JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Article

Machine Learning Integrating Protein Structure, Sequence, and Dynamics to Predict the Enzyme Activity of Bovine Enterokinase Variants

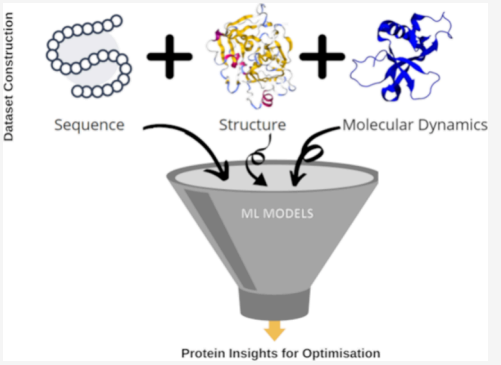
Niccolo Alberto Elia Venanzi, Andrea Basciu, Attilio Vittorio Vargiu, Alexandros Kiparissides, Paul A. Dalby, and Duygu Dikicioglu*

Cite This: https://doi.org/10.1021/acs.jcim.3c00999

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Despite recent advances in computational protein science, the dynamic behavior of proteins, which directly governs their biological activity, cannot be gleaned from sequence information alone. To overcome this challenge, we propose a framework that integrates the peptide sequence, protein structure, and protein dynamics descriptors into machine learning algorithms to enhance their predictive capabilities and achieve improved prediction of the protein variant function. The resulting machine learning pipeline integrates traditional sequence and structure information with molecular dynamics simulation data to predict the effects of multiple point mutations on the fold improvement of the activity of bovine enterokinase variants. This study highlights how the combination of structural and dynamic data can provide predictive insights into protein functionality and address protein engineering challenges in industrial contexts.



INTRODUCTION

Proteins are essential, powerful machines in biology and consequently find a wide range of application areas in biotechnology, including the manufacturing of targeted therapies. However, their development as a functional product is expensive, time-consuming, and frequently yields unsuccessful results.[1] The estimated average cost of bringing a new protein-based therapy to market is between $1 and $3 billion, and the success rate of clinical trials is below 10%. To overcome these obstacles, researchers are investigating new methods for expediting the engineering of proteins with enhanced functionality and manufacturability. Protein engineering entails making precise alterations to the original sequence of a protein to identify variants with desirable properties while minimizing interference with its function. Thus, protein engineering has revolutionized the production and use of protein-based products.[2] However, due to the vast number of possible amino acid combinations, exhaustive experimental exploration of the landscape of protein fitness remains nearly impossible.[3−6] Theoretically, the scope of mutations could be restricted to include only the ostensibly significant fragments of the protein such as the binding sites. However, this remains a heuristic solution applicable to a limited number of cases, as the majority of protein properties depend on the entire sequence and structural conformation, not just a few amino acids, due to the presence of epistatic effects.[7,8]

Biodescriptors are quantitative characteristics that shed light on a protein's chemistry and structure. Algorithms can use the information contained in biodescriptors to predict the effects of amino acid substitutions on the properties of proteins. Recently, machine learning (ML) techniques have been applied to the classification of proteins and the prediction of the stability of protein−ligand complexes.[9−13] Nevertheless, biodescriptors, which incorporate function-related properties associated with the macromolecules, were not employed in the investigation of the role of mutations introduced into the protein sequence on protein performance.[14−16]

Unsupervised ML methods mostly use natural language processing (NLP) to obtain sequence-level information for protein prediction. Recently, generative models that are able to design effective and diverse proteins suitable for various applications have received substantial attention, owing to their promise and potential. These models are all based on the assumption that sequence-based information, i.e., the order of the amino acids in a given length of peptide sequence, encompasses both structure and function information in its entirety. However, a true understanding of the physics and

Special Issue: Machine Learning in Bio-cheminformatics

Received: July 3, 2023
Revised: February 12, 2024
Accepted: February 13, 2024

biochemical properties that go beyond sequence information has promise to be very useful for model interpretability and scientific discovery.

Molecular Dynamics (MD) simulations have emerged as an efficient method for describing a vast array of biomolecular processes, such as protein folding and molecular recognition, and for demonstrating how mutations affect the stability and function of proteins.[17−22] MD simulation techniques allow the investigation of the conformational landscape of a protein and obtain direct information about its flexibility.[23−27] Several approaches coupling ML algorithms with MD-derived data have been developed and successfully implemented over the past few years. MD-derived features were shown to be capable of predicting the stability of protein−protein and protein−ligand complexes[28] and designing effective drug candidates for a known protein target.[29] To our knowledge, however, there is no established or proposed method for predicting the effects of mutations on protein function by using machine learning algorithms that incorporate sequence-based information with MD simulations.

In this work, we present an ML workflow that leverages data generated via MD simulations with sequence- and structural-based features to predict the effects of multiple point mutations on the activity of an enzyme. The model protein employed in this study is bovine enterokinase, an enzyme used to remove affinity tags from high-value biopharmaceuticals.[30] We applied a range of ML models to the 312 variants investigated in this study. The models utilized biodescriptors for sequence, structure, and dynamics-based features for each variant to predict its function and assessed the predictive performance of these models against empirical data available on the functional properties of each variant of the enzyme. We present here an effective machine learning-based strategy for incorporating different levels of information to successfully predict the functional properties of protein variants to enable faster and more powerful routes to protein engineering. We demonstrate the interpretability of these models by identifying the key biodescriptors contributing to the prediction of function and validate how the ML-based models can provide us key insight on the role of specific point mutations introduced to the protein sequence. We also discuss below the challenges and opportunities around incorporating simulation-based data as input for ML algorithms.

METHODS

Experimental Data Set. This study used 312 variants of the engineered template bovine enterokinase (EKB), each containing one to nine mutations randomly introduced at the amino acid level in specific regions of the protein as described previously.[30] The activity of the enzyme was determined upon expressing the protein at 30 °C with and without preincubation heating. The fold in activity was defined as the ratio of the activity of a variant to that of the template EKB. Here we intend to predict the difference between the two experimental settings measured, defined as the fold change in activities (FCA) (Table S1). The experimental data set was constructed by carrying out multiple rounds of error-prone PCR (epPCR), introducing modifications at the nucleotide level, resulting in a total of 312 variants containing zero to nine co-occurring mutations at the amino acid level.

Homology Modeling and Variants Structures Constructions and Evaluations. SWISS-MODEL was used to build a template-based homology model of the engineered form of bovine enterokinase and the 312 variants (PDB ID: 1EKB).[31−33] BLAST and HHBlits were used to search the SWISS-MODEL template library for the sequence of the engineered protein.[34,35] Models were constructed with ProMod3 using the target template 1EKB.[36] The final models were selected to maximize the Qualitative Model Energy Analysis Distance Constraint (QMEANDisco) and the Global Model Quality Estimation (GMQE) scores, evaluating both the stereochemical and energetic features together with structural similarity to 1EKB.[31,37] AlphaFold-2 (AF2) was performed using one model structure generation with three recycles using amber and MMseqs2 (UniRef + Environmental) for Multiple Sequence Alignment (MSA) mode. The structures thus created were then compared to homology model structures using Biopython by superimposing and calculating the root-mean-square deviation per residue (RMSD).

AlphaFold-2. AF2 was utilized by locally executing the publicly available ColabFold script. AF2 employs the predicted local distance difference test (pLDDT) to assess the accuracy of predicted C-alpha locations (on a scale of 0−100) with experimental structures as well as the predicted Template Modeling (pTM) to create projected aligned error (PAE) maps.[38−40]

Molecular Dynamics Simulations. PROPKA's[41] online service was used to set the charge of the amino acids at the experimental pH of 8.0.[30] The model was constructed on the GROMACS 2019.3[42] molecular dynamics engine, and MD simulations were run using the OPLS-AA force field.[43] TIP3P[44] was used as the water model to allow faster computation. All of the proteins were placed in a cubic box, allowing for at least a 1.0 nm protein-edge distance in each dimension. The systems were neutralized using 50 mM Na^+ and Cl^- ions randomly placed to resemble the experimental conditions under which the variants were expressed. The system was then energy minimized over a maximum of 50,000 steps, stopping at 1000 kJ mol^{-1} nm^{-1}. Isothermal-isochoric equilibration at 300 K was performed over 100 ps using the leapfrog integrator with a coordinate update rate of 1.0 ps. The isothermal−isobaric equilibration was carried out using the same settings as the isothermal-isochoric step with the addition of a Parrinello-Rahman pressure coupling and a reference pressure of 1.0 bar. In all simulation steps, the standard Particle-Mesh-Ewald (PME) model[45] was used to treat long-range electrostatic interactions in systems with periodic boundaries. A cutoff distance of 1.0 nm was set for short-range van der Waals and electrostatic interactions.

Simulation Lengths and Evaluation of Variance Analysis among MD Repeats. A subset of 15 variants was randomly selected for the evaluation of the methodology before scaling the simulation protocol to the remaining variants. Following preliminary analysis, the 312 variant molecular structures were simulated five times at the selected length, leading to 1605 trajectories.

Statistical Analysis. A subset of 15 variants was selected for evaluating the methodology. These were simulated across 10 ns spaced simulation lengths ranging from 10 to 200 ns in triplicate. One-way Analysis of Variance (ANOVA)[46] of the root-mean-square deviation (RMSD) values were used to assess the minimum simulation length and the variance between replicate trajectories. The simulation length comparisons focused on the last 10 ns of each simulation. The analysis was performed using SciPy 1.0 through the f_oneway function.[47]

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Path Similarity Analysis (PSA). PSA was performed using MDAnalysis.psa built-in function using Hausdorff distances.[48] Hausdorff distance is an established metric used to compare the geometries of trajectories by comparing two paths P and Q as sequences of conformations. The distance is then calculated as

$$\delta_H(P, Q) = \max(\delta_H(P|Q), \delta_H(Q|P)) \quad (1)$$

where $\delta_H(P|Q)$ is the directed Hausdorff distance from P to Q.

$$\delta_H(P, Q) = \max_{p\epsilon P} \min_{q\epsilon Q} d(p, q) \quad (2)$$

Biodescriptors Data Set. A data set of 192 biodescriptors that provide information on protein sequence, structure, and dynamics was constructed for the 312 variants (Table S2).

Sequence and Structure Descriptors. The sequence embeddings were extracted using the R CRAN package "Peptide", calculating global protein properties in the form of 66 features.[49] These included: Cruciani properties,[50] Kidera factors,[51] zScales,[52] FASGAI vectors,[53] and the BLOSUM indices. Fourteen global properties of the variants were calculated using the module ProtParam of the Biophython package,[53] including molecular weight and isoelectric point.

MD Descriptors. Root-Mean-Squared Deviation (RMSD) and Radius of Gyration (RoG) with respect to the initial structure were extracted for the following selections using the GROMACs package:[42] whole proteins without non-hydrogenic atoms, the backbone, and Cα. These three features were also obtained for the binding site alone, defined as the residues within 3.5 Å of the ligand in the PDB crystal structure. The time series points were preprocessed, keeping the average and standard deviation in the final data set. The Dictionary of Protein Secondary Structure (DSSP)[54] was also extracted as an indicator of structural characteristics of proteins throughout the trajectory, including parallel beta-sheets, antiparallel beta-sheets, alpha helices, 3−10 helices, turns, bent, and random coils using AMBER tools via Pytraj.[55]

MDpocket. MDpocket was used to extract additional features from the binding pocket calculated during the trajectory. The binding site was defined the same way as for the MD descriptors, where only residues within 3.5 Å of the peptide-like ligand in the template model were used for homology modeling. Features included the size and length of the binding site, its depth accessibility, hydrophobicity, the charge of each amino acid, the average number of times an amino acid was encountered in the binding site during the simulation, polarity/apolarity, and the total surface area as well as the normalized B-factor score of the binding site through the trajectory.[56] For practical reasons, we refer to features in the rest of the text with acronyms, as defined in the Supporting Information in Table S2.

Machine Learning Algorithms. The features were standardized using MinMaxScaler from Scikit-Learn.[57] The data set was randomly split such that 80% was used for training and the remaining 20% was used as a test set. 42 supervised machine learning algorithms are present in Scikit-Learn (Table S3)[57] and were fitted with their standard parameters to the data to find the best-performing template model to then carry forward for hyperparameter tuning according to Table S4. FCA was used as the response variable in the models. The models were implemented using the default parameters of Scikit-Learn and evaluated based on root-mean-square error (RMSE) (Eq. 3), R^2 (Eq. 4), and execution time in seconds. Following the removal of those models with negative R^2 from further consideration, the mean absolute error was added as a performance metric (MAE) (Eq. 5)

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (3)$$

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \quad (4)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (5)$$

where $\hat{y}$ is the predicted value of y, $\bar{y}$ is the mean value of y. The normality tests were conducted by calculating the Fisher-Pearson correlation and graphically demonstrating the distribution of FCA with a Q-Q plot.

Model Building. The performance of the selected algorithm, LightGBM, was evaluated when only specific categories of features were employed: sequence/structure descriptors (Seq), MD descriptors, or MDpocket descriptors, resulting in the following combinations: Seq+MD+MDpocket, Seq+MD, MD+MDpocket, Seq+MDpocket, Seq, MD, and MDpocket. All model parameters were tuned using RandomizedSearchCV from Scikit-Learn over 5 cross-validation folds through a defined search space with the negative mean absolute error as a scoring function (Tables S4 and S5). To evaluate the stability and variability of the model, we performed a total of 500 bootstrap iterations. During each iteration, a bootstrap sample was randomly selected with a replacement from the training set. The model was then trained using the optimal hyperparameters obtained by RandomizedSearchCV for each data set, and their performance metrics were calculated for the predictions made on the test set. ANOVA was then performed to determine if there are any statistically significant differences in performance across the different data sets. Subsequently, the Tukey Honestly Significant Difference (HSD) test was conducted to perform pairwise comparisons between the groups, identifying which specific data sets differ significantly from each other.

Model Performance Evaluation and Analysis. The models built using the seven different feature combinations reported above were evaluated based on R^2, MAE, and RMSE as defined in Scikit-Learn.[57] The feature importance values were evaluated by using two methods: Tree SHapley Additive exPlanations (SHAP)[58] and permutation feature importance.[59] SHAP is a method based on game theory that assigns a value to each feature that represents its contribution to the prediction as well as providing insights on feature interactions. It computes Shapley values for each feature by averaging over all possible permutations of the features, resulting in an accurate measure of the feature importance. Permutation feature importance is a method that evaluates the impact of each feature by randomly permuting its values and measuring the resulting decrease or increase in the model's performance, thus providing a direct measure of feature importance. Together, they can provide a comprehensive understanding of their features and their importance. Lastly, hypergeometric distribution testing was performed to identify patterns in the predictions.

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

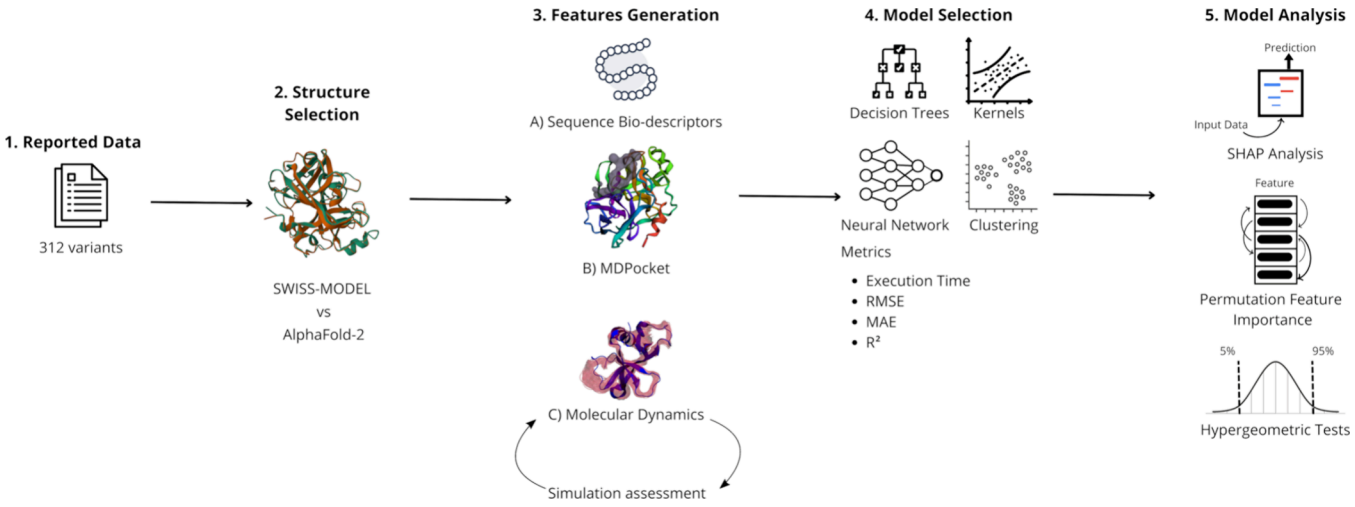


Figure 1. Proposed pipeline comprises MD simulation generated data with sequence and structure features for protein engineering on previously reported enterokinase bovine variants.

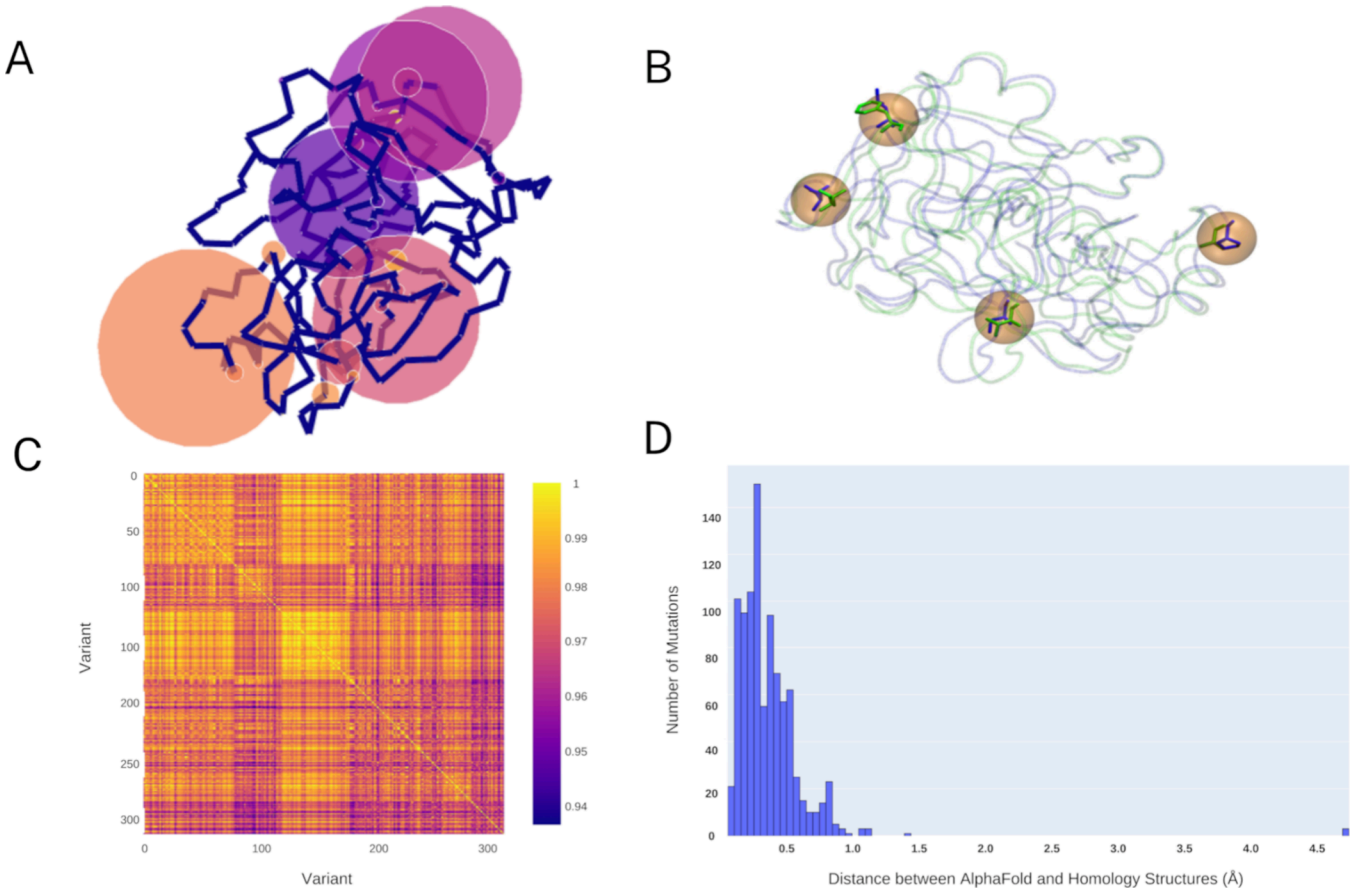


Figure 2. Construction of protein structures and data set descriptions. (A) A graph representing the enzyme's mutation sites. The spheres are the normalized occurrences of mutations at specific sites throughout the data set. (B) Enterokinase bovine and engineered template enterokinase bovine superimposed with the mutation sites shown as licorice (PDB: 1EKB). (C) Heatmap of the identity matrix of the variants across the data set. (D) Histogram representing the RMSD between the mutations of the PDB constructed via AlphaFold-2 and the homology models.

RESULTS AND DISCUSSION

The aim of this work was to establish the role of dynamic features integrated with sequence and structure information in predicting protein function through ML. For this purpose, a case study containing 312 variants of the engineered Chinese Yellow bovine enterokinase light chain (EKL) was selected. The structures of all variants were predicted, and MD simulations were performed for all of the variants following the statistical evaluation of simulation lengths and the variability of the replicate simulations for each sample. The final data set was then constructed by extracting distinct categories of features, including sequence/structure and dynamics-based features, which were then used to predict

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

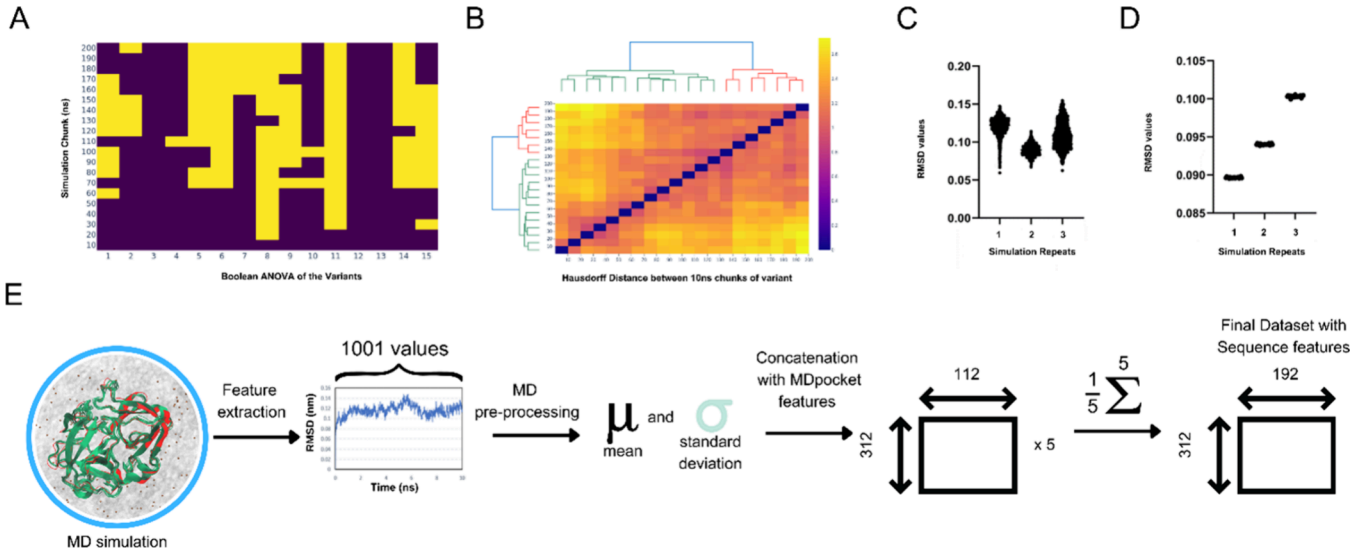


Figure 3. Statistical evaluation of the MD simulations. (A) ANOVA heatmap showing statistical differences across the variants up to 200 ns for 10 ns chunks against the first 10 ns (keys: purple = non statistical difference, yellow = statistically different). (B) Heatmap of Hausdorff distances with ward dendrogram showing clusters and distances between 10 ns simulation chunks across a trajectory. (C) Violin plot of the RMSD values of the simulations for a randomly selected variant from the 15-variant subset. (D) Individual dots plot for each element of the 20-cross fold averages of the same randomly selected simulation shown in (C), confirming the preservation of the differences between repeats of the simulation of the same protein structure after averaging. (E) Through statistical evaluations, the dimension of the final data set was reduced to 112 × 312 (MDpocket features × number of variants) from the initial set of 80 × 5 × 32(1001) × 312 (number MDpocket × replicate trajectories × MD-based biodescriptor entries per trajectory × number of variants). These were then concatenated with 80 sequence and structure biodescriptors, resulting in a data set of 192 features by 312 variants.

FCA using various ML-based algorithms. The best-performing template model was then carried forward to assess the importance of features from different categories and those that contributed the most to the model's performance (Figure 1). Each step of this newly proposed pipeline will be highlighted in the following sections.

Due to the nature of the epPCR protocol implemented, the data set comprised of mutations introduced only at specific sites across the protein structure, as depicted in Figure 2A. This methodology reflects the natural evolutionary pathway of the protein, resulting in a data set with similar amino acid content with an identity index above 90% and 29.7% of positions being at least mutated once through the data set (Figure 2B,C).[30] Activities of the enzyme variants were measured with and without a heat-shock, and the performance of the variants was evaluated by a parameter that we call FCA here, which quantifies the difference in activities between the two experimental settings, with a high FCA indicating a measure of the extent to which the introduced mutations improved the protein's activity.

Building the Protein Structures. The enzyme for which variant enzyme activities were measured does not have a crystal structure available. This limitation was overcome by using the protein structure for the Yellow Bovine Enterokinase (PDB: 1EKB, resolution: 2.30 Å),[32,33] the closest structure to the engineered enterokinase bovine with variants available. This structure and, consequently, the structures of the variants were required for the MD simulations and subsequent modeling steps. The difference between the two proteins is in the four mutation sites: V15Q/R82P/C112S/D176E (Figure 2B). To construct the mutations, we relied on homology modeling, as described earlier. The homology modeling offered quantitative and qualitative measurements for model assessment, resulting in an average GMQE and QMEANDisco of 0.88 and 0.86, respectively, and a standard deviation less than 0.01 in both cases.[37] The process was repeated for the experimental variants, resulting in similar metrics across the 312 variants, which indicated the acquisition of rigorously proposed structures as indicated by these metrics.[37,60] Since the initial protein structures are key elements of the proposed pipeline, it was essential to ensure that high-quality modeled PDB structures were made available before they were carried through the pipeline. For this purpose, an additional quality check was carried out to compare the structures generated via homology modeling to those proposed by AF2.

The pLDDT and pTM of the models obtained via AF2 were found to have an overall score above 90 and 0.90, respectively (Figure S1), in the variant sequence positions except for the N-termini, where the scores dropped below 80 and 0.80 in all of the models. Upon obtaining the AF2 models, the point mutations of the homology structures were compared with those of the AF2 generated ones, with 79% of the mutations displaying close similarity with an RMSD of 0.50 Å and 99% presenting an RMSD below 1.00 Å. The differences in residues were discarded if the distance was determined to be 1.00 Å, as these relate to conformational differences between the amino acids, also indicating possible manifestations of variations during the energy minimization step of the MD simulations. The presence of residues with more than a 1.00 Å difference was recognized as an indication of the fact that both the backbone and the rotamers of the amino acids have been modeled differently. These include mutations in the data set at the sequence positions 47, 49, 83, and 95 with RMSDs of 4.75−4.79, 1.45, 1.00−1.06, and 1.12−1.15 Å, respectively. A closer look at the secondary structure of these sites showed that positions 47 and 49 are part of a loop, whereas positions 83 and 85 belong to a random coil conformation. This finding is in accordance with previously reported findings highlighting

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

AF2's inability to model intrinsically disordered regions, which led to the decision to select the homology model structures for the simulations.[61−63]

Understanding the Role of Simulation Length and Robustness in MD Analysis. The pipeline development sought (i) to determine the minimum informative MD simulation length ranging from 10 to 200 ns and (ii) to evaluate the extent of variability between repeat simulations of the same case. Prior to running 312 computationally costly 200 ns simulations for all variants with replicates, we randomly selected a subset of fifteen variants with an FCA that was statistically representative of the whole data set (two-tailed Welch-corrected test, p-value = 0.78).

Minimum Informative MD Simulation Length. Conducting the MD simulations at the scale proposed here, which comprises more than 1600 trajectories, when considered in conjunction with the possibility of employing the pipeline in studies of even larger scale, necessitates a careful evaluation of the simulation times for the MD simulations. Therefore, the tradeoff between the information gained with extended timeframes and the computational cost of the extended simulation times needed to be carefully evaluated. For this purpose, the effect of simulation lengths on model success parameters was investigated by running three replicates of the simulations for a randomly selected subset of fifteen variants, which were simulated for extended times of up to 200 ns, and the trajectories were investigated in 10 ns increments starting from the shortest trajectory of 10 ns. As RMSD had the highest variance among the MD-based features extracted in this study, it was selected as the main feature for the statistical evaluation of the trajectories.[64]

A number of different approaches were selected to carry out this evaluation, as the findings emerging from the analysis would lead to decisions that would impact the success of the predictions made by the machine-learning models. A method that uses Principal Component Analysis (PCA) was previously proposed to assist decision-making around the similarity between trajectories; however, PCA of this data set did not yield any comparative insight as to the differences observed between trajectories run at different lengths, and therefore, more suitable alternative methods were employed to make conclusive decisions.[65] ANOVA was previously shown to work satisfactorily in comparing MD trajectories.[66] In our analysis, there was no statistically significant difference in the RMSD values obtained from a 10 or a 50 ns trajectory for 80% of the randomly selected variants investigated here. However, statistical differences were observed when the RMSD of the first 10 ns sets were compared with that for trajectories longer than 50 ns. On the other hand, above 50 ns, the simulations were observed not to be significantly different from one another (Figures 3A and S2). While longer simulation times would be essential to capture the dynamic behaviors during large conformational changes, in the case of a compact globular protein, such as that of enterokinase, it was a reasonable assumption that the nature of the dynamics of the protein could be captured by small conformational changes captured at shorter simulation timeframes.[67]

As an alternative to the statistical evaluation of the RMSD values, PSA was employed to compare the 10 ns simulation chunks to 200 ns. PSA enables a quantifiable similarity metric, Hausdorff distance (δ_H) of the different paths identified during an MD simulation. The distance can be used to measure the similarity between the paths and thus provide a metric of comparison between different simulation lengths. We partitioned the trajectories into separate trajectory files of 10 ns each and determined the PSA Hausdorff distances on the data. The simulation subtrajectories had relatively small Hausdorff values, with a maximum δ_H of 2.16 across the triplicates of the 15 simulation sets (totaling 45 sets) (Figure 3B). This small difference between the simulation segments implies a high similarity between the simulation trajectories based on previous reports where trajectories with path lengths of δ_H < 0.5 Å were denoted as identical and δ_H > 3.0 Å as highly different.[68] The similarity of the time segments for each trajectory indicates the similarity in the features to be extracted from these trajectories. This analysis, in conjunction with the statistical evaluation and the globular and compact nature of the bovine enterokinase enzyme used in this study, provided convincing evidence for us to select the first 10 ns of the simulation trajectory to perform further analysis. Consequently, this would allow a reasonable computational time frame within which the MD data set would be generated.

Degree of Variability between Simulation Repetitions. For the ANOVA study, the RMSDs of the MD trajectories were observed to vary significantly between replicates, which is in accordance with previous reports where low p-values were reported.[69,70] For MD simulations run over the course of a fixed time frame, correlations in biodescriptors presented within a given time segment were previously shown not to cause statistically significant differences, leading to vanishing p-values.[71] In line with this observation, the averaging of the MD biodescriptors over a given time segment was proposed.[29,72] Such an approach was implemented in other studies when working with MD simulation data.[29] However, it should be considered that averaging may possibly offset the inherent variation in the data. In order to assess this further, we performed a 20-fold validation to evaluate if the variation embedded in the simulations was retained upon averaging. 80% of the trajectory data points were randomly shuffled, and the RMSD values were thus calculated. These average values remained statistically significant between trajectories (Figure 3C,D) and thus retained the variability of the trajectories within a protein variant. After the significant variation across different trajectories was maintained upon averaging to retain crucial information that can be utilized by ML algorithms, a decision was made to move forward with the averaged values for the MD features.

Since a significant variation between repeated simulations of the same protein model was shown to persist even after averaging the trajectories, a decision was made to increase the number of repeat simulations from three to five with the aim of capturing the inherent variability of the simulation space adequately and effectively, in line with former practice reported.[70] All models were constructed using the average and standard deviation of the averages of the biodescriptors for five trajectories from this point forward (Figure 3E). This ingestion of MD features was shown to be capable of capturing key features linked to protein function and was shown to be a suitable decision for the purposes of the model-based analysis discussed below.[29]

Selection of a Suitable Machine Learning Model. The final data set of 192 observed variables denoted as biodescriptors was used to evaluate the predictive power of 41 machine-learning models for the 312 variants under investigation (Table S5). We applied an extensive range of learning algorithms, including Ensemble models (Decision

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

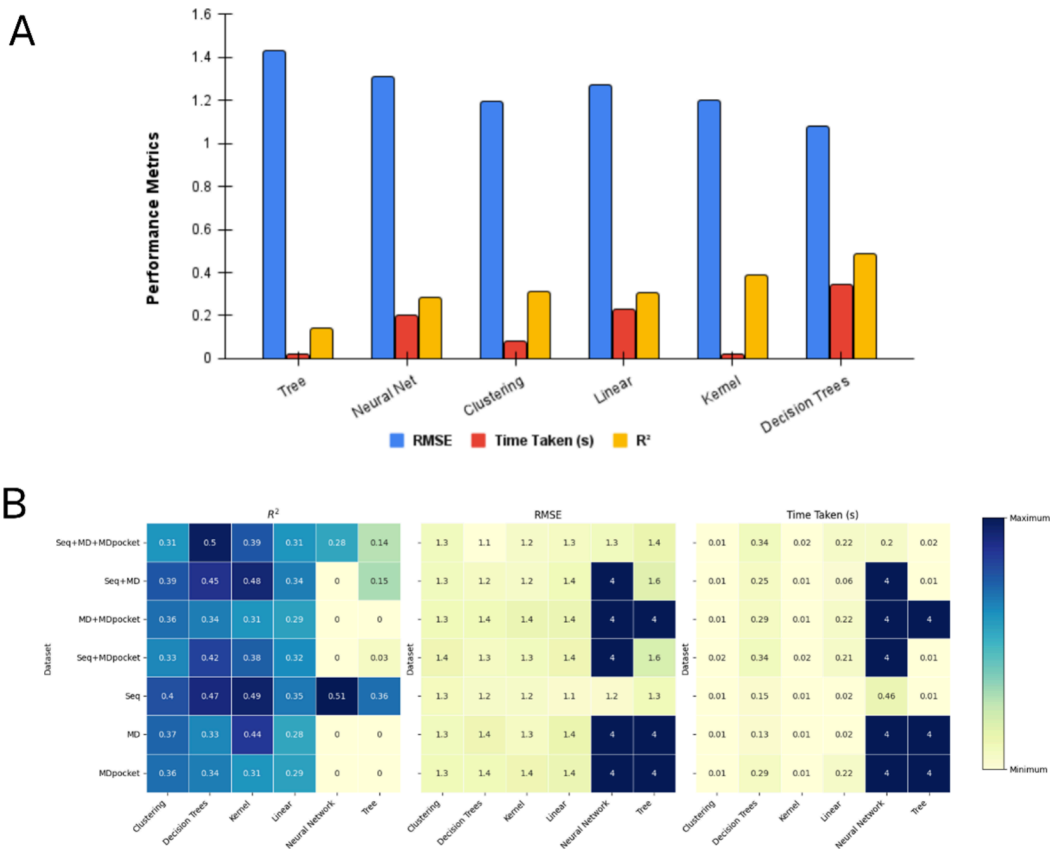


Figure 4. Performance evaluation of the ML models for predicting FCA employing all features from the three biodescriptor categories Seq, MD, and MDpocket. (A) Average performance of the algorithm classes. (B) Heatmap of averaged performance (scaled from 0 to 1) of the algorithm classes when features from specific biodescriptor categories were employed: Seq Seq+MD+MDpocket, Seq+MD, MD+MDpocket, Seq+MDpocket, Seq, MD, and MDpocket.

Trees), Single-Tree-based models, Gaussian Processes, Linear Regression, Clustering Regression, Neural Networks, and Kernel-based approaches. A comprehensive list and description of the models are available in Table S3. We opted to conduct a preliminary screening for model selection using the default model parameters available in scikit-learn or their packages (e.g., XGBoost). This decision to use default parameters to identify the algorithms to be carried forward is in accordance with the consensus that there is no preferable model *a priori*.[73] The models were first ranked and screened according to three metrics: R^2, Root Mean Square Error (RMSE), and time taken in seconds for the execution of the algorithm, as calculated using scikit-learn functions.

Seven of the models tested; RANASCRegressor, Linear Regression, TransformedTargetRegressor, ExtraTreeRegressor, GaussianProcessRegressor, KernelRidge, and PassiveAggressiveRegressor, produced negative values for the coefficient of regression, indicating that these models fit the data worse than a horizontal line would do; in other words, those models were only predicting the average value of the input variables. Consequently, these models were excluded from further consideration. The FCA probability density function manifested a normal-like skewness (0.16) and kurtosis (0.45) in its distribution (Figure S3). The Poisson and Gamma Regression algorithms did not fit the distribution of the response variable, a prerequisite for the models' training, and they were thus excluded from further consideration.

The predictive power of the remaining 32 models was assessed by comparing the predicted and empirical values of the response variable, FCA, for variants of the engineered template form of the bovine enterokinase light chain in the test set. Despite their generally longer runtime compared to other algorithm types, decision trees provide substantially improved model fit with an average unadjusted R^2 score of 0.48. This was 0.10 higher than the R^2 score of the second-best performing class of algorithms, as indicated in Figure 4A. We further performed an analysis for the decomposed data sets that contained features drawn exclusively from the following biodescriptor categories in order to evaluate whether any of the model classes would perform superior for data originating from specific biodescriptor categories: Seq+MD+MDpocket, Seq+MD, MD+MDpocket, Seq+MDpocket, and Seq, MD, and MDpocket. The decision tree-based models yielded the lowest RMSE across all seven data sets as well as ranking high for performance as indicated by R^2 (Figure 4B). This evaluation conclusively showed that the decision tree-based algorithms did not exhibit any bias for modeling a particular group of features while maintaining high algorithm performance.

Among the decision trees tested, the Gradient Boosting Regressor (GBR) generated the resulting models with the lowest RMSE on the Seq+MD+MDpocket data set (Figure S4). Although GBR was praised as a regressor that allowed for a high degree of generalizability and interpretability,[74] GBR's trees are constructed based on the surrogate loss function for minimizing the error of the overall model. As such, this function can only be considered a proxy for the true loss.[75] However, the Light Gradient Boosting Machine (LightGBM) algorithm calculates the second-order derivative of the loss

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

function. This alternative approach was reported to allow for a quick and accurate minimum search of the loss function.[76] LightGBM offers several advantages over other decision trees: (i) its regularization is more complex, thereby preventing overfitting; (ii) it admits sparse features and offers interpretable tools for model analysis; and (iii) it retains performance in high-dimensional data sets.[75,77−79] Lastly, previous reports have shown that LightGBM improved the model's generalizability.[80] In our analysis, the difference between GBR and LightGBM in RMSE was only 3%, and in the current investigation, LightGBM ran noticeably faster than GBR (Figure S4), taking 0.05 compared to 0.38 s. All such advantages led to the decision to select LightGBM to conduct the downstream modeling for predicting FCA in this work. The hyper-parameter tuning of the models was carried out as detailed in the Methods section.

Evaluation of Model Performance. The performance of the models was evaluated based on three metrics: R^2, RMSE, and MAE. RMSE and MAE provided measures of the difference between the observed values and those predicted by the estimator. The evaluation was performed on hyper-parameter-tuned LightGBM decompositions of the data sets iteratively to assess the dependence of the model on specific features. The hyperparameters for each data set are presented in Table S6.

Applying the algorithm to MD and MDPocket features only, we noticed strikingly low R^2 values (Table 1) implying that there was not a strong correlation between these features alone and the predictions made on the performance of the enzyme variant. Since the MD features extracted here are not meaningful unless associated with structural and sequential information, the poor predictive performance would be unsurprising. Furthermore, the limited number of MD features (32 as opposed to 80 sequence-based and structure-based biodescriptors) rendered building models with high predictive capability difficult given the unbalanced number of feature classes in the data set. This hypothesis was further supported and confirmed by the observation that the inclusion of the other two subsets of features together with MD-based biodescriptors remarkably improved prediction accuracy.

Table 1. Key Performance Indicators Evaluating the Predictive Capability of the Models That Used Features from Specific Sets of Bio-Descriptor Categories[a]

Feature Combination	R^2	RMSE	MAE	Approximate time required for data generation per variant
Seq+MD	**0.50**	**1.18**	**0.83**	45 min
Seq	0.47	1.21	0.86	5 s
Seq+MD +MDpocket	0.45	1.24	0.86	2 h
Seq +MDpocket	0.42	1.27	0.86	2 h
MD +MDpocket	0.21	1.48	1.08	2 h
MD	0.22	1.47	1.08	45 min
MDpocket	0.20	1.49	1.16	2 h

[a]Rank of the models that utilized different feature combinations based on the different performance metrics is provided in brackets for the 63 cases of the 312 variants in the test set. In bold: the best-performing feature combination and underlined: the second best.

The models that were trained solely on sequence-based features had the second-best RMSE and MAE values. The predictive accuracy of the models improved when sequence features were used in conjunction with MD features across all metrics assessed here assessed: R^2 (ca. 6.4%), RMSE (ca. 2.5%), and MAE (ca. 3.5%) representing higher predictive capabilities (Table 1). However, such performance differences are small and might be influenced by factors such as the choice of random seed or the limited data set size (312 variants in this study). To address these concerns and assess the model robustness, we performed 500 bootstrap iterations with replacement. There were no significant statistical differences between the Seq and Seq+MD data set across all metrics (Tukey HSD results between Seq and Seq+MD with a threshold of 0.05: R^2 p-value = 0.941, RMSE p-value = 0.9081, MAE p-value = 0.9844) (Figure S5). However, both data sets showed statistical superiority over the remaining four data sets. Given this finding, the Seq+MD model, which integrates sequence- and structure-based features with MD-based features, was selected for subsequent analyses. This decision was driven by the potential for richer insights from the added features that capture the system's dynamics and thus improve the explainability or the interpretability of the models.

Feature Analysis and Feature Importance. Machine learning algorithms were shown to be highly effective in depicting the data landscape they are exposed to.[6,81] This ability to translate a data landscape description into an understanding of the relationships between observable and response variables is of utmost importance in protein engineering, where the connection between protein function and product creation is crucial.[6] Therefore, it is imperative to develop an interpretable system that can use the information gathered by these models. The literature offers various ways to measure the value of a feature, and for this work, we employed two methodologies, permutation feature importance and SHAP analysis.[82]

Permutation feature importance is a highly effective method that plays a pivotal role in identifying the tangible contributions of individual features to the overall performance of a model.[83] This algorithm is designed to break down the complex relationship between various characteristics and their impact on outcomes. By doing so, it provides an accurate and realistic landscape of how each feature contributes to the overall performance of the model. The final model was identified to be heavily dependent upon a specific set of features, including DSSP-derived features such as h-alpha as well as a range of other MD dynamic and sequence-based features from BLOSSUM, VHSE, and zScales (Figure 5A−F). These features were found to be instrumental in predicting FCA. The analysis identified unique connections between features such as the helix composition, the protein's hydrophobicity, and FCA. This discovery can be used as a guiding point for protein engineering research in the future, providing invaluable insights into the development of more effective models. The machine learning modeling of the enzyme variant data showed that the selected model was able to learn from both sequence-based and molecular dynamics-based features. When the model relied solely on MD-based features, it detected a subset of MD features that were different from those used in models that integrated MD descriptors with sequence-based descriptors (Table S7). In the latter scenario, additional features were identified as important, such as the fraction of protein with random coil patterns throughout all of the simulated trajectories. This phenomenon, known as feature domain alteration, illustrated how a single MD-based feature

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

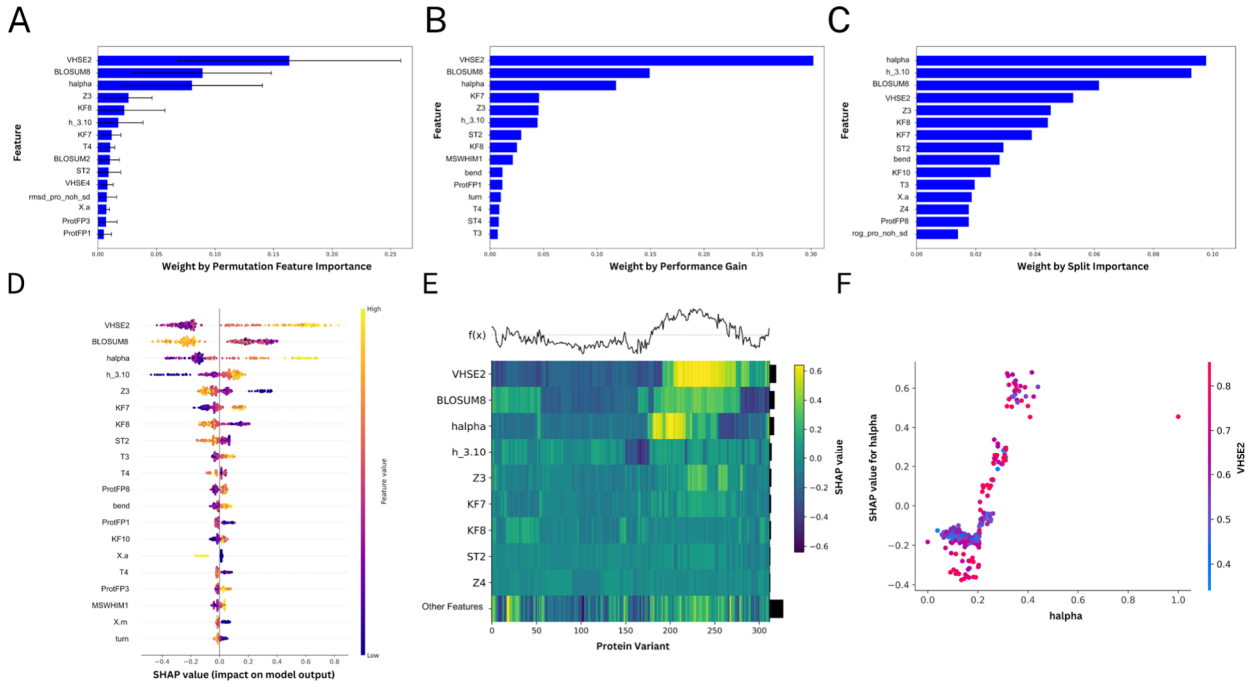


Figure 5. Ranking of features according to the Seq+MD data set using Permutation feature importance and SHAP. (A) Permutation feature importance ranking. (B) Features ranked by the performance gain they provided. (C) Features ranked by the number of times each feature was used to make splits in the data. (D) Feature importance bee swarm plot ranking features according to their impact on model output measured via SHAP. Each dot represents the value of a datapoint in the data set and is colored according to the feature value. The SHAP value related to each datapoint is a measure of how much knowing that datapoint affected the prediction. Negative and positive SHAP values represent, respectively, a decrease and an increase in the predicted value. (E) Top: the function of the model performance ($f(x)$ horizontal line being the true FAC and the oscillating line the predicted value per instance). Bottom: Heatmap of the features according to their SHAP values showing how differently they altered the predictions across the data set. (F) h-alpha dependence plot showing the distribution of the SHAP values against the actual h-alpha values. The coloring is based on a second feature, in this case, VHSE2, with the strongest interaction effect.

can be contextualized when supplemented with sequence-based features as model descriptors. This is because decision trees, which are used by the model employed in this work, can facilitate or disrupt innovative feature interactions as required, allowing different characteristics to be identified as important for accurately representing a data set. It should be noted that the sequence-based descriptors used in this study were not position-based descriptors based on the alignment of the mutant sequences such as one-hot descriptors. One-hot encoding, as a descriptor, could implicitly capture some 3D structure information and thus lead to predictive capability comparable to that of the models that included the MD-based features; they do not universally offer the interpretability we sought after in this study to get meaningful insights for subsequent rounds of protein engineering. By using interpretable features contributed by the MD-data set, we aimed to facilitate a deeper understanding of the underlying mechanisms and guide future engineering efforts through feature importance.

To delve deeper into the impact of different descriptors on the algorithm's decision-making, we modified the ranking criterion from permutation to performance gain and the importance of a feature in the algorithm's decision to split the data set into branches (Figure 5A–C). This in-depth investigation showed that the features that contributed to the formation of decision trees predominantly stemmed from MD features, specifically those related to the binding site, the proteins positioning relative to the backbone and radius of gyration, random coil conformations, and the para- and beta-sheet conformations. Although it is unlikely that the MD characteristics were the primary contributors to the model's numerical performance, as we deduced from the poor predictive performance of models that relied solely on MD-based data (Table 1), they played an important role in the algorithm's decision-making process that ultimately led to a successful prediction of the output response. Consequently, the predictive capability of the models that incorporated MD-based features was greatly enhanced.

After completing the analysis of the model's features using permutation feature importance, we proceeded to employ the SHAP methodology to compare and further evaluate the obtained results. This method was specifically designed to provide an in-depth explanation of individual predictions, offer a comprehensive overview of the model's overall performance, and feature interactions. Notably, over 70% of the features identified as the 20 most influential 20 features contributing to the model's decision-making by permutation feature importance were also identified as the most influential features by SHAP. This discovery confirms the findings through an independent algorithm-based search, as demonstrated by permutation feature importance.

Furthermore, the use of SHAP facilitated the establishment of a sense of concurrence between the response variable and the features, enabling the identification of how these biodescriptors were employed for each prediction. Upon analysis, it was discovered that higher percentages of amino acids under helices conformation, specifically h-alpha and 3.10 helices (h_3.10) (Figure 5D), were found to be related to higher FCA values. Conversely, the prevalence of turn conformations in the protein was demonstrated to have a

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

negative influence on the FCA of the specific variation. This finding suggests that features representing both h-alpha helices and turns conformations were identified as key features, implying that the number of turns may have worked as a control system to reduce the number of reversals in the protein structure. It is worth noting that SHAP enabled the identification of essential aspects in MD that are frequently overlooked as well as the explanation of how these features were related to enzymatic properties, in this case FCA. Overall, this approach provided an in-depth and comprehensive understanding of the relationship between specific features and their impact on the response variable, shedding light on the underlying mechanisms that govern enzymatic properties.

Further, SHAP analysis identified several variables that have an impact on increasing or decreasing the predicted FCA values (Figure 5E). In Figure 5E, it is possible to evaluate each individual prediction and how the features were aggregated to construct the predicted values, as shown by the $f(x)$ showing as a horizontal line the true value and the splines representing the predicted values. Notably, the variants with the largest amount of absolute error are found to be between 200 and 275. Particularly for these sets of variants, VHSE2, BLOSUM8, and h-alpha present large SHAP values, meaning that they contribute extensively to increasing the predicted value of FCA. The magnitude of this impact prompted us to investigate further to see whether there were any potential synergies for the output predictions. Specifically, we examined whether there were any interactions between cognate features with high SHAP values that could affect the accuracy of our predictions. After scrutiny, we found that h-alpha was a potential candidate for such interactions, having strong feature interactions with VHSE2, the second most informative feature (Figure 5F). This finding emphasizes the close correlation between sequence-based and MD-based biodescriptors, highlighting the importance of considering both types of descriptors in predictive models.

Hypergeometric Testing. Hypergeometric testing is a statistical method used to determine the statistical significance of observing a certain number of successes (k) out of a sample size (s) within a larger population size (N), given that the population contains a total number of occurrences (K) with the specific characteristic being studied.[84] In other words, it determines if a sample is random or whether it over- or under-represents a specific population. It is commonly employed in bioinformatics, but it has been applied in other areas of research and analysis.[85]

In the context of model performance analysis, hypergeometric testing can be used to assess the significance of the overlap between the predicted and actual outcomes. It is particularly useful for testing categorization models to assign instances to predefined classes or categories. To facilitate a comprehensive analysis of model performance, we propose a threshold-based approach to assigning labels to predictions. By calculating the threshold at 95% and 5% of the distribution of the difference between the predicted values and the true values, we were able to identify poor predictions beyond these thresholds. This label assignment process enables a fine-grained evaluation of the model performance of a regression task with classification labels.

We utilized this analysis to investigate whether there was any underlying bias created by the number of mutations introduced into the variants that would impact the predictive capability of this modeling pipeline. By constructing a contingency table capturing the counts of observed and predicted outcomes, we calculate the p-value representing the likelihood of observing at least as much overlap as that observed under the null hypothesis of random predictions. A significance threshold of 0.05 was set to determine the statistical significance. In the poor prediction's thresholds, we identified $s = 6$ with $k = 4$ occurrences having the number of co-occurring mutations ≥ 5. According to the hypergeometric distribution testing, the occurrences in the poor thresholds were over-enriched by 3-fold compared to expectations, with a hypergeometric p-value of 0.01. This p-value indicates the statistical significance of the observed overlap between the predicted and actual outcomes. In this context, having the obtained hypergeometric p-value below 0.05 is considered statistically significant. This finding implies that our model has worse performance when predicting the effects of five or more co-occurring mutations, as these are over-enriched observations in the poor predictions. While it was an informative observation, this finding could have been biased since only 24.7% of the data contains more than five mutations, rendering this group under-represented within the data set. On the other hand, the relative scarcity of variants with a high number of mutations could have been caused by the enhancement of undesirable modifications leading to problems in protein folding or function. We further investigated these variants with a high number of mutations and identified that three of the six poor-performing predictions were the sole cases where the predictions were above the 95% threshold and that these variants shared the same four out of five mutations: S38T/L74F/M100K/S127T. As these mutations occurred in only 9% of the variants across the data set, we believe that integrating more data with variants containing large mutation counts could aid in better model performances and provide further assistance to guide improved advice for direct mutagenesis strategies in protein engineering.

CONCLUSIONS

In this work, we present a machine learning-based modeling pipeline that integrates sequence-based and structure-based protein features with dynamics-based features extracted from large-scale data generated by MD simulations to predict protein functionality. The framework addresses some of the current challenges and limitations in protein engineering, particularly those around the identification and prediction of subtle differences between variants of the same protein. We showed that the approach proposed here performed successfully with sufficient predictive power to guide the engineering of novel protein designs and to provide mechanistic insight into the functionality of the protein. The pipeline can allow novel hypotheses to be derived from within a large search space in a feasible time frame to recommend the design of proteins equipped with desirable properties, which can then be realized through site-directed mutagenesis with minimal experimental effort.

Utilizing an information-driven approach to interpret ML-based models, we demonstrated that the information provided by MD-derived, highlighted the essential role protein dynamics plays in predicting function prediction. We highlight the following key attributes of this pipeline development process to guide future efforts in this domain. This study is the first of its kind to integrate replicate MD trajectories into an ML processing pipeline for predicting mutation effects on protein functions, and as such, we were able to highlight some challenges that are inherent in MD data sets. Through

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

statistical analysis, MD simulations were shown to be very noisy, which necessitated dedicated evaluation and preprocessing prior to their integration within the ML pipelines. Model screening highlighted the superior ability of decision tree algorithms to exploit high-dimensional data sets for protein function prediction tasks. Decision trees yielded models that allowed for an interpretable understanding of the predictions.

The relative dominance of alpha helices in protein conformation was identified as a key characteristic impacting the models' predictive performance for the EKB protein. This characteristic, represented by the feature h-alpha, is often overlooked and thus not evaluated during MD analysis. As we demonstrated here, such findings can allow a preassessment of protein engineering features, and using this pipeline, we can identify features from MD that can consequently be used to propose rational designs for protein engineering, possibly substituting residues that favor α helix formation.

One of the most important challenges to the successful implementation of ML-based approaches lies in accessing high-quality data. In the proposed pipeline, MD simulations were coupled with sequences to generate the input data set of protein features. Despite the limited availability of experimental data still being a bottleneck, for MD simulations the outlook is far from bleak, with the increasing availability of protein structures reported through crystallography and more recently in silico methods such as AlphaFold and the development of AI-based MD force fields.[86] Furthermore, advancements in computing power and resources will allow for the extension of the applicability of the proposed protocol.[87] The improvement in the data sets reinforced through MD will inevitably increase the predictive power of approaches such as the pipeline presented here, which will consequently boost efforts toward protein engineering in healthcare applications and in sustainable manufacturing.

■ ASSOCIATED CONTENT

Data Availability Statement

The raw data is available as Supporting Information Tables S1 and S5 at https://zenodo.org/records/10511492. Larger data, such as raw MD simulaitons (>10TB) and ML models, are available upon request. The relevant scripts are publicly shared via GitHub: https://github.com/NAEV95/Engineering_EKB.

Supporting Information

Supporting Information contains The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00999.

Examples of AlphaFold-2 protein structure models, description of ML algorithms and settings, the preprocessed data, and further figures showing ANOVA analysis details and data insights (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Duygu Dikicioglu − *Department of Biochemical Engineering, University College London, WC1E 6BT London, U.K.;* orcid.org/0000-0002-3018-4790; Email: d.dikicioglu@ucl.ac.uk

Authors

Niccolo Alberto Elia Venanzi − *Department of Biochemical Engineering, University College London, WC1E 6BT London, U.K.*

Andrea Basciu − *Department of Physics, University of Cagliari, Cittadella Universitaria, I-09042 Monserrato, Cagliari, Italy*

Attilio Vittorio Vargiu − *Department of Physics, University of Cagliari, Cittadella Universitaria, I-09042 Monserrato, Cagliari, Italy;* orcid.org/0000-0003-4013-8867

Alexandros Kiparissides − *Department of Biochemical Engineering, University College London, WC1E 6BT London, U.K.; Department of Chemical Engineering, Aristotle University of Thessaloniki, 54 124 Thessaloniki, Greece*

Paul A. Dalby − *Department of Biochemical Engineering, University College London, WC1E 6BT London, U.K.;* orcid.org/0000-0002-0980-8167

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00999

Author Contributions

N.A.E.V., A.K., and P.A.D. conceptualized the project. N.A.E.V., supervised by A.B and A.V.V., implemented MD, conducted feature extraction from MD simulations, and built the data set. N.A.E.V. conducted statistical analysis and modeling. D.D. oversaw the project, supervised the statistical analysis and modeling. N.A.E.V. wrote the original draft manuscript. All authors have revised and contributed to the final manuscript.

Funding

This work has been supported by the Engineering and Physical Sciences Research Council (EP/R513143/1). For open access, the authors have applied for a Creative Commons Attribution (CC BY) license for any Author manuscript version arising.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Nishanthi Gangadharan, Cheng Zhang, Bruno Manganelli, and Matthew Banner for useful discussions.

■ REFERENCES

(1) Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009−2018. *Jama* **2020**, *323* (9), 844−853.

(2) Lutz, S.; Iamurri, S. M. Protein Engineering, Methods and Protocols. *Methods Mol. Biol.* **2018**, *1685*, 1−12.

(3) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225* (5232), 563−564.

(4) Orr, H. A. The Distribution of Fitness Effects among Beneficial Mutations in Fisher's Geometric Model of Adaptation. *J. Theor. Biol.* **2006**, *238* (2), 279−285.

(5) Dalby, P. A. Strategy and Success for the Directed Evolution of Enzymes. *Curr Opin Struc Biol* **2011**, *21* (4), 473−480.

(6) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat Rev Mol Cell Bio* **2009**, *10* (12), 866−876.

(7) Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H. Learned Protein Embeddings for Machine Learning. *Bioinformatics* **2018**, *34* (15), 2642−2648.

(8) Yu, H.; Dalby, P. A. Coupled Molecular Dynamics Mediate Long- and Short-Range Epistasis between Mutations That Affect Stability and Aggregation Kinetics. *Proc National Acad Sci* **2018**, *115* (47), E11043−E11052.

(9) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34* (5), 760−769.

(10) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep Learning Enables High-Quality and High-Throughput Prediction of Enzyme Commission Numbers. *Proc National Acad Sci* **2019**, *116* (28), 13996−14001.
(11) Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: A Tool for the Prediction of the Enzymatic Functions of Protein Sequences Based on the EC Nomenclature. *Bmc Bioinformatics* **2018**, *19* (1), 334.
(12) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26* (9), 1169−1175.
(13) Sequeira, A. M.; Lousa, D.; Rocha, M. ProPythia: A Python Package for Protein Classification Based on Machine and Deep Learning. *Neurocomputing* **2022**, *484*, 172.
(14) Lim, H.; Jeon, H.-N.; Lim, S.; Jang, Y.; Kim, T.; Cho, H.; Pan, J.-G.; No, K. T. Evaluation of Protein Descriptors in Computer-Aided Rational Protein Engineering Tasks and Its Application in Property Prediction in SARS-CoV-2 Spike Glycoprotein. *Comput Struct Biotechnology J* **2022**, *20*, 788−798.
(15) Xu, Y.; Verma, D.; Sheridan, R. P.; Liaw, A.; Ma, J.; Marshall, N. M.; McIntosh, J.; Sherer, E. C.; Svetnik, V.; Johnston, J. M. Deep Dive into Machine Learning Models for Protein Engineering. *J Chem Inf Model* **2020**, *60* (6), 2773−2790.
(16) Bhati, A. P.; Wan, S.; Alfe, D.; Clyde, A. R.; Bode, M.; Tan, L.; Titov, M.; Merzky, A.; Turilli, M.; Jha, S.; Highfield, R. R.; Rocchia, W.; Scafuri, N.; Succi, S.; Kranzlmuller, D.; Mathias, G.; Wifling, D.; Donon, Y.; Di Meglio, A.; Vallecorsa, S.; Ma, H.; Trifan, A.; Ramanathan, A.; Brettin, T.; Partin, A.; Xia, F.; Duan, X.; Stevens, R.; Coveney, P. V. Pandemic Drugs at Pandemic Speed: Infrastructure for Accelerating COVID-19 Drug Discovery with Hybrid Machine Learning- and Physics-Based Simulations on High-Performance Computers. *Interface Focus* **2021**, *11* (6), 20210018.
(17) Salmaso, V.; Moro, S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Front Pharmacol* **2018**, *9*, 923.
(18) Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M.; Shaw, D. E. Atomic-Level Characterization of Protein-Protein Association. *Proc National Acad Sci* **2019**, *116* (10), 4244.
(19) Bernardi, R. C.; Cann, I.; Schulten, K. Molecular Dynamics Study of Enhanced Man5B Enzymatic Activity. *Biotechnol Biofuels* **2014**, *7* (1), 83.
(20) Asthana, S.; Shukla, S.; Ruggerone, P.; Vargiu, A. V. Molecular Mechanism of Viral Resistance to a Potent Non-Nucleoside Inhibitor Unveiled by Molecular Simulations. *Biochemistry-us* **2014**, *53* (44), 6941−6953.
(21) Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular Dynamics Simulations: Advances and Applications. *Adv Appl Bioinform Chem Aabc* **2015**, *8*, 37−47.
(22) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr Opin Struc Biol* **2009**, *19* (2), 120−127.
(23) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J Chem Inf Model* **2017**, *57* (4), 726−741.
(24) Van Lommel, R.; Zhao, J.; De Borggraeve, W. M.; De Proft, F.; Alonso, M. Molecular Dynamics Based Descriptors for Predicting Supramolecular Gelation. *Chem Sci* **2020**, *11* (16), 4226−4238.
(25) Wang, D. D.; Ou-Yang, L.; Xie, H.; Zhu, M.; Yan, H. Predicting the Impacts of Mutations on Protein-Ligand Binding Affinity Based on Molecular Dynamics Simulations and Machine Learning Methods. *Comput Struct Biotechnology J* **2020**, *18*, 439−454.
(26) Audagnotto, M.; Czechtizky, W.; De Maria, L.; Kack, H.; Papoian, G.; Tornberg, L.; Tyrchan, C.; Ulander, J. Machine Learning/Molecular Dynamic Protein Structure Prediction Approach to Investigate the Protein Conformational Ensemble. *Sci Rep* **2022**, *12* (1), 10018.
(27) Marchetti, F.; Moroni, E.; Pandini, A.; Colombo, G. Machine Learning Prediction of Allosteric Drug Activity from Molecular Dynamics. *J Phys Chem Lett* **2021**, *12* (15), 3724−3732.
(28) Jandova, Z.; Vargiu, A. V.; Bonvin, A. M. J. J. Native or Non-Native Protein-Protein Docking Models? Molecular Dynamics to the Rescue. *J Chem Theory Comput* **2021**, *17* (9), 5944−5954.
(29) Jamal, S.; Grover, A.; Grover, S. Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease. *Front Pharmacol* **2019**, *10*, 780.
(30) Lee, W.; Pradhan, S.; Zhang, C.; Venanzi, N.; Li, W.; Goldrick, S.; Dalby, P. A. Directed Evolution for Soluble and Active Periplasmic Expression of Bovine Enterokinase in Escherichia Coli. *Research Square*, 2022. .
(31) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: An Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* **2003**, *31* (13), 3381−3385.
(32) Lu, D.; Fütterer, K.; Korolev, S.; Zheng, X.; Tan, K.; Waksman, G.; Sadler, J. E. Crystal Structure of Enteropeptidase Light Chain Complexed with an Analog of the Trypsinogen Activation Peptide. *J. Mol. Biol.* **1999**, *292* (2), 361−373.
(33) Fuetterer, K.; Lu, D.; Sadler, J. E.; Waksman, G. The Serine Protease Domain of Enteropeptidase Bound to Inhibitor VAL-ASP-ASP-ASP-ASP-LYS-Chloromethane. *Protein Data Bank*, 1999. .
(34) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat Methods* **2012**, *9* (2), 173−175.
(35) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. *Bmc Bioinformatics* **2009**, *10* (1), 421.
(36) Studer, G.; Tauriello, G.; Bienert, S.; Biasini, M.; Johner, N.; Schwede, T. ProMod3—A Versatile Homology Modelling Toolbox. *Plos Comput Biol* **2021**, *17* (1), e1008667.
(37) Studer, G.; Rempfer, C.; Waterhouse, A. M.; Gumienny, R.; Haas, J.; Schwede, T. QMEANDisCo - Distance Constraints Applied on Model Quality Estimation. *Bioinformatics* **2020**, *36* (6), 1765−1771.
(38) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(39) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. LDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* **2013**, *29* (21), 2722−2728.
(40) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590−596.
(41) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical PKa Predictions. *J Chem Theory Comput* **2011**, *7* (2), 525−537.
(42) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *Softwarex* **2015**, *1*, 19−25.
(43) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J Chem Theory Comput* **2015**, *11* (7), 3499−3509.
(44) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos,

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J Phys Chem B* **1998**, *102* (18), 3586−3616.
(45) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N · log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.
(46) Johnson, L. W.; Girden, E. R. ANOVA: Repeated Measures. *J Marketing Res* **1995**, *32* (2), 243.
(47) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C J; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Haggstrom, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T; Slavic, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schonberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodriguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kummerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vazquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods* **2020**, *17* (3), 261−272.
(48) Michaud Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32* (10), 2319−2327.
(49) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J* **2015**, *7* (1), 4.
(50) Cruciani, G.; Baroni, M.; Carosati, E.; Clementi, M.; Valigi, R.; Clementi, S. Peptide Studies by Means of Principal Properties of Amino Acids Derived from MIF Descriptors. *J Chemometr* **2004**, *18* (3-4), 146−155.
(51) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. *J Protein Chem* **1985**, *4* (1), 23−55.
(52) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481−2491.
(53) Liang, G.; Li, Z. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *Qsar Comb Sci* **2007**, *26* (6), 754−763.
(54) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.
(55) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, *9* (7), 3084−3095.
(56) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-Source Cavity Detection and Characterization on Molecular Dynamics Trajectories. *Bioinformatics* **2011**, *27* (23), 3276−3285.
(57) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *arXiv* 2012.
(58) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* 2017.
(59) Korobov, M.; Lopuhin, K. *ELI5*. https://eli5.readthedocs.io/en/latest/ (accessed 2022-08-07).
(60) Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T. SWISS-MODEL: Modelling Protein Tertiary and Quaternary Structure Using Evolutionary Information. *Nucleic Acids Res.* **2014**, *42* (W1), W252−W258.
(61) Strodel, B. Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167182.
(62) Lindorff-Larsen, K.; Kragelund, B. B. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167196.
(63) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167208.
(64) Arnittali, M.; Rissanou, A. N.; Harmandaris, V. Structure Of Biomolecules Through Molecular Dynamics Simulations. *Procedia Comput Sci* **2019**, *156*, 69−78.
(65) Peng, J.; Zhang, Z. Simulating Large-Scale Conformational Changes of Proteins by Accelerating Collective Motions Obtained from Principal Component Analysis. *J. Chem. Theory Comput.* **2014**, *10* (8), 3449−3458.
(66) Bruzzese, A.; Dalton, J. A. R.; Giraldo, J. Statistics for the Analysis of Molecular Dynamics Simulations: Providing P Values for Agonist-Dependent GPCR Activation. *Sci Rep-uk* **2020**, *10* (1), 19942.
(67) Damry, A. M.; Mayer, M. M.; Broom, A.; Goto, N. K.; Chica, R. A. Origin of Conformational Dynamics in a Globular Protein. *Commun Biology* **2019**, *2* (1), 433.
(68) Seyler, S. L.; Kumar, A.; Thorpe, M. F.; Beckstein, O. Path Similarity Analysis: A Method for Quantifying Macromolecular Pathways. *Plos Comput Biol* **2015**, *11* (10), e1004568.
(69) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Phys Chem* **1997**, *48* (1), 545−600.
(70) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J Chem Theory Comput* **2018**, *14* (12), 6127−6138.
(71) Farmer, J.; Kanwal, F.; Nikulsin, N.; Tsilimigras, M. C. B.; Jacobs, D. J. Statistical Measures to Quantify Similarity between Molecular Dynamics Simulation Trajectories. *Entropy Basel Switz* **2017**, *19* (12), 646.
(72) Yadava, U.; Gupta, H.; Roychoudhury, M. Stabilization of Microtubules by Taxane Diterpenoids: Insight from Docking and MD Simulations. *J Biol Phys* **2015**, *41* (2), 117−133.
(73) Wolpert, D. H.; Macready, W. G. Coevolutionary Free Lunches. *Ieee T Evolut Comput* **2005**, *9* (6), 721−735.
(74) Konstantinov, A. V.; Utkin, L. V. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowl-based Syst* **2021**, *222*, 106993.
(75) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29* (5), 1189−1232.
(76) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems* **2017**, *30*, 3146−3154.
(77) Konstantinov, A. V.; Utkin, L. V. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowl-based Syst* **2021**, *222*, 106993.
(78) Ransom, K. M.; Nolan, B. T.; Stackelberg, P. E.; Belitz, K.; Fram, M. S. Machine Learning Predictions of Nitrate in Groundwater Used for Drinking Supply in the Conterminous United States. *Sci. Total Environ.* **2022**, *807* (Pt3), 151065.

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(79) Ogunleye, A.; Wang, Q.-G. XGBoost Model for Chronic Kidney Disease Diagnosis. *Ieee Acm Transactions Comput Biology Bioinform* **2020**, *17* (6), 2131−2140.
(80) Anghel, A.; Papandreou, N.; Parnell, T.; Palma, A. D.; Pozidis, H. Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. *arXiv* 2018.
(81) Banner, M.; Alosert, H.; Spencer, C.; Cheeks, M.; Farid, S. S.; Thomas, M.; Goldrick, S. A Decade in Review: Use of Data Analytics within the Biopharmaceutical Sector. *Curr Opin Chem Eng* **2021**, *34*, 100758.
(82) Saarela, M.; Jauhiainen, S. Comparison of Feature Importance Measures as Explanations for Classification Models. *Sn Appl Sci* **2021**, *3* (2), 272.
(83) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26* (10), 1340−1347.
(84) Hwang, F. K.; Song, T. T.; Du, D. Z. Hypergeometric and Generalized Hypergeometric Group Testing. SIAM J. Algebraic Discret. *Methods* **1981**, *2* (4), 426−428.
(85) Simillion, C.; Liechti, R.; Lischer, H. E. L.; Ioannidis, V.; Bruggmann, R. Avoiding the Pitfalls of Gene Set Enrichment Analysis with SetRank. *BMC Bioinform.* **2017**, *18* (1), 151.
(86) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem Rev* **2021**, *121* (16), 10142−10186.
(87) Supinski, B. R. de; Hall, M.; Gamblin, T.; Shaw, D. E.; Adams, P. J.; Azaria, A.; Bank, J. A.; Batson, B.; Bell, A.; Bergdorf, M.; Bhatt, J.; Butts, J. A.; Correia, T.; Dirks, R. M.; Dror, R. O.; Eastwood, M. P.; Edwards, B.; Even, A.; Feldmann, P.; Fenn, M.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Gorlatova, M.; Greskamp, B.; Grossman, J. P.; Gullingsrud, J.; Harper, A.; Hasenplaugh, W.; Heily, M.; Heshmat, B. C.; Hunt, J.; Ierardi, D. J.; Iserovich, L.; Jackson, B. L.; Johnson, N. P.; Kirk, M. M.; Klepeis, J. L.; Kuskin, J. S.; Mackenzie, K. M.; Mader, R. J.; McGowen, R.; McLaughlin, A.; Moraes, M. A.; Nasr, M. H.; Nociolo, L. J.; O'Donnell, L.; Parker, A.; Peticolas, J. L.; Pocina, G.; Predescu, C.; Quan, T.; Salmon, J. K.; Schwink, C.; Shim, K. S.; Siddique, N.; Spengler, J.; Szalay, T.; Tabladillo, R.; Tartler, R.; Taube, A. G.; Theobald, M.; Towles, B.; Vick, W.; Wang, S. C.; Wazlowski, M.; Weingarten, M. J.; Williams, J. M.; Yuh, K. A. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. *Proc Int Conf High Perform Comput Netw Storage Analysis* **2021**, 1−11.

https://doi.org/10.1021/acs.jcim.3c00999
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX