Journal of
Materials Chemistry B

ROYAL SOCIETY
OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *J. Mater. Chem. B*, 2026, **14**, 4059

Discovering naturally occurring antifreeze peptides from microbiome by integrating protein language models and molecular dynamics simulation

Ibrahim A. Imam,†[a] Trevor Morey,†[ab] Yuexu Jiang,[ac] Duolin Wang,[c] Dong Xu[c] and Qing Shao*[a]

Received 10th December 2025,
Accepted 4th March 2026

DOI: 10.1039/d5tb02758f

rsc.li/materials-b

Antifreeze peptides inhibit ice crystal growth and recrystallization, and are promising components of cryoprotective formulations for cell, tissue, and food preservation, as well as anti-icing surface coatings. However, the discovery of new antifreeze peptides has been hindered by their sequence diversity and the limited scalability of experimental screening. In this study, we identify novel antifreeze peptide candidates from a microbiome-derived sequence library using ensemble machine learning and molecular dynamics (MD) simulations. We developed an ensemble classifier composed of 10 adapter-tuned protein-language models and a random forest meta-learner. After training on a curated dataset of 73 766 sequences, we applied this ensemble to 56 008 amino acid sequences from an Arctic microbiome library to identify antifreeze peptide candidates. Structural prediction yields a diverse range of conformations for six selected candidates, including α-helices, coils, and combinations of both. To evaluate their functional relevance, atomistic MD simulations were conducted to assess conformational stability and solvent interactions under freezing conditions. One candidate shows persistent helicity, surface amphipathicity, and an organized hydration pattern consistent with structural signatures reported for ice-binding helices. These findings expand the known landscape of antifreeze peptides and highlight a scalable strategy for discovering functional peptides from complex biological sources.

1. Introduction

Antifreeze proteins (AFPs) and peptides play a crucial role in various applications, including cell cryopreservation, food preservation, and vaccine delivery.[1–5] The uncontrolled nucleation and propagation of ice crystals can cause cellular damage and compromise the structural integrity of materials.[6–8] Such uncontrolled ice crystallization must be addressed to ensure the system can perform its desired function. Diverse organisms evolved ice-binding proteins (IBPs) and their smaller peptide derivatives, antifreeze peptides, to non-colligatively inhibit ice formation.[9–12] These peptides employ a thermal hysteresis mechanism to depress freezing points by several degrees while minimally affecting melting points.[13–15] Antifreeze peptides offer design flexibility as biomimetic scaffolds for next-generation materials, thanks to their relatively small size, high stability, and potent, structurally defined ice-binding properties.[3,16–18]

The functional potency of AFPs arises from their structural diversity and distinct mechanisms of interaction with ice.[19–22] These proteins and peptides are broadly classified into four types based on their structures: (I) short, alanine-rich α-helices defined by a canonical Thr-Ala-X repeating motif that organizes their ice-binding surface (IBS);[23,24] (II) globular, cysteine-rich with complex, non-repetitive IBS stabilized by multiple disulfide bridges;[25] (III) globular with a compact β-sandwich fold to project a flat and highly repetitive IBS;[26,27] and (IV) glutamate and glutamine-rich, composed of multiple α-helices, and featuring a structurally irregular ice-binding surface compared to other types.[28,29] Despite this structural variance, their function converges on a shared mechanistic principle: irreversible adsorption to specific ice crystal planes *via* well-defined, flat, and often amphipathic IBS.[30] This adsorption locally increases the curvature of the ice-water interface *via* the Kelvin effect, inhibiting further crystal growth and recrystallization.[31] While

[a] *Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY 40506, USA. E-mail: qshao@uky.edu*
[b] *Department of Chemistry, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA*

† The two authors contribute equally to this paper.

this structural arrangement determines antifreeze activity, reliance on specific motif patterns such as regularly spaced threonine/alanine arrays on flat ice-binding surfaces constrains the range of biophysical properties (*e.g.*, solubility, thermal stability, ice-binding kinetics) available for material design.[23]

Efforts have been focused on discovering AFPs and antifreeze peptides from bioprospecting organisms in extreme environments. Indeed, this approach has successfully identified numerous antifreeze peptides through phenotypic screening and sequence homology searches.[32,33] The focus is on peptides sharing homology with canonical ice-binding motifs. Exploring uncharacterized sequence space through computational prediction, therefore, offers a powerful complementary strategy to discover peptides with potent ice-inhibiting properties that do not conform to these established structural classes. Recent machine-learning predictors have accelerated AFP discovery by learning sequence-derived representations beyond motif homology.[34,35] Such novel sequences could possess superior biophysical profiles, thereby overcoming material design challenges such as limited solubility, aggregation propensity, and suboptimal thermal stability. Exploring such potential might require moving beyond homology-based methods toward models that identify function independent of sequence similarity.

Protein language models (PLMs) have transformed the process of discovering proteins and peptides with desired properties by learning deep contextual representations directly from sequence data.[36–38] These models vary in their design and training objectives. ProtBERT and ProtT5, members of the ProtTrans[39] family exemplifies this diversity. ProtBERT is optimized for masked language modelling, and ProtT5 employs an encoder–decoder framework for generative tasks. Beyond ProtTrans, other architectures explore distinct directions. ProteinBERT[40] integrates functional annotations into its training to enhance biological interpretability, while ProGen[41] adopts a generative GPT-style architecture to design novel proteins. The Evolutionary Scale Modeling (ESM)[42] family, in contrast, was trained on massive, diverse sequence databases specifically to capture deep evolutionary context and co-variation signals, which are correlated with protein function. The capacity to learn functional representations directly from sequence, without reliance on homology, sets the pace for developing powerful pipelines for screening vast, unannotated metagenomic databases.

ESM2[43] has shown its extraordinary ability to use rotary positional embeddings to efficiently capture long-range amino acid dependencies.[44–46] The 650-million parameter variant (ESM2-650M) provides an optimal balance between predictive accuracy and computational throughput for large-scale screening.[47] However, the ESM2 model presents a significant challenge. Full fine-tuning is not only computationally costly but also risks model degradation by overwriting the powerful, generalizable representations learned during pretraining. To strike a balance between specialization and efficiency, we employed parameter-efficient adapter-based fine-tuning.[48] The strategy preserves the model's core knowledge by freezing the 650M-parameter backbone and training only a small, specialized set of adapter weights. This enables the development of a downstream binary classifier to identify functional peptides.

Ensemble learning is a method that aggregates predictions from multiple independent models to improve overall performance.[49] An ensemble of models can attenuate the bias of individual models. Common approaches to developing an ensemble of models include voting, where predictions are aggregated *via* majority voting or probability averaging; bagging, which trains models on different data subsets to reduce variance; boosting, which sequentially corrects errors from previous models; and stacking, which uses a *meta*-learner to optimally weight individual predictions. These strategies have been applied to the prediction of functional peptides. Studies have reported improved accuracy using ensembles for antimicrobial peptides,[50] anticancer peptides,[51] antibiofouling peptides,[47] and neuropeptides.[52] For antifreeze peptide discovery, where experimentally validated examples are scarce and structural diversity is high, ensemble methods provide a practical way to maximize prediction reliability.

All-atom molecular dynamics (MD) simulation is a powerful method for further analysing the potential of deep learning predicted candidates and characterizing antifreeze protein (AFP) and peptide mechanisms at an atomic resolution that static structural methods cannot access.[53–55] Pioneering studies used MD to elucidate how Type I AFPs pre-order water on their IBS to template ice recognition,[56–58] to quantify the thermodynamic contributions of individual residues to adsorption,[59,60] and to demonstrate that antifreeze glycoproteins bind *via* dynamic hydrogen-bonding networks rather than rigid complementarity.[61,62] Subsequent work has used MD to resolve the hydration dynamics governing binding affinity,[63,64] to define the molecular basis of hyperactive *versus* moderate activity,[65] to identify conformational transitions at the ice interface,[66] and to validate the mechanisms of computationally designed peptides.[67] These studies have established MD as a versatile computational framework for the rational design of AFPs and antifreeze peptides, enabling assessment of structural descriptors like stability, amphipathic profile, and solvent interaction required for ice-binding function.

In this study, we integrated PLM adaptation, ensemble learning and MD simulation to identify and characterize novel antifreeze peptide candidates from a library of amino acid sequences identified in the Arctic microbiome. We first compiled a database of amino acid sequences derived from the Arctic microbiome. We then applied an ensemble of adapter-tuned ESM2-650M classifiers to identify high-confidence candidates with physicochemical properties distinct from those of the known AFP motifs. These candidates were then subjected to a two-step physics-based *in silico* procedure for further analysis of their chemicophysical properties. In the first stage, we predict their possible 3D structures using AlphaFold3. In the second stage, we perform all-atom MD simulations to analyse their conformational stability, quantify their solvent exposure and amphipathic character, and unveil their hydrogen bonding patterns. This computational analysis provides a comprehensive structural and dynamic filter, allowing us to prioritize the candidates and identify a prime candidate for experimental validation. The rest of the paper is organized as follows. Section 2 will focus on the data construction, deep learning methods, and MD simulation details. Section 3 will focus on the results and discussion. Section 4 will present a summary.

2. Methods

2.1. Dataset construction

Protein sequences used for model development were sourced from the UniProt[68] Database. Antifreeze protein (AFP) sequences were retrieved from UniRef[69] using the keyword "antifreeze," followed by manual verification against functional annotations to reduce false positives. Non-AFP sequences were selected from UniProtKB/SwissProt[70] by excluding entries annotated with antifreeze-related taxonomy or function, thereby minimizing inadvertent inclusion of uncharacterized AFPs.

To mitigate bias arising from unequal sequence lengths between AFP and non-AFP proteins, all sequences beyond a maximum length of 200 residues were excluded. This cutoff was chosen based on the observed distribution of AFPs to avoid introducing sequence-length bias into the classifier. In total, 73 766 sequences were used for model development. A balanced subset of 1100 AFP and non-AFP sequences was reserved for training and evaluating the random forest-based ensemble, with a stratified split into training and testing sets at a 4 : 1 ratio.

For adaptation of the protein language models (PLMs), the remaining 6506 AFP and 65 060 non-AFP sequences were partitioned into ten balanced groups. Non-AFP sequences were randomly divided into groups of 6506 sequences each, and the complete AFP set was added to every group to ensure consistent representation and reduce class-imbalance bias. Each group was subsequently clustered using MMSeqs2, and the resulting clusters were split into training, validation, and testing subsets in a 70 : 15 : 15 ratio. Sequence similarity across subsets was restricted to below 30% (global identity), thereby reducing redundancy and ensuring robust evaluation of model generalization. A detailed summary of dataset composition and splitting for all groups is provided in Table S1.

2.2. Adapter tuning ESM-2 model for binary classification

To enable efficient adaptation of the ESM2-650M model for antifreeze protein (AFP) classification, we employ adapter tuning (Fig. 1), a parameter-efficient fine-tuning strategy that injects lightweight, trainable bottleneck modules into the transformer architecture while preserving all pre-trained weights frozen. The adapter modules are implemented according to Houlsby's study.[48] They are positioned twice in one Transformer layer of ESM2: after the self-attention projection and after the two feedforward. The adapter consists of a two-layer feed-forward network: a down-projection compresses the input data into a reduced-dimensional space, followed by a GELU non-linearity, and an up-projection restore it to the original input dimension, with a residual connection added to the original transformer output. This design enables task-specific modulation of internal representations without altering the model's general protein knowledge, thereby mitigating overfitting and preserving transferability. In our implementation, instead of applying adapter modules in all ESM2 layers, we insert them into the top-K Transformer layers of ESM2. K is a hyperparameter equal to 16. In Table 1, we summarize other hyperparameters used in our model.

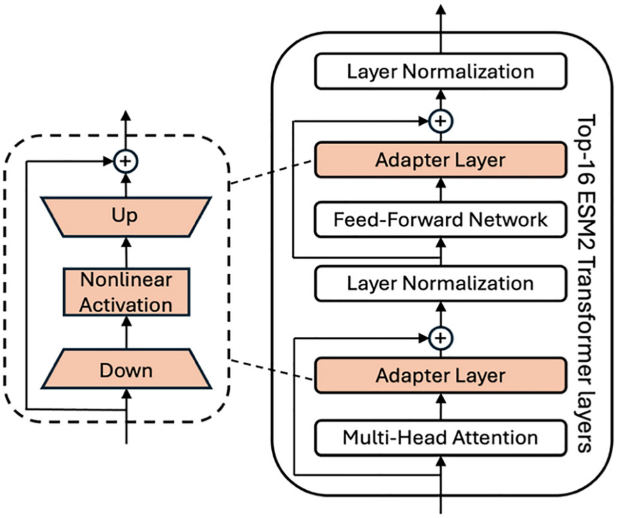

Fig. 1 Schematic of adapter-tuning of the ESM2 classifiers.

The final representation for classification is obtained by averaging the hidden states across all residue positions in each sequence, yielding a fixed-length embedding. This pooled representation is then passed through a single-layer feed-forward classifier with sigmoid activation to produce a binary probability score for AFP. Training is conducted for a maximum of 50 epochs with early stopping based on validation loss, using the curated dataset described in Section 2.1 for monitoring.

2.3. Model ensemble

A random forest (RF) binary classifier was constructed to integrate the ensemble outputs of the ten adapted ESM2-650M models. The RF model received as input the probability scores generated by each adapted ESM2 and produced a final output on a continuous scale from 0 to 1, reflecting the antifreeze potential of the peptides (classified as non-AFP if <0.5 and AFP if ≥ 0.5). The RF classifier was trained and evaluated on the balanced subset of 1100 AFP and 1100 non-AFP sequences described in Section 2.1 and implemented in

Table 1 The summary of model configuration

Hyperparameters	Values
Max seq length	220
Adapter layer number	16
Epoch number	50
Batch size	32
Optimizer	Adam
Learning rate	0.00001

Scikit-learn.[71] Hyperparameters were optimized to balance accuracy and generalization, with the final model employing 188 estimators, a maximum tree depth of 6, a minimum split size of 9, and a minimum leaf size of 2. This implementation was adapted from Imam *et al.*[47]

2.4. Molecular dynamics simulation

2.4.1. Molecular model. The 3D structures of the six selected antifreeze peptides candidates were predicted using the AlphaFold3[72] webserver and visually inspected to ensure structural integrity. Protonation states of titratable residues were assigned at physiological pH 7.0 using the H++ webserver.[73] The GenIce2[74] toolkit was used to construct the hexagonal ice slab representing the basal plane of ice Ih, with its crystallographic *c*-axis oriented normal to the interface.

Each simulation box was arranged along the *z*-axis in a solvent-ice slab-solvent configuration, where the ice layer was sandwiched between two water phases to mimic a quasi-symmetric environment. Along the *z*-axis, the box length was chosen such that the water–ice–water slab was separated from its periodic images by a 3 nm thick vacuum buffer region, preventing interactions between replicated interfaces. The peptide was placed in the solvent 1.5 nm away from the ice surface to prevent steric interference while allowing spontaneous peptide diffusion toward the interface during the simulation run. Fig. 2 presents a representative snapshot of the simulation box, while Table S2 summarizes the dimensions and composition of all simulated systems.

The basal-plane ice-water interface was chosen as a well-characterized, structurally simple model system for probing peptide structure and hydration under sub-zero conditions, rather than as a detailed model of the preferred binding plane of a particular AFP. The basal-plane interface serves as a generic, strongly structured hydration environment for stress-testing the conformational stability of the helical peptide candidates.

All peptide molecules were described using the Amber14SB[75] all-atom force field, which reliably represents both backbone and side-chain conformations. The TIP4P/Ice[76] water model was used to describe the ice and solvent molecules, as it is well-suited for simulations involving ice-water phase equilibria. To maintain overall charge neutrality, sodium (Na^+) and chloride (Cl^-) counter-ions were added to each system.

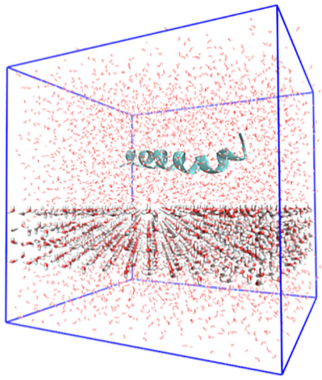

Fig. 2 Representative snapshot of the simulation system. The peptide candidate (cyan) is 1.5 nm away from the ice-slab (licorice) and solvated with water molecules (red).

Non-bonded interactions were calculated as the sum of a Lennard-Jones 12-6 potential and a Coulombic potential, as shown in eqn (1):

$$E_{ij}(r_{ij}) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{e_i e_j}{4\pi\varepsilon_0 r_{ij}} \qquad (1)$$

where E_{ij} is the potential energy between atoms i and j due to non-bonded interactions. r_{ij} is the distance between atoms i and j, ϵ_{ij} is the energetic parameter, σ_{ij} is the geometric parameter, and e_i is the partial charge of atom i.

Bonded interactions, including bond, angle, and dihedral terms, were quantified according to the AMBER force field specifications. Visual Molecular Dynamics (VMD) software[77] was used for comprehensive visual analysis.

2.4.2. Simulation details. All MD simulations were performed using GROMACS 2023.2.[78] Periodic boundary conditions were applied in all three dimensions. Each system was first energy-minimized using the steepest-descent algorithm to remove too close atomic contacts. During all simulations, the oxygen atoms of the ice lattice were position-restrained to maintain the crystalline framework while allowing hydrogen atoms and interfacial water molecules to relax. Molecular dynamics simulations were then carried out in the canonical (NVT) ensemble for 400 ns using the leapfrog integrator with a 2-fs time step.

To maintain a solid ice slab in contact with a supercooled aqueous region under sub-zero conditions, the system was coupled to two independent thermostats: the ice layer was maintained at 200 K, and the liquid water, ions, and peptide (non-ice) were maintained at 250 K. This dual-thermostat scheme keeps the ice slab well below its model melting point while maintaining the surrounding water in a supercooled state with sufficient mobility to sample peptide conformational dynamics at the ice-water interface. Temperature control in both groups employed the V-rescale thermostat[79] with 0.1 ps time constant to ensure stable regulation. All covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm.[80] Lennard-Jones interactions were truncated at 1.2 nm, and long-range electrostatics were computed using the particle-mesh Ewald[81] (PME) method.

Trajectory coordinates were saved every 100 ps, and all analyses were performed on the final 350 ns of each trajectory, which corresponds to the period after an initial equilibration phase where the system reached stability, as confirmed by convergence of temperature and potential energy.

Because the oxygen atoms in the ice slab were position-restrained and separate thermostats were applied to the ice and water-peptide regions, these simulations represent a quasi-static model of an ice-water interface. They are therefore used to probe peptide conformational behaviour and hydration under sub-zero interfacial conditions, rather than to provide quantitative information on ice growth or melting kinetics. Post-processing and quantitative analyses were performed using both GROMACS analysis utilities and custom Python scripts developed in-house to automate time averaging of structural and hydration descriptors and to assess convergence.

To further assess the specific influence of the ice-water interface, we also performed control simulations of all six peptides in bulk supercooled water at 250 K (no ice slab). These bulk-water systems contained only the peptide, water, and ions, and were run with simulation settings identical to those used for the ice–water interface trajectories.

3. Results and discussion

3.1. Model evaluation

The RF-based ensemble integrating the ten adapted ESM-2 models demonstrated robust and balanced performance in AFP classification. Average F1 scores for the individual ESM-2 models were 0.898 (non-AFP) and 0.892 (AFP), and standard deviations of 0.009 and 0.008, respectively. The RF classifier improved overall accuracy, achieving F1 scores of 0.947 and 0.944, with precision of 0.926 (non-AFP) and 0.967 (AFP), and recall of 0.968 (non-AFP) and 0.923 (AFP). These metrics imply that the ensemble does not favour one class disproportionately and maintains a strong precision–recall balance across AFP and non-AFP sequences.

The confusion matrix (Fig. 3) further illustrates this balance: 213 non-AFP sequences and 203 AFP sequences were correctly classified, with only 24 total misclassifications. The ensemble achieved an overall accuracy of 0.945 while maintaining balanced performance across both classes, making it a reliable screener for identifying novel antifreeze peptide candidates. The ensemble's balance is critical for downstream applications, where minimizing bias ensures that promising peptide candidates are not overlooked.

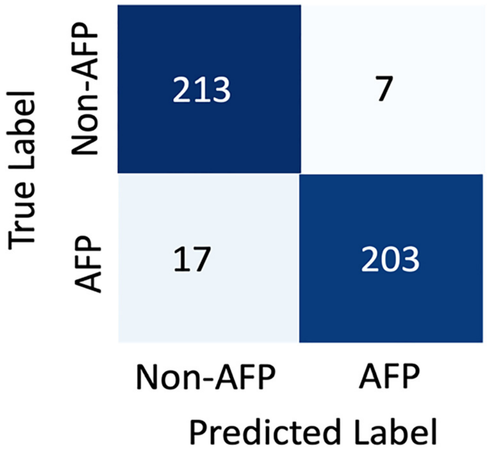

Fig. 3 Confusion matrix showing the performance of the ensemble model on AFP classification, comparing predicted and true labels for AFP and non-AFP sequences.

3.2. Ensemble screening

The ensemble-based high-throughput screening involved three steps. First, the RF-based model ensemble was applied to 56 008 protein sequences derived from an Arctic microbiome[82] library, selected because organisms in extreme cold environments are likely to possess proteomes enriched in antifreeze proteins. This search identified 11 286 candidate antifreeze sequences. Second, the candidate list was refined to retain only sequences with 30–50 amino acid residues, yielding 144 candidates. Third, these 144 sequences were subjected to statistically informed outlier-based filtering to identify candidates most distinct from the AFP training distribution.

To evaluate statistical distinctness, sequences were embedded in a five-dimensional physicochemical descriptor space using the E-descriptor system defined by Venkatarajan and Braun.[83] The system derives five principal components (E_1–E_5) from multidimensional scaling of 237 physicochemical parameters, where E_1 represents hydrophobicity, E_2 captures molecular size and steric properties, E_3 encodes α-helix propensity, E_4, relates to partial specific volume, and E_5 indicates β-strand propensity. The vector value of these E-descriptors for each amino acid are presented in Table S3. For every peptide, per-residue E-values were averaged to yield a single five-component vector describing its position in descriptor space.

E-Descriptor vectors were first computed for AFPs in the training dataset, from which the mean vector and covariance matrix in E-space were estimated. Model-predicted peptide candidates were then embedded in this same space, and the Mahalanobis distance from each candidate to the training distribution mean was calculated. To assess whether candidates represented statistically significant outliers, deviations were evaluated using an upper-tail chi-square test with 5 degrees of freedom, corresponding to the five-dimensional E_1–E_5 representation. Sequences with significant deviation ($p \leq 0.01$) were flagged as physicochemically distinct from the characterized AFP training set.

By focusing on statistically significant outliers in this five-dimensional E-descriptor space, we prioritized candidates that occupy physicochemical regions underrepresented among characterized AFPs. This strategy is designed to probe the periphery of known AFP chemical space and potentially reveal novel sequence solutions to antifreeze function, while recognizing that such outliers may also include false positives that require further computational and experimental scrutiny.

Table 2 presents the six (6) peptide sequences that satisfied this statistical criterion. For visualization, principal component analysis (PCA) was performed on the E-embeddings. The first three components captured more than 75% of the variance, and the selected sequences occupied marginal regions relative to the training cloud. Fig. S1 shows the 3D PCA projection of the AFP training set, the overlapped model-predicted peptides, and the six diverse outlier candidates, highlighting their peripheral positioning and potential as unconventional AFP leads.

Table 2 Amino acid sequences of six structurally diverse antifreeze peptide candidates

#	Amino acid sequence	Sequence length
1	MGNEQKQHHEPEREEHRQKPEEEKPQTWKHPDDGTELSERDQERPLKP	48
2	MVVIMTAVMITNVVMTVAMIGDMTMTAGVIAMVETRIIAGKLLQ	44
3	MLSLNGCTVLAIADVAVATTVKVGGAVVGTAVDVTKAGVGAVTGSAAK	48
4	MSKDRKTVKEVKKQPTVNVNKKQSAYQSGKGSASSDLGKK	40
5	MKKIYISAAVLLAVAITEGCIKQRVAESSAAKHISVRKAL	40
6	MLLIMALITSVQTTLLLMVLTTFLLTGLLLTALITWALKA	40

3.3. Peptide conformational dynamics

AlphaFold3 predictions reveal three distinct secondary structural patterns among the six peptides (Fig. 4). Peptides #2, #3, #5, and #6 adopt extended α-helical conformations spanning their entire length, whereas peptide #4 displays a central α-helix with disordered N- and C-termini. Peptide #1 exhibits a predominantly disordered conformation containing short helical segments. The extended helical secondary structures observed in peptides #2, #3, #5, and #6 are structurally consistent with type I AFP-like scaffolds, which often present amphipathic helices implicated in ice binding to prism/pyramidal faces.[84,85] The structural heterogeneity among these candidates, particularly the contrasting conformations of peptides #1 and #4 compared to the fully helical peptides, may indicate either a distinct functional mechanism or sequence variants with differing structural stabilities. Here, MD is primarily used as a stability/solvation filter at a standardized ice–water interface, not as evidence of plane-specific adsorption.

The biophysical properties and dynamic behaviour of these peptides were investigated using 400 ns all-atom MD simulations. Secondary structure evolution was analysed using the Dictionary of Secondary Structure of Proteins (DSSP) algorithm,[86] which assigns structure based on backbone hydrogen bonding patterns (Fig. 5). This revealed dynamic diversity in conformational stability, particularly within the main helical class. Peptides #2 and #5, demonstrated notable conformational rigidity, maintaining end-to-end α-helical structure, with minimal termini fraying, throughout the simulation. In contrast, peptides #3 and #6, also predicted to be fully helical, exhibited different dynamic profiles. Peptide #3 retained its central helical core but displayed a clear disorder at both its N- and C-termini. Peptide #6 initially adopts a stable helix comparable to peptides #2 and #5, after which it exhibits structural instability, marked by the partial unfolding of its C-terminal residues into a disordered coil after approximately 280 ns. The simulations also affirmed the non-canonical predictions. Peptide #4 maintained its stable central α-helix while its termini remained highly disordered, and peptide 1 behaved as a dynamic disordered system, sampling mostly helical segments and irregular secondary structures. The DSSP dynamic profiles reveal a hierarchy of structural stability within the helical class. The rigid architectures of peptides #2 and #5 contrast sharply with the conditional stability of peptides #3 and #6 and the persistent disorder of peptide #1. Such diversity in conformational behaviour may reflect distinct stability and amphipathic presentation relevant for downstream validation. To illustrate the structural diversity among predicted peptides, representative MD snapshots at multiple time points are provided for two contrasting candidates, peptide #1 and peptide #5 (Fig. S2), showing both peptide conformational evolution and the development of interfacial water structuring during equilibration.

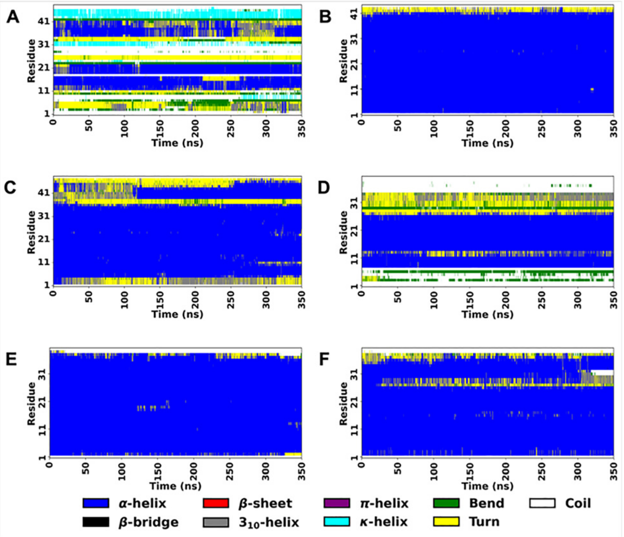

Fig. 5 Secondary structure evolution of (A) peptide #1, (B) peptide #2, (C) peptide #3, (D) peptide #4, (E) peptide #5, and (F) peptide #6, during MD simulations.

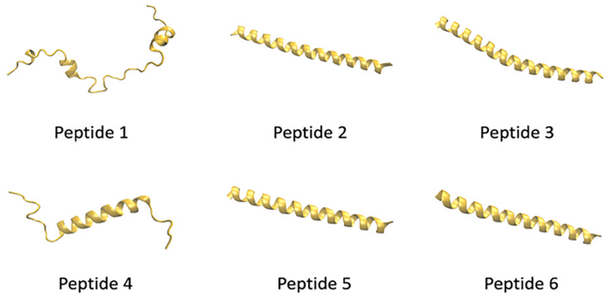

Fig. 4 AlphaFold3 structural predictions for six candidate peptides identified by the ensemble model.

To assess peptide proximity to the ice interface, we quantified the peptide–ice separation (interfacial gap) along the slab normal throughout the trajectories (Fig. S3). All peptides begin at an initial separation of approximately 1.4–1.5 nm and relax over the first 100–150 ns to separations of 1.9–2.2 nm, after which the distance fluctuates around this larger value (Fig. S3). This indicates that there was no persistent adsorption of any of the candidates on the restrained basal surface. This absence of spontaneous approach on the basal plane is consistent with expectations for α-helical candidates and is not interpreted here as evidence against antifreeze potential.

3.4. Peptide surface properties

Solvent exposure patterns reflected the conformational stability hierarchy among the peptides. The solvent exposure of each peptide was quantified by computing the distribution of total solvent-accessible surface area (SASA) over the trajectory (Fig. 6A). This analysis revealed distinct populations corresponding to the different peptide candidates. Peptide #1 exhibited the largest mean SASA (approximately 64 nm^2) with a heterogeneous distribution. Peptides #2 and #5 show a narrow SASA distribution, centred on the lower mean SASA values of about 42 nm^2 and 41 nm^2, respectively. The remaining candidates (peptides #3, #4, and #6) populated an intermediate range, with mean SASA values between 43–46 nm^2, consistent with their partial terminal unfolding or flexibility.

These SASA distributions reflect the peptides' conformational dynamics and helical stability. The large distributed SASA of Peptide #1 is a biophysical effect of its disordered, solvent-exposed conformational ensemble. At the opposite end of the stability spectrum, the minimal SASA values for peptides #2 and #5 imply the formation of compact, stable helical structures that minimize solvent interfacing through efficient residue packing. The intermediate values of peptides #3, #4, and #6 are also non-random; they reflect the partial disorder (flexible termini or C-terminal unfolding) observed in the DSSP analysis (Fig. 5) in Section 3.3, which increases their average solvent interface.

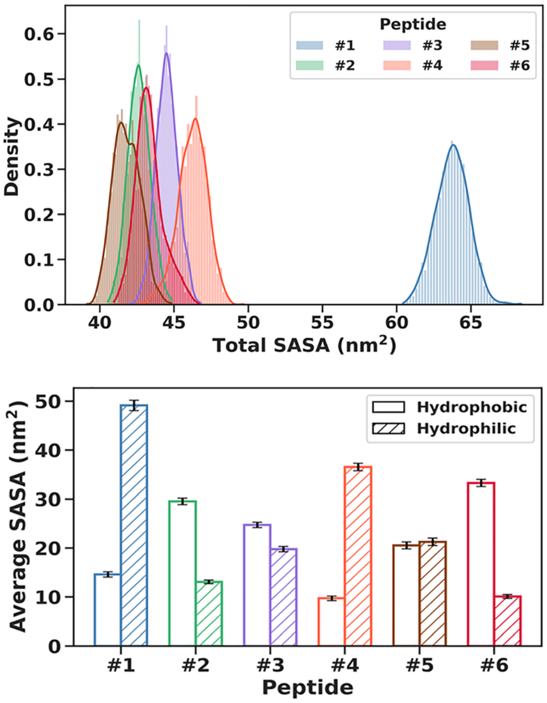

Fig. 6 Peptide solvent accessibility; (A) distribution of total SASA across peptides. (B) Average solvent exposure of hydrophobic and hydrophilic residues.

Decomposition of SASA into hydrophobic and hydrophilic components revealed two surface chemistry profiles among the peptides (Fig. 6B). Peptides #1 and #4, both structurally disordered, exposed predominantly hydrophilic surfaces. Peptide #1 demonstrated 14.6 : 49.1 nm^2 hydrophobic-to-hydrophillic exposed area, while Peptide #4 showed 9.7 : 36.5 nm^2 exposure. The remaining peptides exposed greater hydrophobic surface area, with varying hydrophobic-to-hydrophilic ratios. Peptide #6 displayed the largest hydrophobic exposure (33.3 : 10.1 nm^2 hydrophobic to hydrophilic exposure), followed by Peptide #2 (29.6 : 13.1 nm^2) and Peptide #3 (24.7 : 19.8 nm^2). These variations in surface composition reflect the interplay between sequence composition and conformational stability.

Peptide #5 diverged clearly from both classes, presenting balanced hydrophobic-to-hydrophilic ratio (20.6 nm^2 *vs.* 21.3 nm^2). For a stable and extended helical structure, this near-equal distribution indicates spatial segregation of hydrophobic and hydrophilic residues onto opposing helical faces. Such amphipathic organization is commonly observed in type I antifreeze proteins, where the balanced surface architecture facilitates oriented binding to ice crystal planes. The combination of sustained helical stability and balanced surface chemistry distinguishes peptide #5 from the other candidates and suggests structural compatibility with ice-binding function, though experimental validation remains necessary to confirm activity. Matched bulk-water control simulations at 250 K reproduce highly similar total SASA distributions and hydrophobic/hydrophilic SASA partitioning, preserving the same peptide ordering (Fig. S4A and B), indicating no systematic interface-induced shift in solvent exposure.

3.5. Peptide–solvent interactions

To evaluate the structural solvation of the predicted peptide, hydrogen bonds with surrounding water molecules were quantified. The hydrogen bond was defined based on the acceptor–donor distance (<0.35 nm) and angle ($<30°$) criteria proposed by Chandler's group.[87] The average number of hydrogen bonds formed between peptide and surrounding water molecules is partitioned into side-chain and backbone contributions. The goal was to differentiate polar surface hydration from internal structural organization across the six predicted candidates. Consistent with the SASA controls, backbone–solvent and side-chain–solvent hydrogen-bond counts are similar between slab and bulk-water simulations (Fig. S4C), indicating that the peptide hydration is not measurably perturbed by the presence of the restrained ice interface under our conditions.

As shown in Fig. 7, Peptide #1 exhibited the highest solvent interaction, forming approximately 137 sidechain and 76 backbone hydrogen bonds. This reflects the extended, disordered structure earlier attributed to peptide #1, which lacks internal hydrogen bonding and exposes most polar and nonpolar group. Peptide #4 formed around 73 sidechain and 51 backbone hydrogen bonds with surrounding solvent molecules,

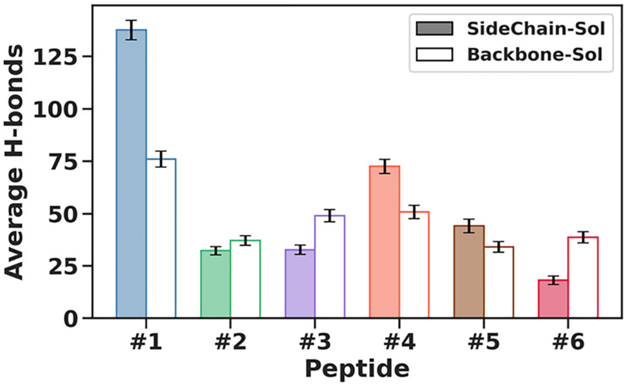

Fig. 7 Average number of hydrogen bonds formed by backbone atoms (C-α, C, N, O) with water molecules compared to those formed by side-chain atoms.

indicating a flexible conformation with partial internal organization.

In contrast, peptides #2, #3, #5, and #6 showed lower backbone hydration, ranging from 34 to 49 hydrogen bonds. This reduction reflects the formation of intramolecular backbone hydrogen bonds in helical or ordered structures, which limits backbone accessibility to solvent. Peptide #6 showed the lowest side-chain hydration at 18 hydrogen bonds, indicating a predominant hydrophobic surface character with minimal polar–solvent interaction. Peptide #3 maintained higher side-chain hydration at around 33 hydrogen bonds, likely reflecting its disordered termini or exposed loop regions identified in the DSSP analysis (Section 3.3). Peptide #2 displayed 32 sidechain and 37 backbone hydrogen bonds. This suggests moderate surface polarity within a stable helical framework.

Peptide #5 displayed a highly distinct hydration profile, forming approximately 44 sidechain and 34 backbone hydrogen bonds. This moderate asymmetry highlights a critical structural equilibrium: the peptide core remains effectively shielded, while polar residues maintain interactions with the solvent. This hydration pattern serves as a key indicator of surface polarity, suggesting amphiphilic nature that is consistent with structural pattern that could, in principle, support an ice interaction in appropriate crystallographic planes.

While these simulations do not directly measure ice-binding activity, the structural properties of peptide #5 like stable helical structure, balanced solvent exposure, and side-chain-dominated hydration consistent with organized polar surface regions, aligns with the biophysical features characteristic of functional Type I antifreeze proteins. This convergence of structural and hydration signatures supports peptide #5 as a high-priority candidate for experimental ice-binding testing.

Conclusions

This work presents an integrated machine learning-molecular dynamics (ML – MD) workflow for the discovery of antifreeze peptides from Arctic microbiome-derived sequence libraries. By combining adapter-tuned ESM2 protein language models into a stacked ensemble with a random forest *meta*-learner, we achieve robust classification performance, enabling the high-throughput identification of physicochemically diverse candidates. Subsequent structural characterization using AlphaFold3 and all-atom molecular dynamics simulations provided a biophysical filter to evaluate conformational stability, solvent exposure, and peptide–water hydrogen-bonding patterns under sub-zero conditions. Among the identified candidates, peptide #5 exhibits a rigid, stable helical core, balanced solvent exposure, and moderate, selective side-chain hydration. This profile aligns with the structural features of functional type I AFPs and supports peptide #5 as a priority target for experimental validation.

This study demonstrates the utility of combining protein language models with MD simulations for functional peptide discovery. The presented approach enables rapid screening of large unlabelled peptide datasets and provides atomic-level insights into structure–function relationships. Future extensions using enhanced sampling or explicit simulations ice nucleation resistance may refine our understanding of antifreeze mechanisms and guide rational peptide optimization for cryopreservation and materials science applications. Because α-helical AFPs often bind non-basal planes, future work will also examine prism and pyramidal interfaces and may compute adsorption free energies to assess plane specificity. Experimental studies will be crucial for validating the predicted peptide candidates and for refining the computational design rules that emerge from this work.

Author contributions

Ibrahim A. Imam: writing – original draft; data curation and analysis; investigation; methodology; visualization. Trevor: writing – review and editing, methodology; data curation and analysis. Yuexu Jiang: writing – review and editing, methodology. Duolin Wang: methodology. Dong Xu: writing – review and editing, methodology; funding acquisition. Qing Shao: conceptualization; writing – review and editing; formal analysis; supervision; methodology, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

The datasets and source code supporting the findings of this study are available at the following GitHub repository: https://github.com/imamabi/Antifreeze-Peptide-Discovery. The simulation input files and analysis data are available from the corresponding upon request.

Supplementary information: Table S1 Summary of AFP and non-AFP sequence counts; Table S2 The details of initial system setup; Table S3 The 5 vector E-descriptors of each amino acid proposed by Venkatarajan & Braun; Fig. S1 The 3D PCA visualization of AFP predicted peptides overlapped with AFP from the training data; Fig. S2 Representative VMD snapshots of the ice–water slab simulation for peptide #5 at 0, 50, 100, 200, and 400 ns for (A) peptide #1, (B) peptide #5. The restrained basal-plane ice slab maintains an ordered lattice, while the adjacent liquid water exhibits interfacial ordering; Fig. S3 Peptide-ice interfacial gap (surface-to-surface separation along the slab normal, z) as a function of simulation time for the six peptide candidates in the basal-plane ice–water slab system (peptides initially placed $\sim$1.5 nm from the ice surface); Fig. S4 Comparison of key structural/solvation descriptors between peptide–ice–water slab simulations and matched bulk-water controls. Left panels: basal-plane ice–water slab; right panels: bulk water (no ice), both at 250 K for the liquid/peptide region (ice lattice restrained at 200 K in slab runs). Metrics computed over the production window ($t \geq 50$ ns). (A) Distributions of total solvent-accessible surface area (SASA). (B) Mean hydrophobic and hydrophilic SASA components. (C) Mean peptide–water hydrogen-bond counts partitioned into backbone–water and sidechain–water contributions. See DOI: https://doi.org/10.1039/d5tb02758f.

Acknowledgements

We acknowledge the National Institutes of Health (grant R01LM014510) for the financial support. D. W and D. X are also partially supported by the National Institutes of Health (grant R35GM126985). T. M and Q. S also acknowledge the support of the National Science Foundation (NO. 2150337). We also thank the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for supporting and using associated research computing resources. This work also used Delta-GPU at NCSA through allocation CIS230053 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which National Science Foundation supports grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1 B. Xia, J. T. Wang, H. H. Chen, S. Y. Lin, B. C. Pan and N. Wang, *Molecules*, 2024, **29**, 4913.
2 A. Eskandari, T. C. Leow, M. B. A. Rahman and S. N. Oslan, *Biomolecules*, 2020, **10**, 1649.
3 D. Zhang, H. Chen, Y. X. Zhang, J. T. Yang, Q. Chen, J. Wu, Y. L. Liu, C. Zhao, Y. J. Tang and J. Zheng, *Chem. Soc. Rev.*, 2025, **54**, 5292–5341.
4 I. K. Voets, *Soft Matter*, 2017, **13**, 4808–4823.
5 H. S. Qi, Y. H. Gao, L. Zhang, Z. X. Cui, X. J. Sui, J. F. Ma, J. Yang, Z. Q. Shu and L. Zhang, *Engineering*, 2024, **34**, 164–173.
6 T. Chang and G. Zhao, *Adv. Sci.*, 2021, **8**, 2002425.
7 J. Saragusty, H. Gacitua, I. Rozenboim and A. Arav, *Biotechnol. Bioeng.*, 2009, **104**, 719–728.
8 S. Bojic, A. Murray, B. L. Bentley, R. Spindler, P. Pawlik, J. L. Cordeiro, R. Bauer and J. P. de Magalhaes, *BMC Biol.*, 2021, **19**, 56.
9 M. Bar Dolev, I. Braslavsky and P. L. Davies, *Annu. Rev. Biochem.*, 2016, **85**, 515–542.
10 G. L. Fletcher, C. L. Hew and P. L. Davies, *Annu. Rev. Physiol.*, 2001, **63**, 359–390.
11 E. J. Smith and A. D. J. Haymet, *J. Mol. Liq.*, 2023, **390**, 359–390.
12 E. Kristiansen and K. E. Zachariassen, *Cryobiology*, 2005, **51**, 262–280.
13 Z. Y. Zhu, H. R. Ma, H. Z. Du, L. N. Zhang, J. H. Wu, C. Gao, W. Li, X. F. Chen, Y. Q. Su, D. Wang, X. T. Chen and Z. Y. He, *Angew. Chem., Int. Ed.*, 2025, **64**, e2025053259.
14 J. A. Raymond and A. L. Devries, *Proc. Natl. Acad. Sci. U. S. A.*, 1977, **74**, 2589–2593.
15 C. A. Knight, *Nature*, 2000, **406**, 249–251.
16 T. J. McPartlon, C. T. Osborne, K. Wang, R. E. Detwiler, K. Meister and J. R. Kramer, *Adv. Mater.*, 2025, e2050410.
17 Y. Gwak, J. I. Park, M. Kim, H. S. Kim, M. J. Kwon, S. J. Oh, Y. P. Kim and E. Jin, *Sci. Rep.*, 2015, **5**, 12019.
18 C. I. Biggs, T. L. Bailey, B. Graham, C. Stubbs, A. Fayter and M. I. Gibson, *Nat. Commun.*, 2017, **8**, 1546.
19 L. L. C. Olijve, K. Meister, A. L. DeVries, J. G. Duman, S. Q. Guo, H. J. Bakker and I. K. Voets, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 3740–3745.
20 R. P. Tas, M. M. R. M. Hendrix and I. K. Voets, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2212456120.
21 Q. L. Ye, K. Basu, R. Eves, R. L. Campbell, S. Phippen and P. L. Davies, *Acta Crystallogr., Sect. A: Found. Adv.*, 2018, **74**, A163–A163.
22 A. A. Antson, D. J. Smith, D. I. Roper, S. Lewis, L. S. D. Caves, C. S. Verma, S. L. Buckley, P. J. Lillford and R. E. Hubbard, *J. Mol. Biol.*, 2001, **305**, 875–889.
23 F. Sicheri and D. S. C. Yang, *Nature*, 1995, **375**, 427–431.
24 M. M. Harding, L. G. Ward and A. D. J. Haymet, *Eur. J. Biochem.*, 1999, **264**, 653–665.
25 W. Gronwald, M. C. Loewen, B. Lix, A. J. Daugulis, F. D. Sönnichsen, P. L. Davies and B. D. Sykes, *Biochemistry*, 1998, **37**, 4712–4721.
26 T. P. Ko, H. Robinson, Y. G. Gao, C. H. C. Cheng, A. L. DeVries and A. H. J. Wang, *Biophys. J.*, 2003, **84**, 1228–1237.
27 E. K. Leinala, P. L. Davies, D. Doucet, M. G. Tyshenko, V. K. Walker and Z. C. Jia, *J. Biol. Chem.*, 2002, **277**, 33349–33352.
28 K. V. Ewart, Q. Lin and C. L. Hew, *Cell. Mol. Life Sci.*, 1999, **55**, 271–283.
29 G. J. Deng, D. W. Andrews and R. A. Laursen, *FEBS Lett.*, 1997, **402**, 17–20.
30 C. A. Knight, C. C. Cheng and A. L. Devries, *Biophys. J.*, 1991, **59**, 409–418.
31 U. S. Midya and S. Bandyopadhyay, *J. Phys. Chem. B*, 2018, **122**, 3079–3087.
32 A. L. Devries and D. E. Wohlschlag, *Science*, 1969, **163**, 1073–1075.

33 J. C. Lopes, C. T. Kinasz, A. M. C. Luiz, M. G. Kreusch and R. T. D. Duarte, *J. Appl. Microbiol.*, 2024, **135**(6), Ixae140.

34 A. Khan, J. Uddin, F. Ali, A. Ahmad, O. Alghushairy, A. Banjar and A. Daud, *Sci. Rep.*, 2022, **12**, 20672.

35 S. Dhibar and B. Jana, *J. Phys. Chem. Lett.*, 2023, **14**, 10727–10735.

36 Y. C. Qiu and G. W. Wei, *Brief Bioinform.*, 2023, **24**(5), 1–13.

37 L. C. Vieira, M. L. Handojo and C. O. Wilke, *Sci. Rep.*, 2025, **15**, 21400.

38 N. Ferruz and B. Höcker, *Nat Mach Intell*, 2022, **4**, 521–532.

39 A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik and B. Rost, *IEEE Trans. Pattern Anal.*, 2022, **44**, 7112–7127.

40 N. Brandes, D. Ofer, Y. Peleg, N. Rappoport and M. Linial, *Bioinformatics*, 2022, **38**, 2102–2110.

41 A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. J. R. Olmos, C. M. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser and N. Naik, *Nat. Biotechnol.*, 2023, **41**, 1099–1115.

42 A. Rives, J. Meier, T. Sercu, S. Goyal, Z. M. Lin, J. S. Liu, D. M. Guo, M. Ott, C. L. Zitnick, J. Ma and R. Fergus, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2016239118.

43 Z. M. Lin, H. Akin, R. S. Rao, B. Hie, Z. K. Zhu, W. T. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. D. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.

44 M. Ertelt, J. Meiler and C. T. Schoeder, *ACS Synth. Biol.*, 2024, **13**, 1085–1092.

45 L. Xu, *Comm. Com. Inf. Sc.*, 2024, **2058**, 98–111.

46 L. Zhao, Q. He, H. J. Song, T. Q. Zhou, A. Luo, Z. G. Wen, T. Wang and X. Z. Lin, *Molecules*, 2024, **29**, 4965.

47 I. A. Imam, S. Bailey, D. L. Wang, S. Zeng, D. Xu and Q. Shao, *Langmuir*, 2024, **41**, 811–821.

48 N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan and S. Gelly, *Proc. Mach. Learn. Res.*, 2019, **97**, 1–10.

49 A. Chandra and X. Yao, *Neurocomputing*, 2006, **69**, 686–700.

50 H. W. Lv, K. Yan, Y. C. Guo, Q. Zou, A. Hesham and B. Liu, *Comput. Biol. Med.*, 2022, **146**, 105577.

51 Q. T. Yuan, K. Y. Chen, Y. M. Yu, N. Q. K. Le and M. C. H. Chua, *Briefings Bioinf.*, 2023, **24**(1), 1–10.

52 Y. N. Bin, W. Zhang, W. D. Tang, R. Y. Dai, M. L. Li, Q. Z. Zhu and J. F. Xia, *J. Proteome Res.*, 2020, **19**, 3732–3740.

53 A. Kuffel, D. Czapiewski and J. Zielkiewicz, *J. Chem. Phys.*, 2015, **143**(3), 1–8.

54 M. J. Kuiper, C. J. Morton, S. E. Abraham and A. Gray-Weale, *Elife*, 2015, **4**, e05142.

55 J. D. Madura, M. S. Taylor, A. Wierzbicki, J. P. Harrington, C. S. Sikes and F. Sonnichsen, *J. Mol. Struct.: THEOCHEM*, 1996, **388**, 65–77.

56 A. Wierzbicki, P. Dalal, T. E. Cheatham, J. E. Knickelbein, A. D. J. Haymet and J. D. Madura, *Biophys. J.*, 2007, **93**, 1442–1451.

57 J. Cui, K. Battle, A. Wierzbicki and J. D. Madura, *Int. J. Quantum Chem.*, 2009, **109**, 73–80.

58 R. K. Kar and A. Bhunia, *J. Phys. Chem. B*, 2015, **119**, 11485–11495.

59 K. Battle, E. A. Salter, R. W. Edmunds and A. Wierzbicki, *J. Cryst. Growth*, 2010, **312**, 1257–1261.

60 M. Schauperl, M. Podewitz, T. S. Ortner, F. Waibl, A. Thoeny, T. Loerting and K. R. Liedl, *Sci. Rep.*, 2017, **7**, 11901.

61 K. Mochizuki and V. Molinero, *J. Am. Chem. Soc.*, 2018, **140**, 4803–4811.

62 W. J. Zhang, H. Liu, H. H. Fu, X. G. Shao and W. S. Cai, *J. Phys. Chem. B*, 2022, **126**, 10637–10645.

63 A. D. Biswas, V. Barone and I. Daidone, *J. Phys. Chem. Lett.*, 2021, **12**, 8777–8783.

64 T. Jin, F. Q. Long, Q. Zhang and W. Zhuang, *Phys. Chem. Chem. Phys.*, 2022, **24**, 21165–21177.

65 Y. X. Zhou, H. W. Tan, Z. Y. Yang, Z. C. Jia, R. Z. Liu and G. J. Chen, *Sci. China, Ser. B: Chem.*, 2007, **50**, 266–271.

66 Y. C. Liao, W. T. Yang and Z. R. Sun, *Langmuir*, 2024, **41**, 663–670.

67 H. Nada and Y. Furukawa, *Polym. J.*, 2012, **44**, 690–698.

68 C. UniProt, *Nucleic Acids Res.*, 2025, **53**, D609–D617.

69 B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder and C. H. Wu, *Bioinformatics*, 2007, **23**, 1282–1288.

70 E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret and I. Xenarios, *Methods Mol. Biol.*, 2016, **1374**, 23–54.

71 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

72 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C. C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Zemgulyte, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Zídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **636**, 493–516.

73 R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic Acids Res.*, 2012, **40**, W537–W541.

74 M. Matsumoto, T. Yagasaki and H. Tanaka, *J. Chem. Phys.*, 2024, **160**, 094101.

75 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

76 J. L. F. Abascal, E. Sanz, R. G. Fernández and C. Vega, *J. Chem. Phys.*, 2005, **122**, 234511.

77 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics Modell.*, 1996, **14**, 33–38.

78 T. M. Mark James Abraham, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *Software X*, 2015, 19–25.

79 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

80 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

81 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

82 S. W. Tighe, D. L. Vellone, K. M. Tracy, D. B. Lynch, K. H. Finstad, M. C. McLlelan and J. A. Dragon, *J. Biomol. Tech.*, 2025, **36**, 1–21.

83 M. S. Venkatarajan and W. Braun, *J. Mol. Model.*, 2001, **7**, 445–453.

84 J. G. Duman and A. L. de Vries, *Comp. Biochem. Physiol., Part B: Biochem. Mol. Biol.*, 1976, **54**, 375–380.

85 W. Zhang and R. A. Laursen, *J. Biol. Chem.*, 1998, **273**, 34806–34812.

86 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.

87 A. Luzar and D. Chandler, *Nature*, 1996, **379**, 55–57.