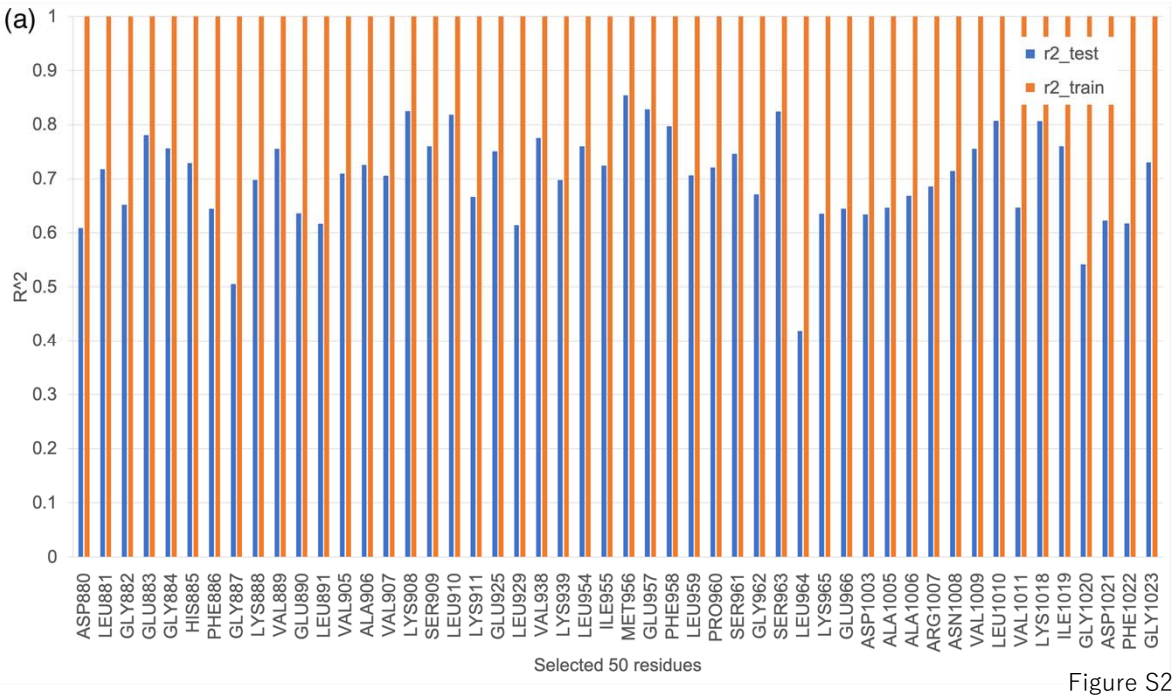


Figure S2 (a)

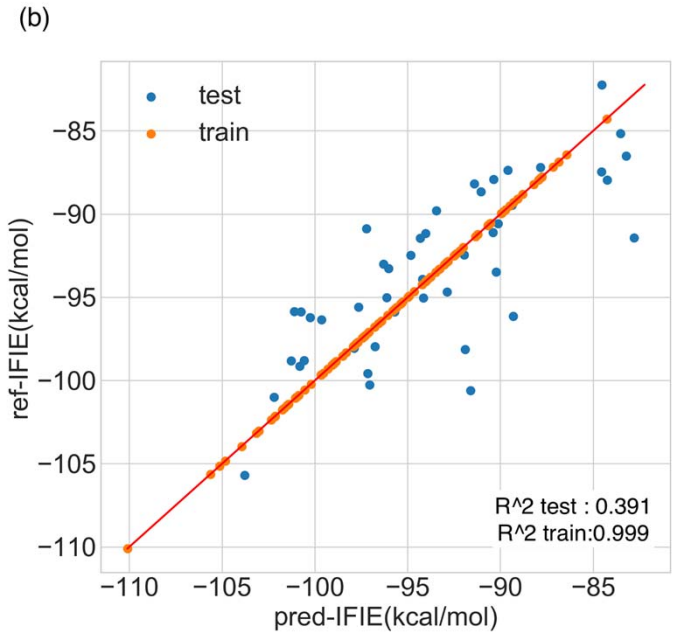


Figure S2 (b)

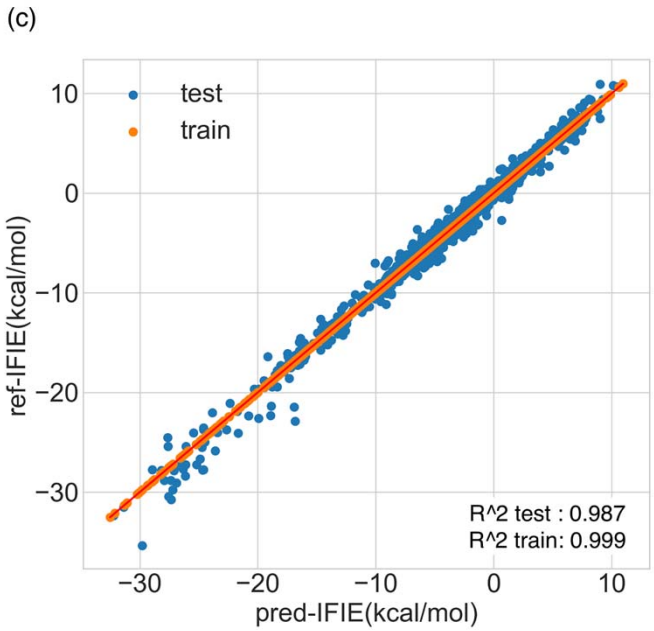


Figure S2 (c)

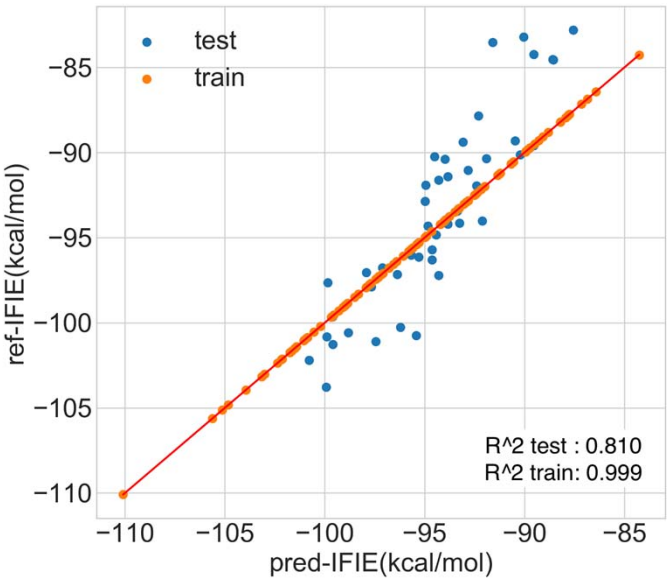


Figure S3

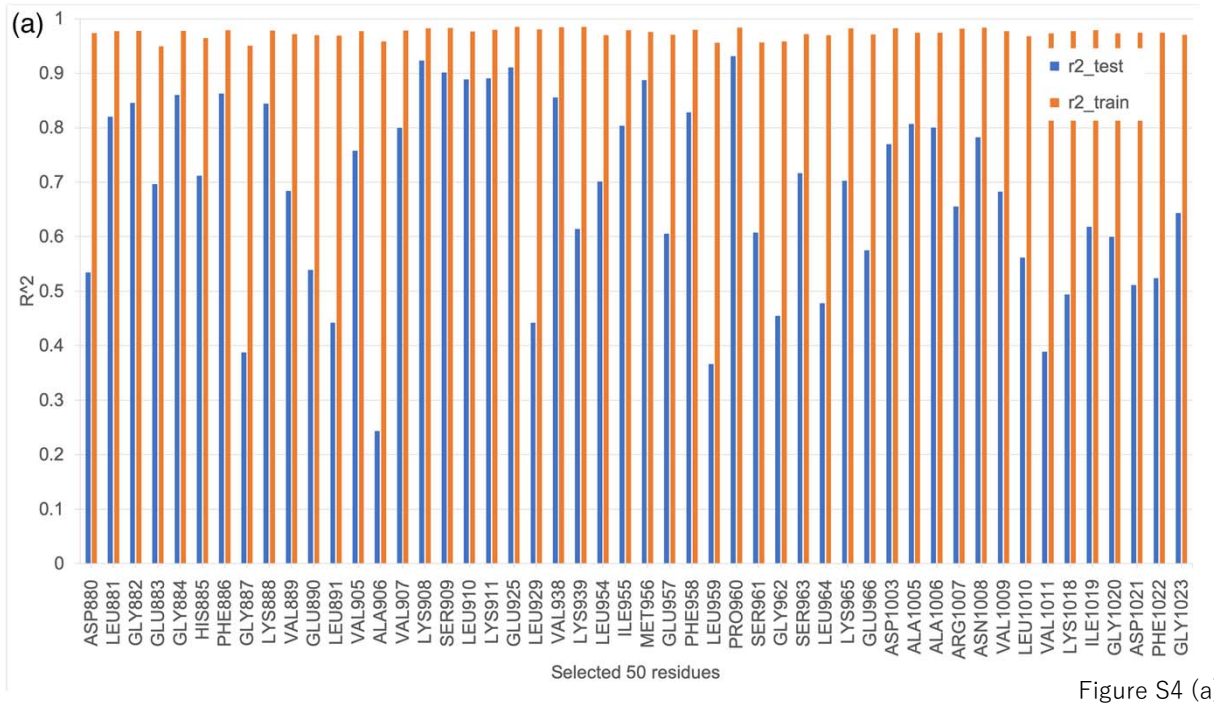


Figure S4 (a)

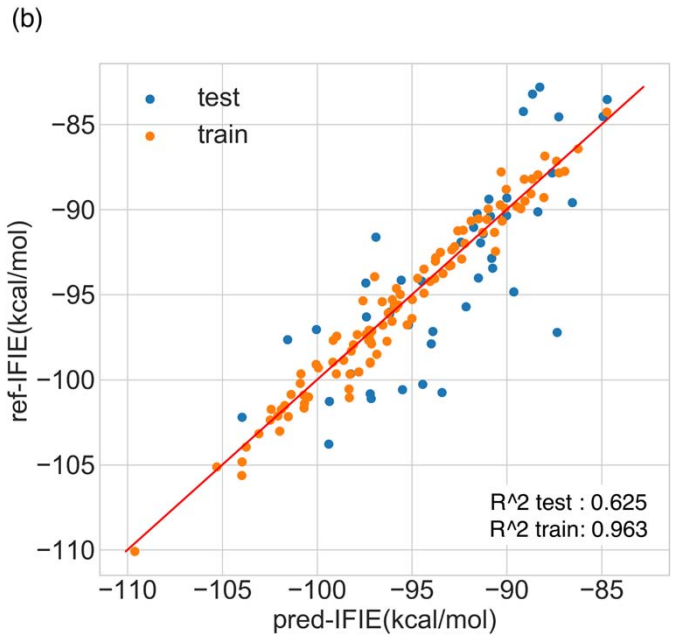


Figure S4 (b)

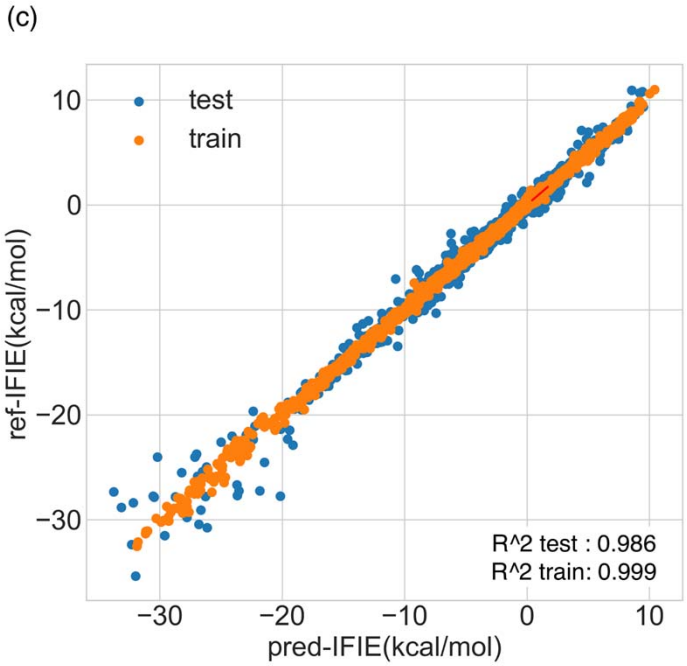

(c)
test
train
ref-IFIE(kcal/mol)
pred-IFIE(kcal/mol)
10
0
−10
−20
−30
−30
−20
−10
0
10
R^2 test : 0.986
R^2 train: 0.999

Figure S4 (c)