JCTC Journal of Chemical Theory and Computation

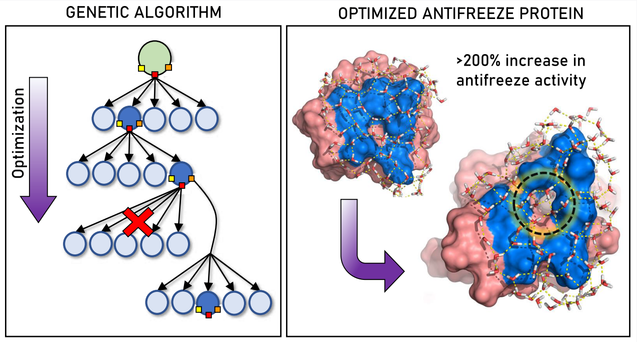
pubs.acs.org/JCTC Article

Genetic Algorithm Approach for the Optimization of Protein Antifreeze Activity Using Molecular Simulations

Daniel J. Kozuch, Frank H. Stillinger, and Pablo G. Debenedetti*

Cite This: *J. Chem. Theory Comput.* 2020, 16, 7866−7873 Read Online

ACCESS | Metrics & More | Article Recommendations


GENETIC ALGORITHM
Optimization
OPTIMIZED ANTIFREEZE PROTEIN
>200% increase in antifreeze activity

ABSTRACT: Antifreeze proteins (AFPs) are of much interest for their ability to inhibit ice growth at low concentrations. In this work, we present a genetic algorithm for the *in silico* design of AFP mutants with improved antifreeze activity, measured as the predicted thermal hysteresis at a fixed concentration, ΔT_C. Central to the algorithm is our recently developed neural network method for predicting ΔT_C from molecular simulations [Kozuch et al., *PNAS*, *115*, 13252 (2018)]. Applying the algorithm to three structurally diverse AFPs, *wf*AFP, rQAE, and *Ri*AFP, we find that significantly improved mutants are discovered for rQAE and *Ri*AFP. Testing of the optimized mutants shows an increase in ΔT_C of 0.572 ± 0.11 K (262 ± 50.6%) and 1.33 ± 0.14 K (39.9 ± 4.19%) over the native structures for rQAE and *Ri*AFP, respectively. Structural analysis of the optimized mutants reveals that the algorithm is able to exploit two pathways for enhancing the predicted antifreeze activity of the mutants: (1) increasing the local order of surface waters by encouraging the formation of internal water channels in the protein and (2) increasing the total ice-binding area by improving the planar structure of the ice-binding surface. Additionally, analysis of all mutants explored by the algorithm reveals that a subset of residues, mainly nonpolar, are particularly helpful in improving antifreeze activity at the ice-binding surface.

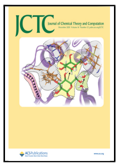
■ INTRODUCTION

Antifreeze proteins (AFPs) are a broad class of proteins that noncolligatively depress the freezing point of water by binding to nascent ice crystals. Once bound, the proteins increase the local curvature of the ice surface, thereby depressing the freezing point through the Gibbs−Thomson effect.[1,2] The resulting difference between the melting temperature and the freezing temperature is referred to as thermal hysteresis, ΔT.

AFPs have already been widely used in the food industry, and they are also being explored for use in cryopreservation, agriculture, and de-icing.[3−7] However, for many applications, the native antifreeze activity of AFPs is too low for practical implementation. For example, it was found that transgenic plants expressing AFPs could withstand temperatures as much as 3 °C colder than the wild type, but that this was insufficient for crop protection, which required plants to remain unfrozen at temperatures of −5 or −6 °C.[8] While there are of course many factors at play in the cold survival of plants, increasing the antifreeze activity of transgenically expressed AFPs is one potential route for achieving the necessary cold endurance.

As discussed in our recent paper,[9] many methods have been explored for studying and designing new AFPs. Early experiments used site-specific mutations to probe which residues were involved in ice binding by determining which mutations led to the greatest loss in antifreeze activity.[10−12] Later work studied the effect of joining AFPs into dimers,[13] resulting in a significant increase in ΔT on a molar basis, although little to no increase in ΔT was observed on a per-mass basis (which we suggest for more equal comparisons of differently sized AFPs). In a similar fashion, Marshall et al.[14] increased the ice-binding surface (IBS) of AFPs by extending a repeating loop structure already present in the native AFP, demonstrating that these enlarged structures possessed higher antifreeze activity than the wild type. Fusion constructs have also been a useful tool in studying AFP behavior[15] and have been used to design increasingly complex antifreeze particles.[16,17] In contrast to earlier work, more recent site-directed mutagenesis of a highly active insect AFP was used to produce mutants with higher antifreeze activity than the wild type by specifically mutating residues that were suspected to interfere with ice binding.[18] In addition to these experimental techniques, computational methods have also been widely used

Received: July 23, 2020
Published: November 17, 2020



https://dx.doi.org/10.1021/acs.jctc.0c00773

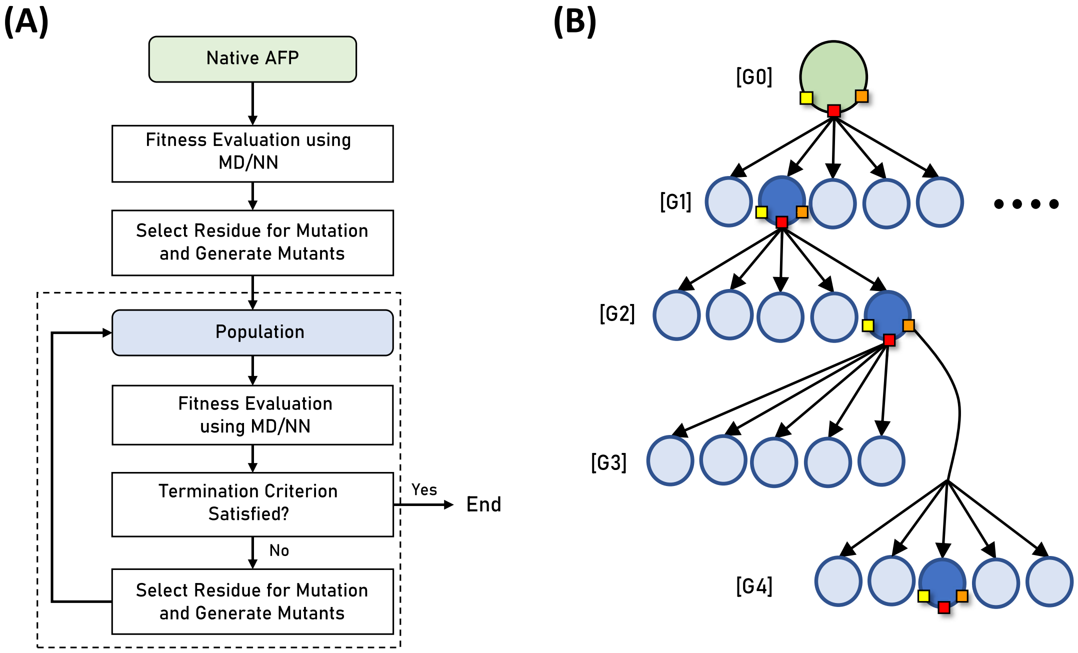

(A)
Native AFP
Fitness Evaluation using MD/NN
Select Residue for Mutation and Generate Mutants
Population
Fitness Evaluation using MD/NN
Termination Criterion Satisfied?
Yes
End
No
Select Residue for Mutation and Generate Mutants
(B)
[G0]
[G1]
[G2]
[G3]
[G4]

Figure 1. (A) Procedure for the genetic algorithm explained in the text. (B) Schematic illustrating how residues are selected for mutation. [G0]: Generation zero, the native antifreeze protein (large green circle) is evaluated with the fitness function and the residue with the lowest L_i (red square) is mutated to form the next generation (only 5 of the 17 shown for each generation). [G1/G2]: Individuals in the generation (blue circles) are evaluated and the best-performing individual (dark blue circle) is selected to produce the next generation through the same mutation scheme. [G3/G4]: None of the mutants in G3 outperform the best mutant in G2, so the algorithm returns to the best mutant in G2 and mutates at the residue with the second lowest L_i (orange square) to produce G4.

to study AFP behavior[2,3,19−23] and identify potential AFPs from growing genetic and structural databases.[24−27] However, these efforts have, until now, not yet been directed toward the optimization of AFP antifreeze activity at the atomistic level.

We propose here a general method for optimizing naturally occurring AFPs *in silico* using a genetic algorithm (GA) that successively mutates the protein of interest. Evaluation of the resulting mutants is performed using our recently developed technique for predicting protein antifreeze activity from molecular simulations.[9] Details of the GA and corresponding molecular simulations are provided in the Methods section. Performance of the GA and discussion of the optimized AFPs are presented in the Results and Discussion section.

■ METHODS

Genetic Algorithm. In this work, we aim to develop a strategy for optimizing the antifreeze activity of naturally occurring antifreeze proteins *in silico* through the implementation of a genetic algorithm.[28] Fundamental to all genetic algorithms is the fitness function, $f(\mathbf{X})$, which is used to evaluate the fitness of each individual in the population, where $\mathbf{X}$ is a vector describing a given individual. For our fitness function, we employ our recently developed neural network method for predicting protein antifreeze activity from molecular simulations.[9] In this work, antifreeze activity is specifically defined as ΔT at a fixed concentration of 0.3 g/L AFP, referred to as ΔT_C.

Briefly, this method relies on the observation that high-activity antifreeze proteins possess a planar ice-binding surface (IBS), near which the hydrogen bond lifetime, L, of the solvating water is longer than elsewhere near the protein. To quantify this property, an all-atom molecular dynamics simulation is first performed for the protein of interest. From this simulation, each solvent-accessible residue, i, is assigned a mean hydrogen bond lifetime value, L_i, calculated from the nearby solvation water. Our algorithm then automatically identifies a set of planar residues, S, that form the IBS. S is selected by maximizing the area of the IBS, A, and the mean L_i for residues in the IBS, L_B, while minimizing the mean L_i for residues not in the IBS, L_N. The neural network then predicts ΔT_C from A, L_B, and L_N. The function is positively correlated with A and L_B and negatively correlated with L_N. For simulation details and complete description, see our previous work.[9]

Given that our chosen f is expensive to compute (~24 h on 1 GPU/14CPUs, mainly for the molecular simulation) and the sequence space is immense, it would be impractical to employ a standard genetic algorithm. Instead, we choose to implement a directed and simplified genetic algorithm with targeted mutation, zero crossover, and elitist selection.[29] This design limits the population of each generation to a manageable number, but it should be noted that, given more resources or a less expensive f, the full search space would likely be explored better with the inclusion of some selective crossovers. The following describes our algorithm in detail, and a procedure/schematic is provided in Figure 1.

The algorithm is initialized by first evaluating the native antifreeze protein of interest (obtained from rcsb.org[30]) with the fitness function. This is generation $g = 0$, and it provides the score (ΔT_C) for the native protein as well as identifying the native IBS and providing L_i for each residue in the IBS. Since the fitness function is positively correlated with L_B, it is logical to mutate the residue in or adjacent to the IBS with the lowest L_i, referred to as i_L^{min}, since we seek to eliminate low L_i on the IBS and to expand the IBS to encompass nearby residues when possible. Here, "adjacent to the IBS" is defined as all solvent-exposed residues with a geometric center within 1 nm of the

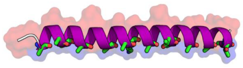
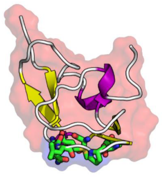
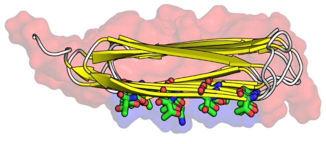
Common Name	Organism	Type	RCSB PDB	Native Structure
*wf*AFP	Winter Flounder	I	1WFA	
rQAE (HPLC-12)	Ocean Pout	III	1HG7	
*Ri*AFP	Longhorn Beetle	V	4DT5	

Figure 2. Properties of the three AFPs studied in this work. Type refers to the structural classification of each AFP.[3] Structures are shown with α-helices in purple, β-sheets in yellow, and unstructured loops in white. Ice-binding residues are explicitly shown with carbons in green, nitrogen atoms in blue, and oxygen atoms in red (hydrogen atoms not shown for clarity). Transparent protein surfaces shown with the IBS in blue and the non-IBS in red. PDB structures are obtained from www.rcsb.org with refs 46, 47 and 48 for *wf*AFP, rQAE, and *Ri*AFP, respectively. Visualization is done with PyMOL.[49]

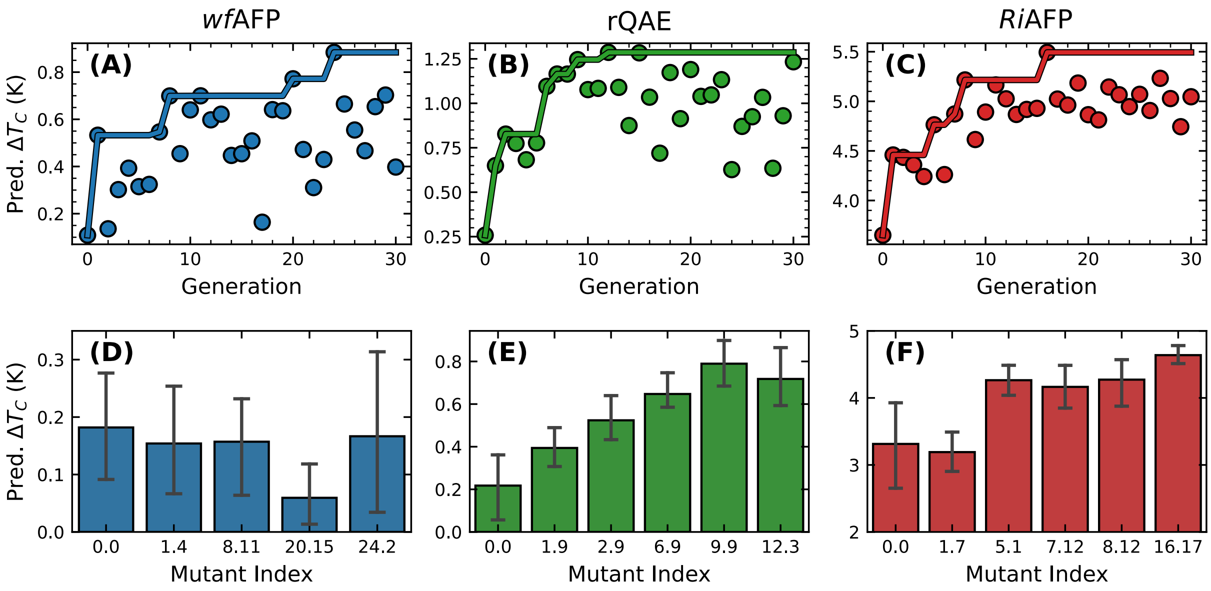


Figure 3. Evolution results for *wf*AFP, rQAE, and *Ri*AFP over 30 generations. (A−C) Points are the maximum ΔT_C predicted for a given generation, and the solid line represents the maximum ΔT_C observed over all previous generations. (D−F) Block averaged results for longer simulations of selected mutants for each AFP. Error bars are 95% confidence intervals (1.96xSE).

geometric center of any residue in the IBS. As such, the next generation ($g = 1$) is populated by mutating $i_L^{\min}$ to all other nonaromatic residues using the SCWRL4[31] program to generate new protein structures (aromatic residues were considered too bulky to assist in ice binding). We note that the length of the simulations employed during the optimization is not enough to fully "refold" the proteins. Instead, the method assumes that a point mutation is a small enough perturbation in most cases that a short equilibration period is enough to resolve changes to the IBS. This yields a population of 17 individuals (all mutants), which are then simultaneously evaluated using the fitness function.

Elitist selection is then performed by finding the best-scoring individual, $I_{\max}$, observed so far in *any* generation. The same mutation procedure is then repeated, except that now we mutate at the residue in $I_{\max}$ that has the lowest L_i on the IBS and has not yet been chosen for mutation. In this way, a new generation is always produced by mutating the best-observed individual, and no generations will be unnecessarily repeated. This procedure is then repeated until 30 generations are completed.

Molecular Simulation. All molecular dynamics simulations were performed using GROMACS 2016.4.[32−35] Native protein structures were obtained from the RCSB Protein Data Bank and solvated in at least 1.5 nm of water in all directions, employing periodic boundary conditions for a protein−protein self-image distance of at least 3 nm. Na^+/Cl^- ions were added for charge neutralization. Water was modeled using the

Table 1. Mutations, Properties, and Scores for Selected Mutants in Figure 3D−F[a]

results for selected mutants of *wf*AFP

generation	mutant	mutations	binding residues	A (nm^2)	±	L_B (ps)	±	L_N (ps)	±	pred. ΔT_C (K)	±
0	0	N/A	13	2.63	0.49	127.98	3.93	110.04	1.11	0.18	0.10
1	4	R36E	13	3.14	0.53	128.26	3.59	115.53	2.10	0.15	0.09
8	11	T1N, A33N, R36E	12	3.16	0.48	122.18	4.42	107.84	1.56	0.16	0.08
20	15	T1N, S3K, K17S, A33N, R36E	14	3.08	0.71	108.51	9.96	112.10	4.76	0.06	0.05
24	2	T1N, S3K, L11C, K17S, A33N, R36E	13	3.12	0.14	118.81	4.57	104.94	1.72	0.17	0.14

results for selected mutants of rQAE

generation	mutant	mutations	binding residues	A (nm^2)	±	L_B (ps)	±	L_N (ps)	±	pred. ΔT_C (K)	±
0	0	N/A	8	2.15	0.16	124.02	20.56	125.15	3.32	0.22	0.16
1	9	K61L	8	2.20	0.23	141.00	4.47	123.80	3.02	0.39	0.09
2	9	N8L, K61L	9	2.36	0.03	149.13	3.76	122.60	1.72	0.52	0.11
6	9	N8G, Q9L, K61L	8	2.18	0.15	152.23	6.31	132.83	1.02	0.65	0.08
9	9	N8G, Q9L, N14L, L40I, K61L	8	2.10	0.09	161.40	5.33	131.23	1.17	0.79	0.11
12	3	N8G, Q9L, N14L, L40I, Q44D, K61L	8	2.23	0.12	157.04	6.61	129.76	1.83	0.72	0.14

results for selected mutants of *Ri*AFP

generation	mutant	mutations	binding residues	A (nm^2)	±	L_B (ps)	±	L_N (ps)	±	pred. ΔT_C (K)	±
0	0	N/A	23	6.60	1.28	132.80	3.12	105.26	0.74	3.31	0.68
1	7	G10I	26	7.20	0.25	124.36	4.09	104.57	0.92	3.19	0.33
5	1	G10I, T108A	26	8.47	0.35	132.31	1.65	105.70	2.19	4.26	0.22
7	12	G10I, T108A, T114P	27	7.55	0.51	141.15	1.43	106.61	1.74	4.16	0.33
8	12	G10I, K23P, T108A, T114P	29	8.57	0.78	141.69	6.99	110.88	1.41	4.27	0.36
16	17	G10I, K23P, T108A, T114V	30	8.50	0.27	139.22	1.85	105.48	0.55	4.64	0.14

[a] Mutations are given as a list of mutated residues with the form [Original Residue][Residue Number][New Residue], using the single-letter amino acid code. Residue numbers begin at zero. Binding residues refers to the number of residues in the IBS. A, L_B, and L_N are inputs to the neural network used in the fitness function. Errors (±) are 95% confidence intervals (1.96xSE).

TIP4P/Ice model[36] for its realistic melting temperature of ∼270 K,[37] and proteins/ions were modeled by the Amber03w force field[38] for its compatibility with 4-site water models.[39,40]

The temperature was maintained at 265 K to mimic subfreezing conditions using a v-rescale thermostat.[41] All systems were first energy minimized using the steepest descent and then equilibrated for a short period of 100 ps at 1 bar using the Berendsen barostat.[42] The pressure was maintained at 1 bar using the Parrinello−Rahman barostat[43] during sampling. A time step of 2 fs was used and trajectories were saved every 10 ps. Bonds were constrained using the LINCS algorithm.[44] Short-range interactions were truncated at 1.0 nm, and long-range electrostatics were handled by particle mesh Ewald (PME) summation.[45]

■ RESULTS AND DISCUSSION

For this study, three naturally occurring AFPs, shown in Figure 2, were selected to represent several commonly occurring types of AFPs, covering a diverse set of tertiary structures and a wide range of measured thermal hysteresis values. To avoid confusion with different isoforms, the RCSB PDB code (www.rcsb.org) for each AFP is also listed in Figure 2. Optimization of the predicted thermal hysteresis, ΔT_C, was performed using the GA outlined in the Methods section. Each AFP was evolved over 30 generations, and results are presented in Figure 3A−C. Mutants are referred to by their index in the form of {gen}.{num}. For example, mutant number 11 from generation 8 would have an index of 8.11, and index 0.0 is the native structure.

Maximum observed ΔT_C values for mutants of the three AFPs represent an increase of approximately 700, 400, and 50 percent over the native structure for *wf*AFP, rQAE, and *Ri*AFP, respectively. While these are encouraging results, given the large number of tests (17 mutants per generation and 30 generations give over 500 mutants per AFP) and relatively short simulation times, these values are likely statistical outliers.

To provide a more accurate estimate of the expected ΔT_C, structures with high ΔT_C across several generations were selected from each optimization and subjected to longer simulations (200 ns). The first 100 ns of these simulations were discarded for equilibration, and the last 100 ns were divided into 10 blocks of 10 ns each. During this time, the tertiary structure for rQAE and *Ri*AFP remained stable, with the α-carbon root-mean-square-deviation (C-α RMSD) with respect to the starting structures less than 0.3 nm for all mutants. *wf*AFP remained semistable, with several mutants reaching a maximum C-α RMSD of 0.4−0.6 nm. These blocks were then independently scored using the same fitness function employed by the GA to obtain error estimates for the predicted ΔT_C. Results are shown in Figure 3D−F and Table 1.

From the data presented in Figure 3D−F, it is evident that the GA was successful for rQAE and *Ri*AFP, where the predicted ΔT_C of the optimal mutants had improvements of 0.572 ± 0.11 K (262 ± 50.6%) and 1.33 ± 0.14 K (39.9 ± 4.19%) over the native structures, respectively. However, the GA failed to provide a statistically improved candidate for *wf*AFP. This is likely due to the more unstable nature of *wf*AFP as compared to rQAE and *Ri*AFP, quantitatively characterized by the mean α-carbon root-mean-square-fluctuation (C-α RMSF). The C-α RMSF was found to be more than twice as high for the native *wf*AFP (0.127 ± 0.055 nm) than for the native rQAE and *Ri*AFP (0.058 ± 0.042, and 0.048 ± 0.036 nm, respectively). This also holds true if we restrict the calculation to residues in the IBS, where the C-α RMSF is 0.122 ± 0.036, 0.026 ± 0.003, and 0.057 ± 0.030 nm for *wf*AFP, rQAE, and *Ri*AFP, respectively. Additionally, *wf*AFP is

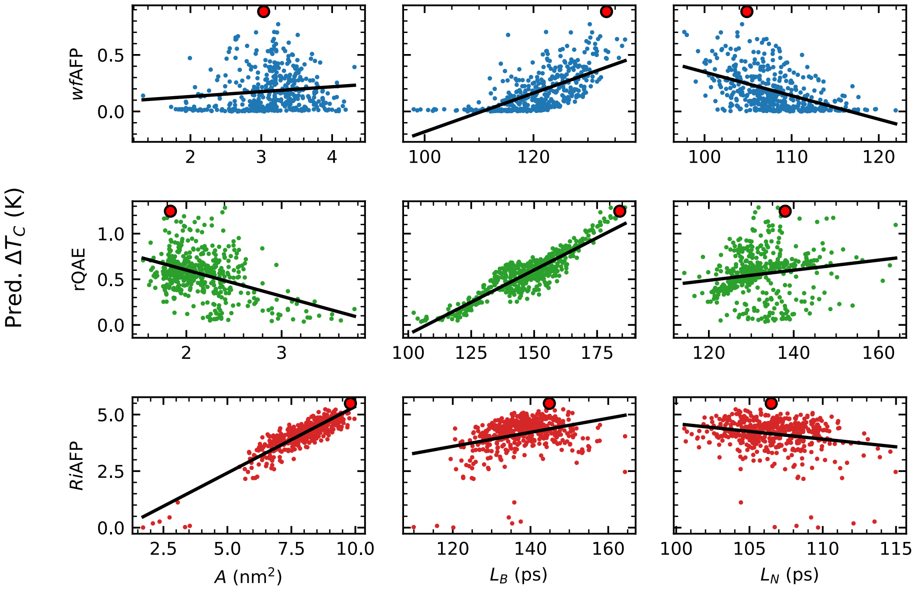


Figure 4. Predicted thermal hysteresis, ΔT_C, as a function of the three underlying variables used by the neural network (see text for details). Each point is a mutant produced by the genetic algorithm. Black lines are linear fits to the data. Larger red points are the optimal mutants selected from the results in Figure 3, with mutant index 24.2, 9.9, and 16.17 for *wf*AFP, rQAE, and *Ri*AFP, respectively.

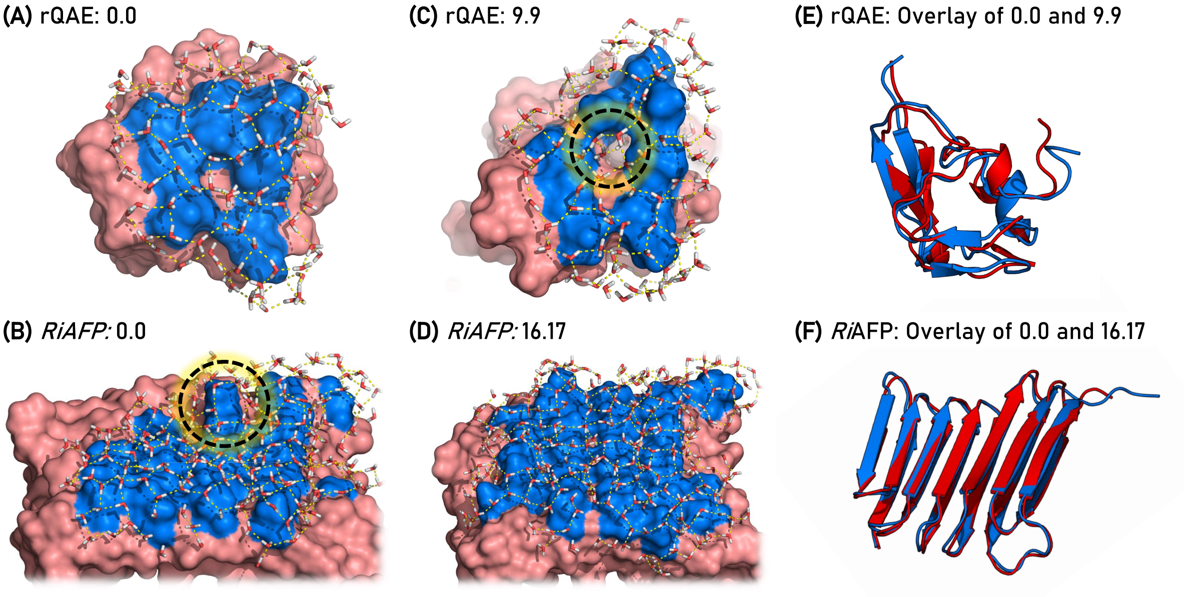

(A) rQAE: 0.0
(C) rQAE: 9.9
(E) rQAE: Overlay of 0.0 and 9.9
(B) RiAFP: 0.0
(D) RiAFP: 16.17
(F) RiAFP: Overlay of 0.0 and 16.17

Figure 5. (A, B) IBS of native structures for rQAE and *Ri*AFP. Ice-binding surfaces are colored blue, and non-ice-binding surfaces are colored light red. Water molecules are shown near the IBS, with hydrogen bond networks shown with yellow dashed lines. The dashed circle in (B) indicates the exposed lysine residue that is mutated in the optimized protein. (C, D) IBS of optimized structures for rQAE and *Ri*AFP. The dashed circle in (C) indicates the location of a solvent channel formed in the IBS of the optimized protein. (E, F) Overlay of the native (red) and optimized (blue) secondary/tertiary structures showing that the mutations do not significantly perturb the overall structure of the protein. Visualization is done in PyMOL.[49]

composed of a single α-helix that can be significantly disturbed by a single point mutation, indicated by the higher mutant C-α RMSD (discussed above), while the larger rQAE and *Ri*AFP are composed of more robust tertiary structures that retain their three-dimensional structure upon a single point mutation.

We now consider how the GA chose to optimize the three underlying variables that are fed to the neural network in our fitness function: (1) the area of the IBS, A, (2) the mean hydrogen bond lifetime of the solvent near the IBS, L_B, and (3) the mean hydrogen bond lifetime of the solvent near the non-ice-binding surface, L_N. For details concerning the determination of these variables, please see the Methods section and our previous work.[9] From Figure 4, it is evident that the GA chooses to optimize different variables for different AFPs. For example, the GA maximized L_B and selected a low A for rQAE, while a large A and medium L_B were selected for *Ri*AFP. This divergence can most likely be attributed to the differences in tertiary structures between rQAE and *Ri*AFP; rQAE is more globular than *Ri*AFP (which has a highly planar IBS), making it more efficient for the algorithm to optimize the L_B of rQAE

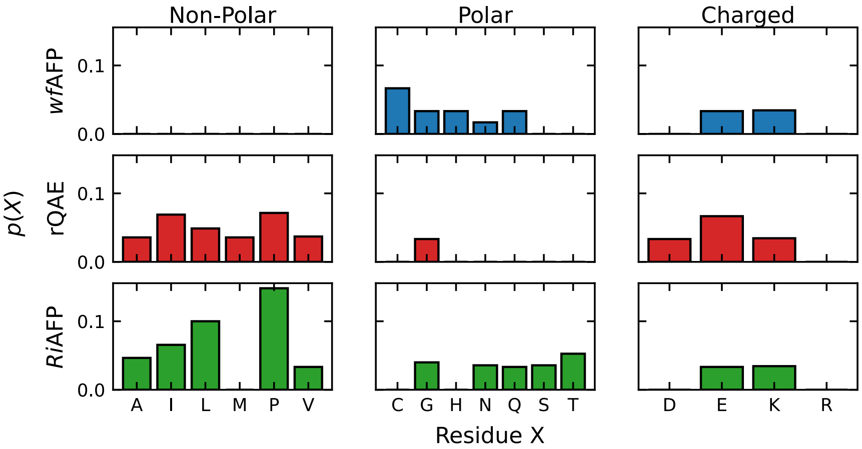


Figure 6. Probability, $p(X)$, that mutation from any residue to a specific residue X results in an increase in the predicted ΔT_C greater than 0.2 K. Results are shown for each AFP studied and residues are split by their classification as nonpolar, polar, or charged. While histidine (H) can have different protonation near physiological pH depending on its environment, it was modeled as neutral (polar) for all simulations. Residues with no bar have a $p(X)$ = 0. Aromatic groups are not included as no mutations to aromatic residues were made.

than to completely rearrange a significant number of residues to achieve a higher A.

Additionally, there also exists a natural competition between increasing A and maintaining a structured binding surface with high L_B. This is because the likelihood that the best mutation for increasing A is also the best mutation for increasing L_B is quite small, considering the size of the search space. Furthermore, mutations that expand the IBS by allowing for additional residues (which are more likely to have lower L_B than those already in the IBS) to be considered part of the IBS can be expected to decrease L_B while increasing A. The algorithm must balance these properties to discover structures that will be scored highest by the fitness function. This balance is, of course, greatly influenced by the structure of the native AFP that provides the starting point for the algorithm.

The difference in optimization paths for different AFPs is also evident in the structure of the optimized AFPs, shown in Figure 5. Comparing the native and optimized structures for rQAE, a significant pore is opened in the center of the ice-binding surface where water molecules can reside. This allows for much longer hydrogen bond lifetimes, as the location/orientation of the water molecules is stabilized by the surrounding protein. This property of encouraging stable hydrogen bond networks by allowing interior channels of the AFP to be occupied by the solvent has actually been observed in a naturally occurring, high-activity AFP known as Maxi,[50,51] demonstrating that this is likely a viable route for optimization.

In contrast, the evolution of *Ri*AFP proceeds by expanding the ice-binding surface. This is in part achieved by removing an exposed lysine residue in the center of the binding face and replacing it with a more compact proline residue. This allows for several additional residues to be included in the now more planar binding face (see the back left of the ice-binding surface in Figure 5D). Interestingly, a similar mutation in which a lysine was removed from a planar IBS was employed by Friis et al. in their optimization of an insect AFP (similar to *Ri*AFP), resulting in a significant increase in ΔT.[18] This again indicates that the mutations discovered by the algorithm are likely feasible pathways for improving antifreeze activity, while the algorithm also has the benefit of being fully automated, customizable, and transferable to other AFPs.

The resulting benefit of each mutation performed during optimization was also analyzed by considering the outcome of mutations to each possible residue. This was done by calculating the probability, $p(X)$, that a given mutation from any original residue to the new residue, X, resulted in an increase in the predicted ΔT_C of more than 0.2 K. Therefore, a higher $p(X)$ indicates that mutating to residue X is, on average, more efficient than mutating to a residue with a lower $p(X)$. Results are shown in Figure 6 for each AFP, but since the evolution of *wf*AFP failed to produce any statistically improved candidate, we will again focus on rQAE and *Ri*AFP.

The residues with the highest $p(X)$ for rQAE and *Ri*AFP are all nonpolar amino acids, with proline (P) and leucine (L) both having $p(X) \geq 0.10$ for *Ri*AFP. This is likely because the mean L_i for nonpolar residues (141 ± 13 ps) is greater than the mean L_i near polar (128 ± 12 ps) or charged residues (107 ± 7.4 ps) as measured for all rQAE and *Ri*AFP mutants. See the Methods section for details concerning the calculation of hydrogen bond lifetimes near protein residues. As a result of this difference, the algorithm can easily increase L_B by mutating to nonpolar residues. Intriguingly, the charged residues glutamic acid (E) and lysine (K) also have consistently high $p(X)$. Since these residues are not compact and do not encourage long hydrogen bond lifetimes in the solvation shell, these results likely indicate that glutamic acid and lysine residues are important for maintaining structural stability in the simulations. Furthermore, even though lysine has a modest $p(X)$, it is often replaced by other residues in the highest performing mutants (see Table 1), suggesting that while it may provide a small increase in ΔT_C during optimization, it is not a desirable residue to retain in the IBS.

In contrast, mutations to polar amino acids are significantly less likely to result in an increase in ΔT_C, especially for rQAE where $p(X)$ = 0 for many polar residues. This is somewhat surprising considering the strong role played by the polar residue threonine (T) in several AFPs,[12,52,53] although mutations to threonine have been observed to cause a loss of antifreeze activity.[11] One reason for this outcome may be that, by design, mutations were localized to residues with poor (low) hydrogen bond lifetimes in their solvation shell, and that these locations may not benefit from the addition of threonine. It is worth mentioning that previous work by Midya et al.

predicts that both polar and nonpolar residues are likely required for efficient ice binding of a highly active insect AFP, *Tm*AFP.[54] This prediction is not at odds with the data presented in Figure 6, as our mutants contain at most 6 mutations, and all still retain a significant number of polar residues on the binding face. Further work would be required to demonstrate whether a balance of polar/nonpolar residues on the IBS is a strict requirement for antifreeze activity.

There are several residues that have a $p(X)$ of zero for both rQAE and *Ri*AFP, i.e., mutation to these residues never provided a significant increase in ΔT_C. These residues are cysteine (C), histidine (H), and arginine (R), and all of them are either polar or charged. Additionally, methionine (M) and aspartic acid (D) have $p(X)$ of zero for *Ri*AFP but not rQAE. Regardless, these results indicate that, in the future, a more efficient algorithm may be designed by eliminating mutations to residues that have a very low likelihood of improving the expected antifreeze activity, such as cysteine and histidine, and focusing on mutation to residues that have a high likelihood of improving antifreeze activity.

■ CONCLUSIONS

In this work, we explored the use of a genetic algorithm to computationally optimize the predicted activity of three antifreeze proteins. In two of the three cases, a significantly improved sequence was designed, suggesting that this method could be a useful pathway to producing higher activity antifreeze proteins. Additionally, analysis suggests that for the proteins studied, there are particular residues that are more beneficial than others to include in the ice-binding surface of the antifreeze protein. We hope that this work will encourage further study in this area and that experimental testing of our *in silico* mutants demonstrates improved activity *in vitro*. Furthermore, we hope that the design strategy presented here inspires future theoretical work in the space of protein sequence optimization, such as how changes in sequence length or introduction of crossover (which were excluded here) might impact convergence.

■ AUTHOR INFORMATION

Corresponding Author

Pablo G. Debenedetti − *Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0003-1881-1728; Email: pdebene@princeton.edu

Authors

Daniel J. Kozuch − *Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-9671-8396

Frank H. Stillinger − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-1225-8186

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.0c00773

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656466 awarded to D.J.K.

■ REFERENCES

(1) Karlsson, J. O. M.; Braslavsky, I.; Elliott, J. A. W. Protein−Water−Ice Contact Angle. *Langmuir* **2018**, *35*, 7383−7387.

(2) Naullage, P. M.; Qiu, Y.; Molinero, V. What Controls the Limit of Supercooling and Superheating of Pinned Ice Surfaces? *J. Phys. Chem. Lett.* **2018**, *1*, 1712−1720.

(3) Bar Dolev, M.; Braslavsky, I.; Davies, P. L. Ice-Binding Proteins and Their Function. *Annu. Rev. Biochem.* **2016**, *85*, 515−542.

(4) Voets, I. K. From ice-binding proteins to bio-inspired antifreeze materials. *Soft Matter* **2017**, *13*, 4808−4823.

(5) Li, Q.; Guo, Z. Fundamentals of icing and common strategies for designing biomimetic anti-icing surfaces. *J. Mater. Chem. A* **2018**, *6*, 13549−13581.

(6) Surís-Valls, R.; Voets, I. K. Peptidic Antifreeze Materials: Prospects and Challenges. *Int. J. Mol. Sci.* **2019**, *20*, 5149.

(7) Mangiagalli, M.; Brocca, S.; Orlando, M.; Lotti, M. The "cold revolution". Present and future applications of cold-active enzymes and ice-binding proteins. *New Biotechnol.* **2020**, *55*, 5−11.

(8) Duman, J. G.; Wisniewski, M. J. The use of antifreeze proteins for frost protection in sensitive crop plants. *Environ. Exp. Bot.* **2014**, *106*, 60−69.

(9) Kozuch, D. J.; Stillinger, F. H.; Debenedetti, P. G. Combined molecular dynamics and neural network method for predicting protein antifreeze activity. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 13252−13257.

(10) Chao, H.; DeLuca, C. I.; Davies, P. L.; Sykes, B. D.; Sönnichsen, F. D. Structure-function relationship in the globular type III antifreeze protein: Identification of a cluster of surface residues required for binding to ice. *Protein Sci.* **1994**, *3*, 1760−1769.

(11) Graether, S. P.; DeLuca, C. I.; Baardsnes, J.; Hill, G. A.; Davies, P. L.; Jia, Z. Quantitative and Qualitative Analysis of Type III Antifreeze Protein Structure and Function. *J. Biol. Chem.* **1999**, *274*, 11842−11847.

(12) Graether, S. P.; Kuiper, M. J.; Gagné, S. M.; Walker, V. K.; Jia, Z.; Sykes, B. D.; Davies, P. L. Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature* **2000**, *406*, 325−328.

(13) Baardsnes, J.; Kuiper, M. J.; Davies, P. L. Antifreeze protein dimer: when two ice-binding faces are better than one. *J. Biol. Chem.* **2003**, *278*, 38942−38947.

(14) Marshall, C. B.; Daley, M. E.; Sykes, B. D.; Davies, P. L. Enhancing the Activity of a β-Helical Antifreeze Protein by the Engineered Addition of Coils. *Biochemistry* **2004**, *43*, 11637−11646.

(15) DeLuca, C. I.; Comley, R.; Davies, P. L. Antifreeze proteins bind independently to ice. *Biophys. J.* **1998**, *74*, 1502−1508.

(16) Phippen, S. W.; Stevens, C. A.; Vance, T. D. R.; King, N. P.; Baker, D.; Davies, P. L. Multivalent Display of Antifreeze Proteins by Fusion to Self-Assembling Protein Cages Enhances Ice-Binding Activities. *Biochemistry* **2016**, *55*, 6811−6820.

(17) Wilkins, L. E.; Hasan, M.; Fayter, A. E. R.; Biggs, C.; Walker, M.; Gibson, M. I. Site-specific conjugation of antifreeze proteins onto polymer-stabilized nanoparticles. *Polym. Chem.* **2019**, *10*, 2986−2990.

(18) Friis, D. S.; Kristiansen, E.; von Solms, N.; Ramløv, H. Antifreeze activity enhancement by site directed mutagenesis on an antifreeze protein from the beetle Rhagium mordax. *FEBS Lett.* **2014**, *588*, 1767−1772.

(19) Baardsnes, J.; Jelokhani-Niaraki, M.; Kondejewski, L. H.; Kuiper, M. J.; Kay, C. M.; Hodges, R. S.; Davies, P. L. Antifreeze protein from shorthorn sculpin: Identification of the ice-binding surface. *Protein Sci.* **2009**, *10*, 2566−2576.

(20) Garnham, C. P.; Campbell, R. L.; Davies, P. L. Anchored clathrate waters bind antifreeze proteins to ice. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 7363−7367.

(21) Duboué-Dijon, E.; Laage, D. Comparative study of hydration shell dynamics around a hyperactive antifreeze protein and around ubiquitin. *J. Chem. Phys.* **2014**, *141*, No. 22D529.

(22) Schauperl, M.; Podewitz, M.; Ortner, T. S.; Waibl, F.; Thoeny, A.; Loerting, T.; Liedl, K. R. Balance between hydration enthalpy and entropy is important for ice binding surfaces in Antifreeze Proteins. *Sci. Rep.* **2017**, *7*, No. 11901.

(23) Lee, H. Structures, dynamics, and hydrogen-bond interactions of antifreeze proteins in TIP4P/Ice water and their dependence on force fields. *PLoS One* **2018**, *13*, No. e0198887.

(24) Doxey, A. C.; Yaish, M. W.; Griffith, M.; McConkey, B. J. Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat. Biotechnol.* **2006**, *24*, 852−855.

(25) Yang, R.; Zhang, C.; Gao, R.; Zhang, L. An Effective Antifreeze Protein Predictor with Ensemble Classifiers and Comprehensive Sequence Descriptors. *Int. J. Mol. Sci.* **2015**, *16*, 21191−21214.

(26) Pratiwi, R.; Malik, A. A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J. E.; Nantasenamat, C.; Shoombuatong, W. CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins. *J. Chem.* **2017**, *2017*, No. 9861752.

(27) Usman, M.; Khan, S.; Lee, J.-A. AFP-LSE: Antifreeze Proteins Prediction Using Latent Space Encoding of Composition of k-Spaced Amino Acid Pairs. *Sci. Rep.* **2020**, *10*, No. 7197.

(28) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed.; Addison-Wesley Longman Publishing Co., Inc.: USA, 1989.

(29) Baluja, S.; Caruana, R. In *Removing the Genetics from the Standard Genetic Algorithm*, Proceedings of the International Conference on Machine Learning; Elsevier, 1995; pp 38−46.

(30) Berman, H. M.; et al. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899−907.

(31) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, *77*, 778−795.

(32) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(33) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(34) Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS*, International Conference on Exascale Applications and Software; Springer: Cham, 2015; pp 3−27.

(35) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(36) Abascal, J. L.; Sanz, E.; Fernández, R. G.; Vega, C. A potential model for the study of ices and amorphous water: TIP4P/Ice. *J. Chem. Phys.* **2005**, *122*, No. 234511.

(37) García Fernández, R.; Abascal, J. L. F.; Vega, C. The melting point of ice Ih for common water models calculated from direct coexistence of the solid-liquid interface. *J. Chem. Phys.* **2006**, *124*, No. 144506.

(38) Best, R. B.; Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* **2010**, *114*, 14916−14923.

(39) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.

(40) Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A. Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. *J. Chem. Theory Comput.* **2015**, *11*, 2−7.

(41) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, No. 014101.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(43) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(44) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(45) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(46) Sicheri, F.; Yang, D. S. C. Ice-binding structure and mechanism of an antifreeze protein from winter flounder. *Nature* **1995**, *375*, 427−431.

(47) Antson, A. A.; Smith, D. J.; Roper, D. I.; Lewis, S.; Caves, L. S.; Verma, C. S.; Buckley, S. L.; Lillford, P. J.; Hubbard, R. E. Understanding the mechanism of ice binding by type III antifreeze proteins. *J. Mol. Biol.* **2001**, *305*, 875−889.

(48) Hakim, A.; Nguyen, J. B.; Basu, K.; Zhu, D. F.; Thakral, D.; Davies, P. L.; Isaacs, F. J.; Modis, Y.; Meng, W. Crystal Structure of an Insect Antifreeze Protein and Its Implications for Ice Binding. *J. Biol. Chem.* **2013**, *288*, 12295−12304.

(49) Schrödinger, L. *The PyMOL Molecular Graphics System*, version 1.8.; Schrödinger, LLC, 2015.

(50) Sun, T.; Lin, F. H.; Campbell, R. L.; Allingham, J. S.; Davies, P. L. An Antifreeze Protein Folds with an Interior Network of More Than 400 Semi-Clathrate Waters. *Science* **2014**, *343*, 795−798.

(51) Parui, S.; Jana, B. Molecular Insights into the Unusual Structure of an Antifreeze Protein with a Hydrated Core. *J. Phys. Chem. B* **2018**, *122*, 9827−9839.

(52) Yeh, Y.; Feeney, R. E. Antifreeze Proteins: Structures and Mechanisms of Function. *Chem. Rev.* **1996**, *96*, 601−618.

(53) Liou, Y. C.; Tocilj, A.; Davies, P. L.; Jia, Z. Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein. *Nature* **2000**, *406*, 322−324.

(54) Midya, U. S.; Bandyopadhyay, S. Role of Polar and Nonpolar Groups in the Activity of Antifreeze Proteins: A Molecular Dynamics Simulation Study. *J. Phys. Chem. B* **2018**, *122*, 9389−9398.