Combined molecular dynamics and neural network method for predicting protein antifreeze activity

Daniel J. Kozuch[a], Frank H. Stillinger[b], and Pablo G. Debenedetti[a,1]

[a]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; and [b]Department of Chemistry, Princeton University, Princeton, NJ 08544

Contributed by Pablo G. Debenedetti, October 31, 2018 (sent for review August 30, 2018; reviewed by Ido Braslavsky and Valeria Molinero)

Antifreeze proteins (AFPs) are a diverse class of proteins that depress the kinetically observable freezing point of water. AFPs have been of scientific interest for decades, but the lack of an accurate model for predicting AFP activity has hindered the logical design of novel antifreeze systems. To address this, we perform molecular dynamics simulation for a collection of well-studied AFPs. By analyzing both the dynamic behavior of water near the protein surface and the geometric structure of the protein, we introduce a method that automatically detects the ice binding face of AFPs. From these data, we construct a simple neural network that is capable of quantitatively predicting experimentally observed thermal hysteresis from a trio of relevant physical variables. The model's accuracy is tested against data for 17 known AFPs and 5 non-AFP controls.

proteins | antifreeze | molecular dynamics | neural networks | simulation

Antifreeze proteins (AFPs) have been identified from a variety of sources, including fish, insects, bacteria, plants, and fungi (1). The antifreeze activity of these proteins is characterized by the difference between the nonequilibrium melting and freezing points, referred to as thermal hysteresis (ΔT) (2). ΔT values span a wide range from <1 K for most alanine-rich α-helical AFPs in fish (3, 4) to more than 6 K in hyperactive threonine-rich β-helical proteins found in insects (5). At lower concentrations (<0.5 g/L), hyperactive AFPs greatly outperform more traditional antifreeze agents, making them of potential interest for use in medicine, agriculture, food processing, and surface protection (6).

The most widely accepted theory for the origin of ΔT was put forward by Raymond and DeVries (7) in 1977. They postulated that AFPs first bind irreversibly to the surface of a nascent ice crystal. The ice surface is then forced to adopt an increased curvature as a cap grows between the bound AFPs. This increased surface curvature then depresses the freezing point through the Gibbs–Thomson (Kelvin) effect (8, 9):

$$\Delta T = \alpha_p \left(\frac{\gamma_{sl}\, T_m \nu}{\Delta H_m} \right) \cos\theta / d. \qquad [1]$$

Here, α_p is a geometric constant (two for cylindrical ice cap, four for spherical), γ_{sl} is the ice–liquid surface tension, T_m is the bulk freezing point, ν is the molar volume of ice, ΔH_m is the molar latent heat of fusion, θ is the ice cap contact angle, and d is distance between adsorbed AFPs. This theory has recently been supported via molecular simulation work by Naullage et al. (9), who accurately calculated ΔT from θ and d for a model system. Additionally, Kuiper et al. (10) confirmed that the binding of an AFP to the ice front is nearly irreversible in microsecond-long simulations, agreeing with earlier experimental evidence (11).

However, how does the AFP first recognize and bind to a small quantity of solid water in a vast reservoir of liquid water? Nutt and Smith (12) suggested that AFPs accomplish this feat by preorganizing a "quasi ice-like layer" of water on the ice binding surface (IBS) of the protein. This layer can then be easily incorporated into the growing ice crystal, binding the protein to the solid–liquid interface (12). Recent work by Hudait et al. (13) has shown that water near the IBS is not truly ice like, since its structural order is much lower than ice, but simulations do show that water near the IBS displays exceptionally slower hydrogen bond reorientation dynamics compared with other protein surfaces (14). The presence of slow hydrogen bond dynamics near the IBS was also confirmed experimentally by Meister et al. (15).

Despite a growing body of literature on the topic of AFPs, there has been little progress in successfully engineering new AFPs. Many studies have shown that single mutations often lead to decreased antifreeze activity, with the best-performing mutants often showing little to no advantage over the naturally occurring protein (16–19). In contrast, Marshall et al. (20) showed that the antifreeze activity of a commonly studied AFP, isolated from the spruce budworm beetle, could be enhanced by the addition of coils already found in the AFP. While encouraging, this technique relies on copying an existing structure, and therefore is not a viable route for designing improved antifreeze functionality. Similar efforts have been made by linking two AFPs together, but on a per mass basis, this showed very little improvement (21).

Development of nonbiological thermal hysteresis molecules has been similarly challenging. Synthetic polymers, while often good ice recrystallization inhibition agents, possess a ΔT less than 1 K at relevant concentrations (22). Thermal hysteresis has also recently been observed with a red synthetic dye, Safranin, but its activity was found to be considerably smaller than the best naturally occurring AFPs (23).

Given the growing use of computation in materials design, the availability of a quantitative method for predicting AFP activity in silico would clearly be of interest. Since ice nucleation and

Significance

Antifreeze proteins offer a technologically underutilized approach for controlling the freezing of water, a process intrinsically important in broad areas, such as medicine, agriculture, and food engineering, among others. To harness this capability, a better understanding of the measurable properties involved and their quantitative contribution to the observed antifreeze effect is needed. Here, we present a physically motivated method for the prediction of antifreeze activity purely from simulation, opening routes for the design of computationally optimized antifreeze materials.

Author contributions: D.J.K. and P.G.D. designed research; D.J.K. performed research; D.J.K., F.H.S., and P.G.D. analyzed data; and D.J.K. wrote the paper with assistance from F.H.S. and P.G.D.

Reviewers: I.B., The Hebrew University of Jerusalem; and V.M., The University of Utah.

The authors declare no conflict of interest.

Published under the PNAS license.

[1]To whom correspondence should be addressed. Email: pdebene@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814945115/-/DCSupplemental.

Published online December 7, 2018.

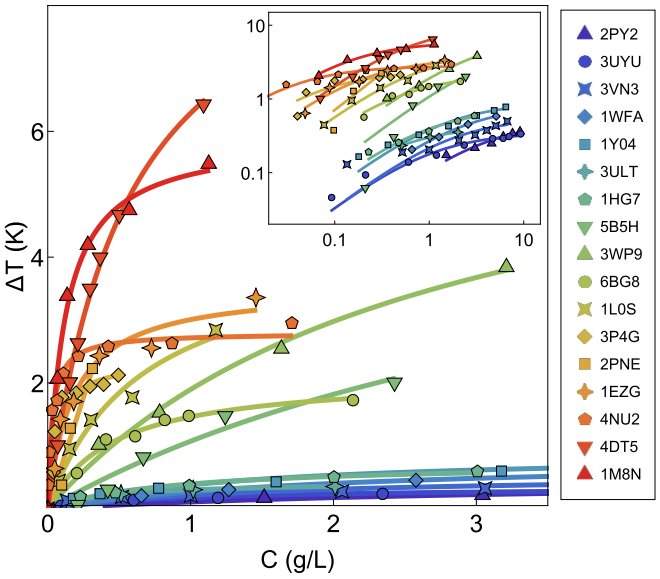

Fig. 1. Experimental thermal hysteresis (ΔT) as a function of mass concentration for 17 selected AFPs collected from the literature. Proteins are referred to by their PDB ID codes: 2PY2 from ref. 25; 3UYU from ref. 26; 3VN3 and 5B5H from ref. 18; 1Y04, 1WFA, 1HG7, and 3P4G from ref. 27; 3ULT from ref. 28; 3WP9 from ref. 29; 6BG8 from ref. 30; 1L0S from ref. 31; 2PNE from ref. 32; 1EZG from ref. 20; 4NU2 from ref. 33; 4DT5 from ref. 5; and 1M8N from ref. 34. Solid lines are fits to the data. *Inset* shows the full dataset in log–log representation.

freezing occur on time scales much too long for efficient simulation of molecular models of water, this task must be done in an indirect way. More than 15 y ago, Graether et al. (17) attempted to predict the antifreeze activity resulting from single mutations using a neural network (NN) approach. While the strategy was sound, their analysis was limited to the static properties of a single protein, yielding results that were not particularly accurate. A more general approach was undertaken by Doxey et al. (24), who managed to successfully separate AFP structures from non-AFP structures using the fraction of ordered surface carbons, but their method was only intended for classification and actually "scored" rather low-activity AFPs much higher than hyperactive AFPs.

To address this lack of an accurate and transferable method for predicting AFP activity in silico, we use molecular dynamics simulations to analyze the geometric properties of AFPs and the dynamics of the surrounding water. From these data, we construct both a linear model (LM) and an NN, and we demonstrate that the NN is able to accurately predict the activity of a wide variety of AFPs. Our results identify key parameters necessary for antifreeze activity and should be of use in designing antifreeze materials.

Results and Discussion

Experimental Data. For this study, we attempt to include all available nonmutant AFPs that have both a structure deposited in the Protein Data Bank (PDB) (www.rcsb.org) and a published thermal hysteresis curve as a function of concentration. We also avoid any multisubunit or linked AFPs. As a result, we selected 17 proteins with a spectrum of antifreeze activities along with 3 non-AFPs containing exposed planar faces and 2 homologs (proteins with similar structure to AFPs but no known antifreeze activity) to be used as a control. The AFPs come from a wide variety of organisms (fish, insects, plants, bacteria, and fungi) and cover all basic classes of AFP (e.g., types I–III and hyperactive) (22).

To avoid confusion with different isoforms, we reference the proteins by their PDB ID codes. Experimental data for ΔT were collected from the literature and are shown in Fig. 1. We plot ΔT as a function of mass concentration to ensure equal treatment between AFPs of different sizes.

Assuming a square lattice of adsorbed AFPs, Raymond and DeVries (7) suggested that the data could be fit with a function of the form

$$\Delta T = \alpha C^{1/2}, \qquad [2]$$

where α is a constant and C is the bulk concentration of AFP. While this form is acceptable for low-activity AFPs, it fits the data rather poorly for hyperactive AFPs. The same is true for the kinetic pinning model suggested by Sander and Tkachenko

Table 1. Data for all 22 proteins used in this work

PDB ID code	Source	M_p (kDa)	N	N_B	H (nm)	D (nm)	R_g (nm)	A (nm^2)	L_B (ps)	L_N (ps)	LM ΔT_C (K)	NN ΔT_C (K)	Exp. ΔT_C (K)
3CM4*	Non-AFP	7.50	60	8.00 ± 0.00	1.79 ± 0.16	1.84 ± 0.41	1.08 ± 0.01	2.05 ± 0.66	97.6 ± 5.55	98.1 ± 1.82	−0.08 ± 0.22	0.06 ± 0.12	0.00
3WHD†	Non-AFP	18.3	137	8.57 ± 0.53	3.03 ± 0.18	1.28 ± 0.21	1.48 ± 0.01	2.63 ± 0.35	111. ± 6.08	112. ± 2.42	−0.02 ± 0.13	0.00 ± 0.00	0.00
1AKI	Non-AFP	14.3	129	8.43 ± 0.53	2.82 ± 0.05	2.36 ± 0.29	1.42 ± 0.01	2.68 ± 0.25	103. ± 4.17	98.8 ± 1.52	0.30 ± 0.13	0.11 ± 0.06	0.00
1UBQ	Non-AFP	8.58	76	9.43 ± 1.51	2.01 ± 0.06	2.07 ± 0.45	1.17 ± 0.01	2.53 ± 0.65	117. ± 5.89	108. ± 2.76	0.27 ± 0.21	−0.00 ± 0.00	0.00
3Q6L	Non-AFP	16.9	131	9.57 ± 1.81	2.73 ± 0.34	2.17 ± 0.84	1.42 ± 0.00	2.98 ± 0.86	117. ± 2.61	126. ± 4.85	−0.25 ± 0.42	0.04 ± 0.02	0.00
2PY2	Fish	15.5	127	9.67 ± 1.63	2.70 ± 0.14	2.16 ± 0.27	1.36 ± 0.00	3.68 ± 0.86	125. ± 9.94	122. ± 2.40	0.49 ± 0.60	0.05 ± 0.10	0.03
3UYU	Fungus	25.2	231	9.14 ± 0.90	4.44 ± 0.70	2.28 ± 0.40	1.73 ± 0.01	2.62 ± 0.48	122. ± 2.40	122. ± 2.07	−0.08 ± 0.23	0.07 ± 0.03	0.08
3VN3	Fungus	22.5	222	12.4 ± 4.50	3.26 ± 0.45	1.60 ± 0.91	1.69 ± 0.00	2.82 ± 1.58	135. ± 5.79	125. ± 1.78	0.28 ± 0.69	0.34 ± 0.07	0.08
1WFA	Fish	3.24	37	12.7 ± 1.50	0.65 ± 0.13	1.05 ± 0.32	1.66 ± 0.01	3.23 ± 0.44	125. ± 6.67	115. ± 3.04	0.58 ± 0.29	0.09 ± 0.13	0.10
1Y04	Fish	3.09	35	11.7 ± 2.07	0.79 ± 0.23	1.05 ± 0.17	1.56 ± 0.04	3.14 ± 0.52	126. ± 8.88	105. ± 3.60	0.98 ± 0.42	0.32 ± 0.36	0.16
3ULT	Plant	27.6	114	12.4 ± 3.10	1.72 ± 0.02	1.68 ± 0.41	1.36 ± 0.00	3.37 ± 1.21	121. ± 11.3	92.6 ± 1.46	1.44 ± 0.24	0.98 ± 0.15	0.17
1HG7	Fish	7.14	66	8.14 ± 0.37	2.23 ± 0.16	1.67 ± 0.22	1.09 ± 0.01	1.96 ± 0.35	137. ± 18.8	122. ± 6.10	0.09 ± 0.68	0.33 ± 0.14	0.18
5B5H	Fungus	22.4	223	9.57 ± 2.51	3.32 ± 0.62	1.91 ± 0.58	1.70 ± 0.00	2.74 ± 0.92	138. ± 5.25	122. ± 1.90	0.47 ± 0.44	0.40 ± 0.10	0.36
3WP9	Bacteria	23.5	224	15.7 ± 1.60	3.32 ± 0.10	3.32 ± 0.20	1.67 ± 0.00	4.37 ± 0.38	145. ± 4.48	120. ± 3.00	1.54 ± 0.29	0.94 ± 0.30	0.70
6BG8	Bacteria	26.3	239	8.86 ± 1.07	3.52 ± 0.07	1.41 ± 0.22	1.75 ± 0.00	2.11 ± 0.26	140. ± 4.77	128. ± 2.29	0.00 ± 0.25	0.52 ± 0.07	0.84
1L0S	Insect	9.07	88	12.0 ± 1.73	2.16 ± 0.04	1.57 ± 0.26	1.14 ± 0.00	3.07 ± 0.56	133. ± 3.30	101. ± 1.58	1.36 ± 0.19	1.18 ± 0.25	1.34
3P4G	Bacteria	33.9	301	21.0 ± 3.61	2.86 ± 0.07	1.53 ± 0.41	2.40 ± 0.01	5.29 ± 1.44	163. ± 7.75	127. ± 1.83	2.26 ± 0.45	1.77 ± 0.32	1.98
2PNE	Insect	6.49	81	15.3 ± 2.75	1.05 ± 0.09	1.54 ± 0.26	1.38 ± 0.01	4.38 ± 0.65	149. ± 6.23	106. ± 4.58	2.24 ± 0.57	2.36 ± 0.85	2.13
1EZG	Insect	8.39	82	15.0 ± 2.08	1.46 ± 0.02	1.57 ± 0.28	1.17 ± 0.00	3.73 ± 0.48	107. ± 3.57	89.8 ± 1.51	1.28 ± 0.27	1.77 ± 0.24	2.22
4NU2	Bacteria	25.5	216	17.7 ± 1.50	3.16 ± 0.07	2.97 ± 0.56	1.65 ± 0.00	5.24 ± 0.35	136. ± 4.68	114. ± 1.59	1.91 ± 0.15	2.19 ± 0.13	2.54
4DT5	Insect	14.5	143	21.7 ± 2.63	1.43 ± 0.05	3.78 ± 0.44	1.56 ± 0.01	6.48 ± 0.75	136. ± 6.39	108. ± 3.06	2.71 ± 0.38	3.37 ± 0.49	3.43
1M8N	Insect	12.5	120	15.3 ± 0.75	2.18 ± 0.02	2.24 ± 0.32	1.34 ± 0.00	5.04 ± 0.54	150. ± 4.82	96.6 ± 1.13	2.96 ± 0.20	4.06 ± 0.23	4.28

M_p is the molecular mass of the protein. N is the number of solvent-accessible residues in the protein (defined in *Materials and Methods*), and N_B is the number of residues in the IBS. Height, H, is the maximum distance from the ice binding plane S (defined in the text) to the geometric center of any residue in the protein measured in the direction normal to S. The width of the IBS, D, is the minimum edge length of a rectangle that bounds all coordinates in S. R_g is the radius of gyration measured using the GROMACS 2016.4 tool gmx_polystat. A, L_B, and L_N (defined in the text) are computed from molecular simulation. LM ΔT_C and NN ΔT_C are predicted values using the LM and the NN, respectively. Exp. ΔT_C at 0.3 g/L was interpolated from experimental data using Eq. **3**. All errors are one SD.

*Only the C-terminal domain of 3CM4 was used for the study.

† Residue numbers 61–68 were excluded from PDB ID code 3WHD, as they are not part of the main backbone.

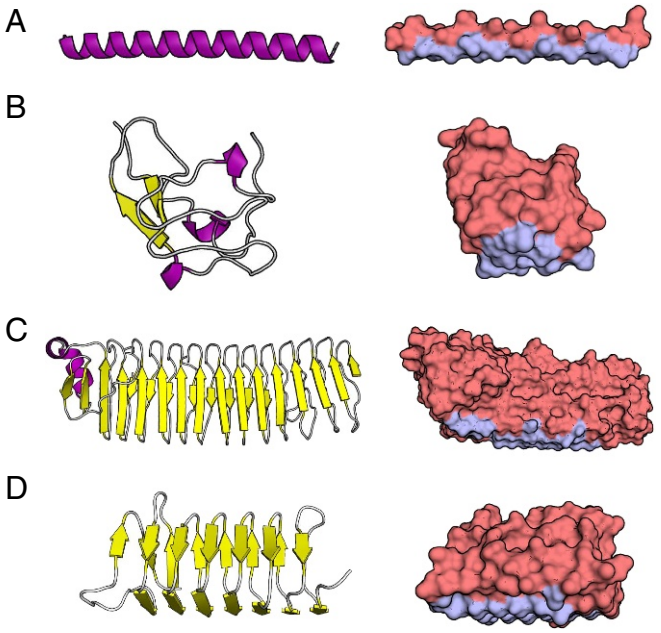

Fig. 2. Secondary/tertiary structure (*Left*) and surface structure with the IBS shown in light blue (*Right*). (*A*) Protein ID code 1WFA. Type I α-helical fish protein. Predicted IBS matches that identified by ref. 3. (*B*) Protein ID code 1HG7. Type III globular fish protein. Predicted IBS matches that identified by ref. 16. (*C*) Protein ID code 3P4G. Ca^{2+}-dependent bacterial protein. Predicted IBS matches that identified by ref. 39. (*D*) Protein ID code 1M8N. Hyperactive AFP from the spruce budworm beetle. Predicted IBS matches that determined by ref. 40. Visualization done with PyMOL (41).

(35). Since a fitting function is still desired to avoid statistical error in the experimental data, we use an expression based on the Langmuir adsorption model (36):

$$\Delta T = \alpha \left(\frac{KC}{1+KC} \right), \qquad [3]$$

where K is a constant. This model fits the data for all AFPs much better (solid lines in Fig. 1), but we emphasize that this is a phenomenological choice, not a rigorous derivation in this context. Additionally, our treatment (as well as Eq. **2**) ignores the interesting sigmoidal behavior of AFP activity noted at extremely low concentrations (37). We consider this approximation quantitatively sufficient for all relevant concentrations and use it to compare ΔT at a selected concentration.

Which Properties Characterize Antifreeze Activity? To construct a predictive model, input variables must first be selected. We begin by assuming that a given AFP is characterized by two surfaces: the planar IBS, which faces into the ice crystal after binding, and the nonplanar, nonice-binding surface (NBS), which faces into liquid water and prevents additional ice growth. A detailed justification for this assumption is given by Knight and Wierzbicki (38). Next, we quantify the slow hydrogen bond dynamics (mentioned above) using a hydrogen bond lifetime, L (defined below). Three variables were then chosen for our analysis.

i) A: the area of the predicted IBS.
ii) L_B: L measured near the IBS.
iii) L_N: L measured near the NBS.

The logic for selecting each parameter is as follows. (*i*) The larger the area of the IBS, the more likely that contact with the ice will lead to binding and the greater the surface coverage will be. (*ii*) Slow hydrogen bond dynamics near the IBS has long been recognized as a signature of AFPs, and therefore, we hypothesize that AFPs with larger L_B will be better able to recognize/bind ice. (*iii*) By the same argument, if L_N is large, this indicates that the protein will be more susceptible to ice overgrowth on the NBS, destroying the antifreeze effect. Therefore, we expect ΔT to be positively correlated with A and L_B and negatively correlated with L_N.

Work by Naullage et al. (9) suggests that bulkier AFPs should display increased ΔT. We, therefore, also tested the inclusion of the following variables: protein height, H; distance across the binding plane, D; molecular mass of the protein, M_p; and radius of gyration, R_g. The inclusion of these variables in the models discussed below did not significantly increase prediction accuracy and led to overfitting. As such, they are not considered in the following work. This is not to say that the bulkiness of the protein is irrelevant to antifreeze activity. It could be that A captures this property appropriately or that bulkiness is less important at fixed mass concentration. Nevertheless, we provide the values of the variables in Table 1 in case this information could prove helpful to the reader.

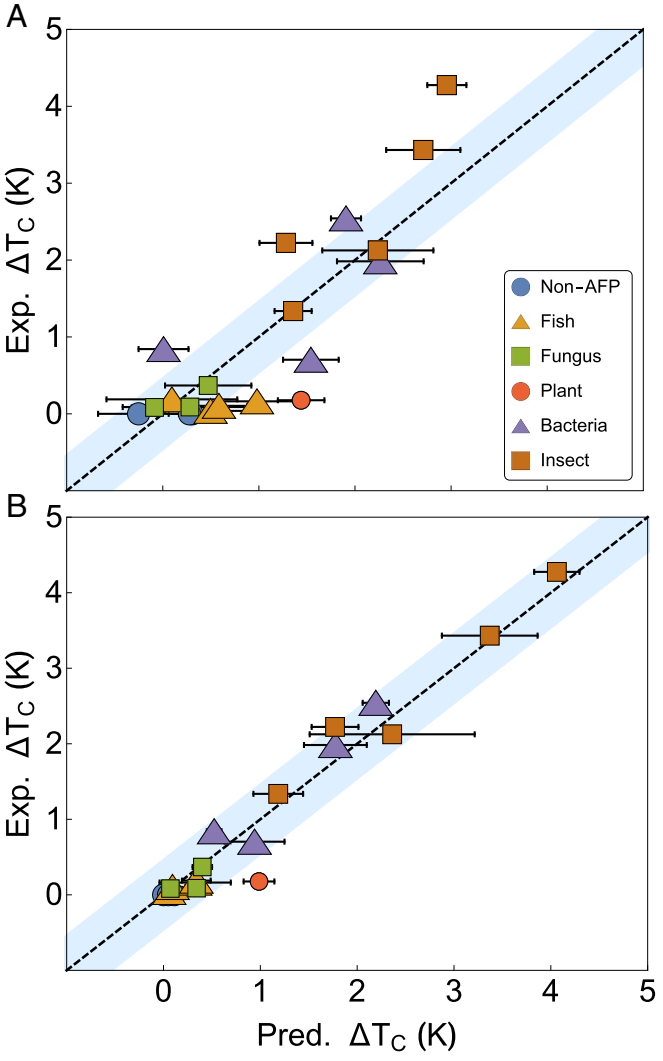

Fig. 3. ΔT_C values predicted (Pred.) by (*A*) LM and (*B*) NN compared with values from experimental (Exp.) data in Table 1. Dashed lines are $y = x$, and the blue regions represent an error of ± 0.5 K. Error bars represent one SD in the predicted value. Error is a result of uncertainty in the measured variables due to thermal motion.

Automated Detection of the IBS. The determination of the IBS is critical to our analysis. Since we intend our method for materials design, we keep this calculation as general as possible, selecting the IBS based on geometric requirements and the optimization of our three physical variables as described next.

For each protein, standard molecular dynamics simulations were performed as described in *Materials and Methods*. From the simulation data, an average protein structure composed of residue coordinates is generated, and a water–water hydrogen bond lifetime, L_i, is assigned to each residue (the procedure is described in *Materials and Methods*). The residues are then grouped into planes using the following procedure.

i) Using any three noncollinear residues, define a plane P.
ii) Identify a set S of all residues within 0.11 nm of P.
iii) If S contains eight or more residues and all residues have at least two neighbors in S within 1 nm, continue. Otherwise, discard the set. This ensures a connected set with a relevant number of residues.
iv) Calculate the number of residues not in S found above or below P, defined as n_1 and n_2. If $\min(n_1, n_2) = 0$, continue. Otherwise, discard the set. This effectively eliminates all planes "inside" the protein structure.
v) Add set S to the set of possible IBSs, S_{all}.
vi) Repeat for all possible sets of three noncollinear residues.

For each S in S_{all}, we then calculate $L_{S,B}$, the average of L_i for all residues in S weighted by the solvent-accessible surface area (SASA) of each residue, and $L_{S,N}$, being the same as $L_{S,B}$ but for all residues not in S. We also calculate A_S, which is defined as the area enclosed by the convex envelope containing the projection of all of the solvent-accessible atoms corresponding to the points in S onto plane P—essentially the planar area of the IBS.

Since the scoring function is not known a priori, we choose the IBS simply as the S that maximizes the expression

$$A_S^* + (L_{S,B} - L_{S,N})^*. \quad [4]$$

Here, the asterisks denote that the value was normalized by the maximum of the respective quantity observed in S_{all} so that all values can be compared on a one-to-one basis. A_S, $L_{S,B}$, and $L_{S,N}$ of the S selected as the IBS then become A, L_B, and L_N, respectively, for the scoring functions described below. If no S is found during evaluation (i.e., S_{all} is an empty set), the protein can be considered a non-AFP and scored as having a ΔT of zero. A schematic of this process is included in *SI Appendix*, Fig. S1.

We note that the IBS predicted by our method shows excellent agreement with the IBS suggested by experiment and computationally by Doxey et al. (24). Fig. 2 shows a graphical display of the IBS for selected AFPs and additional details. This automated method successfully identifies the IBS based on geometry, and on the dynamic properties of water near the protein.

Predicting Antifreeze Activity Using an NN. We measured our selected variables for each protein using 10-ns blocks from 30 to 100 ns for a total of 7 measurements per protein and 154 measurements overall. Averages are shown in Table 1. We first fit the data (excluding the two homologs) to the simplest possible equation, a linear combination:

$$\Delta T_C = a_0 + a_1 A + a_2 L_B + a_3 L_N. \quad [5]$$

Here, a_0 through a_3 are fitted constants, and ΔT_C is the experimental ΔT evaluated at a concentration of 0.3 g/L using Eq. **3**. This concentration was selected to avoid saturation effects (where ΔT is nearly constant with increasing concentration) and extrapolating the data. While this treatment is somewhat biased against AFPs like PDB ID code 4DT5 that show better performance at higher concentrations, some compromise is necessary, since an accurate single-variable equation for ΔT as a function of concentration is not available.

The results of this LM are shown in Fig. 3*A*, and the values of the fitted coefficients in Eq. **5** are as follows: $a_0 = -0.167$ K, $a_1 = 0.456$ K/nm^2, $a_2 = 0.032$ K/ps, $a_3 = -0.0411$ K/ps. The performance of the LM is quite mediocre, with a mean error from experiment of 0.51 K and a correlation coefficient of $R^2 = 0.80$. Clearly, Eq. **5** is not complex enough to accurately capture ΔT_C. We show it here, however, as it does exhibit an important qualitative insight: our coefficients match the physical intuition discussed earlier. A and L_B are positively correlated with ΔT_C, and L_N is negatively correlated with ΔT_C.

Given the quantitative shortcomings of the LM, we turn to a non-LM in the form of an NN. The NN is trained on the same 154 data points as the LM. To minimize overfitting, we use a validation holdout set along with L2 regularization (42) and model averaging (43). Details are given in *Materials and Methods*. Five-fold cross-validation for the NN shows a mean deviation from experiment of 0.36 K, suggesting that the NN is quite robust. When trained on 70% of the data, the NN shows a mean error from experiment of 0.19 K and an R^2 of 0.97, a significant improvement compared with the LM. Results are given in Fig. 3 and Table 1. This NN is remarkable in its accuracy given the

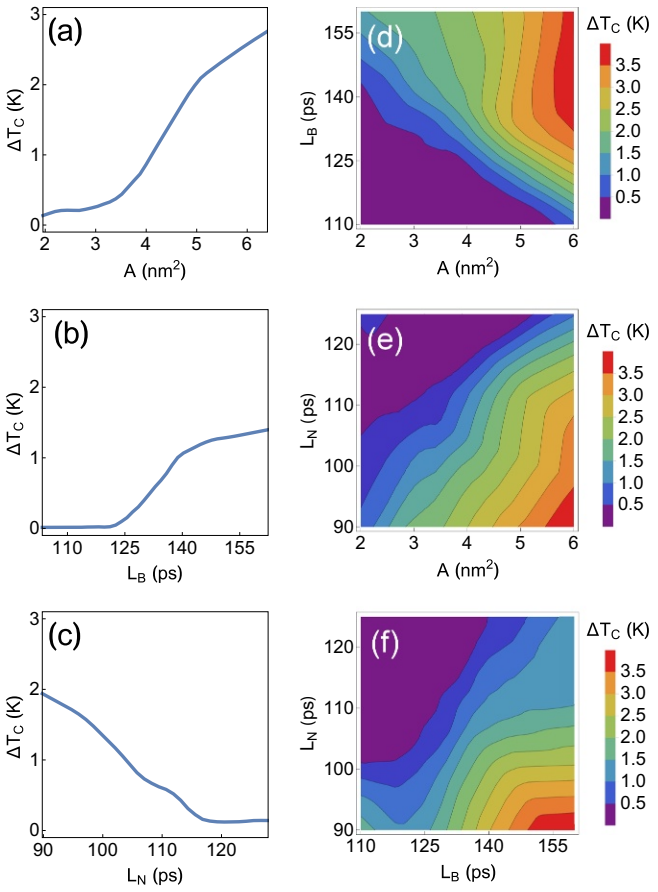

Fig. 4. Average behavior of the NN: (*A*–*C*) holding two variables at their average value and (*D*–*F*) holding one variable at its average value ($\langle A \rangle$ = 3.57 nm^2, $\langle L_B \rangle$ = 131 ps, $\langle L_N \rangle$ = 113 ps).

diversity of the proteins in the dataset and that it only requires three variables (A, L_B, and L_N) as input.

While the NN performs extremely well for almost all proteins in the dataset, it does struggle to accurately predict our one plant AFP, PDB ID code 3ULT (also known as *Lolium perenne* ice-binding protein, *Lp*IBP), isolated from ryegrass. This error may be due to the lack of other plant AFP samples in the model, but interestingly, this is not the first instance where the ryegrass AFP has underperformed expectation. Work by Middleton et al. (44) showed that, while PDB ID codes 3ULT and 1EZG (an insect AFP; also known as *Tenebrio molitor* antifreeze protein, *Tm*AFP) are very similar in structure, PDB ID code 1EZG exhibits a ΔT more than 10 times that of PDB ID code 3ULT. Importantly, ice growth experiments in the same study showed that PDB ID code 1EZG binds to all ice faces, while PDB ID code 3ULT binds almost exclusively to the basal plane. This difference would not be captured by our model given that we make no distinction regarding which ice face the AFP prefers. Incorporating this information may lead to increased accuracy in future work.

As an additional test of the NN, we also score the two homologs in our dataset that were not included in any of the training procedures. The selected homologs are (*i*) the C-terminal domain of sialic acid synthase (PDB ID code 3CM4), which is a homolog of type III AFPs like PDB ID code 1HG7, and (*ii*) a C-type lectin (PDB ID code 3WHD), which is a homolog of type II AFPs like PDB ID code 2PY2. Both homologs were evaluated in the same manner as the rest of the AFPs and received T_C scores of nearly zero (Table 1), confirming that the NN can discriminate even against AFP homologs.

For a better understanding how the NN scores AFPs, we hold two of three variables constant at their average values in the dataset and vary the third. An effective trend for each variable is, therefore, calculated as shown in Fig. 4. Overall, the curves in Fig. 4 are very clear: ΔT_C is positively correlated with A and L_B and negatively correlated with L_N, in agreement with the LM.

Importantly, the NN additionally shows that there are certain thresholds at which ΔT_C changes dramatically with respect to the input variables. Moving from $A = 2$ to 3.5 nm^2 results in very little change, but moving from $A = 3.5$ to 5 nm^2 results in a nearly fourfold increase in ΔT_C. Similarly, for L_B between 110 and 125 ps, ΔT_C is roughly zero, but for L_B between 125 and 140 ps, there is a significant gain in activity, which then slows after 140 ps. With respect to L_N, there is a nearly linear drop in activity with increasing L_N until ΔT_H is nearly zero at $L_N > 115$ ps. We imagine that this information might be of use when deciding how to construct, run, and attain convergence of a computational design process. We also include the behavior of the NN as a function of two variables (holding the third constant at its average value in the dataset) in Fig. 4.

Conclusions

This work presents a straightforward and physically motivated method for predicting the antifreeze activity of AFPs from molecular simulation. The method supports current understanding that AFPs recognize and bind ice with a water layer defined by long hydrogen bond lifetimes near the IBS. We show that a simple NN produces quantitatively accurate predictions of thermal hysteresis. Furthermore, the NN suggests that short hydrogen bond lifetimes on the NBS are also quite important for producing high-activity AFPs. We hope that this information will aid in the development of advanced antifreeze materials.

Materials and Methods

Molecular Dynamics. All molecular dynamics simulations were performed using GROMACS 2016.4 (45–48). Protein structures were obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) PDB and solvated in at least 1.5 nm of water in all directions using periodic boundary conditions for a protein–protein self-image distance of at least 3 nm. Water was modeled using the Transferable Intermolecular Potential, 4-point, Ice (TIP4P/Ice) model (49) for its realistic melting temperature of ∼270 K (50), and proteins were modeled by the Amber03w force field (51) for its compatibility with four-site water models (52, 53). Additional details and comments on computational efficiency are included in *SI Appendix*.

Protein Structure Coordinates. An average protein structure was first generated by averaging over atomic coordinates for the simulation window. This structure was then reduced to a set of N residue coordinates using the geometric center of all atoms with an SASA $\geq$0.01 nm^2 in each residue. If a residue has no atoms with SASA $\geq$ 0.01 nm^2, it is eliminated. SASA is determined using GROMACS 2016.4 following the method of Eisenhaber et al. (54).

Hydrogen Bond Lifetimes. We define a hydrogen bond lifetime, L_i, as the average time that it takes for the hydrogen bond autocorrelation function (HBAF_i) to decay to 0.1. HBAF_i was defined for water–water hydrogen bonds occurring between water molecules within 0.8 nm of any atom in residue i. HBAFs were calculated over 1-ns blocks and averaged over the sample simulation window. Calculations were performed using the Python package MDAnalysis (55, 56), which uses the definition provided by Rapaport (57). For a select few partially buried residues, there exist trapped water molecules that do not influence ice binding but nevertheless, drive $L_i \to \infty$. We, therefore, ignore any L_i longer than 1 ns by setting it to zero.

NN. The NN was trained using the machine learning suite in Mathematica 11.3 (58). It was composed of one three-node batch normalization layer, four fully connected hidden layers with six nodes each, and a final single-node output layer. Linear, Ramp, Linear, Ramp activation functions, respectively, were used for the hidden layers. A minimum of 10,000 training rounds using the ADAM algorithm (59) were used to minimize a mean squared error loss function. L2 regularization coefficient 0.01 was used for all training. Standard fivefold cross-validation was used to check performance (60). The final NN is an average of five separately trained NNs using different random samplings of 70% of the data, with the remaining 30% of the data acting as a validation holdout set to reduce the overfitting and error from outliers. Experimentation with larger/deeper NNs and more complex activation functions lead to worse cross-validation scores. A grid of points scored by the NN is included in Dataset S1.

ACKNOWLEDGMENTS. D.J.K. acknowledges support from NSF Graduate Research Fellowship Grant DGE-1656466. Calculations were performed at the Terascale Infrastructure for Groundbreaking Research in Engineering and Science (TIGRESS) at Princeton University.

1. Venketesh S, Dayananda C (2008) Properties, potentials, and prospects of antifreeze proteins. *Crit Rev Biotechnol* 28:57–82.
2. Celik Y, et al. (2010) Superheating of ice crystals in antifreeze protein solutions. *Proc Natl Acad Sci USA* 107:5423–5428.
3. Harding MM, Ward LG, Haymet AD (1999) Type I 'antifreeze' proteins. Structure-activity studies and mechanisms of ice growth inhibition. *Eur J Biochem* 264:653–665.
4. Fletcher GL, Hew CL, Davies PL (2001) Antifreeze proteins of teleost fishes. *Annu Rev Physiol* 63:359–390.
5. Kristiansen E, et al. (2011) Structural characteristics of a novel antifreeze protein from the longhorn beetle Rhagium inquisitor. *Insect Biochem Mol Biol* 41:109–117.
6. Bar Dolev M, Braslavsky I, Davies PL (2016) Ice-binding proteins and their function. *Annu Rev Biochem* 85:515–542.
7. Raymond JA, DeVries AL (1977) Adsorption inhibition as a mechanism of freezing resistance in polar fishes. *Proc Natl Acad Sci USA* 74:2589–2593.
8. Karlsson JOM, Braslavsky I, Elliott JAW (July 6, 2018) Protein-water-ice contact angle. *Langmuir*, 10.1021/acs.langmuir.8b01276.
9. Naullage PM, Qiu Y, Molinero V (2018) What controls the limit of supercooling and superheating of pinned ice surfaces?. *J Phys Chem Lett* 9:1712–1720.
10. Kuiper MJ, Morton CJ, Abraham SE, Gray-Weale A (2015) The biological function of an insect antifreeze protein simulated by molecular dynamics. *eLife* 4:1–14.
11. Celik Y, et al. (2013) Microfluidic experiments reveal that antifreeze proteins bound to ice crystals suffice to prevent their growth. *Proc Natl Acad Sci USA* 110:1309–1314.
12. Nutt DR, Smith JC (2008) Dual function of the hydration layer around an antifreeze protein revealed by atomistic molecular dynamics simulations. *J Am Chem Soc* 130:13066–13073.
13. Hudait A, et al. (2018) Preordering of water is not needed for ice recognition by hyperactive antifreeze proteins. *Proc Natl Acad Sci USA* 115:8266–8271.

14. Duboué-Dijon E, Laage D (2014) Comparative study of hydration shell dynamics around a hyperactive antifreeze protein and around ubiquitin. *J Chem Phys* 141: 22D529.
15. Meister K, et al. (2013) Long-range protein-water dynamics in hyperactive insect antifreeze proteins. *Proc Natl Acad Sci USA* 110:1617–1622.
16. Chao H, DeLuca CI, Davies PL, Sykes BD, Sönnichsen FD (1994) Structure-function relationship in the globular type III antifreeze protein: Identification of a cluster of surface residues required for binding to ice. *Protein Sci* 3:1760–1769.
17. Graether SP, et al. (1999) Quantitative and qualitative analysis of type III antifreeze protein structure and function. *J Biol Chem* 274:11842–11847.
18. Cheng J, Hanada Y, Miura A, Tsuda S, Kondo H (2016) Hydrophobic ice-binding sites confer hyperactivity of an antifreeze protein from a snow mold fungus. *Biochem J* 473:4011–4026.
19. Wang C, Pakhomova S, Newcomer ME, Christner BC, Luo BH (2017) Structural basis of antifreeze activity of a bacterial multi-domain antifreeze protein. *PLoS One* 12:e0187169.
20. Marshall CB, Daley ME, Sykes BD, Davies PL (2004) Enhancing the activity of a β-helical antifreeze protein by the engineered addition of coils. *Biochemistry* 43:11637–11646.
21. Baardsnes J, Kuiper MJ, Davies PL (2003) Antifreeze protein dimer: When two ice-binding faces are better than one. *J Biol Chem* 278:38942–38947.
22. Haridas V, Naik S (2013) Natural macromolecular antifreeze agents to synthetic antifreeze agents. *RSC Adv* 3:14199.
23. Drori R, et al. (2016) A supramolecular ice growth inhibitor. *J Am Chem Soc* 138: 13396–13401.
24. Doxey AC, Yaish MW, Griffith M, McConkey BJ (2006) Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat Biotechnol* 24:852–855.
25. Liu Y, et al. (2007) Structure and evolutionary origin of Ca2+-dependent herring type II antifreeze protein. *PLoS One* 2:e548.
26. Lee JH, et al. (2012) Structural basis for antifreeze activity of ice-binding protein from arctic yeast. *J Biol Chem* 287:11460–11468.
27. Olijve LLC, et al. (2016) Blocking rapid ice crystal growth through nonbasal plane adsorption of antifreeze proteins. *Proc Natl Acad Sci USA* 113:3740–3745.
28. Lauersen KJ, Brown A, Middleton A, Davies PL, Walker VK (2011) Expression and characterization of an antifreeze protein from the perennial rye grass, Lolium perenne. *Cryobiology* 62:194–201.
29. Hanada Y, Nishimiya Y, Miura A, Tsuda S, Kondo H (2014) Hyperactive antifreeze protein from an Antarctic sea ice bacterium Colwellia sp. has a compound ice-binding site without repetitive sequences. *FEBS J* 281:3576–3590.
30. Vance TD, Graham LA, Davies PL (2018) An ice-binding and tandem beta-sandwich domain-containing protein in Shewanella frigidimarina is a potential new type of ice adhesin. *FEBS J* 285:1511–1527.
31. Graether SP, et al. (2003) Spruce budworm antifreeze protein: Changes in structure and dynamics at low temperature. *J Mol Biol* 327:1155–1168.
32. Graham LA, Davies PL (2005) Glycine-rich antifreeze proteins from snow fleas. *Science* 310:461–461.
33. Do H, Kim SJ, Kim HJ, Lee JH (2014) Structure-based characterization and antifreeze properties of a hyperactive ice-binding protein from the Antarctic bacterium Flavobacterium frigoris PS1. *Acta Crystallogr Section D Biol Crystallogr* 70:1061–1073.
34. Leinala EK, et al. (2002) A beta-helical antifreeze protein isoform with increased activity. Structural and functional insights. *J Biol Chem* 277:33349–33352.
35. Sander LM, Tkachenko AV (2004) Kinetic pinning and biological antifreezes. *Phys Rev Lett* 93:128102.
36. Langmuir I (1918) The adsorption of gases on plane surfaces of glass, mica and platinum. *J Am Chem Soc* 40:1361–1403.
37. Marshall CB, Chakrabartty A, Davies PL (2005) Hyperactive antifreeze protein from winter flounder is a very long rod-like dimer of α-helices. *J Biol Chem* 280:17920–17929.
38. Knight CA, Wierzbicki A (2001) Adsorption of biomolecules to ice and their effects upon ice growth. 2. A discussion of the basic mechanism of "antifreeze" phenomena. *Cryst Growth Des* 1:439–446.
39. Garnham CP, et al. (2008) A Ca 2+ -dependent bacterial antifreeze protein domain has a novel β-helical ice-binding fold. *Biochem J* 411:171–180.
40. Graether SP, et al. (2000) Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature* 406:325–328.
41. Schrödinger L (2015) The PyMOL Molecular Graphics System (Schrödinger, LLC, Portland, OR), Version 1.8.
42. Ng AY (2004) Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *Proceedings of the 21st International Conference on Machine Learning*. Available at https://dl.acm.org/citation.cfm?doid=1015330.1015435. Accessed October 31, 2018.
43. Hashem S (1997) Optimal linear combinations of neural networks. *Neural Networks* 10:599–614.
44. Middleton AJ, et al. (2012) Antifreeze protein from freeze-tolerant grass has a beta-roll fold with an irregularly structured ice-binding site. *J Mol Biol* 416:713–724.
45. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.
46. Pronk S, et al. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–854.
47. Szilárd P, Abraham MJ, Kutzner C, Hess B, Lindahl E (2015) Tackling exascale software challenges in molecular dynamics simulations with GROMACS. *Solving Software Challenges for Exascale*, Lecture Notes in Computer Science, eds Markidis S, Laure E (Springer, Cham, Switzerland), Vol 8759, pp 3–27.
48. Abraham MJ, et al. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19–25.
49. Abascal JL, Sanz E, Fernández RG, Vega C (2005) A potential model for the study of ices and amorphous water: TIP4P/Ice. *J Chem Phys* 122:234511.
50. García Fernández R, Abascal JLF, Vega C (2006) The melting point of ice Ih for common water models calculated from direct coexistence of the solid-liquid interface. *J Chem Phys* 124:144506.
51. Best RB, Mittal J (2010) Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse. *J Phys Chem B* 114:14916–14923.
52. Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 8:1409–1414.
53. Palazzesi F, Prakash MK, Bonomi M, Barducci A (2015) Accuracy of current all-atom force-fields in modeling protein disordered states. *J Chem Theory Comput* 11:2–7.
54. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M (1995) The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J Comput Chem* 16:273–284.
55. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* 32:2319–2327.
56. Gowers RJ, et al. (2016) MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. *Proceedings of the 15th Python in Science Conference (Scipy)*. Available at https://conference.scipy.org/proceedings/scipy2016/oliver_beckstein.html. Accessed October 31, 2018.
57. Rapaport D (1983) Hydrogen bonds in water. *Mol Phys* 50:1151–1162.
58. Wolfram Research Inc. (2018) Mathematica (Wolfram Research, Inc, Champaign, IL).
59. Kingma DP, Ba JL (2014) Adam: A method for stochastic optimization. arXiv:1412.6980v9. Preprint, posted December 22, 2014.
60. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence*. Available at http://ai.stanford.edu/∼ronnyk/accEst.pdf. Accessed October 31, 2018.