

# Chemistry-informed Machine Learning Explains Calcium-binding Proteins' Fuzzy Shape for Communicating Changes in the Atomic States of Calcium Ions

Pengzhi Zhang,<sup>1,\*</sup> Jules Nde,<sup>2,\*</sup> Yossi Eliaz,<sup>3,4</sup> Nathaniel Jennings,<sup>3</sup> Piotr Cieplak,<sup>5</sup> Margaret. S. Cheung<sup>2,6</sup>

1. Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA

2. Department of Physics, University of Washington, Seattle, WA, USA

3. Department of Physics, University of Houston, Houston, TX, USA

4. Computer Science Department, HIT Holon Institute of Technology, Holon, Israel

5. Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA

6. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA

\* These authors contributed equally

Keywords: ion charge state, calcium-binding protein, EF-hand motif, graph theory, machine learning explanation, many-body interactions, calmodulin

## Abstract

Proteins' fuzziness are features for communicating changes in cell signaling instigated by binding with secondary messengers, such as calcium ions, associated with the coordination of muscle contraction, neurotransmitter release, and gene expression. Binding with the disordered parts of a protein, calcium ions must balance their charge states with the shape of calcium-binding proteins and their versatile pool of partners depending on the circumstances they transmit, but it is unclear whether the limited experimental data available can be used to train models to accurately predict the charges of calcium-binding protein variants. Here, we developed a chemistry-informed, machine-learning algorithm that implements a game theoretic approach to explain the output of a machine-learning model without the prerequisite of an excessively large database for high-performance prediction of atomic charges. We used the *ab initio* electronic structure data representing calcium ions and the structures of the disordered segments of calcium-binding peptides with surrounding water molecules to train several explainable models. Network theory was used to extract the topological features of atomic interactions in the structurally complex data dictated by the coordination chemistry of a calcium ion, a potent indicator of its charge state in protein. With our designs, we provided a framework of explainable machine learning model to annotate atomic charges of calcium ions in calcium-binding proteins with domain knowledge in response to the chemical changes in an environment based on the limited size of scientific data in a genome space.

## I. Introduction

Proteins' fuzzy structures are features for communicating in biological processes,<sup>1-3</sup> instigated by secondary messengers such as calcium ions ( $\text{Ca}^{2+}$ ). The binding of  $\text{Ca}^{2+}$  almost always induces local or global conformational changes to the protein due to the alteration of its atomic charge or electrostatic interaction. These conformational changes and the dynamics of calcium-binding proteins are intimately linked to the function of a calcium-binding protein, indicating that the surrounding environments influence its structural dynamics, thereby impacting the accessibility of its several calcium-binding loops.<sup>2,4,5</sup> The versatility of calcium-binding proteins, often characterized by disordered domains, underscores the need of unraveling the causal relationship in determining the atomic charge states of  $\text{Ca}^{2+}$  and the corresponding protein configurations. *Ab initio* calculations have been employed for this purpose, showing that the determination of the atomic charge of  $\text{Ca}^{2+}$  requires information on its coordination geometry in the binding motif of a calcium-binding protein.<sup>6</sup> However, this approach is not only labor intensive due to in part the complexity of the protein environment involving water molecules, but also computationally expensive, making it infeasible to explore a broad configurational space of calcium-binding proteins in a changing environment. Because determining the atomic charge states is resource-intensive, it is extremely valuable to develop a machine learning (ML) algorithm that quickly identifies and explains the molecular descriptors influencing the atomic charge states without the necessity of a large database particularly in a genome space.

Calcium is a versatile signaling ion that plays a pivotal role in regulating downstream signaling targets in numerous biological functions, including muscle contraction, neurotransmitter release, and gene expression.<sup>2,4,5,7-10</sup> Calcium ion performs these biological functions through its

interactions with numerous calcium-binding proteins, including calmodulin (CaM), which serves as the main calcium sensor and effectors.<sup>11</sup> Upon calcium binding, these proteins undergo conformational changes, enabling them to perform specific biological functions.<sup>12,13</sup> Additionally, the conformational changes reciprocally alter the ion charge state, which deviates from the conventional divalent charges (or +2e) observed in the crystal structures of calcium-binding proteins, due to the binding affinity that is influenced by the local protein environment and the presence of other ions or molecules.<sup>14,15</sup> The conformational changes affect the charge state, as atomic interactions are primarily dictated by the coordination number of a divalent ion,<sup>16-18</sup> in comparison to the atomic charges or the polarization effects.<sup>19</sup> It is essential for comprehending the intricacies of the causal functions,<sup>14</sup> such as the reciprocity observed among  $\text{Ca}^{2+}$ , CaM, and the CaM-target proteins in the context of synaptic plasticity.<sup>12</sup>

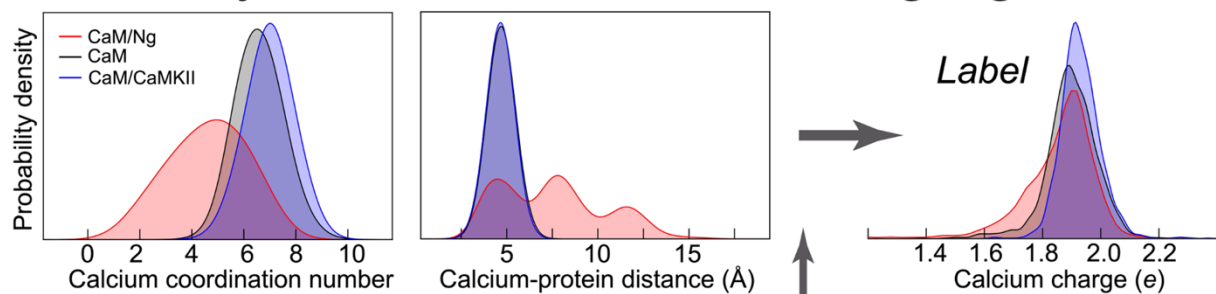
Here, we used a “sandbox” system of  $\text{Ca}^{2+}$  and calcium-binding loops from shape-shifting CaM<sup>20</sup> as a use case to investigate how  $\text{Ca}^{2+}$  brings the loops to adapt into distinct geometrical shapes dynamically, illustrating the role of their chemistry in the interplay of electrostatic forces, conformational changes, and molecular recognition within the biologically complex milieu. The calcium-binding loop is an evolutionarily conserved helix-coil-helix EF-hand motif that chelates  $\text{Ca}^{2+}$  with its oxygen-rich amino acids, such as aspartate and glutamate acids, into a three-dimensional global form.<sup>7</sup> One of the amino acids from the EF-hand motif coordinates the  $\text{Ca}^{2+}$  through a water molecule. This water molecule complicates the shape of the EF-hand motif<sup>14</sup> as if it rolls like a ball bearing, transitioning the EF-hand motif into various coordination geometries,<sup>21-23</sup> from octahedral to pentagonal bipyramidal. Alternating through these distinctive geometries of a calcium binding loop,  $\text{Ca}^{2+}$  interacts with a combinatorial set of conserved and non-conserved amino acids that affect the charge state of the  $\text{Ca}^{2+}$ .

Despite a proliferation of ML models that establish the quantitative relations between configurational space and atomic charge or between configuration and energy in materials,<sup>24-29</sup> investigations into divalent ions such as Ca<sup>2+</sup> in protein environments are limited. This is due to a lack of accumulative data that covers a broad parameter space in biomolecules or materials, describing the causal relations for model training at the interface of organic and inorganic compounds. Among the burgeoning landscape of ML applications in molecular biology,<sup>28,30-33</sup> the interpretability of model outputs remains a critical challenge because changes in the input features nonlinearly contribute to the output. This creates a growing tension between the accuracy and interpretability of model predictions.<sup>34,35</sup>

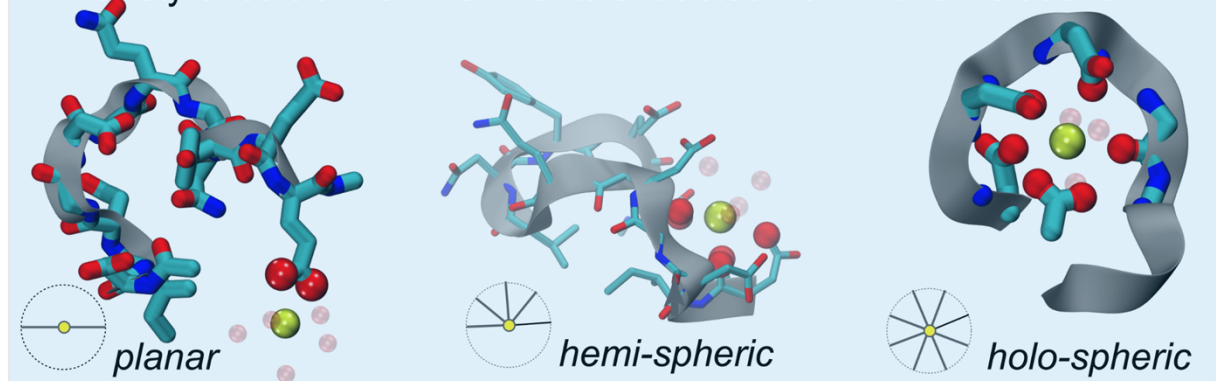
To bridge the technical gap, we developed an ML algorithm (**Fig. 1**) that implements a game theoretic approach to explain the output of an ML model with a limited training dataset of over 7,000 configurations in our *sandbox* system. This dataset was curated from a combined *ab initio* calculations<sup>14</sup> and the structures of calcium-binding peptides from molecular dynamics simulations guided by the chemical coordination of Ca<sup>2+</sup>.<sup>36,37</sup> We deployed network theory to uncover topological features that best capture the molecular descriptors to the charge states of Ca<sup>2+</sup> within the chemically complex datasets. We evaluated the SHAP (SHapley Additive exPlanations) values of these explainable features,<sup>34,38</sup> a concept rooted in cooperative game theory, for increasing transparencies of individual features that affect the ML models (see *Supplementary* for further description). The chemistry-informed ML algorithm explains the causal relationship between the unique characteristics of Ca<sup>2+</sup> embedded in various protein environments and their charge states.

We found that the coordination number of a  $\text{Ca}^{2+}$  contributes the most to the determination of the calcium atomic charge state in various protein environments. With several other additive features, we applied the explainable models to annotate the calcium charge states of  $\text{Ca}^{2+}$  binding motifs for another dataset of over 400 curated experimentally derived calcium-binding loops (**Table S1** in *Supplementary information*). The model annotates the atomic charge state of calcium according to the unique molecular descriptors of a configuration. Although the principle of target selection in calcium-binding proteins remains unknown due to the astronomical possibilities for a shape-shifting calcium-binding protein to connect with a versatile pool of downstream partners, we harnessed an explainable, chemistry-informed ML algorithm, to annotate calcium's atomic charge state by unraveling the causal features from the coordination chemistry that influence the shape of the loop in diverse protein environments, giving rise to robust regulations in calcium signaling.

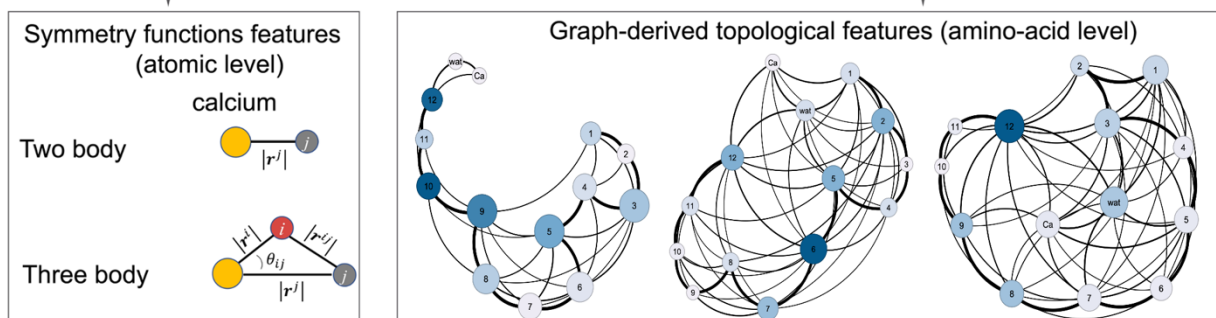
# Chemistry-informed machine learning algorithm



A variety of calcium environments extracted from MD simulations.

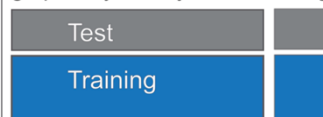


Chemistry-informed feature extraction



XGBoost predictive model

graph + symmetry features



Explanation

Important features by ML model lead to better understanding of calcium binding in protein.

Feature	Score
deg_Ca	1.1
clos_Ca	0.7
bet_r10	0.5
...	

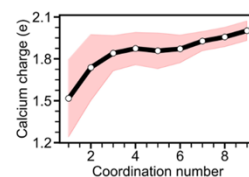


Figure 1. Chemistry-informed ML enhances predictive power of atomic configuration-to-charge relation for  $\text{Ca}^{2+}$ . A dataset of over 7000 configurations was curated from a combined *ab initio* calculations

of calcium-binding peptides and the structures of calmodulin-binding targets from molecular dynamics simulations guided by the chemical coordination of calcium<sup>36,37</sup> (the middle panel in light blue background). The shapes of the loops are categorized as planar, hemi-spheric, and holo-spheric coordination geometries. Variation of the loops shifts with the coordination numbers of calcium ions as well as the calcium-protein distances, tuning the atomic charge of the Ca<sup>2+</sup> (top panel). The graph-derived topological features were trained for explainable ML models to increase the transparencies and interpretability of features accounted by the many-body interactions. The additive features from many-body interactions further improved the accuracy of the ML algorithm compared to those that used only the two-body and three-body interactions as training features. The explainable XGBoost model computed and ranked the key features and relevant physical quantities in the charge state of calcium ions that impact most on the constructure of ML model.

## II. Results

**Overview: Chemistry-informed ML algorithm enhances the predictive power of a configuration-to-charge analysis by computing graph-derived topological features.**

A single calcium ion is chelated by three to eight oxygen ligand atoms with an average of six ligands for all Ca<sup>2+</sup> binding loops or seven ligands for only EF-hand loops or by oxygen atoms from the nearby water molecules<sup>37,39</sup> (**Fig. 2a**). The chemical coordination of a calcium ion involves chelation with various classes of oxygen atoms, including those from the carboxyl groups (Asp, Glu), the carboxamide groups (Asn, Gln), and the hydroxyl groups (Ser, Thr) in the sidechains, as well as the carbonyl oxygen atoms of most residues in the main chain or cofactors, and water molecules, predominantly through electrostatic interactions. These oxygen atoms, in turn, are covalently linked to other heavy atoms such as carbon and nitrogen, forming a secondary shell. Additionally, a tertiary shell of carbon or nitrogen atoms can be delineated, consisting of atoms covalently bonded to the secondary shell.<sup>37,39,40</sup>



The configurations of ligands' oxygen atoms can be roughly classified into three categories based on their geometries: planar, hemi-directed (or hemi-spheric), and holo-directed (or holo-spheric) (**Fig. 1**). In a typical energy-minimized calcium-binding loop where the  $\text{Ca}^{2+}$  is tightly bound holo-directedly, seven coordinating oxygen atoms are arranged to form a bipyramidal pentagonal coordination geometry, as shown in the crystal structure of  $\text{Ca}^{2+}/\text{CaM}$  (**Fig. 2a**). These concentric arrangements of atoms form a framework that defines the central binding site of calcium. Consequently, a set of physical parameters describing the spatial arrangement of atoms comprising the binding site can be delineated by the " $\alpha_{r_c}$ " / " $\alpha\beta_{r_c}$ " for the radial/angular distribution functions. " $\alpha$ " or " $\beta$ " is the element name (H, C, N, O) of atom surrounding the  $\text{Ca}^{2+}$ , and " $r_c$ " is the cutoff distance within the concentric shell. The geometrical arrangement of the atom clusters also provides information about the closeness and betweenness centrality of  $\text{Ca}^{2+}$ , allowing for selective differentiation between  $\text{Ca}^{2+}$  and other relevant divalent cations (see Supplementary **Fig. 1** for a complete description of features and the *Supplementary Information* for the definition of parameters).

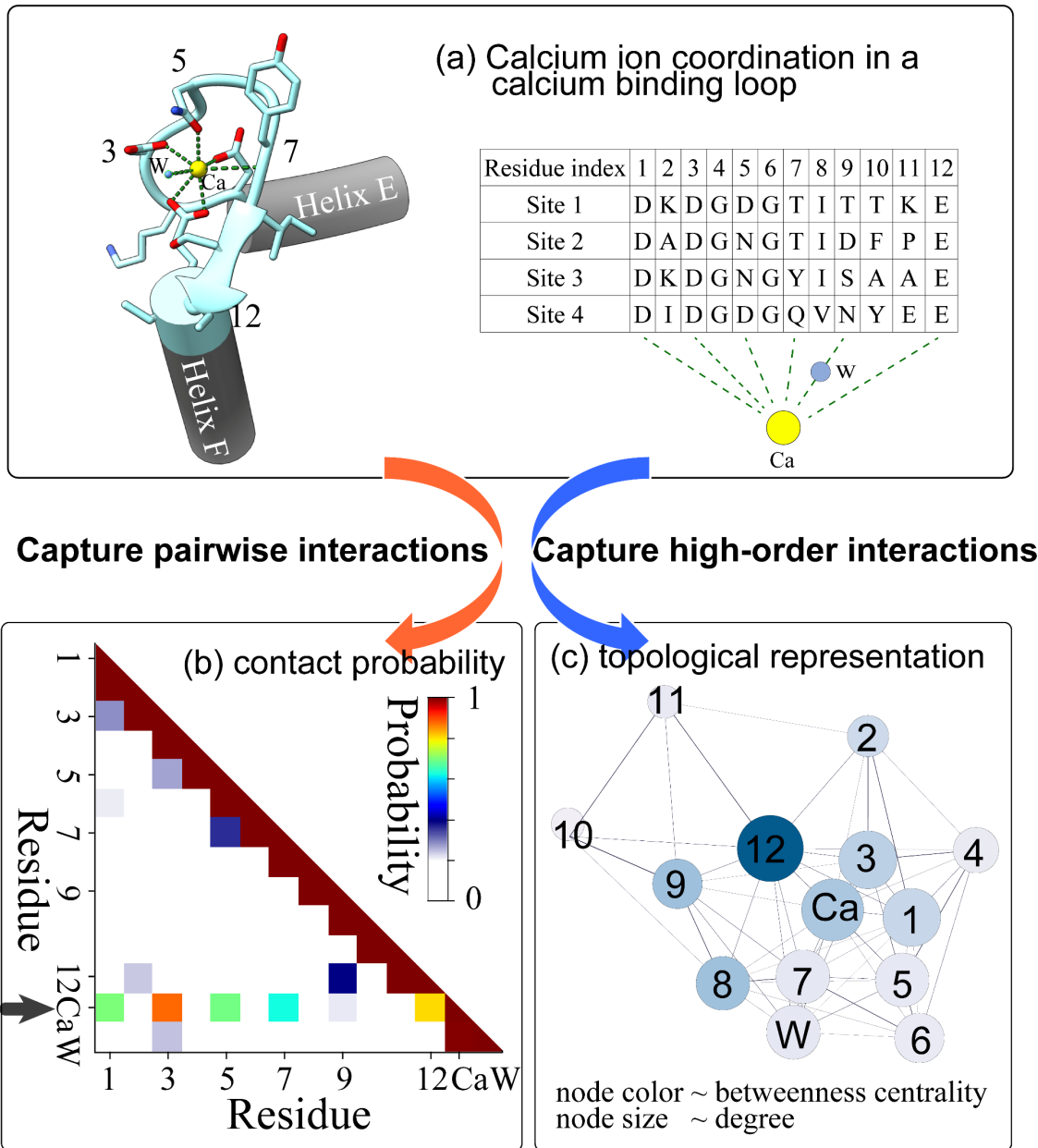
We aim to use the curated conformations of the calcium-binding motifs in CaM and CaM-target bound complexes, which provide examples of the coordination geometries as a sandbox dataset, to train a chemistry-informed ML model. The domain knowledge of chemical principles (or multiple geometries of the calcium motifs) provided explainable molecular descriptors that influence the atomic charge state of calcium in various protein microenvironments. The proposed algorithm includes three major steps to capture these unique geometries from the structures:

Step 1. Compute the feature of  $\text{Ca}^{2+}$  and its coordination with a calcium-binding loop from the curated datasets: With the sandbox dataset, we calculated the distribution functions to account for the spatial arrangement of the atoms surrounding the  $\text{Ca}^{2+}$  and graph-based topological parameters to include high-order interactions. **Fig. 2b** shows pairwise interactions in terms of the contact probability map that captures probable interactions between the  $\text{Ca}^{2+}$  and the conserved residues at the first, third, fifth, seventh, and twelfth positions, and the interaction between the ninth and twelfth residues. Featuring two-body interactions, we also capture the radial distribution functions around the  $\text{Ca}^{2+}$ . To capture the higher order interactions in the atom representation involved in the calcium coordination in a calcium-binding loop, we used a mathematical graph representation of the system and computed their topological parameters. The advantage of using measurements from graph theory over atomic contacts is that they capture the shape of the data in a complex system with voids and gaps in the structures.<sup>41</sup> The betweenness centrality of a node contains convoluted information about the entire graph of the calcium-binding loop, which may not be easily derived from a pairwise interaction map. For example, high betweenness centrality of the twelfth residue node suggests that the node is critical to maintain the connectivity of the network (**Fig. 2c**).

Step 2. Establish explainable ML algorithms to evaluate additive features. To assess the effectiveness of the extracted features, we designed three algorithms that learn the chemistry of many-body interactions in models. Model I: The radial distribution functions around the  $\text{Ca}^{2+}$ , featuring only the two-body pairwise interactions; Model II: In addition to Model I, the angular distribution functions around the  $\text{Ca}^{2+}$ , featuring the three-body

interactions; and Model III: in addition to Model II, the topological order parameters from the graphs, featuring the higher order interactions. Then, we trained them with the XGBoost<sup>42</sup> regression models that were evaluated by the mean absolute error (MAE) in the calcium atomic charges, the coefficient of determination ( $r^2$ ), and Pearson correlation coefficient (R) (see *Supplementary Figs. 2 and 3*).

*Step 3. Explain the feature importance for the interpretation of outcomes from the ML models.* The XGBoost regression model calculates the importance score according to the averaged contribution of a specific feature to the relevant decision trees. We explained the importance of each feature by their SHAP values.<sup>34</sup> The SHAP value, widely used in game theory to evaluate the marginal contribution of each player in a cooperative multiplayer game (see *Supplementary* for the equation and further details), provides a principled way to explain the predictions of nonlinear models common in ML. Positive (negative) SHAP values indicate that the feature positively (negatively) contributes to the model's prediction; thus, a higher positive (or negative) SHAP value suggests that the feature has a stronger influence on increasing (or decreasing) the prediction for a specific instance. Zero SHAP values imply that the feature does not contribute to the prediction for the given instance.

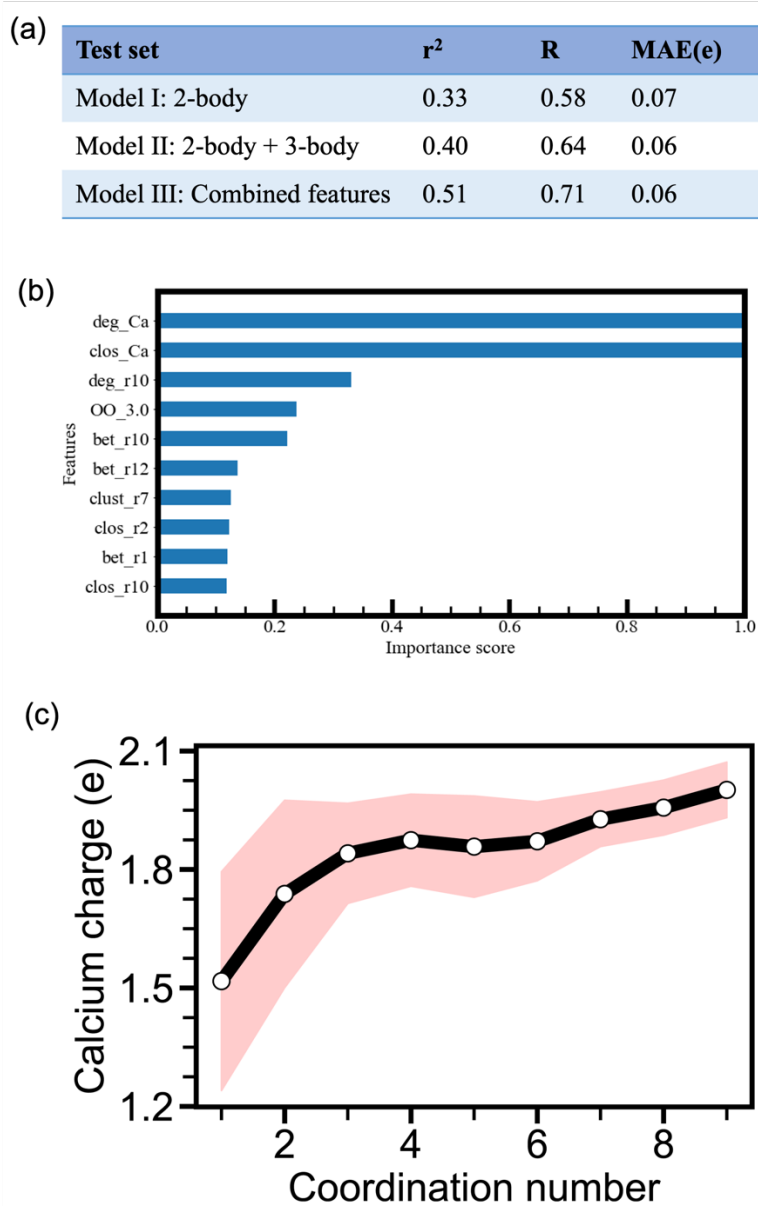


**Figure 2. Calcium-coordination features extracted to include pairwise and high-order interactions among residues in the calcium-binding loops.** (a) A schematic representation of the calcium-coordination structure. The calcium-binding loop is composed of 12 amino acids, and we relabeled the residue indices according to the amino-acid alignment. In this geometry, the coordination oxygen atoms come from the carbonyl groups in the sidechains of the glutamate/glutamine/aspartate/asparagine at the first, third, fifth, and twelfth position and the backbones of the amino acid at the seventh position (middle of the loop). The residues at the ninth position usually coordinate a  $\text{Ca}^{2+}$  through a water molecule (W). (b) A contact probability map is

used for capturing the pairwise interactions between residues/ion/water molecules in the system. (c) A mathematical graph representation of the system and selected topological parameters captures the higher order interactions involved in the calcium coordination in a calcium-binding loop.

### **Explainable machine learning reveals important additive features in training models**

To explain the importance of additive features, we constructed three explainable ML models while using the Pearson correlation coefficient (R) to test the accuracy of each model. We randomly divided the sandbox dataset into subsets for training (70%) and testing (30%), such as shown in Supplementary **Fig. 2a**. We evaluated the performance of the three explainable ML models with additive features in predicting charges in the test dataset **Fig. 3a**. When we used only the two-body pairwise interactions as features in Model I, the yielded ML model predicts the calcium atomic charge with the mean absolute error (MAE)  $\sim 0.07$  e; there is a positive correlation between the predicted charges and the *ab initio* charges (R = 0.58). When the three-body interactions are also introduced as features in Model II, the accuracy of the ML model improved significantly, where R increases to 0.64. In the third explainable machine learning model in which the network topological parameters were included as features in Model III, there is a significant improvement of the accuracy on the test dataset (R increases to 0.71).



**Figure 3. ML performance and feature importance in determining the calcium charge.** (a) Evaluation statistics of models in predicting the calcium charge,  $r^2$  is the coefficient of determination in the linear fit, R the Pearson correlation coefficient, and MAE mean absolute error. Addition of the network features noticeably improves the predicting performance in the test dataset when comparing Model I, II and III. (b) For Model III, we showed the ranking of the features that contribute most to the prediction of the calcium charge state; (c) for Model III, we showed the dependence of calcium charge on the coordination number, which reflects the top two most important features plotted against each other; mean and standard deviation (pink shaded area) of the calcium charge are plotted against coordination number of  $\text{Ca}^{2+}$ .

For the rest of analyses, we used only the Model III. We sorted the features by their importance scores (**Fig. 3b**) from the best performing explainable machine learning model with network order parameters as additive features. Among the top ten most important features, eight are network order parameters. The top two are closely related to the calcium-coordination number. The network order parameters about the  $\text{Ca}^{2+}$  such as the closeness (i.e. *clos\_Ca*), which describes the average distance between the calcium ion and other coordination residues, and degree centrality (*deg\_Ca*), which describes the total number of connections between the calcium ion and other coordination residues, appear most decisive to the charges of calcium ions. These two features show that calcium plays a central role in the loop by bringing in positive charges while other elements in the loop mostly contribute with negative charge. Surprisingly, the third most important feature is the network order parameter about the tenth residue node (i.e. *deg\_r10*), a non-conserved residue in the calcium loops. This residue is not directly involved in the calcium coordination.

We then investigated the dependence of calcium charges on the coordination number (**Fig. 3c**), as the charges monotonically increases with the calcium coordination number. When the coordination number is less than 3, the mean of the calcium charges is less than 1.80 e and the standard deviation is large  $\sim 0.30$  e. With such low coordination numbers,  $\text{Ca}^{2+}$  is loosely bound in the loop forming a planar coordination geometry. There exist combinatorial ways to bind with oxygen candidates, contributing to the large deviation in the calcium charges. When the coordination number is between three and six, corresponding to a hemi-directed coordination

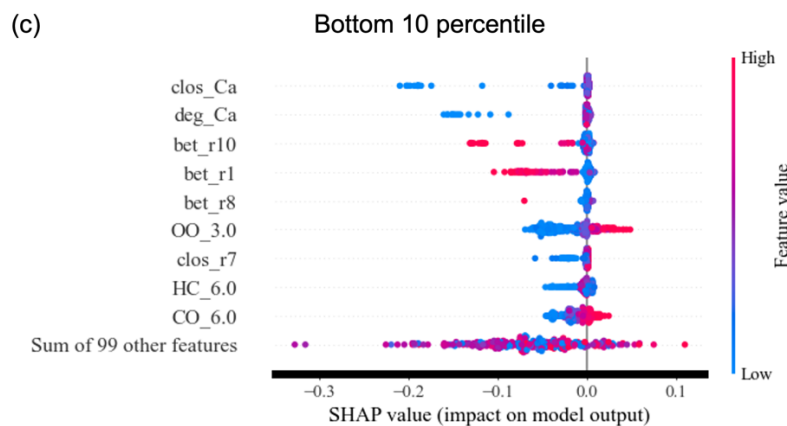
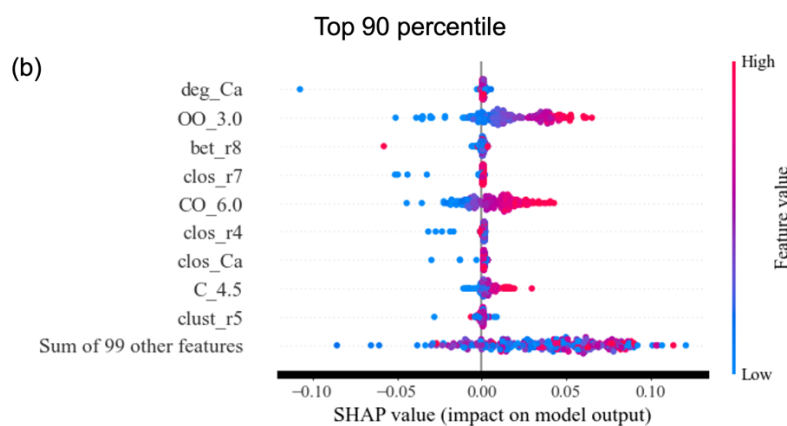
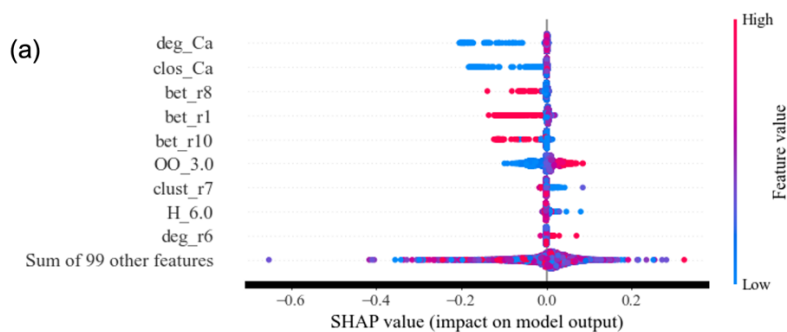
geometry, the mean of the calcium charges is  $\sim 1.85$  e and the standard deviation is  $\sim 0.15$  e. When the coordination number is greater than or equal to six, the mean of the calcium charges monotonically increases with the coordination number from 1.87 e to 2.00 e with a small standard deviation  $\sim 0.07$  e, forming a rigid holo-directed coordination geometry.

### **SHAP values explain the contribution of additive features in annotating the calcium atomic charge states**

By interpreting a model trained on a set of features as a value function on a coalition of players in a theoretic game, SHAP values provide an intuitive way to evaluate the contribution of additive features to a prediction or to the uncertainty of a prediction. We aim to evaluate the contributions of the most influential descriptors in predicting the charge states not only globally (i.e. across the entire dataset), but also for the groups of high and low charges, which will further our functional understanding of the geometries of the calcium-binding loops, such as holo-directed and planar, observed experimentally. Therefore, we computed SHAP values on the entire dataset (**Fig. 4a**) and ranked the features based on their importance in predicting the calcium atomic charge states. The negative (or positive) Shapley values, indicating an effect of decreasing (or increasing) the calcium atomic charge state relative to the mean value. Low values of the `deg_Ca` and `Clos_Ca` variables have high negative contributions in predicting the calcium charge states, while high positive values have almost no contribution. These observations further solidify the central role of the calcium in the binding motif. We further computed the SHAP values of the structures of which the calcium charge is beyond 90 percentile (**Fig. 4b**), and the structures of which the calcium charge is below 10 percentile (**Fig. 4c**) to determine the overall importance of each feature in annotating the calcium atomic charge states in these subgroups. The top 90 percentile of the



structures constitute a category of structures with “charges well above the mean value”, while the bottom 10 percentile of the structures constitute a category of structures with “charges well below the mean value”.



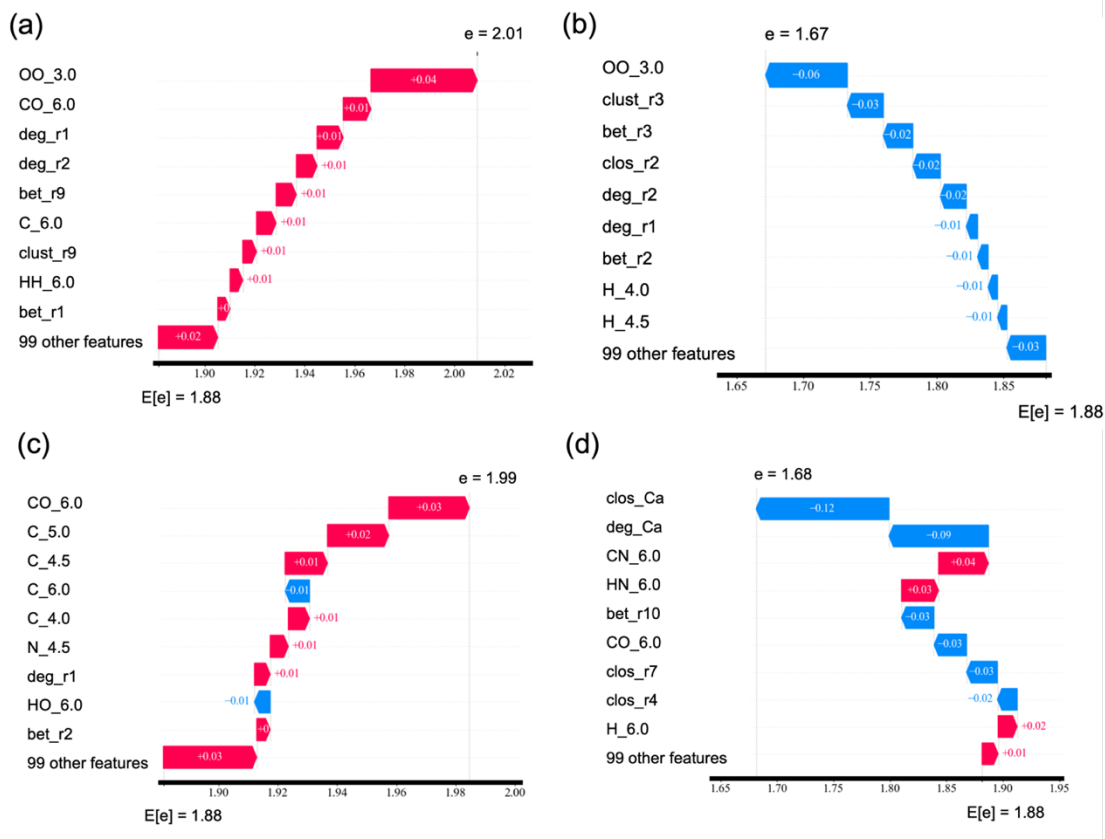
**Figure 4. Ranking of the features based on the SHAP values for their global contribution to the prediction of the calcium atomic charge state.** (a) The SHAP values of most important features influencing the calcium charge state prediction; (b) The SHAP values of the most important features influencing the charges that are above the 90<sup>th</sup> percentile; (c) The SHAP values of the most important features influencing the charges that are below the 10<sup>th</sup> percentile.

In examining **Fig. 4(a, c)**, we noticed a striking similarity among the most important features. The closeness centrality of calcium ion (how central is the calcium ion from the other atoms in the binding loop) together with its degree centrality (how connected the calcium ion is to other atoms in the binding loop) are the most dominant features. Both contribute to decreasing the calcium atomic charge state with lower feature values. This is expected, since calcium ion is the most valuable player in contributing its own charge state. Interestingly, the next top features are those derived from the topological graph network for residues 1, 8 and 10, whose removal in the game theoretic approach contributes to decreasing the calcium atomic charge state annotation. Similarly, in **Fig. 4b**, the topological order parameters for residues 7 and 8 contribute to decreasing the calcium atomic charge state annotation. These examples signify the importance of the geometry of the calcium binding loop to the determination of calcium charge states.

### **Features that influence calcium charge states can be quite diverse for individual charge predictions**

In this section, we unraveled the local impact of each feature on the variability and patterns observed in calcium ion charge states. We calculated the SHAP values for each feature from Model

III (in **Fig. 4**) for a single instance, i.e., a specific calcium atomic charge state. The features that contribute the most to calcium atomic charge above an average value in the database are dominated by diverse features derived from the symmetry functions (**Fig. 5(a, c)**). These features mostly contribute to increasing the charges, except the  $\text{Ca}^{2+}$  and carbon atoms interactions (carbon atoms from the third concentric shell) as well as the angular distribution between hydrogen, oxygen, and  $\text{Ca}^{2+}$ , which both contribute to decreasing the value of the annotated charge (**Fig. 5c**). On the other hand, the features that contribute the most to the charges below an average value (**Fig. 5(b, d)**) are dominated by diverse features derived from the topological order parameters. These selections of the topological order parameters that contribute to decreasing the calcium atomic charge states, however, are quite diverse. These analyses signifying that it's the overall shape of the calcium-binding loop, rather than the exact residues, determining the value of calcium charges.



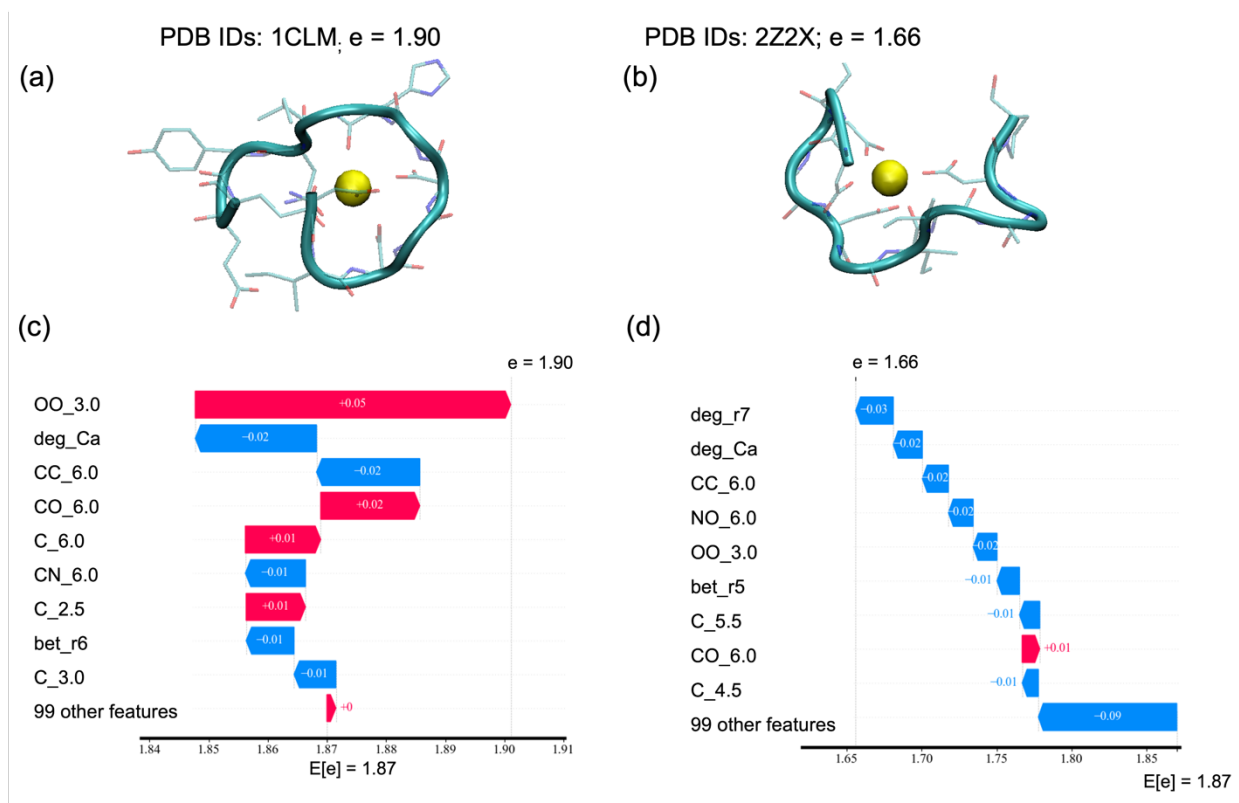
**Figure 5. Local contribution of the most importance features on the charge state of a single conformation.** (a, c) represent the structures whose calcium charges are above the averaged charge, (b, d) represent the structures whose calcium charges are below the averaged charge.

### Using the explainable chemistry-informed ML Model III to annotate the calcium charge states for the experimentally derived calcium-binding motifs

Next, we used the chemistry-informed ML Model III to annotate the calcium charge states of the experimentally determined calcium-binding loops. We extracted 475 calcium-binding motifs from 217 protein structures downloaded from the Protein Data Bank (see *Supplementary Table 1* and *Supplementary Fig. 4*). Among these structures, five were solved using the nuclear magnetic resonance experimental technique and 212 were solved using the X-ray crystallography

experimental technique. Each extracted motif has 12 residues plus  $\text{Ca}^{2+}$  (see *Supplementary information* on how to extract EF-hand motifs from the Protein Data Bank structures) to match the data format of conformations in the *Sandbox* (<https://github.com/Cheung-group/caXML>). We provided the annotated charges of all the experimentally derived structures as *Supplementary Table 1* with the associated calcium motif structures.

We next evaluated the importance features that influence individual atomic charge state annotations. **Fig. 6(a, b)** show the cartoon representations of the calcium-binding loops annotated with an atomic charge above averaged value (**Fig. 6a**) and another one below the average value (**Fig. 6b**). (see *Supplementary Table 1* for the rest of the annotated calcium atomic charges). For the conformation in **Fig. 6a**, the geometry of the calcium binding loop is holo-spheric and its charge is close to 2.00 e. We used SHAP to access the most importance features influencing the charge in **Fig. 6c**. We observed the same trend in annotating individual calcium atomic charge state in **Fig. 5**, where the most influential features in a calcium binding are the symmetry functions capturing a holo-directed geometry. We analyzed the SHAP values for the other calcium binding loop whose calcium charge states is far below average in **Fig. 6d**. The shape of the calcium binding loop follows a planar geometry. The most influential features contributing to decrease of charges in this case are mostly topological order parameters. They exemplify the use of the geometry of a calcium binding loop to annotate its calcium charge state.



**Figure 6. Impact of input features on individual calcium atomic charge state annotation of the experimentally derived calcium-binding loops varies from charge to another.** Cartoon representations of the calcium-binding loops showing an annotated charge state (a) above or (b) below the average value. (c and d) The most influential features for representation in (a and b).

### III. Discussion

#### Calcium-coordination number dictates the geometrical arrangement of the calcium-binding loop.

Within calcium-binding loops, certain residues exhibit remarkable evolutionary conservation across diverse species. These conserved residues, typically found at positions one, three, five, seven, and twelve, play pivotal roles in coordinating  $\text{Ca}^{2+}$  effectively<sup>21,43,44</sup> (**Fig. 2(a, c)**). Their preservation highlights their indispensable contribution to the structural integrity and functionality of calcium-dependent proteins. At position one, acidic residues such as glutamate

(Glu) or aspartate (Asp) are prevalent, providing oxygen atoms that form strong electrostatic interactions with  $\text{Ca}^{2+}$ . Similarly, positions three, five, and seven often host residues with oxygen-containing functional groups, like Asp or serine (Ser), which enhance calcium coordination through hydrogen bonding.<sup>43</sup> Position twelve specifically features Glu with oxygen-rich side chain.<sup>18,43</sup> This residue at the twelfth position with a bidentate interaction contributes additional coordination sites, ensuring tight binding and specificity of the calcium-binding complex. These canonically conserved interactions are captured in the proposed chemistry-informed ML using the symmetry function features. The calcium binding loops coordinated by the five conserved residues adopt holo-spheric geometries (**Figs. 1 and 3**). From the cartoon representation shown in **Fig. 6a**,  $\text{Ca}^{2+}$  is coordinated by Asp, Asp, Asp, His, and Glu at positions one, three, five, seven, and twelve respectively, which explains the holo-directed representation of such structure.

In contrast to the conserved core residues, calcium-binding loops often exhibit variability in other positions, where non-conserved amino acids are prevalent. These non-conserved residues introduce diversity into the binding loop, potentially modulating calcium binding specificity, affinity, or regulatory properties. Non-conserved amino acids may confer adaptability to environmental or physiological changes, allowing calcium-dependent proteins to adjust their function accordingly. Their presence could influence binding kinetics, coordination geometry, and overall structural dynamics of the binding loop, contributing to the versatility and functional diversity of calcium-binding proteins.<sup>21</sup> These non-conserved amino acids interact with  $\text{Ca}^{2+}$  through the high order interactions, which are captured in the explainable chemistry-informed ML model using the topological graph features. In such representations, calcium binding loops are coordinated by at most three of the conserved residues. They may adopt either hemi-spheric or planar geometries (**Figs. 1 and 3**) depending on the number of coordinating ligands. From the

cartoon representation shown in **Fig. 6b**,  $\text{Ca}^{2+}$  is coordinated by Asp at positions one and three but is missing the conserved residue at positions five, seven, as well as Glu at the twelfth position. Therefore, the structure adopts either a hemi-directed or a planar geometry (**Fig. 1**). *The balance between evolutionary conservation and functional diversification not only maintains the essential functions of calcium-dependent proteins but also provides the flexibility necessary for adaptive responses to cellular cues and environmental stimuli.*

The relative mobility of the calcium reverses the intuitive physical effect of the chelating residues. In addition to the effect of mobile sidechains on the calcium charge, the loop backbone also presents a strong effect. For example, the oxygen atom from the ninth residue affects calcium through a mobile water molecule. The mobility of the water molecules contribute variations to the calcium local environment.. Similarly, the high mobility of water renders its contact with calcium almost invariant, so it cannot affect the charge. Residues at positions three and twelve, with the highest probability of contact (**Fig. 2b**), hold the calcium between them, and those residues determine the local environment of the calcium.

Since the twelfth residue in the loop, Glu, contributes to selectivity and specificity of calcium binding in proteins with EF-hand motifs, extensive studies have been conducted to understand its involvement in the structural stability of the EF-hand motif as well as the calcium charge state.<sup>1,43,45</sup> However, unlike the conserved twelfth residue, the tenth non-conserved residue has received less attention. The explainable chemistry-informed ML model showed that the tenth residue plays a topological role in maintaining the motif stability due to its position in the loop (**Fig. 3b**). Because this tenth position can be occupied by any other residue, mostly Phe, Val, Glu, Ala, Leu, Tyr, or Lys such as shown by Marsden et al.,<sup>43</sup> the contribution of this residue to the calcium atomic charge is determined by its topological importance rather than its chemistry.



### **Topological network order parameters reveal hidden geometrical arrangement of residues in calcium-binding loops.**

It is advantageous to use network theory order parameter over atomic contacts or radial/angular distribution (i.e. symmetry functions) for capturing the hierarchical patterns in a complex data. In particular, betweenness centrality shows its importance as a critical bridge in connecting different nodes, while clustering coefficient indicates how tightly connected its immediate neighbors are. Topological network parameters play a critical role in distinguishing the calcium-binding loops from holo-spheric, planar or hemi-spheric geometries, such as evidenced in **Figs. 4c, 5(b, d), and 6d**. These topological features signified the importance of the connectivity among  $\text{Ca}^{2+}$  and the surrounding atoms, the closeness centrality of the  $\text{Ca}^{2+}$ , and the betweenness centrality of some of the non-conserved residues such as in the tenth position. With explainable features, we reveal the role of non-evolutionarily conserved residues in tuning the shape of a calcium-binding loop. The tenth residue is not directly involved in the coordination of  $\text{Ca}^{2+}$  (**Fig. 2a**).<sup>18</sup> It is, however, critical in determining the geometrical arrangement of residues through the shape of the loop that can be modulated by CaM's interaction with its binding target.<sup>11</sup>

### **SHAP explains the significance of additive features influential to the calcium atomic charge state.**

The explainable chemistry-informed ML represents a promising alternative to tackle the challenges presented by the *ab initio* representation of electronic structure data. One of the advantages of the chemistry-informed ML model over data-driven ML models (such as deep neural network) is that the features are tightly related to physical properties and thus are readily

interpretable.<sup>33,46,47</sup> In the case of calcium atomic charge state, the algorithm learns from *ab initio* calculations data and the disordered fragments from the molecular dynamic simulations that contain information about  $\text{Ca}^{2+}$  interactions within the different calcium-binding loops in the lobes of CaM and CaM/target complexes (**Fig. 2**). Through the recognition of patterns and relationships in the data, the model can be constructed to annotate the charge states of  $\text{Ca}^{2+}$  based on features such as coordination chemistry and the different geometries of the calcium-binding loops accounted by the topological network parameters (**Fig. 1**).

In this ever-expanding landscape of ML, the interpretability of models remains a paramount concern.<sup>48,49</sup> In this context, the application of SHAP<sup>34</sup> emerges as a transformative tool for shedding light on the opaque inner workings of the algorithm (see *Supplementary* for further details). By decomposing complex predictions into understandable components, SHAP unravels the intricate relationships between additive input features and model outputs. Adding features from high-order interactions (topological network features) and from the symmetry function (radial/angular distribution functions) allows us to capture the loosely bound calcium geometry (planar or hemi-spheric) and the tightly bound calcium geometry (holo-spheric),

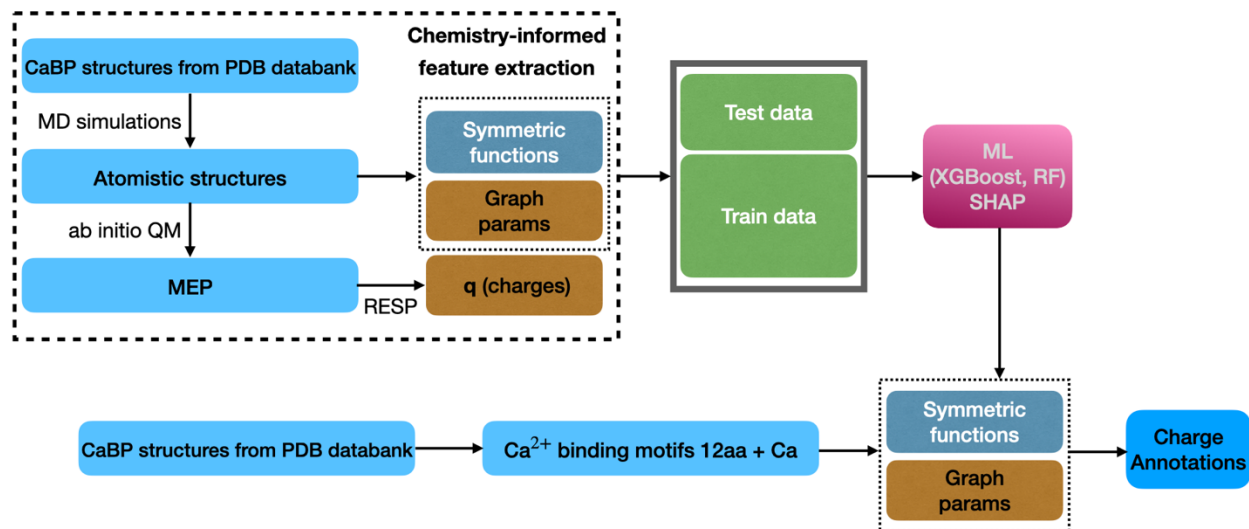
One of the most compelling aspects of SHAP lies in its ability to provide both global and local explanations<sup>34,35</sup> to understand how the combination of these features leads to an accurate calcium atomic charge state annotation. Global insights offer a holistic view of feature importance across an entire dataset (**Fig. 4(a-c)**), while local interpretations elucidate the rationale behind individual charge predictions (**Figs. 5(a-d)** and **6(c, d)**). This dual perspective enhances trust in the construction of ML model for charge annotation.

#### **IV. Conclusion**

The integration of chemistry-informed ML techniques explains the significant additive features in constructing models to annotate the atomic charge states of calcium ions in calcium-binding proteins, despite a small dataset in the genome or materials space. Combining the domain knowledge of chemistry, explainable ML, our approach identifies the explainable molecular descriptors in the fuzzy shape of a calcium binding loop that influence the calcium atomic charge state. It unveils the intricate patterns and correlations in additive features in a structurally complex dataset. With transparent machine learning models, subtle variations in the coordination chemistry of calcium ion in a calcium binding loop intricately modulate its charge states. This interdisciplinary approach not only enhances our comprehension of fundamental biological processes, but also offers invaluable insights into the design and engineering of metalloprotein-based systems with a limited genome space.

## V. Method

The procedures of the explainable ML models are shown in the flowchart (**Fig. 7**), including the molecular dynamics data preparation, *ab initio* quantum-mechanics-based label generation, chemistry-informed feature extraction, experimentally derived calcium motifs data preparation, and construction and analyses of the ML model.



**Figure 7.** Flowchart of the chemistry-informed feature extraction and ML interpretation. The algorithm was trained using data from the combined *ab initio* electronic calculations and the structures of the calcium-binding loops from molecular dynamics simulations. We subsequently used one of the constructed ML models to annotate the calcium charge states from the calcium binding motifs extracted from the experimentally derived structures from the Protein Data Bank (Supplementary **Table 1** and Supplementary **Fig. 4** in the *Supplementary*).

### **Data preparation: calcium-binding environments extraction and *ab initio* calcium charge determination.**

We consider a dataset consisting of selected calcium-binding loop structures with bound  $\text{Ca}^{2+}$  ions and nearby water molecules from molecular dynamics simulations. Simulations of  $\text{Ca}^{2+}/\text{CaM}$  in explicit water in the following three conditions were examined: (i) Neat  $\text{Ca}^{2+}/\text{CaM}$ ; (ii)  $\text{Ca}^{2+}$ -retaining environment:  $\text{Ca}^{2+}/\text{CaM}/\text{CaMKII}$  peptide; and (iii)  $\text{Ca}^{2+}$ -releasing environment:  $\text{Ca}^{2+}/\text{CaM}/\text{Ng}$  peptide. Please refer to the *Supplementary* for details of the molecular dynamics simulations. In total, 8,000 snapshots of  $\text{Ca}^{2+}$  in EF-hand loops were selected for further analyses. Atomic charges were determined by fitting to the electrostatic potential obtained from single-point

quantum mechanical calculations of the calcium-binding loop configurations. Detailed workflow of the charge determination is described in the *Supplementary* and in our previous study.<sup>14</sup>

### **Topological parameters in the calcium-coordination network highlight many-body interactions at the amino-acid level.**

The topology of the calcium-coordination network was described analytically in the form of graph theory. The nodes are formed by an amino-acid,  $\text{Ca}^{2+}$ , or the closest water molecule. An edge is assigned when the distance between the closest pair of atoms of C, N, or O in each residue. For each residue in a loop, an undirected edge exists between the node representing it and its contacting residue with a weight equal to the reciprocal distance of the contact. Analysis of the graphs formed under this definition is accomplished in Gephi<sup>50</sup> and inherits its definitions of graph theoretic measurements.

Degree centrality,  $C_d(v)$ , quantifies how well the node connects with the rest of the graph. It is calculated as the degree or the connection number ( $k$ ) of a node  $v$  divided by the number of other nodes in the graph ( $n-1$ ), such as shown in Equation (1)

$$C_d(v) = \frac{k}{n-1} \quad (1)$$

Betweenness centrality,  $C_B(v)$ , is used to describe how pivotal a node is to bridge the rest of the graph. It is calculated as the fraction that the node shortcuts any two other nodes in the graph. The mathematical definition of the betweenness centrality  $C_B(v)$  of the node  $v$  is defined in Equation (2)

$$C_B(v) = \sum_{\substack{s \in V \\ s \neq v}} \sum_{\substack{t \in V \\ t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

with  $\sigma_{st}$  representing the number of shortest paths from node  $s$  to node  $t$ , and  $\sigma_{st}(v)$  representing the number of those paths which go through the node  $v$ . The sum is over all pairs of nodes in the graph.  $V$  is the set of nodes.

Clustering coefficient,  $CC(v)$ , quantifies how well the neighbors of a node are interconnected. By definition, the clustering coefficient of a node  $v$ , is the fraction of edges to possible edges between its neighbors for the node  $v$ , as defined in Equation 3

$$CC(v) = \frac{T(v)}{\binom{k}{2}} \quad (3)$$

where  $T(v)$  is the number of edges between the neighbors of the node  $v$  and  $k$  is the cardinality of the neighborhood (degree) of the node  $v$ .

Closeness centrality,  $C_c(v)$ , quantifies the average distance between the node and the rest of the graph. Mathematically, the definition is given in Equation 4

$$C_c(v) = \frac{n - 1}{\sum_{u \in V, u \neq v} d(u, v)} \quad (4)$$

where  $d(u, v)$  is the shortest path distance between nodes  $u$  and  $v$ .

### **Construction of predictive ML algorithms.**

We used a tree model with gradient boosting, XGBoost<sup>42</sup> and Sci-Kit Learn Python package,<sup>51</sup> to train and test the *ab initio* charge estimator models. We split the data into training (70%) and testing (30%) sets. Quality of the data splitting was justified by the overlap between distributions of the representative features and the label generated from the training/testing data (Supplementary **Figs. 2** and **3**). We used fivefold cross-validation by searching exhaustively on a hyperparameters grid on high-performance computers to determine the optimized hyperparameters. Specifically, we tuned  $n\_estimators$  (number of gradient boosted decision trees),

*colsample\_bytree* (the proportion of features for constructing each tree), *subsample* (the portion of training data samples for training each additional tree), and *learning\_rate* (shrinkage of feature weights to make the boosting process more conservative and prevent overfitting, also called " $\eta$ "). Subsequently, we made predictions on the calcium atomic charges using the optimized XGBoost regression model, assessed the model accuracy by calculating the coefficient of determination, Pearson correlation coefficient, and the MAE between the predicted and *ab initio*-derived reference values.

## Acknowledgment

MSC and PZ thank the support from the National Institutes of Health (2R01GM097553). JN thanks the support from the National Science Foundation (MCB 2221824). We are grateful for the computing resources from XSEDE (TG-MCB190109), and the resources from the Research Computing Data Core at the University of Houston.

## Author contributions

M.S.C., P.Z., Y.E., P.C. designed the research. P.Z., J.N., N.J. performed the research. M.S.C., P.Z., J.N., Y.E., P.C. analyzed data. All authors contributed to writing the paper.

## References

- 1 Carafoli, E. & Krebs, J. Why Calcium? How Calcium Became the Best Communicator. *J Biol Chem* **291**, 20849-20857 (2016). <https://doi.org:10.1074/jbc.R116.735894>
- 2 Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodelling. *Nature Reviews Molecular Cell Biology* **4**, 517-529 (2003). <https://doi.org:10.1038/nrm1155>

- 3 Fan, C. *et al.* Calcium-gated potassium channel blockade via membrane-facing fenestrations. *Nature Chemical Biology* **20**, 52-61 (2024). <https://doi.org:10.1038/s41589-023-01406-2>
- 4 Berridge, M. J., Lipp, P. & Bootman, M. D. The versatility and universality of calcium signalling. *Nature Reviews Molecular Cell Biology* **1**, 11-21 (2000). <https://doi.org:10.1038/35036035>
- 5 Carafoli, E. Calcium signaling: A tale for all seasons. *Proceedings of the National Academy of Sciences* **99**, 1115-1122 (2002). <https://doi.org:10.1073/pnas.032427999>
- 6 Zhang, P., Han, J., Cieplak, P. & Cheung, M. S. Determining the atomic charge of calcium ion requires the information of its coordination geometry in an EF-hand motif. *J. Chem. Phys.* **154**, 124104 (2021).
- 7 Clapham, D. E. Calcium Signaling. *Cell* **131**, 1047-1058 (2007). <https://doi.org:10.1016/j.cell.2007.11.028>
- 8 Bootman, M. D. Calcium signaling. *Cold Spring Harb Perspect Biol* **4**, a011171 (2012). <https://doi.org:10.1101/cshperspect.a011171>
- 9 Moustakas, M. The Role of Metal Ions in Biology, Biochemistry and Medicine. *Materials (Basel)* **14** (2021). <https://doi.org:10.3390/ma14030549>
- 10 Zheng, X., Cheng, W., Ji, C., Zhang, J. & Yin, M. Detection of metal ions in biological systems: A review. *Reviews in Analytical Chemistry* **39**, 231-246 (2020). <https://doi.org:doi:10.1515/revac-2020-0118>
- 11 Bayer, K. U. & Schulman, H. CaM Kinase: Still Inspiring at 40. *Neuron* **103**, 380-394 (2019). <https://doi.org:https://doi.org/10.1016/j.neuron.2019.05.033>
- 12 Nde, J., Zhang, P., Waxham, M. N. & Cheung, M. S. Experiment and Simulation Reveal Residue Details for How Target Binding Tunes Calmodulin's Calcium-Binding Properties. *The Journal of Physical Chemistry B* **127**, 2900-2908 (2023). <https://doi.org:10.1021/acs.jpcc.2c08734>
- 13 Nde, J. *et al.* Coarse-Grained Modeling and Molecular Dynamics Simulations of Ca<sup>2+</sup>-Calmodulin. *Frontiers in Molecular Biosciences* **8** (2021). <https://doi.org:10.3389/fmolb.2021.661322>
- 14 Zhang, P., Han, J., Cieplak, P. & Cheung, M. S. Determining the atomic charge of calcium ion requires the information of its coordination geometry in an EF-hand motif. *J Chem Phys* **154**, 124104 (2021). <https://doi.org:10.1063/5.0037517>
- 15 Bootman, M. D. & Bultynck, G. Fundamentals of Cellular Calcium Signaling: A Primer. *Cold Spring Harb Perspect Biol* **12** (2020). <https://doi.org:10.1101/cshperspect.a038802>
- 16 Denessiouk, K., Permyakov, S., Denesyuk, A., Permyakov, E. & Johnson, M. S. Two Structural Motifs within Canonical EF-Hand Calcium-Binding Domains Identify Five Different Classes of Calcium Buffers and Sensors. *PLOS ONE* **9**, e109287 (2014). <https://doi.org:10.1371/journal.pone.0109287>
- 17 Chazin, W. J. Relating form and function of EF-hand calcium binding proteins. *Acc Chem Res* **44**, 171-179 (2011). <https://doi.org:10.1021/ar100110d>
- 18 Lewit-Bentley, A. & Réty, S. EF-hand calcium-binding proteins. *Curr Opin Struct Biol* **10**, 637-643 (2000). [https://doi.org:10.1016/s0959-440x\(00\)00142-1](https://doi.org:10.1016/s0959-440x(00)00142-1)
- 19 Prajapati, J. D., Mele, C., Aksoyoglu, M. A., Winterhalter, M. & Kleinekathöfer, U. Computational Modeling of Ion Transport in Bulk and through a Nanopore Using the Drude Polarizable Force Field. *Journal of Chemical Information and Modeling* **60**, 3188-3203 (2020). <https://doi.org:10.1021/acs.jcim.0c00389>



- 20 Waxham, M. N. & Cheung, M. S. in *Encyclopedia of Computational Neuroscience* (eds D. Jaeger & R. Jung) (Springer, 2013).
- 21 Wang, X., Kirberger, M., Qiu, F., Chen, G. & Yang, J. J. Towards predicting Ca<sup>2+</sup>-binding sites with different coordination numbers in proteins with atomic resolution. *Proteins* **75**, 787-798 (2009). <https://doi.org/10.1002/prot.22285>
- 22 Cates, M. S., Teodoro, M. L. & Phillips, G. N. Molecular Mechanisms of Calcium and Magnesium Binding to Parvalbumin. *Biophysical Journal* **82**, 1133-1146 (2002). [https://doi.org/10.1016/S0006-3495\(02\)75472-6](https://doi.org/10.1016/S0006-3495(02)75472-6)
- 23 Tan, Q., Ding, Y., Qiu, Z. & Huang, J. Binding Energy and Free Energy of Calcium Ion to Calmodulin EF-Hands with the Drude Polarizable Force Field. *ACS Phys Chem Au* **2**, 143-155 (2022). <https://doi.org/10.1021/acspchemau.1c00039>
- 24 Noe, F., Tkatchenko, A., Muller, K. R. & Clementi, C. Machine Learning for Molecular Simulation. *Annu Rev Phys Chem* **71**, 361-390 (2020). <https://doi.org/10.1146/annurev-physchem-042018-052331>
- 25 Deringer, V. L., Caro, M. A. & Csanyi, G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat Commun* **11**, 5461 (2020). <https://doi.org/10.1038/s41467-020-19168-z>
- 26 Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463-477 (2019). <https://doi.org/10.1038/s41573-019-0024-5>
- 27 Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* **13**, 8-17 (2015). <https://doi.org/10.1016/j.csbj.2014.11.005>
- 28 Shi, Y.-F. *et al.* Machine Learning for Chemistry: Basics and Applications. *Engineering* (2023). <https://doi.org/10.1016/j.eng.2023.04.013>
- 29 Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *The Journal of Physical Chemistry Letters* **11**, 9656-9658 (2020). <https://doi.org/10.1021/acs.jpcllett.0c03130>
- 30 Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **6**, 161-169 (2024). <https://doi.org/10.1038/s42256-023-00788-1>
- 31 Case, M., Smith, M., Vinh, J. & Thurber, G. Machine learning to predict continuous protein properties from binary cell sorting data and map unseen sequence space. *Proceedings of the National Academy of Sciences* **121**, e2311726121 (2024). <https://doi.org/10.1073/pnas.2311726121>
- 32 Jung, H. *et al.* Machine-guided path sampling to discover mechanisms of molecular self-organization. *Nature Computational Science* **3**, 334-345 (2023). <https://doi.org/10.1038/s43588-023-00428-z>
- 33 Zhang, B. *et al.* Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning. *Proceedings of the National Academy of Sciences* **119**, e2212711119 (2022). <https://doi.org/10.1073/pnas.2212711119>
- 34 Lundberg, S. M. & Lee, S.-I. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Long Beach, California, USA, 2017).
- 35 Lundberg, S. A., Gabriel%AChen, Hugh%ADeGrave, Alex%APrutkin, Jordan%ANair, Bala%AKatz, Ronit%AHimmelfarb, Jonathan%ABansal, Nisha%ALee, Su-In%BJournal Name:

- Nature machine intelligence & 2, J. V. From local explanations to global understanding with explainable AI for trees. *Journal Name: Nature machine intelligence; Journal Volume: 2, Medium: X* (2020).
- 36 Paknejad, N., Sapuru, V. & Hite, R. K. Structural titration reveals Ca<sup>2+</sup>-dependent conformational landscape of the IP3 receptor. *Nature Communications* **14**, 6897 (2023). <https://doi.org:10.1038/s41467-023-42707-3>
- 37 Zhao, K. *et al.* Predicting Ca<sup>2+</sup> -binding sites using refined carbon clusters. *Proteins* **80**, 2666-2679 (2012). <https://doi.org:10.1002/prot.24149>
- 38 Shapley, L. S. & CALIF., R. C. S. M. *A Value for N-person Games*. (Defense Technical Information Center).
- 39 Wang, X. *et al.* Analysis and prediction of calcium-binding pockets from apo-protein structures exhibiting calcium-induced localized conformational changes. *Protein Sci* **19**, 1180-1190 (2010). <https://doi.org:10.1002/pro.394>
- 40 Kirberger, M. *et al.* Statistical analysis of structural characteristics of protein Ca<sup>2+</sup>-binding sites. *J Biol Inorg Chem* **13**, 1169-1181 (2008). <https://doi.org:10.1007/s00775-008-0402-7>
- 41 Porter, M. A., Feng, M. & Katifori, E. The topology of data. *Physics Today* **76**, 36-42 (2023). <https://doi.org:10.1063/pt.3.5157>
- 42 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).
- 43 Marsden, B. J., Shaw, G. S. & Sykes, B. D. Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. *Biochem Cell Biol* **68**, 587-601 (1990). <https://doi.org:10.1139/o90-084>
- 44 Snyder, E. E., Buoscio, B. W. & Falke, J. J. Calcium(II) site specificity: effect of size and charge on metal ion binding to an EF-hand-like site. *Biochemistry* **29**, 3937-3943 (1990). <https://doi.org:10.1021/bi00468a021>
- 45 Schwaller, B. Cytosolic Ca<sup>2+</sup> buffers. *Cold Spring Harb Perspect Biol* **2**, a004051 (2010). <https://doi.org:10.1101/cshperspect.a004051>
- 46 Bradford, G. *et al.* Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery. *ACS Central Science* **9**, 206-216 (2023). <https://doi.org:10.1021/acscentsci.2c01123>
- 47 Zhang, L. V., Marani, A. & Nehdi, M. L. Chemistry-informed machine learning prediction of compressive strength for alkali-activated materials. *Construction and Building Materials* **316**, 126103 (2022). <https://doi.org:https://doi.org/10.1016/j.conbuildmat.2021.126103>
- 48 Chen, H., Lundberg, S. M. & Lee, S.-I. Explaining a series of models by propagating Shapley values. *Nature Communications* **13**, 4512 (2022). <https://doi.org:10.1038/s41467-022-31384-3>
- 49 Aas, K., Jullum, M. & Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* **298**, 103502 (2021). <https://doi.org:https://doi.org/10.1016/j.artint.2021.103502>
- 50 Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. (2009).
- 51 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).