PNAS Nexus, 2024, **3**, pgae538
https://doi.org/10.1093/pnasnexus/pgae538
Advance access publication 26 November 2024
Research Report

Molecular dynamics and machine learning stratify motion-dependent activity profiles of S-layer destabilizing nanobodies

Adam J. Cecil[a], Adrià Sogues[b,c], Mukund Gurumurthi[d], Kaylee S. Lane[e], Han Remaut[b,c] and Alexander J. Pak[a,d,f,*]

[a]Department of Chemical and Biological Engineering, Colorado School of Mines, Golden, CO 80401, USA
[b]Structural and Molecular Microbiology, VIB-VUB Center for Structural Biology, Pleinlaan 2, 1050 Brussels, Belgium
[c]Structural Biology, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[d]Quantitative Biosciences and Engineering Program, Colorado School of Mines, Golden, CO 80401, USA
[e]Computer Science and Software Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN 47803, USA
[f]Materials Science Program, Colorado School of Mines, Golden, CO 80401, USA
*To whom correspondence should be addressed: Email: apak@mines.edu
Edited By: Nikolay Dokholyan

Abstract

Nanobody (Nb)-induced disassembly of surface array protein (Sap) S-layers, a two-dimensional paracrystalline protein lattice from *Bacillus anthracis*, has been presented as a therapeutic intervention for lethal anthrax infections. However, only a subset of existing Nbs with affinity to Sap exhibit depolymerization activity, suggesting that affinity and epitope recognition are not enough to explain inhibitory activity. In this study, we performed all-atom molecular dynamics simulations of each Nb bound to the Sap binding site and trained a collection of machine learning classifiers to predict whether each Nb induces depolymerization. We used feature importance analysis to filter out unnecessary features and engineered remaining features to regularize the feature landscape and encourage learning of the depolymerization mechanism. We find that, while not enforced in training, a gradient-boosting decision tree is able to reproduce the experimental activities of inhibitory Nbs while maintaining high classification accuracy, whereas neural networks were only able to discriminate between classes. Further feature analysis revealed that inhibitory Nbs restrain Sap motions toward an inhibitory conformational state described by domain–domain clamping and induced twisting of domains normal to the lattice plane. We believe these motions drive Sap lattice depolymerization and can be used as design targets for improved Sap-inhibitory Nbs. Finally, we expect our method of study to apply to S-layers that serve as virulence factors in other pathogens, paving the way forward for Nb therapeutics that target depolymerization mechanisms.

Significance Statement

Surface-layer proteins (SLPs) are increasingly being recognized as virulence factors in pathogens that can be dispensable for viability, an attribute desirable for therapeutic targets to mitigate the risk of antimicrobial resistance. As recently shown for *Bacillus anthracis*, prevention of virulence and full clearing of infections can be achieved through nanobody (Nb)-aided SLP depolymerization. To enable the broader development of SLP-targeting Nb therapeutics, it is critical that the mechanism of Nb -induced depolymerization is identified. We explore the utility of machine learning analysis of all-atom molecular dynamics simulations to predict the activity of SLP-binding Nbs. Regularization through model architecture and feature engineering enabled implicit learning of relative Nb inhibitory activities, while feature importance analysis guided the identification of the mechanism for Nb-aided depolymerization.

Introduction

Many bacteria and almost all archaea express SLPs that assemble into paracrystalline arrays known as S-layers (1, 2). *Bacillus anthracis*—the bacterium responsible for lethal anthrax infections—is one such organism. The surface array protein (Sap) is one of the two SLPs identified in *B. anthracis* and is expressed during the exponential phase of growth. After secretion, Sap noncovalently attaches to the peptidoglycan layer via the surface-layer homology motif (3). The assembly domain of Sap (Sap^{AD}) consists of six Ig-like domains connected by flexible linkers. The monomers assemble into a two-dimensional lattice with p2 symmetry (4, 5). While the many functions of such S-layers are still being uncovered, several pathogenic bacteria have been shown to lose their pathogenicity upon knockout of their S-layer genes (2, 4, 6, 7), suggesting that S-layer protein

Competing interests: H.R. is named as an inventor on a patent describing the use of S-layer disrupting nanobodies as antibacterial agents. The other authors declare no competing interests.
Received: August 15, 2024. **Accepted:** November 4, 2024

assembly may be a viable target for therapeutic intervention. One potential avenue for targeting these S-layers is through the use of nanobodies (Nbs), a type of single-domain antibody (4, 8, 9).

Camelid Nbs were recently demonstrated to depolymerize and inhibit the formation of Sap S-layers in *B. anthracis* (4). Infected mouse models treated with Sap-inhibitory Nbs completely cleared their infections. Of the 11 isolated Nbs in this study, five caused lattice depolymerization and inhibited subsequent lattice formation. The other six Nbs bound with high affinity but did not cause lattice depolymerization, suggesting that binding affinity is not the primary driving force for S-layer depolymerization. To date, X-ray structures of only two Sap-binding Nbs have been solved (4), which show that their complementarity determining regions (CDRs), specifically CDR3, are responsible for Sap epitope recognition. However, there are no clear sequence patterns differentiating inhibitory Nbs from noninhibitory Nbs (Table S1). Outside of CDR3, Nb sequences are highly conserved; within CDR3, we observe chemically similar mutations in both inhibitory and noninhibitory Nbs (see Table S2), which make narrowing down specific point-mutations that may be responsible for inhibitory action nontrivial if they even exist. More likely is that mutations leading to inhibitory action are epistatic in nature.

At present, the mechanism of Nb-induced depolymerization is unknown, and experimental methods are unable to access the time- and length-scales necessary to observe this mechanism at molecular resolution (10). While all-atom (AA) molecular dynamics (MD) simulations can provide the needed spatial resolution, the accessible time scales are too short to observe Nb binding and lattice-wide depolymerization, which occurs on the order of minutes (4, 10). Low-resolution coarse-grained (CG) MD simulations could extend the accessible time- and length-scales; however, there is no existing CG model for Sap, and current bottom-up CG methods are ill suited for the configurational diversity of large multidomain proteins (11–13). Instead, we hypothesized that we could take advantage of the hierarchical nature of protein assembly (14, 15) by directing our attention to only the binding region of Sap (i.e. D1D2) and using all-atom molecular dynamics (AAMD) simulations combined with machine learning (ML) to detect differences in small time- and length-scale motions that may lead to collective S-layer disassembly.

Several studies have already attempted to couple MD simulations with ML to elucidate molecular driving forces behind target protein properties (16–21). One study used ML to predict protein antifreeze activity using AAMD data (22, 23). Antifreeze activity was approximated using hydrogen-bond lifetimes of water molecules to solvent-accessible residues to quantify the degree to which the protein restrains water. Another recent study coupled sequence descriptors, structural descriptors, and averaged MD features to predict the activities of a library of enterokinase enzymes (24). By analyzing varying combinations of these features trained using different model architectures, the authors found that the dynamical information encoded by the MD features augmented model predictions when combined with sequence descriptors. Together, these studies show that MD and ML can be coupled to better study protein behavior.

To our knowledge, MD-informed ML has not yet been used to study the functional response of an antigen to libraries of antibodies (Abs) or Nbs bound to the same epitope when competitive inhibition is not likely as a mechanism of action. Existing ML models developed for Abs/Nbs typically attempt to predict structure (25–27) or binding pose (28–30), improve binding affinity and specificity (31–36), or predict properties such as thermostability, toxicity, and nativeness (37, 38). The implicit assumption behind these models is that binding affinity, specificity, and protein stability are the primary drivers of Ab/Nb function. However, these prior approaches are unsuitable in the present context since experiments have already shown that binding affinity and binding pose are incomplete predictors of Nb-induced S-layer depolymerization. To enhance S-layer targeting Nbs for therapeutic use, a new quantifiable objective must be developed that directly correlates to the depolymerization mechanism.

In this work, we set out to design an interpretable predictor for the S-layer depolymerization activity of Nbs that could also be used to identify microscopic correlations related to the mechanism of action for depolymerization. We performed AAMD simulations of isolated D1D2 (the Nb binding site of Sap) with and without a bound Nb across all known inhibitory and noninhibitory Nbs that bind to the epitope. We then trained a series of ML models using D1D2 conformations to predict inhibitory activity and identified the architecture that most closely reproduces experimental observations. Through feature importance analysis and feature engineering, we refined our ML models to enhance learning of inhibitory activities and simplified feature analysis to suggest mechanisms for Sap depolymerization (Fig. 1). Our analysis reveals that the inhibitory activity of Nbs corresponds to Nb-promoted clamping and twisting motions near the binding site that we reason lead to Sap depolymerization. This finding demonstrates the utility of an interpretation framework that leverages both MD simulations and ML analysis to identify functional mechanisms of Nbs based on antigen dynamical responses.

Results

Experimental study of Nb–Sap depolymerization

We first performed an S-layer depolymerization assay to probe which Nbs cause lattice depolymerization and to quantify the depolymerization activity of each Nb. Beginning with polymerized Sap^{AD}, varying concentrations (between 0 and 8 μM) of each inhibitory and noninhibitory Nb (Table S1) were added and allowed to incubate. The mixtures were then subjected to a size-based separation method to isolate monomeric Sap from S-layers, followed by quantification using immunodetection (see “Methods”). The resulting “Sap signal” measures the amount of free Sap monomer, which directly indicates the effective depolymerizing activity of the Nb. Figure 2A compares inhibitory Nbs and shows that Nb692 and Nb702 exhibit the highest depolymerization activity (>60% of Sap depolymerized), Nb683 with intermediate activity (25%), and Nb707 and Nb704 the lowest activity (<13%). As a control, we also included the noninhibitory Nbs which showed no measurable depolymerization activity (Fig. S1). We then solved the X-ray atomic structure of D1D2 bound to Nb692, the most inhibitory Nb from our assay, and Nb694, a noninhibitory Nb that binds to an alternate site on D1 and was previously used as a crystallization aid (4). Figure 2B shows the alignment of the monomeric D1D2 structure when bound to noninhibitory Nb684 (solved in (4, 5)), crystallization aid Nb694, and inhibitory Nb692. The RMSD of D1D2 α-C atoms from Nb694-bound to Nb684- and Nb692-bound is 7.8 and 7.6 Å, respectively. However, RMSD between Nb692- and Nb684-bound is 1.7 Å, revealing that binding of inhibitory Nb692 induces subtle conformational differences in D1D2 compared with binding of noninhibitory Nb684, but not enough to explain inhibitory action.

Identification of Nb binding sites on Sap

Before investigating the inhibitory activity of our Nbs using AAMD simulations, we first sought to determine the binding site for each

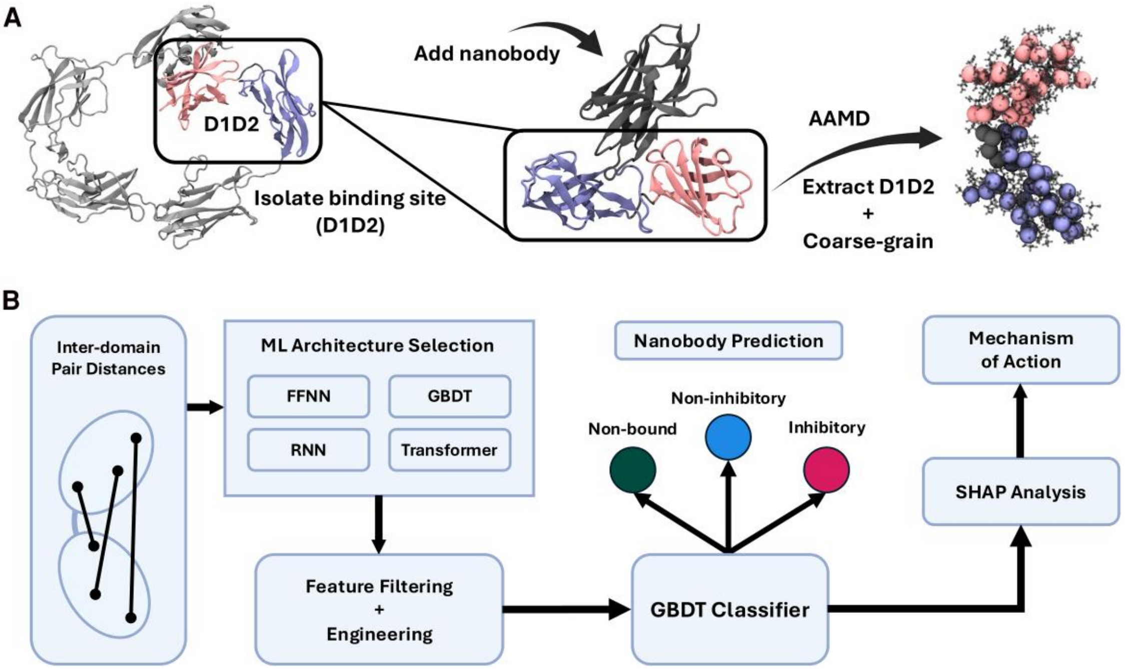

Fig. 1. Overview of the computational workflow. A) Schematic of the AAMD pre- and post-processing steps. The binding site (D1D2) is first isolated from the Sap monomer, each Nb is aligned to the binding position, each system is simulated using AAMD, and D1D2 is extracted from the simulations and CG-mapped to reduce the dimensionality of the system while maintaining dynamic information. B) Schematic of the ML workflow. Processed trajectories are transformed into inter-domain pair distances and used to train a collection of ML classifiers to predict the inhibitory character of a given bound Nb based on D1D2 motions alone. The best model is selected, unimportant features are removed, improved classifiers are trained on the remaining engineered features, then feature analysis is used to pinpoint mechanisms for Nb-induced Sap depolymerization.

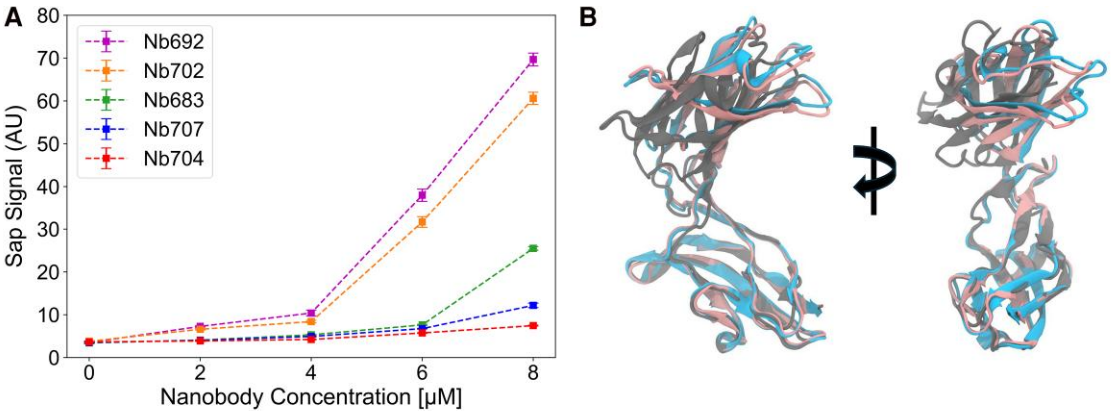

Fig. 2. Experimental characterization of Nb-induced depolymerization. A) Sap depolymerization assay in the presence of each of the listed inhibitory Nbs at varying concentrations; higher Sap signal indicates more depolymerized Sap monomer, and therefore higher Nb activity. B) Overlay of crystallization-aid Nb694-bound D1D2 represented in gray, Nb684-bound D1D2 extracted from Sap monomer form (5) represented in blue, and Nb692-bound D1D2 represented in pink; the D2 domains are aligned to emphasize the change in D1.

of our Nbs. Our solved structure of Nb692-bound D1D2 suggests a common binding site at the hinge between D1 and D2 that is shared by both inhibitory Nb692 and noninhibitory Nb684, which is contrasted by the alternate binding site on D1 by crystallization aid Nb694. We performed multisequence alignment across all of our Nbs and focused on the CDR3 sequence (Table S2). While the majority of Nb sequences were consistent with both Nb692 and Nb684, both Nb703 and Nb704 were notably different enough to suggest an alternate binding site. However, when we aligned Nb703 and Nb704 to the alternate binding site of Nb694 and performed AAMD simulations, we found that both Nbs released within 10 ns of every replica, suggesting that the Nbs do not bind there. Instead, all our Nbs were observed to be stable during AAMD simulations when bound to the Nb692 (and Nb684) binding site (described next), suggesting that the hinge region between D1 and D2 (as depicted in Fig. 1A) is indeed the correct Nb binding site.

ML to classify Nb inhibitory activity

We performed extensive AAMD simulations of D1D2 bound to each of the Nbs to identify conformational motions that might explain inhibitory action. We hypothesized that the activity of these Nbs depended only on the resultant inter-domain conformational motions of D1D2. To describe these motions, we grouped amino

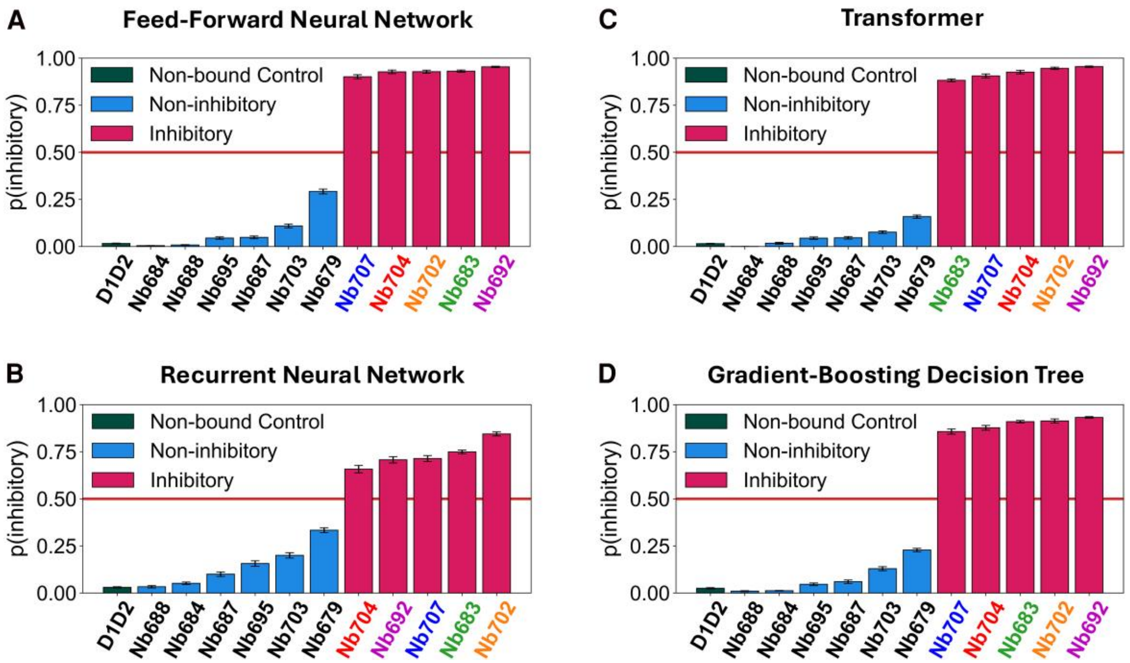

Fig. 3. Ranking of Nb systems across ML architecture. A–D) Probability that a given nanobody (*x*-axis) is inhibitory (*p*(*inhibitory*), *y*-axis), as predicted by a FFNN (A), RNN (B), transformer (C), and GBDT (D) ML models trained on inter-domain pair distances. The horizontal red line in each graph is placed at *p*(*inhibitory*) = 0.5 (50%); predictions higher than 50% are classified as inhibitory, while predictions lower are classified as either noninhibitory or nonbound.

acids with highly correlated motions into CG sites following the Essential Dynamics Coarse-Graining method (39). We then computed pairwise distances between D1 and D2 CG sites for a resultant 900-dimensional dataset. Using this dataset, we tested the use of ML models to predict the two activity classes (inhibitory and noninhibitory) of bound Nbs identified experimentally; as a control, we also included a third nonbound class representing D1D2 without a Nb present. Four ML architectures were tested for their effectiveness in this classification task: a feed-forward neural network (FFNN) for its simplicity, a recurrent neural network (RNN) for its ability to learn correlations across time-sequences, a transformer for its ability to learn relationships across distant features within sequences, and a gradient-boosting decision tree (GBDT) for its interpretability. We evaluated all trained models on both their classification accuracy and their ability to rank inhibitory Nbs according to experimental activity.

Figure 3 presents the average predicted probability that a given Nb-D1D2 system is inhibitory across the four ML architectures. All models showed similar accuracies between 92 and 94% (Table S4) and successful delineation of inhibitory dynamics from both noninhibitory and nonbound dynamics. However, only the GBDT (Fig. 3D) correctly placed inhibitory Nbs close to the experimental order of inhibitory activity, with the predicted ranking following Nb692 > Nb702 and Nb683 > Nb704 > Nb707; recall that the experimental ranking follows Nb692 > Nb702 > Nb683 > Nb707 > Nb704 (Fig. 2A).

Our results suggest that the decision-tree architecture is capable of identifying the correlations driving inhibition. While encouraging, we desired a model whose predictions more closely follow the experimental differences in activity between inhibitory Nbs. Because we are only optimizing discrimination between nonbound, noninhibitory, and inhibitory classes, the ability to correctly rank Nbs within a single class is not strictly guaranteed. However, we believe that there is a common underlying mechanism driving depolymerization that increases in strength with increasing inhibitory activity. Therefore, we expect to see increased sampling of “inhibitory conformations” in high-activity systems. If the classifier correctly identifies those conformations, described by correlations hidden within the input features, experimental ranking could arise naturally as more samples of “inhibitory conformations” improve model confidence.

Given the high-dimensional nature of our dataset, we hypothesized that noisy or redundant features could be impairing the ranking. To identify possible redundant features, we computed Shapley Additive Explanation (SHAP) (40) values that represent the importance of each of the 900 pairwise distance features for the final GBDT predictions. We then trained a series of new GBDTs with a subset of features filtered in decreasing order of importance to find the cutoff between information-dense features and unnecessary features. For each model, we computed a simplified ranking score that represents the deviation between the predicted ranking and the experimental ranking of inhibitory activity; here, a lower ranking score is better, with zero indicating a qualitatively perfect match. As shown in Fig. 4A, we found that the GBDT effectively learned the inhibitory behavior using the top 200 pairwise distance features (in terms of SHAP values). Adding more features did not improve accuracy or decrease the ranking score. While the GBDT model retained the same accuracy of 92% when using only the top 200 pairwise distances as features, the ranking score did not decrease beyond two due to the fact that the predicted ranking for Nb707 and Nb704 (the two least inhibitory Nbs) was consistently flipped with respect to the experimental ranking, shown in Fig. 4B. We therefore sought to improve the GBDT with additional feature engineering.

We reasoned that emphasizing correlated motions within our top 200 pairwise distance features might remove additional noise and further regularize the model to learn inhibitory behavior. We investigated two transformations: principal component analysis (PCA) and time-lagged independent component analysis (TICA). PCA and TICA are linear projections of features that represent

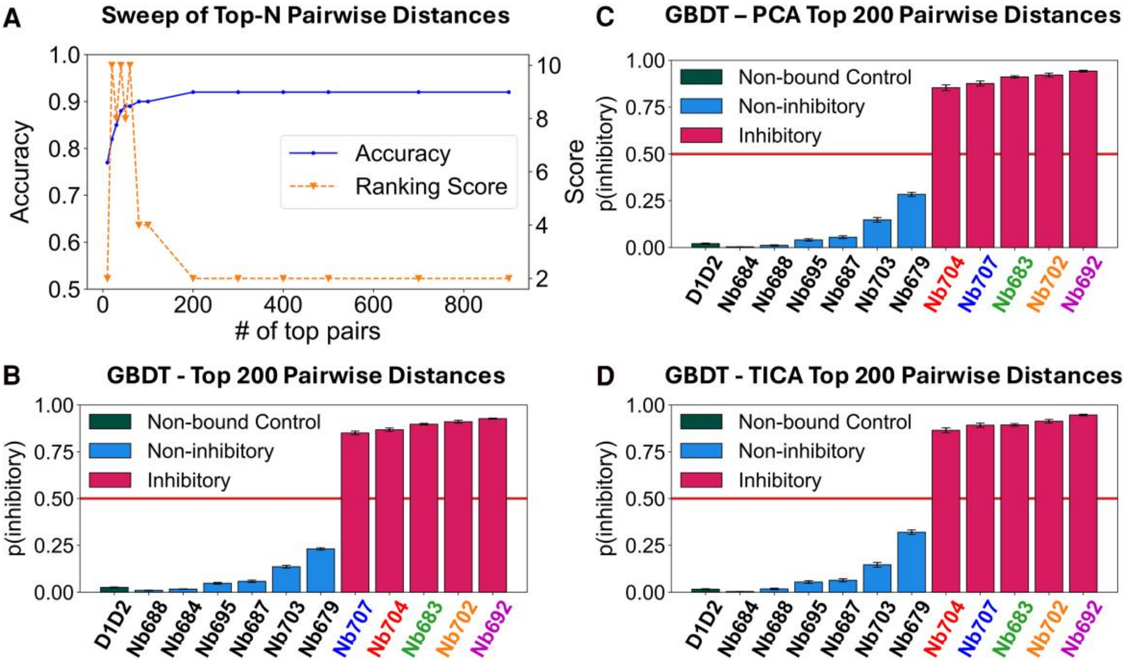

Fig. 4. Feature engineering to improve predictions of inhibitory Nb ranking. A) Accuracy and ranking scores of the GBDTs trained on the listed numbers of top pairwise distance features, as defined by SHAP analysis. B–D) Probability that a given Nb (*x*-axis) is inhibitory (*p*(*inhibitory*), *y*-axis), as predicted by a GBDT model trained on the top 200 pairwise distance features (B), PCA features projected from the top 200 pairwise distance features (C), and TICA features projected from the top 200 pairwise distance features (D). The horizontal red line in each graph is placed at *p*(*inhibitory*) = 0.5 (50%); predictions higher than 50% are classified as inhibitory, while predictions lower are classified as either noninhibitory or nonbound.

the directions of largest variance and autocorrelation, respectively. We trained the PCA and TICA transformations based on the nonbound D1D2 data only, as we expect the unrestrained statistics to include motions characteristic of all three classes. Figure 4C and D show the average predicted probabilities that each Nb is inhibitory using the transformed PCA and TICA features, respectively. In comparison to the previous top 200 pairwise distance GBDT (Fig. 4B), both linear projections improve the ranking prediction to a ranking score of 0 and still retain accuracies of 92%. However, we do observe that the difference between the predicted inhibitory probabilities of Nb707 and Nb683 are not statistically significant in the TICA model unlike in the PCA model, which suggests that the PCA transformation is the most successful at improving the relative ranking of inhibitory Nbs. This analysis shows that by using SHAP analyses and feature engineering to remove unnecessary features and transform information-dense features, we can retain model accuracy while improving its ability to learn the driving force behind the target behavior.

Feature analysis to identify the mechanism of inhibition

With a GBDT architecture and PCA-transformed feature set that aligns excellently with our experimental data, we sought to use this model (hereafter referred to as the top-200-PCA-GBDT model) to identify collective motions that explain the differences in Nb activity. We expect these motions to reveal the mechanism of action that drives depolymerization of the Sap lattice and inhibition of Sap lattice assembly.

First, we investigated class-level feature differences, i.e. delineation of inhibitory, noninhibitory, and nonbound trajectories. In Fig. 5A, we compare potential of mean force (PMF) profiles projected on the first two principal components (PC1 and PC2) from statistics aggregated across each D1D2 trajectory within each class. We find that both the noninhibitory and inhibitory Nb-bound systems exhibit a free energy minimum around PC1 = −21.5 and PC2 = 6.0 that does not exist in the nonbound case. This new free energy minimum is commensurate with the D1D2 X-ray structure when bound to inhibitory Nb692 (the purple triangle in Fig. 5A). There is also a reduction in sampling of other conformations for inhibitory-bound D1D2 than for noninhibitory-bound, suggesting that conformational restriction may play a significant part in the mechanism of depolymerization. Motivated by this insight, we calculated the RMSD of each CG site in D1D2 for each class relative to the X-ray Nb692-bound D1D2 structure. In Fig. 5B, the RMSD distribution across nonbound, noninhibitory, and inhibitory systems indicates that conformational restriction of D1D2 occurs upon Nb binding, with further restriction from inhibitory Nbs, indicated by the heightening of the peak at 1.4 Å (Fig. 5B). This restriction of conformational fluctuations suggests that rigidification of key sites may drive lattice inhibition. However, this analysis alone is not enough to explain the differences in activity across inhibitory Nbs, as separating RMSD statistics by Nb did not show consistent overall rigidification that scales with inhibitory activity (Fig. S3).

Next, we analyzed the PCA-transformed top 200 pairwise distances at the individual Nb level, expecting to see significant localization of sampling within the first few PCs to explain variations in inhibitory activity. However, the spread in conformational sampling at low PCs did not consistently decrease with increasing inhibitory activity (Fig. S4). The Nb with intermediate inhibitory activity (Nb683) appeared to restrict D1D2 the most in this low-dimensional projection, suggesting that the conformational restriction driving lattice inhibition is encoded in collective motions with lower variance observed in free D1D2—motions described by higher PCs.

Because the top-200-PCA-GBDT model is able to correctly place Nb683 in the ranking of inhibitory Nbs with these PCA-transformed features, we performed SHAP analysis on this model to reveal the

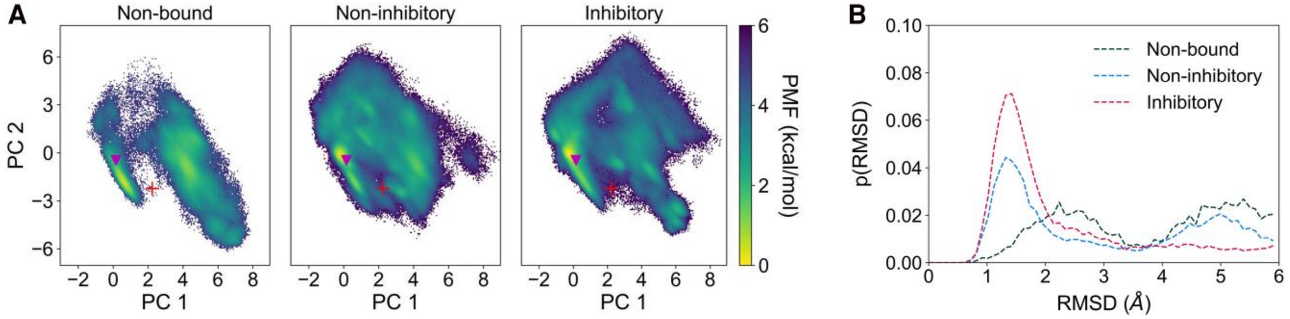

Fig. 5. Configurational differences that stratify the three classes. A) PMF profiles projected onto PCs 1 and 2, separated by class. The PCA model is transformed from the top 200 pairwise distances. The purple triangle indicates the X-ray structure of Nb692-bound D1D2. The red plus (+) indicates the X-ray structure of Nb694-bound D1D2. B) Probability distributions of the RMSD of D1D2 averaged per CG site. The X-ray structure of Nb692-bound D1D2 serves as the reference.

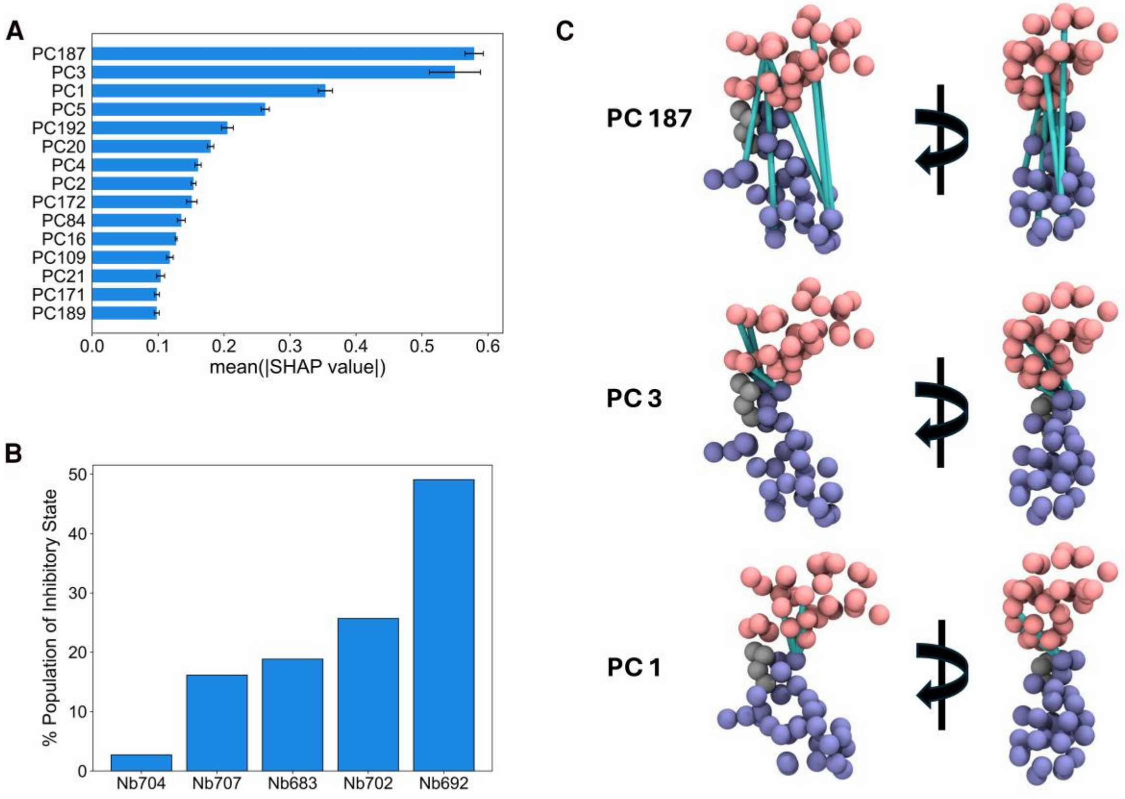

Fig. 6. Analysis of collective D1D2 motions contributing to nanobody inhibitory activity. A) Top 15 PCs identified by SHAP analysis of the top-200-PCA-GBDT model. Larger mean absolute SHAP values indicate higher impact on the model output induced by that feature. Error bars indicate the standard deviation of mean SHAP values across five randomized SHAP analyses. B) Percent sampling of the inhibitory state described by PC187/PC3/PC1 across inhibitory Nbs. C) Depiction of PC187, PC3, and PC1 motions by highlighting the five highest-weighted pair distances for each PC eigenvector (cyan lines) on CG-mapped D1D2 (D1 in pink, D2 in blue, and the linker in gray).

PCs most responsible for the prediction of inhibitory activity. Intriguingly, the top three most important PCs are PC187, PC3, and PC1 (see Fig. 6A). The inclusion of low- and high-PCs suggests that the inhibitory motions of interest are a combination of motions that exhibit low-variance (PC187) and high-variance (PC3 and PC1) in nonbound D1D2. To further explore these important PCs, the Nb-separated inhibitory dataset was clustered using agglomerative clustering in all 200 PCA dimensions and projected onto PC187, PC3, and PC1 (Fig. S9). The resulting clusters were filtered to be within the free energy minimum near the Nb692-D1D2 X-ray structure and have a high average probability of inhibition (>0.9) of the inclusive samples, as predicted by the top-200-PCA-GBDT model. Figure 6B shows the percentage of samples from each inhibitory Nb trajectory found within this conformationally restricted state, revealing a stark increase in restricted state population with increasing inhibition.

Because we are studying such a small library of Nbs, we wanted to ensure as much as possible that our top-200-PCA-GBDT model is not simply overfitting the dynamics of a small set of five inhibitory Nbs despite convergence of its test loss. In the absence of other Nbs to independently test, we reasoned that an ensemble of models trained on subsets of our data may be sufficient to test for generalizability—if the models are only memorizing, we would expect considerable fluctuations in the prediction accuracy and probabilities of inhibition. We trained a series of five subset-top-200-PCA-GBDT models where one noninhibitory Nb (Nb688) and one of each of the inhibitory Nbs was entirely withheld from training, then re-introduced for the final inhibitory predictions. While the predicted probability of inhibition for withheld inhibitory Nbs was <40% in most cases, the experimental ranking of remaining inhibitory Nbs was preserved (Fig. S6). The inability of each subset model to correctly rank Nbs it has not seen does indicate that the GBDT model is at the cusp of being too data-scarce to be generalizable to new Nbs. Yet, because each subset model retains the correct ranking on remaining inhibitory Nbs—which, we reiterate, is not enforced during training—we believe the GBDT model is still robust enough to sufficiently learn the underlying mechanism of inhibition. SHAP analyses on each of these new subset-top-200-PCA-GBDT models show that the set of most important features are largely retained upon removal of Nbs during training (Fig. S7). Therefore, because important features remain effectively unperturbed with removal of training data, we concluded that the learned correlations characteristic of the mechanism of inhibition are indeed generalizable across Nbs despite the small size of our Nb library.

Finally, we extracted CG-site pairs corresponding to the largest magnitude eigenvector components from PC187, PC3, and PC1. Figure 6C illustrates the five top-weighted interdomain pairs for each of the three PCs, highlighted by cyan lines. The placement and direction of these pairs suggest that stratification of motions along these vectors likely results in a complex combination of end-to-end clamping, twisting, and out-of-lattice pulling, driving lattice depolymerization and inhibition of further assembly. Indeed, traversing along PC1 while restraining PC187 and PC3 (Movie S1) shows significant restriction of end-to-end clamping, induction of D1 twisting, and pulling out of the lattice plane. Restraining PC187 and PC1 while traversing along PC3 further confirms end-to-end rigidification and D1 twisting (Movie S2). From this analysis, we conclude that motions characteristic of Nb-induced Sap depolymerization are strongly encoded within the features identified through SHAP analysis of our top-performing GBDT. The inhibitory activity of each Nb is not simply a function of whether it binds to the hinge between D1 and D2. Rather, the efficacy of Nbs as lattice-inhibitors is determined by how well the Nb induces and maintains these inhibitory conformations in D1D2. Finally, our study also demonstrates the utility of using feature importance analysis on well-trained ML models to identify collective motions associated with a target property that is not easily identifiable.

Discussion

This study presents an explainable ML framework that uses AAMD data of the Sap/Nb binding site (i.e. D1D2) to distinguish between inhibitory Nbs that cause Sap S-layer depolymerization and noninhibitory Nbs that have no macroscopic effect on the lattice. Although experiments clearly rank the inhibitory activities of Sap-binding Nbs, there are no obvious patterns in structure or sequence that explain the mechanism driving these differences.

In this work, we leveraged ML classifiers to connect microscopic conformations at the Nb-binding site (e.g. expressed as CG-mapped pairwise distances between domains) directly to the observed efficacy for lattice depolymerization behavior. In order for this model to be useful for uncovering the mechanism of action for depolymerization, the model must first be shown to accurately reproduce experiments. We initially tested simplified descriptors of the binding site conformations (Fig. S2), similar to prior studies that leveraged ML with AAMD data (18, 19, 24, 41). However, we found that higher fidelity descriptors of conformational motions, such as those encoded within inter-domain pairwise distances, were necessary to delineate systems that included inhibitory Nbs from those that did not.

While we observed that successful prediction was agnostic to several different ML architectures, only the GBDT model was able to qualitatively rank the probabilities of inhibitory Nbs similar to their experimental activities even though ranking of inhibitory Nbs was not explicitly enforced during training. Recently, ML models of various architectures have been used to analyze complex biomolecular conformational motions (24, 42, 43), yet relatively few studies have analyzed the benefits of different ML models across one system. One study suggested that neural networks are better suited to learn complex conformational motions compared to simpler models, including decision trees (44), while another argued the opposite (24). At first glance, given the reduction in accuracy from the NNs to the GBDTs (Table S4), our results would agree with the conclusions of the former study. Yet, we suspect that the arguably simpler decision tree architecture serves as a form of regularization that helps the model learn generalizable behavior at the minor expense of accuracy. Along these lines, and in an effort to enhance the inhibitory ranking predictions, we filtered out noisy and unimportant features based on feature importance, simplifying the feature landscape while preserving class-defining correlations. Transforming these pairs into PC vectors further simplified the landscape, allowing the GBDT model to learn the inhibitory ranking behavior with striking similarity to experimental Nb activities. Our findings suggest that a combination of “regularization” through feature and architecture engineering is beneficial when characterizing complex configurational motions on the basis of discrete functional outcomes.

Several assumptions were made that could impact the validity of our conclusions. First, in order to make MD simulations of Nb-bound Sap tractable, we assumed that the isolated Nb-binding site (i.e. D1D2) explores conformational states similarly to the binding site in context with the rest of the monomer and lattice. This assumption is likely reasonable when Sap is monomeric because the domains are connected by flexible linkers and conformationally decoupled from the rest of the monomer (see the atomic model of the monomer (4, 5)). Indeed, comparison of PMFs of AAMD simulations of a whole Sap^{AD} monomer unbound and Nb692-bound reveal that D1D2 does sample remarkably similar conformational states (with probabilities of inhibition of 0.07 ± 0.01 and 0.71 ± 0.04 for non-bound and Nb692-bound, respectively) as simulations when D1D2 is isolated (see Fig. S5), suggesting that analysis of the conformational sampling of isolated D1D2 is warranted. In addition, we have recently resolved an atomic model of the Sap lattice (45) that shows D1, the first domain of the binding site, noncovalently binding to D6 in the lattice, which may affect the conformational sampling of D1 (and D2). However, we performed AAMD simulations of a single Sap^{AD} monomer in its lattice conformation unbound and Nb692-bound, which reveal that the binding of Nb692 destabilizes the D1–D6 interface (Fig. S10), likely due to our observed rigidification of D1D2 induced by inhibitory Nbs. We intend to test whether this behavior leads to Sap^{AD} lattice depolymerization in future CGMD work.

The second major assumption was that features important for GBDT predictions correspond to the physical forces driving inhibition. To minimize the likelihood that important features do not correlate with inhibitory action, it was of critical importance that we first show that our GBDT model reproduces our experimental observations as closely as possible, even though experimental ranking was not explicitly enforced. By comparing SHAP-defined top features (i.e. the most important for prediction) across models with various Nbs withheld from training, we found that top features are relatively conserved, supporting our conclusion that these important features do indeed correlate with inhibitory action. In the future, it may be possible to test the importance of identified pair correlations by cross-linking residues with cross-linking agents of tuned lengths. Instead, we plan to develop CG models of Sap^{AD} (46–48) and deploy lattice-scale CGMD simulations to rigorously test how perturbing the GBDT-identified pair correlations influence lattice stability, assembly, and depolymerization. In parallel, we plan to iteratively use and retrain our GBDT model to engineer new Nbs with enhanced inhibitory activity against Sap and in the process elucidate the connection between Nb sequence and inhibitory activity with a larger curated library of Nbs.

Our primary conclusion is that inhibitory Nbs promote the sampling of a D1D2 conformational state that is rarely observed in Nb-free D1D2 (Fig. S8), with increased inhibitory action associated with increased population. The conformational state is associated with a restriction of clamping and induction of twisting motions acting on D1 at the D1D2 hinge compared to Nb-free D1D2. It is worth discussing if similar motions can be leveraged as a therapeutic target in other S-layers. Within the genus *Bacillus*, S-layer proteins share assembly domains composed of 6 to 8 Ig-like (or β-strand rich) domains despite the low sequence identity across species (4, 7, 49, 50). Other S-layer proteins from *Haloferax volcanii* (51), *Deinococcus radiodurans* (52), *Caulobacter crescentus* (53), and *Clostridium difficile* (6) also contain assembly domains consisting of at least two β-strand rich domains. The prevalence of multiple β-strand rich domains as a common structural motif across known S-layer structures, in which we presume the conformational flexibility between domains is essential for productive lattice assembly, implies that Nb-induced clamping of inter-domain hinges may generalize beyond the Sap S-layer protein from *B. anthracis*. We anticipate that our computational framework that combines high-dimensional AAMD data with explainable ML classifiers will extend to these other S-layer systems, and furthermore, be adaptable to investigate stratified motions in other protein–ligand or antibody–antigen complexes.

Methods

AA molecular dynamics

The atomic structure of D1D2 was isolated from the solved atomic model of Sap^{AD} in the Protein Data Bank (PDB: 6HHU) (5). The Nb Nb684 was included in this structure, and its binding position was used as the template for all subsequent Nbs except for Nb692, which used the corresponding X-ray structure solved in this study. Atomic structures of all remaining Nbs were obtained using homology modeling using Modeller (54). Nbs were aligned to the Nb684 binding position using VMD 1.9.4 (55) by superposing all Cα atoms over their corresponding Cα in Nb684. Overlapping atoms were perturbed using an in-house Python script to reduce steric clashes before energy minimization.

Nb-D1D2 AAMD simulations were run with GROMACS 2021 (56) using the CHARMM36m forcefield (57) and TIP3P water (58) with a timestep of 2 fs. Each system was placed in a box with periodic boundary conditions and a 2 nm buffer distance between the protein complex and all sides of the box. The systems were solvated in water with 150 mM NaCl, then energy-minimized using the steepest descent algorithm with a force tolerance of 500 kJ/mol/nm. A constant NVT (i.e. canonical) equilibration was performed for 5 ns at 300 K using the V-rescale thermostat (59) with a damping constant of 0.1 ps applied to the entire system. Then, a constant NPT (i.e. isobaric-isothermal) equilibration was performed for 1 ns at 300 K using the same thermostat but a damping constant of 0.5 ps for both protein and solvent and held at 1 bar using a Berendsen barostat (60) with a damping constant of 5.0 ps. In both equilibrations, all Cα atoms were restrained with a force constant of 1,000 kJ/mol, allowing the solvent to relax. A 50 ns constant NVT equilibration step was performed restraining D1D2 and nonbinding-end Nb Cα atoms to allow the CDRs to relax into the binding site. Finally, 1 μs constant NVT production simulations were conducted at 300 K with a damping constant of 2.0 ps for six independent replicas of every Nb-D1D2 system and eight replicas for nonbound D1D2. Protein configurations were saved every 10 ps for each trajectory.

Data featurization and ML

Pairwise distance featurization

Prior to featurization, D1D2 was isolated from each AAMD simulation using GROMACS, then mapped into CG resolution using the OpenMSCG python package (61). Each domain was mapped to 30 CG sites using the Essential Dynamics Coarse-Graining method (39), and the linker region was mapped with a 1:1 residue to CG site resolution. The first 200 ns of each trajectory were discarded to remove any nonequilibrated statistics from the training data. After processing, each (Nb-)D1D2 trajectory was transformed into 900-dimensional datasets of inter-domain pair distances using MDTraj (62).

The 200-component PCA model was trained on the top 200 (via SHAP) inter-domain pair distances from all nonbound D1D2 simulations. Prior to training, the distances were transformed using the StandardScaler from Scikit-learn (63), trained on the nonbound dataset. The explained variance per eigenvalue revealed two spectral gaps (the first after PC1 and the second after PC3), with the first three PCs capturing 84% of the total explained variance. The 200-component TICA model was also trained on the same scaled top 200 inter-domain pair distances using DeepTime (64). A lag time of 10 ns was chosen for model training.

Machine learning

The GBDTs were trained built using LGBMClassifier from LightGBM (65). All other models were built using TensorFlow 2 (66). Training data for each model were balanced by Nb class and randomly sampled with replacement to obtain a 60/20/20 split of train, validation, and test samples, resulting in just over 100,000 train samples for each dataset. Three classes were predicted by the models: first that the sample came from a simulation with no Nb, second that there was a noninhibitory Nb bound, and third that there was an inhibitory Nb bound. These classes were one-hot encoded for all models except the GBDT, which were integer encoded. The output of every ML model was the probability that the sample was taken from any one of the three classes. The categorical cross entropy loss was minimized in all cases except for the GBDT, which minimized multi-logloss. Additional details on each of the four ML models are provided in Tables S3 and S7 and as text in the Supplementary Material.

ML model analysis

SHAP analysis

SHAP analyses were conducted using the TreeExplainer from the SHAP python library (67). Here, 5,000 samples were used for the background, and 50,000 samples were used for the SHAP analysis.

Table 1. True positions and scores associated with each inhibitory Nb in order of increasing inhibitory activity.

	Nb704	Nb707	Nb683	Nb702	Nb692
Position	1	2	3	4	5
Score	−2	−1	0	1	2

Each SHAP analysis was run five times, each with different randomized datasets to ensure reproducibility of SHAP values.

Nb ranking

Nb systems were ranked by block-averaging the inhibitory class predictions per frame in blocks of 25 ns to remove autocorrelations, then averaging those blocked samples and calculating SEM using SciPy (68). Ranking score (as reported in Fig. 4A) was calculated using Equation 1:

$$\text{Ranking score} = \sum (\text{predicted position score} - \text{true position score})^2 , \quad (1)$$

where the "position score" is the score associated with a given ranking position (1 to 5) in the order of inhibitory Nbs from lowest to highest activity. The assigned positions and scores are shown in Table 1.

Cloning for recombinant production in *Escherichia coli*

The domains 1 and 2 of *B. anthracis* Sap (from E125 to G384; UniprotKB: P49051) were cloned into a pASK-IBA3C plasmid and Hisx6 N-terminally tagged. We used the synthetic codon-optimized Sap^{AD} previously described in Fioravanti *et al.* (4) as a PCR template using oligos p373 and p374. The DNA fragment was cloned using Gibson assembly into a linearized pASK-IBA3C vector using oligos p321 and p322. For cloning purposes, all the strains were grown in lysogeny broth at 37 °C and supplemented with 100 μg/mL of Ampicillin when required. All plasmids were sequence-verified (Eurofins) using primers p305 and p306. All plasmid and primers used in this study are listed in Table S5.

Protein expression and purification of Nbs and Sap domains

The Sap-binding Nbs have been expressed and purified as previously described (4). Sap D1D2 was expressed in *E. coli* BL21 (DE3) grown in Terrific broth (TB) supplemented with 100 μg/mL of ampicillin at 37 °C and induced with 200 μg/L anhydrotetracycline when OD_{600} reached 0.6. At this point, the temperature was set to 23 °C and cells were left to express overnight. The next day, cells were harvested by centrifugation and pellets were kept at −20 °C. Frozen pellets were resuspended in 100 mL of lysis buffer (50 mM Hepes pH 8, 300 mM NaCl, 1 mM MgCl2, DNase, lysozyme and ethylenediaminetetraacetic acid-free protease inhibitor cocktails [ROCHE]) at 4 °C and lysed by five cycles of 15 s sonication. The lysate was centrifuged for 30 min at 30,000×*g* and 4 °C. The cleared supernatant containing D1D2 was loaded onto a 1 mL Ni-NTA affinity chromatography column (HisTrap FF crude, GE Healthcare). The column was washed with five column volumes with Buffer A (30 mM NaCl, 10 mM Hepes pH 8, 10 mM Imidazole) and eluted with a linear gradient of Buffer B (300 mM NaCl, 10 mM Hepes pH 8, 1 M Imidazole). Eluted protein was concentrated and loaded onto a Superdex 200 16/60 size-exclusion column (SEC) (GE Life Sciences) that was equilibrated with SEC buffer (10 mM Hepes pH 8 and 100 mM NaCl) at 4 °C. The fractions corresponding to the D1D2 protein were concentrated, flash-frozen in small aliquots in liquid nitrogen, and stored at −80 °C. After purification, the samples were run on a sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) to evaluate their purity.

Crystallization, structure determination, and analysis

Crystallization screens were performed using freshly purified Sap D1D2 alone or in combination with nanobody Nb692 or Nb694 using 1.2-fold molar excess and the mixture was concentrated using an AMICON 10 kDa molecular weight cut-off. Optimal crystals of D1D2 alone appeared after 14 days in a condition of the Morpheus kit (Molecular dimensions) containing 0.09 M Halogens, 0.1 M Buffer System 1 6.5 (pH) and 37.5% v/v Precipitant Mix 41 at 144 mg/mL. Crystals of the complex D1D2 with Nb694 appeared after 30 days in a condition of the Midas plus screen (Molecular dimensions) containing 0.2 M Ammonium Chloride, 0.1 M Hepes (pH 7.5), and 25% v/v Glycerol Ethoxylate at 60 mg/mL. Finally, crystals of the D1D2 in complex with Nb692 appeared after 5 days in a condition of the Midas plus kit containing 0.2 M Ammonium chloride, 0.1 M Hepes (pH 7.5) and 25% v/v Glycerol ethoxylate at 50 mg/mL. In all cases, the drop containing the crystals was supplemented with 15% glycerol and the crystals were mounted in nylon loops and flash-cooled in liquid nitrogen. X-ray diffraction data were collected at 100 K using the Beamlines Proxima 2 and Proxima 1 at the Soleil synchrotron (Gif-sur-Yvette, France) and Diamond Light Source (Didcot, UK) on beamline I04. Data were processed with Autoproc (69) and the structure was determined by molecular replacement using phaser from the Phenix suite (70) and the D1D2 and Nbs from PDB 6QX4 as a search model. The structure was refined through iterative cycles of manual model building with COOT (71) and reciprocal space refinement with phenix.refine (72) and Buster (73). The crystallographic statistics are reported in Table S6.

Sap depolymerization assay

To classify the Nbs based on their depolymerization activity, we incubated 5 μM of purified Sap^{AD} with varying concentrations of Nbs (2, 4, 6, and 8 μM) in phosphate buffered saline buffer for 30 min at 25 °C with shaking at 100 rpm. A control sample with Sap^{AD} incubated without Nbs was included to estimate the monomeric Sap. Following incubation, 200 μL of each mixture was placed into a 100 kDa concentrator (Merck Amicon Ultra Centrifugal Filter, #UFC5100) and centrifuged for 5 min at 10,000×*g* to recover 10 μL of the flow-through containing depolymerized (monomeric) Sap. The sample was then diluted 1:10, and 2 μL was spotted onto a nitrocellulose membrane and allowed to dry for 5 min. The membrane was blocked with 5% skimmed milk for 1 h, and then incubated with mouse serum anti-Sap polyclonal antibody (previously characterized in Fioravanti *et al.* (4)) at a 1:1,000 dilution for 1 h with shaking. This was followed by three washes in TBS-Tween buffer (Tris-HCl pH 8, 10 mM; NaCl 150 mM; Tween 20, 0.05% v/v) for 5 min each. As a secondary antibody, we used IRDye 800CW Goat anti-Mouse IgG Secondary Antibody (Licorbio #926-32210) at a 1:10,000 dilution and incubated for 1 h at room temperature with shaking. The washing steps were repeated as described above. Finally, the membrane was developed using Odyssey M Imagers, and the results were analyzed using the Li-cor Empiria Studio Software to measure the intensity of the fluorescent signal (Sap signal AU). The experiments were repeated three independent times.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work was partially supported by the National Institutes of Health through grant R21-AI168838 (A.J.C., M.G., and A.J.P.), the National Science Foundation through grant DMR-2244331 (K.S.L.), and VIB and FWO Flanders through project grant no. G065220N (H.R.). A.S. was supported by post-doctoral fellowships of the European Molecular Biology Organization (EMBO; ALTF-709-2021) and the Marie Skłodowska-Curie Actions (MSCA). This work used high-performance computing resources from the Cyber Infrastructure and Advanced Research Computing Group at Colorado School of Mines and from Anvil at the Rosen Center for Advanced Computing through allocation BIO220015 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which was supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Author Contributions

Adam J. Cecil (Conceptualization, Data curation, Software, Formal Analysis, Investigation, Visualization, Methodology, Writing—original draft, Writing—review & editing), Adrià Sogues (Conceptualization, Data curation, Investigation, Methodology, Writing—review & editing), Mukund Gurumurthid (Software, Formal Analysis, Investigation, Methodology), Kaylee S. Lane (Software, Investigation), Han Remaut and Alexander J. Pak (Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing—review & editing).

Data Availability

The data underlying this study, including simulation files, analysis scripts, and processed data, are openly available at: https://gitlab.com/pak-group/sap_mdml.

References

1 Fagan RP, Fairweather NF. 2014. Biogenesis and functions of bacterial S-layers. *Nat Rev Microbiol.* 12(3):211–222.

2 Ravi J, Fioravanti A. 2021. S-layers: the proteinaceous multifunctional armors of Gram-positive pathogens. *Front Microbiol.* 12: 663468.

3 Kern J, *et al.* 2011. Structure of surface layer homology (SLH) domains from *Bacillus anthracis* surface array protein. *J Biol Chem.* 286(29):26042–26049.

4 Fioravanti A, *et al.* 2019. Structure of S-layer protein Sap reveals a mechanism for therapeutic intervention in anthrax. *Nat Microbiol.* 4(11):1805–1814.

5 Remaut H, Fioravanti A. 2019. *Structure of the Bacillus anthracis Sap S-layer assembly domain*. Protein Data Bank.

6 Lanzoni-Mangutchi P, *et al.* 2022. Structure and assembly of the S-layer in *C. difficile*. *Nat Commun.* 13(1):970.

7 Sogues A, *et al.* 2023. Structure and function of the EA1 surface layer of *Bacillus anthracis*. *Nat Commun.* 14(1):7051.

8 Yang EY, Shah K. 2020. Nanobodies: next generation of cancer diagnostics and therapeutics. *Front Oncol.* 10:1182–1182.

9 Wu Y, Jiang S, Ying T. 2017. Single-domain antibodies as therapeutics against human viral diseases. *Front Immunol.* 8:1802–1802.

10 Frederix PWJM, Patmanidis I, Marrink SJ. 2018. Molecular simulations of self-assembling bio-inspired supramolecular systems and their connection to experiments. *Chem Soc Rev.* 47(10): 3470–3489.

11 Jin J, Pak AJ, Durumeric AEP, Loose TD, Voth GA. 2022. Bottom-up coarse-graining: principles and perspectives. *J Chem Theory Comput.* 18(10):5759–5791.

12 Pak AJ, Voth GA. 2018. Advances in coarse-grained modeling of macromolecular complexes. *Curr Opin Struct Biol.* 52:119–126.

13 Sharp ME, Vázquez FX, Wagner JW, Dannenhoffer-Lafage T, Voth GA. 2019. Multiconfigurational coarse-grained molecular dynamics. *J Chem Theory Comput.* 15(5):3306–3315.

14 Zottig X, Côté-Cyr M, Arpin D, Archambault D, Bourgault S. 2020. Protein supramolecular structures: from self-assembly to nanovaccine design. *Nanomaterials.* 10(5):1008.

15 Rest C, Kandanelli R, Fernández G. 2015. Strategies to create hierarchical self-assembled structures via cooperative non-covalent interactions. *Chem Soc Rev.* 44(8):2573–2573.

16 Abdelbaky I, Tayara H, Chong KT. 2021. Prediction of kinase inhibitors binding modes with machine learning and reduced descriptor sets. *Sci Rep.* 11(1):706.

17 Eshak F, *et al.* 2024. Epitope identification of an mGlu5 receptor nanobody using physics-based molecular modeling and deep learning techniques. *J Chem Inf Model.* 64(11):4436–4461.

18 Jamal S, Grover A, Grover S. 2019. Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against Alzheimer's disease. *Front Pharmacol.* 10:780.

19 Wang F, *et al.* 2019. Machine learning classification model for functional binding modes of TEM-1 β-lactamase. *Front Mol Biosci.* 6:47.

20 Kaptan S, Vattulainen I. 2022. Machine learning in the analysis of biomolecular simulations. *Adv Phys X.* 7(1):2006080.

21 Seshadri K, Krishnan M. 2023. Molecular dynamics and machine learning study of adrenaline dynamics in the binding pocket of GPCR. *J Chem Inf Model.* 63:4291–4300.

22 Kozuch DJ, Stillinger FH, Debenedetti PG. 2020. Genetic algorithm approach for the optimization of protein antifreeze activity using molecular simulations. *J Chem Theory Comput.* 16(12):7866–7873.

23 Kozuch DJ, Stillinger FH, Debenedetti PG. 2018. Combined molecular dynamics and neural network method for predicting protein antifreeze activity. *Proc Natl Acad Sci U S A.* 115(52): 13252–13257.

24 Venanzi NAE, *et al.* 2024. Machine learning integrating protein structure, sequence, and dynamics to predict the enzyme activity of bovine enterokinase variants. *J Chem Inf Model.* 64(7): 2681–2694.

25 Abanades B, *et al.* 2023. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun Biol.* 6(1):575.

26 Ruffolo JA, Sulam J, Gray JJ. 2022. Antibody structure prediction using interpretable deep learning. *Patterns.* 3(2):100406.

27 Ruffolo JA, Chu L-S, Mahajan SP, Gray JJ. 2023. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat Commun.* 14(1):2389.

28 Tubiana J, Schneidman-Duhovny D, Wolfson HJ. 2022. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat Methods.* 19(6):730–739.

29 Krapp LF, Abriata LA, Cortés Rodriguez F, Dal Peraro M. 2023. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun*. 14(1):2175.

30 Khan SH, Tayara H, Chong KT. 2022. ProB-Site: protein binding site prediction using local features. *Cells*. 11(13):2117.

31 Hacisuleyman A, Erman B. 2020. ModiBodies: a computational method for modifying nanobodies in nanobody-antigen complexes to improve binding affinity and specificity. *J Biol Phys*. 46(2):189–208.

32 Tam C, Kumar A, Zhang KYJ. 2021. Nbx: machine learning-guided re-ranking of nanobody-antigen binding poses. *Pharmaceuticals*. 14(10):968.

33 Huang Y, Zhang ZD, Zhou Y. 2022. AbAgIntPre: a deep learning method for predicting antibody-antigen interactions based on sequence information. *Front Immunol*. 13:1053617.

34 Shroff R, *et al*. 2020. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth Biol*. 9(11):2927–2935.

35 Makowski EK, *et al*. 2022. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat Commun*. 13(1):3788.

36 Yuan Y, Chen Q, Mao J, Li G, Pan X. 2023. DG-affinity: predicting antigen-antibody affinity with language models from sequences. *BMC Bioinformatics*. 24(1):430.

37 Ramon A, *et al*. 2024. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nat Mach Intell*. 6(1):74–91.

38 Zhou Y, *et al*. 2023. AB-Amy: machine learning aided amyloidogenic risk prediction of therapeutic antibody light chains. *Antib Ther*. 6(3):147–156.

39 Zhang Z, *et al*. 2008. A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys J*. 95(11):5073–5083.

40 Chen H, Covert IC, Lundberg SM, Lee S-I. 2023. Algorithms to estimate Shapley value feature attributions. *Nat Mach Intell*. 5(6):590–601.

41 Geisel D, Lenz P. 2022. Machine learning classification of trajectories from molecular dynamics simulations of chromosome segregation. *PLoS One*. 17(1):e0262177.

42 Ahalawat N, Sahil M, Mondal J. 2023. Resolving protein conformational plasticity and substrate binding via machine learning. *J Chem Theory Comput*. 19(9):2644–2657.

43 Chen JF, *et al*. 2024. Exploring biased activation characteristics by molecular dynamics simulation and machine learning for the μ-opioid receptor. *Phys Chem Chem Phys*. 26(14):10698–10710.

44 Fleetwood O, Kasimova MA, Westerlund AM, Delemotte L. 2020. Molecular insights from conformational ensembles via machine learning. *Biophys J*. 118(3):765–780.

45 Sogues A, Remaut H. 2024. Architecture of the Sap S-layer of *Bacillus anthracis* revealed by integrative structural biology. *Proc Natl Acad Sci U S A*. Online ahead of print

46 Christians LF, Halingstad EV, Kram E, Okolovitch EM, Pak AJ. 2024. Formalizing coarse-grained representations of anisotropic interactions at multimeric protein interfaces using virtual sites. *J Phys Chem B*. 128(6):1394–1406.

47 Pak AJ, Gupta M, Yeager M, Voth GA. 2022. Inositol hexakisphosphate (IP6) accelerates immature HIV-1 gag protein assembly toward kinetically trapped morphologies. *J Am Chem Soc*. 144(23):10417–10428.

48 Pak AJ, Yu A, Ke Z, Briggs JAG, Voth GA. 2022. Cooperative multivalent receptor binding promotes exposure of the SARS-CoV-2 fusion machinery core. *Nat Commun*. 13(1):1002.

49 Baranova E, *et al*. 2012. SbsB structure and lattice reconstruction unveil Ca2+ triggered S-layer assembly. *Nature*. 487(7405):119–122.

50 Pavkov T, *et al*. 2008. The structure and binding behavior of the bacterial cell surface layer protein SbsC. *Structure*. 16(8):1226–1237.

51 von Kügelgen A, Alva V, Bharat TAM. 2021. Complete atomic structure of a native archaeal cell surface. *Cell Rep*. 37(8):110052.

52 von Kügelgen A, *et al*. 2023. Interdigitated immunoglobulin arrays form the hyperstable surface layer of the extremophilic bacterium *Deinococcus radiodurans*. *Proc Natl Acad Sci U S A*. 120(16):e2215808120.

53 Bharat TAM, *et al*. 2017. Structure of the hexagonal surface layer on *Caulobacter crescentus* cells. *Nat Microbiol*. 2(7):17059.

54 Webb B, Sali A. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinform*. 54:5.6.1–5.6.37.

55 Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph*. 14(1):33–38.

56 Abraham MJ, *et al*. 2015. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 1–2:19–25.

57 Huang J, *et al*. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*. 14(1):71–73.

58 Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 79(2):926–935.

59 Bussi G, Zykova-Timan T, Parrinello M. 2009. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J Chem Phys*. 130(7):074101.

60 Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. 1984. Molecular-dynamics with coupling to an external bath. *J Chem Phys*. 81(8):3684–3690.

61 Peng YX, *et al*. 2023. OpenMSCG: a software tool for bottom-up coarse-graining. *J Phys Chem B*. 127(40):8537–8550.

62 McGibbon RT, *et al*. 2015. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys J*. 109(8):1528–1532.

63 Pedregosa F, *et al*. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 12:2825–2830.

64 Hoffmann M, *et al*. 2022. Deeptime: a Python library for machine learning dynamical models from time series data. *Mach Learn Sci Technol*. 3(1):015009.

65 Ke G, *et al*. 2017. LightGBM: a highly efficient gradient boosting decision tree. Paper presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).

66 Abadi M, *et al*. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467. Available from: https://doi.org/10.48550/arXiv.1603.04467, preprint: not peer reviewed.

67 Lundberg SM, *et al*. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2(1):56–67.

68 Virtanen P, *et al*. 2020. Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17(3):261–272.

69 Vonrhein C, *et al*. 2011. Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr D Biol Crystallogr*. 67:293–302.

70 Bunkóczi G, *et al*. 2013. Phaser.MRage: automated molecular replacement. *Acta Crystallogr D Biol Crystallogr*. 69:2276–2286.

71 Emsley P, Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 60:2126–2132.

72 Afonine PV, *et al*. 2012. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr*. 68:352–367.

73 Smart OS, *et al*. 2012. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr D Biol Crystallogr*. 68:368–380.