Machine-learning prediction of protein function from the portrait of its intramolecular electric field

Santiago Vargas [a], Shobhit S. Chaturvedi [a], Anastassia N. Alexandrova [a,*]
[a] Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States.
* Corresponding author email: ana@chem.ucla.edu

Abstract
We introduce a machine learning framework designed to predict enzyme functionality directly from the heterogeneous electric fields inherent to protein active sites. We apply this method to a curated dataset of Heme-Iron Oxidoreductases, spanning three enzyme classes: monooxygenases, peroxidases, and catalases. Conventional analysis, focused on simplistic, point electric fields along the Fe-O bond, are shown to be inadequate for accurate activity prediction. Our model demonstrates that the enzyme's heterogenous 3-D electric field, alone, can accurately predict its function, without relying on additional protein-specific information. Through feature selection, we uncover key electric field components that not only validate previous studies but also underscore the crucial role of multiple components beyond the traditionally emphasized electric field along the Fe-O bond in heme enzymes. Further, by integrating protein dynamics, principal component analysis, clustering, and QM/MM calculations, we reveal that while dynamic complexities in protein structures can complicate predictions, accounting for this increased dynamic variability can substantially enhance model performance. This research significantly advances our understanding of how protein scaffolds possess signature electric fields that are tailored to their functions at the active site. Moreover, it presents a novel electrostatics-based tool to harness these signature electric fields for predicting enzyme function.

Introduction
Warshel's groundbreaking analysis on enzymology asserts that catalysis in enzymes is partly governed by the charge distribution within the protein.[1,2] This concept, electrostatic preorganization, asserts that protein charges imbue an electrostatic potential favorable to reaction transition states (TS) over reactants and, thus, catalyzes a reaction. Despite the fact that electrostatic potentials are largely local, it was postulated that the large, complex structure of proteins serves the dual purpose of positioning neighboring atoms (charges) optimally while shielding the active site from the environment. One way to quantify the electrostatic preorganization is by analyzing the electric fields generated by enzymes.[3]

The notion that electric fields can act as catalytic components deviates from the framework that catalysts must be purely chemical. Numerous studies have demonstrated that electric fields significantly influence both chemical reactivity and selectivity across a wide range of proteins, both metal-free and metal-dependent.[4–18] Ketosteroid isomerases (KSI) became a key example demonstrating this effect through several *in silico* studies.[19–21] Following these numerous computational demonstrations, Boxer provided experimental validation by showing

that in ketosteroid isomerases, the electric field acting on the charged enolate intermediate correlated with the reaction's free energy barrier.[10] Subsequently, the quantum theory of atoms in molecules (QTAIM) was employed to examine how electric fields impact the reactivity of KSI, where it revealed that fields manipulate electron density throughout the substrate. [16,22,23]

In the realm of metal-containing enzymes, significant attention has been devoted to exploring electric fields within Fe-heme containing enzymes and their model systems.[24–27] Even with identical Fe-heme coordination, mere variation in axial ligands such as cysteine, histidine, and tyrosine, heme enzymes exhibit diverse reactivity. Our prior research unearthed a pivotal revelation: beyond the axial ligand, the electric field from the surrounding protein (excluding the heme and the axial ligand) strongly influences reactivity.[24] This underscores the heme scaffold's role as a molecular capacitor, where specific configurations of charged amino acids generate a characteristic electric field along the Fe(IV)=O bond in Compound I (F_z). Notably, we predicted that a heme equipped with the suitable axial ligand for its intended function yet situated within a protein environment typical of a different class of oxidoreductases may acquire an unintended function, such as off-pathway oxidation.[24] In a recent study of laboratory evolved protoglobin for the catalysis of carbene transfer reactions, we furthermore showed that it is the catalytic component of the electric field in the active site that the evolution develops in its course.[26] We infer the impact of fields within protein active sites on chemical reactivity, and thus offer another avenue in the pursuit of effective protein design.[28] Such insights could bridge the existing gap between computationally designed proteins and genuinely effective enzymes, whether naturally occurring or laboratory evolved.

Since electric fields are so prominent in governing enzyme reactivity, here we flip the problem and explore whether machine learning can predict enzyme reactivity solely based on the electric field of the protein scaffold. For this purpose, we use the previously reported dataset of ~200 Hemoglobin proteins[24] and, with the electric field as a sole input, classify these proteins as catalases, peroxidases, or monooxygenases (**Figure 1**). In other words, we train a ML model that would predict the heme Fe reactivity strictly from the heterogeneous field that the protein produces. Indeed, the task is analogous to a classic image recognition problem where spatial field components act as pixel components for machine learning algorithms.

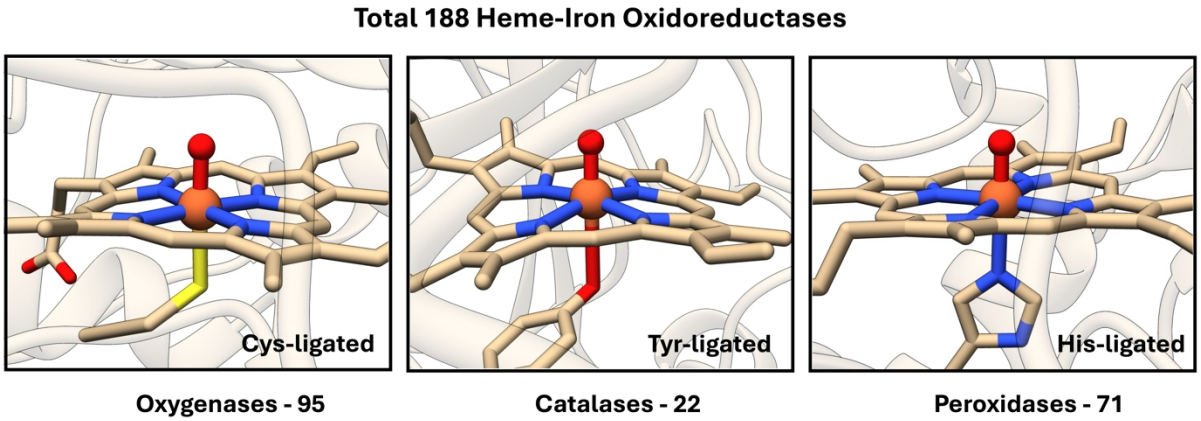

Figure 1. The dataset includes three classes of hemes: oxygenases, catalases, and peroxidases, each with distinct axial ligands. The total number of examples for each class is indicated on the figure, highlighting the representation of each class within the dataset.

Methods

Despite the broad success of theoretical analyses of electrostatic preorganization, they often have two shortcomings: firstly, they lack dynamic information, in the sense of the dynamics of the field itself. Naturally, the structural dynamics of the protein is included via molecular dynamics (MD) simulations and subsequent averaging of computed properties, such as reaction barriers and electric fields. Some exceptions exist; for example, the effects of KSI conformational changes on the electric field have been tracked to explain transition state stability.[29] Secondly, analysis is generally reduced to a field at a single point in an enzyme. The reason that the single point analysis is incomplete is that, for many systems, the reaction mechanism is not localized to a single bond. For example, the ubiquitous Diels-Alder reaction as an example where reactivity is delocalized across a number of atoms and bonds. Recently, a second dimension was added to field analysis, mitigating the problem to an extent.[9,18] Here, we analyze the field in the active site in its entirety, considering also field dynamics, and then use the fundamental components of the field from dimensional reduction and machine learning, for protein function recognition.

The issue of ingesting raw heterogeneous electric fields is dimensionally daunting - a coarse sampling of electric field values can lead to tens of thousands of input dimensions as each spatial point is associated with three directional components. This scaling leads to an intractable problem for manual analysis where we simply cannot separate signal from noise in such a high dimensional space. In addition, even statistical/machine-learned (ML) methods can struggle to find meaningful descriptors without a large enough dataset for either supervised or unsupervised machine learning tasks. We address this by using dimensionality reduction, via principal component analysis (PCA), to create a more manageable, data-informed set of input dimensions. PCA is often used as a preprocessing step before supervised machine learning tasks to reduce noise in data and simplify learning tasks. For our use case, PCA was highly attractive as it is a data-first representation scheme where prior knowledge of a system is not necessary. This establishes our framework as a universal scheme that could be used to study and explain families of proteins where domain knowledge is lacking or where representative fields are simply too complex to construct *a priori*. We envision using this methodology to recognize the functions of active sites of newly discovered proteins, distinguish active sites from areas in proteins that look like active sites but are not, and attributing selectivity to an enzyme without lengthy mechanistic investigations.

System set up and field calculations and analysis

To represent each protein, we take crystal structures from the Protein Data Bank, remove co-crystalized water molecules and ions, and zero the charges on the axial ligands, and the hemoglobin itself. First, we develop and ML algorithms that operate on the point field at the Fe, then – the 3-D field in a volume around the Fe without dynamics, and then extend this study

to include the dynamics and clustering of the field. The fields are computed classically using the point charges of the protein, and thus excluding the Fe(IV)=O moiety, the heme, and the axial ligand. The 3-D fields were constructed on the grid over a cubic box centered at the Fe atom in the CpdI intermediate (**Figure 2**), the box (dimensions: 3 Å x 3 Å x 3 Å) is visualized in **Figure 2**. The grid spacing was 21 sampling points along each dimension for a total of ~9,200 points. In the context of molecular dynamics, the fields are compared to each other using the global distribution of streamlines method.

In detail, our group previously adapted a distance metric from fluid dynamics to study the differences between complex, heterogenous electric fields.[16] This method constructs a global distribution of slipstreams within a vector field, yielding histograms that describe an electric field. The formulation enjoys important mathematical properties such as rotational, scalar, and translational invariance. Here, within the 3 x 3 x 3 Å cube, random points are sampled to create linearizations, known as slipstreams. Random points along a given slipstream are selected to compute mean curvature and distance of a line where curvature is defined as

$$\kappa = \frac{\left\| r'(t) \times r''(t) \right\|}{\left\| r'(t) \right\|^3}$$

A histogram of L2 distance to curvature can thereby be compiled and the distance between two discrete distributions can be computed via the χ^2 distance:

$$\chi^2 : D(f, g) = \frac{1}{2} \sum_{i=1}^{N} \frac{(f[i] - g[i])^2}{f[i] + g[i]}$$

With a defined distance comparing electric fields, we can then create a graph where the edge weights are the distances between two electric fields. This graph encoding is ripe for graph compression algorithms, notably affinity propagation, to aid in the selection of a few representative frames entirely on the basis of the 3-D heterogeneous electric field. Our group has previously used this protocol to interpret the dynamic heterogeneous field differences along a directed evolution pathway of catalytic protoglobin complexes.[26] We used these compressed representations of the electric field along the entire MD trajectory to further study the effects of the 3-D electric field on electronic populations within the active site. With this, we demonstrate the relationship between induced fields at the active site and the overall protein activity.

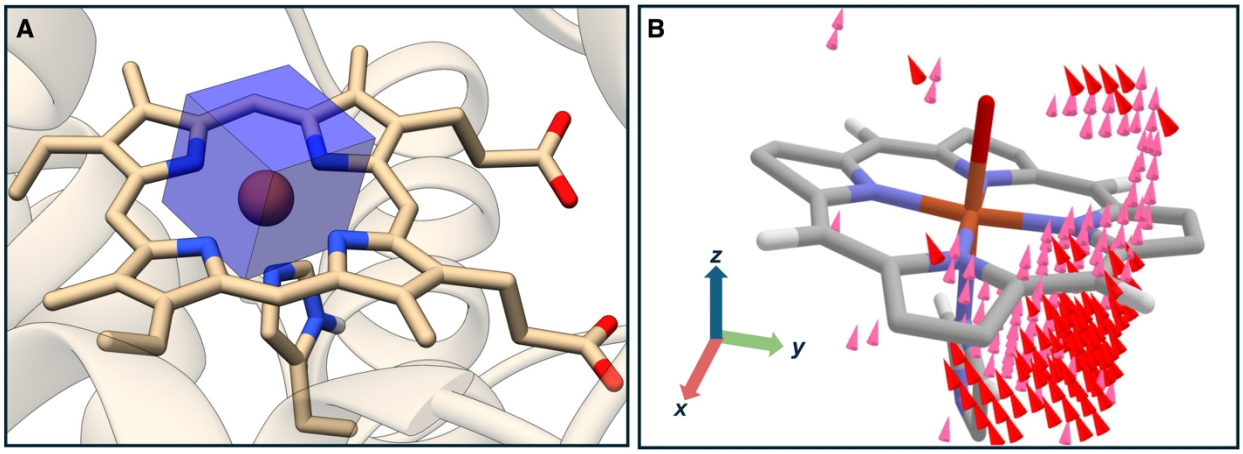

Figure 2. (a) The cubic box centered on Fe, used for computing the electric field on the grid. (b) An example of typical principal component computed on the dataset, plotted on the exponential scale for clarity.

PCA

To determine the proper number of PCA components, we swept the number of PCA components of the electric field in the model from 5 to 25 PCA components and used cross-fold validation to select the optimal number of components. We found that 9 components were optimal for performance on validation data. For validation and testing, we split our dataset into an 80-20 train-test set and used k-folds (k=5) training-validation splits to tune model parameters, PCA components were constructed entirely from the training split to avoid data leakage into the test set.

Molecular Dynamics

We parametrized the Fe-containing heme active site for MD simulation with the Metal Centre Parameter Builder (MCPB.py).[30] We modeled the remainder of the protein using the Amber FF19SB force field.[31] The leap module in AMBER 22 was utilized to introduce Na^+ counterions to neutralize protein systems.[32] These systems were then placed in a rectangular box, surrounded by OPC water molecules[33] extending at least 10 Å beyond the outermost boundary of the protein. We applied periodic boundary conditions throughout the simulations. The particle mesh Ewald method was used to calculate long-range electrostatic interactions, with both the direct space and the van der Waals interactions capped at a 10 Å cutoff. The protein systems was minimizated, initially with 5,000 steps of steepest descent followed by another 5,000 steps using the conjugate gradient method, all under a 100 kcal mol^{-1} $Å^2$ restraint on the solute molecules. This was succeeded by another round of full system minimization employing the same descent and gradient steps. Subsequently, the systems were gradually heated from 0 to 300 K in an NVT ensemble, controlled by a Langevin thermostat with a collision frequency of 1 ps-1 over 250 ps, while the solute molecules were held under a 50 kcal mol-1 Å2 harmonic restraint. Bonds involving hydrogen were constrained by the SHAKE algorithm.[34] Following this, a 1 ns lightly restrained MD simulation was conducted to stabilize the density under periodic boundary conditions. All systems were equilibrated at 300 K for 3 ns in an NPT ensemble, using the Berendsen barostat to maintain pressure at 1 bar, without

restraints. A 100 ns productive MD simulation was carried out for each system in an NPT ensemble, maintaining a constant pressure of 1 bar with a 2 ps pressure coupling, using the GPU-accelerated version of AMBER 22.[32] The trajectories are subjected to field topology calculation (using the CPET code)[16] via embedding the active site in the point charges. The fields were then compared to each other along the trajectory and clustered by the topology similarity.[26]

Quantum Mechanics/Molecular Mechanics (QM/MM) calculations

Quantum mechanics/molecular mechanics (QM/MM) calculations were conducted using the ChemShell[35] software suite, integrating Turbomole[36] for quantum mechanics calculations and DL_POLY[37] for molecular mechanics. For these calculations, water molecules beyond a 10 Å solvation layer surrounding the protein were removed using CPPTRAJ,[38] leaving the protein optimally hydrated. The QM region encompassed the heme iron center, the intermediate oxo or hydroxo groups, and the axial ligand located at the active site, similar to our earlier study.[24] The unrestricted B3LYP functional,[39] as previously shown to be reasonable for these systems,[24] was employed for the QM calculations. The molecular mechanics region was defined as the protein area within 8 Å of the QM zone, while the remaining system components were held static. The Amber FF19SB force field was applied to the molecular mechanics region. Hydrogen link atoms capped the QM/MM boundaries, and a charge shift model was utilized. Electrostatic embedding accounted for the polarizing effects of the protein environment on the QM region. Geometry optimization and frequency analyses utilized the def2-TZVP basis set, with the exception that hydrogens were treated using the def2-SVP basis set. The CpdI Fe(IV)=O ($Por^{+\bullet}$) complex was modelled as a doublet while the CpdII Fe(IV)-OH was modelled as a triplet for all systems.

Results and Discussion

Single point fields. We used a host of traditional machine learning models due to the relatively middling amount of data, including, XGBoost, Random Forests, Ridge Regression, and K-nearest Neighbors (**Figure 2A**). To tackle imbalanced data, present by the underrepresentation of catalases (21 proteins in training vs. roughly triple the number of monooxygenases, peroxidases) — we trained Balanced Random Forests algorithms.[40] For hyperparameter tuning, we employed a 5-fold cross-validation method combined with an 80-20 train-test split for both single point and complete, heterogeneous training. To optimize parameter selection further, we used Bayesian optimization techniques in WanDB.[41] The detailed model parameter dictionaries can be found in the supplementary information.

Model	F1 Score	Accuracy
XGBoost (Single Point, 3-Comp)	0.42	0.44
Balanced Random Forest (3-D Fields, PCA)	0.75	0.82
XGBoost (3-D Fields, PCA)	0.84	0.84

Table 1. Performance of the two top performing ML models benchmarked against the top model to predict on a single point (x,y,z components at the Fe in Heme). This is a proxy for the

previous mapping of F_z at the Fe to axial ligand. Note the dramatic improvement in performance with a richer set of electric field features.

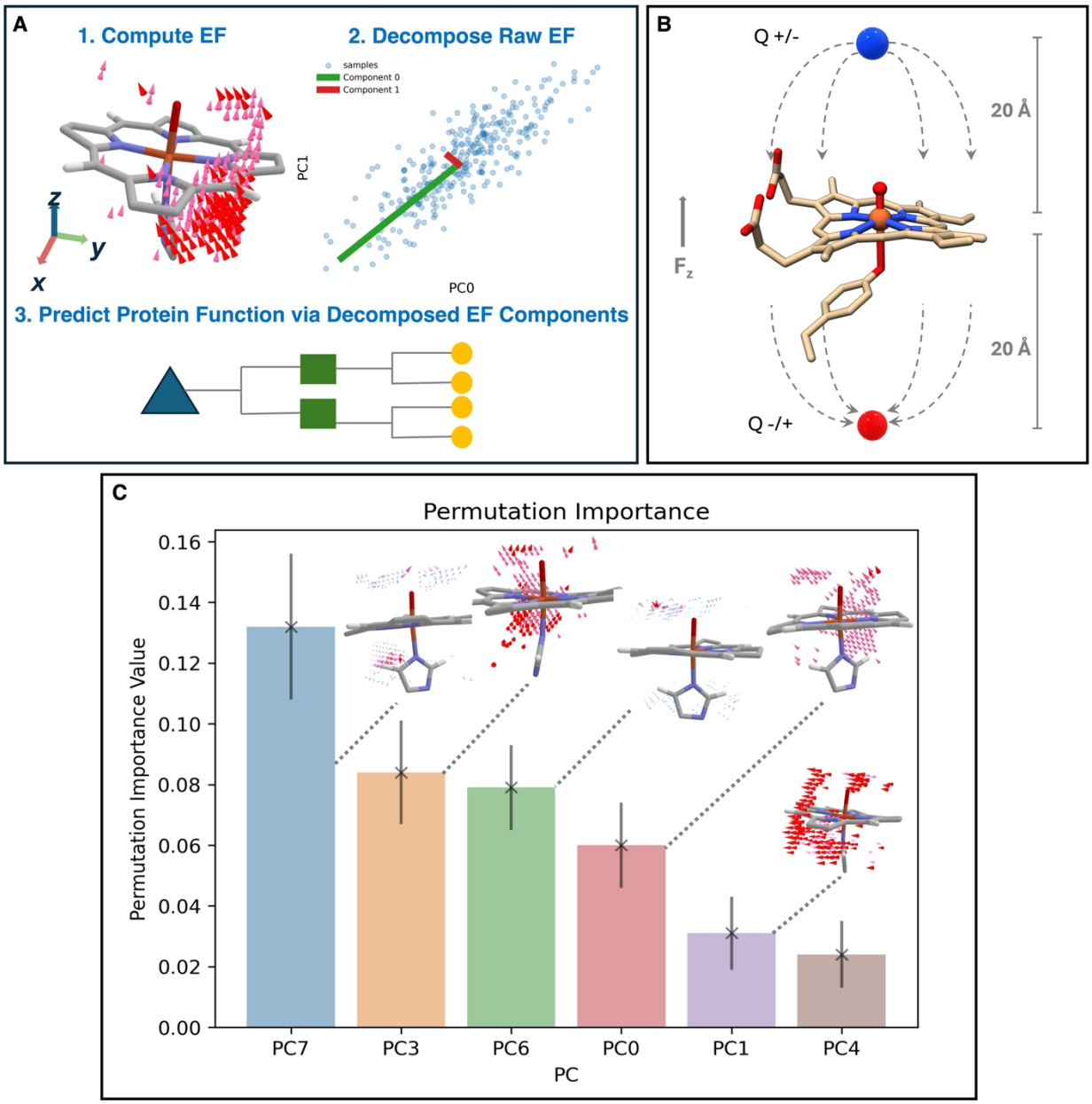

Figure 3. (A) Workflow for predicting protein function using Machine Learning models (B) Surrogate model to test ML machinery with applied fields. (C) Principal components selected by permutation importance and Boruta. Visualized structures (PC7, PC3, PC6, and PC4) were also flagged by Boruta as important.

Performance evaluations were conducted using accuracy and F1-scores, providing a holistic view of model effectiveness (**Table 1**). Considering our dataset's label ratio of roughly 4:3:1, we prioritized the F1-score as a fairer performance metric. All the above-mentioned models were applied to single-point electric field data, with XGBoost emerging as the best performer among them. Focusing on the three components: F_x, F_y, and F_z at the Fe atom,

XGBoost achieved an underwhelming F1-score of 0.42 and an accuracy of 0.44, illustrating the limitations when relying solely on point electric fields for predicting protein functions. The results indicate that while point electric fields offer a straightforward interpretation, they are insufficient for capturing the comprehensive detail required for accurate model predictions.

3-D fields. In stark contrast, incorporating a full 3-D heterogeneous electric field representation, through PCA, significantly enhances model performance, achieving accuracy and F1 scores of up to 84% and 0.84 respectively. This contrasting difference underlines the inadequacy of point electric fields as simplistic, whereas 3-D heterogeneous electric fields offer more representative depictions of the enzymatic environment (**Table 1**). Moreover, the ability of a machine learning model to predict functions from electric field data of a protein scaffold suggests that the scaffold is evolutionarily optimized to provide the specific fields necessary for efficient catalysis.

Applied Uniform Fields. Given a machine-learning model trained on compressed electric field representations, we aim to identify which components from the heterogeneous electric field are critical for the model predictions. For this, we utilized the trained, heterogeneous electric field models to predict changes in predicted heme activity with externally applied fields. We aimed to test a crucial hypothesis: whether the magnitude of the applied F_z electric field is decisive in determining their catalytic function. Specifically, we sought to understand if changes along the F_z direction alone could flip the predicted activity of the enzyme. To explore this, we positioned positive and negative charges 20 Å away from the Fe center of the active site, aligned along the F_z axis on each side of the heme plane (**Figure 3B**). Here we selected a Tyr-ligated complex (PDB code 2j2m) as a test subject, allowing us to determine if the model could be biased to predict Cys-ligated/oxygenases for positive fields of large magnitudes and His-ligated/peroxidases for significant negative electric fields. This choice of protein, an unseen test example, also belongs to the category of Tyr-ligated proteins that exhibit intermediate, near-zero F_z values. We tested four distinct electric field strengths: +50, +10, -10, and -50 MV/cm along the iron-oxy bond, with the direction of the field indicated by the black arrow in **Figure 3B**. These field intensities were informed by our prior research,[24] which categorized Cys, Tyr, and His-ligated heme Fe proteins, under average vertical fields of 28.5 MV/cm, 3 MV/cm, and -8.7 MV/cm, respectively.

Applied Field (MV/cm)	**Predicted Ligand/Activity**
+50	Cys-ligated Oxygenases
+10	Tyr-ligated Catalases
0 (Original)	Tyr-ligated Catalases
-10	His-ligated Peroxidases
-50	Cys-ligated Oxygenases

Table 2*: Table illustrates how inducing an electric field along the oxy-iron bond modifies the predicted activity of the protein. Notably, large negative fields along the bond led to*

categorization as C/oxygenases—an outcome that seems unlikely based on our previous studies and thus suggesting a limitation of the low-dimensional, uniform electric field applied.

Our most effective model seemingly shows mixed success in predicting enzyme activity with applied electric fields, as illustrated by the results presented in **Table 2**. The model correctly altered its predictions for most cases: a large (+50 MV/cm) positive field switched the accurate prediction from a Tyr-ligated catalase to a Cys-ligated oxygenase, while a moderate (-10 MV/cm) negative field led to a prediction of a His-ligated peroxidase. However, the model's limitations became apparent in certain cases; notably, a strong negative field along the Fe(IV)=O bond incorrectly predicted a Cys-ligated oxygenase—an outcome that seems unlikely considering the typically moderate F_z component magnitudes observed in this family of proteins. These discrepancies suggest that the model might be utilizing more than just the F_z electric field component from the heterogeneous 3-D electric field of the protein in making its predictions.

Feature Importance. Therefore, we conducted a feature importance analysis to identify all the crucial features (i.e., the key principal components) involved in the model's accurate decision-making process. A naive approach would be to consider the % explained variance of each PCA component and assert that the most variable components impact activity more. This is imperfect for several reasons. First, correlation is not causation and this signifies that components with a large variance determine ligand specificity. Looking at the correlation or variance in a single component also ignores the effects that multiple vector field components might have in conjunction. Finally, PCA does not intake labels in a supervised manner, thus these components have no mapping to function directly. To address this, we utilized Boruta[42] and permutation importance[43] feature selection. Boruta is built on top of permutation feature importance, where individual variables are shuffled between examples and the resulting change in performance gives a quantitative measure of how important that feature was to a model's prediction. Boruta extends this idea by constructing "shadow features" that are Gaussian noise with the same mean and variance as true variables in the input of a model. These features, which by construction are random, serve as a benchmark of importance for other variables; if a variable is more important in permutation importance than a shadow feature it is more likely to be of importance in predicting a target label. This process is repeated a fixed number of times and these trials, in conjunction, creating a binomial distribution where features eventually fall into the tails of the distribution - important or not important. The resulting components from this feature selection step were studied by backtracking PCA components to their original electric field motifs.

Between Boruta and permutation importance, PC0, PC3, PC4, PC6, and PC7 were the most informative to the model. Visualizing these features (**Figure 3C**), we can summarize that a rich host of electric field features inform model predictions. Important components such as the field along the iron-oxy bond emerge, corroborating previous findings and supporting the notion that fields will shift electron distribution along this bond to promote the activation of substrates and control the selectivity. Combined, PC0 and PC3 have strong components along the Fe(IV)=O bonds, but opposite lateral components - suggesting that they together could explain the strong "vertical" component also previously proposed. PC4 is an entirely lateral field

component - not previously established as an important motif in Heme selectivity. PC4 might contribute to the placement and delocalization of the radical on the porphyrin (versus on the nearby Trp residue), particularly in the His-ligated proteins. Components PC6 and PC7 are harder to decipher visually - they have strong compressive/expansive features that shift electric fields into or out of the heme center and might control the access to the active site. These components are undoubtedly complex and underscore the difficulty of fully interpreting the effect of electric field processes *a priori* without a statistical, high-throughput approach. It is also noteworthy that the most variable field components, as indicated by percentage explained variability, were not necessarily the most informative for the models. Thus, our findings demonstrate that features of the 3-D electric field, extending beyond just the F_z component, are crucial for enhancing the accuracy of model predictions related to enzyme activity. This underscores that enzymes utilize these diverse directionalities within the 3-D field at the active site to drive their catalytic functions.

Dynamic 3-D fields. To build upon our static, single-frame analysis, we expanded our approach to incorporate temporal information via molecular dynamics (MD) trajectories of known proteins from each class. The premise here is that the field, as much as the protein producing it, is not static and that particularly functional fields may emerge dynamically. We selected a training set consisting of the proteins 1dgh and 1gwf (Tyr-ligated), 1ebe and 1hch (His-ligated), and 4g3j(Cys-ligated), and designated one protein from each class for the test set: 1u5u (Tyr-ligated), 3xvi (His-ligated), and 1jio (Cys-ligated) – again, ligation being linked to catalase, peroxidases, and monoxygenase activity, respectively. Employing the same suite of models, we optimized parameters using subsets of the electric fields from the training set and implemented a simple majority voting system to determine the protein class/activity. The results reveal that while the models performed well in the static single-frame analysis with a high F1-score of 0.84 using 3-D fields, their performance declined in the dynamic setting, as evidenced by the XGBoost model achieving an F1-score of 0.35 and an accuracy of 0.43 (Table 4), signifying a drop in the ability of the models to generalize to the dynamic regime. We do note that taking a majority-vote approach to predicting activity from MD trajectories, we are able to predict the activity of 2 out of 3 protein classes correctly.

MD Trials	**Test F1**	**Test Acc**
XGBoost, MD	**0.35**	**0.43**
XGBoost, MD (Combined PCAs)	**0.59**	**0.59**

Table 4*. The data illustrates the performance outcomes for XGBoost models tailored to molecular dynamics simulations.*

Protein	***Ground Value***	***Majority Prediction***	***Majority Prediction(Combined PCAs)***

1u5u	*Y*/catalase	*Y*/catalase	*Y*/catalase
3abb	*C*/oxygenase	*Y*/catalase	C/oxygenases
1apx	*H*/peroxidase	*H*/peroxidase	*H*/peroxidase

Table 5: *Predicted activities for proteins in our test set, utilizing two distinct approaches: predictions made with PCA components just from the training set and those using combined PCA components constructed from the training and testing set. The comparative analysis highlights that employing combined PCA components leads to improved prediction accuracy. This improvement suggests that the previously observed poor performance was likely due to the dynamics introducing a broader variety of components.*

To better understand why models extended on 3-D electric fields from MD simulations did not perform as expected, we examined the differences between electric fields derived from crystal structures and those obtained from MD simulations. **Figure 4A** presents the PCA explained variability, which measures the amount of variance each principal component captures from the dataset. This metric is commonly used to assess dataset dimensionality and complexity. Our analysis revealed significant differences in the cumulative variance between PCA results from crystal structures and those from dynamic simulations. This suggests that dynamic electric fields encapsulate more complex patterns and interactions, which are not as prevalent in the static fields derived from crystal structures. The increased complexity in dynamic fields likely reflects the continual conformational changes and interactions within the protein environment.

Further complicating our model training, there was a noticeable difference in the explained variance between the PCA components derived from our training set (MD simulations) and our test set. Specifically, the training set demonstrated a higher explained variance, with fewer PCA components, compared to the test set. This indicates that the PCA components from the training set may be over fit to a small set of dynamical degrees-of-freedom. Consequently, when these PCA components used to reduce dimensionality in the test set, they may not adequately capture the essential features needed for accurate predictions, leading to a mismatch in the model's ability to generalize. The model trained on less variable and comparatively simpler data from the MD training set struggles to accurately interpret and predict the behavior of complex test data. This issue highlights the need for developing strategies that can better account for and adapt to the variations in electric field complexity between different sets of molecular dynamics data.

To enhance the interpretability of our MD based 3-D electric field analysis and reduce its complexity, we have recently developed a protocol that captures dynamic information regarding the electric fields experienced by the active site of a protein,[26] as illustrated in **Figure 4B**, followed by mapping these clusters onto the principal components identified as critical. By capturing the complex dynamic fluctuations within the enzyme's active site, we aim to elucidate how these variations complicate the model's ability to accurately predict enzyme activity.

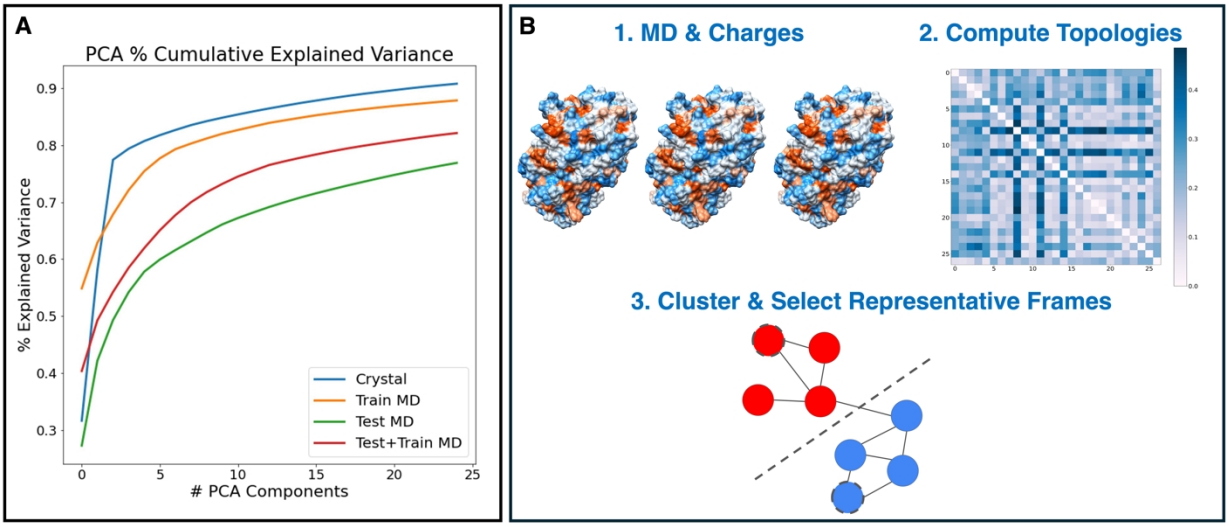

Figure 4. (A) Cumulative explained variance between PCAs constructed from crystal structure fields show these fields require fewer components to explain dataset variability. (B) An outline of our method for selecting representative frames based on electric field topologies.

Here we focused on the components that Boruta and permutation importance determined to be critical: PC0, PC3, PC4, PC6, and PC7. PC0, recognized as the most vertical component along the Fe-O bond, exhibited clustering trends that align with our previous studies. The ordering of His < Tyr < Cys within these clusters suggests that cysteine-binding proteins tend to exhibit the most positive electric field components along this direction (**Figure S2**). However, the presence of both Tyr and His-ligated complexes in the most positive clusters of this component might affect model's accuracy. PC4 exhibits a strong vertical orientation. Notably, clusters representing 1jio are among the most positively positioned on PC4 (**Figure S3**). While the overall trend of His < Tyr < Cys is maintained, there is significant overlap among the data points of the three protein classes in the projection onto this principal component. In the case of PC7, which introduces a vertical component with some compressive characteristics toward the active site, 1jio is distinctly the most positive, suggesting preorganization of the electric field to enhance activity at the active site (**Figure S4**). Contrarily, 1u5u and 3vxi show mixed projections on this component, aligning with prior observations of comparable F_z components between these protein categories. Our analysis on PC6 revealed a lack of clear separation between protein types, indicating that this component is less interpretable compared to others (**Figure S5**). PC3, characterized by its predominantly horizontal orientation orthogonal to many other significant components, uniquely identified the most positive cluster associated with 1jio(C) (**Figure S6**). This specificity did not extend to 1u5u and 3vxi, which did not separate distinctly along this component. This structured approach of clustering and principal component mapping has revealed that among the most important principal components for the model, certain components, such as PC0, can distinctly separate the three protein systems within dynamic data, while other components like PC6 and PC3 complicate the clarity of these separations.

System Description	E_H^o **(kcal/mol)**

1jio (Cys/Oxygenase – 2 Clusters)	68.5 – 92.3
1u5u (Tyr/Catalase – 3 Clusters)	64.7 – 68.5
3vxi (His/Peroxidase – 4 Clusters)	27.6 – 83.7

Table 6. Proton-coupled electron transfer potential (E_H^o) ranges for enzyme systems analyzed using QM/MM methods, highlighting variations across different clusters within each enzyme category.

Finally, we aim to explore whether the dynamic complexity, identified through PCA, truly influences factors critical to enzyme activity. To this end, we decipher how the dynamic 3-D heterogeneous electric field affects the electronic structure of the CpdI Fe(IV)=O ($Por^{+\bullet}$) and CpdII Fe(IV)-OH complex by employing quantum mechanics/molecular mechanics (QM/MM) calculations. For these calculations, we have chosen specific model systems that are representative of the enzyme classes under study: 1jio for monoxygenases, 3vxi for peroxidases, and 1u5u for catalases. The selection of structures such that they represent unique electric field configurations, is vital. Random or field-agnostic selection methods may fail to capture variations caused by heterogeneous electric fields, potentially overlooking critical dynamic interactions that influence enzyme activity. Therefore, we used above identified cluster centers, via electric field clustering, for these calculations. For each major cluster (>10% representation), we computed the free energies of the CpdI Fe(IV)=O and CpdII Fe(IV)-OH variants to assess the relative activity of each cluster along the putative reaction pathway. The computed proton-coupled electron transfer potential (E_H^o) ranges for these clusters are as follows: 68.5 – 92.3 kcal/mol for the Cys-ligated oxygenase system 1jio, 64.7 – 68.5 kcal/mol for 1u5u, and 27.6 – 83.7 kcal/mol for 3vxi (**Table 6**). These values align with the expected trend where Cys-ligated oxygenases exhibit higher reactivity compared to Tyr-ligated catalases and His-ligated peroxidases. Thus, the results indicate that the range of E_H^o values becomes less distinct between the three systems, suggesting that the introduction of dynamics extends the ranges of catalytically relevant properties and diminishes the clear segregation between them. This blurring effect might help explain why dynamics affects the machine learning model's ability to accurately classify the different systems using 3-D electric fields.

In response to these findings, we hypothesized that a model constructed with combined principal components from both the test and training datasets, providing a broader spectrum of variability for the model to learn from, might enhance classification accuracy. Indeed, this approach resulted in improved performance, where our F1 and test accuracy improved to 0.59 and majority vote approach correctly predicts all three test protein categories (**Table 4** and **5**). Here, we note that mixing train and test components between the sets is neither completely valid nor entirely invalid. On one hand, it introduces bias that can obscure the evaluation of the model's generalization. Therefore, our initial approach avoided this combination. On the other hand, in practical applications, combining electric fields to construct PCAs does not require prior knowledge of protein activity. Consequently, this approach remains a valuable tool for protein analysis via electric fields.

For future improvements in handling dynamic electric field data, the implementation of highly efficient, sparse neural network architectures and advanced signal processing techniques could be beneficial. Equivariant neural networks, which have rapidly gained traction in scientific fields, are particularly promising due to their efficiency in learning with less data. When integrated with robust data augmentation schemes, these networks can directly process raw electric fields, minimizing data demands while ensuring that key physical symmetries are preserved. Additionally, embracing methods that intrinsically manage structured, temporal data will be essential for extending the analysis to include dynamics natively. Architectures borrowed from natural language processing, such as Long Short-Term Memory (LSTM) networks, or those that incorporate geometric learning, like message-passing graph neural networks, are well-suited for this purpose. These techniques can effectively interpret the temporal variations observed in MD trajectories, potentially enhancing the ability to predict protein behavior based on dynamic electric fields.

Conclusions

In this study, we have developed a machine-learning pipeline that ingests electric fields, reduces dimensionality via PCA, and applies these fields in a supervised learning task. Our tests on a well-studied family of Fe heme enzymes demonstrated that traditional lower-dimensional analyses of electric fields along the Fe(IV)=O bond are insufficient for accurate activity prediction. This underscores the necessity for analytical techniques capable of parsing the more complex, heterogeneous fields that are actually present at protein active sites. Our findings reveal that point electric field calculations, despite their simplicity and ease of interpretation, do not accurately reflect the true nature of electric fields within these sites. Additionally, when we applied a uniform electric field using our trained model, it failed to induce the predicted changes in a test protein, highlighting the importance of multidirectional fields in enzyme function. Importantly, our trained machine learning model demonstrated that the enzyme's 3-D heterogeneous electric field alone can predict its function without any other protein-specific information. Through feature selection techniques such as Boruta and permutation importance, we identified key electric field components that not only corroborated previous studies but also emphasized the critical influence of several components alongside the F_z value along the Fe-O bond. Expanding our analysis to include MD trajectories and employing PCA, clustering, and QM/MM calculations, we observed that the inherent complexity in protein dynamics can complicate model predictions. However, we show that if the model is exposed to sufficient dynamic variability, its performance can improve significantly.

This research marks a significant advancement in our understanding of electrostatics in proteins. We have shown that natural enzyme scaffolds have evolved to optimize the electric field at the active site, tailored to their function. This insight offers a powerful tool for predicting potential enzyme functions based solely on their electric fields. Although our analysis focused on heme Fe proteins, the methodology is broadly applicable to any study involving electric fields at largely conserved active sites, even where there is no prior knowledge of crucial field components. Overall, the approach presented here provides a robust framework for not only

understanding but also predicting enzyme functions across diverse biological systems based solely on electric field analysis.

Acknowledgement
This work was supported by the NSF-CHE grant 2203366 to A.N.A. S. V. was supported by the Department of Energy Computational Science Graduate Fellowship under grant DE-SC0021110.

Data Availability
Code and datasets for this work can be found at https://github.com/santi921/HEML.

References
1. Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci.* **75**, 5250–5254 (1978).
2. Warshel, A. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* **273**, 27035–27038 (1998).
3. Chiang, A. H., Kastrup, C. & Vogan, J. M. Electric Fields and Enzyme Catalysis. in (2017).
4. Marshall, N. M. *et al.* Rationally tuning the reduction potential of a single cupredoxin beyond the natural range. *Nature* **462**, 113–116 (2009).
5. Bím, D. & Alexandrova, A. N. Electrostatic regulation of blue copper sites. *Chem. Sci.* **12**, 11406–11413 (2021).
6. Dubey, K. D., Stuyver, T. & Shaik, S. Local Electric Fields: From Enzyme Catalysis to Synthetic Catalyst Design. *J. Phys. Chem. B* **126**, 10285–10294 (2022).
7. Chaturvedi, S. S. *et al.* Can an external electric field switch between ethylene formation and L -arginine hydroxylation in the ethylene forming enzyme? *Phys. Chem. Chem. Phys.* **25**, 13772–13783 (2023).
8. Waheed, S. O., Chaturvedi, S. S., Karabencheva-Christova, T. G. & Christov, C. Z. Catalytic Mechanism of Human Ten-Eleven Translocation-2 (TET2) Enzyme: Effects of Conformational Changes, Electric Field, and Mutations. *ACS Catal.* **11**, 3877–3890 (2021).
9. Ji, Z. & Boxer, S. G. β-Lactamases Evolve against Antibiotics by Acquiring Large Active-Site Electric Fields. *J. Am. Chem. Soc.* **144**, 22289–22294 (2022).
10. Fried, S. D., Bagchi, S. & Boxer, S. G. Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science* **346**, 1510–1514 (2014).
11. Morgenstern, A., Jaszai, M., Eberhart, M. E. & Alexandrova, A. N. Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chem. Sci.* **8**, 5010–5018 (2017).
12. Shaik, S., Mandal, D. & Ramanan, R. Oriented electric fields as future smart reagents in chemistry. *Nat. Chem.* **8**, 1091–1098 (2016).
13. Shaik, S., Danovich, D., Joy, J., Wang, Z. & Stuyver, T. Electric-Field Mediated Chemistry: Uncovering and Exploiting the Potential of (Oriented) Electric Fields to Exert Chemical Catalysis and Reaction Control. *J. Am. Chem. Soc.* **142**, 12551–12562 (2020).
14. Isaksen, G. V., Hopmann, K. H., Åqvist, J. & Brandsdal, B. O. Computer Simulations Reveal Substrate Specificity of Glycosidic Bond Cleavage in Native and Mutant Human Purine Nucleoside Phosphorylase. *Biochemistry* **55**, 2153–2162 (2016).
15. Sharma, P. K., Chu, Z. T., Olsson, M. H. M. & Warshel, A. A new paradigm for electrostatic catalysis of radical reactions in vitamin B_{12} enzymes. *Proc. Natl. Acad. Sci.* **104**, 9661–9666 (2007).
16. Hennefarth, M. R. & Alexandrova, A. N. Direct Look at the Electric Field in Ketosteroid Isomerase and Its Variants. *ACS Catal.* **10**, 9915–9924 (2020).

17. Adamczyk, A. J., Cao, J., Kamerlin, S. C. L. & Warshel, A. Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Proc. Natl. Acad. Sci.* **108**, 14115–14120 (2011).
18. Jabeen, H. *et al.* Electric Fields Are a Key Determinant of Carbapenemase Activity in Class A β-Lactamases. *ACS Catal.* **14**, 7166–7172 (2024).
19. Warshel, A., Sharma, P. K., Chu, Z. T. & Åqvist, J. Electrostatic Contributions to Binding of Transition State Analogues Can Be Very Different from the Corresponding Contributions to Catalysis: Phenolates Binding to the Oxyanion Hole of Ketosteroid Isomerase. *Biochemistry* **46**, 1466–1476 (2007).
20. Feierberg, I. & Åqvist, J. The Catalytic Power of Ketosteroid Isomerase Investigated by Computer Simulation. *Biochemistry* **41**, 15728–15735 (2002).
21. Kamerlin, S. C. L., Sharma, P. K., Chu, Z. T. & Warshel, A. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization. *Proc. Natl. Acad. Sci.* **107**, 4075–4080 (2010).
22. Wilson, T. R., Morgenstern, A., Alexandrova, A. N. & Eberhart, M. E. Bond Bundle Analysis of Ketosteroid Isomerase. *J. Phys. Chem. B* **126**, 9443–9456 (2022).
23. Fuller, J., Wilson, T. R., Eberhart, M. E. & Alexandrova, A. N. Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization. *J. Chem. Inf. Model.* **59**, 2367–2373 (2019).
24. Bím, D. & Alexandrova, A. N. Local Electric Fields As a Natural Switch of Heme-Iron Protein Reactivity. *ACS Catal.* **11**, 6534–6546 (2021).
25. Lai, W., Chen, H., Cho, K.-B. & Shaik, S. External Electric Field Can Control the Catalytic Cycle of Cytochrome P450cam: A QM/MM Study. *J. Phys. Chem. Lett.* **1**, 2082–2087 (2010).
26. Chaturvedi, S. S., Vargas, S., Ajmera, P. & Alexandrova, A. N. Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field. *J. Am. Chem. Soc.* jacs.4c03914 (2024) doi:10.1021/jacs.4c03914.
27. Stuyver, T., Ramanan, R., Mallick, D. & Shaik, S. Oriented (Local) Electric Fields Drive the Millionfold Enhancement of the H-Abstraction Catalysis Observed for Synthetic Metalloenzyme Analogues. *Angew. Chem. Int. Ed.* **59**, 7915–7920 (2020).
28. Chaturvedi, S. S., Bím, D., Christov, C. Z. & Alexandrova, A. N. From random to rational: improving enzyme design through electric fields, second coordination sphere interactions, and conformational dynamics. *Chem. Sci.* **14**, 10997–11011 (2023).
29. Welborn, V. V. & Head-Gordon, T. Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase. *J. Am. Chem. Soc.* **141**, 12487–12492 (2019).
30. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).
31. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
32. Case, D. A. *et al.* AMBER 2018. University of California.
33. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
34. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
35. Metz, S., Kästner, J., Sokol, A. A., Keal, T. W. & Sherwood, P. Chemshell—a modular software package for QM/MM simulations. *WIREs Comput. Mol. Sci.* **4**, 101–110 (2014).
36. Ahlrichs, R., Bär, M., Häser, M., Horn, H. & Kölmel, C. Electronic structure calculations on workstation computers: The program system turbomole. *Chem. Phys. Lett.* **162**, 165–169 (1989).

37. Smith, W., Yong, C. W. & Rodger, P. M. DL_POLY: Application to molecular simulation. *Mol. Simul.* **28**, 385–471 (2002).
38. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
39. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
40. Chen, C. Using Random Forest to Learn Imbalanced Data. *Univ. Calif. Berkeley* **110**, 1–12 (2004).
41. Biewald, L. Experiment Tracking with Weights and Biases. (2020).
42. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the **Boruta** Package. *J. Stat. Softw.* **36**, (2010).
43. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

TOC