Food Research International 221 (2025) 117322

Contents lists available at ScienceDirect

Food Research International

journal homepage: www.elsevier.com/locate/foodres

Unraveling the physicochemical differences among Osborne protein classes via bioinformatics and AI

Hyukjin Kwon [a], Yixiang Xu [b], Xuan Xu [c,d], Yonghui Li [a,*]

[a] Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA
[b] Healthy Processed Foods Research Unit, Western Regional Research Center, USDA-ARS, Albany, CA, USA
[c] Department of Statistics, Kansas State University, Manhattan, KS 66056, USA
[d] 1DATA Consortium, www.1DATA.life, Kansas State University Olathe, Olathe, KS 66061, USA

ARTICLE INFO

Keywords:
Seed storage protein
Osborne classification
Machine learning classification
Graph convolutional networks
Molecular dynamics
Protein solubility

ABSTRACT

Osborne fractionation remains a cornerstone in food science for categorizing seed storage proteins (SSPs), yet molecular distinctions among the classes remain unclear. This study employs a computational framework integrating structural modeling, AI (artificial intelligence)-driven classification, and molecular dynamics (MD) simulations to elucidate these underlying physicochemical differences. Using a dataset of 1039 SSPs from 215 species, sequence and structural-based features were extracted and compared to identify class-specific characteristics, such as low hydrophobic patch area of albumins. Machine learning (ML) classifiers, including binary support vector machines and graph convolutional networks were trained on these features, achieving validation and test accuracies ranging from 96.0 % to 100.0 %. Model interpretations using SHapley Additive exPlanations and saliency mapping revealed key distinguishing features between albumin/prolamin and globulin/glutelin, respectively. For the albumin and prolamin classes, physicochemical feature comparisons and ML classifiers identified factors underlying their solubility differences, such as the low abundance of charged residues in prolamins. On the other hand, although certain features, such as mean surface electric potential, distinguished globulins from glutelins, no clear association was found between these features and experimental solubility trend. Notably, saliency analysis of globulins and glutelins highlighted loop and helical regions outside the conserved β-barrel motifs, where compositional differences in glutamic acid, glycine, serine, and glutamine residues were observed. MD simulations explored solvent-specific conformational changes in representative SSPs, with all-atomic simulations performed on single monomers and coarse-grained simulations conducted with multiple monomers. For 2S soy albumin and 19 kDa maize prolamin, distinct hydrogen bonding patterns was observed during their adaptations to 70 % ethanol environment, and the expected aggregation tendency was reproduced in the multiple-monomer simulation. Taken together with the highlighted features in ML classification, these results suggest that the experimental solubility of albumins and prolamins can be explained at the monomeric level. However, for pea legumin A (globulin) and rice glutelin A1, no clear differences in structural and aggregation dynamics were observed, and monomeric properties alone failed to account for their distinct solubility. These findings suggest that glutelin insolubility is likely dictated by inter-protein disulfide networks rather than intrinsic monomeric characteristics, aligning with previous experimental observations.

1. Introduction

Seed storage proteins (SSPs) are essential for seed development and serve as primary dietary protein sources for humans. Broadly, seed proteins can be grouped into housekeeping and storage proteins (Zhou et al., 2016). Housekeeping proteins function to maintain cellular activity such as translation or metabolism, while SSPs work as a source of nutrition during the early growth of the seed. Between the two, storage proteins composites the major portion of the proteins in seeds (Radhika & Rao, 2015).

As plant proteins are gaining increasing interests in the food industry, a comprehensive understanding of the structural and molecular characteristics of SSPs is crucial not only for elucidating their biological roles but also for optimizing protein extraction, processing, and

* Corresponding author.
E-mail address: yonghui@ksu.edu (Y. Li).

https://doi.org/10.1016/j.foodres.2025.117322
Received 22 April 2025; Received in revised form 8 August 2025; Accepted 10 August 2025
Available online 13 August 2025

functionality in food applications (Bera et al., 2023). While the traditional solubility-based Osborne classification of SSPs remains widely used, the molecular distinctions among these groups are not yet fully understood. In fact, the classification of SSPs remains an area with unresolved complexities and grey areas, where solubility, structure, and evolutionary relationships do not always align.

One of the earliest and most widely used classification systems for SSPs was introduced by T. B. Osborne in 1924 (Osborne, 1924). Employing the classification system from the American committee on protein nomenclature, as well as his findings, Osborne suggested categorizing the major types of SSPs into four groups (Shewry & Casey, 1999) through serial extraction. These fractions included albumins, globulins, prolamins, and glutelins based on their solubility in water, salt solutions, alcohol, and acid/base, respectively. The four groups encompass all storage protein categories operationally defined by the Osborne system.

Up to date, this solubility-based classification remains the standard in food science and industry due to its practicality and ease of adaptation (Boulter & Derbyshire, 2014). However, subsequent biochemical studies have revealed its limitations. For instance, equilibrium centrifugation and size-exclusion chromatography demonstrated that globulins, initially classified as a single group based on salt solubility, comprise of two structurally distinct subfamilies of 7 ~ 8S vicilin and 11 ~ 12S legumin (Koshiyama, 1972). Similarly, further studies have led to multiple classification criteria for prolamins, such as molecular weight (e.g., 22 kDa prolamin) (Esen, 1986) or amino acid composition (e.g., cysteine-rich and poor prolamins) (Tatham & Shewry, 1995). These findings highlight the limitation of solubility-based classification, as it fails to capture the structural and functional diversity of SSPs.

With advances in sequencing technologies, researchers introduced an alternative classification framework based on evolutionary relationships, leading to the recognition of two major superfamilies of SSPs: prolamins and cupins. Under this sequence-based method, SSPs are grouped according to gene structure, sequence homology, and conserved structural motifs (Fukushima, 1991). However, this approach does not always align with the distinct functional properties of SSPs. For example, some glutelins, such as rice oryzenin, are grouped together with globulins within the cupin superfamily (Tan-Wilson & Wilson, 2012). Similarly, 2S albumins, traditionally classified as water-soluble proteins, are now considered members of the prolamin superfamily (Mills & Shewry, 2004). These discrepancies highlight a critical limitation: evolutionary similarity does not necessarily reflect physicochemical similarity (Tan-Wilson & Wilson, 2012), which is often more relevant to food processing and protein extraction.

Beyond such classification inconsistencies, practical extraction methods also diverge from Osborne-defined boundaries. In industrial settings, protein extraction is typically performed using alkaline solubilization, enzymatic hydrolysis, or physical disruption methods (e.g., high-pressure homogenization), which often co-extract overlapping fractions of albumins, globulins, and glutelins. For example, a recent study on lentil protein extraction showed that alkaline- and enzyme-extracted proteins shared physicochemical and functional properties with both albumin- and globulin-rich fractions, indicating a breakdown of Osborne boundaries in extraction workflows (Dias et al., 2024). Moreover, as food science is shifting toward precision design of plant proteins for specific structural and functional roles, the limitations of Osborne-based categorization become increasingly evident. For example, proteomic profiling of the pea globulin fraction has revealed over 200 distinct protein species by 2D-gel electrophoresis (Dziuba et al., 2014), underscoring the chemical heterogeneity within a single Osborne class. This highlights the inability of conventional solubility-based separation to resolve protein subclasses relevant to functionality. Therefore, while Osborne classification provides a useful conceptual scaffold, a more physicochemically informed and data-driven approach is essential for guiding the functional formulation of plant-based food systems.

To address the discrepancies in the Osborne classification system, some researchers have turned to machine learning (ML) methods for automated protein classification, primarily using sequence-derived physicochemical features. For instance, Marla et al. utilized multiple properties extracted from the sequences of 170 rice SSPs, including isoelectric point and molar extinction coefficients (Marla et al., 2010). Employing a neural network to classify them according to the respective Osborne classes, the group achieved 95.3 % accuracy for rice SSPs. Similarly, Radhika and Rao utilized features such as individual amino acid compositions and applied neural network and support vector machine (SVM) to classify storage proteins from five different species: rice, wheat, maize, thale cress, and castor bean, with accuracies ranging from 82.1 % to 98.6 % (Radhika & Rao, 2015). While these studies highlighted the potential for accurate classification of SSPs based on physicochemical properties, they suffered from three major limitations that must be addressed for broader applicability. First, they relied exclusively on sequence data, overlooking critical structural properties such as secondary structure motifs or surface hydrophobicity. Moreover, the models were species-specific, limiting generalizability across different protein sources. Lastly, the models lacked interpretability, providing little or no insight into the molecular determinants of each Osborne class.

Over the past five years, the fields of artificial intelligence and computational biology have undergone dramatic advancements, leading to the adoption of new algorithms and technologies (Kumar & Srivastava, 2024). A key aspect of these developments lies in the significant refinement of *ab-initio* protein modeling and explainable machine learning models. Specifically, the near-experimental accuracy of the AlphaFold series (Jumper et al., 2021) has enabled *in-silico* studies of proteins that could not be crystallized or remained structurally unresolved, including most SSPs. Moreover, advancements in interpretation algorithms, such as SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and saliency mapping, have begun to mitigate the previously perceived black-box nature of ML models, providing insight into the key features driving model predictions. In this context, earlier models that classified SSPs solely based on sequence data are insufficient. More modern *in-silico* techniques could reveal molecular-level differences among the four groups, which are critical for understanding SSP functionality in food applications.

This study explores the physicochemical characteristics of Osborne classes through AI and bioinformatics techniques such as molecular dynamics (MD) simulation. More specifically, it aims to integrate sequence, structural, and physicochemical information to refine our current understanding in SSP classification. The sequence information of 1039 SSPs from 215 species was collected, and their 3D structures were predicted using AlphaFold 3.0. The molecular features, including global (entire sequence or protein) and residual descriptors were extracted from their sequences and structures. To explore different feature spaces, SVM with SHAP was employed to identify global sequence patterns, while graph convolutional networks (GCN) with saliency mapping were utilized to highlight local structural patterns that differentiate each group. Furthermore, to assess whether Osborne classes exhibit distinct, solvent-specific physicochemical dynamics, all-atomic (AA) and coarse-grained (CG) MD simulations of representative proteins, such as 2S soy albumin and 11S pea globulin, were conducted in different solvent environments. By providing an integrative molecular perspective through interpretable AI and comprehensive physicochemical analysis, this research aims not only to refine our understanding of SSP classification but also to enhance its applicability in food processing and functional optimization.

2. Methods

2.1. Dataset

2.1.1. Data acquisition

The protein sequences of SSPs were retrieved from UniProt (The UniProt Consortium, 2023) using the keyword KW-0708 (nutrient reservoir) along with their Osborne classes (albumin, globulin, prolamin, and glutelin). Additionally, their trivial names were also considered, including napin, conglutin (albumin); glycinin, conglycinin, legumin, vicilin, convincilin, cruciferin, phaseolin, edestin (globulin); zein, kaffirin, gliadin, hordein, secalin, coixin, avenin (prolamin); oryzenin, glutenin, hordenin, secalinin, aveninin, and zeinin (glutelin) (Day, 2013; Fukushima, 1991). Accordingly, the final query string was: "(KW-0708) AND (albumin OR napin OR conglutin) AND NOT (globulin OR glycinin OR [trivial names]) ..."

The acquired dataset was then manually curated using several filtering criteria to ensure data quality. Proteins labeled as "hypothetical," "putative," "-like," or "uncharacterized" or sequences without signal peptide annotation were excluded. These filtering criteria were based on Uniprot's standardized protein annotation guideline: (https://www.uniprot.org/help/annotation_guidelines), which follows the International Protein Nomenclature Guideline. According to these standards, descriptors such as "putative," "hypothetical," "-like," and "uncharacterized" denote functional ambiguity and are typically used only when no definitive functional assignment is possible. In this context, the proteins carrying these terms were excluded to ensure annotation reliability. In addition, proteins labeled "recombinant" were excluded to focus on native forms of SSPs. Proteins lacking annotated signal peptides, residue information, or named "precursor" were also removed, as the study targeted mature protein structures following signal sequence cleavage. No subjective filtering was done outside these criteria.

After applying these filtering steps, the final dataset comprised 1039 sequences from 215 different source organisms, including both manually reviewed and automatically annotated sequences to preserve dataset diversity. Using this filtered dataset, AlphaFold 3.0 (Abramson et al., 2024) was employed for *ab-initio* modeling of 3D protein structures, where signal peptides were removed prior to modeling. For each protein, five models were generated, and the structure with the highest overall predicted local distance difference test (PLDDT) score, a confidence metric indicating prediction reliability (Jumper et al., 2021), was selected to ensure the most reliable structural prediction. To provide confidence level in the predicted protein structures, all structures were colored based on their PLDDT score and deposited in https://github.com/john94kwon/Osborne-class-classification.

2.1.2. Data structure

While the above method enabled the collection of sequences and structures of SSPs, additional post-processing was necessary to resolve classification ambiguities between globulin and glutelin classes. Specifically, certain protein entries were annotated generically as "cupin type-1 domain-containing protein," making it unclear whether they belonged to globulin or glutelin, as both classes contain the cupin domain. For instance, the UniProt entry A0A0E0M7E8 (from rice, 56.32 kDa) was labeled as such, leading to ambiguity in classification. Given that rice glutelin constitutes 70–80 % of the total rice protein (Katsube-Tanaka et al., 2004) and that the protein bands of approximately 50 kDa are absent from SDS-PAGE gels of rice globulin fractions (Kim & Jeong, 2002), it is likely that the protein belongs to glutelin class. To address this inconsistency, based on established domain knowledge that glutelins are mainly from grass family such as wheat, barley, rice, or rye (Shewry et al., 1995), the cupin domain-containing proteins were annotated as glutelin if their source was from grass family. The family of the source organism was identified using USDA database of plant families (https://plants.usda.gov/). The summary of the source variation, number of species, and the total counts of data per each Osborne class is listed as Table 1.

2.2. Statistical testing

Statistical analyses were performed using a one-way analysis of variance (ANOVA) model implemented in Python 3.10 with the scikit-learn package (version 1.3.0) (Buitinck et al., 2013). ANOVA residuals were checked for normality using Shapiro-Wilk test. Bonferroni multiplicity adjustment was applied when comparing the 38 extracted physicochemical features among the four groups, with adjusted p-values <0.05 considered statistically significant. If the normality of residuals assumption for ANOVA was not satisfied, the non-parametric Kruskal-Wallis test was used instead. Probability values of $p < 0.05$ (2-tailed) were considered statistically significant for all comparisons. For features that passed the residual normality test, results are presented as mean $\pm$ standard deviation. For non-normally distributed data, results are presented as median with interquartile range (IQR).

2.3. Multivariate analysis

Multivariate analysis was conducted following MinMax scaling to normalize feature distributions across variables. Principal component analysis (PCA) was performed to reduce dimensionality while retaining the majority of variance in the data. In addition, linear discriminant analysis (LDA) was employed to maximize class separability, leveraging class labels to extract the most discriminative components. To further explore feature correlations and their contributions to classification, partial least squares discriminant analysis (PLS-DA) was applied, capturing both variance within predictors and their relationship to the response variable. All analysis and relevant visualization were conducted with Python 3.10 with sklearn library.

2.4. Machine learning classification of proteins

Machine learning models were utilized to categorize seed storage proteins based on their Osborne classification. For global feature-based classification, SVM was selected as it achieved the highest interpretation stability and performance among commonly used tabular classifiers. This point is further discussed in the later section (Section 3.2.2). For residue-level classification, GCN was implemented, as it effectively captures local structural patterns within protein structure. GCN model was developed using the PyTorch library (ver 2.6.0) (Paszke et al., 2019). To ensure an even class distribution, the dataset was split into training and testing sets using stratified train_test_split() from scikit-learn (80:20 ratio). MinMax Scaling was applied for global features (SVM), and standard scaling was used for residue-level features (GCN). To enhance model generalizability and prevent overfitting, 5-fold stratified cross-validation was performed during hyperparameter

Table 1
Summary of collected seed storage protein data (215 species and 1039 proteins) organized by Osborne classification.

Class	Source (number of data)	Species	Counts
Albumin	Field mustard (24), Ethiopian mustard (12), Rapeseed (12), False flax (10), Mouse-ear cress (9), Wild cabbage (8), Radish (8)…	77	250
Globulin	Garden pea (42), Rapeseed (24), Field mustard (17), Soybean (14), False flax (12), Wild soy (10), Peanut (10), Lupine (10)…	112	336
Prolamin	Maize (106), Wheat (46), Sorghum (30), Hall's panicgrass (24), Job's tear (23), Foxtail millet (7), Goatgrass (7), Sugarcane (6)…	27	294
Glutelin	Japonica rice (35), Red rice (11), Malo sina (10), African rice (10), Indica rice (10), Oryza glumaepatula (6), *Oryza barthii* (6)…	32	159

optimization.

2.4.1. Global (sequence + structural) feature extraction

Feature extraction was performed on both sequence-based and structure-based properties. Biopython (ver 1.81) (Cock et al., 2009) was used to extract sequence-based features, including molecular weight, aromaticity, instability index, hydrophobic index, aliphatic index, absolute charge per residue, and hydrophilic index. For structure-based features, Quilt (ver 1.3) (Lijnzaad et al., 1996) and ChimeraX (ver 1.6.1) (Meng et al., 2023) were employed. Quilt was used to compute hydrophobic patch area, hydrophilic surface area, and the hydrophobic surface ratio, while ChimeraX provided additional structural descriptors, such as mean lipophilicity potential, mean Coulombic potential, total surface area, molecular volume, solvent-accessible surface area (SASA), number of hydrogen bonds, helix/coil/strand propensities, and the number of favorable contacts. To ensure structural consistency, all PDB structures were converted to PQR format using the pdb2pqr plugin (ver 3.6.1) (Dolinsky et al., 2004) under the CHARMM force field at pH 7.0. The number of disulfide linkages was determined by identifying cysteine pairs within a 2.0 Å threshold. It should be clarified that disulfide linkage refers to covalent bonds between cysteine residues within a single polypeptide (intra-protein), while disulfide network denotes higher-order crosslinking involving multiple polypeptides (inter-protein). The methodology and codes to extract these features, including sequence descriptors and structure-based parameters, were adapted from our previous work (Kwon et al., 2024).

2.4.2. Residue-level feature extraction

To explore fine-grained structural differences, GCN models were trained using residue-based node features: BLOSUM62 matrix (Henikoff & Henikoff, 1992), AAPHY7 (seven physicochemical properties of amino acids) (Meiler et al., 2001), and alpha carbon XYZ coordinates, combined with protein contact maps(Fig. 1). Prior to coordinate extraction, all protein structures were centered at the origin using VMD (version 1.9.4) (Humphrey et al., 1996). Graph-based neural networks utilize nodes (vertices), which represent individual elements in a graph, and edge, which defines the relationship or connection between each node. For the scope of this study, nodes represented each amino acid residue of the protein, and the three node features defined the characteristics of the nodes. Protein contact maps, constructed using a 12.0 Å threshold for alpha carbon distances, formed the edge matrix.

2.4.3. GCN architecture

All the contact maps and residue embeddings of a protein with L length were padded with P zeros to match the longest sequence in the dataset (L + P). During backpropagation, these padded nodes were ignored, ensuring they had no effect on the gradients (Gama et al., 2018). Additionally, the corresponding edge matrix entries were also set to zero, ensuring that artificially added nodes did not introduce spurious connections. This padding operation ensured that graphs of different proteins had a uniform shape, reducing the influence of variable edge structures, and making node-based features more comparable across proteins. The output dimension of GCN model was 32 resulting in a node-wise vector of size $[(L + P) \times 32]$ after feature aggregation through the GCN layers. With the node-wise vector, self-attention pooling mechanism developed by Lin et al. (Lin et al., 2017) was employed with 4 attention heads, further aggregating the vector by scoring the importance of individual nodes (residues) compared to other nodes, thereby producing a $[1 \times 4]$ vector (Fig. 1). This aggregated vector then passed through the output layer, where the final classification was performed using BCEWithLogitsLoss function, which inherently incorporates a sigmoid activation. More detailed mathematical formulation of the GCN architecture can be found in our previous work (Kwon et al., 2024). All sequence, structure, extracted features, and relevant scripts were deposited in the same Github link.

2.5. Molecular dynamics simulation

Four manually reviewed proteins, selected for their structural and functional relevance in food and industrial applications, were chosen to represent each Osborne class: 2S albumin from soy (P19594), pea 11S legumin A (P02857), 19 kDa maize zein (P06678), and rice (*japonica*) glutelin A1 (P07728). Proteolytic cleavage sites, such as residue 306 of glutelin A1 that naturally split the protein into two chains, were reflected by removing the annotated covalent bond, while leaving the N- and C-terminal residues in their default charged states. All disulfide linkages were assigned based on post-translational modification data from UniProt.

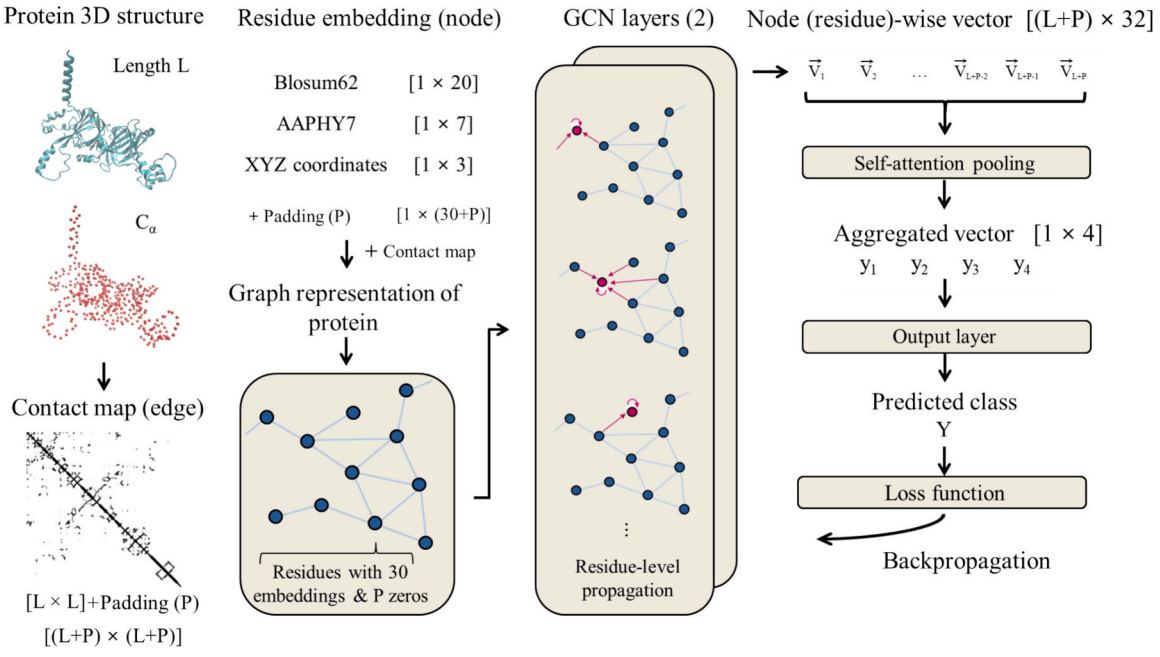

Fig. 1. Graphical representation of protein structure embedding and architecture of the GCN model.

The protonation states of the residues were predicted based on their isoelectric points. This choice was motivated by the unconventional environments in our simulations, such as 70 % ethanol and 1 M NaCl. Structure-based pKa tools such as PROPKA (Olsson et al., 2011) and H++ (Anandakrishnan et al., 2012) are not designed to handle such non-aqueous or high-salt conditions, and their assumptions (e.g., dielectric constant, implicit solvent, static conformation) do not hold for modeling such systems. Specifically, PROPKA uses experimental pKa and H++ employs electrostatic potential model (Poisson-Boltzmann), making both models highly specific to aqueous condition. Therefore, we selected a standard pKa-based protonation scheme as the most universal and interpretable approach for protonation state assignments. pdb2gmx module in GROMACS was used to assign protonation states.

Molecular dynamics simulations were performed using GROMACS 2021.5 (Abraham et al., 2015). All simulated systems underwent charge neutralization, steepest descent energy minimization in vacuum and solvent, followed by extensive NVT and NPT equilibration before production runs. Simulation setup was described with technical details below, in line with established modeling practices (e.g. Wang et al., 2025).

2.5.1. All atomic MD simulation

The prepared protein structures were placed in a cubic simulation box with a minimum distance of 1.2 nm between the protein and the box boundaries. Initial energy minimization in vacuum was performed under the CHARMM36 force field (Huang & MacKerell Jr., 2013) using the steepest descent algorithm for 5000 steps, employing a Verlet cutoff scheme with 1.2 nm cutoffs for both Coulombic and van der Waals interactions. These cutoff values were used consistently in all subsequent simulations. Afterward, the system was solvated with TIP3P water using the gmx_solvate module, and counterions were added to neutralize the net charge using the gmx_genion module. A second round of energy minimization was then conducted on the solvated system using the same parameters to remove unfavorable contacts between protein, solvent, and ions. For solvated simulations, LINCS algorithm (Hess et al., 1997) was applied to constrain hydrogen bonds, Particle mesh Ewald algorithm (Darden et al., 1993) was used for electrostatics, and force-switching was applied to van der Waals interactions.

NVT equilibration was then carried out at 300 K using the Nose-Hoover thermostat (Evans & Holian, 1985), with separate coupling groups assigned to protein and solvent (temperature coupling constant = 1.0 ps) with a time step of 1.0 fs over 125.0 ps. This was followed by 125.0 ps NPT equilibration using the same thermostat and Parrinello-Rahman barostat (Parrinello & Rahman, 1981) at 1.0 bar, with a compressibility of 4.5×10^{-5} bar^{-1} and a 5.0 ps pressure coupling constant. The protein structure was position-restrained with restraint strength of 1000 kJ/mol/nm^2 for both equilibration steps. The production run was conducted for 200.0–400.0 ns using a 2.0 fs time step, depending on the root mean square deviation (RMSD) convergence, maintaining the same thermostat and barostat settings as the NPT equilibration.

2.5.2. Coarse-grained MD simulation

A set of CG-MD simulations was conducted using the Martini 3.0 force field (Souza et al., 2021) to explore large-scale conformational dynamics and protein-protein interactions over an extended timescale. The CG representation of the proteins was generated using the martinize2 Python library (Kroon et al., 2022). Backbone-based elastic network models were constructed with a force constant of 700.0 kJ/mol/nm^2, and secondary structure was assigned using DSSP (Kabsch & Sander, 1983). Five monomers of prepared protein were randomly inserted into the simulation box using gmx_insert function, ensuring an approximate 1 % protein concentration while maintaining a minimum inter-protein distance of 2.0 nm to prevent initial steric clashes. The system was initially minimized in vacuum with a steepest descent integrator for 10,000 steps. Electrostatic interactions were treated using a reaction-field scheme (ε_r = 15, r_cutoff = 1.1 nm), and van der Waals interactions were shifted to zero at 1.1 nm using the Verlet cutoff scheme. These cutoff distances were retained throughout subsequent simulations.

The minimized system was then solvated with standard Martini water beads (W) and neutralized with gmx_insert function. The box size was determined based on the molar ratio between water and protein, using the Martini model's 4:1 mapping of water molecules per W bead. In ethanol simulations, the standard Martini ethanol bead (SP1) was found to be overly hydrophilic (octanol-water transfer free energy = −5.23 kcal/mol) compared to the experimental value (−1.76 kcal/mol) (Souza et al., 2021). To better approximate ethanol's hydrophobicity, the 1-propanol bead (N6) was used instead, as it shares a similar molecular structure and exhibits transfer free energy of −1.95 kcal/mol (Souza et al., 2021), closely matching the experimental ethanol value.

After solvation, a second round of energy minimization was performed using the same settings as in vacuum. The LINCS algorithm was applied to constrain bond lengths in the system. NVT equilibration was conducted using the velocity-rescaling thermostat at 300 K with a 1.0 ps coupling constant for 50.0 ns (Δt = 20.0 fs). This was followed by NPT equilibration using the Parrinello–Rahman barostat (τ_p = 5.0 ps, compressibility = 4.5×10^{-5} bar^{-1}) for another 50.0 ns. During equilibration, position restraints were applied to the protein heavy atoms with a force constant of 4000 kJ/mol/nm^2. The production run was conducted without restraints and continued until notable aggregation was observed in at least one solvent system (1.0–4.0 μs), using the same simulation parameters as the NPT equilibration phase without position restraints.

3. Results and discussion

3.1. Protein characterization

3.1.1. Comparative analysis of 38 physicochemical properties

To highlight differences among the four Osborne classes, 38 physicochemical properties were compared (Table S1). Among the features, only the turn-forming residue fraction met the residual normality criterion. Therefore, a non-parametric Kruskal-Wallis test was employed for the remaining features. The non-parametric analysis identified key distinguishing features among the classes, such as the highest ratio of methionine (rM) in albumins (2.400, 1.344) (median, IQR) or leucine (rL) in prolamins (16.000, 10.036). However, in some cases, small numerical differences reached statistical significance. For example, albumins and glutelins exhibited no meaningful difference in relative hydrophobic area (0.543, 0.049 and 0.549, 0.015, respectively). This might be attributed to the increased sensitivity of statistical tests in large datasets (Sullivan & Feinn, 2012). While statistically significant, these differences were likely not biologically meaningful, as the absolute differences were minimal. Therefore, only notably distinctive features that uniquely defined each class (i.e., those with no IQR overlap across the four classes) were provided in Table 2a. For instance, cysteine content (rC) in albumins was classified as a meaningful feature, as its distribution was entirely separated from that of other classes (Table 2a). In contrast, leucine content, although statistically different, had overlapping IQRs and was therefore not considered biologically distinctive (Table S1).

Table 2a lists the eight most distinctive features observed among the extracted properties. Albumins were characterized by their high rC (5.594 %), disulfide linkage counts (4.000), along with low hydrophobic patch area (5.689 nm^2). These values aligned with previous reports on the high amount of cysteine residues in 2S albumins and their canonical disulfide bonding patterns (Clement et al., 2005; Shewry et al., 2002). Prolamins, in contrast, were distinguished by their notably low hydrophilic index (−0.635 a.u.) and reduced proportions of rR (Arg) (1.361 %), rD (Asp) (0.357 %), rE (Glu) (0.670 %), rG (Gly) (1.653 %), and rK (Lys) (0.000 %). The relatively small number of the charged residues

Table 2
Key physicochemical properties distinguishing Osborne protein classes (a), including features with notable differences between globulin and glutelin (b). Values are expressed as mean ± standard deviation and median with interquartile range. Medians with different superscript letters indicate statistically significant differences ($p < 0.05$).

(a)

Distinctive features (no interquartile range (IQR) overlap across the four classes)								
Class	rC (%)	Disulfide linkage count	Hydrophobic patch area (nm^2)	Hydrophilic index (a.u.)	rR (%)	rD (%)	rE (%)	rK (%)
Albumin	**5.838 ±1.236/ (5.594, 1.627a)**	**4.196±1.510/ (4.000, 0.000a)**	**5.918±1.599/ (5.689, 1.654b)**	0.232±0.173/ (0.203, 0.253a)	7.719±3.383/ (6.711, 4.464a)	4.689±1.531/ (4.576, 1.341a)	7.913±2.836/ (7.000, 3.735a)	4.192±2.251/ (4.505, 3.388a)
Globulin	0.945±0.608/ (1.040, 0.558c)	1.642±1.280/ (2.000, 1.000b)	16.941±3.478/ (16.043, 3.658a)	0.224±0.223/ (0.227, 0.350a)	8.031±2.529/ (7.551, 3.983a)	4.327±0.866/ (4.329, 1.091a)	8.987±3.290/ (8.924, 4.356a)	3.864±2.030/ (3.178, 2.713a)
Prolamin	1.408±1.213/ (0.939, 2.084c)	1.044±1.575/ (0.000, 2.000c)	15.591±3.576/ (16.684, 4.878a)	**−0.556±0.162/ (−0.635, 0.318c)**	**1.460±0.836/ (1.361, 0.808b)**	**0.398±0.533/ (0.357, 0.649c)**	**0.735±0.677/ (0.670, 0.410c)**	**0.246±0.473/ (0.000, 0.404b)**
Glutelin	1.297±0.337/ (1.261, 0.602b)	1.956±0.235/ (2.000, 0.000b)	16.760±2.073/ (16.453, 0.897a)	−0.004±0.077/ (−0.015, 0.048b)	7.504±1.293/ (7.361, 1.089a)	3.243±0.825/ (3.158, 1.151b)	5.945±1.051/ (5.732, 0.684b)	3.078±0.598/ (3.104, 0.728a)

(b)

Features with notable difference between globulin and glutelin				
Class	rE (%)	rY (%)	Aromaticity (a.u.)	Predicted isoelectric point
Albumin	7.913 ± 2.836/ (7.000, 3.735a)	2.523 ± 1.193/ (2.210, 1.235c)	6.447 ± 2.241/ (6.708, 2.319c)	6.480 ± 1.100/ (6.391, 1.116c)
Globulin	8.987 ± 3.290/ (8.924, 4.356a)	2.552 ± 0.656/ (2.434, 0.797c)	7.433 ± 0.979/ (7.425, 1.126b)	6.346 ± 1.020/ (6.067, 1.402c)
Prolamin	0.735 ± 0.677/ (0.670, 0.410c)	2.844 ± 1.365/ (3.077, 1.632b)	8.026 ± 2.030/ (7.377, 2.369b)	7.933 ± 0.846/ (8.089, 0.825b)
Glutelin	5.945 ± 1.051/ (5.732, 0.684b)	3.461 ± 0.556/ (3.619, 0.662a)	9.098 ± 0.916/ (8.956, 1.263a)	8.674 ± 1.040/ (9.090, 0.546a)

(Grumezescu & Holban, 2018), as well as high contents of hydrophobic residues (Xing et al., 2023) (Table S1), such as alanine or valine, were consistent with the previous reports.

On the other hand, globulins and glutelins lacked uniquely distinguishing features (i.e., no IQR overlap) that set them apart from the other classes. Instead, they exhibited a similar range of values across multiple properties, including molecular weight (53.382 vs 53.668 kDa) (globulin vs glutelin), solvent-accessible surface area (29.917 vs. 29.281 nm^2), and rI (Ile) (4.762 vs 4.873 %) (Table S1). However, some features displayed notable differences between the two SSP groups. For instance, globulins exhibited higher rE (8.924 % vs 5.732 %) but lower aromaticity (7.425 vs 8.956), rY (Tyr) (2.434 vs 3.619 %), and predicted

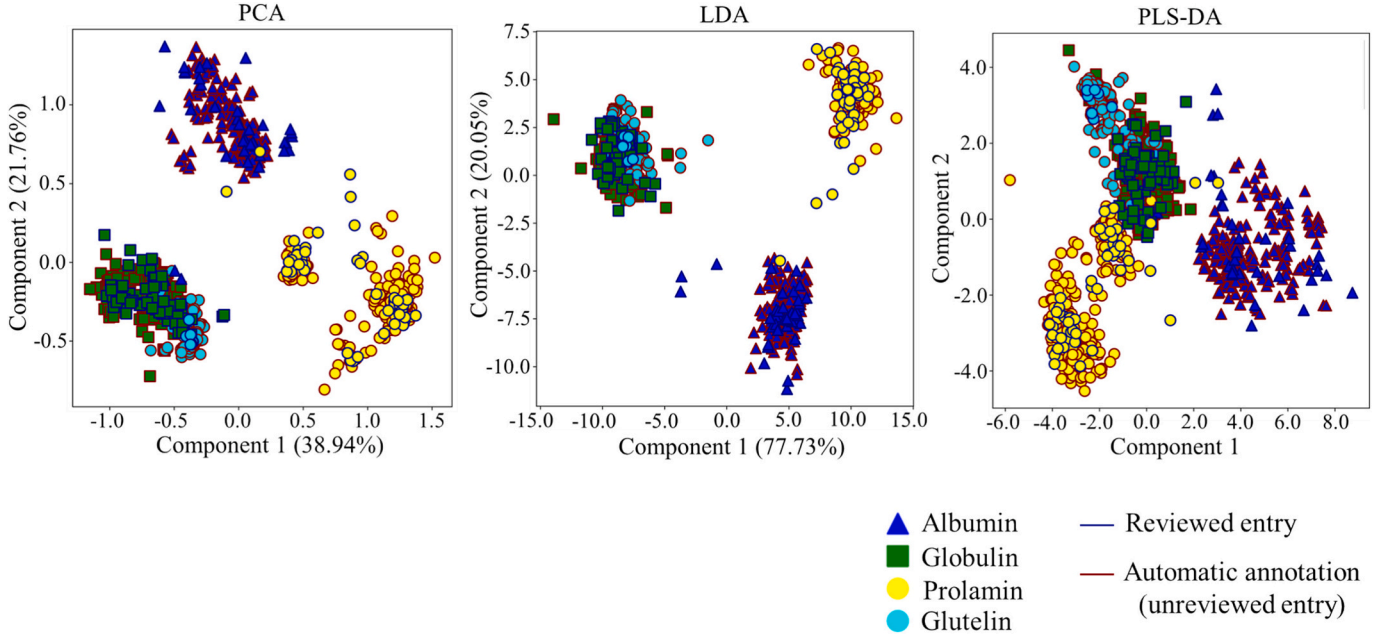

Fig. 2. Projection of seed storage proteins by Osborne classification using principal component analysis (PCA), linear discriminant analysis (LDA), and partial least squares discriminant analysis (PLS-DA).

isoelectric point (6.067 vs 9.090) relative to glutelin (Table 2b).

3.1.2. Multivariate analysis of physicochemical features

Although certain individual features exhibited class-specificity, a single feature alone might be insufficient to robustly distinguish the four classes, especially between globulin and glutelin. To address this, multiple projection methods: principal component analysis (PCA), linear discriminant analysis (LDA), and partial least squares discriminant analysis (PLS-DA), were applied to extract and visualize the most informative feature combinations (Fig. 2). As presented in the figure, all three methods effectively distinguished albumins and prolamins, but globulins and glutelins displayed substantial overlaps across all the projection methods. This similarity between globulins and glutelins was expected, given their well-known sequence homology. For instance, rice glutelin shares 34 % sequence identity with chickpea legumin by sequence alignment (Chang & Alli, 2012), and 70 % identity with 12S oat globulin (Shotwell et al., 1990), surpassing the threshold for homologous proteins (25 %) (Rost, 1999). The reported cross-reactivity between 12S oat globulin antibodies and rice glutelin, further supports their molecular resemblance (Robert et al., 1985). Additionally, it is believed that rice glutelin originated from the identical ancestral gene to 11S globulins (Furuta et al., 1986; Krishnan & Okita, 1986). The similarity of the extracted feature (Table S1) also aligned with this idea. Across all three projection methods, manually reviewed and automatically annotated (unreviewed) entries were similarly positioned within each class across all projection methods, demonstrating that both data contributed consistently to the classification patterns.

Among the three techniques, LDA provided the clearest boundaries between albumin, prolamin, and the globulin/glutelin group. To further interpret these boundaries, the two LDA axes were decomposed, with the explained variance ratio of 77.26 % and 20.05 % for axis 1 and 2, respectively. LDA axis decomposition revealed that SASA, molecular weight, and hydrophobic patch area were the primary determinants of class separation along both axes (Fig. S1). The prominence of these contributors was consistent with the feature distributions in Table S1, as these features exhibited strong differences among the three groups. The clear boundaries of LDA suggested that simple and linear techniques could be sufficient for broad classification across albumin, prolamin, and globulin/glutelin.

3.2. Machine learning classification and interpretation of the model

3.2.1. Classification using global features: SHAP analysis & interpretability

As demonstrated in Fig. 2, albumins, prolamins, and the globulin/glutelin group could be effectively distinguished using simple linear classification methods, without requiring complex non-linear models for multiclassification. Instead, binary classification could provide a more in-depth interpretation by focusing on the differences between two groups without diluting information. Therefore, two binary SVM classifiers were trained to differentiate between albumin/prolamin and globulin/glutelin. The models achieved high accuracy: 0.995/0.995/1.000 (training/cross-validation/testing) for albumin/prolamin and 0.987/0.972/0.960 for globulin/glutelin. Additional evaluation metrics, such as precision and recall, are provided in Table S2.

To justify the selection of SVM over alternative classifiers, multiple models were tested on the binary classification tasks using consistent validation protocols. Since SVM achieved near-perfect accuracy in distinguishing albumins from prolamins, it was evaluated on the more challenging globulin/glutelin dataset. To this end, five commonly used tabular classifiers—SVM, decision tree (DT), random forest (RF), k-nearest neighbors (KNN), and multilayer perceptron (MLP)—were compared using a consistent random seed and 5-fold stratified cross-validation for hyperparameter tuning. Each model was then retrained with five different random seeds to assess generalization stability. Performance metrics and the consistency of the top five SHAP features were evaluated using Jaccard similarity (Jaccard, 1901). As shown in Tables S3 and S4, SVM outperformed the other models in both predictive performance and interpretive stability and was therefore selected for further analysis.

To better understand how the models differentiated between groups, SHAP analysis was applied to quantify the contributions of the features in classification (Fig. 3). Briefly, SHAP assigns each feature a numerical score that reflects its influence on the model's prediction. By summing these scores across proteins, it can identify which physicochemical traits most strongly drive the model's decision. In a SHAP scatter plot, feature values are color-coded from blue (low values) to red (high values). SHAP values closer to zero indicate minimal influence, whereas more negative or positive values reflect stronger associations with a specific class. Negative SHAP values correspond to class 0 (albumin and glutelin), while positive values indicate class 1 (prolamin and globulin). SHAP values revealed that the most significant contributors for distinguishing albumin from prolamin were hydrophilic index, rK, rD, rE, and hydrophobic patch area. Higher hydrophilicity, rK, rD, and rE increased the likelihood of albumin classification, whereas a larger hydrophobic area was more indicative of prolamins. For globulin/glutelin, key contributors included mean Coulombic potential, predicted isoelectric point, hydrophobic index, rW (Trp), and rS (Ser). Among these features, higher mean Coulombic potential, predicted PI, hydrophobic index and rS favored glutelin classification. Conversely, higher rW was associated with globulin predictions.

While the top contributors to binary classification were identified, it would be important to connect these features with the expected solubility differences among the groups. Based on previous studies, two of the key factors affecting protein solubility in aqueous solutions are surface charge and the balance between hydrophilicity and hydrophobicity (Trevino et al., 2008; Van Oss, 1997). Briefly, proteins with higher net surface charge tend to be more soluble in water, as charged residues form strong charge-dipole interaction with water molecules, stabilizing protein in solution. Conversely, hydrophobic residues assume low affinity for water, and the hydrophobic patches on the surface could drive self-association to minimize exposure to water, leading to aggregation and even precipitation. In this context, the distinguishing features for albumin and prolamin aligned with their known solubility behavior in water and alcohol solutions. A lower abundance of charged residues (K, D, E) of prolamins would reduce its hydrophilicity and decrease charge-charge repulsion between them, promoting aggregation. Additionally, the significantly more hydrophobic nature of prolamins (lower relative hydrophilic index and higher hydrophobic patch area ratio) would promote hydrophobic interactions, which would also lead to self-aggregation in polar aqueous environments.

Although these charge and hydrophobicity-based trends aligned well with albumin and prolamin solubility, the case for globulin and glutelin was less straightforward. Although mean Coulombic potential was identified as a distinguishing feature between globulin and glutelin, it did not fully explain their solubility differences. The median Coulombic potential values were − 0.740 for globulin and 1.350 for glutelin, suggesting that glutelin carried more positive surface charges. However, the total charge magnitude, considering both positive and negative charges, was comparable between the two. If charge-charge screening by salt had been the primary driver of solubility, it would stabilize both positive and negative charges, making it difficult to argue that this alone explained why globulin was salt-soluble while glutelin was acid/alkali-soluble. Similarly, the second top contributor isoelectric point (pI), did not fully clarify the solubility difference. The median predicted isoelectric points of globulin and glutelin were 6.067 and 9.090, respectively. If charge repulsion (due to deprotonation of positively charged residues) were the primary driver of glutelin's alkaline solubility, a more negative charge would be expected under mildly alkaline conditions. However, its significantly higher value suggested that it would remain relatively neutral or even positively charged in such conditions, making simple charge-based repulsion an unlikely explanation. Furthermore, rS and rW, while identified as key features, would not have direct contributions

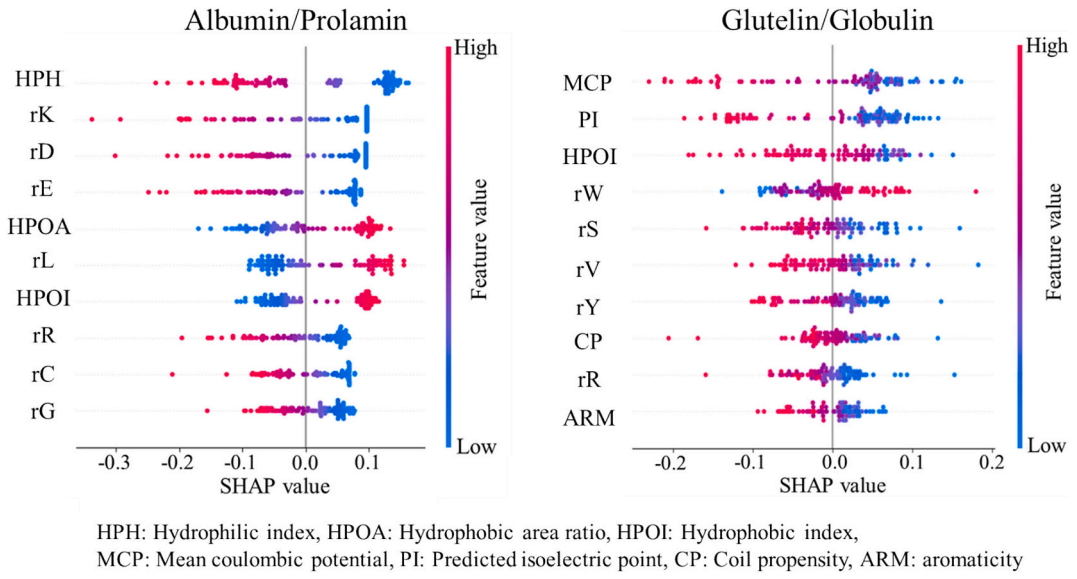

Fig. 3. SHapley Additive exPlanations (SHAP) scatter plots showing feature importance in seed storage protein classification. Feature values are color-coded from low (blue) to high (red), and SHAP values indicate each feature's impact on the model output for individual predictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to salt or alkaline solubility mechanisms, as they are not protonatable. Taken together, these findings suggested that the information obtained here was insufficient to fully explain the solubility differences between globulin and glutelin. Other factors, such as structural dynamics, hydration effects, or intermolecular interactions, might play a more significant role, which will be discussed in the following sessions.

3.2.2. Residue-level classification via GCN: Saliency mapping & interpretability

While global feature-based classification identified key distinguishing features, relying solely on global properties might limit the ability to investigate local structural variations, hindering deeper insights. Therefore, to explore more fine-grained structural differences, GCN models were trained, achieving accuracies of 1.000/1.000/1.000 (training/cross-validation/testing) for albumin/prolamin and 0.974/0.987/0.968 for globulin/glutelin. GCN model also yielded high average accuracy of 0.971 ± 0.009 across different random seeds (Table S3 and S4).

To identify the most important residues for classification, saliency scores were computed on the testing set. These scores quantify the influence of each input feature by analyzing gradient sensitivity, namely the partial derivative of the model's output with respect to each input feature (Simonyan et al., 2013). In this context, residues with high saliency values are those where small perturbations significantly affect the model's output, indicating strong influence on the prediction. Unlike SHAP, which captures global feature importance, saliency mapping provides fine-grained, spatially resolved insights, allowing visualization of critical regions on the protein surface that distinguish between classes. The magnitudes of the scores across all features were summed per node, identifying residues with the highest influence on classification. A representative saliency map with residue annotations is provided as Fig. 4a.

In order to examine the positional trends among these residues, their surface exposure was analyzed using the FreeSASA Python package (Mitternacht, 2016), with a threshold of 0.25 to define exposed residues. From analyzing the peaks in the saliency scores (Fig. 4a), it was found that surface-exposed residues L, Q (Gln), C, A (Ala), V (Val), and P (Pro) were most important for distinguishing albumin from prolamin. In the case of globulin/glutelin, the most significant residues were surface-exposed E, Q, S, G, and R (Fig. 4b).

3.2.3. Globulin vs. glutelin: Distinct structural & residue-level differences

Compared to albumin and prolamin, globulin and glutelin exhibit more well-defined protein structures. Specifically, they share two cupin domains, which contain β-barrel motifs, enabling a more in-depth comparison of local features between the two SSP classes. Analysis of the regions with saliency scores >0.5 revealed a consistent positional bias outside the β-barrel structural motifs in both testing and entire datasets. These highlighted regions suggested that globulin and glutelin differentiation was primarily driven by variations in loop and helical regions rather than core β-strand structures, which remained relatively stable across different environmental and evolutionary pressures (Abrusán & Marsh, 2016; Eswar et al., 2003). The absence of highlighted regions within the β-barrel motifs also aligned with evolutionary expectations, as these structural motifs are highly conserved to maintain functionality.

To further contextualize these findings, saliency maps were compared to the annotated positions of the two cupin domains, as retrieved from InterPro (Blum et al., 2025). Seven entries (P13917, Q8RVH5, A0A072VR, A0A7J0GR, A0A8S9FU, A0A0E0M7E9, and A0A444DIN8) were excluded due to the absence of domain annotations. This comparison revealed three distinct patterns. Pattern 1, the most frequent (61/99 in the testing set; 238/493 overall), highlighted the regions between the two cupin domains. Pattern 2 focused on regions within each cupin domain (24/99; 129/487), while pattern 3 showed outside the domains near the termini (14/99; 120/487). Representative saliency examples for Patterns 2 and 3 are provided in Fig. 5a and b, respectively. No clear association between pattern type and protein origin or functional subclass was observed.

The saliency pattern for globulin/glutelin exhibited a clear positional trend, prompting a comparison of amino acid compositions within the highlighted regions of the two classes. In the testing set, the highlighted regions of globulins contained notably higher rE (23.07 %) compared to glutelins (13.64 %). Other major differences included rG, rQ, and rS. This trend persisted across the entire dataset, with the magnitude of differences exceeding those observed in full-sequence comparisons (Fig. 6a). Moreover, combining the highly divergent amino acid ratios (rE, rG, rQ, and rS) further amplified the distinction in amino acid composition between globulin and glutelin (Fig. 6b). These findings suggested that local loop and helical regions outside the barrel motif exhibit marked differences in E, Q, G, and S residues between globulins

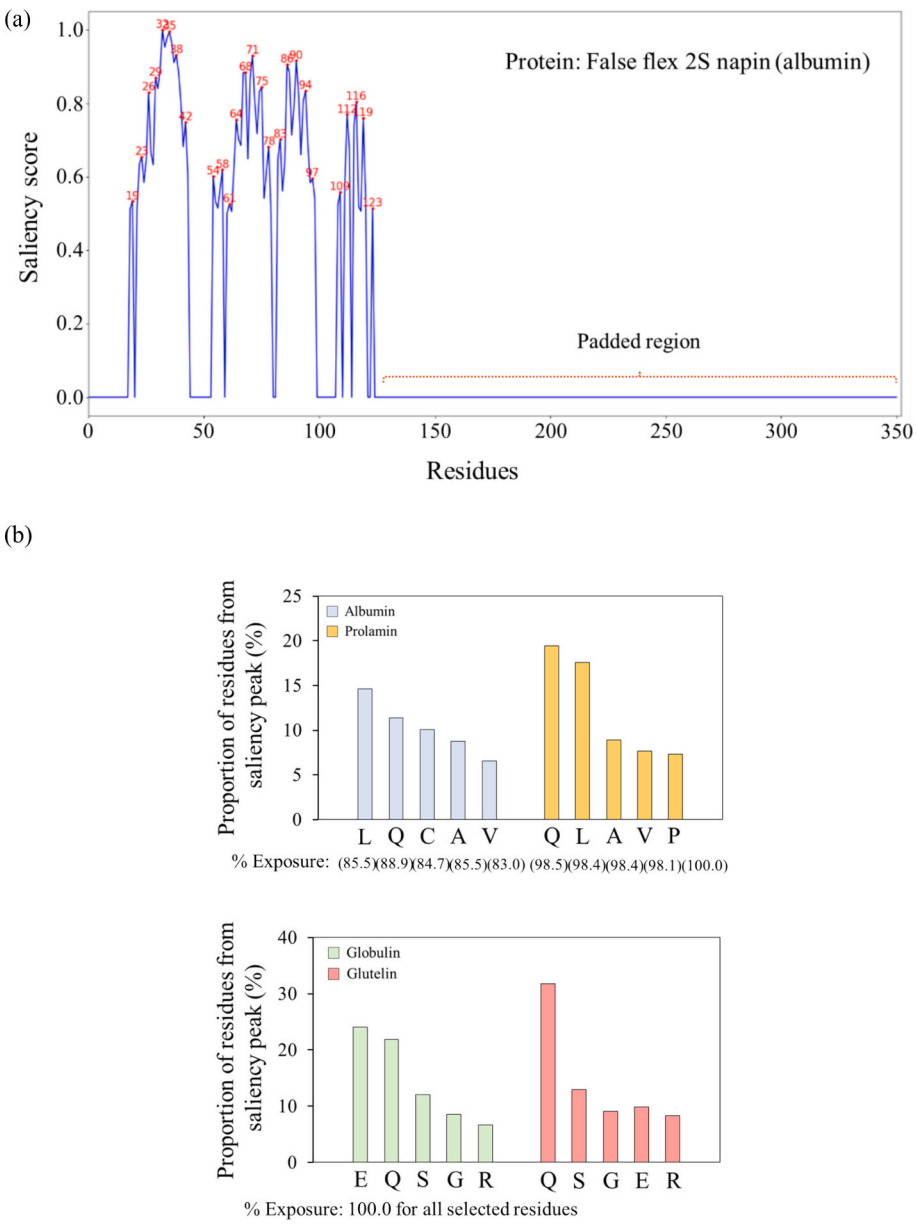

Fig. 4. Representative saliency score peaks illustrating the identification of key residues for seed storage protein classification (a) and top 5 residue types identified from the score peaks and their relative exposures (b).

and glutelins.

Given that GCN captures both structural connectivity and residue interactions that differentiate the two classes, the significance of these residues could be interpreted in two ways. First, the model could have identified inherent structural and compositional differences, where surface residues simply reflected distinct amino acid distributions and folding patterns across different classes without necessarily driving functional differences (Lapuschkin et al., 2019). Alternatively, these residues might actively contribute to solubility differences by influencing solvent interactions, hydration dynamics, and charge distribution. In this sense, the extent to which these selected residues for globulin and glutelin influenced solubility differences remained unclear. For example, one of the key residues, E, is already negatively charged at neutral environment, meaning it does not significantly alter the surface charge under basic conditions. Similarly, the other important residues for distinguishing globulin from glutelin, namely, G, Q, and S, are not protonatable, making it uncertain how they would directly affect glutelin solubility at high pH.

It should be noted that model outputs were interpreted in two distinct ways depending on the protein classes. For albumins and prolamins, the top-ranked SHAP features (e.g., charged residues, hydrophilic residue content, hydrophobic patch area) were biophysically consistent with known mechanisms of solubility in aqueous environments. In contrast, for globulins and glutelins, although both SHAP and saliency mapping identified discriminative features (e.g., Coulombic potential, surface loop composition), these did not align clearly with known salting-in mechanisms. To further explore this, a set of MD simulations was performed.

3.3. All atomic molecular dynamic simulation

3.3.1. AA-simulation of 2S soy albumin & 19 kDa maize zein in water & 70 % ETOH

To explore the interaction between SSPs and solvent environment, a set of AAMD simulations were performed on 2S soy albumin and 19 kDa maize zein (prolamin). When the 3D structure of 2S soy albumin was

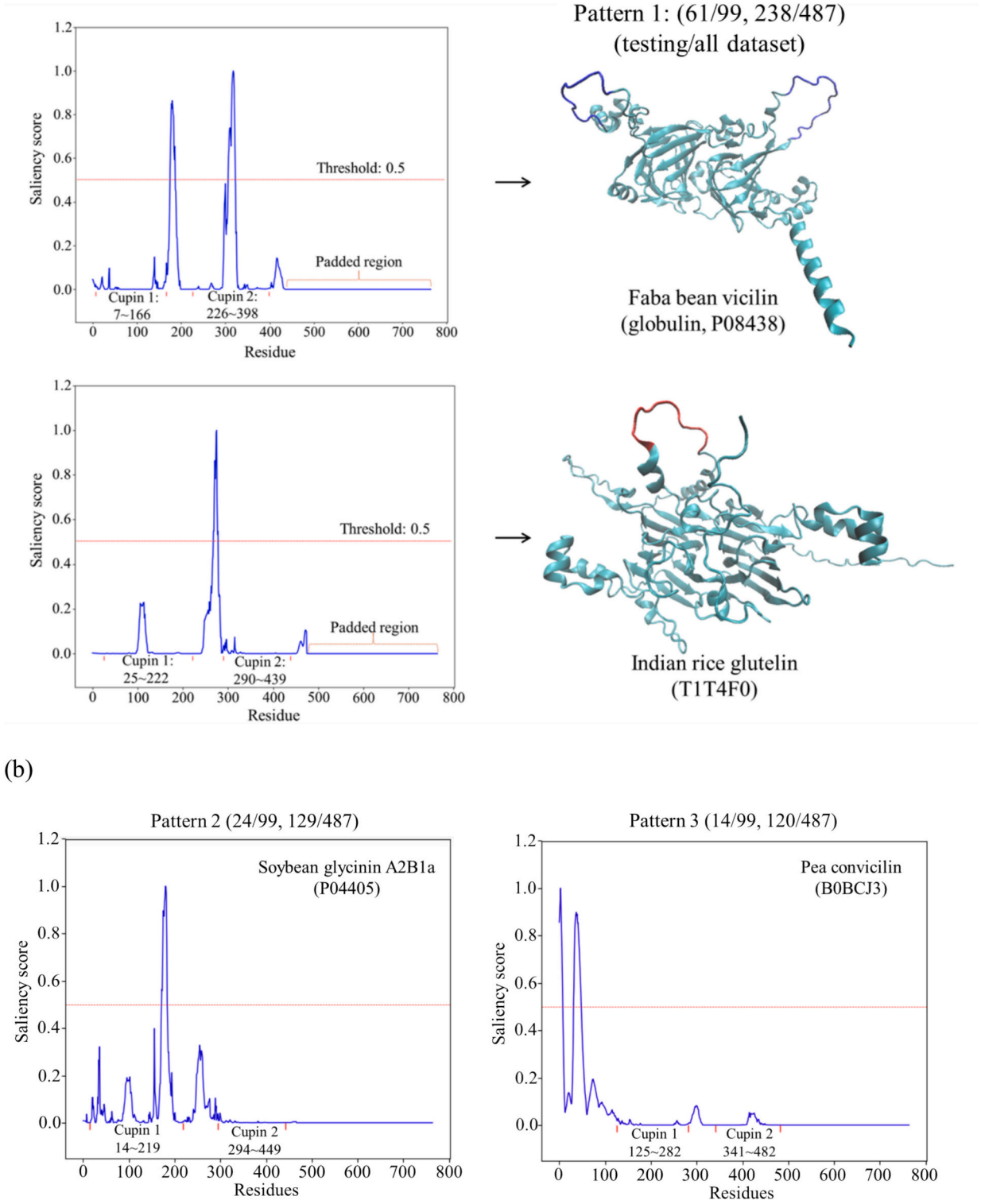

Fig. 5. Representative saliency scores highlighting high-saliency regions (>0.5), shown with 3D protein structure mapping for pattern 1 (a), and saliency profiles for patterns 2 and 3 (b).

subjected to water, only a minor change in the tertiary structure from the initial structure was observed (Fig. 7a). In 70 % ETOH, on the other hand, exposure of the hydrophobic core was noted, which aligned with the evolution of RMSD and hydrophobic SASA over the trajectory (Fig. 7b). In the case of 19 kDa maize zein, more dramatic structural changes were detected compared to those of albumin. In water, the initial zein structure collapsed to form a compact structure. On the other hand, the prolamin displayed markedly expanded structure in 70 % ETOH (Fig. 7c). This expanded structure of 19 kDa zein was similar to those observed from the small angle x-ray scattering (22/19 kDa zein mixture in methanol solution) (Tatham et al., 1993) and other MD simulation study for α-zein in 100 % ethanol (Christensen, 2024). Moreover, the increased hydrophobic SASA in 70 % ETOH also aligned with the previous report that zein composites displayed higher surface hydrophobicity in ethanol compared to in water (Anvar et al., 2025).

To identify the factors driving the observed exposures of

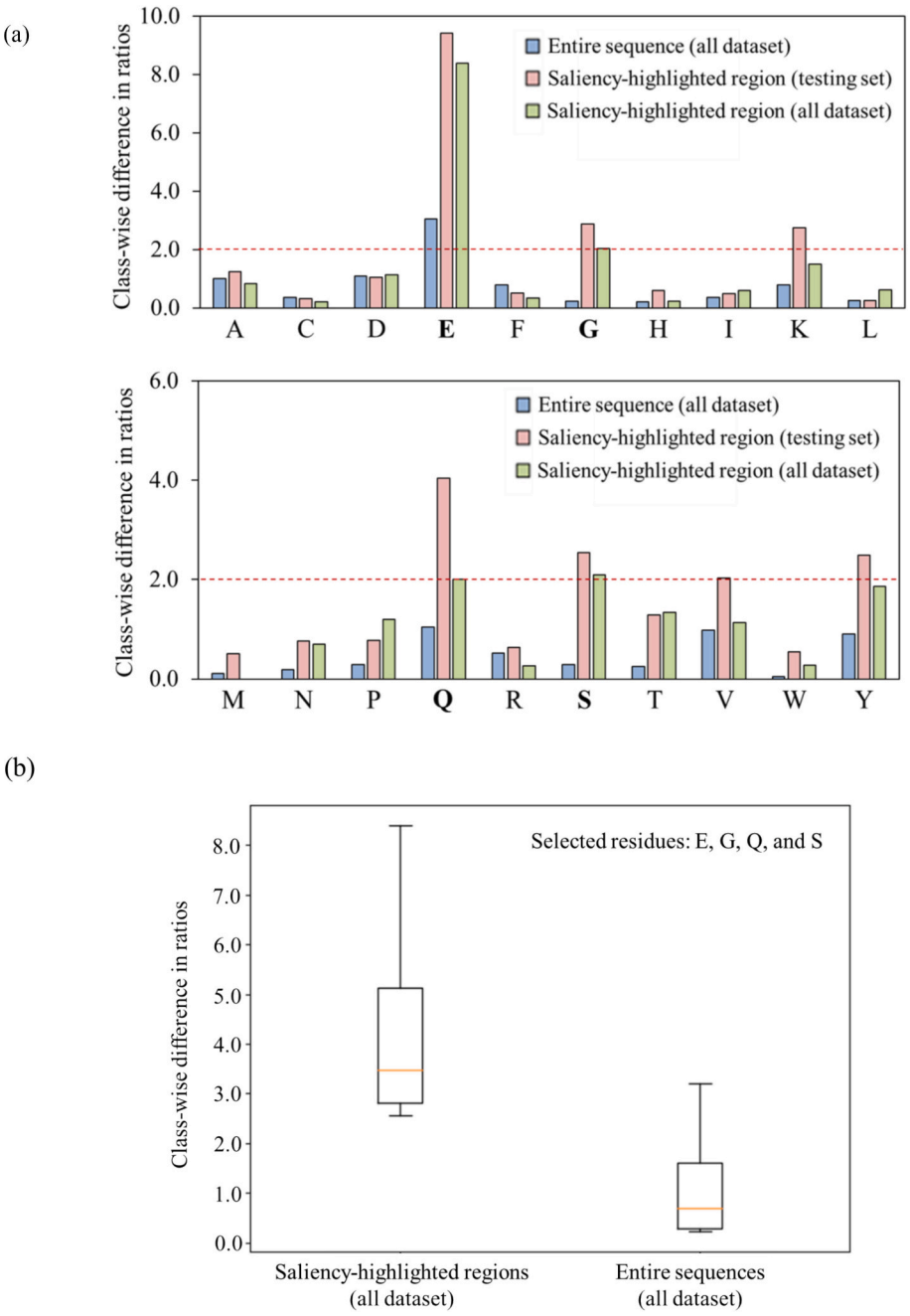

Fig. 6. Comparison of amino acid composition between globulins and glutelins, focusing on regions highlighted by saliency scores (>0.5) (a). Class-wise differences in the relative abundance of selected residues (E, G, Q, S) based on saliency-highlighted regions versus entire sequences (b).

hydrophobic residue in 70 % ETOH, energy decomposition and hydrogen bonding analysis were conducted. Energy calculations were performed using the gmx_energy function, which isolates the energy contributions of different entities throughout the trajectory under the applied force field. In the CHARMM force field series, non-bonded interactions are governed by Coulombic interactions and the Lennard-Jones (LJ) potential (Zhu et al., 2012). Between the two, hydrophobic interactions are dictated by LJ potential, as nonpolar molecules lack significant partial charges and primarily interact via van der Waals forces rather than electrostatics (Brooks et al., 2009). When the average LJ potentials of hydrophobic residues with the rest of the system were compared, those in 70 % ETOH exhibited a significant increase in the van der Waals interaction energy in both albumin and prolamin, suggesting the shift of the system in a way of promoting hydrophobic-hydrophobic interactions (Fig. 8a). On the other hand, when the number of intra-molecular hydrogen bonding was compared, interestingly, despite the observed denaturation (exposed hydrophobic residues), albumin formed an increased number of internal hydrogen bonds compared to ethanol, while no clear difference was observed in the prolamin (Fig. 8b).

As the observed changes in the number of hydrogen bonds (nohb) were counterintuitive to the idea that alcohol would disrupt intra-molecular hydrogen bonding (Thomas & Dill, 1993), the evolution of nohb between the solvent and proteins was computed (Fig. 9a-d). For better visibility, Savitski-Golay filter (Savitzky & Golay, 1964) with polyorder 3 was applied (dark). As illustrated in the figure, for both albumin and prolamin, the nohb between solvent and protein was noticeably lower in 70 % ETOH than in water. This was expected, as

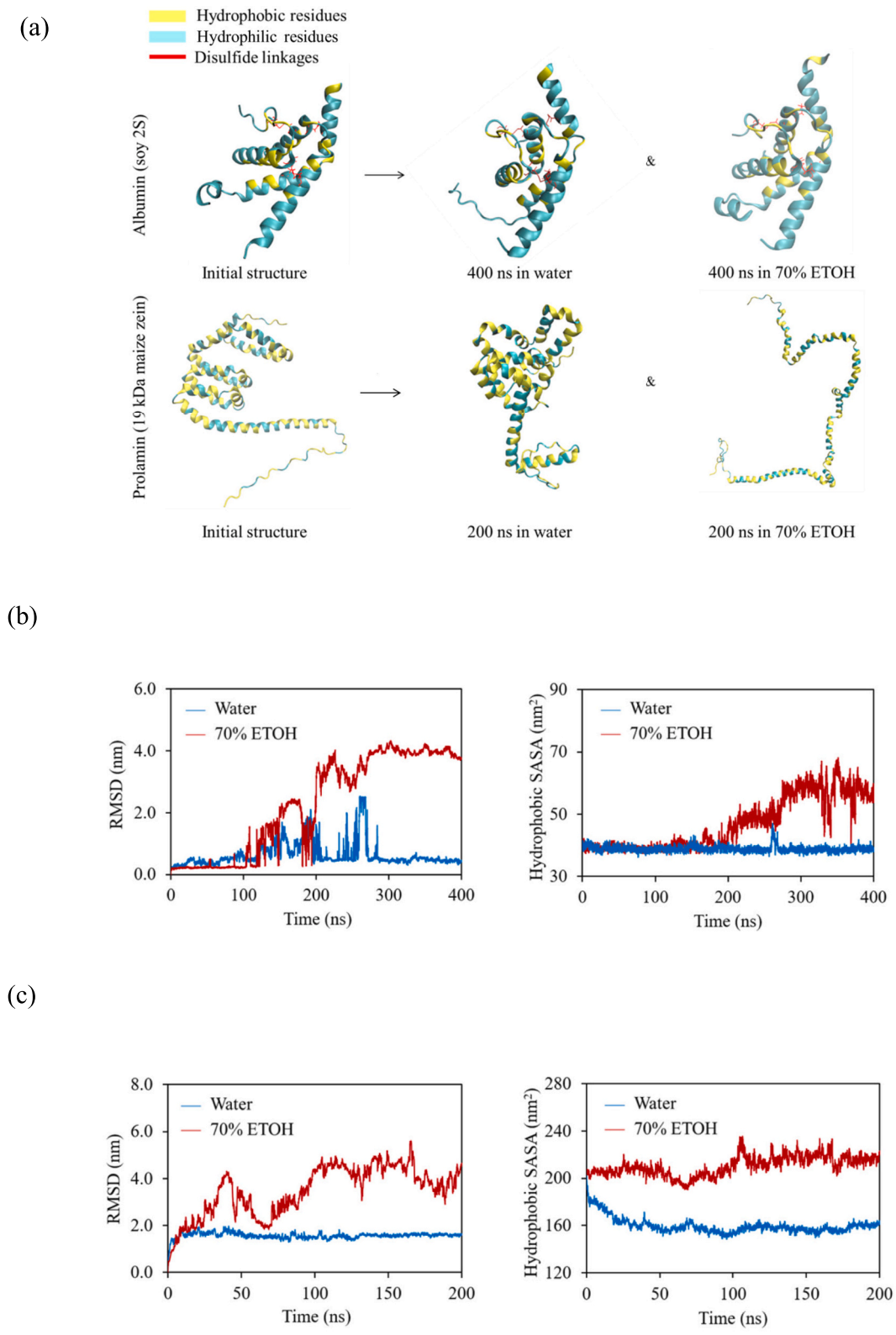

Fig. 7. Conformational changes of soy 2S albumin and 19 kDa zein in water and 70 % ethanol (a) and evolution of root-mean-square deviation (RMSD) and hydrophobic solvent-accessible surface area (SASA) of the albumin (b) and prolamin (c).

ethanol is larger than water and possesses only a single hydrogen bond donor, limiting its ability to form extensive hydrogen bond networks with proteins. For albumin, the protein-solvent nohb experienced an abrupt decrease in 70 % ETOH, while it remained stable in water (Fig. 9b, c). Moreover, in 70 % ETOH, the decrease in the solvent-protein nohb accompanied an increase in the intra-protein nohb (Fig. 9e), suggesting that hydrophilic albumin rearranged to form more intramolecular hydrogen bonds to maintain structural stability. This observed increase in internal hydrogen bonding in albumin under 70 % ethanol aligned with recent MD and CD studies on biomacromolecules

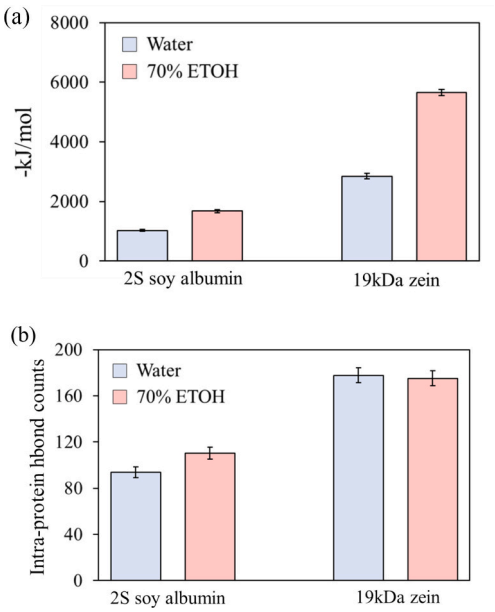

Fig. 8. Average Lennard-Jones potential energy between protein and system in water and 70 % ethanol (a) and intra-protein hydrogen bond counts (b).

such as spidroins, where ethanol exposure led to enhanced intra-molecular hydrogen bonding, helix/turn stabilization, and aggregation (Tolmachev et al., 2023). Furthermore, CD spectroscopy on human serum albumin revealed a recovery of α-helical content at ethanol concentrations above 55 %, alongside reversible aggregation behavior (Taboada et al., 2007), supporting the idea that ethanol can promote internal, structural rearrangement rather than complete destabilization.

In the case of prolamin, on the other hand, the solvent-protein nohb displayed abrupt decrease in water but not in 70 % ETOH (Fig. 9b, d). However, unlike the case of albumin, this decrease did not accompany gain in the intra-protein hydrogen bonds (Fig. 9f). These observations highlighted the differences in solvent adaptation between 2S albumin and 19 kDa zein: unlike the hydrophilic albumin, which reorganized its hydrogen bonding network against less polar environment, the hydrophobic prolamin instead relied on stronger hydrophobic interactions with solvent, which became significantly more stabilizing in ethanol. Moreover, prolamin did not compensate for the loss of solvent-protein nohb with intra-molecular hydrogen bonds, likely due to its intrinsically lower hydrogen bond-forming potential (higher proportion of hydrophobic residues). Taken together, the observed exposure of hydrophobic residues (Fig. 7b, c), the increase in van der Waals interactions (Fig. 8a), and the stability of the internal hydrogen bonding network (Fig. 8b), collectively suggested that hydrophobic interactions, rather than the disruption of internal hydrogen bonds, were the dominant factor driving the observed denaturation in 70 % ethanol for the two tested SSPs.

Notably, fluctuations were observed in the RMSD profiles of the 70 % ethanol systems, particularly for 19 kDa zein. These variations were more likely attributable to solvent-induced structural transitions rather than numerical instability. For example, both RMSD and hydrophobic SASA increased in albumin and prolamin under ETOH, indicating partial unfolding and hydrophobic core exposure. This behavior was consistent with prior findings that ethanol disrupts intra-molecular hydrophobic interactions and promotes protein expansion (Halder & Jana, 2021), as well as with the classical hydrophobic effect. These trends were supported by Fig. 7b and 9, where increased surface exposure of hydrophobic residues and rearrangement of internal hydrogen bonds seemed evident. Thus, the observed RMSD fluctuations likely reflected adaptation to the ethanol-rich environment. Given that the study aimed to compare solvent-induced structural changes rather than achieve full

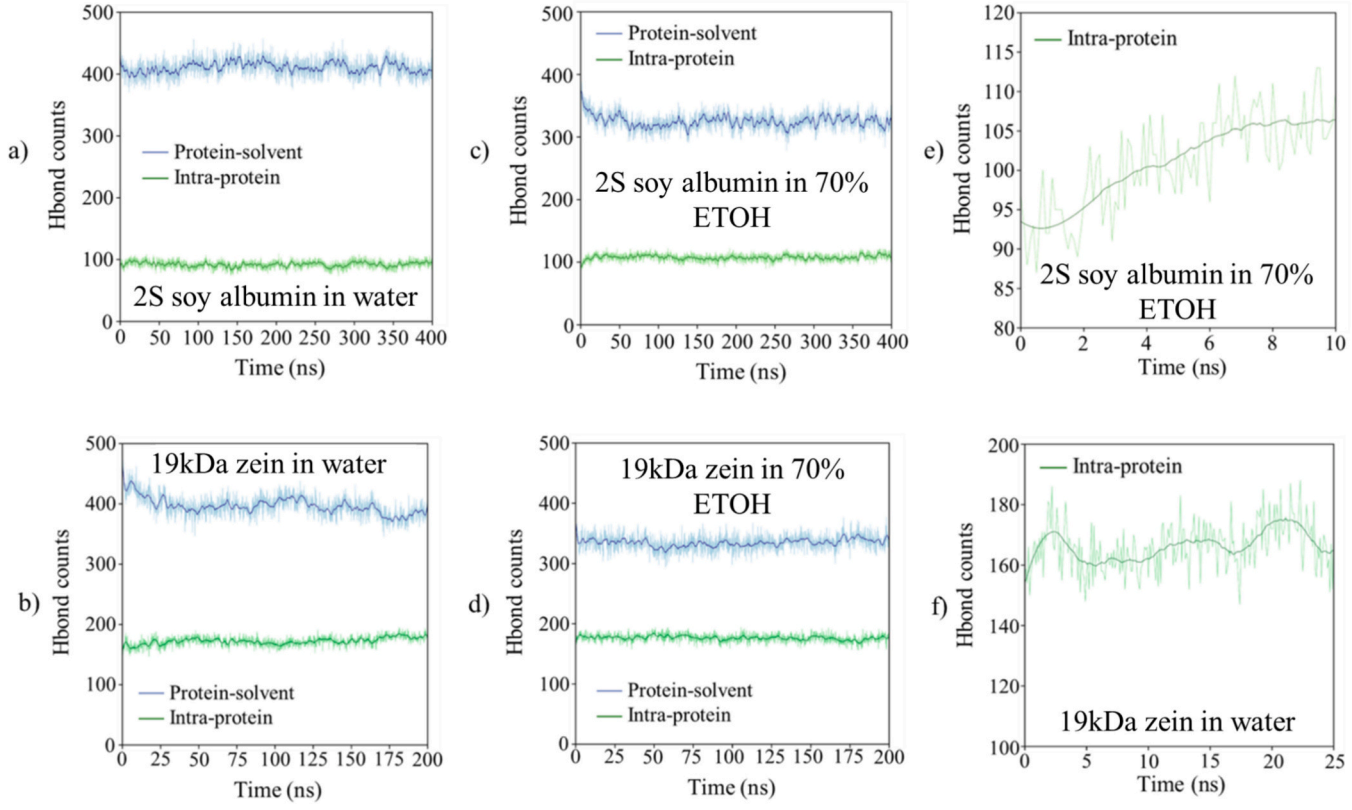

Fig. 9. Time evolution of protein-solvent and intra-protein hydrogen bonding in water and 70 % ethanol (a ~ d). Initial changes in the number of hydrogen bonds (e ~ f) for 2S soy albumin and 19 kDa zein.

equilibration, and considering the substantial differences observed across solvent conditions, it was regarded that the current trajectories were sufficient to support our conclusions.

3.3.2. AA-simulation of 11S pea globulin & rice glutelin in water & 1 M NaCl

When the structures of pea legumin A (globulin) and rice glutelin A1 were simulated in water and 1 M NaCl, no dramatic changes in the tertiary structures were observed for both proteins (Fig. 10a). Globulin displayed an increased RMSD in 1 M NaCl compared to in water (Fig. 10b), while no significant difference was observed for glutelin. However, the increase in RMSD for globulin appeared to originate from the highly charged loop region (Fig. 10a), rather than from global unfolding, as the RMSD of the β-barrel core and the hydrophobic SASA of the entire protein remained largely unchanged across the simulation in both water and saline solution (Fig. 10b, c).

Unlike the addition of hydrophobic solvents, the most direct effect of adding salt would be the disruption of electrostatic interactions between charged residues. To quantify the degree of disrupted electrostatic interactions, the Coulombic potential between the charged residues, as well as the number of salt bridges were computed (Fig. 11a,b). As can be seen from the figures, both globulin and glutelin displayed less stabilized coulombic potential in 1 M NaCl compared to water. This was also reflected in the decreased number of salt bridges in the saline solution. On the other hand, while 1 M NaCl effectively decreased the magnitude of the interaction between charged residues, no significant changes in the nohb of the barrel roll was observed (Fig. 11c). Therefore, considering the minimal changes in hydrophobic SASA (Fig. 10 b,c) and the

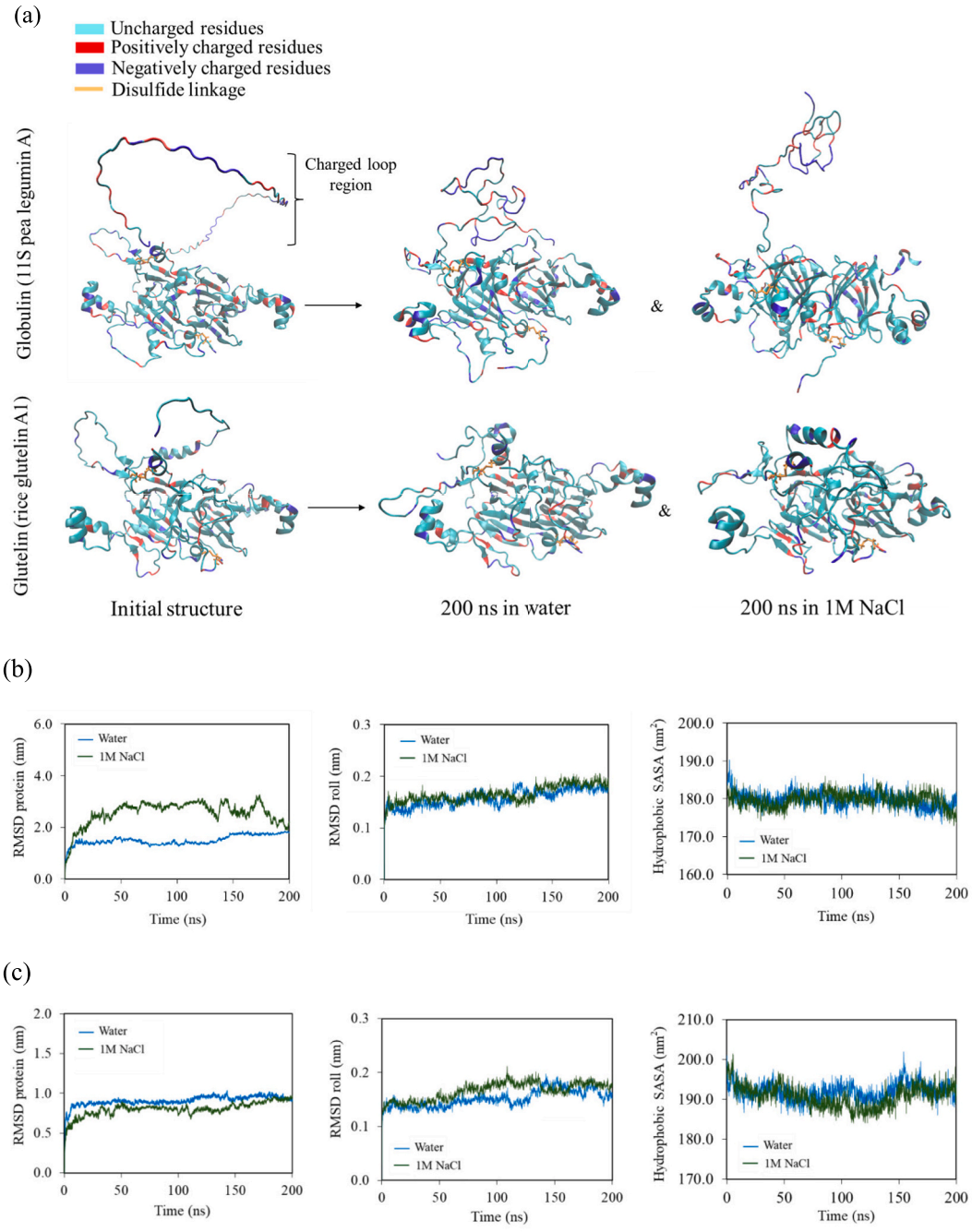

Fig. 10. Conformational changes of pea legumin A and glutelin A1 in water and 1 M NaCl (a), and the evolution of root-mean-square deviation (RMSD) for the protein and barrel roll, along with hydrophobic solvent-accessible surface area (SASA), for globulin (b) and glutelin (c) under the same conditions.

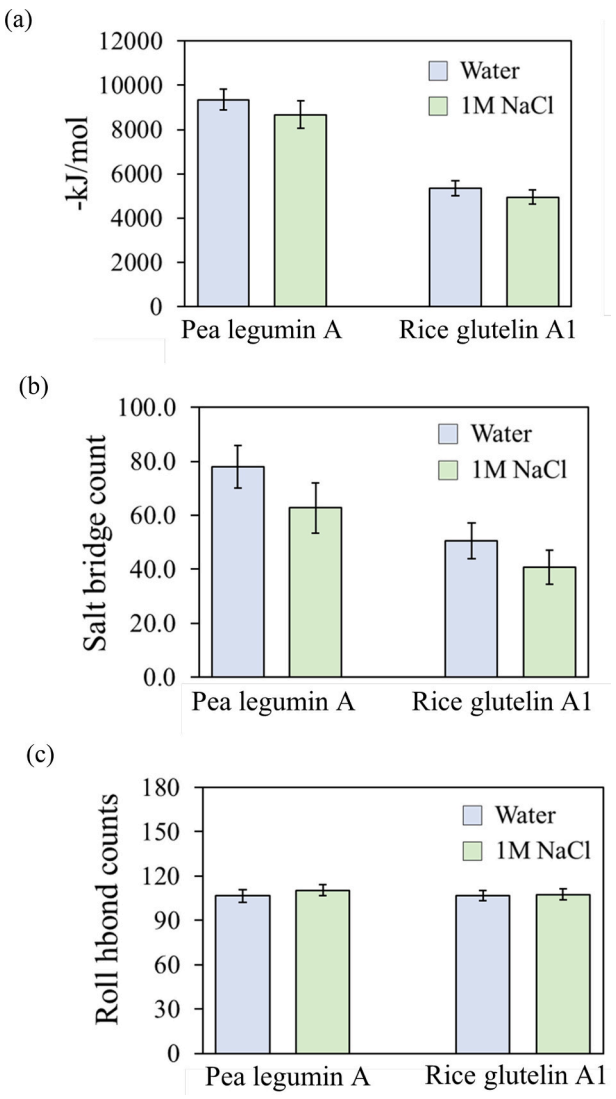

Fig. 11. The average Coulombic energy between protein and system (a), salt bridge count (b), and hydrogen bonding count at barrel rolls in water and 1 M NaCl (c) for pea legumin A and rice glutelin A1.

preserved hydrogen bonding network in the β-barrel core, our results suggested that although intra-protein electrostatic interactions were significantly weakened in saline conditions, this did not induce global conformational changes in the tertiary structures of the SSPs analyzed. This outcome should be interpreted within the context of the limited simulation timescale (200.0 ns), which may not capture slower structural transitions occurring over longer periods.

In addition to the structural changes induced by salt addition, one more question remained to be answered: whether the regions highlighted by the GCN model would have any meaningful attribution to the solubility difference or not. Fig. 12 presents the regions highlighted by the GCN model, and the root mean square fluctuation (RMSF) of residues. As can be seen from the figure, these regions were located at the regions with higher tendency to fluctuate. This higher variability was sensible, as they were exposed to the surface, and since these regions were located at mostly loop and helical regions, which are more susceptible to fluctuations than strand structures in most cases (Pathak et al., 2021). When the fluctuation in water and in salt was compared, only a slight increase in the overall fluctuation was observed in 1 M NaCl for both globulin and glutelin. The exception was at the region 3 of pea legumin, where notably higher movements in 1 M NaCl were found. This high fluctuation likely stemmed from the highly charged nature of region 3 and the residues nearby (Fig. 10a). More importantly, it was highly questionable whether the movement of this local loop region would trigger any meaningful difference in protein structure that would alter solubility. Instead, given the minimal changes in the overall tertiary structure (Fig. 10a), it was more plausible that the GCN model primarily identified regions with distinct residue compositions rather than those directly contributing to solubility differences. This result suggested that while the GCN model effectively captured sequence variation, additional factors such as intermolecular interactions must be considered to fully understand the solubility differences between globulin and glutelin.

3.4. Coarse-grained molecular dynamics simulation

3.4.1. CG-simulation of 2S soy albumin & 19 kDa maize zein in water & 70 % ETOH

While simulating a single protein under different solvent environments provides insights into structural changes, protein solubility is a property arising from multiple protein-protein interactions. To investigate the role of intermolecular interactions, CG simulations were conducted with five monomers at 1 % concentration. The effective simulation time was computed using the standard Martini speed conversion factor of 4.0 (Marrink et al., 2007). As shown in Fig. 13a, soy 2S albumin and 19 kDa zein exhibited distinct solubility behaviors, forming aggregates in solvents known to reduce their solubility. Additionally, the number of residual contacts was computed using the MDAnalysis package (Gowers et al., 2019) with a 4.5 Å distance contact threshold (Fig. 13b). For albumins, while some prolonged contacts were observed in water, the degree of contacts in 70 % ETOH was notably higher and formed more abruptly under the given simulation time. The partial aggregation of albumins in water was also observed in our previous CG simulation of multiple albumins in water, which did not incorporate PTMs (i.e., signal peptide and propeptide regions were not removed) (Kwon et al., 2024). In contrast to albumins, prolamins exhibited an opposite trend, where the number of contacts increased in water. Notably, in 70 % ETOH, prolamins initially formed contacts but subsequently dissociated after 500.0 ns, further highlighting the ability of hydrophobic solvents to prevent prolamin aggregation.

To determine the residues contributing the most to the observed aggregation, residue-specific contacts were computed throughout the trajectory, and the top five residues with the highest degree of contacts were identified (Fig. 13c). For albumins in 70 % ethanol, charged (Asp, Glu, Lys) and polar (Ser, His) residues were the primary contributors to the observed aggregation in 70 % ETOH, indicating that electrostatic and hydrophilic interactions drove the clustering in the less polar solvent.

Conversely, for prolamins in water, hydrophobic residues (Phe, Leu, Pro, Ala) dominated, with only a minor contribution from the polar residue Gln. The identified residues highlighted that hydrophobic interactions predominantly governed the aqueous aggregation of 19 kDa zein. The aqueous aggregation behavior of zeins, dominated by hydrophobic residues, aligned with previous findings that hydrophobic dehydration-driven clustering facilitates aggregate formation (Thirumalai et al., 2012). In summary, the simulations suggested the solvent-dependent aggregation behaviors of the two SSPs, where hydrophilic albumin aggregated via electrostatic interactions in less polar solvents, while hydrophobic prolamin agglomerated through hydrophobic interactions in aqueous environments.

3.4.2. CG-simulation of 11S pea legumin A & rice glutelin A1 in water & 1 M NaCl

Unlike albumins and prolamins, globulins and glutelins did not exhibit the expected solubility behavior in different solvents. Specifically, both protein types aggregated in water but remained soluble in 1 M NaCl (Fig. 14a), suggesting that electrostatic interactions played a crucial role in their clustering. On the other hand, residue-specific

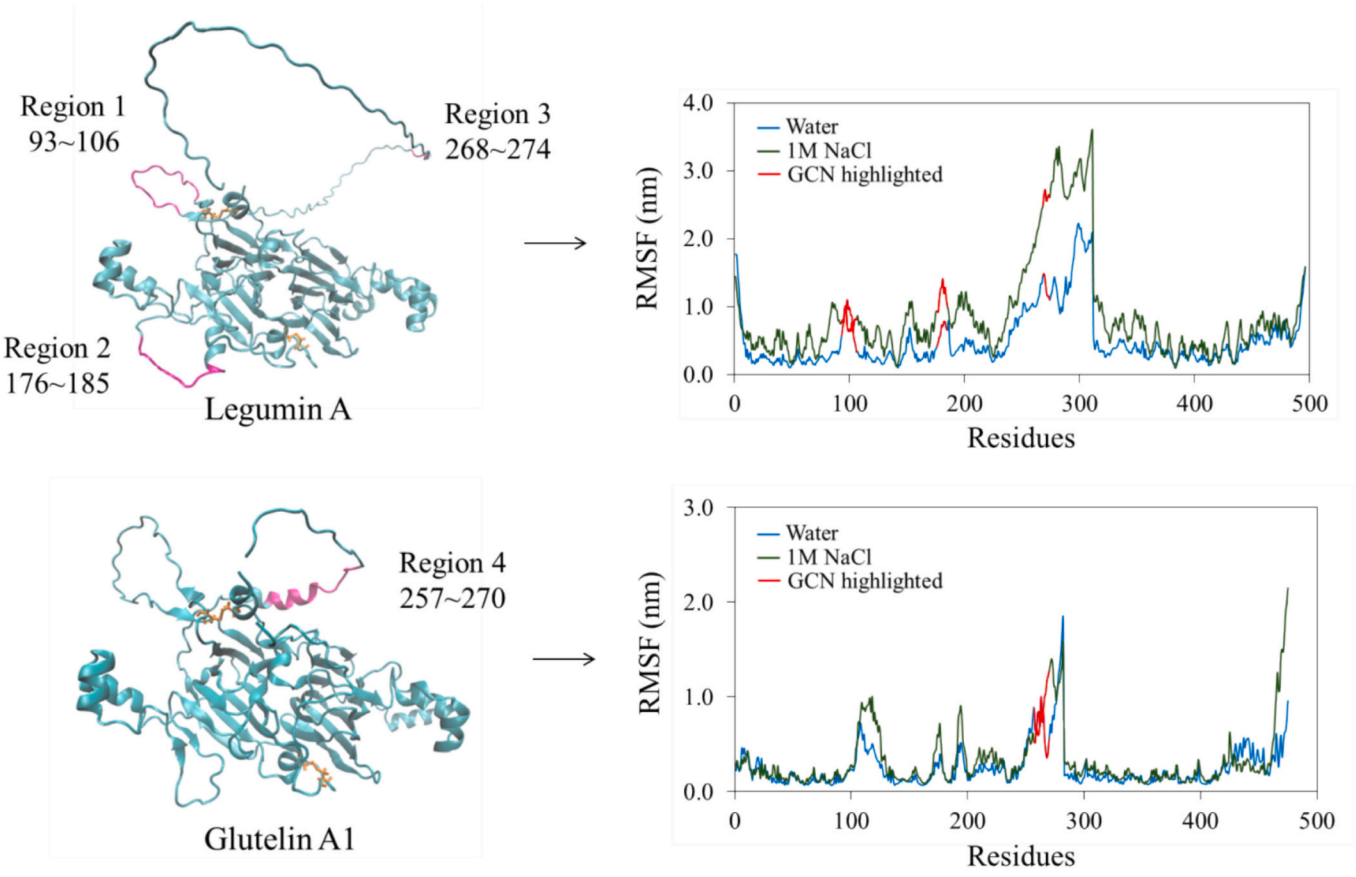

Fig. 12. Saliency-highlighted regions of pea legumin A and rice glutelin A1 with corresponding root-mean-square fluctuation (RMSF) profiles.

contact analysis in water revealed a more diverse interaction profile for globulins and glutelins (Fig. 14b). These interactions included electrostatic attractions (Lys, Arg), π-π stacking (Phe, Tyr), hydrogen bonding (Ser, His), and hydrophobic effects (Phe, Ala). Thus, the contact-prone residue composition of aggregated globulins and glutelins contrasted with that of 2S soy albumins and 19 kDa zein. While albumin aggregation was primarily driven by electrostatic interactions and prolamin aggregation by hydrophobic interactions, pea legumin A and rice glutelin A1 aggregation appeared to be influenced by a more complex interplay of multiple interactions.

To quantify the role of the saliency-highlighted regions in the observed aggregations, the average number of contacts formed by these regions was compared to that of all other residues. For a clear comparison, three percentiles of contact-forming residues (top 10 %, 25 %, and 50 %) were analyzed. As shown in Fig. 14c, for both globulin and glutelin, the saliency-highlighted regions exhibited a significantly lower number of contacts throughout the simulation compared to other contact-forming residues, suggesting that these regions did not directly contribute to aggregation.

Contrary to the experimentally observed solubility differences between globulin and glutelin, simulations of their monomeric states displayed similar aggregation behavior—both were insoluble in water but remained soluble in 1 M NaCl. Furthermore, none of the three additional analysis protocols (global features, residue-level features, and all-atom simulations) provided clear evidence to explain their solubility differences. The lack of a direct correlation between specific features and the pH-dependent solubility of glutelin, along with the similar dynamic and aggregation behaviors observed in MD simulations, suggested that these solubility differences might not stem from monomeric properties alone. Instead, it was likely that higher-order structures, particularly inter-protein disulfide networks, would play a crucial role in glutelin insolubility. This hypothesis aligns with the fact that glutelins are typically extracted under reducing conditions (e.g., DTT) in addition to alkaline solvents (e.g., NaOH), indicating that disulfide-mediated supramolecular assemblies significantly influence their solubility properties. Moreover, glutelins and 11S globulins are thought to share a common ancestral gene and display considerable sequence and structural similarity. Our BLAST pairwise alignment (Altschul et al., 1990) yielded 36 % sequence identity and 57 % similarity (both identical residues and those with similar physicochemical properties) between the two proteins used in this study. The observed similarity between the two groups is further supported by previous work showing high sequence identity between rice glutelins and other 11S-type globulins, such as chickpea legumin (Chang & Alli, 2012). Notably, related storage proteins like oat 12S globulin have also shown significant sequence similarity to rice glutelins (Shotwell et al., 1990), reinforcing their evolutionary and structural relatedness.

Additional support for this idea comes from the work of Katsube et al., where soy glycinin (globulin) gene was expressed in rice (Katsube et al., 1999). Their study found that while glycinin was generally solubilized by salt solutions, it partially formed insoluble complexes with endogenous glutelins through disulfide networks. Further experimental evidence includes multiple reports where incorporation of reducing agents significantly increased extraction yields for glutelins (Roy et al., 2023; Wilson et al., 1981). Moreover, it was reported that rice glutelin fractions are extensively cross-linked via disulfide networks (Amagliani et al., 2017). Consistent with this, our computational findings suggested that monomer-based analysis alone might be insufficient to explain the experimental solubility differences in globulin and glutelin classes, highlighting the critical role of higher-order supramolecular structures in glutelin insolubility.

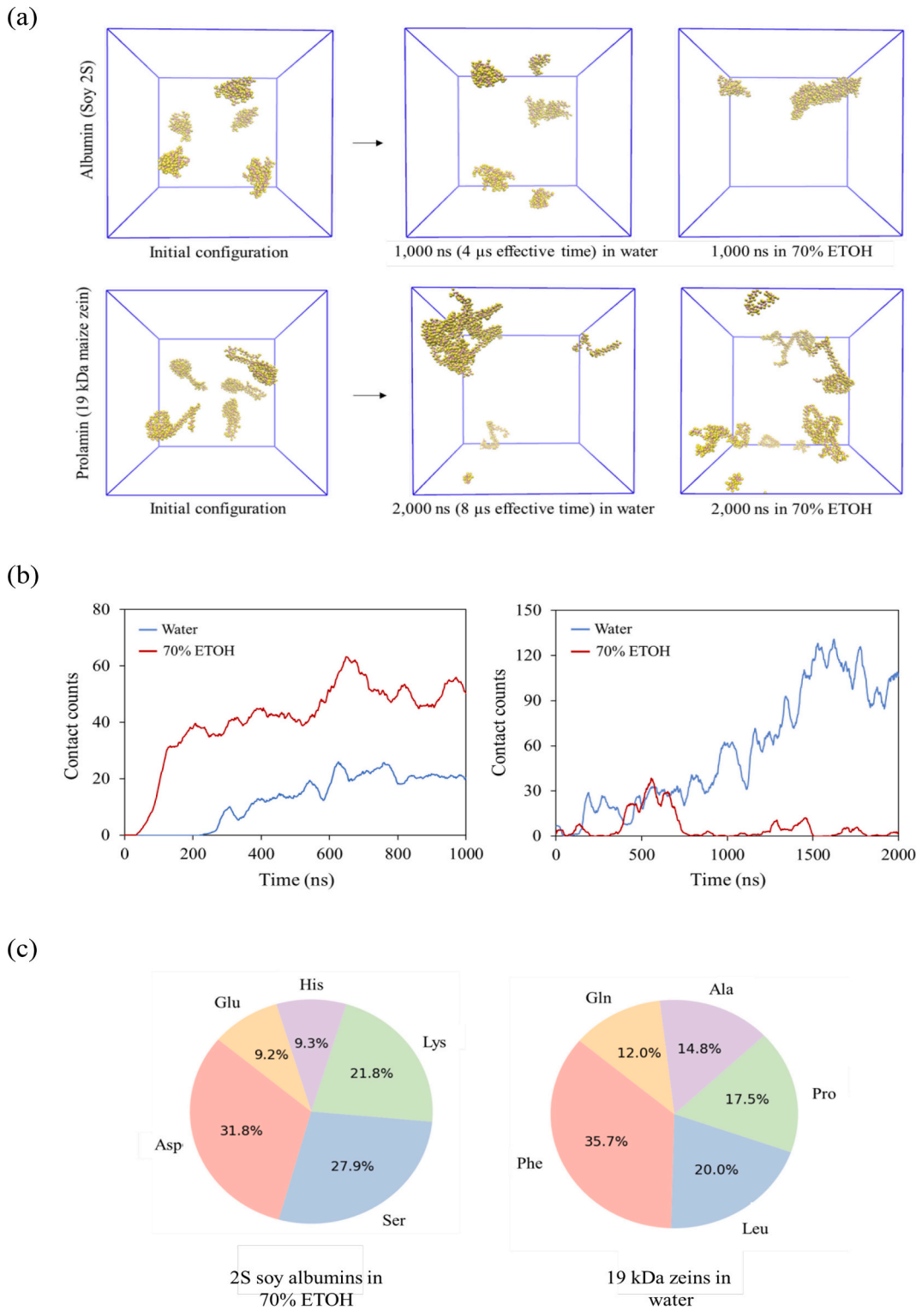

Fig. 13. Coarse-grained molecular dynamics simulation of multiple soy 2S albumin and 19 kDa zein (a), evolution of contacts (b), and top 5 residues involved in aggregation (c).

4. Conclusion

This study systematically analyzes the monomeric properties of Osborne-classified seed storage proteins (SSPs) through a multi-faceted computational approach, integrating structural modeling, machine learning, and molecular dynamics simulations with the largest dataset to date. Distinctive physicochemical features defining each class were identified, such as the high cysteine content of albumins and the low abundance of charged amino acids in prolamins. Compared to albumins and prolamins, globulins and glutelins exhibited similar physicochemical properties, reflecting their evolutionary and structural closeness. Furthermore, interpretation of support vector classifiers trained on global physicochemical features successfully delineated key differences among the protein classes. Molecular dynamics simulations further

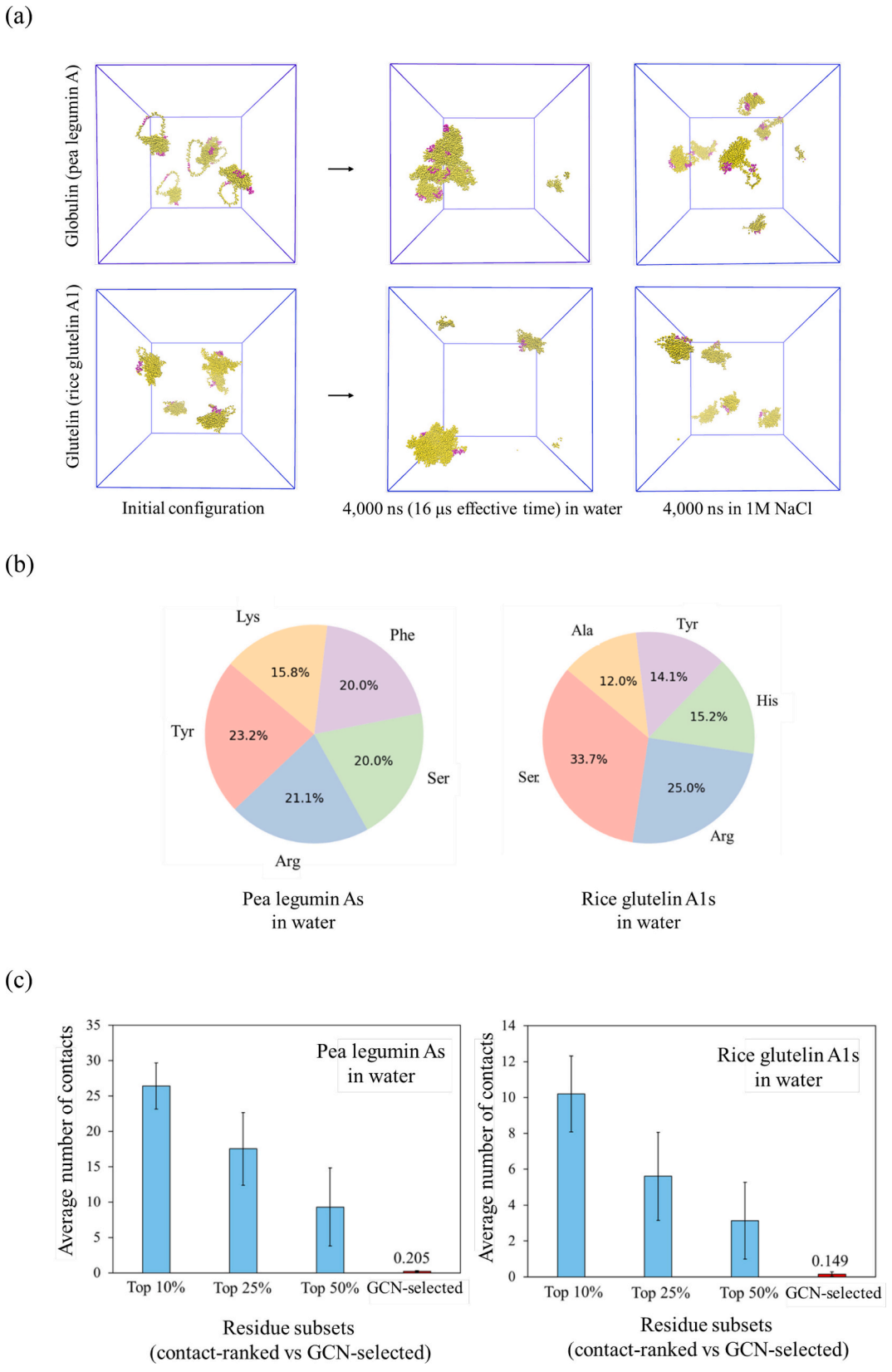

Fig. 14. Coarse-grained molecular dynamics simulation of multiple 11S legumin A and glutelin A1 with saliency-highlighted regions (purple) (a), top 5 residues involved in aggregation (b), and comparison of contact frequencies between residues with high contact ranks and GCN-selected regions (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

elucidated distinct dynamic and aggregation behaviors in representative model proteins, where albumins (2S soy albumin) exhibited greater aqueous solubility, whereas prolamins (19 kDa zein) preferentially dissolved in ethanol, reflecting their reduced solubility in water.

For globulins and glutelins, saliency mapping from a graph convolutional network classifier revealed distinct residue-level variations in the loop and helical regions outside their β-barrel rolls, where significant differences in the distributions of E (Glu), Q (Gln), S (Ser), and G (Gly) residues were observed between the two classes. However, all-atomic and coarse-grained molecular dynamics simulations of model proteins (pea legumin A and rice glutelin A1) did not reveal significant differences in their dynamics and aggregation patterns. Combined with similar physicochemical profiles, these findings provide computational support for the experimental evidence suggesting that the solubility differences between globulins and glutelins are primarily dictated by supramolecular interactions, particularly disulfide-mediated network formation.

Overall, this study integrates modern computational approaches to provide molecular insights into the distinct properties of each Osborne class and their solubility under different solvent conditions. The proposed framework and key findings not only advance our understanding of seed storage protein behavior but also have valuable implications for food science. Moreover, by elucidating the molecular basis of distinct behaviors of each Osborne class, this study establishes a foundation for improving protein dispersibility, solubility, and stability in food hydrocolloid systems, ultimately contributing to the development of plant-based food products with desired functional properties. For example, the findings regarding globulin and glutelin solubility may inform future process design by emphasizing the need to critically evaluate whether conventional extraction conditions, such as low pH or high ionic strength, are sufficient to disrupt supramolecular disulfide-linked networks, or whether additional strategies (e.g., reducing agents, enzymatic cleavage) are required.

The limitations of this study should be addressed in future research. While SHAP and saliency mapping improved interpretability for both machine learning and deep learning models, the inner workings of the GCN models could remain largely opaque. Although GCN captures residue-level connectivity through both node and edge matrices, our current models focus only on node features due to the complexity of analyzing 2D contact maps. Despite this limitation, the consistent localization of high-saliency regions across ~500 proteins suggested that the model's decisions were biologically meaningful and non-random. Moreover, given that native SSPs often exist as oligomeric assemblies, it is important to note that the present study was limited to analyzing monomeric properties. For future studies, it would be valuable to systematically examine glutelin solubility under different salt types and concentrations, particularly if high-purity isolation and monomeric stabilization are achievable. Additionally, simulations of oligomeric assemblies and interprotein interactions (e.g., legumin–vicilin systems) could offer deeper insights into aggregation-prone interfaces and solvation dynamics beyond the monomeric level.

CRediT authorship contribution statement

Hyukjin Kwon: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yixiang Xu:** Writing – review & editing, Resources, Investigation. **Xuan Xu:** Writing – review & editing, Methodology, Investigation. **Yonghui Li:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This is Contribution No. 25-195-J from the Kansas Agricultural Experiment Station. This work was in part supported by the USDA Agricultural Research Service Non-Assistance Cooperative Project (Grant Accession No. 443993, 439200) and USDA National Institute of Food and Agriculture Hatch project (Grant Accession No. 7003330). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2025.117322.

Data availability

Data will be made available on request.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX, 1*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., … J. M.. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w

Abrusán, G., & Marsh, J. A. (2016). Alpha helices are more robust to mutations than beta strands. *PLoS Computational Biology, 12*(12), Article e1005242. https://doi.org/10.1371/journal.pcbi.1005242

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amagliani, L., O'Regan, J., Kelly, A. L., & O'Mahony, J. A. (2017). Composition and protein profile analysis of rice protein ingredients. *Journal of Food Composition and Analysis, 59*, 18–26. https://doi.org/10.1016/j.jfca.2016.12.026

Anandakrishnan, R., Aguilar, B., & Onufriev, A. V. (2012). H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research, 40*(W1), W537–W541. https://doi.org/10.1093/nar/gks375

Anvar, A., Azizi, M. H., & Gavlighi, H. A. (2025). Exploring the effect of natural deep eutectic solvents on zein: Structural and functional properties. *Current Research in Food Science, 10*, Article 100965. https://doi.org/10.1016/j.crfs.2024.100965

Bera, I., O'Sullivan, M., Flynn, D., & Shields, D. C. (2023). Relationship between protein digestibility and the proteolysis of legume proteins during seed germination. *Molecules, 28*(7), 3204. https://doi.org/10.3390/molecules28073204

Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., … A.. (2025). InterPro: The protein sequence classification resource in 2025. *Nucleic Acids Research, 53*(D1), D444–D456. https://doi.org/10.1093/nar/gkae1082

Boulter, D., & Derbyshire, E. (2014). The general properties, classification, and distribution of plant proteins. In G. Norton (Ed.), Plant proteins: Proceedings of the easter School in Agricultural Science (pp. 1–10). Butterworth-Heinemann.

Brooks, B. R., Brooks, C. L., III, MacKerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., … M.. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry, 30*(10), 1545–1614. https://doi.org/10.1002/jcc.21287

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G., Passos, A., Weiss, R., Dubourg, V., Fabijan, C., … Varoquaux, G. (2013). *API design for machine learning software: Experiences from the scikit-learn project. arXiv preprint* arXiv:1309.0238. https://doi.org/10.48550/arXiv.1309.0238.

Chang, Y. W., & Alli, I. (2012). In silico assessment: Suggested homology of chickpea (*Cicer arietinum L.*) legumin and prediction of ACE-inhibitory peptides from chickpea proteins using BLAST and BIOPEP analyses. *Food Research International, 49*(1), 477–486. https://doi.org/10.1016/j.foodres.2012.07.006

Christensen, N. J. (2024). Conformations of a highly expressed Z19 α-zein studied with AlphaFold2 and MD simulations. *PLoS One, 19*(5), Article e0293786. https://doi.org/10.1371/journal.pone.0293786

Clement, G., Boquet, D., Mondoulet, L., Lamourette, P., Bernard, H., & Wal, J. M. (2005). Expression in *Escherichia coli* and disulfide bridge mapping of PSC33, an allergenic 2S albumin from peanut. *Protein Expression and Purification, 44*(2), 110–120. https://doi.org/10.1016/j.pep.2005.05.015

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *Journal of Chemical Physics, 98*, 10089–10092. https://doi.org/10.1063/1.464397

Day, L. (2013). Proteins from land plants–potential resources for human nutrition and food security. *Trends in Food Science & Technology, 32*(1), 25–42. https://doi.org/10.1016/j.tifs.2013.05.005

Dias, F. F., Yang, J. S., Pham, T. T. K., Barile, D., & de Moura Bell, J. M. L. (2024). Unveiling the contribution of Osborne protein fractions to the physicochemical and functional properties of alkaline and enzymatically extracted green lentil proteins. *Stainable Food Proteins, 2*(2), 61–77. https://doi.org/10.1002/sfp2.1026

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: An automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research, 32*(suppl_2), W665–W667. https://doi.org/10.1093/nar/gkh381

Dziuba, J., Szerszunowicz, I., Nałęcz, D., & Dziuba, M. (2014). Proteomic analysis of albumin and globulin fractions of pea (*Pisum sativum L.*) seeds. *Acta Scientiarum Polonorum. Technologia Alimentaria, 13*(2), 181–190. https://doi.org/10.17306/j.afs.2014.2.7

Esen, A. (1986). Separation of alcohol-soluble proteins (zeins) from maize into three fractions by differential solubility. *Plant Physiology, 80*(3), 623–627. https://doi.org/10.1104/pp.80.3.623

Eswar, N., Ramakrishnan, C., & Srinivasan, N. (2003). Stranded in isolation: Structural role of isolated extended strands in proteins. *Protein Engineering, 16*(5), 331–339. https://doi.org/10.1093/protein/gzg046

Evans, D. J., & Holian, B. L. (1985). The nose–hoover thermostat. *Journal of Chemical Physics, 83*(8), 4069–4074. https://doi.org/10.1063/1.449071

Fukushima, D. (1991). Structures of plant storage proteins and their functions. *Food Reviews International, 7*(3), 353–381. https://doi.org/10.1080/87559129109540916

Furuta, M., Yamagata, H., Tanaka, K., Kasai, Z., & Fujii, S. (1986). Cell-free synthesis of the rice glutelin precursor. *Plant and Cell Physiology, 27*(6), 1201–1204. https://doi.org/10.1093/oxfordjournals.pcp.a077206

Gama, F., Marques, A. G., Leus, G., & Ribeiro, A. (2018). Convolutional neural network architectures for signals supported on graphs. *IEEE Transactions on Signal Processing, 67*(4), 1034–1049. https://doi.org/10.1109/TSP.2018.2889921

Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., … Beckstein, O. (2019). *MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations* (Vols. No. LA-UR-19-29136). Los Alamos National Laboratory. https://doi.org/10.25080/Majora-629e541a-00e

Grumezescu, A. M., & Holban, A. M. (Eds.). (2018). *19. Role of materials science in food bioengineering*. Academic Press.

Halder, R., & Jana, B. (2021). Exploring the role of hydrophilic amino acids in unfolding of protein in aqueous ethanol solution. *Proteins: Structure, Function, and Bioinformatics, 89*(1), 116–125. https://doi.org/10.1002/prot.25999

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, 89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hess, B., Bekker, H., Berendsen, H. J., & Fraaije, J. G. (1997). LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry, 18*(12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Huang, J., & MacKerell, A. D., Jr. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry, 34*(25), 2135–2145. https://doi.org/10.1002/jcc.23354

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics, 14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin De La Société Vaudoise Des Sciences Naturelles, 37*, 547–579. https://doi.org/10.5169/seals-266450

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules, 22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Katsube, T., Kurisaka, N., Ogawa, M., Maruyama, N., Ohtsuka, R., Utsumi, S., & Takaiwa, F. (1999). Accumulation of soybean glycinin and its assembly with the glutelins in rice. *Plant Physiology, 120*(4), 1063–1074. https://doi.org/10.1104/pp.120.4.1063

Katsube-Tanaka, T., Duldulao, J. B. A., Kimura, Y., Iida, S., Yamaguchi, T., Nakano, J., & Utsumi, S. (2004). The two subfamilies of rice glutelin differ in both primary and higher-order structures. *Biochimica et Biophysica Acta, Proteins and Proteomics, 1699*(1–2), 95–102. https://doi.org/10.1016/j.bbapap.2004.02.001

Kim, M. S., & Jeong, Y. H. (2002). Extraction and electrophoretic characterization of rice proteins. *Preventive Nutrition and Food Science, 7*(4), 437–441. https://doi.org/10.3746/jfn.2002.7.4.437

Koshiyama, I. (1972). Purification and physico-chemical properties of 11S globulin in soybean seeds. *International Journal of Peptide and Protein Research, 4*(3), 167–176. https://doi.org/10.1111/j.1399-3011.1972.tb03416.x

Krishnan, H. B., & Okita, T. W. (1986). Structural relationship among the rice glutelin polypeptides. *Plant Physiology, 81*(3), 748–753. https://doi.org/10.1104/pp.81.3.748

Kroon, P. C., Grünewald, F., Barnoud, J., van Tilburg, M., Souza, P. C., Wassenaar, T. A., & Marrink, S. J. (2022). Martinize2 and vermouth: Unified framework for topology generation. *eLife*. https://doi.org/10.7554/eLife.90627.2

Kumar, N., & Srivastava, R. (2024). Deep learning in structural bioinformatics: Current applications and future perspectives. *Briefings in Bioinformatics, 25*(3), Article bbae042. https://doi.org/10.1093/bib/bbae042

Kwon, H., Du, Z., & Li, Y. (2024). AlphaFold 2-based stacking model for protein solubility prediction and its transferability on seed storage proteins. *International Journal of Biological Macromolecules, 278*, Article 134601. https://doi.org/10.1016/j.ijbiomac.2024.134601

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*(1), 1096. https://doi.org/10.1038/s41467-019-08987-4

Lijnzaad, P., Berendsen, H. J., & Argos, P. (1996). A method for detecting hydrophobic patches on protein surfaces. *Proteins: Structure, Function, and Bioinformatics, 26*(2), 192–203. https://doi.org/10.1002/(SICI)1097-0134(199610)26:2<192::AID-PROT9>3.0.CO;2-I

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A structured self-attentive sentence embedding*. arXiv preprint arXiv:1703.03130. https://doi.org/10.48550/arXiv.1703.03130

Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30. arXiv. 10.48550/arXiv.1705.07874.

Marla, S., Bharatiya, D., Bala, M., Singh, V., & Kumar, A. (2010). Classification of rice seed storage proteins using neural networks. *Journal of Plant Biochemistry and Biotechnology, 19*, 123–126. https://doi.org/10.1007/BF03323450

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & De Vries, A. H. (2007). The MARTINI force field: Coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B, 111*(27), 7812–7824. https://doi.org/10.1021/jp071097f

Meiler, J., Müller, M., Zeidler, A., & Schmäschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular Modeling Annual, 7*(9), 360–369. https://doi.org/10.1007/s008940100038

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science, 32*(11), Article e4792. https://doi.org/10.1002/pro.4792

Mills, E. C., & Shewry, P. R. (Eds.). (2004). *Plant food allergens*. Blackwell Science. https://doi.org/10.1002/9780470995174.

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research, 5*, 189. https://doi.org/10.12688/f1000research.7931.1

Olsson, M. H., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation, 7*(2), 525–537. https://doi.org/10.1021/ct100578z

Osborne, T. B. (1924). *The vegetable proteins*. Longmans, Green and Company.

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics, 52*(12), 7182–7190. https://doi.org/10.1063/1.328693

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*. *arXiv preprint* arXiv: 1912.01703. https://doi.org/10.48550/arXiv.1912.01703.

Pathak, R. K., Singh, D. B., Kuntal, A. K., Srivastava, S., & Kesharwani, R. K. (2021). Molecular dynamics simulation in drug discovery. In K. Roy (Ed.), *Vol. Chapter 10. Recent advances in computer aided drug designing* (pp. 228–247). Bentham Science Publishers.

Radhika, V., & Rao, V. S. H. (2015). Computational approaches for the classification of seed storage proteins. *Journal of Food Science and Technology, 52*, 4246–4255. https://doi.org/10.1007/s13197-014-1500-x

Robert, L. S., Nozzolillo, C., & Altosaar, I. (1985). Homology between rice glutelin and oat 12 S globulin. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, 829*(1), 19–26. https://doi.org/10.1016/0167-4838(85)90063-9

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, 12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Roy, T., Singh, A., Sari, T. P., & Homroy, S. (2023). Rice protein: Emerging insights of extraction, structural characteristics, functionality, and application in the food industry. *Journal of Food Composition and Analysis, 123*, Article 105581. https://doi.org/10.1016/j.jfca.2023.105581

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.

Shewry, P. R., Beaudoin, F., Jenkins, J., Griffiths-Jones, S., & Mills, E. N. C. (2002). Plant protein families and their relationships to food allergy. *Biochemical Society Transactions, 30*(6), 906–910. https://doi.org/10.1042/bst0300906

Shewry, P. R., & Casey, R. (1999). Seed proteins. In P. R. Shewry, & R. Casey (Eds.), *Seed proteins* (pp. 1–10). Springer Netherlands. https://doi.org/10.1007/978-94-011-4431-5_1.

Shewry, P. R., Napier, J. A., & Tatham, A. S. (1995). Seed storage proteins: Structures and biosynthesis. *The Plant Cell, 7*(7), 945–956. https://doi.org/10.1105/tpc.7.7.945

Shotwell, M. A., Boyer, S. K., Chesnut, R. S., & Larkins, B. A. (1990). Analysis of seed storage protein genes of oats. *Journal of Biological Chemistry, 265*(17), 9652–9658. https://doi.org/10.1016/S0021-9258(19)38719-8

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. arXiv preprint. https://doi.org/10.48550/arXiv.1312.6034.

Souza, P. C., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., … Marrink, S. J. (2021). Martini 3: A general purpose force field for coarse-grained molecular dynamics. *Nature Methods, 18*(4), 382–388. https://doi.org/10.1038/s41592-021-01117-4

Sullivan, G. M., & Feinn, R. (2012). Using effect size—Or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Taboada, P., Barbosa, S., Castro, E., Gutiérrez-Pichel, M., & Mosquera, V. (2007). Effect of solvation on the structure conformation of human serum albumin in aqueous–alcohol mixed solvents. *Chemical Physics, 340*(1–3), 59–68. https://doi.org/10.1016/j.chemphys.2007.07.027

Tan-Wilson, A. L., & Wilson, K. A. (2012). Mobilization of seed protein reserves. *Physiologia Plantarum, 145*(1), 140–153. https://doi.org/10.1111/j.1399-3054.2011.01535.x

Tatham, A. S., Field, J. M., Morris, V. J., I'Anson, K. J., Cardle, L., Dufton, M., & Shewry, P. R. (1993). Solution conformational analysis of the alpha-zein proteins of maize. *Journal of Biological Chemistry, 268*(35), 26253–26259. https://doi.org/10.1016/S0021-9258(19)74308-7

Tatham, A. S., & Shewry, P. R. (1995). The S-poor prolamins of wheat, barley and rye. *Journal of Cereal Science, 22*(1), 1–16. https://doi.org/10.1016/S0733-5210(05)80002-5

The UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research, 51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Thirumalai, D., Reddy, G., & Straub, J. E. (2012). Role of water in protein aggregation and amyloid polymorphism. *Accounts of Chemical Research, 45*(1), 83–92. https://doi.org/10.1021/ar2000869

Thomas, P. D., & Dill, K. A. (1993). Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. *Protein Science, 2*(12), 2050–2065. https://doi.org/10.1002/pro.5560021206

Tolmachev, D. A., Malkamäki, M., Linder, M. B., & Sammalkorpi, M. (2023). Spidroins under the influence of alcohol: Effect of ethanol on secondary structure and molecular level solvation of silk-like proteins. *Biomacromolecules, 24*(12), 5638–5653. https://doi.org/10.1021/acs.biomac.3c00637

Trevino, S. R., Scholtz, J. M., & Pace, C. N. (2008). Measuring and increasing protein solubility. *Journal of Pharmaceutical Sciences, 97*(10), 4155–4166. https://doi.org/10.1002/jps.21327

Van Oss, C. J. (1997). Hydrophobicity and hydrophilicity of biosurfaces. *Current Opinion in Colloid & Interface Science, 2*(5), 503–512. https://doi.org/10.1016/S1359-0294(97)80099-4

Wang, G., Wu, H., Wang, Y., Liu, X., Peng, S., Wang, W., & Hou, T. (2025). Discovery of novel Nav1.7-selective inhibitors with the 1H-indole-3-propionamide scaffold for effective pain relief. *Research, 8*, Article 0599. https://doi.org/10.34133/research.0599

Wilson, C. M., Shewry, P. R., Faulks, A. J., & Miflin, B. J. (1981). The extraction and separation of barley glutelins and their relationship to other endosperm proteins. *Journal of Experimental Botany, 32*(6), 1287–1293. https://doi.org/10.1093/jxb/32.6.1287

Xing, J., Li, Z., Zhang, W., & Wang, P. (2023). The composition, structure, and functionalities of prolamins from highland barley. *Molecules, 28*(14), 5334. https://doi.org/10.3390/molecules28145334

Zhou, M., Kreft, I., Woo, S. H., Chrungoo, N., & Wieslander, G. (2016). *Molecular breeding and nutritional aspects of buckwheat*. Academic Press.

Zhu, X., Lopes, P. E., & MacKerell, A. D., Jr. (2012). Recent developments and applications of the CHARMM force fields. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 2*(1), 167–185. https://doi.org/10.1002/wcms.74