

Plant protein behavior in food systems: Data-driven, molecular, and multiscale approaches

by

Hyukjin Kwon

B.S., University of California, Berkeley, 2019

M.S., Sejong University, 2022

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Grain and Food Science
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Abstract

Plant proteins play important roles in food systems, yet their functional behavior remains difficult to understand and predict because it is governed by interactions across multiple length and time scales. Their properties in food systems are not dictated by a single molecular feature or processing variable, but emerge from coupled effects across scales. These effects include contributions from amino acid sequence, protein structure, solvent environment, intermolecular assembly, and other system-specific factors.

Conventional experimental approaches are essential for characterizing food protein functionality, but they often provide indirect or ensemble-averaged measurements of systems that are structurally heterogeneous and source-specific. This challenge has become increasingly significant as food protein research moves from descriptive characterization toward more predictive and mechanism-oriented approaches. Data-driven modeling, bioinformatics, and structure-based computation offer new opportunities to identify patterns that are difficult to extract from experiments alone.

The work presented in this dissertation takes this opportunity as its starting point. Chapter 1 first established the methodological basis for a simulation-based approach by examining the principles, capabilities, and limitations of all-atom and coarse-grained molecular dynamics simulations in food protein research. This chapter emphasized that molecular simulation is most useful when its resolution, force-field assumptions, and sampling limitations are matched to the question being asked. Chapter 2 then extended this framework toward a data-driven prediction approach that integrated global sequence- and structure-level descriptors with residue-level graph representations. Although the model was trained on microbial protein datasets, its application to seed storage proteins suggested that solubility-related sequence and structural patterns could be partially transferred to food and agricultural proteins. In particular, the seed storage protein analysis indicated that predicted solubility was strongly associated with surface electrostatic properties.

Chapter 3 applied this broader computational perspective to Osborne protein classification, a classical framework in cereal and seed protein chemistry. By combining physicochemical descriptors, interpretable AI, molecular-level analysis, and simulations, this chapter examined whether empirical protein fractions could be understood at the monomeric scale. The results suggested that albumins and prolamins show relatively distinct monomer-level characteristics consistent with their solvent preferences, whereas the distinction between globulins and glutelins requires consideration of higher-order organization.

Chapter 4 provided a more focused experimental and computational case study using zein–lutein encapsulation. In this system, molecular dynamics simulations, dilute-solution measurements, suspension-level characterization, and powder thermal stability tests were used to examine how ethanol-dependent protein behavior translated into encapsulation performance. The

findings showed that although monomer-level simulations and dilute-solution measurements provided important insight into ethanol-dependent zein–lutein interactions, these lower-level observations did not directly translate to functional outcomes at higher structural scales, such as lutein retention in suspension or thermal stability in the powder. This finding suggests that molecular-level prediction alone may be insufficient for food-scale applications of plant proteins, particularly when functionality depends on higher-order assembly and processing kinetics.

Chapter 5 further extended the findings of Chapter 4 by applying the zein–lutein encapsulation system to corn gluten meal-derived zein fractions. The feasibility of using the ethanol-soluble CGM extract directly for powder preparation was evaluated, and the effects of solvent-reduction pathway on powder thermal stability were compared with those observed for commercial zein. The resulting powders were also incorporated into cookies to assess whether their protective functionality was retained during baking.

Beyond the specific findings of individual chapters, a recurring perspective of this dissertation is transferability: how information obtained from one scale, protein source, model system, or measurement can be extended to another context in food and agricultural protein research. Across the chapters discussed, data-driven approaches based on sequence- or structure-level information revealed useful patterns, but their transferability depended on protein source and structural organization. Similarly, molecular simulations provided useful mechanistic insight, but their relevance depended on whether the modeled scale corresponded to the scale of the functional context. Together, these chapters clarify that computational approaches can be valuable in food and agricultural protein research when they are applied to questions and contexts for which their predictions remain meaningful.

Overall, this dissertation shows that computational approaches can provide an additional layer of analysis in plant protein research, enabling complex protein behavior to be examined beyond what can be obtained from empirical characterization alone. The findings further demonstrate the potential to extend such approaches to agricultural coproducts and practical food systems.

Plant protein behavior in food systems: Data-driven, molecular, and multiscale approaches

by

Hyukjin Kwon

B.S., University of California, Berkeley, 2019

M.S., Sejong University, 2022

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Grain and Food Science
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Yonghui Li

Copyright

© Hyukjin Kwon 2026.

Abstract

Plant proteins play important roles in food systems, yet their functional behavior remains difficult to understand and predict because it is governed by interactions across multiple length and time scales. Their properties in food systems are not dictated by a single molecular feature or processing variable, but emerge from coupled effects across scales. These effects include contributions from amino acid sequence, protein structure, solvent environment, intermolecular assembly, and other system-specific factors.

Conventional experimental approaches are essential for characterizing food protein functionality, but they often provide indirect or ensemble-averaged measurements of systems that are structurally heterogeneous and source-specific. This challenge has become increasingly significant as food protein research moves from descriptive characterization toward more predictive and mechanism-oriented approaches. Data-driven modeling, bioinformatics, and structure-based computation offer new opportunities to identify patterns that are difficult to extract from experiments alone.

The work presented in this dissertation takes this opportunity as its starting point. Chapter 1 first established the methodological basis for a simulation-based approach by examining the principles, capabilities, and limitations of all-atom and coarse-grained molecular dynamics simulations in food protein research. This chapter emphasized that molecular simulation is most useful when its resolution, force-field assumptions, and sampling limitations are matched to the question being asked. Chapter 2 then extended this framework toward a data-driven prediction approach that integrated global sequence- and structure-level descriptors with residue-level graph representations. Although the model was trained on microbial protein datasets, its application to seed storage proteins suggested that solubility-related sequence and structural patterns could be partially transferred to food and agricultural proteins. In particular, the seed storage protein analysis indicated that predicted solubility was strongly associated with surface electrostatic properties.

Chapter 3 applied this broader computational perspective to Osborne protein classification, a classical framework in cereal and seed protein chemistry. By combining physicochemical descriptors, interpretable AI, molecular-level analysis, and simulations, this chapter examined whether empirical protein fractions could be understood at the monomeric scale. The results suggested that albumins and prolamins show relatively distinct monomer-level characteristics consistent with their solvent preferences, whereas the distinction between globulins and glutelins requires consideration of higher-order organization.

Chapter 4 provided a more focused experimental and computational case study using zein–lutein encapsulation. In this system, molecular dynamics simulations, dilute-solution measurements, suspension-level characterization, and powder thermal stability tests were used to examine how ethanol-dependent protein behavior translated into encapsulation performance. The

findings showed that although monomer-level simulations and dilute-solution measurements provided important insight into ethanol-dependent zein–lutein interactions, these lower-level observations did not directly translate to functional outcomes at higher structural scales, such as lutein retention in suspension or thermal stability in the powder. This finding suggests that molecular-level prediction alone may be insufficient for food-scale applications of plant proteins, particularly when functionality depends on higher-order assembly and processing kinetics.

Chapter 5 further extended the findings of Chapter 4 by applying the zein–lutein encapsulation system to corn gluten meal-derived zein fractions. The feasibility of using the ethanol-soluble CGM extract directly for powder preparation was evaluated, and the effects of solvent-reduction pathway on powder thermal stability were compared with those observed for commercial zein. The resulting powders were also incorporated into cookies to assess whether their protective functionality was retained during baking.

Beyond the specific findings of individual chapters, a recurring perspective of this dissertation is transferability: how information obtained from one scale, protein source, model system, or measurement can be extended to another context in food and agricultural protein research. Across the chapters discussed, data-driven approaches based on sequence- or structure-level information revealed useful patterns, but their transferability depended on protein source and structural organization. Similarly, molecular simulations provided useful mechanistic insight, but their relevance depended on whether the modeled scale corresponded to the scale of the functional context. Together, these chapters clarify that computational approaches can be valuable in food and agricultural protein research when they are applied to questions and contexts for which their predictions remain meaningful.

Overall, this dissertation shows that computational approaches can provide an additional layer of analysis in plant protein research, enabling complex protein behavior to be examined beyond what can be obtained from empirical characterization alone. The findings further demonstrate the potential to extend such approaches to agricultural coproducts and practical food systems.

Table of Contents

List of Figures	xiv
List of Tables	xxi
Acknowledgements	xxiv
Dedication	xxv
Chapter 1 - All-atom and coarse-grained simulation of food proteins: Principles, force fields, and applications	1
Abstract	1
1.1. Introduction.....	2
1.2. Molecular dynamics simulations: Principles, force fields, and practice.....	5
1.2.1. From forces to motion: The basic MD workflow	5
1.2.2. Controlling thermodynamic conditions: Ensembles, thermostats, and barostats.....	8
1.2.3. Atomistic force fields.....	11
1.2.3.1. Bonded energy terms for atomistic force fields: CHARMM36.....	12
1.2.3.2. Non-bonded energy terms for atomistic force fields: CHARMM36	15
1.2.4. Coarse-grained force fields	19
1.2.4.1. Coarse-grained force field: Martini 3	21
1.2.4.2 Potential energy terms for CG force fields: Martini 3	23
1.2.5. AA and CG-MD simulation: Time and length scale	25
1.2.6. Available tools for MD simulation	28
1.2.6.1. Simulation engines.....	28
1.2.6.2. System preparation and visualization	30
1.2.7. Extracting physical insights of proteins from MD simulations	31
1.2.7.1. Hydrogen bonding	31
1.2.7.2. Secondary structure analysis.....	32
1.2.7.3. Solvent-accessible surface area (SASA).....	33
1.2.7.4. Contact frequency analysis	34
1.3. Application scenarios.....	35
1.3.1. Solvent- and co-solute-induced protein responses.....	35
1.3.2. Process-induced structural evolution	38

1.3.3. Bioactive ligand interactions and redistribution	40
1.4. Practical limitations and methodological caveats in applying MD	42
1.4.1. Force-field-consistent simulation protocols.....	42
1.4.2. Availability and reliability of structural information.....	43
1.4.3. Stochastic variability and sampling convergence	44
1.4.4. Limitations of Martini coarse-grained models and strategies for refinement.....	45
1.5. Conclusion	46
Acknowledgement	47
References.....	47
Chapter 2 - AlphaFold 2-based stacking model for protein solubility prediction and its	
transferability to seed storage proteins	67
Abstract.....	67
2.1. Introduction.....	69
2.2. Materials and methods	74
2.2.1. Protein solubility definition	74
2.2.2. Datasets	74
2.2.2.1. <i>E. coli</i> dataset	74
2.2.2.2. <i>S. cerevisiae</i> dataset	75
2.2.2.3. Seed storage protein dataset.....	75
2.2.3. Feature extraction for MLP model.....	76
2.2.4. Feature extraction for GCN model	78
2.2.4.1. Node features for GCN model	78
2.2.4.2. Protein contact map (edge for GCN model)	79
2.2.5. MLP model structure	79
2.2.6. GCN model structure	82
2.2.6.1. GCN layers.....	82
2.2.6.2. Pooling & output layer.....	83
2.2.7. Benchmark models: MLP and GCN with ESM-2 embedding.....	84
2.2.8. Stacking model.....	85
2.2.9. Model hyperparameters and evaluation metrics	87
2.2.10. Identification of influential features/residues	87

2.2.11. Case study: molecular dynamics simulation	88
2.3. Results and discussion	89
2.3.1. Comparison of extracted features among datasets	89
2.3.2. Linear correlation analysis between structure/sequence features with solubility	90
2.3.3. Solubility prediction using MLP model.....	91
2.3.4. Solubility prediction using GCN model	92
2.3.5. Performance of stacking model.....	93
2.3.6. Model interpretation.....	96
2.3.7. Comparison with other methods using external <i>S. cerevisiae</i> test set	98
2.3.8. Transferability of the stacking model to seed storage proteins.....	101
2.3.9. Case study on seed storage proteins: identification of important residues	104
2.3.10. Case study on seed storage proteins: CG-MD simulation	105
2.4. Conclusion	109
Acknowledgements.....	110
References.....	110
Chapter 3 - Unraveling the physicochemical differences among Osborne protein classes via	
bioinformatics and AI	121
Abstract	121
3.1. Introduction.....	122
3.2. Materials and methods	127
3.2.1. Dataset.....	127
3.2.1.1. Data acquisition	127
3.2.1.2. Data structure	128
3.2.2. Statistical testing	130
3.2.3. Multivariate analysis.....	130
3.2.4. Machine learning classification of proteins	131
3.2.4.1. Global (sequence + structural) feature extraction.....	131
3.2.4.2. Residue-level feature extraction	132
3.2.4.3. GCN architecture	132
3.2.5. Molecular dynamics simulation.....	134
3.2.5.1. All-atom MD simulation.....	135

3.2.5.2. Coarse-grained MD simulation.....	136
3.3. Results and discussion	138
3.3.1. Protein characterization.....	138
3.3.1.1. Comparative analysis of 38 physicochemical properties.....	138
3.3.1.2. Multivariate analysis of physicochemical features	142
3.3.2. Machine learning classification and interpretation of the model.....	143
3.3.2.1. Classification using global features: SHAP analysis & interpretability	143
3.3.2.2. Residue-level classification via GCN: Saliency mapping & interpretability	147
3.3.2.3. Globulin vs. Glutelin: Distinct structural & residue-level differences	149
3.3.3. All-atom molecular dynamics simulation	154
3.3.3.1. Simulation of 2S soy albumin & 19 kDa maize zein in water & 70% EtOH....	154
3.3.3.2. Simulation of 11S pea globulin & rice glutelin in water & 1M NaCl.....	160
3.3.4. Coarse-grained molecular dynamics simulation	164
3.3.4.1. Simulation of 2S soy albumin & 19 kDa maize zein in water & 70% EtOH....	164
3.3.4.2. Simulation of 11S pea legumin A & rice glutelin A1 in water & 1M NaCl.....	166
3.4. Conclusion	170
Acknowledgement	172
References	172
Chapter 4 - Ethanol-dependent multiscale transitions in zein–lutein encapsulation: From	
monomer collapse to path-dependent thermal stability	186
Abstract	186
4.1. Introduction.....	187
4.2. Materials and methods	190
4.2.1. Materials	190
4.2.2. Pretreatment of zein	191
4.2.3. Preparation of dilute zein–lutein solutions	192
4.2.4. UV–vis absorption spectra	193
4.2.5. Particle size measurement.....	193
4.2.6. Preparation of zein-lutein suspension	193
4.2.7. Measurement of solvent extraction resistance and turbidity.....	194
4.2.8. Confocal laser scanning microscopy.....	195

4.2.9. Molecular dynamics simulation	195
4.2.9.1. All-atom molecular dynamics simulation	196
4.2.9.2. Coarse-grained molecular dynamics simulation	198
4.2.10. Preparation of zein-lutein powder	200
4.2.11. Measurement of thermal stability	201
4.2.12. Scanning electron microscopic analysis	201
4.2.13. Statistical analysis	201
4.3. Results and discussion	202
4.3.1. Lower-scale structural transition of zein-lutein across ethanol concentration.....	202
4.3.1.1. All-atom molecular dynamics simulation of α -zein	202
4.3.1.2. All-atom molecular dynamics simulation of ethanol-dependent association between α -zein and lutein	203
4.3.1.3. UV–Vis spectral transitions in dilute zein–lutein systems	207
4.3.1.4. Particle size analysis of low-ethanol ranges in dilute zein–lutein solution	210
4.3.2. Collective assembly behavior of zein–lutein systems across ethanol concentrations	211
4.3.2.1. Lutein retention under solvent extraction	212
4.3.2.2. Ethanol-dependent particle size distribution in zein–lutein suspensions.....	213
4.3.2.3. Coarse-grained molecular dynamics simulation.....	217
4.3.3. Powder-level thermal stability of zein-lutein system.....	220
4.3.3.1. Thermal stability of zein–lutein powders prepared under different ethanol conditions.....	220
4.3.3.2. Powder morphology observed by SEM	225
4.3.3.3. Possible explanation for differences in powder thermal stability	227
4.4. Conclusions.....	230
Acknowledgement	231
References	232
Chapter 5 - Feasibility of direct lutein encapsulation using corn gluten meal ethanol extract for application in cookies	240
Abstract.....	240
5.1. Introduction.....	241

5.2. Materials and methods	243
5.2.1. Materials	243
5.2.2. Zein fractionation.....	244
5.2.3. Proximate analysis	245
5.2.4. SDS-PAGE	246
5.2.5. Solubility assay	246
5.2.6. Molecular dynamics simulation.....	247
5.2.7. Powder formulation	247
5.2.8. Powder thermal stability	248
5.2.9. Cookie formulation	249
5.2.10. Apparent lutein retention in cookie.....	249
5.2.11. Statistical analysis	250
5.3. Results and discussion	251
5.3.1. Proximate analysis and yield.....	251
5.3.2. SDS-PAGE profiles of CGM-derived fractions	253
5.3.3. Protein solubility in 70% EtOH	254
5.3.4. Molecular dynamics simulation.....	256
5.3.5. Powder thermal stability	257
5.3.6. Lutein retention before and after baking.....	262
5.4. Conclusion	264
References.....	267
Chapter 6 - Conclusion and future perspectives	269
6.1. Conclusion	269
6.2. Future perspectives	270
Appendix A - Supplementary data.....	272
Supplementary data for Chapter 1.	273
Supplementary data for Chapter 2.	277
Supplementary data for Chapter 3.	287
Supplementary data for Chapter 4.	300

List of Figures

Figure 1.1. Schematic workflow of a molecular dynamics simulation. The force calculation, integration scheme, and particle update follow the formulations in Equations (1), (3), and (4), respectively. Note: System preparation steps such as energy minimization and equilibration are omitted for clarity.	8
Figure 1.2. Schematic representation of the primary bonded interaction terms in the CHARMM36 force field: bond stretching, angle bending, dihedral and improper torsions, corresponding to Equations (5) – (8), respectively.	13
Figure 1.3. TIP3P _{mod} water model and its defined parameters in CHARMM36 (a); Coulombic interaction potentials for H–H and O–H pairs ($\epsilon_r = 97$) (b). Lennard–Jones potential for an O–H pair (c). Comparison of Coulombic and Lennard–Jones interaction ranges under a nonbonded cutoff of 1.20 nm (d).	17
Figure 1.4. Atomistic and coarse-grained representations of β -lactoglobulin monomer (PDB ID: 1BEB) (a). The protein structure is shown as (left to right) an all-atom representation, a cartoon representation of the same structure, a bead-mapped coarse-grained model overlaid on the cartoon backbone, and the coarse-grained bead representation including elastic networks. Visualization was performed using VMD (Humphrey et al., 1996) and Martiniglass (Brasnett and Marrink, 2025). Representation of a cysteine residue and its bond stretching parameters in CHARMM36 and Martini 3 force fields (b). The notation follows Equations (5) and (11).	23
Figure 1.5. Approximate length and time scale comparison of molecular simulations and experimental techniques. The scales shown are approximate and are intended to reflect practical use cases rather than strict boundaries.	27
Figure 2.1. The structure of the MLP (a) and GCN (b) model constructed.	81
Figure 2.2. Overall workflow for the construction and validation of the stacking solubility model.	86
Figure 2.3. The scatter plots of predicted and experimental solubility (a) and receiver operating characteristic (ROC) curve (b) for each model on <i>E. coli</i> test dataset.	95
Figure 2.4. The relative importance of features determined from saliency maps of the trained MLP (a) and GCN (b). The blosum62 mappings for charged residues are bolded in (b). ...	97

Figure 2.5. The scatter plots of predicted and experimental solubility on <i>S. cerevisiae</i> validation dataset.	101
Figure 2.6. Stacking model (SVR) predictions for the solubility of seed storage proteins by their Osborne classifications.	103
Figure 2.7. Selected regions with high saliency and electrostatic potential maps of soybean 2S albumin and maize 22 kDa prolamin.	105
Figure 2.8. Snapshots of CGMD simulations for soybean 2S albumin and maize 22 kDa prolamin at different intervals (a) and initial aggregation stage (b).	108
Figure 3.1. Graphical representation of protein structure embedding and architecture of the GCN model.	134
Figure 3.2. 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).	142
Figure 3.3. 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.	145
Figure 3.4. 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).	149
Figure 3.5. 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).	151
Figure 3.6. 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).	153
Figure 3.7. 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).	156

Figure 3.8. Average Lennard-Jones potential energy between protein and system in water and 70% ethanol (left) and intra-protein hydrogen bond counts (right).....	157
Figure 3.9. 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.	159
Figure 3.10. Conformational changes of pea legumin A and glutelin A1 in water and 1M 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. The average Coulombic potential between protein and system, salt bridge count, and hydrogen bonding count at barrel rolls in water and 1M NaCl (d).	162
Figure 3.11. Saliency-highlighted regions of pea legumin A and rice glutelin A1 with corresponding root-mean-square fluctuation (RMSF) profiles. The highlighted saliency region is shown in magenta, and disulfide bonds are shown as orange sticks.....	163
Figure 3.12. 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). ...	165
Figure 3.13. 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).	167
Figure 4.1. Structural evolution of zein–lutein complexes across ethanol concentrations from all-atom molecular dynamics simulations (a). Each system was initialized from the final configuration of the preceding ethanol condition to reflect the gradual solvent transition encountered during antisolvent processing. Endpoint comparison of α -zein–lutein conformations at 50% and 40% EtOH (b). Red arrows indicate the top 20 largest C α displacement vectors between the two endpoint conformations. Nonlocal residue–residue intra-protein contacts calculated over the final 50 ns of all-atom MD simulations at different EtOH concentrations (c). Contacts between residues within six sequence positions were excluded. * The 70% EtOH condition (panel c) is based on a single simulation trajectory, whereas all other conditions are shown for two independent replicas.....	205

Figure 4.2. UV–Vis absorbance spectra of dilute lutein, zein, and zein–lutein systems at 70–35% EtOH. Left panels show raw spectra of lutein-only, zein-only, zein–lutein, and EtOH blank samples, while right panels show blank-subtracted lutein spectra and zein-subtracted zein–lutein spectra. (a) 70%, (b) 60%, (c) 50%, (d) 40%, and (e) 35% EtOH.....	209
Figure 4.3. Hexane extraction resistance (left) and turbidity (right) of zein–lutein suspension across ethanol concentrations (a). CLSM images of the suspensions across ethanol concentrations (b).....	213
Figure 4.4. Particle size distributions of zein–lutein suspensions and control systems at different EtOH concentrations. Distributions are shown for zein–lutein (a), zein-only (b), and lutein-only (c) systems.	215
Figure 4.5. Structural evolution of zein–lutein assemblies across ethanol concentrations from coarse-grained molecular dynamics simulations (a). Each system was initialized from the final configuration of the preceding ethanol condition to reflect the gradual solvent transition encountered during antisolvent processing. Individual zein monomers are shown in distinct colors. Relative burial of lutein averaged over the final 500 ns of simulation, defined as $100 \times [1 - (\text{exposed solvent-accessible surface area (SASA) of lutein} / \text{SASA of lutein in the absence of zein})]$ (b). Values were not calculated for ethanol concentrations $\geq 50\%$ due to the absence of stable complexes.	218
Figure 4.6. Thermal stability of zein–lutein powders under different processing histories: physical mixture (a), fast antisolvent addition from 70% ethanol (b), and stepwise ethanol reduction to 20% with equilibration followed by rapid dilution (c). All values are expressed as μg of lutein per 15 mg of zein–lutein powder.	223
Figure 4.7. SEM morphology of zein–lutein powders prepared under different EtOH processing histories and the physical mixture (PM). Images are shown at 100 $\times$ magnification (a) and 1000 $\times$ magnification (b).	226
Figure 4.8. Configurations of zein–lutein complex and spatial distribution of lutein after water quenching under different processing conditions, predicted from all-atom (a) and coarse-grained molecular dynamics simulations (b).	229
Figure 5.1. Schematic representation of zein extraction and sequential fractionation from corn gluten meal.....	245

Figure 5.2. Visual appearance of CGM, CGME, IP, BCP, ACP, and CZ powders. CGM, corn gluten meal; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction; and CZ, commercial zein ..	253
Figure 5.3. SDS-PAGE profiles of commercial zein and CGM-derived fractions. Lanes: M, molecular-weight marker; CZ, commercial zein; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; and ACP, after-cold-precipitation fraction.	254
Figure 5.4. Apparent soluble protein concentrations of commercial zein and CGM-derived fractions in 70% EtOH. CZ, commercial zein; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction	255
Figure 5.5. Ethanol-dependent association of zein with lutein and linoleic acid. (a) Representative configurations obtained after 200 ns of simulation at each ethanol concentration. Zein is shown in cyan, lutein in orange, and linoleic acid in green. (b) Average numbers of zein–lutein and zein–linoleic acid contacts calculated over the final 50 ns of each 200 ns simulation.	257
Figure 5.6. Thermal stability and sequential extractability of lutein in encapsulated powders prepared by fast and stepwise quenching during heating at 60°C: (a) CZ, (b) CGME, (c) CGMD, (d) ACP, and (e) BCP. The left and right panels represent powders prepared by fast and stepwise quenching, respectively. Values are presented as means ± standard deviations. Different lowercase and uppercase letters within each panel indicate significant differences across heating times in the hexane-extractable and ethanol-released fractions, respectively. The amount of lutein is expressed as µg per 100 mg of powder. Retention was calculated by dividing the total apparent lutein content at 72 h by that at 0 h × 100. CZ, commercial zein; CGME, corn gluten meal extract; CGMD, corn gluten meal direct powder; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction	262
Figure 5.7. Visual appearance and apparent lutein retention of cookie formulations before and after baking. Representative images of (a) unbaked doughs and (b) baked cookies, and (c) apparent lutein retention after baking. PM, physical mixture; ACP, after-cold-precipitation fraction; CGME, corn gluten meal extract; CGMD, corn gluten meal direct powder; and CZ, commercial zein.	264

Appendix Figure A.1. Mean square displacement of the selected proteins over the simulation time	286
Appendix Figure A.2. Top 10 relative contributions of input features to class separation in linear discriminant analysis (LDA).....	291
Appendix Figure A.3. Gravimetric monitoring of solvent evaporation during zein–lutein suspension preparation. Total mass change over time during rotary evaporation (left). Linear relationship between mass loss and time within the 70–20% ethanol range (right).	300
Appendix Figure A.4. Hexane-extractable lutein in ethanol solutions without zein across different ethanol concentrations.....	301
Appendix Figure A.5. Time evolution of zein–lutein contacts in coarse-grained simulations at 50% (a) and 40% ethanol (b).	302
Appendix Figure A.6. Comparison of zein backbone conformations between all-atom and coarse-grained simulations with varying elastic network (EN) strength. The helical region where pronounced expansion was observed (residues 100–245) is colored in blue.....	303
Appendix Figure A.7. AlphaFold 3-predicted α -zein monomer structure colored by pLDDT score (left) and the corresponding PAE matrix (right).....	304
Appendix Figure A.8. Conformational behavior of zein in 70% ethanol from all-atom molecular dynamics simulations. RMSD of zein over the 1 μ s simulation (a). Representative structure obtained from RMSD-based clustering of equilibrated frames (500–1000 ns) (b).	305
Appendix Figure A.9. Average radius of gyration (a) and protein–lutein contacts (b) during the last 50 ns of all-atom MD simulations at different ethanol concentrations, shown separately for each replica.....	306
Appendix Figure A.10. Final conformations of the α -zein–lutein system at 50% and 40% EtOH from the original simulation and after the 200 ns extension. (a) Replica 1; (b) replica 2. .	307
Appendix Figure A.11. Residue–residue intra-protein contact distributions from the final 50 ns and subsequent 200 ns extension of all-atom MD simulations at 50% and 40% EtOH. (a) Replica 1; (b) replica 2.....	308
Appendix Figure A.12. Residue-level contact frequency of zein–lutein interactions at different ethanol concentrations of 60% (a), 40% (b), and 20% (c). Contact frequencies were	

calculated over the final 50 ns of each trajectory. Results from independent replicas are shown.	309
Appendix Figure A.13. Particle size distributions of dilute solution samples of zein–lutein, zein-only, and lutein-only systems at 40% and 35% EtOH.	311

List of Tables

Table 1.1. Comparison between all-atom and coarse-grained molecular dynamics simulations.	28
Table 2.1. Prediction performances of the MLP and benchmark models (cross-validation / testing). For binary classification evaluation, regression outputs were converted to class labels using a threshold of 0.5. Abbreviations are: Linear regression (LR), stochastic gradient descent (SGD), support vector regression (SVR), random forest (RF), gradient boosting (GB), multilayer perceptron (MLP), area under curve (AUC), and Matthews correlation coefficient (MCC).....	92
Table 2.2. Prediction performances of stacking models with different meta-regressors and benchmark/base models (cross-validation / testing).	94
Table 2.3. Prediction performance of the stacking models and existing prediction models on <i>S. cerevisiae</i> dataset.	100
Table 2.4. Predictions of existing models on seed storage protein dataset.	104
Table 3.1. Summary of collected seed storage protein data (215 species and 1039 proteins) organized by Osborne classification.	129
Table 3.2. Key physicochemical properties distinguishing Osborne protein classes (a), including features with notable differences between globulin and glutelin (b). Values are expressed as mean $\pm$ standard deviation and median with interquartile range. Medians with different superscript letters indicate statistically significant differences ($p < 0.05$).	140
Table 4.1. EtOH-dependent particle size and polydispersity of zein–lutein, zein-only, and lutein-only suspensions at different ethanol fractions. Values are presented as mean $\pm$ SD. Different lowercase letters within the same column indicate significant differences among ethanol fractions for each sample type ($p < 0.05$). Asterisks indicate no significant difference. Statistical analysis of Z-average values was performed after log transformation because particle-size values differed by orders of magnitude across ethanol fractions. PDI, polydispersity index; ND, not determined.	216
Table 4.2. Thermal stability and final lutein retention of zein–lutein powders prepared under different ethanol-dependent processing routes. Thermal stability was evaluated at 60 °C for 72 h. PM: physical mixture; FQ: fast quenching from 70% to 10% ethanol by rapid water addition; GQ: gradual quenching with equilibration at the indicated ethanol concentrations	

prior to final quenching to 10% ethanol (e.g., GQ 50% indicates equilibration at 50% ethanol for 10 min before final quenching). Values are presented as mean $\pm$ standard deviation. The upper and lower values represent hexane-extractable and ethanol-released lutein fractions, respectively. Different lowercase (hexane-extractable fraction) and uppercase (ethanol-released fraction) letters indicate significant differences over time within each sample ($p < 0.05$). All values are expressed as μg of lutein per 15 mg of zein–lutein powder. The final lutein retention column indicates differences among processing routes ($p < 0.05$). Final retention (%) was calculated as the amount of lutein at 72 h relative to 0 h.

..... 222

Table 5.1. Proximate composition and recovery yields of commercial zein, corn gluten meal, and corn gluten meal-derived fractions. Different lowercase letters within the same column indicate significant differences among samples ($p < 0.05$). Yield was calculated relative to the initial total dry mass of CGM. CZ, commercial zein; CGM, corn gluten meal; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction. 251

Table 5.2. Changes in apparent lutein content and final retention of physical mixtures during heating at 60°C. Within each cell, the upper and lower values represent the hexane-extractable and ethanol-released lutein contents, respectively. Different lowercase and uppercase letters within the same row indicate significant differences across heating times in the hexane-extractable and ethanol-released fractions, respectively. Different letters in the retention column indicate significant differences among physical mixtures. Retention (%) was calculated as the total apparent lutein content at 72 h divided by that at 0 h $\times$ 100. The amount of lutein is expressed as μg per 100 mg of powder. CZ, commercial zein; CGME, corn gluten meal extract; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction 260

Appendix Table A.1. Representative applications of all-atom and coarse-grained molecular dynamics simulations in food protein systems. 273

Appendix Table A.2. Source organisms of seed storage dataset and their numbers by each Osborne class. 277

Appendix Table A.3. Effect of feature selection methods on the performance of MLP and benchmark models.	278
Appendix Table A.4. Selected hyperparameters for the MLP (a), GCN (b), and stacking (SVR) meta-regressor (c).....	279
Appendix Table A.5. Comparison of extracted MLP features across the datasets used.	280
Appendix Table A.6. Comparison of amino acids composition across the datasets used.	282
Appendix Table A.7. Comparison of selected seed proteins to reference human proteins. Means with different letters indicate significant differences within each row ($p < 0.05$).	283
Appendix Table A.8. Linear correlation coefficient of MLP features with experimental solubility	285
Appendix Table A.9. Comparison of 38 extracted physicochemical features across Osborne classes. Values are presented as mean $\pm$ std and median, interquartile range.	287
Appendix Table A.10. Performances of the models trained (training/cross-validation/testing).	292
Appendix Table A.11. Generalizability performances of the models tested on globulin/glutelin set and respective top 5 SHAP features selected using different train_test split seeds. The first column indicates the random seeds tested. All hyperparameters were selected via cross-validation at random seed = 98. The hyperparameters were fixed for the other 5 train_test split seeds. SHAP was not applied to GCN model. Model abbreviations: SVM (support vector machine), DT (decision tree), RF (random forest), KNN (K-nearest neighbor), MLP (multilayer perceptron).	293
Appendix Table A.12. Average generalizability performances of the models tested on globulin/glutelin set across 6 different train_test split random seeds.	299

Acknowledgements

I would first like to express my sincere gratitude to my major advisor, Dr. Yonghui Li, for his guidance, support, and encouragement throughout my doctoral studies. His mentorship provided me with valuable opportunities to develop as an independent researcher and to address research questions through multidisciplinary approaches. I am especially grateful for the trust and flexibility he gave me in pursuing my research interests, as well as for his thoughtful advice throughout the development of this dissertation.

I would also like to thank my committee members, Dr. Jeffery Comer, Dr. Yong-Cheng Shi, and Dr. Mike Tilley, for their time, expertise, and constructive feedback. Their diverse perspectives greatly strengthened both this dissertation and my development as a researcher. I am particularly thankful for their insightful questions, suggestions, and encouragement throughout my doctoral program.

I am grateful to all current and former members of the Li Laboratory for their assistance, companionship, and support. Their help with laboratory work, discussions, and critical feedback made my doctoral experience both more productive and more enjoyable. I sincerely appreciate the collaborative and supportive environment they helped create.

Finally, I would like to thank everyone who supported me, directly or indirectly, throughout this journey. This dissertation would not have been possible without their encouragement and generosity.

Dedication

To my parents and my brother, for their unconditional love and support.

Chapter 1 - All-atom and coarse-grained simulation of food proteins: Principles, force fields, and applications

Abstract

Despite the growing application of MD simulations in food protein analysis, they are often applied with limited consideration of underlying mechanics, force-field assumptions, and resolution constraints. Among these, resolution constraint is particularly critical in food protein modeling, because protein functionality arises from coupled phenomena across multiple length scales. In this context, all-atom (AA) and coarse-grained (CG)-MD simulations can serve complementary roles in probing food protein behavior across scales. This review summarizes key principles of AA and CG-MD simulations, with a focus on their force field architecture and application in food protein research. Methodological considerations, resolution-dependent limitations, and selected analytical approaches relevant to food protein simulations are discussed to support appropriate model selection and interpretation. Overall, this review clarifies how resolution-aware hybrid MD simulations can improve interpretation of protein structure–function relationships in food systems.

Keywords: Multiscale modeling, molecular dynamics methodology, protein aggregation and self-assembly, structure–function relationships

1.1. Introduction

A comprehensive understanding of food proteins necessitates not only conventional experimental tools but also the ability to resolve structural and functional details at the molecular level. Instruments such as rheometers and texture analyzers have long been used to probe bulk properties of food proteins. However, unraveling how the structure of individual protein molecules changes and contributes to the macroscopic properties, such as storage modulus, remains challenging with conventional experimental measures. The inherent complexity of food matrices, often comprising heterogeneous mixtures of proteins, polysaccharides, and other components, further complicates interpretation. Fourier-transform infrared (FTIR) spectroscopy, for instance, estimates protein secondary structure by deconvoluting the amide I band ($1600 - 1700 \text{ cm}^{-1}$), but is frequently hampered by overlapping water absorption (Barth, 2007) and sample impurities. More broadly, such experimental techniques provide ensemble information that is often only semi-quantitative, particularly for samples where extensive protein purification is not feasible. These limitations hinder the precise elucidation of structure–function relationships in food proteins.

As food science increasingly shifts toward mechanistic understanding and quantitative modeling, there is a growing demand for tools that can selectively analyze target molecules from complex systems, and provide molecular-level resolution. Molecular dynamics (MD) simulation, a computational method that generates time-resolved trajectories of target molecules, fulfills both criteria. By revealing how protein structures evolve over time, MD simulations offer powerful insights into various protein complexes and supra-molecular structures, such as protein aggregation, protein–ligand complex formation, and peptide assembly (Strodel, 2021; Kumar et al., 2022; Williams-Noonan et al., 2023). Reflecting this versatility, MD simulations have increasingly been adopted in food protein research.

Classical all-atom (AA) simulations enable detailed tracking of individual atoms over time, providing access to dynamic trajectories, energy landscapes, and the conformational flexibility of proteins. These features make AA simulations particularly relevant for food scientists aiming to understand the behavior of food proteins at a mechanistic level. The rising interest in this approach is also reflected in recent review papers that highlight the expanding role of AA simulations in food science (Yu et al., 2024; Cen et al., 2026).

While AA simulations provide atomic-level insights into protein dynamics, they require continuous computation involving all inter-atomic interactions, resulting in significant computational costs. This restricts the feasible timescales and system sizes, unless high-performance computing resources are available. For instance, simulating a large protein such as myosin (~1900 amino acids) poses challenges due to its elongated tail domain, which demands an extensive simulation box and long simulation time to capture meaningful changes in conformation (Zhang et al., 2020). Moreover, understanding mesoscale phenomena in food proteins, such as colloidal behavior, aggregation kinetics, or phase separation, often mandates models that extend beyond quaternary structures and operate on greater temporal and spatial scales.

Coarse-grained (CG) molecular dynamics simulations address these limitations by representing groups of atoms as single interaction sites (bead). This reduction in degrees of freedom significantly lowers computational demands, enabling the simulation of larger and more complex systems within practical time frames. Moreover, the simplified potential energy surface in CG models allows for the use of longer integration timesteps (typically 5–20 fs, compared to 1–2 fs in AA simulations), thereby extending the accessible simulation timescales from hundreds of microseconds to millisecond range and beyond. These capabilities make CG-MD particularly suitable for bridging molecular-level understanding with macroscopic behaviors observed in food

systems. Despite its potential, CG-MD remains underutilized in food protein research, with little to no dedicated review literature available on its principles and applications.

Against this background, this work is intended to support food scientists seeking to apply hybrid MD simulations in their work to investigate structure–function relationships at the protein level. As mentioned, while AA simulations are increasingly applied in food science, CG models remain underutilized despite their clear advantages for modeling mesoscale phenomena. This gap is addressed by a dedicated discussion on CG-MD principles and applications in food protein systems. Furthermore, existing reviews often focus primarily on software tutorials or application case studies, with limited attention given to the physical principles underlying MD simulations and the structural composition of force fields. We aim to clarify these theoretical foundations through accessible explanations, making them approachable for researchers without prior expertise in computational modeling.

The paper is organized into three main sections. The first section introduces the fundamental principles of MD simulations, with particular emphasis on the distinctions between AA and CG approaches and their underlying physical assumptions. The second section discusses the types of molecular descriptors that can be extracted from MD simulations of proteins, with an emphasis on interpreting simulation outputs in the context of food protein structure–function relationships. The last section focuses on the application of both AA and CG simulations in food science research, concluding with a discussion on the practical limitations, assumptions, and common pitfalls of MD-based approaches in the field. More advanced techniques, including enhanced sampling and free energy calculations, are covered in dedicated supplementary material.

1.2. Molecular dynamics simulations: Principles, force fields, and practice

1.2.1. From forces to motion: The basic MD workflow

A classical MD simulation is a computational technique based on Newtonian mechanics, a classical physics framework that describes how particles move in response to applied forces. In this context, each atom (or coarse-grained bead) in a system is treated as a particle with mass, and the forces acting on each particle are iteratively calculated. Given a defined molecular system and environmental conditions (e.g., a solvated egg ovalbumin structure in a periodic box at 300 K and 1 atm), the simulation begins by loading atomic coordinates and initializing velocities according to the target temperature and random seed. It should be noted that, during energy minimization, a standard preparatory step before time-resolved simulation, velocities are not yet assigned, as the system is first brought to a nearby local energy minimum at 0 K (no thermal motion). This minimization removes unnatural conditions introduced during system construction (e.g., atoms placed closer than their allowed van der Waals radius) and reduces the initial forces, helping ensure numerical stability when the simulation is started. Once the system velocity is initialized, forces acting on each atom are computed from the negative gradient of the system's potential energy, as shown in Equation (1). These gradients are typically taken analytically (i.e. an exact mathematical expression) and implemented directly in the computer code, rather than being computed numerically.

$$\mathbf{F}_i = -\nabla_i E_{potential}(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (1)$$

Here, $\mathbf{F}_i$ is the three-dimensional force vector acting on particle i , ∇_i denotes the gradient with respect to the position of particle i ($\mathbf{x}_i$), $E_{potential}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is the potential energy of the system as a function of all atomic coordinates, and $\mathbf{x}_i$ is the position vector of particle i .

The motion of the atoms is then simulated by integrating Newton’s second law, as shown in Equation (2).

$$\mathbf{F}_i = m_i \frac{d^2 \mathbf{x}_i}{dt^2} \quad (2)$$

where m_i is the mass of particle i , $\mathbf{x}_i$ is the position vector of particle i , and t is time.

To compute the position and velocity of each atom over time, the differential Equation (2) must be solved. However, for systems with three or more interacting particles, there is no general analytical solution to describe their trajectories. As a result, MD simulations rely on numerical integration methods, where time is divided into small steps (Δt) and atomic position and velocity are updated accordingly.

Many numerical integration methods can be used to solve Equation (2); however, only a few methods are typically used in MD simulations. Perhaps surprisingly, MD simulations do not typically employ integration methods that yield the highest-accuracy particle trajectories, such as the high-order Runge–Kutta–Nyström methods used for tracking satellites in orbit (Papanikolaou and Tsoulis, 2016). Instead, for molecular systems, we are usually interested in statistical and thermodynamic properties for systems of particular conditions (e.g., concentration), rather than the trajectories of individual atoms. Exact atomic trajectories are mostly irrelevant since these depend delicately on the initial conditions and environmental influences, which cannot be precisely controlled or measured experimentally.

For correct reproduction of statistical properties, it is essential that the simulations can reach thermodynamic equilibrium, which requires an integrator that is time reversible and accurately conserves energy and position–momentum phase space volume (“symplectic”) (Gray et al, 1994; Kamberaj, 2020). Time reversibility means that if a simulation is run forward and then backward (inverting the timestep to $-\Delta t$ or, equivalently, inverting all velocities $\mathbf{v} \rightarrow -\mathbf{v}$), the

system must in principle return to its original state. Conservation of phase-space volume indicates that the time evolution of the coordinate–momentum space (possible combinations of coordinates and momentum) preserves its geometric volume, meaning that the set of accessible microstates does not compress or expand over time.

The most widely used symplectic integration methods are variants of the Verlet algorithm (Verlet, 1967) particularly the leapfrog method (default in GROMACS) (Abraham et al., 2026), and velocity Verlet (default in NAMD) (Phillips et al, 2005). GROMACS and NAMD are among the most widely used MD engines and will be discussed in Section 2. The update rules for the velocity Verlet algorithm are given by:

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{F}(t)}{2m}\Delta t^2 \quad (3)$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{1}{2m}[\mathbf{F}(t) + \mathbf{F}(t + \Delta t)]\Delta t \quad (4)$$

where $\mathbf{x}$ is the position; t is the time; $\mathbf{v}$ is the velocity; Δt is the timestep; $\mathbf{F}$ is the force; m is the mass. Note that $\mathbf{F}(t+\Delta t)$ in Equation (4) is calculated using the new coordinates $\mathbf{x}(t+\Delta t)$ given in Equation (3). The iteratively updated positions and velocities are written to a file at intervals (e.g., every 5000 time steps), generating a trajectory that describes the time evolution of the molecular system (Fig. 1.1).

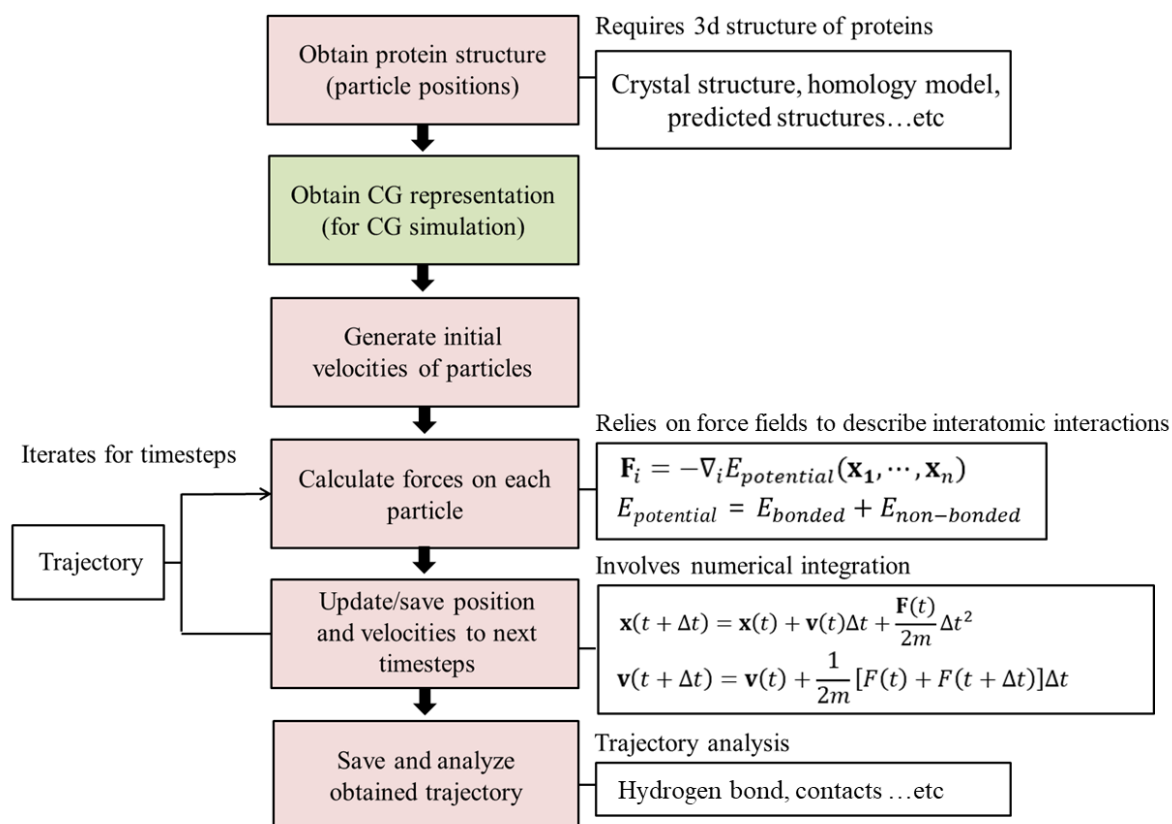


Figure 1.1. Schematic workflow of a molecular dynamics simulation. The force calculation, integration scheme, and particle update follow the formulations in Equations (1), (3), and (4), respectively. Note: System preparation steps such as energy minimization and equilibration are omitted for clarity.

1.2.2. Controlling thermodynamic conditions: Ensembles, thermostats, and barostats

At thermodynamic equilibrium, a system does not occupy a single conformation but fluctuates among many possible microstates. The collection of all such microstates, together with their energy-dependent probabilities, constitutes a statistical ensemble. MD produces a time-ordered trajectory of microstates; over sufficiently long times, the frequency with which these microstates are visited reproduces the probability distribution of the specific ensemble chosen by the user.

Newtonian dynamics formulated as in Equations (1) and (2) naturally produce a thermodynamic ensemble with constant particle number, volume, and energy, known as the microcanonical or *NVE* ensemble. However, molecular systems in solution are not isolated: they rapidly exchange energy with a macroscopic solvent environment. Under experimentally relevant conditions, energy is therefore not conserved but exchanged with the surroundings until thermal equilibrium is reached. Consequently, molecular simulations aiming to emulate experimental setting employ constant-temperature ensembles rather than *NVE*.

Two constant-temperature ensembles widely used in biomolecular and food-protein simulations are the *NVT* (canonical) and *NPT* (isothermal–isobaric) ensembles. The *NVT* ensemble preserves constant particle number, volume, and temperature, whereas the *NPT* ensemble maintains constant particle number, pressure, and temperature. In typical food-protein simulations, *NVT* is used during the equilibration stage to stabilize the temperature, while *NPT* is used during the production run, the primary data-collection phase, to mimic experimentally relevant conditions (e.g., room temperature and atmospheric pressure). However, in other systems, production simulations are sometimes performed under an *NVT* ensemble for various reasons (e.g., faster performance, better numerical stability, or simpler implementation of some enhanced sampling algorithms), which is acceptable because *NVT* and *NPT* ensembles are often approximately identical for simulations of biomolecules in aqueous solutions due to the low compressibility of water (Hansen and Gunsteren, 2014).

To approximate the effects of temperature, algorithms known as thermostats are employed. Specifically, the equations of motion are extended with additional terms or constraints introduced by the algorithms. Typically, thermostats regulate the system’s kinetic energy, usually by adjusting particle velocities to maintain the desired temperature, resulting in properties consistent with the

Maxwell–Boltzmann distribution and energy equipartition (Jain et al., 2012). Popular choices for thermostats include the Nosé-Hoover thermostat (Hoover, 1985), the Bussi-Donadio-Parrinello stochastic velocity rescaling thermostat (V-rescaling) (Bussi et al., 2007), and the Langevin thermostat (Izaguirre et al., 2001). It is worth noting that the Berendsen thermostat (Berendsen et al., 1984), once popular for its simplicity and effective temperature control, is now generally discouraged for NVT ensemble sampling, as it fails to generate the correct thermodynamic distribution (Bussi et al., 2007).

The pressure of the simulated systems is maintained through algorithms called barostats. Barostats control the simulation box volume to maintain pressure, enabling NPT ensemble simulations (Ke et al., 2022). Widely used barostats include the Andersen barostat (Andersen, 1980), the Parrinello-Rahman barostat (Parrinello and Rahman, 1981), the Langevin piston (Feller et al., 1995), and the Bernetti–Bussi stochastic cell rescaling (Bernetti and Bussi, 2020). Because barostats act on the box geometry rather than directly on atomic motion, they do not possess a simple force-based form like the Langevin thermostats.

Importantly, controlling thermodynamic conditions alone does not guarantee physically meaningful dynamics unless the forces governing particle motion are also accurately defined. As noted above, these forces are derived from the system’s potential energy. To define and compute this potential energy surface, a set of equations and parameters that describe atomic interactions (e.g., bond stretching, angle bending, torsional rotation, electrostatic, and van der Waals forces) are used. Together, these equations and parameters define the force field, which governs how atomic interactions are represented and ultimately determines the physical realism of the simulated system.

1.2.3. Atomistic force fields

Atomistic force fields are mathematical frameworks that describe interatomic interactions through predefined functional forms and relevant parameters. These parameters are typically derived from both empirical data (e.g., crystal and NMR structures, vibrational spectra, and solvation free energies) and quantum mechanical calculations (often density functional theory (DFT) applied to small model compounds) (Cornell et al., 1995; MacKerell et al., 1998). Through this parameterization strategy, atomistic force fields approximate the structure, dynamics, and thermodynamic behavior of biomolecules.

To implement these interactions in practice, force fields assign atoms to specific ‘types’ that determine how they interact (e.g., alcohol oxygen atoms and amide oxygen atoms have different equilibrium C–O bond lengths, atomic charges, and van der Waals radii). Despite being limited to physical rather than chemical changes, conventional biomolecular force fields are able to capture conformational dynamics, formation and breakage of hydrogen bonds, and hydrophobic interactions, which are the essential components of phenomena such as protein folding (Lane et al., 2013), protein-ligand interaction (Schindler et al., 2020), and protein aggregation (Carballo-Pacheco, 2016).

Depending on their applications and the level of details captured, atomistic force fields can broadly fall into practical classes: additive (fixed charge) and polarizable force fields (Lemkul et al., 2016), and recently, machine-learning force fields, which could be classified as polarizable force fields, but are different enough in the mathematical representation to deserve their own category. As these latter two are still rarely used in food-protein simulations and fall outside the scope of this review, the remainder of this section will focus exclusively on additive force fields.

Popular additive atomistic force fields include CHARMM (Chemistry at Harvard Molecular Mechanics) (Mackerell et al., 2004; Huang et al., 2016), Amber (Assisted Model Building and Energy Refinement) (Cornell et al., 1995; Maier et al., 2015), and OPLS (Optimized Potentials for Liquid Simulation) (Jorgensen et al., 1996). The history of these force fields has been reviewed elsewhere (Best, 2019). Among these, the CHARMM force field family has been widely adopted in protein-related molecular dynamics due to its comprehensive parameter libraries and continuous development. Moreover, the availability of the semi-automated CHARMM General Force Field (CGenFF) (Vanommeslaeghe and Mackerelle, 2012), combined with optimization tools such as FFParam (Kumar et al., 2020), enables streamlined parameter generation for ligands and modified residues. This makes CHARMM particularly useful for studying complex food systems where non-standard molecules are common. Given its widespread adoption, robust parameterization, and accessibility for modeling non-standard molecules, this review adopts CHARMM as the representative atomistic force field for food protein simulations.

1.2.3.1. Bonded energy terms for atomistic force fields: CHARMM36

The bonded energy terms in CHARMM36 consist of four primary interactions: bond stretching, angle bending, dihedral torsion, and improper torsion (Fig. 1.2). In addition, CHARMM36 incorporates two cross terms: the Urey–Bradley and CMAP (correction map) potentials, which enhance the accuracy of structural representations of proteins. A cross term refers to a potential energy component that couples multiple internal coordinates (e.g., bond lengths and angles, or pairs of torsion angles), capturing interactions not accounted for by standard bonded terms.

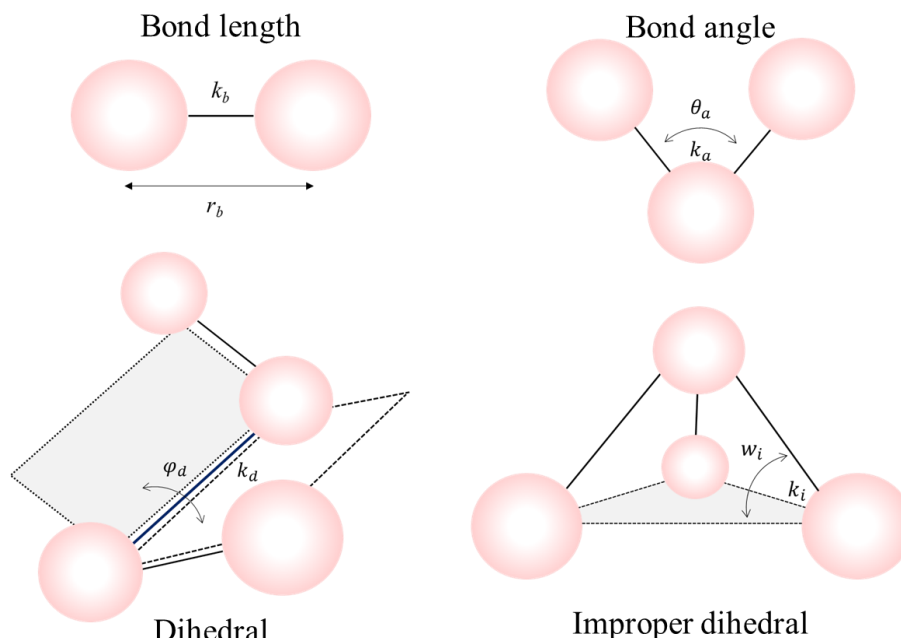


Figure 1.2. Schematic representation of the primary bonded interaction terms in the CHARMM36 force field: bond stretching, angle bending, dihedral and improper torsions, corresponding to Equations (5) – (8), respectively.

In molecular mechanics, bonded interactions describe the energetic penalties associated with deviations from ideal geometric configurations between covalently bonded atoms. The bond stretching potential accounts for the energy required to extend or compress a bond from its equilibrium length. In CHARMM36, similar to many other force fields, bond stretching is represented as a harmonic potential (Equation 5). The angle bending term penalizes deviations from the ideal bond angle and is also expressed as a harmonic function (Equation 6). The dihedral torsion term captures the rotational energy around a bond axis involving four sequentially bonded atoms (Equation 7). The improper torsion term is used to maintain structural constraints, such as planarity or chirality, by applying potential to four atoms that are not sequentially bonded (Equation 8). The mathematical expressions for each bonded interaction terms are given by:

$$E_{bond} = \sum_{bonds b} k_b (r_b - R_b)^2 \quad (5)$$

$$E_{angle} = \sum_{angles a} k_a (\theta_a - \Theta_a)^2 \quad (6)$$

$$E_{dihedral} = \sum_{dihedrals d} k_d [1 + \cos(n_d \varphi_d - \delta_d)] \quad (7)$$

$$E_{improper} = \sum_{improvers i} k_i (w_i - W_i)^2 \quad (8)$$

where k_b is the force constant for bond b ; r_b is the distance between the bonded atoms; R_b is the reference (equilibrium) bond length; k_a is the angle force constant; θ_a is the bond angle; Θ_a is the reference bond angle; k_d is the dihedral force constant; n_d is the periodicity of the dihedral angle; φ_d is the dihedral angle value; δ_d is the phase shift angle; k_i is the improper dihedral force constant; w_i is the improper dihedral angle value; and W_i is the equilibrium improper dihedral angle. It should be noted that while CHARMM-format parameter files use the conventions given in Equations (5) and (6), the native format of GROMACS assumes a form that includes a factor of 1/2 before k_b and k_a (Abraham et al., 2026).

As mentioned, CHARMM36 includes two correction terms beyond the standard bonded potentials, namely Urey-Bradley and CMAP potentials. The Urey-Bradley potential applies a harmonic interaction between 1–3 atom pairs (i.e., atoms separated by two bonds) within a bond angle, using the same functional form as bond-stretching (Urey and Bradley, 1931). While its mathematical form is identical to a bond-stretching term, the 1–3 interaction does not correspond to a conventional covalent bond. Conceptually, it serves to refine angle bending behavior, particularly in systems where simple angular potentials are insufficient to reproduce experimental geometries. Its explicit expression is omitted here, as it can be subsumed into the bond terms.

Moreover, CMAP accounts for dihedral angle coupling (Mackerell et al., 2004). While conventional dihedral terms model individual torsional angles (ϕ and ψ), the conformational behavior of protein backbones is fundamentally governed by their coupled interactions. For example, a ϕ angle of -60° may be associated with either an α -helical or β -sheet conformation, depending on the corresponding ψ angle. Such conformational preferences emerge not from each dihedral angle alone, but from their specific combinations, a relationship clearly depicted in the Ramachandran map. Unlike standard energy terms that use two or three global parameters, CMAP applies a two-dimensional correction table based on the combination of ϕ and ψ dihedral angles in the protein backbone. The values in this table are typically obtained from quantum mechanical calculations or from statistical analysis of experimentally determined protein structures (Mackerell et al., 2004).

1.2.3.2. Non-bonded energy terms for atomistic force fields: CHARMM36

The non-bonded potential energy in CHARMM36 force field consists of two primary terms: the Coulombic (electrostatic) interaction and the Lennard-Jones (van der Waals) interaction (Fig. 1.3). The Coulombic term depends on the partial charges (q_i and q_j) assigned to the atoms, their interatomic distance (r_{ij}), and the Coulomb constant ($k_{Coulomb}$), in appropriate units ($332.064 \text{ kcal/mol } \text{\AA} \text{ e}^{-2}$ or $138.935 \text{ kJ/mol nm e}^{-2}$) (Equation 9). For AA simulations with explicit water molecules (e.g. TIP3P model), the permittivity ϵ_r is set to be 1 and the dielectric properties of the medium emerge from the electrostatic interactions, rather than being enforced. It should be noted that the dielectric constant that stems from the TIP3P_{mod} water model (modified TIP3P water to include nonzero van der Waals interactions on the hydrogen atoms, the standard water model of CHARMM) is somewhat higher than real water (97 in simulation versus 78 in experiments near room temperature) (Höchtel et al., 1998).

$$E_{elec} = \sum_{atompairs i \neq j} \frac{k_{Coulomb} q_i q_j}{\epsilon_r r_{ij}} \quad (9)$$

$$E_{vdW} = \sum_{atompairs i \neq j} \epsilon_{ij} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^6 \right] \quad (10)$$

$$E_{potential} = E_{bond} + E_{angle} + E_{dihedral} + E_{improper} + E_{vdW} + E_{elec}$$

Where $k_{Coulomb}$ is the Coulomb constant; q_i is the charge of atom i , and ϵ_r is the relative permittivity (dielectric constant); r_{ij} is the current distance between atoms i and j ; ϵ_{ij} is the depth of van der Waals potential energy well; R_{ij}^{min} is the distance of minimum van der Waals energy for atoms i and j .

The Lennard-Jones potential describes short-range repulsion and dispersion forces between atoms. It depends on the distance between atoms (r_{ij}) and the depth of their potential well (ϵ_{ij}) and the distance of minimum van der Waals potential energy (R_{ij}^{min}), as shown in the Equation (10) and Fig. 1.3c. Most force fields, including CHARMM, give only single values of ϵ_i and R_i^{min} for each atom type, and the interactions between atoms of different types (types i and j) are calculated according to the Lorentz-Berthelot mixing rules: $\epsilon_{ij} = (\epsilon_i \epsilon_j)^{1/2}$ and $R_{ij}^{min} = (R_i^{min} + R_j^{min})/2$. This reduces the work of parameterizing the force field, since only N pairs of parameters are needed for N types, while $N(N+1)/2$ parameters would be needed for all pairs. However, these mixing rules do not result in correct behavior for some pairs of atoms, especially ions and other highly charged groups (Luo and Roux, 2010). In this case, the mixing rule parameters are overridden by specific parameters for pairs of atom type, which are referred to as “NBFIX” terms in the CHARMM documentation and also sometimes called “bespoke” parameters. For instance, in the CHARMM force field, Na^+ and Cl^- ions are given a larger R_{ij}^{min}

than is calculated by the mixing rules above to reduce the spurious formation of ion pairs in solution (Luo and Roux, 2010).

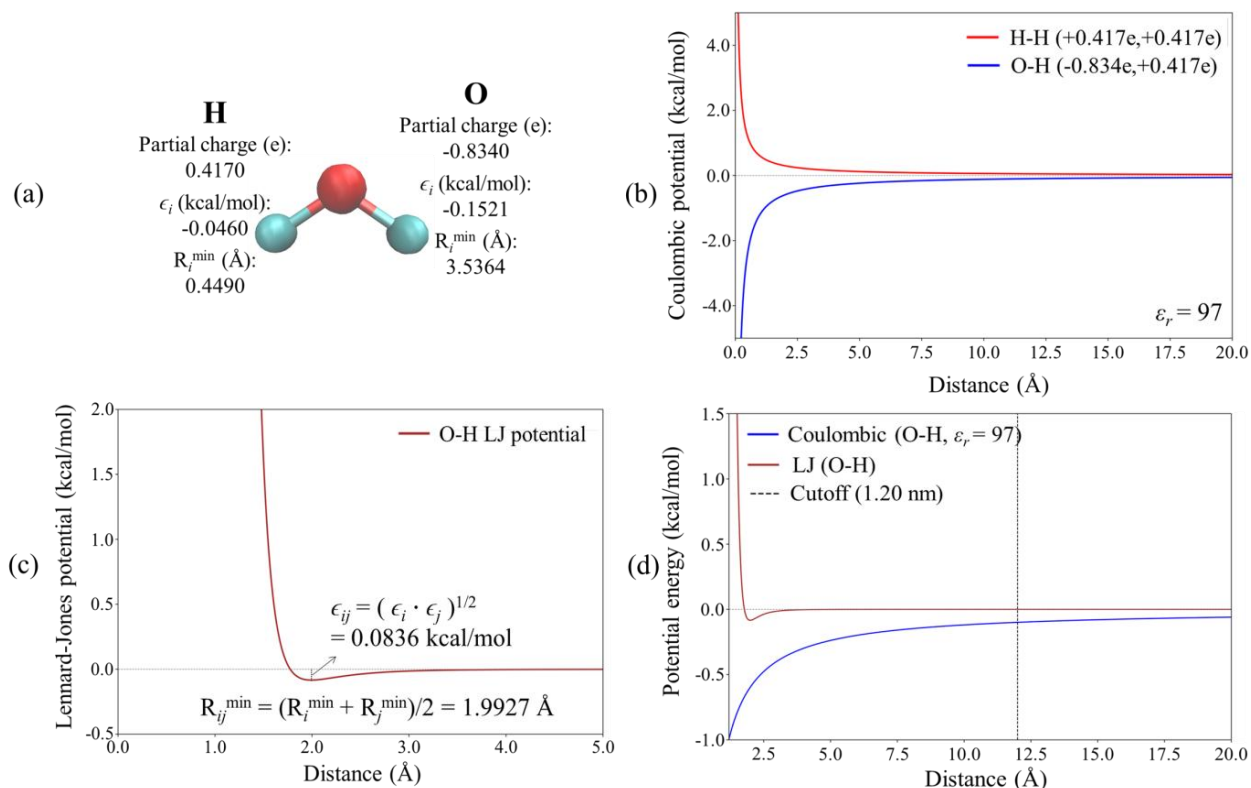


Figure 1.3. TIP3P_{mod} water model and its defined parameters in CHARMM36 (a); Coulombic interaction potentials for H–H and O–H pairs ($\epsilon_r = 97$) (b). Lennard–Jones potential for an O–H pair (c). Comparison of Coulombic and Lennard–Jones interaction ranges under a nonbonded cutoff of 1.20 nm (d).

In practice, van der Waals interactions are inherently short-ranged and decay rapidly with distance. This allows their contributions to be truncated beyond a specified cutoff (typically 10–12 Å) without significant loss of accuracy (Fig. 1.3d), although fully long-range methods like those discussed below for Coulombic interactions may yield greater accuracy (Yu et al., 2021). In contrast, Coulombic interactions are long-ranged and decay more slowly, so simple truncation

can lead to unrealistic behavior, especially near charged surfaces like charged lipid membranes, where the electric field should be nearly constant with distance rather than decaying.

Therefore, long-range electrostatics are typically handled using advanced methods such as the particle mesh Ewald (PME) method (Darden et al., 1993), which efficiently computes long-range interactions under periodic boundary conditions. Periodic boundary conditions are commonly used in molecular dynamics simulations, replicating the simulation box along all three axes to mimic an infinite system, thereby avoiding edge effects caused by finite box sizes. The combination of periodic boundary conditions and PME is considered the standard setup in most modern atomistic simulations. Under PME, electrostatic interactions are decomposed into a short and long-range component. The short-range component can be directly calculated between pairs within the cutoff distance. The long-range component of the electrostatic interaction varies smoothly with distance, and such slowly varying functions are efficiently and accurately evaluated in reciprocal (Fourier) space using fast Fourier transforms. This reduces computational complexity from $O(N^2)$, required for brute-force pairwise calculations, to $O(N \log N)$, significantly improving scalability for large systems.

It should be noted that the length of covalent bonds is much shorter than the van der Waals minimum and is in the deep repulsive part of the Lennard-Jones potential. In CHARMM and other conventional biomolecular force fields, atoms that covalently bonded to one another (1–2 interactions) or separated by two covalent bonds (1–3 interactions) interact only via bonded interactions and nonbonded interactions are excluded.

1.2.4. Coarse-grained force fields

To mitigate the computational burden of AA simulations, CG simulation groups multiple atoms into single beads. This simplification reduces the number of interacting particles in the system but necessitates a different class of force fields specifically designed for coarse-grained representations. Unlike atomistic force fields, which operate on a fixed, atomic resolution, CG force fields define both the mapping scheme (i.e. how atoms are grouped) and the interaction parameters that govern the behavior of the resulting beads. Therefore, CG force fields constitute an integrated framework rather than a standalone parameter set and must be used in conjunction with their prescribed mapping conventions to ensure physical accuracy and transferability.

A variety of coarse-grained force fields have been developed for biomolecular systems, including SPICA (surface property fitting coarse graining) (Seo and Shinoda, 2018), UNRES (united-residue) (Liwo et al., 2004), SIRAH (Southamerican Initiative for a Rapid and Accurate Hamiltonian) (Darré et al., 2015) and Martini (Marrink et al., 2007). These models differ in their levels of resolution and targets specific molecular categories such as lipids, membranes, or proteins. In practice, arguably the two most widely employed CG force fields for proteins are Martini and SIRAH. Both models have undergone extensive validation and are actively maintained with established mapping tools and parameter libraries.

Coarse-grained force fields are typically parameterized through a combination of bottom-up and top-down approaches. In bottom-up strategies, interaction potentials are derived to match data from atomistic molecular dynamics or quantum mechanical calculations, thereby preserving detailed molecular-level physics. In contrast, top-down approaches adjust parameters to reproduce macroscopic or experimental observables, such as solvation free energies, partition coefficients, or diffusion rates. Both Martini and SIRAH employ hybrid parameterization

schemes that integrate both approaches, but they differ fundamentally in their mapping philosophies.

Martini aims to provide a general-purpose and broadly applicable force field by grouping chemically similar atoms into coarse-grained beads (Marrink et al., 2023). In contrast, SIRAH adopts a structure-aware mapping scheme, where beads are assigned not only on chemical similarity but also according to their structural roles. This philosophical distinction is most evident in how they treat backbone structures. In Martini, all backbone atoms of each residue are coarse-grained into a single particle. In Martini 2.x, the type of this particle (e.g., P5 or N0 bead) is assigned based on the predefined secondary structure (Monticelli et al., 2008). In the latest Martini 3, the default backbone particle type is P2, with specific exceptions such as SP1 for glycine, SP2 for alanine, and SP1a for proline (Souza et al., 2021).

In contrast, SIRAH 1.0 and 2.0 both adopt a finer-grained representation of the protein backbone, assigning separate beads to backbone atoms such as N, C α , and O, enabling a more detailed reconstruction of geometry and hydrogen bonding patterns (Darré et al., 2015; Machado et al., 2019). This finer granularity allows SIRAH to more accurately reconstruct hydrogen bonding geometries and backbone dihedrals. As a result, SIRAH can intrinsically model secondary structure transitions without the need for external restraints. On the other hand, Martini model does not support spontaneous changes in secondary structure, and requires additional stabilization mechanisms, such as elastic network models (Periole et al., 2009) or Gō-type potentials (Poma et al., 2017) to preserve structural integrity. Elastic networks consist of harmonic bond potentials that hold native contacts together, while Gō models are special LJ potentials that reinforce native contacts.

In food protein research, it is often necessary to consider the effects of co-solvents and additives used in food processing, such as ethanol or salts. Martini stands out in this regard for its extensive parameter coverage across a wide chemical space. For instance, even in its latest version (v2.2) SIRAH does not include a parameter set for generic solvents such as methanol, while Martini 3 includes over 50 solvent models by default. In addition, Martini is supported by a robust ecosystem of automated pipelines—including martinize2.py (Kroon et al., 2025), INSANE.py (INSert membrANE) (Wassenaar et al., 2015), and polyply.py (Grünewald et al., 2022), which streamline bead assignment, system construction, and input generation without requiring users to have extensive coarse-graining experience. The existence of centralized repositories like the MAD (MArtini Database) (<https://mad.ibcp.fr/>) (Hilpert et al., 2023) further enhances accessibility by offering validated topologies and ready-to-use models. Given that generating a coarse-grained mapping is highly technical and requires numerous calculations to ensure balanced coarse-grained parameters, Martini offers a pragmatic and scalable solution for non-experts in molecular simulation. For these reasons, the latest Martini 3 force field will be discussed in detail as the representative CG model in this review.

1.2.4.1. Coarse-grained force field: Martini 3

Originally developed for lipids (Martini 1.0 and 2.0) (Marrink et al., 2004; Marrink et al., 2007), the Martini force field was later extended to model a broader range of biomolecules such as DNA and proteins (Martini 2.1 and 2.2) (Monticelli et al., 2008; De Jong et al., 2013). While Martini 2.x was widely used, it suffered from overstabilized interactions and artificial aggregation tendencies, a phenomenon commonly referred to as the “stickiness problem” (Alessandri et al., 2019).

To address these issues, the latest Martini 3 model was introduced in 2021 (Souza et al., 2021). The developers employed different bead types and reparameterization strategies,

demonstrating improved representations of interactions between proteins and peptides compared to earlier versions. The force field employs a mapping scheme that typically groups two to four non-hydrogen atoms (i.e. heavy atoms) into a single coarse-grained bead. Specifically, Martini 3 combines heavy atoms and their attached hydrogens into a single bead, classifying them as polar (P), nonpolar (N), charged (Q), apolar (C), halogen (X), water (W), and divalent (D). Additionally, it categorizes them into subgroups based on hydrogen bonding capability and the magnitude of polarity, resulting in a total of 29 bead types. The model has been widely applied across biomolecular systems, including the coacervation of poly-lysine and poly-glutamate peptides (Tsanai et al., 2021), oligomerization of poxvirus proteins (Thomasen et al., 2023), and spontaneous membrane insertion of melittin (a membrane-penetrating peptide) (Sun et al., 2022).

As Martini 3 model is parameterized for a coarse-grained representation, an initial mapping from all-atom structure to CG beads is required. This mapping process, particularly for protein, can be automatized using `martinize2.py` script in the `vermouth` python library without extensive user intervention. In earlier Martini 2.x versions, secondary structure information was required as an input, as the assignment of backbone bead types depended on the local secondary structure. The Martini 3 model has dispensed with this requirement; however, protein structure is maintained by predefined Gō or elastic network models (Fig. 1.4a). Both structural restraints can be applied using `martinize2.py` script during CG mapping.

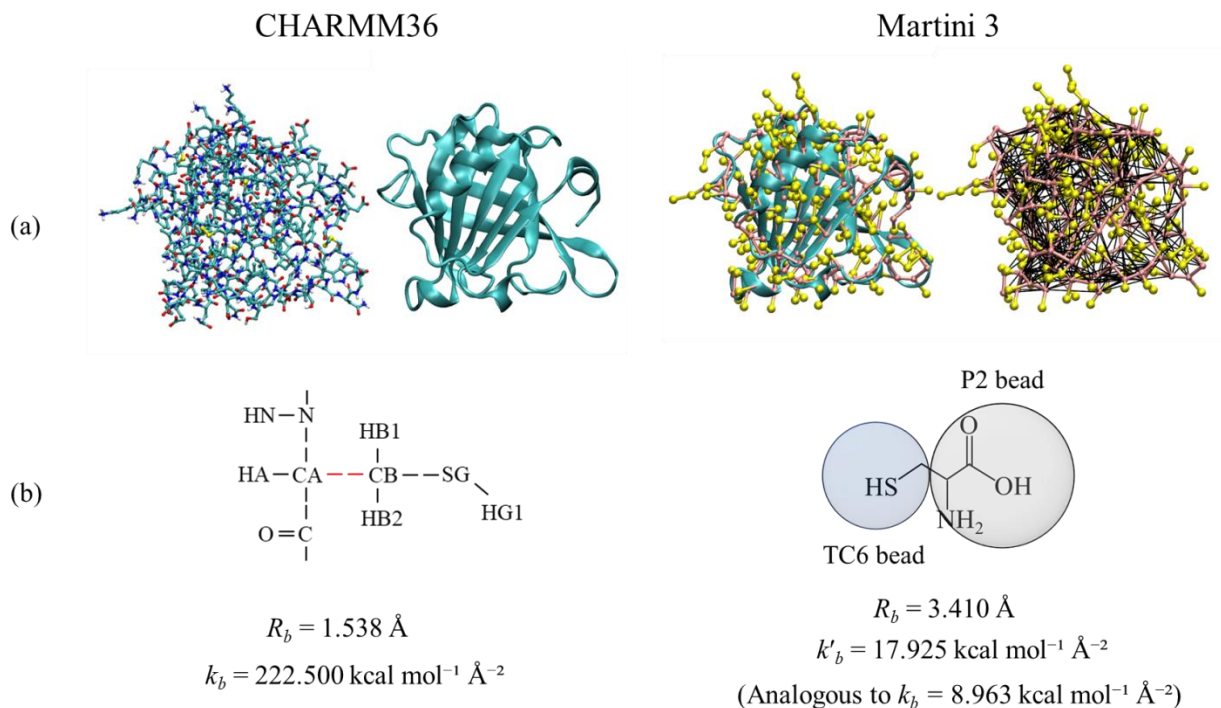


Figure 1.4. Atomistic and coarse-grained representations of β -lactoglobulin monomer (PDB ID: 1BEB) (a). The protein structure is shown as (left to right) an all-atom representation, a cartoon representation of the same structure, a bead-mapped coarse-grained model overlaid on the cartoon backbone, and the coarse-grained bead representation including elastic networks. Visualization was performed using VMD (Humphrey et al., 1996) and Martiniglass (Brasnett and Marrink, 2025). Representation of a cysteine residue and its bond stretching parameters in CHARMM36 and Martini 3 force fields (b). The notation follows Equations (5) and (11).

1.2.4.2 Potential energy terms for CG force fields: Martini 3

As in AA simulations, Martini CG simulations rely on classical mechanics and Newton's second law to propagate particle motion through potential energy functions. The forces acting on CG beads are derived from bonded and non-bonded interactions, analogous to those in AA force fields. The bond stretching potential in Martini 3 adopts a form similar to that of CHARMM36. However, Martini literature and official Martini GROMACS-format parameter files use the convention where an explicit $\frac{1}{2}$ is included in the energy function, while CHARMM-format parameter files do not include this factor of $\frac{1}{2}$ (the conversion is simply: $k_{\text{Martini}} = 2k_{\text{CHARMM}}$). Martini differs from CHARMM36 in its angle bending formulation, using cosine-based potentials

(Equation (13)) and restricted bending potential (Equation (14)) (Abraham et al., 2026) rather than the harmonic form. The choice of angular potential form (cosine-based or restricted bending) follows the predefined angle type assignments in the force field parameter files. The terms for dihedral angle and improper dihedral are identical to those of CHARMM36. On the other hand, Martini framework does not contain the cross-terms in CHARMM36, such as CMAP, reflecting its simplified design.

$$E_{bond} = \frac{1}{2} \sum_{bonds b} k'_b (r_b - R_b)^2 \quad (11)$$

$$E_{angle} = \sum_{angles a \in A_{cos}} E_{cos}(\theta_a) + \sum_{angles a \in A_{Reb}} E_{Reb}(\theta_a) \quad (12)$$

$$E_{cos} = \frac{1}{2} \sum_{angles a \in A_{cos}} k'_a \{cos(\theta_a) - cos(\theta_a)\}^2 \quad (13)$$

$$E_{Reb} = \frac{1}{2} \sum_{angles a \in A_{Reb}} k'_a \frac{\{cos(\theta_a) - cos(\theta_a)\}^2}{sin^2(\theta_a)} \quad (14)$$

A prime (') is used to denote Martini-specific parameters. Note that bond lengths and angles in this context refer to values between coarse-grained beads, not atoms. k'_b is the force constant for bond b ; r_b is the distance between the bonded beads; R_b is the reference bond length; k'_a is the angle force constant; θ_a is the current bond angle; A_{cos} and A_{Reb} denote subsets of angles assigned cosine-based and restricted bending potentials, respectively; θ_a is the reference bond angle.

While the potential energy functions in CHARMM and Martini share similar mathematical forms, the actual parameter values, such as equilibrium bond lengths and angles differ substantially between the two force fields. This is because Martini parameters describe interactions between beads, rather than between individual atoms. Hence, generally bond distances and Lennard-Jones radii are larger and the bond, angle, and dihedral force constants are weaker to reflect the smoother,

averaged potential energy surface characteristic of coarse-grained models. For instance, in the case of a cysteine residue, the bond distance and bond force constant connecting the backbone and side-chain beads in MARTINI 3 are 3.410 Å and 8.963 kcal mol⁻¹ Å⁻² (following the convention of Equation (5)), respectively. In contrast, the distance and force constant between the C_α and C_β atoms in the CHARMM36 force field are 1.538 Å and 222.500 kcal mol⁻¹ Å⁻² (Fig. 1.4b).

The mathematical forms of the non-bonded potentials (Coulombic and van der Waals interaction) in Martini 3 are retained from those of CHARMM36. However, key differences exist in their parameterization. For Coulombic interactions, CHARMM assumes a dielectric constant of 1, with the dielectric constant emerging naturally from the behavior of the explicit polar solvent molecules. In contrast, the default Martini 3 (as well as previous Martini versions) employs a fixed dielectric constant of 15 (dimensionless) (Souza et al., 2021), conceptually representing the screening effect of polar groups that are not explicitly modeled. For instance, the Martini beads for atomic ions are supposed to represent the ion together with its hydration shell (Marrink et al., 2007). For van der Waals interactions, Martini 3 explicitly defines the depths of the Lennard-Jones energy well (ϵ_{ij}) for each pair of bead types. On the other hand, in AA force fields, they are more often assigned according to the Berthelot mixing rule. These distinctions underscore that, despite using similar mathematical forms, AA and CG force fields differ fundamentally in what their parameters represent: explicit atomic interactions in AA models versus effective, resolution-dependent interactions in coarse-grained models such as Martini.

1.2.5. AA and CG-MD simulation: Time and length scale

Understanding the processing of protein-based foods requires insight into protein interactions across multiple length and time scales. For example, during tofu production, thermal

processing induces individual soy proteins to unfold at the nanometer scale, exposing hydrophobic regions and triggering conformational transitions, such as α -helices converting into β -sheets or random coils (Guan et al., 2021). With continued heating, disulfide bonds form between proteins, promoting aggregation and formation of filamentous structure assemblies (Chen et al., 2023), that may span tens of nanometers or larger. Under the action of coagulants, these filaments further associate into a macroscopic gel network, ultimately forming the tofu curd observed at the micrometer scale and beyond (Guan et al., 2021). This example illustrates how food protein functionality emerges from molecular level events and evolves into bulk material properties, emphasizing the need to study interactions at different spatial and temporal resolutions.

As demonstrated in Fig. 1.5, different simulation techniques span different spatio-temporal scales. Starting from the lowest time/length scale, Quantum mechanical (QM) methods, particularly in the form of hybrid quantum mechanics/molecular mechanics (QM/MM) approaches, are used to model electronic interactions with high accuracy in localized regions of a molecular system. These methods are especially useful for studying bond formation and cleavage, charge transfer, or protonation events within a larger biomolecular context. Intermediate approaches, including reactive force fields and machine-learned interatomic potentials, extend classical simulations to regimes where bond breaking/forming are important, while retaining atomistic resolution. Moving up the scale, AA models eschew explicit representation of electrons, averaging their interactions into atom-centered force terms and allowing a much more computationally efficient classical physics description. On the other hand, CG models average the motion of multiple atoms into single beads, leading to a smoother free energy surface than AA models. A smoother energy surface translates to less abrupt changes in forces acting on beads. This allows for a larger integration time step in CG simulation and accelerated kinetics even when accounting

for the increased time step (Marrink et al., 2007). The key similarities and differences between AA and CG-MD simulations in terms of resolution, accessible time and length scales, and typical research interests are summarized in Table 1.

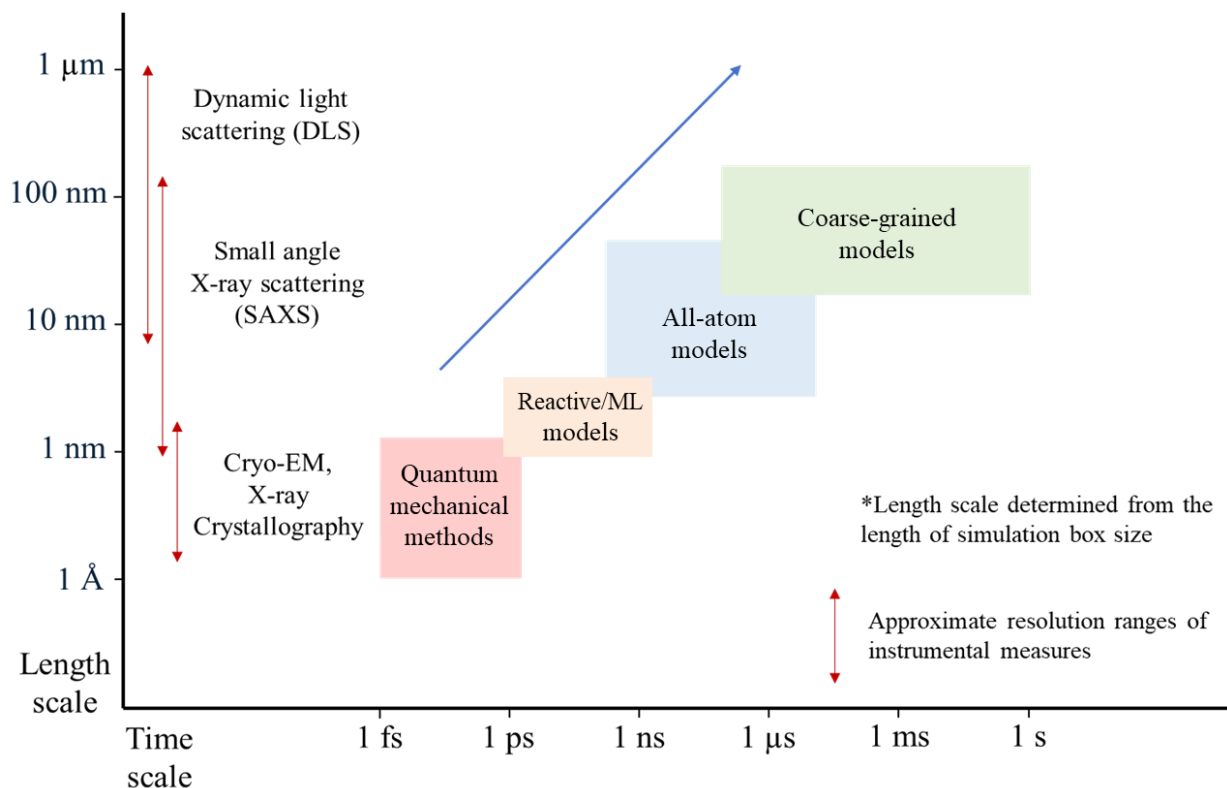


Figure 1.5. Approximate length and time scale comparison of molecular simulations and experimental techniques. The scales shown are approximate and are intended to reflect practical use cases rather than strict boundaries.

Table 1.1. Comparison between all-atom and coarse-grained molecular dynamics simulations.

	Characteristics	All-atom (AA) MD simulation	Coarse-grained (CG) MD simulation
Similarity	Mechanics	Numerical integration of Newtonian mechanics	
	Energy function	Generally consists of bonded and non-bonded terms	
	Integration method	Verlet-type integration schemes are commonly employed	
	Software (engines)	GROMACS, AMBER, CHARMM ...etc	
Difference	Resolution	Atomistic	Beads (groups of atoms)
	Typical timesteps	1 – 2 fs	5 – 20 fs
	Typical timescale	Hundreds of ns – low μ s	Low of μ s – hundreds of ms
	Typical system size	Tens of nm	Up to few hundred of nm
	Force field structures	Explicit atomic representation and interaction parameters	Defined by mapping schemes and effective interaction parameters
	Typical interests	Local conformational dynamics and solvent-mediated interactions at atomic resolution	Collective dynamics, mesoscale assembly, and long-time kinetic processes

1.2.6. Available tools for MD simulation

1.2.6.1. Simulation engines

Many simulation engines have been developed for performing MD simulations.

Currently, engines commonly used by academic groups are GROMACS (Groningen Machine for Chemical Simulations) (Abraham et al., 2015), NAMD (Nanoscale Molecular Dynamics) (Phillips et al., 2020), Amber (Assisted Model Building with Energy Refinement) (Case et al.,

2023), CHARMM (Chemistry at Harvard Molecular Mechanics) (Brooks et al., 2009), OpenMM (Open Molecular Mechanics) (Eastman et al., 2023), and LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) (Thompson et al., 2022). Such engines and their source codes are generally distributed freely under open-source licenses for academic users, or in the case of GROMACS, OpenMM, and LAMMPS, any user. These software platforms calculate forces according to a supplied molecular model and numerically integrate the equations of motion.

It is important to distinguish between simulation engines from force fields, as some simulation engine packages share names with force fields (e.g., Amber and CHARMM), owing to shared development history. While development of the CHARMM and Amber MD engines was originally linked to development of the CHARMM and Amber force fields, respectively, force fields and MD engines have become more independent over time. Most modern engines support all common force fields. For instance, the Amber MD engine can be used for simulations with the CHARMM force field (Crowley et al., 2009) and the NAMD MD engine, which was originally developed to support the CHARMM force field, can read native Amber input files (Phillips et al., 2005), and validated conversions of the Amber force field to the CHARMM format for use with NAMD have been published (Antolínez et al., 2024). Moreover, tools have been developed to generate the input and output file formats expected by different engines. For example, CHARMM-GUI (Jo et al., 2008) provides input structures and configuration files for performing simulations under CHARMM force fields with CHARMM, NAMD, Amber, GROMACS, OpenMM, and LAMMPS and GROMACS input files for Martini simulations (Lee et al., 2020; Qi et al., 2015). Other tools, such as charmm2gmx (Wacha et al., 2023), ParmEd

(Swails et al., 2010), and VMD (Humphrey et al., 1996) can facilitate converting structure and force field file formats used by different simulation engines.

However, some limitations and technical barriers remain for using force fields outside their “native” simulation engine. For instance, the Martini force field has mostly been developed using GROMACS and the official releases from the Marrink group use GROMACS formats (Marrink et al., 2023). While complete OpenMM implementation is also available (MacCallum et al., 2023), it does not seem to be currently possible to fully implement Martini 3 protein models in NAMD owing to an inability to apply rigid bond constraints to rings such as those used in phenylalanine side chains.

1.2.6.2. System preparation and visualization

Efficient MD simulations require not only a simulation platform but also supporting tools for system construction and input preparation. For constructing initial systems, such as solvated proteins, micelles, or bilayers, software like Packmol (Martínez et al., 2009) can arrange molecules in defined spatial configurations. Several higher-level tools offer automation and graphical interfaces to simplify the input generation process. CHARMM-GUI (Jo et al., 2008), for instance, is a widely used web-based platform capable of producing topology and coordinate files. For CG simulations with the Martini force field, the INSANE (Wassenaar et al., 2015) Python library enables custom construction of bilayers, vesicles, and complex solvated systems, while Martinize2 and Vermouth (Kroon et al., 2025) produce coarse-grained models and simulation inputs from atomic structures. Moreover, visualization and trajectory inspection are essential throughout the simulation process. Popular tools for this purpose include VMD (Humphrey et al., 1996), PyMOL (DeLano, 2002), and ChimeraX (Meng et al, 2023).

1.2.7. Extracting physical insights of proteins from MD simulations

After completing an MD simulation, the resulting trajectory provides a wealth of structural and energetic information that can be analyzed using various techniques. While standard metrics such as RMSD (root mean square deviation) and RMSF (root mean square fluctuation) are widely used, they have been thoroughly reviewed elsewhere (Yu et al., 2024) and will not be discussed in detail here. Likewise, approximate free energy methods such as MM/PBSA (molecular mechanics/Poisson–Boltzmann surface area) and more rigorous ones such as umbrella sampling, adaptive biasing force, metadynamics, and FEP (free energy perturbation) are excluded in this section owing to its basic scope and the advanced nature these approaches. Instead, this section highlights four analytical approaches that are particularly well-suited for food protein simulations: hydrogen bond analysis, solvent-accessible surface area, secondary structure profiling, and contact frequency mapping. These methods provide interpretable and robust metrics for assessing structural dynamics, hydration properties, and inter-residue interactions, offering insight into the behavior of food proteins under simulated conditions.

1.2.7.1. Hydrogen bonding

Hydrogen bonds play a crucial role in determining key functional properties of food proteins, such as solubility and gel strength. For example, in denaturant solubility assays, the addition of urea often increases protein solubility by disrupting internal hydrogen bonding networks. Similarly, during the thermal gelation of soy protein isolate, heat-induced unfolding breaks intra-molecular hydrogen bonds and enables the formation of new intermolecular hydrogen bonds, thereby promoting the development of a three-dimensional gel matrix (Fu et al., 2023).

In AA simulations, hydrogen bonds are typically identified using geometrical criteria involving the donor and acceptor atoms. For instance, GROMACS uses a distance cutoff of 3.5 Å

and a donor–hydrogen–acceptor angle of 30° by default (Abraham et al., 2026), while the VMD Hbonds plugin uses slightly stricter values of 3 Å and 20° (Humphrey et al., 1996). In CG simulations, the reduced resolution of the beads eliminates explicit hydrogen atoms, making standard hydrogen bond definitions inapplicable. One workaround is to reverse-map the CG structure back to an all-atom representation and then perform hydrogen bond analysis using standard AA tools. For instance, in studying actin filaments, the number of hydrogen bonds was calculated using a model backmapped from the Martini to GROMOS force field (Kim et al., 2016). Alternatively, a CG-specific geometric criterion, such as a donor-acceptor distance of 4.0 Å and a donor-hydrogen-acceptor angle of 40°, can be defined to approximate the number of hydrogen bonds from the CG model (Piskorz et al., 2021). Lastly, users may consider conducting multi-scale modeling by performing both AA and CG simulations and focusing on the AA simulation results for hydrogen bonding analysis.

1.2.7.2. Secondary structure analysis

Changes in protein secondary structure are often closely linked to functional transitions such as gelation. For instance, during dense phase CO₂-induced gelation of surimi, the contents of α -helix and random coil decrease, while β -sheet and β -turn increase, correlating with enhanced gel strength (Duan et al., 2023). Moreover, monitoring secondary structure throughout a simulation can provide insight not only into conformational transitions but also into the convergence and stability of the modeled system. These computational results are often compared against experimental data obtained from circular dichroism (CD), FTIR, or NMR to evaluate structural realism.

Commonly employed tools for secondary structure annotation include DSSP (Dictionary of secondary structure in protein) (Kabsch and Sander, 1983) and STRIDE (Structural

identification) (Frishman and Argos, 1995). DSSP identifies secondary structure based on hydrogen bond patterns and backbone geometry, while STRIDE incorporates dihedral angles information (Zhang et al., 2015). In CG simulations using Martini 3, although no predefined secondary structure is required as input, the framework currently does not support natural secondary structure transitions (Souza et al., 2021). To approximate such transitions, one strategy involves reverse-mapping techniques to convert CG trajectories into atomistic models, followed by further AA simulations (Joshi and Deshmukh, 2021). Alternatively, certain CG force fields such as SIRAH allow more dynamic handling of secondary structures without artificial restraints (Latham and Zhang, 2022).

1.2.7.3. Solvent-accessible surface area (SASA)

Solvent-accessible surface area (SASA) quantifies the extent to which protein residues are exposed to the surrounding solvents, providing a useful measure of solvation and molecular surface accessibility (Ooi et al., 1987). In many food protein systems, total SASA is often less informative compared to hydrophobic SASA (i.e., SASA obtained by summing solvent-accessible surface areas of hydrophobic residues such as Ala, Val, Leu, Ile, Met, Phe, Trp, and Pro), which more directly correlates with key behaviors, such as lipophilic binding, interfacial adsorption, and aggregation. Increased exposure of hydrophobic residues enhances a protein's affinity for nonpolar compounds, promotes adsorption at oil–water or air–water interfaces, and under specific processing conditions, facilitates protein–protein aggregation through hydrophobic interactions. These structural changes are closely tied to macroscopic properties such as emulsification capacity, turbidity, and gelation induced by heat or ethanol treatment.

SASA is typically computed by rolling a spherical probe, commonly with a radius of 0.14 nm (approximating a water molecule), across the van der Waals surface of the protein (Durham et

al., 2009). In CG simulations, lower resolution necessitates larger probe sizes, such as 0.56 nm to mimic four water molecules (Vuorela et al., 2010) or 0.37 nm derived from parameter fitting against all-atom results (Chakraborty et al., 2021). As mentioned, SASA calculations can also be restricted to specific residue types, such as hydrophobic, polar, or charged residues, offering targeted insight into solvent exposure relevant to specific functional properties.

1.2.7.4. Contact frequency analysis

In molecular simulations, simply monitoring the distance between two molecules or domains often fails to capture the persistence and statistical reliability of their interactions. Especially in food protein systems, where transient aggregation or ligand rebinding may occur, a more comprehensive method is required to assess contact stability over time. Beyond static contact maps, contact frequency analysis provides a more quantitative approach to assessing residue-residue interactions across an entire simulation trajectory. Rather than capturing a single structural snapshot, this method calculates the fraction of frames in which a given residue pair remains within a specified cutoff distance (e.g., 0.4 nm for protein-ligand interaction), resulting in a contact probability matrix.

This analysis is particularly useful for studying protein aggregation and ligand binding behavior in food protein systems. For instance, in recent simulations of bovine serum albumin under heat, acidic pH, and shear stress, contact frequency analysis revealed distinct domain-level behaviors despite identical RMSD values. Acidic pH caused complete inter-domain separation with preserved secondary structure, while heat led to local unfolding without domain dissociation (Jia et al., 2025). These findings highlight the value of contact maps in resolving structural nuances that RMSD alone cannot capture. Likewise, contact frequency maps between ligand and protein residues help pinpoint stable binding regions even in cases where hydrogen bonds are transient or

absent. These frequency-based maps offer a time-resolved statistical measure, which can be clustered or compared across different conditions (e.g., solvent environments, pH, or temperature). Tools such as MDAnalysis (Gowers et al., 2019) and MDTraj (McGibbon et al., 2015) facilitate this analysis, and visualization can be enhanced using heatmaps or network graphs.

1.3. Application scenarios

In practice, the choice between AA and CG simulations essentially depends on the required resolution and the nature of the research question. If the goal is to investigate detailed molecular behavior, such as hydrogen bonding networks, conformational transitions, or solvation effects, AA simulations are preferred due to their atomistic precision. Conversely, if the study focuses on mesoscale phenomena such as protein aggregation, phase separation, or long-timescale self-assembly, CG simulations offer a pragmatic balance between computational efficiency and physical insight. However, food protein systems inherently span multiple length scales, from local solvent-mediated interactions to network formation. In this perspective, neither AA nor CG-MD simulations alone are sufficient to capture their behavior. Hybrid AA–CG strategies are therefore particularly relevant in food protein research, where molecular interactions can be interpreted at two different scales. Accordingly, this section highlights representative application scenarios and case studies in which AA and CG simulations can be used in a complementary manner to investigate food protein systems, with selected studies summarized in Appendix Table A.1.

1.3.1. Solvent- and co-solute-induced protein responses

Solvent composition and the presence of co-solutes play a central role in determining the conformational stability, solubility, and aggregation propensity of food proteins. In many food

processing contexts, including alcohol-based extraction, antifreeze oligosaccharides addition, and salt-mediated structuring, protein behavior is governed by subtle changes in protein–solvent or solute interactions. AA simulations have therefore been widely applied to quantify solvent and solute-dependent changes in secondary structure, intra-molecular interactions, and residue-level flexibility. For instance, simulations of α -gliadin across ethanol–water mixtures with varying concentration (0%, 30%, 70%, and 100% v/v ethanol) have provided a detailed molecular picture of gliadin rearrangement (Kumar and Li, 2025). As ethanol concentration increased, α -gliadin exhibited global conformational expansion while largely preserving its α -helix and β -sheet content, in good agreement with x-ray scattering and CD measurements. Moreover, solvation shell analyses demonstrated that ethanol-induced conformational expansion of α -gliadin arises not from secondary-structure disruption, but from a progressive remodeling of the hydration shell. Structured water layers that stabilize compact conformations in aqueous environments were selectively displaced by ethanol, while water remained localized around polar and charged residues. Together, these results indicate that solvent-induced conformational expansion can occur without major disruption of secondary structure, and is instead governed by remodeling of the protein hydration shell.

While atomistic simulations clarify how solvent environments reshape individual protein conformations, protein solubility fundamentally emerges from inter-protein interactions. To bridge this gap, CG simulations can be applied to directly probe solvent-dependent aggregation behavior in multi-protein systems. Recent CG studies of soy 2S albumin and 19 kDa maize zein demonstrated that solvent quality drives distinct aggregation pattern across different seed storage protein classes (Kwon et al., 2025). Specifically, 2S soy albumins exhibited enhanced clustering in less polar 70% ethanol (w/w) environments, driven primarily by electrostatic and polar

interactions, whereas 19 kDa zeins showed the opposite trend, aggregating preferentially in aqueous conditions through hydrophobic residue-mediated contacts. These contrasting aggregation patterns highlight that solvent quality modulates protein-protein interactions in a class-specific manner, which cannot be inferred from monomer-level behavior alone.

Beyond solvent composition, the presence of co-solutes can also directly reshape protein conformations through specific non-covalent interactions. Recently, the interaction between β -lactoglobulin and three food grade solutes with different chemical nature (galactooligosaccharides (hydrophilic), sucrose fatty acid esters (amphiphilic), and casein phosphopeptides (charged)) was investigated (Song et al., 2025). AA-MD showed that ligand chemistry strongly governs protein-ligand interaction modes. Amphiphilic ligands tended to insert deeply into hydrophobic cavities, whereas more hydrophilic ligands preferentially interacted at pocket entrances or protein surfaces, leading to enhanced exposure of aromatic residues. This increased exposure was correlated with the red shifts observed in UV-Vis and fluorescence spectra. Despite these ligand-dependent differences in the surface exposure, DSSP analyses consistently demonstrate that the secondary structure of β -lactoglobulin remained highly stable over MD trajectories, suggesting that such interactions primarily involve local tertiary rearrangements.

CG simulations have also been applied to resolve solute-driven self-assembly and interfacial behavior in multicomponent systems. For instance, CG-MD simulations revealed that zein, epigallocatechin-3-gallate (EGCG), and polyethylene glycol (PEG) undergo progressive self-assembly in aqueous environments, forming protein aggregates that remained intact without dissociation (Wang et al., 2025). The aggregation process was accompanied by increased packing density, PEG-mediated entanglement, and convergence of multimer-level geometries.

When the pre-formed aggregate was transferred to an oil–water interface, it underwent only minor size adjustments associated with interfacial reorientation but rapidly stabilized within ~500 ns, maintaining their structural integrity. These results demonstrated that zein–EGCG–PEG nanoparticles achieve stability through polymer-assisted organization, and retain their integrity upon interfacial adsorption, supporting their suitability as emulsion stabilizers. Together, these examples illustrate that atomistic simulations elucidate solvent- and solute-driven conformational mechanisms at the molecular level. In contrast, CG simulations capture emergent aggregation, collective organization, and interfacial stability that arise only at the multimeric scale.

1.3.2. Process-induced structural evolution

Food processing directly imposes perturbations such as high pressure or temperature on proteins, driving structural evolution along the processing history. AA-MD has been applied to reveal such structural fluctuations upon various processing conditions. For instance, in studying the high hydrostatic pressure processing of egg ovalbumin, AA-MD was applied to investigate the effect of pressure on the protein conformation (Yang et al., 2024). Simulation identified pressure-sensitive regions, particularly helix–loop segments and selected β -sheets that underwent pronounced conformational rearrangements under high pressure. These pressure-induced changes were accompanied by a monotonic decrease in radius of gyration, suggesting that high pressure induces molecular compaction, in contrast to the expanded conformations typically associated with thermal unfolding. This example illustrates that pressure-induced structural responses follow mechanistic pathways distinct from thermal unfolding, underscoring the need for processing-specific molecular interpretation.

While AA-MD captures localized structural responses, many processing phenomena involve multimeric reorganization that extends beyond atomistic length scales. In a recent study on the effect of ultrasound treatment on walnut protein isolate (Zhang et al., 2025), square wave pressure emulating the ultrasound wave was applied in both AA and CG simulation. Across multiple structural descriptors in AA-MD simulations, ultrasound consistently induced partial destabilization and reorganization of the protein structure rather than complete unfolding. Atomistic simulations further identified an intermediate frequency (500 MHz) at which structural perturbation was maximized, providing a rational basis for subsequent CG-MD simulations. CG-MD simulations conducted at this frequency revealed a decreased SASA of walnut protein aggregate, indicating a more compact structure under ultrasound treatment, linking microscale aggregation behavior to functionality differences observed in the study. This multiscale strategy demonstrates how AA-MD can identify processing-relevant perturbation windows, which can then be propagated to CG models to capture aggregation-level consequences.

In another study on heat-treated zein systems (Devnani et al, 2024), CG-MD showed that heating promotes chain extension and alignment, while unheated zeins formed globular conformation. During subsequent cooling, these aligned conformations became kinetically trapped, leading to the formation of fiber-like protein networks. This temperature-dependent reorganization was accompanied by a decrease in SASA, indicating increased burial of hydrophobic regions making contacts, despite overall molecular elongation. Tensile pulling simulations further suggested that such orientation can withstand higher pulling stress compared to unheated aggregate (Devnani et al, 2024). Such fiber-like orientation of heated zein provides molecular insights for increased network cohesion and texture-forming properties of zein upon thermal process. These studies again highlight the distinct roles of AA and CG-MD simulation in

food protein studies, where detailed structural responses are resolved via AA-MD while reorganization events hinting emergent functional properties can be captured through CG-MD.

1.3.3. Bioactive ligand interactions and redistribution

Bioactive ligands and their association with protein constitute important application domains for food proteins, underpinning the development of functional foods and delivery systems. Accordingly, this section focuses on studies in which bioactive ligands were analyzed in terms of binding modes, spatial distribution, and release behavior relative to protein structures.

AA-MD simulations have been widely employed to resolve ligand binding modes, stability, and residue-level interaction patterns within individual protein matrices. For example, detailed analyses revealed that resveratrol binds stably within a hydrophobic β -sheet cavity of peanut arachin primarily through van der Waals interactions (Yang et al., 2025). Binding energetics were further quantified using MM/PBSA analysis, which yielded a moderately favorable binding free energy ($\Delta G \approx -19 \text{ kcal mol}^{-1}$) for the resveratrol–protein complex. Residue-level decomposition identified aromatic and hydrophobic residues, including Phe405, Phe464, and Val426, as major contributors, with additional stabilization provided by polar residues such as His425. In a study investigating the diffusion kinetics of ferulic acid and epicatechin in the presence of β -casein, the diffusion coefficient of the two ligands were first measured from dialysis escape (Condict et al., 2024). The presence of β -casein markedly reduced the diffusion coefficients of both phenolic compounds with sharper decrease in epicatechin. Umbrella sampling coupled with steered-MD simulation revealed stronger binding for epicatechin than for ferulic acid (-6.7 and $-4.6 \text{ kcal mol}^{-1}$, respectively), accompanied by more complex dissociation pathways due to ligand entanglement with β -casein. These studies

demonstrate how all-atom MD, when combined with appropriate free-energy frameworks, can quantitatively connect binding energetics and dissociation pathways to ligand retention and release behavior in food protein–ligand systems.

However, most AA-MD studies on ligand-protein interaction remain restricted to monomer-level descriptions, motivating CG consideration. Recent CG-MD studies have begun to explicitly resolve the fate of bioactive ligands during the formation of protein-ligand composites. For example, in a study involving black rice anthocyanins encapsulated within bovine serum albumin (Zhao et al., 2025), AA-MD simulations first revealed that anthocyanins initially associate with multiple binding sites on the protein surface through hydrogen bonding and van der Waals interactions. Subsequent CG-MD simulation demonstrated protein–protein aggregation upon ethanol addition, driving protein self-assembly. While anthocyanin was not explicitly included in the CG-MD simulation, these results suggest that anthocyanins can be kinetically trapped within the emerging protein matrix. Notably, the study suggests that ligand encapsulation and release behavior in food protein systems are governed not solely by equilibrium binding affinity, but by the temporal coupling between ligand adsorption and protein self-assembly during processing.

In a study where CG-MD was employed without AA consideration, mixed κ -casein/ β -casein self-assembled in water and adopted loose structures containing water-filled cavities (Gao et al., 2025). Interestingly, even when curcumin was initially placed within these cavities, spontaneous retention was not observed but redistributed on the surface of the assembly. This result underscores the importance of multiscale simulation strategies. Protein–ligand associations observed at the coarse-grained level reflect effective interactions defined within a reduced bead representation. As such, validation at atomistic resolution is required to determine whether these

apparent associations correspond to physically meaningful binding modes and realistic interaction energetics. Accordingly, interpretation of ligand retention or encapsulation solely from CG-level association should be treated with caution unless supported by atomistic or experimental validation. In this perspective, multiscale strategies are not merely complementary but necessary to avoid overinterpretation of either atomistic binding or coarse-grained association alone in studying ligand-protein interaction.

1.4. Practical limitations and methodological caveats in applying MD

Despite the growing popularity of MD simulations in food science, several limitations must be considered when interpreting their results or designing simulation-based studies.

1.4.1. Force-field-consistent simulation protocols

It should be noted that correct realizations of force fields such as CHARMM and Amber require choosing simulation parameters identical or sufficiently similar to those used in the parameterization of the force field. For example, the CHARMM36m protein force field (Huang et al., 2017) is typically used with a smooth truncation (cutoff) of the Lennard-Jones interactions from 10 to 12 Å, long-range electrostatics (typically by the PME method), rigid TIP3P_{mod} water, rigid covalent bonds to hydrogen, and exclusions on 1–3 atoms. Such simulation engine parameters should be thought of as part of the force field. Simulations omitting any force field terms (e.g., cross-terms or special Lennard-Jones parameters), using different exclusion schemes, water models, or cutoffs cannot be guaranteed yield the intended results. Furthermore, users should avoid combining force fields developed under different parameterization philosophies, such as using an Amber-based protein model with a ligand parameterized by CGenFF. Although

such combinations may appear technically feasible, they can lead to significant inconsistencies due to differences in charge derivation, dihedral angle definitions, and water model assumptions. Mixing incompatible force fields undermines the internal physical coherence of the system and may produce artifacts that are difficult to trace.

1.4.2. Availability and reliability of structural information

MD simulations require an initial 3D structure. In food protein systems, where crystallographic or NMR data is often unavailable, prediction methods such as homology modeling or *ab-initio* modeling tools (e.g., AlphaFold series) can assist in structure acquisition. Nevertheless, these models should be critically assessed, particularly for proteins with disordered regions, low sequence homology, or multimeric cases. For instance, it was reported that the dimer of shrimp myosin dimer predicted from AlphaFold multimer generates unreasonable conformation at the tail (Zhang et al., 2024). In such case, homology modeling using existing experimental structure would largely be preferred. Moreover, post-translational modifications, such as cleavage, signal peptide removal, disulfide linkage, or glycosylation should also be considered if the relevant information is available. Furthermore, from our literature review, most of the simulation study in food protein relies on monomer structures. This could be potentially misleading and interpreted with caution, especially for pulse proteins, where their major proteins legumin and vicilin assume hexameric and trimeric conformation, respectively.

In addition, the generation of parameters for unconventional small molecules, such as EGCG could also be challenging. Automated protocols such as cgenff or databases for published parameters like MAD are available to meet such requirements. In the case of cgenff, the penalty scores for generated parameters are demonstrated, which can be further parameterized using

tools such as ffparam. Lastly, free energy calculation methods such as MM/PBSA, umbrella sampling, or PMF are predicated on the assumption of a well-defined and stable binding pose or reaction coordinate. In food protein systems, where binding pockets are often unknown or highly flexible, the application of these techniques is discouraged unless prior structural information, either experimental or through validated docking, is available.

1.4.3. Stochastic variability and sampling convergence

While MD trajectories are generated using deterministic equations of motion, the assignment of initial velocities from the Maxwell–Boltzmann distribution introduces stochastic variation. Especially in short simulations, this can affect trajectory outcomes. Therefore, replicate simulations with different random seeds, or extended trajectories, are advised to ensure convergence and reproducibility. This stochasticity becomes increasingly problematic as the precision of the question increases. For instance, it could be misleading to identify ligand binding pockets or compute binding free energies based solely on a single MD run or molecular docking result. Moreover, MD simulation is fundamentally a sampling technique aiming to reproduce thermodynamic equilibrium. In theory, all available states should be sampling sufficiently to accurately provide the insight we expect from experiments. While this is typically not feasible due to limited time window, users should be aware that the seemingly stable results obtained (e.g., RMSD convergence) may simply reflect trapping in a local free-energy basin rather than true equilibration.

1.4.4. Limitations of Martini coarse-grained models and strategies for refinement

CG simulations introduce new limitations beyond those of AA models. In the case of the Martini 3 force field, it has two inherent limitations: (1) an oversimplified interaction landscape due to its low-resolution bead representation, and (2) limited transferability that often requires system- and environment-specific reparameterization. Martini sacrifices atomistic detail and relies on surrogate interaction terms fitted to experimental trends such as octanol–water partitioning free energies. Because the Martini family uses spherical beads and isotropic interaction potentials, the directionality of key interactions, such as hydrogen bonds or aromatic T-shaped contacts, is not captured. As a consequence of the missing directionality, the model cannot reliably represent protein folding or subtle secondary-structure rearrangements. Moreover, the effective timescale in CG simulations does not directly match physical time, for instance diffusion rates are often exaggerated, requiring correction factors (Martini speed-up factor ≈ 4) (Marrink et al., 2007).

While the oversimplification problem is rather inherent property of CG-modeling and philosophy, the second limitation could be refined using AA simulation. First, AA simulations can be used to refine bonded parameters. When building or adjusting a CG mapping for proteins or small molecules, AA trajectories naturally provide distributions for bond lengths, angles, and dihedrals. Classical CG parameterization matches not the atomistic geometries themselves but the probability distributions (via Boltzmann inversion), ensuring that the CG bonded landscape reproduces the correct flexibility. Moreover, the Martini protein model incorporates structural-bias terms that stabilize the tertiary structure. This two layered architecture can, in principle, benefit from reference AA trajectories, because different proteins can require different force constants choices (Korshunova et al., 2024). In practice, the optimal elastic-network strength is

protein- and region-dependent; flexible linkers or intrinsically disordered segments often require weaker or separate restraints (Thomassen et al., 2024). AA simulations thus serve as a structural reference for calibrating or weakening these biasing potentials. Such AA-guided refinement does not eliminate the fundamental limitations of coarse-graining, but can substantially improve the physical consistency and interpretability of CG simulations in food protein systems.

1.5. Conclusion

Over the past five years, the application of molecular dynamics simulations in food protein research has grown substantially. However, MD simulations are often treated as black-box, routine tools, which limits their interpretability and refinement space in complex and nonstandard systems. Accordingly, this review discusses the mechanics of MD simulations, force-field architectures, and key methodological assumptions that influence simulation behavior. This information is intended to help readers assess simulation setups, identify potential sources of error, and make informed adjustments when applying MD to food protein systems.

This review further highlights the multiscale nature of food protein systems, for which no single simulation resolution can fully capture their structural and functional behavior. All-atom and coarse-grained molecular dynamics simulations therefore serve distinct yet complementary roles in food protein research. Atomistic simulations provide mechanistic insight into solvation dynamics, residue-level interactions, and ligand binding specificity, whereas coarse-grained simulations capture emergent collective phenomena such as aggregation, phase separation, and interfacial stabilization at multimeric length scales. While many studies still employ atomistic or coarse-grained simulations separately, recent work increasingly adopts hybrid and multiscale strategies to connect molecular mechanisms with collective reorganization. These hybrid

approaches offer a practical way to address the intrinsic multiscale nature of food protein systems while accounting for the limitations of each modeling resolution.

Finally, the reliability of simulation-based results depends on careful attention to force-field consistency, sampling, and the limits of coarse-grained representations, as discussed in this review. When these conditions are met, MD simulations provide a physically grounded framework for interpreting food protein behavior across multiple length scales. Within this framework, molecular-level insights can be coherently related to emergent structural and functional outcomes relevant to food protein systems.

Acknowledgement

This is Contribution No. 26-142-J from the Kansas Agricultural Experiment Station. This work was in part supported by the USDA National Institute of Food and Agriculture Hatch project (Grant Accession No. 7003330).

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>

Abraham, M., Alekseenko, A., Andrews, B., Bauer, P., Bergh, C., Bird, H., Briand, E., Brown, A., Chen, Y., Doijade, M., Fiorin, G., Fleischmann, S., Gorelov, S., Gouaillardet, G., Gray, A., Jalalypour, F., Johansson, P., Kutzner, C., Lazarski, G., ... Lindahl, E. (2025).

GROMACS 2026.rc manual (v2026-rc).

Zenodo. <https://doi.org/10.5281/zenodo.17944198>

Alessandri, R., Souza, P. C., Thallmair, S., Melo, M. N., De Vries, A. H., & Marrink, S. J.

(2019). Pitfalls of the Martini model. *Journal of Chemical Theory and*

Computation, 15(10), 5448-5460. <https://doi.org/10.1021/acs.jctc.9b00473>

Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or

temperature. *The Journal of Chemical Physics*, 72(4), 2384-2393.

<https://doi.org/10.1063/1.439486>

Antolínez, S., Jones, P. E., Phillips, J. C., & Hadden-Perilla, J. A. (2024). AMBERff at scale:

Multimillion-atom simulations with AMBER force fields in NAMD. *Biophysical*

Journal, 123(3), 422a. <https://doi.org/10.1021/acs.jcim.3c01648>

Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta (BBA)-*

Bioenergetics, 1767(9), 1073-1101. <https://doi.org/10.1016/j.bbabi.2007.06.004>

Berendsen, H. J., Postma, J. V., Van Gunsteren, W. F., DiNola, A. R. H. J., & Haak, J. R. (1984).

Molecular dynamics with coupling to an external bath. *The Journal of Chemical*

Physics, 81(8), 3684-3690. <https://doi.org/10.1063/1.448118>

Bernetti, M., & Bussi, G. (2020). Pressure control using stochastic cell rescaling. *The Journal of*

Chemical Physics, 153(11). <https://doi.org/10.1063/5.0020514>

Best, R. B. (2019). Atomistic force fields for proteins. In M. Bonomi & C. Camilloni (Eds.),

Biomolecular Simulations: Methods and Protocols (pp. 3-19). Springer.

https://doi.org/10.1007/978-1-4939-9608-7_1

- Brasnett, C., & Marrink, S. J. (2025). MartiniGlass: A tool for enabling visualization of coarse-grained Martini topologies. *Journal of Chemical Information and Modeling*, 65(7), 3137-3141. <https://doi.org/10.1021/acs.jcim.4c02277>
- Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoseck, M., Im, W., ... Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), 1545-1614. <https://doi.org/10.1002/jcc.21287>
- Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1). <https://doi.org/10.1063/1.2408420>
- Cao, C., Zhu, Z., Liang, X., Kong, B., Xu, Z., Shi, P., Li, Y., Ji, Y., Ren, Z., & Liu, Q. (2023). Elucidation of interactions between myofibrillar proteins and κ -carrageenan as mediated by NaCl level: Perspectives on multiple spectroscopy and molecular docking. *International Journal of Biological Macromolecules*, 248, 125903. <https://doi.org/10.1016/j.ijbiomac.2023.125903>
- Carballo-Pacheco, M., & Strodel, B. (2016). Advances in the simulation of protein aggregation at the atomistic scale. *The Journal of Physical Chemistry B*, 120(12), 2991-2999. <https://doi.org/10.1021/acs.jpcb.6b00059>
- Case, D. A., Aktulga, H. M., Belfon, K., Cerutti, D. S., Cisneros, G. A., Cruzeiro, V. W. D., Forouzes, N., Giese, T. J., Götz, A. W., Gohlke, H., Izadi, S., Kasavajhala, K., Kaymak, M. C., King, E., Kurtzman, T., Lee, T.-S., Li, P., Liu, J., Luchko, T., ... Merz, K. M., Jr. (2023). AmberTools. *Journal of Chemical Information and Modeling*, 63(20), 6183-6191. <https://doi.org/10.1021/acs.jcim.3c01153>

- Cen, C., Yu, X., Shentu, X., Miao, S., & Wang, Y. (2026). Application of protein modeling and molecular simulation techniques in the field of food microbial contamination: principles, research status and future prospects. *Critical Reviews in Food Science and Nutrition*, 66(1), 129-144. <https://doi.org/10.1080/10408398.2025.2519709>
- Chakraborty, S., Wagh, K., Gnanakaran, S., & López, C. A. (2021). Development of Martini 2.2 parameters for N-glycans: A case study of the HIV-1 Env glycoprotein dynamics. *Glycobiology*, 31(7), 787-799. <https://doi.org/10.1093/glycob/cwab017>
- Chen, C. C., Hsieh, J. F., & Kuo, M. I. (2023). Insight into the processing, gelation and functional components of tofu: A review. *Processes*, 11(1), 202. <https://doi.org/10.3390/pr11010202>
- Chen, D., Zhu, X., Ilavsky, J., Whitmer, T., Hatzakis, E., Jones, O. G., & Campanella, O. H. (2021). Polyphenols weaken pea protein gel by formation of large aggregates with diminished noncovalent interactions. *Biomacromolecules*, 22(2), 1001-1014. <https://doi.org/10.1021/acs.biomac.0c01753>
- Condict, L., Elliott, S., Hung, A., Ashton, J., & Kasapis, S. (2024). Interfacing β -casein–phenolic compound interactions via molecular dynamics simulations with diffusion kinetics in delivery vehicles. *Food Chemistry*, 435, 137595. <https://doi.org/10.1016/j.foodchem.2023.137595>
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19), 5179-5197. <https://doi.org/10.1021/ja955032e>

- Crowley, M. F., Williamson, M. J., & Walker, R. C. (2009). CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *International Journal of Quantum Chemistry*, 109(15), 3767-3772. <https://doi.org/10.1002/qua.22372>
- 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, 10089-10089. <https://doi.org/10.1063/1.464397>
- Darré, L., Machado, M. R., Brandner, A. F., González, H. C., Ferreira, S., & Pantano, S. (2015). SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *Journal of Chemical Theory and Computation*, 11(2), 723-739. <https://doi.org/10.1021/ct5007746>
- de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., & Marrink, S. J. (2013). Improved parameters for the Martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1), 687-697. <https://doi.org/10.1021/ct300646g>
- DeLano, W. L. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, 40(1), 82-92.
- Devnani, B., Tan, N. K., Sanders, J. M., Shelley, J. C., Moraca, F., & Grossmann, L. (2024). Molecular insights into the structure forming properties of zein and a rheological comparison with hordein. *Future Foods*, 10, 100503. <https://doi.org/10.1016/j.fufo.2024.100503>
- Duan, W., Qiu, H., Htwe, K. K., Wei, S., Liu, Y., Wang, Z., Sun, Q., Han, Z., & Liu, S. (2023). Changes in advanced protein structure during dense phase carbon dioxide induced gel

- formation in golden pompano surimi correlate with gel strength. *Frontiers in Sustainable Food Systems*, 7, 1189149. <https://doi.org/10.3389/fsufs.2023.1189149>
- Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., & Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of Molecular Modeling*, 15(9), 1093-1108. <https://doi.org/10.1007/s00894-009-0454-9>
- Eastman, P., Galvelis, R., Peláez, R. P., Abreu, C. R. A., Farr, S. E., Gallicchio, E., Gorenko, A., Henry, M. M., Hu, F., Huang, J., Krämer, A., Michel, J., Mitchell, J. A., Pande, V. S., Rodrigues, J. P., Rodriguez-Guerra, J., Simmonett, A. C., Singh, S., Swails, J., . . . Markland, T. E. (2024). OpenMM 8: Molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1), 109-116. <https://doi.org/10.1021/acs.jpcb.3c06662>
- Feller, S. E., Zhang, Y., Pastor, R. W., & Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics*, 103(11), 4613-4621. <https://doi.org/10.1063/1.470648>
- Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4), 566-579. <https://doi.org/10.1002/prot.340230412>
- Fu, H., Li, J., Yang, X., Swallah, M. S., Gong, H., Ji, L., Meng, X., Lyu, B., & Yu, H. (2023). The heated-induced gelation of soy protein isolate at subunit level: Exploring the impacts of α and α' subunits on SPI gelation based on natural hybrid breeding varieties. *Food Hydrocolloids*, 134, 108008. <https://doi.org/10.1016/j.foodhyd.2022.108008>

- Gao, J., Dong, R., Yang, S., Zhao, Z., & Zhang, Y. W. (2025). Coarse-grained molecular dynamics investigation on the interaction between κ -and β -casein aggregates and curcumin. *PloS One*, 20(7), e0328010. <https://doi.org/10.1371/journal.pone.0328010>
- Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., Domanski, J., Dotson, D. L., Buchoux, S., Kenney, I. M., & Beckstein, O. (2019). MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations (Technical Report No. LA-UR-19-29136). Los Alamos National Laboratory.
- Gray, S. K., Noid, D. W., & Sumpter, B. G. (1994). Symplectic integrators for large scale molecular dynamics simulations: A comparison of several explicit methods. *The Journal of Chemical Physics*, 101(5), 4062-4072. <https://doi.org/10.1063/1.467523>
- Grünewald, F., Alessandri, R., Kroon, P. C., Monticelli, L., Souza, P. C., & Marrink, S. J. (2022). Polyply: A python suite for facilitating simulations of macromolecules and nanomaterials. *Nature Communications*, 13(1), 68. <https://doi.org/10.1038/s41467-021-27627-4>
- Guan, X., Zhong, X., Lu, Y., Du, X., Jia, R., Li, H., & Zhang, M. (2021). Changes of soybean protein during tofu processing. *Foods*, 10(7), 1594. <https://doi.org/10.3390/foods10071594>
- Hansen, N., & Van Gunsteren, W. F. (2014). Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, 10(7), 2632-2647. <https://doi.org/10.1021/ct500161f>
- Hilpert, C., Beranger, L., Souza, P. C. T., Vainikka, P. A., Nieto, V., Marrink, S. J., Monticelli, L., & Launay, G. (2023). Facilitating CG Simulations with MAD: The MArtini Database

- Server. *Journal of Chemical Information and Modeling*, 63(3), 702-710.
<https://doi.org/10.1021/acs.jcim.2c01375>
- Höchtel, P., Boresch, S., Bitomsky, W., & Steinhauser, O. (1998). Rationalization of the dielectric properties of common three-site water models in terms of their force field parameters. *The Journal of Chemical Physics*, 109(12), 4927-4937.
<https://doi.org/10.1063/1.477104>
- Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3), 1695. <https://doi.org/10.1103/PhysRevA.31.1695>
- Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., & MacKerell, A. D. (2017). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1), 71-73.
<https://doi.org/10.1038/nmeth.4067>
- 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](https://doi.org/10.1016/0263-7855(96)00018-5)
- Izaguirre, J. A., Catarello, D. P., Wozniak, J. M., & Skeel, R. D. (2001). Langevin stabilization of molecular dynamics. *The Journal of Chemical Physics*, 114(5), 2090-2098.
<https://doi.org/10.1063/1.1332996>
- Jain, A., Park, I. H., & Vaidehi, N. (2012). Equipartition principle for internal coordinate molecular dynamics. *Journal of Chemical Theory and Computation*, 8(8), 2581-2587.
<https://doi.org/10.1021/ct3002046>
- Jia, Y., Cocker, C., & Sampath, J. (2025). Insights into protein unfolding under pH, temperature, and shear using molecular dynamics simulations. *Biomacromolecules*, 26(4), 2095-2105.
<https://doi.org/10.1021/acs.biomac.4c00991>

- Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11), 1859-1865. <https://doi.org/10.1002/jcc.20945>
- Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of The American Chemical Society*, 118(45), 11225-11236. <https://doi.org/10.1021/ja9621760>
- Joshi, S. Y., & Deshmukh, S. A. (2021). A review of advancements in coarse-grained molecular dynamics simulations. *Molecular Simulation*, 47(10-11), 786-803. <https://doi.org/10.1080/08927022.2020.1828583>
- 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>
- Kamberaj, H. (2020). *Molecular dynamics simulations in statistical physics: Theory and applications*. Springer. <https://doi.org/10.1007/978-3-030-35702-3>
- Ke, Q., Gong, X., Liao, S., Duan, C., & Li, L. (2022). Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations. *Journal of Molecular Liquids*, 365, 120116. <https://doi.org/10.1016/j.molliq.2022.120116>
- Kim, J. I., Kwon, J., Baek, I., & Na, S. (2016). Steered molecular dynamics analysis of the role of cofilin in increasing the flexibility of actin filaments. *Biophysical Chemistry*, 218, 27-35. <https://doi.org/10.1016/j.bpc.2016.08.002>
- Korshunova, K., Kiuru, J., Liekkinen, J., Enkavi, G., Vattulainen, I., & Bruininks, B. M. (2024). Martini 3 OliGōmers: A scalable approach for multimers and fibrils in

- GROMACS. *Journal of Chemical Theory and Computation*, 20(17), 7635-7645.
<https://doi.org/10.1021/acs.jctc.4c00677>
- Kroon, P. C., Grünewald, F., Barnoud, J., van Tilburg, M., Brasnett, C., Souza, P. C. T., Wassenaar, T. A., & Marrink, S. J. (2025). Martinize2 and Vermouth provide a unified framework for molecular topology generation. *eLife*, 12, RP90627.
<https://doi.org/10.7554/eLife.90627>
- Kumar, A., Yoluk, O., & MacKerell Jr, A. D. (2020). FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *Journal of Computational Chemistry*, 41(9), 958-970.
<https://doi.org/10.1002/jcc.26138>
- Kumar, N., & Li, Y. (2025). Solvent-dependent structural and dynamic behavior of α -gliadin in aqueous ethanol systems: Insights from molecular dynamics simulations. *Food Chemistry*, 497, 147044. <https://doi.org/10.1016/j.foodchem.2025.147044>
- Kumar, N., Sarma, H., & Sastry, G. N. (2022). Repurposing of approved drug molecules for viral infectious diseases: A molecular modelling approach. *Journal of Biomolecular Structure and Dynamics*, 40(17), 8056-8072. <https://doi.org/10.1080/07391102.2021.1905558>
- Kwon, H., Xu, Y., Xu, X., & Li, Y. (2025). Unraveling the physicochemical differences among Osborne protein classes via bioinformatics and AI. *Food Research International*, 221, 117322. <https://doi.org/10.1016/j.foodres.2025.117322>
- Lane, T. J., Shukla, D., Beauchamp, K. A., & Pande, V. S. (2013). To milliseconds and beyond: challenges in the simulation of protein folding. *Current Opinion in Structural Biology*, 23(1), 58-65. <https://doi.org/10.1016/j.sbi.2012.11.002>

- Latham, A. P., & Zhang, B. (2022). Unifying coarse-grained force fields for folded and disordered proteins. *Current Opinion in Structural Biology*, 72, 63-70.
<https://doi.org/10.1016/j.sbi.2021.08.006>
- Lee, J., Hitzenberger, M., Rieger, M., Kern, N. R., Zacharias, M., & Im, W. (2020). CHARMM-GUI supports the Amber force fields. *The Journal of Chemical Physics*, 153(3).
<https://doi.org/10.1063/5.0012280>
- Lemkul, J. A., Huang, J., Roux, B., & MacKerell Jr, A. D. (2016). An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications. *Chemical Reviews*, 116(9), 4983-5013.
<https://doi.org/10.1021/acs.chemrev.5b00505>
- Liwo, A., Ołdziej, S., Czaplewski, C., Kozłowska, U., & Scheraga, H. A. (2004). Parametrization of backbone— electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. *The Journal of Physical Chemistry B*, 108(27), 9421-9438.
<https://doi.org/10.1021/jp030844f>
- Luo, Y., & Roux, B. (2010). Simulation of osmotic pressure in concentrated aqueous salt solutions. *The Journal of Physical Chemistry Letters*, 1(1), 183-189.
<https://doi.org/10.1021/jz900079w>
- MacCallum, J. L., Hu, S., Lenz, S., Souza, P. C., Corradi, V., & Tieleman, D. P. (2023). An implementation of the Martini coarse-grained force field in OpenMM. *Biophysical Journal*, 122(14), 2864-2870. <https://doi.org/10.1016/j.bpj.2023.04.007>

- Machado, M. R., Barrera, E. E., Klein, F., Sónora, M., Silva, S., & Pantano, S. (2019). The SIRAH 2.0 force field: Altius, fortius, citius. *Journal of Chemical Theory and Computation*, 15(4), 2719-2733. <https://doi.org/10.1021/acs.jctc.9b00006>
- Mackerell Jr, A. D., Feig, M., & Brooks III, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25(11), 1400-1415. <https://doi.org/10.1002/jcc.20065>
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), 3696-3713. <https://doi.org/10.1021/acs.jctc.5b00255>
- Marrink, S. J., De Vries, A. H., & Mark, A. E. (2004). Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2), 750-760. <https://doi.org/10.1021/jp036508g>
- Marrink, S. J., Monticelli, L., Melo, M. N., Alessandri, R., Tieleman, D. P., & Souza, P. C. (2023). Two decades of Martini: Better beads, broader scope. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 13(1), e1620. <https://doi.org/10.1002/wcms.1620>
- 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>

- Martínez, L., Andrade, R., Birgin, E. G., & Martínez, J. M. (2009). PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry*, 30(13), 2157-2164. <https://doi.org/10.1002/jcc.21224>
- McGibbon, Robert T., Beauchamp, Kyle A., Harrigan, Matthew P., Klein, C., Swails, Jason M., Hernández, Carlos X., Schwantes, Christian R., Wang, L.-P., Lane, Thomas J., & Pande, Vijay S. (2015). MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8), 1528-1532. <https://doi.org/10.1016/j.bpj.2015.08.015>
- 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), e4792. <https://doi.org/10.1002/pro.4792>
- Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J. (2008). The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5), 819-834. <https://doi.org/10.1021/ct700324x>
- Nnyigide, O. S., & Hyun, K. (2023). Charge-induced low-temperature gelation of mixed proteins and the effect of pH on the gelation: A spectroscopic, rheological and coarse-grained molecular dynamics study. *Colloids and Surfaces B: Biointerfaces*, 230, 113527. <https://doi.org/10.1016/j.colsurfb.2023.113527>
- Ooi, T., Oobatake, M., Nemethy, G., & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proceedings of the National Academy of Sciences*, 84(10), 3086-3090. <https://doi.org/10.1073/pnas.84.10.3086>

- Papanikolaou, T. D., & Tsoulis, D. (2016). Assessment of numerical integration methods in the context of low Earth orbits and inter-satellite observation analysis. *Acta Geodaetica et Geophysica*, 51(4), 619-641. <https://doi.org/10.1007/s40328-016-0159-3>
- 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>
- Periole, X., Cavalli, M., Marrink, S. J., & Ceruso, M. A. (2009). Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition. *Journal of Chemical Theory and Computation*, 5(9), 2531-2543. <https://doi.org/10.1021/ct9002114>
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16), 1781-1802. <https://doi.org/10.1002/jcc.20289>
- Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C. R., Radak, B. K., Skeel, R. D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., ... Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, 153(4). <https://doi.org/10.1063/5.0014475>
- Piskorz, T. K., De Vries, A. H., & Van Esch, J. H. (2021). How the choice of force-field affects the stability and self-assembly process of supramolecular CTA fibers. *Journal of Chemical Theory and Computation*, 18(1), 431-440. <https://doi.org/10.1021/acs.jctc.1c00257>

- Poma, A. B., Cieplak, M., & Theodorakis, P. E. (2017). Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *Journal of Chemical Theory and Computation*, 13(3), 1366-1374.
<https://doi.org/10.1021/acs.jctc.6b00986>
- Qi, Y., Ingólfsson, H. I., Cheng, X., Lee, J., Marrink, S. J., & Im, W. (2015). CHARMM-GUI martini maker for coarse-grained simulations with the martini force field. *Journal of Chemical Theory and Computation*, 11(9), 4486-4494.
<https://doi.org/10.1021/acs.jctc.5b00513>
- Schindler, C. E. M., Baumann, H., Blum, A., Böse, D., Buchstaller, H.-P., Burgdorf, L., Cappel, D., Chekler, E., Czodrowski, P., Dorsch, D., Eguida, M. K. I., Follows, B., Fuchß, T., Grädler, U., Gunera, J., Johnson, T., Jorand Lebrun, C., Karra, S., Klein, M., ... Kuhn, D. (2020). Large-scale assessment of binding free energy calculations in active drug discovery projects. *Journal of Chemical Information and Modeling*, 60(11), 5457-5474.
<https://doi.org/10.1021/acs.jcim.0c00900>
- Seo, S., & Shinoda, W. (2018). SPICA force field for lipid membranes: Domain formation induced by cholesterol. *Journal of Chemical Theory and Computation*, 15(1), 762-774.
<https://doi.org/10.1021/acs.jctc.8b00987>
- Song, R., He, Y., Ding, Y., Li, F., & Han, W. (2025). Interaction mechanisms between β -lactoglobulin and food-grade molecules: Insights from multispectroscopy and molecular dynamics simulations. *Journal of Dairy Science*. <https://doi.org/10.3168/jds.2025-27555>
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C., Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-

- Seara, H., Reuter, N., ... 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/s41592-021-01098-3>
- Strodel, B. (2021). Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20), 167182.
<https://doi.org/10.1016/j.jmb.2021.167182>
- Sun, L., Wang, S., Tian, F., Zhu, H., & Dai, L. (2022). Organizations of melittin peptides after spontaneous penetration into cell membranes. *Biophysical Journal*, 121(22), 4368-4381.
<https://doi.org/10.1016/j.bpj.2022.10.002>
- Swails, J., Hernandez, C., Mobley, D. L., Nguyen, H., Wang, L. P., & Janowski, P. (2010). *ParmEd* [Computer software]. <https://github.com/ParmEd/ParmEd>
- Tao, Y., Xu, J., Zhang, N., Jiao, X., Yan, B., Zhao, J., Zhang, H., Chen, W., & Fan, D. (2024). Unraveling the binding mechanisms of transglutaminase and substrate subjected to microwaves: Molecular docking and molecular dynamic simulations. *Food Chemistry*, 443, 138568. <https://doi.org/10.1016/j.foodchem.2024.138568>
- Thomasen, F. E., Cuneo, M. J., Mittag, T., & Lindorff-Larsen, K. (2023). Conformational and oligomeric states of SPOP from small-angle X-ray scattering and molecular dynamics simulations. *Elife*, 12, e84147. <https://doi.org/10.7554/eLife.84147>
- Thomasen, F. E., Skaalum, T., Kumar, A., Srinivasan, S., Vanni, S., & Lindorff-Larsen, K. (2024). Rescaling protein-protein interactions improves Martini 3 for flexible proteins in solution. *Nature Communications*, 15(1), 6645. <https://doi.org/10.1038/s41467-024-50647-9>

- Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C., & Plimpton, S. J. (2022). LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271, 108171. <https://doi.org/10.1016/j.cpc.2021.108171>
- Tsanai, M., Frederix, P. W., Schroer, C. F., Souza, P. C., & Marrink, S. J. (2021). Coacervate formation studied by explicit solvent coarse-grain molecular dynamics with the Martini model. *Chemical Science*, 12(24), 8521-8530. <https://doi.org/10.1039/D1SC00374G>
- Urey, H. C., & Bradley Jr, C. A. (1931). The vibrations of pentatonic tetrahedral molecules. *Physical Review*, 38(11), 1969. <https://doi.org/10.1103/PhysRev.38.1969>
- Vanommeslaeghe, K., & MacKerell Jr, A. D. (2012). Automation of the CHARMM General Force Field (CGenFF) I: Bond perception and atom typing. *Journal of Chemical Information and Modeling*, 52(12), 3144-3154. <https://doi.org/10.1021/ci300363c>
- Verlet, L. (1967). Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1), 98. <https://doi.org/10.1103/PhysRev.159.98>
- Vuorela, T., Catte, A., Niemelä, P. S., Hall, A., Hyvönen, M. T., Marrink, S.-J., Karttunen, M., & Vattulainen, I. (2010). Role of lipids in spheroidal high density lipoproteins. *PLoS Computational Biology*, 6(10), e1000964. <https://doi.org/10.1371/journal.pcbi.1000964>
- Wacha, A. F., & Lemkul, J. A. (2023). Charmm2gmx: An automated method to port the CHARMM additive force field to GROMACS. *Journal of Chemical Information and Modeling*, 63(14), 4246-4252. <https://doi.org/10.1021/acs.jcim.3c00860>

- Wang, Q., Rao, Z., Jiang, L., Lei, X., Zhao, J., Lei, L., Zeng, K., & Ming, J. (2025). The assembly mechanism of Zein/EGCG/PEG nanoparticles in a water system and their adsorption behavior at the oil-water interface. *Food Chemistry*, 463, 141051. <https://doi.org/10.1016/j.foodchem.2024.141051>
- Wassenaar, T. A., Ingólfsson, H. I., Bockmann, R. A., Tieleman, D. P., & Marrink, S. J. (2015). Computational lipidomics with insane: A versatile tool for generating custom membranes for molecular simulations. *Journal of Chemical Theory and Computation*, 11(5), 2144-2155. <https://doi.org/10.1021/acs.jctc.5b00209>
- Williams-Noonan, B. J., Kamboukos, A., Todorova, N., & Yarovsky, I. (2023). Self-assembling peptide biomaterials: Insights from spontaneous and enhanced sampling molecular dynamics simulations. *Chemical Physics Reviews*, 4(2). <https://doi.org/10.1063/5.0142302>
- Yang, J., Kuang, H., Kumar, N., Song, J., & Li, Y. (2024). Changes of structure properties and potential allergenicity of ovalbumin under high hydrostatic pressures. *Food Research International*, 190, 114658. <https://doi.org/10.1016/j.foodres.2024.114658>
- Yang, Y., Gao, Q., Yue, S., Peng, X., Wang, N., Xin, J. L., Yu, M., Rao, J., & Xue, L. (2025). Investigating the interaction mechanisms between arachin and resveratrol: Utilizing multi-spectroscopy and computational chemistry. *Food chemistry*, 463, 141435. <https://doi.org/10.1016/j.foodchem.2024.141435>
- Yu, D., Li, H., Liu, Y., Yang, X., Yang, W., Fu, Y., Zuo, Y., & Huang, X. (2024). Application of the molecular dynamics simulation GROMACS in food science. *Food Research International*, 190, 114653. <https://doi.org/10.1016/j.foodres.2024.114653>

- Yu, Y., Krämer, A., Venable, R. M., Simmonett, A. C., MacKerell, A. D., Jr., Klauda, J. B., Pastor, R. W., & Brooks, B. R. (2021). Semi-automated optimization of the CHARMM36 lipid force field to include explicit treatment of long-range dispersion. *Journal of Chemical Theory and Computation*, 17(3), 1562-1580.
<https://doi.org/10.1021/acs.jctc.0c01326>
- Zhang, L., Wang, P., Yang, Z., Du, F., Li, Z., Wu, C., Fang, A., Xu, X., & Zhou, G. (2020). Molecular dynamics simulation exploration of the interaction between curcumin and myosin combined with the results of spectroscopy techniques. *Food Hydrocolloids*, 101, 105455. <https://doi.org/10.1016/j.foodhyd.2019.105455>
- Zhang, M., Zhu, Z., Pan, F., Zhou, Q., Zhao, L., & Zhao, L. (2025). Enhancing walnut protein isolate functionality with ultrasound treatment: An integrated experimental and molecular dynamics simulation study. *Food Hydrocolloids*, 163, 111135.
<https://doi.org/10.1016/j.foodhyd.2025.111135>
- Zhang, Y., & Sagui, C. (2015). Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI. *Journal of Molecular Graphics and Modelling*, 55, 72-84. <https://doi.org/10.1016/j.jmgm.2014.10.005>
- Zhang, Y., Chen, Z., Wang, Y., Dong, H., Sun, J., Li, J., & Mao, X. (2024). Molecular modelling studies reveal cryoprotective mechanism of L-Proline during the frozen storage of shrimp (*Litopenaeus vannamei*). *Food Chemistry*, 441, 138259.
<https://doi.org/10.1016/j.foodchem.2023.138259>
- Zhao, L., Fan, Q., Zhu, Z., Zhang, M., Zhou, N., Pan, F., Wang, O., & Zhao, L. (2025). Synthesis and characterization of anthocyanin-loaded bovine serum albumin

nanoparticles: unveiling the encapsulation mechanisms with computational insights. *Journal of the Science of Food and Agriculture*, 105(9), 4771-4781.

<https://doi.org/10.1002/jsfa.14199>

Chapter 2 - AlphaFold 2-based stacking model for protein solubility prediction and its transferability to seed storage proteins¹

Abstract

Accurate protein solubility prediction is crucial in screening suitable candidates for food application. Existing models often rely only on sequences, overlooking important structural details. In this study, a regression model for protein solubility was developed using both the sequences and predicted structures of 2,983 *E. coli* proteins. The sequence and structural level properties of the proteins were bioinformatically extracted and subjected to multilayer perceptron (MLP). Moreover, residue level features and contact maps were utilized to construct a graph convolutional network (GCN). The out-of-fold predictions of the two models were combined and fed into multiple meta-regressors to create a stacking model. The stacking model with support vector regressor (SVR) achieved R^2 of 0.502 and 0.468 on test and external validation datasets, respectively, displaying higher performance compared to existing regression models. Based on the improved performance compared to its based models, the stacking model effectively captured the strength of its base models as well as the significance of the different features used. Furthermore, the model's transferability was indirectly validated on a dataset of seed storage proteins using Osborne definition as well as on a case study using molecular dynamic simulation, showing potential for application beyond microbial proteins to food and agriculture-related ones

¹ This chapter was adapted from: 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, 134601. <https://doi.org/10.1016/j.ijbiomac.2024.134601>

Keywords: Deep learning, AlphaFold, protein solubility prediction, alternative food protein, seed proteins

2.1. Introduction

The food industry's concern about the sustainability of animal protein has grown significantly in the past decade, given the approximately 60% increase in meat demand over 20 years (Whitnall & Pitts, 2019). Recognizing the need for a more sustainable food system, researchers are exploring alternative protein sources, including plants, insects, and transgenic organisms (Westerhout et al., 2019). However, selecting food proteins requires considering specific techno-functional properties. Aqueous solubility, in particular, is vital as it directly impacts various protein functionalities and key food characteristics like mouthfeel and digestibility (Deng et al., 2020). Consequently, there is a growing demand for computational tools that can effectively predict or screen protein solubility, which would aid in selecting the optimal protein for a desired food product (Wood & Tavan, 2022).

Protein solubility is a concept that researchers often approach with different definitions. While various methods for measuring solubility exist, such as induced precipitation using ammonium sulfate (Trevino et al., 2008) or concentration via ultrafiltration (Evans et al., 2004), protein solubility can be broadly categorized into relative and maximum solubility. Relative solubility refers to the fraction of the protein that remains in the supernatant after centrifugation (Schnepf, 1992). On the other hand, maximum solubility pertains to the amount of a protein that can be dissolved in water without observable precipitation (Grossmann & McClements, 2023). While maximum solubility aligns with the thermodynamic definition of solubility, it poses practical challenges in measurements, especially in the case of food proteins where multiple protein types are present. Therefore, a significant portion of research conducted in the fields of food and agriculture science relies on the relative definition of solubility (Grossmann & McClements, 2023).

With the advent of deep learning, researchers have adapted various model architectures to effectively predict the physicochemical properties of proteins. Initially, the basic multilayer perceptron (MLP) was used to predict protease cleavage sites (Yang & Thomson, 2005) and to classify enzymes into six families (Weinert & Lopes, 2004). Improvements in MLP, such as dropout (Srivastava et al., 2014), Xavier initialization (Glorot & Bengio, 2010), and diverse activation functions, have further expanded its application in modeling the behaviors of proteins. Examples include classification of seed storage proteins by Osborne fractionation (Radhika & Rao, 2015), prediction of DNA and RNA binding sites (Arican & Gumus, 2023), and identification of amyloid protein (Li et al., 2020).

While MLP enables modeling complex, non-linear phenomena, architecture struggles to handle the spatial arrangements of amino acids in proteins. To address this, spatial models like graph convolutional network (GCN) have emerged as powerful tools. Originally proposed for node classification tasks (Kipf & Welling, 2016), GCN learns on graph-structured data by aggregating feature information from neighboring nodes. This allows the architecture to capture both local graph topology and node attributes and makes GCN effective for representing protein structures, where nodes represent amino acids and edges represent their interactions. The advancement of attention mechanism (Vaswani et al., 2017) further enhanced GCN models by allowing them to dynamically emphasize local patterns of the input data, thereby assigning levels of importance to different elements. This selective focus improved the learning of node representations, enhancing the model's performance and robustness on proteins. For instance, Baranwal et al. proposed a structure-based prediction model for protein-protein interaction sites using a graph attention network (Baranwal et al., 2022). Similarly, Cheng et al. utilized a multi-head, self-attention graph network to predict drug-target protein interactions (Cheng et al., 2022).

Another drawback of MLP is their inability to effectively handle context-aware data, which could be crucial for modeling protein sequences. The biological meaning of specific amino acids and motifs can heavily depend on their context within the sequence, which MLP struggles to capture without complex and impractical modifications. Recently, sequential models, particularly transformer-based protein language models, have been increasingly applied in the field. Pretrained language models (PLM) like evolutionary scale modeling (ESM) (Lin et al., 2023) and ProtBert (Brandes et al., 2022) have gained attention for their computational efficiency and exceptional performance in representing protein sequences (Zhang et al., 2024). Such PLMs provide embedding per each amino acid in a given protein sequence. These representations capture the contextual information of each amino acid within the sequence, similar to how embeddings in natural language processing capture the context of words in sentences (Tran et al., 2023). The embeddings generated from pre-trained language models have been applied to predict allergenic proteins (Du et al., 2024) and to identify alkaliphilic proteins (Susanty et al., 2024).

Extensive research and development have been devoted to protein solubility prediction models due to its importance across various fields. Given the complex nature of protein solubility, various machine learning and deep learning architectures have been widely employed in these models (Hebditch et al., 2017; Khurana et al., 2018; Rawi et al., 2018). Such tools can be classified based on several criteria; one of which is whether they perform binary classification or regression. In binary classification models, proteins are categorized as either soluble or insoluble, and the majority of the developed models fall into this category, including PROSO (Smialowski et al., 2012), PaRSnIP (Rawi et al., 2018), and PLM_Sol (Zhang et al., 2024). The prevalence of binary classifiers can be attributed to the fact that the most available datasets on

protein solubility, including the widely used TargetTrack database (Berman et al., 2017), define solubility as either soluble or insoluble. On the other hand, regression models provide continuous values for protein solubility, typically ranging from 0 to 1. GraphSol (Chen et al., 2021), and SoLart (Hou et al., 2020) are examples of regression models. Although fewer in number compared to binary counterparts, regression models provides significant engineering advantages as they provide insights into the predicted values' magnitude and offer easier outlier detection.

Another criterion for categorizing solubility prediction tools is the type of features they utilize. Sequence-based models exclusively rely on features derived from the protein sequence itself, including sequence length, amino acid composition, and isoelectric point, etc. For example, SVM-based CCSol combines six different sequence information such as secondary structure propensity, hydrophobicity, and hydrophilicity to define solubility parameters for each region by a sliding window of 21 amino acids (Agostini et al., 2014). Moreover, self-defined features based on domain knowledge are also employed. Another sequence-based model SOLpro (Magnan et al., 2009) utilizes different combinations of the ratio of amino acid residues, such as K-R or D-E, employing 35 sequence features in total. On the other hand, structure-based or structure-aware models leverage features extracted from the 3D structure of the proteins. These structural features can include the fraction of exposed residues or surface hydrophobicity. For instance, SoLart employs solubility-dependent statistical potentials derived from the type of residue, interatomic distance and angles as well as other structural features like solvent accessibility or the fraction of buried residues. While incorporating 3D information into the model has the potential to enhance its performance, the availability of crystal structures for proteins is not always guaranteed. As a result, most existing models were developed based solely on protein sequence information. However, the emergence of modern ab-initio structure

prediction tools like AlphaFold (AF) (Jumper et al., 2021) offers promise for addressing this limited structure availability.

Recently, PLMs have been increasingly integrated into solubility prediction tools. PLMs typically produce embeddings for each residue, allowing graph-based models like GCN to be applied with predicted protein structures. For instance, Wang et al. (2024) used embeddings generated from ProtTrans PLM as node features. Using AF-predicted structures, they constructed the binary GCN classifier DeepMutSol, which outperformed all state-of-the-art (SOTA) classifiers. Moreover, PLM-generated embeddings can be combined with other residue-based features. For example, Chen et al. (2023) utilized combined features from ESM-1v embeddings and evolutionary features like position-specific scoring matrix (PSSM) or hidden Markov model (HMM). With the predicted protein contact map from SPIDER3, they developed a solubility regressor HybridGCN, achieving ~10 % higher coefficient of determination compared to the previous SOTA model GraphSol on an external validation dataset. More recently, Li and Ming (2024) combined the blosum62 matrix with ESM-1b embeddings to construct a GCN solubility regressor, GATSol, which is the current SOTA regressor based on AF-predicted structures.

In this study, regression models for predicting protein solubility were trained by employing two distinct physicochemical feature sets generated by both protein sequences and AF computed structures. The first set encompassed features derived from the protein sequences and structures, including hydrophobic index (GRAVY) (Kyte & Doolittle, 1982) and hydrophobic patch area (Lijnzaad et al., 1996). The second set comprised protein contact maps and residue-level features, such as seven physicochemical properties of amino acids (AAPHY7) (Gligorijević et al., 2021). The sequence + structural feature set was processed using MLP, while the residual feature + contact map set was managed through GCN. Furthermore, to harness the information

provided by both approaches, a stacking method was employed, integrating the outputs of the MLP and GCN models. The performance of the stacking model was compared with existing regression models, and two regressors trained with ESM-2 pre-trained model. The relative importance of each features used in the base models MLP and GCN was computed from their saliency map. Additionally, the stacking model's transferability to plant proteins was indirectly validated using the Osborne definition of seed storage proteins. A case study involving coarse-grained molecular dynamic (CGMD) simulation was also conducted on two proteins with low and high predicted solubility.

2.2. Materials and methods

All datasets and the codes utilized to train and validate the models in this study are available at: <https://github.com/john94kwon/Stacking-model-for-solubility-predicition>

2.2.1. Protein solubility definition

In this study, relative solubility of *E. coli* proteins was utilized to develop prediction models. Solubility was defined as the proportion of supernatant obtained after centrifugation divided by the initial quantity of overexpressed protein (Niwa et al., 2009).

2.2.2. Datasets

2.2.2.1. *E. coli* dataset

A dataset consisting of solubility of proteins from the K-12 *E. coli* strain was utilized for model development. The dataset encompassed solubility measurements for 3147 proteins that were expressed in a cell-free system (Niwa et al., 2009). Using the provided entry information in the data, the corresponding protein structures were retrieved from the AF protein structure

database (<https://alphafold.ebi.ac.uk/>) as pdb format. In order to maintain structural reliability, structures with a pLDDT score below 70 for at least half of the total residues were discarded (Ros-Lucas et al., 2022). Furthermore, to mitigate biases arising from evolutionarily similar proteins, HHblits (Remmert et al., 2012) was utilized to filter out those with <25 % sequence similarity (Shapovalov et al., 2020). To maintain consistency, only structures computed by AF were collected even for the sequences with experimentally determined structures. A total of 2983 protein structure files remained for analysis. This data was further split into training and testing sets, with 75 % (2237) allocated for training and 25 % (746) reserved for testing purposes. 5-fold cross validation along with grid search was employed for hyperparameter tuning.

2.2.2.2. *S. cerevisiae* dataset

For external validation of trained models, another solubility dataset consisting of *S. cerevisiae* (brewer's yeast) proteins was selected. This dataset included the solubility measurements of 108 yeast proteins that were overexpressed in the same system as *E. coli* proteins above (Uemura et al., 2018). We specifically chose the *S. cerevisiae* dataset to compare the performance of our models to existing ones, as several prediction models have already undergone external validation with the dataset (Hou et al., 2020; Chen et al., 2021).

2.2.2.3. Seed storage protein dataset

To investigate the model's relevance in the solubility of plant proteins, a dataset comprising AF structures of a range of seed storage proteins was gathered. A dataset of 200 structures was compiled, covering 50 distinct structures per each Osborne classifications: albumin, globulin, prolamin, and glutelin. Specifically, the UniProt database was searched using the tag "seed storage protein" along with the Osborne classification (e.g., "seed storage protein glutelin"). The protein structures and sequences reviewed by UniProt were exclusively used for

albumin, globulin, and prolamin. In the case of glutelin class, due to limited data availability, both reviewed and automatically annotated proteins were utilized. The number of source organisms for each class was 28 for albumin, 29 for globulin, 13 for prolamin, and 28 for glutelin (Appendix Table A.2.). Moreover, the properties of five seed storage proteins from single sources, each belonging to a different Osborne class (2S albumin from mouse-ear cress, 11S legumin from pea, zein from maize, and glutelin from rice), were compared with those of human reference proteins. The statistical significance between the selected proteins was computed using analysis of variance (ANOVA) with the R statistical package (ver 3.5.0), followed by Duncan's multiple range test with a confidence level of 95 %.

2.2.3. Feature extraction for MLP model

To train the MLP model, a feature set comprising a range of protein descriptors utilized in bioinformatics was employed. The sequence-level features included molecular weight, aromaticity, instability index, hydrophobic index, aliphatic index, absolute charge per residue, and hydrophilic index. These features were extracted with biopython module ver 1.81 (Cock et al., 2009). In addition to these sequence features, an assortment of structural characteristics was extracted from the protein 3D structures computed from AF structures. Specifically, the QUILT software ver 1.3 (Lijnzaad et al., 1996) was employed to calculate the hydrophobic patch area, hydrophilic surface area, and the ratio of hydrophobic surface to total surface area. QUILT calculates the solvent-accessible surface area (SASA) of a protein by rolling a probe that emulates the van der Waals radius of a solvent molecule. The surface is defined by the path of the probe's center and the number of discrete sampling points. The software then identifies contiguous apolar surface areas, namely around sulfur and carbon atoms. In this study, a probe

radius of 1.4 Å and 252 sampling points were used, following the suggested parameters from the author of the software (<https://github.com/plijnzaad/quilt>). The hydrophilic surface area was determined by subtracting the hydrophobic patch area from the total SASA, while the ratio of hydrophobic patches was calculated by dividing the total hydrophobic patch area by the total SASA.

Furthermore, with the ChimeraX software ver 1.6.1, a range of additional structural features was generated. These features encompassed mean lipophilicity potential, mean coulombic potential, surface area, volume, solvent-accessible surface area, number of hydrogen bonding, helix propensity, coil propensity, strand propensity, and number of favorable contacts. Prior to the feature extraction, all pdb structures files were converted into pqr format files using the pbd2pqr plugin ver 3.6.1 (Dolinsky et al., 2004) with pH 7.0 under built-in CHARMM force field parameters. This conversion step was essential to account for the protonation states of the proteins, as they may significantly impact the coulombic potential or number of hydrogen bonding. Similar to QUILT's calculation of SASA, ChimeraX also determines the molecular surface of a protein, but it defines the surface by the trajectory of the probe sphere's outer edge instead of its center. This molecular surface is used to calculate the mean coulombic potential, mean lipophilicity potential, and molecular volume. The mean coulombic potential is calculated as the average electrostatic potential using atomic partial charges, coordinates, and distances from the surface. The mean lipophilicity potential is computed using atomic hydrophobicity values, coordinates, and distances from the surface. Molecular volume is determined by the volume enclosed by the molecular surface. The number of hydrogen bonds was calculated using default geometric criteria in ChimeraX (Mills & Dean, 1996). Favorable contacts were determined by subtracting the number of clashes from the total contacts. In ChimeraX, contacts

include all direct interactions (polar and nonpolar), while clashes are defined as unfavorable interactions where atoms are too close, considering their van der Waals radii. All feature extractions in ChimeraX were computed using default parameters. In sum, a total of 7 sequence features and 13 structural features were gathered. No exclusive feature selection was conducted, as the utilization of common feature selection methods such as Pearson correlation, K-best selection, and recursive feature elimination resulted in inferior performance (Appendix Table A.3.)

2.2.4. Feature extraction for GCN model

In a graph neural network, a graph is represented as a collection of interconnected vertices. These vertices, commonly referred to as nodes, establish the foundation of the graph, while the connections between them are represented as edges (Jia & Benson, 2020). When considering proteins, which are composed of linked residues, a similar graph-based representation can be applied. In this context, the amino acid residues of a protein are the nodes, and their interconnections, whether involved in physical contacts or actual peptide bonds, serve as the edges.

2.2.4.1. Node features for GCN model

In contrast to the features employed in the MLP model, where each feature denoted a characteristic of the entire protein, the features utilized in the GCN model represented the attributes of individual residues. Three distinct node features, including AAPHY7, Blosom62 (Henikoff & Henikoff, 1992), and the coordinates of the alpha carbon were utilized to capture the characteristics of each amino acid residue. Briefly, AAPHY7 is a list of seven physicochemical properties of each amino acids, which includes steric parameter, residue

hydrophobicity, residue volume, polarizability, isoelectric point, helix propensity, and sheet propensity (Meiler et al., 2001). Blosom62 is a substitution matrix derived from conserved regions of protein families. The higher score in the blosum matrix indicates favorable conservation, and physicochemical similarity between two residues.

2.2.4.2. Protein contact map (edge for GCN model)

In graph neural networks, the edge of a graph is often represented by its adjacency matrix $\mathbf{A}$ (Yuan et al., 2020), which is a $[n \times n]$ matrix defined by Equation (1). When dealing with proteins, their contact maps can be used to produce the adjacency matrix of the protein. The generation of contact maps involves calculating the distances between the alpha carbon residues and binarily mapping these distances based on a specified threshold. Such contact maps, by definition, are consistent with the adjacency matrix used in graph neural networks and can be employed as edge matrix. It should be noted that contact maps are equivalent of $\tilde{\mathbf{A}}$, the adjacency matrix added with identity matrix (Equation (2)). Typically, the threshold distance for protein contact maps falls within the range of 6 to 12 Å (Pollastri & Baldi, 2002). In the context of this research, multiple contact maps were created using python script, and after evaluation, the contact map generated with a 12 Å threshold exhibited the highest performance and was therefore selected as the optimal choice.

$$[\mathbf{A}]_{ij} = 1 \text{ if node } i \text{ and } j \text{ are connected, } = 0 \text{ otherwise} \quad (1)$$

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I} \quad (2)$$

2.2.5. MLP model structure

The structure and parameters of the constructed MLP model were displayed in Fig. 2.1

a. For each protein, the input to the model was a feature matrix, $\mathbf{X}$, consisting of 20 sequence-

and structure-derived features. Initially, $\mathbf{X}$ was fed into the input layer, and underwent a series of transformations as it passed through the hidden layers. These transformations were determined by the adjustable layer weights and activation functions, as defined by Equation (3). Moving toward the right side of the figure, the resulting vector from the third hidden layer was further processed at the output layer, where it was transformed into the predicted solubility. The hidden layers used in the MLP model consisted of 128, 64, and 32 weights, respectively. The LeakyReLU activation function (Xu et al., 2015) was applied to these hidden layers. Moreover, a dropout layer with the rate 0.2 was employed at the first hidden layer to alleviate overfitting, and batch normalization was utilized in all three hidden layers. To scale the predicted solubility within the range of 0 to 1, the sigmoid activation function was used at the output layer. The root mean square error (RMSE) and Adamax optimizer (Kingma & Ba, 2014) were used as the loss function and the optimizer, respectively.

$$\mathbf{H}^{(i+1)} = \sigma(\mathbf{H}^i \mathbf{W}^i + \mathbf{b}) \quad (3)$$

Where $\mathbf{W}^i$ denotes the trainable weight of the layer, $\mathbf{H}^i$ represents the transformed feature matrix at the i^{th} hidden layer, σ denotes the activation function, and $\mathbf{b}$ is the bias vector. At $i=0$, the feature matrix $\mathbf{X}$ is $\mathbf{H}^0$.

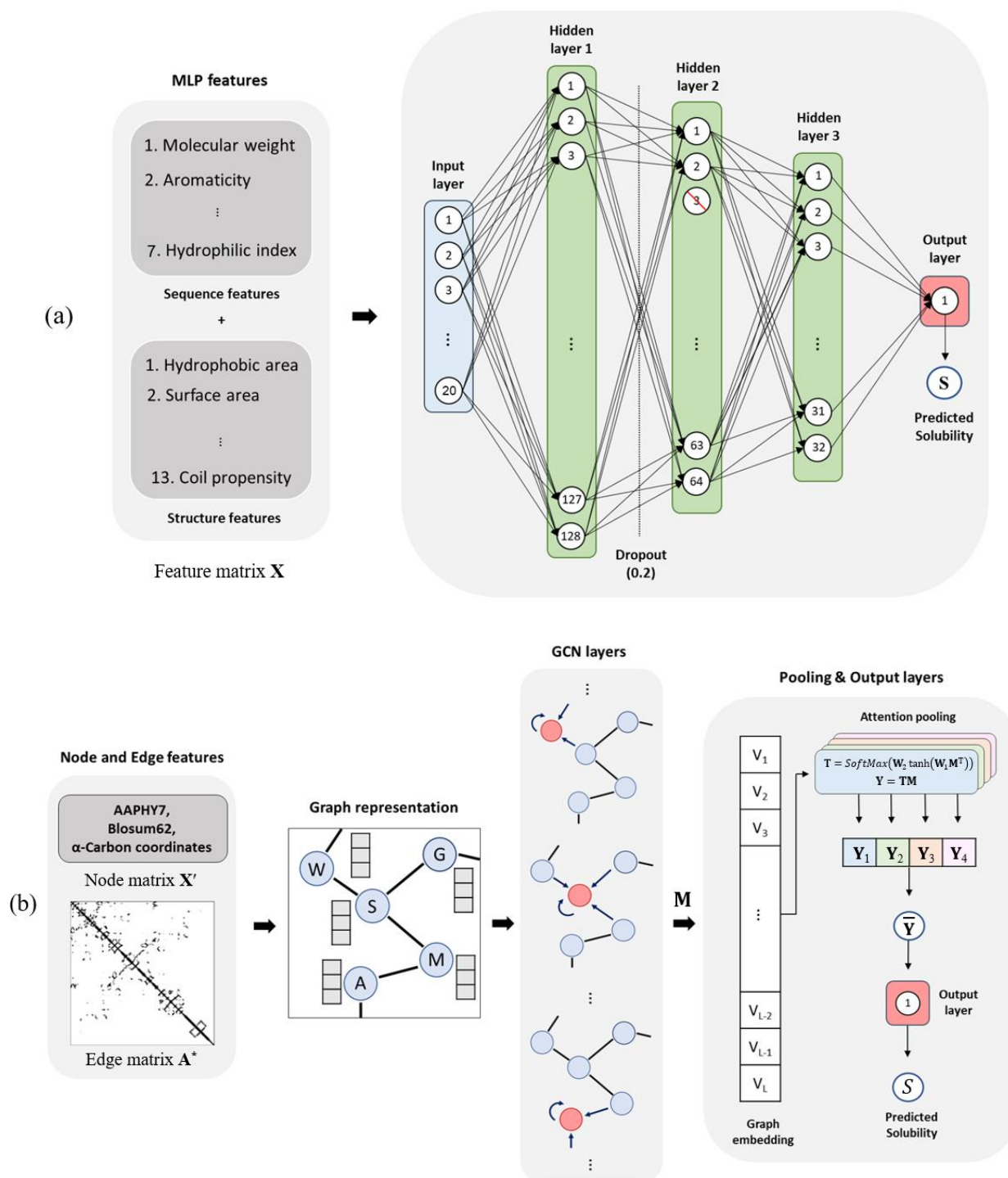


Figure 2.1. The structure of the MLP (a) and GCN (b) model constructed.

2.2.6. GCN model structure

The basic structure of the GCN model used in this experiment was adapted from the work of Chen et al (2021). The model architecture and its process were displayed in Fig. 2.1b. The protein graph was represented by the combination of the node feature matrix $\mathbf{X}'$ and the generated contact map $\tilde{\mathbf{A}}$. These two inputs were then fed into the GCN model, which consists of GCN layers and a pooling & output layer.

2.2.6.1. GCN layers

In the GCN layer, the node features of individual residues were updated based on the features of their neighboring residues as well as their own. This process is illustrated in the “GCN layer” of Fig. 2.1b and can be mathematically represented by Equation (4). The key difference between MLP and GCN stems from the additional element $\mathbf{A}'$, which corresponded to $\tilde{\mathbf{A}}$ normalized by the degree matrix $\mathbf{D}$ (Equation (5)). The degree matrix was used to normalize the adjacency matrix so that feature aggregation was not dominated by residues with a larger number of contacts (Kipf and Welling, 2016). This normalization balanced the contribution of neighboring residues and stabilized graph convolution across nodes with different degrees.

The normalized adjacency matrix $\mathbf{A}'$ encoded the connectivity between the nodes, and the matrix multiplication of the normalized adjacency matrix and node feature matrix, $\mathbf{A}'\mathbf{X}'$ represented the aggregation of neighboring node features at the first layer. The aggregated node features were then transformed by the trainable weight matrix $\mathbf{W}^i$, whose values were optimized during model training to learn the relationship between residue-level graph representations and protein solubility. Thus, unlike the MLP, which used protein-level feature matrices without explicit residue–residue connectivity, the GCN incorporated contact-map-based structural relationships between neighboring residues. The GCN model employed two hidden layers with

LeakyReLU activation functions, consisting of 256 and 64 weights, respectively. Batch normalization was applied to normalize the layers to mitigate overfitting.

$$\mathbf{H}^{(i+1)} = \sigma(\mathbf{A}'\mathbf{H}^i\mathbf{W}^i + \mathbf{b}) \quad (4)$$

$$\mathbf{A}' = \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}} \quad (5)$$

Where $\mathbf{A}'$ is the normalized adjacency matrix, $\mathbf{H}^i$ is the node feature matrix at the i^{th} layer, $\mathbf{W}^i$ is the trainable weights of the i^{th} layer, σ is the activation function, and $\mathbf{b}$ is the bias vector. At $i=0$, the node feature matrix $\mathbf{X}'$ is $\mathbf{H}^0$. $\mathbf{D}$ is degree matrix of contact maps, and $\tilde{\mathbf{A}}$ is the generated contact map.

2.2.6.2. Pooling & output layer

A practical issue with the GCN layer presented would be the variation in protein length. If a protein's length was L and had f number of node features, the size of $\mathbf{A}'$ and $\mathbf{X}'$ would be $[L \times L]$ and $[L \times f]$, respectively. This would make their multiplication an $[L \times f]$ matrix. If the size in the first and second GCN layers were set to be $[f \times p]$ and $[p \times q]$, respectively, then the resulting matrix $\mathbf{M}$ from passing two GCN layers would have the dimension of $[L \times q]$. Consequently, the matrix $\mathbf{M}$ would then have a variable size dependent on L .

To address this variability in protein size, the multi-head self-attention pooling mechanism proposed by Lin et al. was employed (Lin et al., 2017). In Equation (6), $\mathbf{T}$ represents the attention matrix generated by the self-attention pooling layer. Each row of $\mathbf{T}$ corresponds to one attention head, and each column corresponds to a residue in the protein sequence. Therefore, each element of $\mathbf{T}$ represents the relative contribution of a residue-level GCN representation to the pooled protein-level representation. The softmax function normalizes the attention scores across residues, allowing the model to assign larger weights to residues that are more informative. By multiplying $\mathbf{T}$ and $\mathbf{M}$, the residue-level representation matrix $\mathbf{M}$ is converted

into a fixed-size pooled representation $\mathbf{Y}$, regardless of the protein length L (Equation (7)). The number of attention heads was set to four. The four resulting feature vectors were averaged and passed through the output layer to produce the predicted solubility. A sigmoid activation function was used in the output layer.

$$\mathbf{T} = \text{SoftMax}(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{M}^T)) \quad (6)$$

$$\mathbf{Y} = \mathbf{T}\mathbf{M} \quad (7)$$

Where $\mathbf{T}$ is the attention matrix, $\mathbf{M}$ is the residue-level node representation matrix obtained from the GCN layers, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices used to calculate attention scores. $\mathbf{Y}$ represents the fixed-size pooled protein-level representation obtained by multiplying $\mathbf{T}$ and $\mathbf{M}$.

2.2.7. Benchmark models: MLP and GCN with ESM-2 embedding

To compare the performance of models derived from the features in Sections 2.2.3 and 2.2.4, another set of parameters was obtained using the pre-trained protein language model ESM-2. ESM-2 is the latest version of ESM and has outperformed many existing models in various structure forecasting tasks (Lin et al., 2023). Given the relatively small size of the training dataset and the number of features in section 2.2.3 (20 features), the esm2_t6_8M_UR50D model, which generates the lowest dimension embeddings (320 features), was selected. The generated 320 features per residues were averaged by the sequence length to ensure equal feature dimension $[1 \times 320]$. The generated sequence embeddings were then fed into the MLP structure described in section 2.2.5. Moreover, the embeddings were directly fed into the GCN model as node features without taking their average. In the case of the GCN model, the number of epochs was reduced from 10 to 5 to prevent model overfitting.

2.2.8. Stacking model

The two feature sets used for MLP and GCN might offer different perspectives on the specific protein. In the MLP model, features derived from the protein sequences and unmodified structures were utilized. Conversely, the GCN model trained on graph structures reconstructed from residual features and contact maps. To leverage the information provided by both approaches, different meta-regressors including linear regression (LR), support vector regressor (SVR), decision tree (DT), and k-nearest neighbor (KNN) were employed to combine the outcomes of the MLP and GCN models (Fig. 2.2). Specifically, the meta-regressors were trained using the out-of-fold cross validation results from each 5-fold cross-validation from individual base models to prevent data leakage and model overfitting.

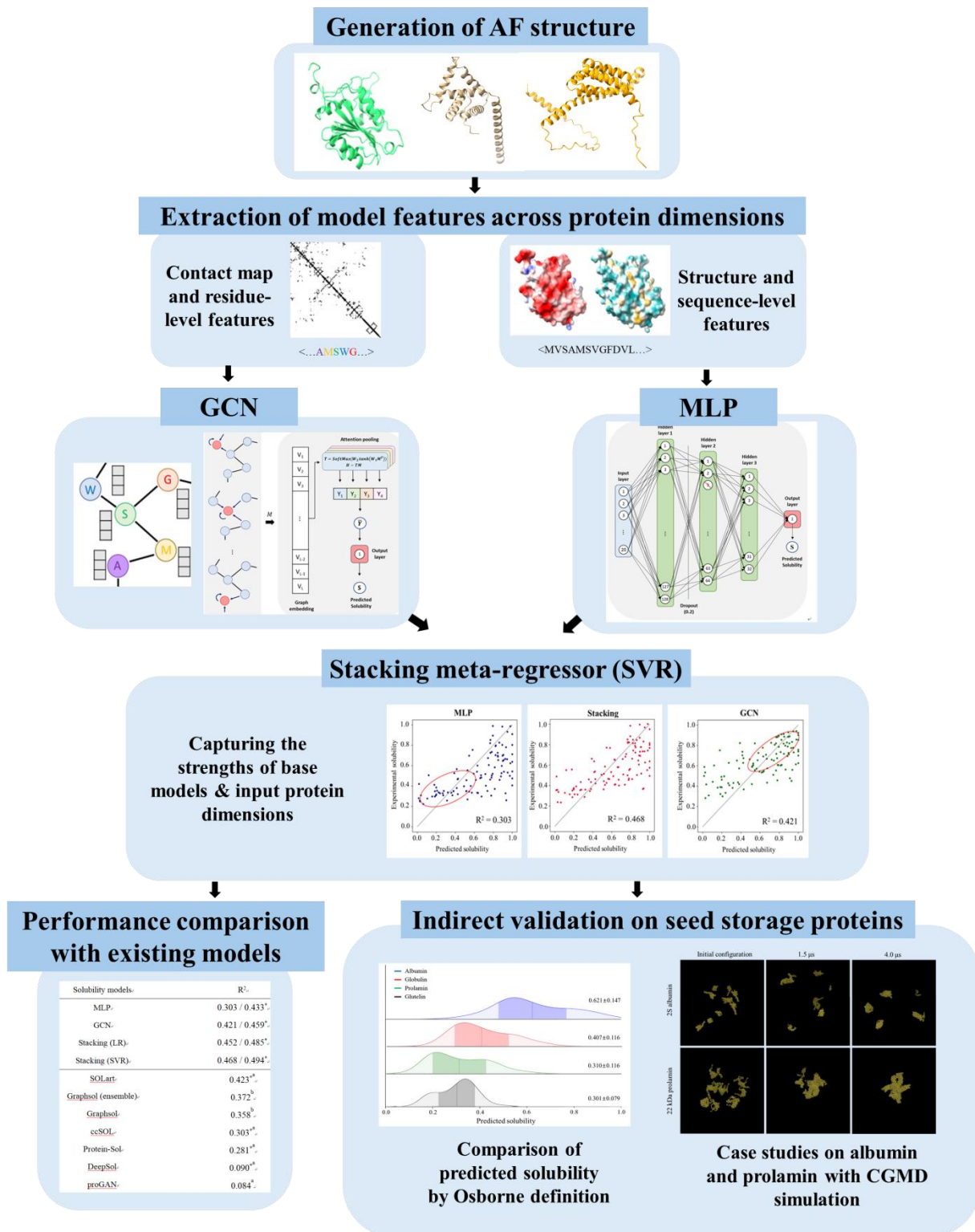


Figure 2.2. Overall workflow for the construction and validation of the stacking solubility model.

2.2.9. Model hyperparameters and evaluation metrics

To ensure optimal performance, all model hyperparameters were tuned using a 5-fold cross-validation technique on the training dataset. The specific hyperparameters, along with their respective values, can be found in Appendix Table A.4. In order to assess the predictive power of the models constructed, coefficient of determination (R^2) and root mean square error (RMSE) were utilized. In an extension to the prediction of continuous solubility values, a binary classification was also performed with threshold solubility value of 0.5. Proteins with solubility values above 0.5 were classified as soluble, while those below were classified as insoluble. The performance of the classification was assessed by accuracy, F1 score, AUC (area under curve), and MCC (Matthews correlation coefficient). For the external validation using the *S.cerevisiae* dataset, both the coefficient of determination and the square of Pearson's correlation were utilized.

2.2.10. Identification of influential features/residues

To assess the importance of features in the base models MLP and GCN, saliency maps of the trained models were generated with respect to the input features. Training datasets were fed into the trained models with PyTorch's `requires_grad` function enabled, allowing the measurement of gradients during backpropagation. The models were set to evaluation mode to negate the effects of dropout and batch normalization. The magnitude of the saliency for each feature was averaged across individual training data to determine the overall influence of each feature. In the case study on seed storage proteins, the relative importance of each residue for the GCN model was identified by summing the magnitudes of all features per residue. The most important regions of residues were selected with thresholds of 0.20 for 2S albumin and 0.15 for

22kda prolamin. The electrostatic potential maps for the two proteins were generated using the adaptive Poisson-Boltzmann solver (APBS) web server (Jurrus et al., 2018).

2.2.11. Case study: molecular dynamics simulation

The aggregation behavior of seed storage proteins was studied using Coarse-grained molecular dynamics (CG-MD) simulation using Martini 3 force field (Souza et al., 2021). AF predicted structures of 2S-albumin from soybean (Uniprot entry: P19594) and 22 kDa alpha zein prolamin from maize (Uniprot entry: P04700) were downloaded to prepare the initial coordinates and topologies of the proteins. Ten molecules of each protein were randomly placed in cubic water boxes using the gmx insert-molecules function in GROMACS software (Abraham et al., 2015). The box dimensions were $39.43 \times 39.43 \times 39.43 \text{ nm}^3$ for albumin and $45.54 \times 45.54 \times 45.54 \text{ nm}^3$ for prolamin, which were selected to approximate a protein concentration of 5% (w/w). The prepared atomistic systems were then converted to coarse-grained models using Martinize 2 (Kroon et al., 2025). The coarse-grained systems were solvated and charge-neutralized with sodium and chloride beads, followed by 5000 steps of energy minimization and 500 ns of equilibration under NVT and NPT ensembles. Production simulations were performed for 4 μs with a 20 fs time step. The temperature and pressure were maintained at 300 K and 1 atm using the velocity-rescaling thermostat and Parrinello–Rahman barostat, respectively. All simulations were conducted using GROMACS (version 2022.5) (Bauer et al., 2023), and snapshots of the trajectories were generated using VMD software (version 1.9.3) (Humphrey et al., 1996).

2.3. Results and discussion

2.3.1. Comparison of extracted features among datasets

Due to the different source organisms, significant variance among the proteins in each dataset could exist. To gauge the variation in the extracted feature values within each dataset, the average and standard deviation of each MLP feature were presented in Appendix Table A.5. As shown in the table, most values of the extracted MLP features were fairly consistent across the three datasets. The two most prominent differences were in the instability index (38.011 ± 10.269 for *E.coli*, 36.929 ± 7.648 for *S.cerevisiae*, and 61.186 ± 24.673 for seed storage) and SASA (16151.322 ± 8318.966 for *E.coli*, 16763.76 ± 6839.81 for *S.cerevisiae*, and 23930.924 ± 10064.226 for seed storage). However, even for these features, high variability within each set was observed, as indicated by their high standard deviations. For the GCN features, a direct comparison of proteins was not possible due to the nature of the graph-based model. However, the relative proportion of each amino acid was presented in Appendix Table A.6. Similar to the comparison of MLP features, no apparent difference was observed across the datasets. While structural differences in proteins across source organisms, particularly in conserved domains or motifs, likely exist, the proteins used in this study did not show strong deviations in the features employed to train the models.

Although the three datasets displayed similar range of feature values on average, it could be worthwhile to compare the characteristics of seed storage proteins to some reference proteins, especially of the four Osborne classes. Therefore, five AF predicted structures of three different human proteins, namely keratin, hemoglobin, and lysozyme were collected as reference proteins. The proteins were selected to represent a broad range of properties, as they are of different origins: connective tissues, blood, and saliva, respectively. Moreover, five AF structures of

albumin, globulin, prolamin, and glutelin with respective single source organism (mouse-ear cress, garden pea, maize, and rice, respectively) were compared (Appendix Table A.7). As shown in the table, the four Osborne groups exhibited significantly different physicochemical properties from each other and the reference proteins. For example, maize zein had the highest ratio of hydrophobic surface (0.707 ± 0.008), while 11S pea globulin exhibited the lowest (0.544 ± 0.012). Rice glutelin had the highest ratio of coiled secondary structures ($53.208 \pm 1.706\%$), whereas zein had the lowest ($26.086 \pm 4.081\%$). The differences among the proteins were also highlighted by the relative amino acid fractions, such as the highest Leu ($18.838 \pm 0.870\%$) and Cys ($5.790 \pm 0.534\%$) contents in zein and 2S albumin, respectively. This comparison suggested that, despite being categorized as seed storage proteins, the four Osborne fractions exhibit substantially different physicochemical properties at monomer-level.

2.3.2. Linear correlation analysis between structure/sequence features with solubility

Prior to MLP model construction, a linear correlation analysis was conducted between each sequence/structure descriptor and solubility within *E.coli* dataset (Appendix Table A.8). As discussed in section 2.3.1, significant deviations in the descriptors compared to their mean values were observed, indicating the diversity of proteins in the dataset. The highest linear correlation with solubility ($r = 0.361$) was exhibited by the hydrophilic index, followed by absolute charge per residue ($r = 0.357$) and aliphatic index ($r = -0.335$). These correlations are in line with the expected behavior, as the presence of hydrophilic or charged residues can facilitate hydrogen bond and electrostatic attractions between the protein and surrounding water (Han et al., 2020). These interactions can effectively prevent aggregation resulting from hydrophobic interactions

and contribute to the promotion of protein solubility. In contrast, certain parameters, such as the instability index ($r = -0.042$) or strand propensity ($r = 0.097$), showed very weak or no linear influence.

2.3.3. Solubility prediction using MLP model

The predictive performances of the MLP model along with those of benchmark models were displayed in Table 2.1. On the test *E. coli* dataset, the MLP model exhibited testing R^2 of 0.469, followed closely by gradient boosting (GB) method at 0.441 and random forest (RF) model at 0.436. SVR, which has been extensively used in non-deep learning solubility prediction models, achieved R^2 of 0.443. The slightly higher performance of the MLP model may reflect its ability, as the only deep-learning model evaluated, to capture nonlinear relationships among the input features. In contrast, SGDRegressor and LR showed lower predictive performances, with test R^2 values of 0.365 and 0.368, respectively. Because both approaches are based on linear regression in this setting, their weaker performance compared with MLP, SVR, GB, and RF suggests that the dataset may contain nonlinear relationships between the input features and protein solubility that are not fully captured by linear models.

Table 2.1. Prediction performances of the MLP and benchmark models (cross-validation / testing). For binary classification evaluation, regression outputs were converted to class labels using a threshold of 0.5. Abbreviations are: Linear regression (LR), stochastic gradient descent (SGD), support vector regression (SVR), random forest (RF), gradient boosting (GB), multilayer perceptron (MLP), area under curve (AUC), and Matthews correlation coefficient (MCC).

Models	Regression metrics		Binary classification metrics			
	R ²	RMSE	Accuracy	F1	AUC	MCC
LR	0.350/0.368	0.256/0.260	0.743/0.742	0.728/0.728	0.822/0.831	0.486/0.488
SGD	0.353/0.365	0.255/0.260	0.740/0.736	0.724/0.721	0.824/0.830	0.481/0.475
SVR	0.472/0.443	0.231/0.244	0.776/0.772	0.748/0.745	0.854/0.849	0.546/0.548
RF	0.431/0.436	0.240/0.245	0.765/0.784	0.732/0.753	0.838/0.848	0.525/0.561
GB	0.436/0.441	0.239/0.244	0.761/0.785	0.732/0.758	0.838/0.849	0.517/0.565
MLP	0.455/0.469	0.234/0.238	0.768/0.795	0.737/0.764	0.845/0.858	0.530/0.583

2.3.4. Solubility prediction using GCN model

The GCN model was constructed and trained using the generated node features and contact maps from AF structures. Along with the GCN model, the latest regression model, GraphSol (Chen et al., 2021) was trained as a reference using the same training set. While both the GCN and GraphSol employed the same solubility dataset for training, GraphSol incorporated a greater number of node features (91 features compared to the GCN model's 30 features). Furthermore, GraphSol utilized adjacency matrices generated from SPIDER3 instead of AF. Table 2 displayed the performances of the two models; the GCN model exhibited inferior predictive power compared to GraphSol in regression task (R² 0.435 vs 0.476) but comparable classification accuracy (0.786 vs 0.787).

The better performance of the GraphSol was likely due to the higher number of features, which would promote a more finely-trained model. Unlike the GCN model, Graphsol incorporated evolutionary features, namely hidden markov matrix (HMM) and position-specific

scoring metrics (PSSM). Both HMM and PSSM are matrices derived from multiple sequence alignment (MSA) of a specific protein sequence, where related protein sequences are aligned based on similarities (Ismi & Pulungan, 2022). Using the aligned sequences, HMM and PSSM summarize the characteristics of a set of related protein. Specifically, they assign probabilities, or scores to residues based on the observed alignment. These scores convey evolutionary information and can be utilized to predict protein functionality and structure (Rashid et al., 2007). However, because this study aimed to develop a model applicable to diverse food and agricultural proteins, HMM- and PSSM-based evolutionary features were excluded from the present GCN model to avoid introducing lineage- or homolog-specific information that may be overrepresented in an *E. coli*-based benchmark dataset, although this exclusion may have contributed to the lower regression performance on the *E. coli* test dataset.

2.3.5. Performance of stacking model

Stacking model, also known as ensemble model, is a powerful approach that combines the prediction results of multiple models to create a more accurate and generalized model (Sagi & Rokach, 2018). As demonstrated in Table 2.2, the stacking model with SVR regressor constructed in this research achieved R^2 of 0.502, outperforming ESM-2 embedded MLP and GCN ($R^2 = 0.364$ and 0.453 , respectively), its base models MLP ($R^2 = 0.469$), GCN ($R^2 = 0.435$) as well as GraphSol ($R^2 = 0.476$) in the regression task. Among the meta-regressors used, SVR displayed the highest performance ($R^2 = 0.502$) followed by LR ($R^2 = 0.498$), KNN ($R^2 = 0.481$) and DT ($R^2 = 0.458$). The better performance of the stacking model compared to its base models suggested that the meta-regressors successfully captured crucial aspects from the respective base models. To visually depict the prediction performance of the models, plots were generated to

compare the actual solubility with the predicted solubility (Fig. 2.3). From the figure, it was observable that the predicted solubility were more closely aligned with the diagonal line $y = x$ compared to the LR model or the base models, indicating higher linearity and a higher R^2 value. This linearity was particularly evident in the region where predicted solubility falls between 0.2 and 0.4. Moreover, the stacking model outperformed the other models in binary classification task as well (Table 2.2, Fig. 2.3).

Table 2.2. Prediction performances of stacking models with different meta-regressors and benchmark/base models (cross-validation / testing).

Models	Regression metrics		Binary classification metrics			
	R^2	RMSE	Accuracy	F1	AUC	MCC
ESM-2	0.372	0.251	0.740	0.691	0.825	0.481
(MLP)	/0.364	/0.261	/0.750	/0.752	/0.841	/0.522
ESM-2	0.439	0.239	0.762	0.723	0.839	0.519
(GCN)	/0.453	/0.241	/0.773	/0.726	/0.859	/0.538
GCN	0.414	0.262	0.747	0.762	0.838	0.558/
	/0.435	/0.246	/0.787	/0.774	/0.857	0.577
MLP	0.455	0.234	0.768	0.737	0.845	0.530
	/0.469	/0.238	/0.795	/0.764	/0.858	/0.583
Graphsol	0.469	0.236	0.779	0.752	0.860	0.545
	/0.476	/0.237	/0.786	/0.759	/0.869	/0.557
Stacking	0.488	0.227	0.784	0.755	0.856	0.563
(LR)	/0.498	/0.231	/0.801	/0.782	/0.874	/0.600
Stacking	0.455	0.235	0.766	0.729	0.839	0.525
(DT)	/0.458	/0.240	/0.792	/0.749	/0.848	/0.577
Stacking	0.473	0.231	0.775	0.745	0.851	0.545
(KNN)	/0.481	/0.235	/0.803	/0.781	/0.867	/0.602
Stacking	0.487	0.227	0.783	0.751	0.856	0.560
(SVR)	/0.502	/0.230	/0.804	/0.783	/0.876	/0.604

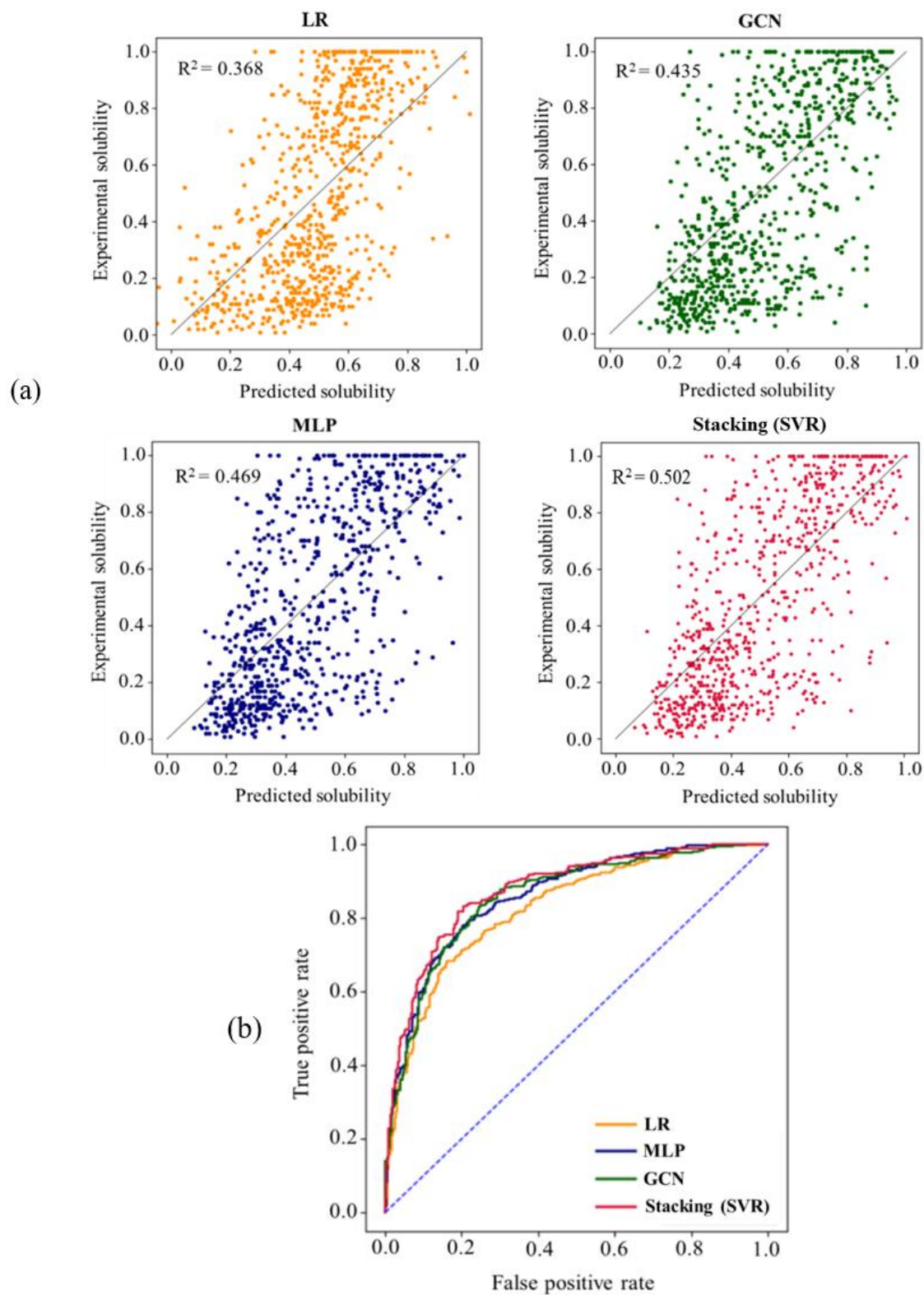
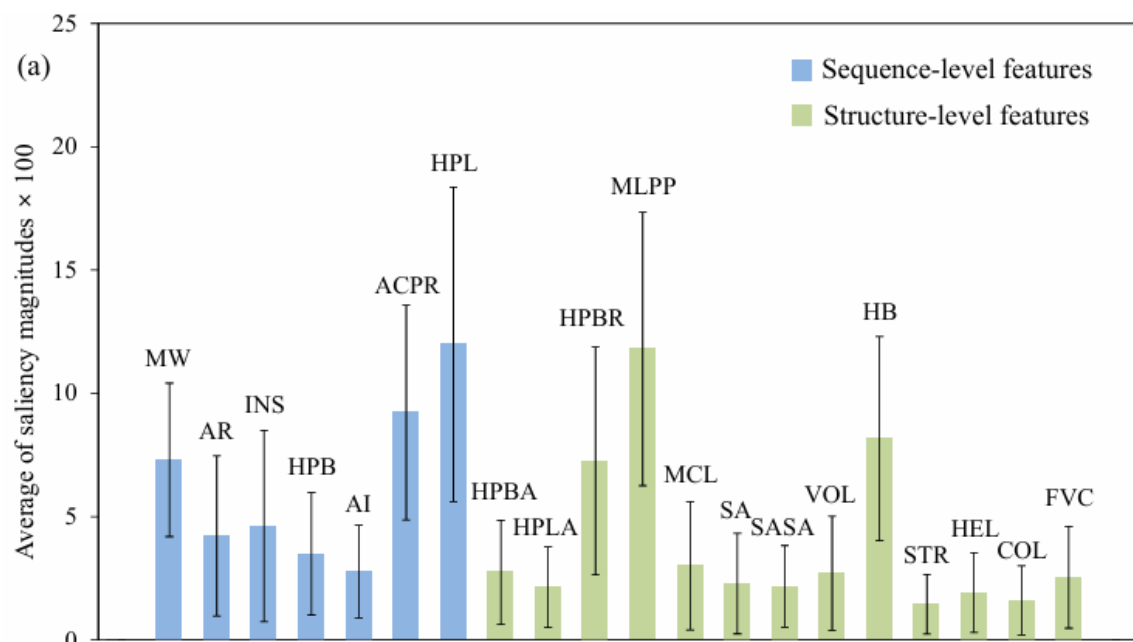


Figure 2.3. The scatter plots of predicted and experimental solubility (a) and receiver operating characteristic (ROC) curve (b) for each model on *E. coli* test dataset.

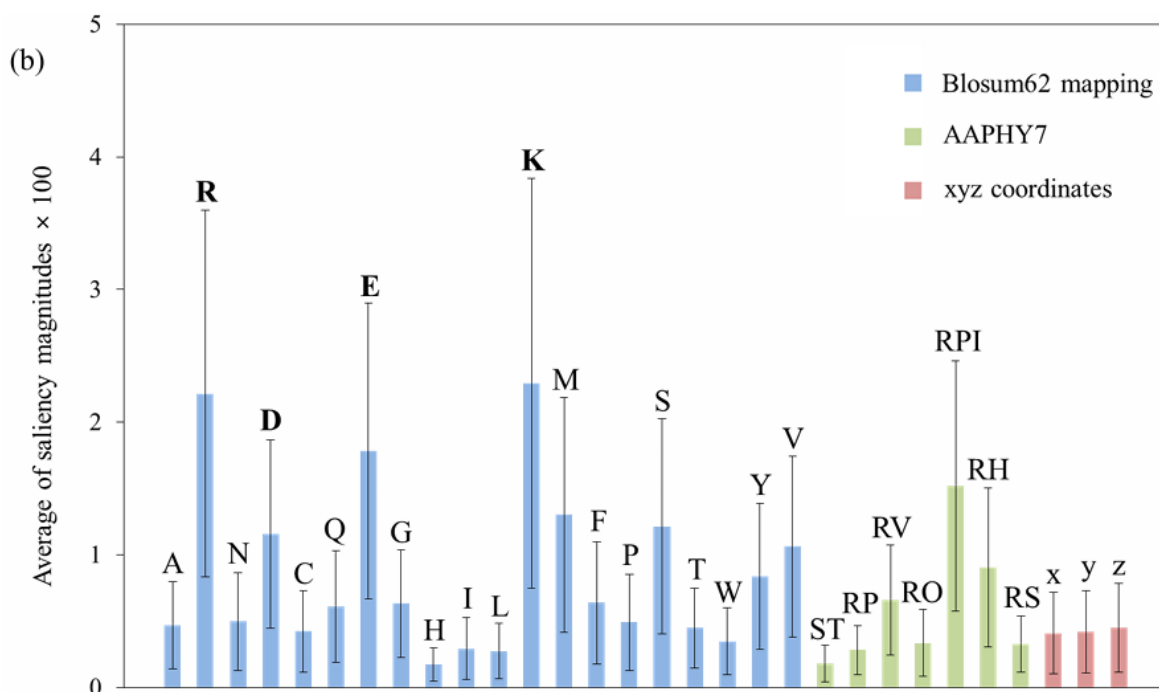
2.3.6. Model interpretation

In order to determine the most influential features used in the base models MLP and GCN, the saliency maps of each model were generated. Saliency map assigns the significance scores of each feature used by computing the gradient of the score function in respect to each inputs (Ismail et al., 2021). As shown in Fig. 2.4a, the most influential features of the MLP models were the hydrophilicity index and mean lipophilicity potential, followed by absolute charge per residue, number of hydrogen bonds, molecular weight, and the ratio of hydrophilic area. These results are consistent with existing reports on the strong influence of hydrophilic, hydrophobic, and charged nature of proteins, as well as molecular weight, on solubility (Trevino et al., 2008; Van Oss, 1997). Based on the magnitudes of computed saliency, sequence-level and structure-level features appeared to demonstrate a similar level of importance in the MLP model.

In the case of the GCN model, the blosum62 mapping displayed higher level of influence compared to AAPHY7 and coordinates features (Fig. 2.4b). Notably, the blosum62 mapping for the charged residues, including R, D, E, and K showed more significant saliencies compared to other residues. Derived from the observation of substitutions in conserved blocks of protein sequences, blosum62 provides the similarity of a residue compared to another residue (Rudnicki et al., 2024). Therefore, GCN seemed to put strong emphasis on the similarity of each residue to the charged ones. Among the AAPHY7 features used, the residual isoelectric point displayed the highest impact, which could correlate with the degree of charged residues. From the high saliency of the residual charge-related features, it was conceivable that the GCN model puts the most emphasis on the charge states of the protein residues.



MW: molecular weight, AR: aromaticity, INS: instability index, HPB: hydrophobic index, AI: aliphatic index, ACPR: absolute charge per residue, HPI: hydrophilic index, HPBA: hydrophobic area, HPLA: hydrophilic area, HPBR: ratio of hydrophobic area, MLPP: mean lipophilic potential, MCL: mean Coulombic potential, SA: surface area, SASA: solvent accessible surface area, VOL: volume, HB: number of hydrogen bonds, STR: strand propensity, HEL: helix propensity, COL: coil propensity, FVC: favorable contacts



ST: steric parameter, RP: residue polarizability, RV: residue volume, RO: residue hydrophobicity, RPI: residue isoelectric point, RH: residue helix propensity, RS: residue strand propensity

Figure 2.4. The relative importance of features determined from saliency maps of the trained MLP (a) and GCN (b). The blosum62 mappings for charged residues are bolded in (b).

2.3.7. Comparison with other methods using external *S. cerevisiae* test set

An external validation was conducted to assess the robustness of the stacking model in comparison to previous models using *S. cerevisiae* dataset. Table 2.3 displayed the prediction performance of the base models, the stacking model, and reference models in terms of the square of Pearson's correlation (r^2) and coefficient of determination (R^2). As shown in the table, the stacking model with SVR regressor outperformed all the existing competitors in both metrics ($r^2 = 0.494$ compared to 0.423 of SOLart, and $R^2 = 0.468$ compared to 0.424 of GATSol). The stacking model with LR regressor displayed slightly lower performance compared to using SVR regressor ($r^2 = 0.485$ and $R^2 = 0.452$). Notably, the stacking models, and the top competitors SOLart, GraphSol, HybridGCN, and GATSol were all structure-based models, while the underperforming models were sequence-based. The better performances of structure-based models indicated the crucial role of structural information in predicting protein solubility. Interestingly, our stacking model, derived from AF-predicted structures, yielded higher performance than SOLart, which was constructed upon experimentally determined structures, by approximately 17%. This improvement could be attributed to the larger number of protein structures in the training set (2237 compared to 406). Furthermore, it was reported that the accuracy of AlphaFold predictions is comparable to experimental structures. Compared to GraphSol, our stacking model demonstrated a performance improvement of around 30%. This difference was more pronounced in the *S. cerevisiae* test set compared to the *E. coli* test set (5.46%). The larger performance gap was believed to stem from the differences in the node features used. As mentioned earlier, GraphSol incorporated evolutionary information represented by HMM and PSSM. Excluding these evolutionary features was thought to help avoid overfitting the stacking model to the proteins similar to *E. coli* family. This notion was further supported by

the fact that the performance of the base model GCN was lower than that of GraphSol in the *E. coli* dataset but higher in the *S. cerevisiae* dataset. Moreover, when compared with PLM-based prediction tools, the stacking model (SVR) outperformed HybridGCN and GATSol by 23.80% and 10.37%, respectively. Derived from the authors of Graphsol, HybridGCN employs ESM-1b language model as well as all the node features used in Graphsol. On the other hand, GATSol utilizes ESM-1v (an updated version of ESM-1b) on par with blosum 62 mapping as its features. While direct comparison was not feasible, as the stacking model does not employ language model embedding, a possible explanation for its better performance could be the high feature dimensions of the two transformers used (1280 per residue).

To illustrate the improvement of performance achieved by stacking the MLP and GCN models, scatter plots were generated for the predictions made in the *S. cerevisiae* set (Fig. 2.5). As seen from the scatter plots of the base models, the major deviations in GCN model results were at the region where the predicted solubility is between 0.0 and 0.4. On the other hand, the MLP model exhibited the most deviation in the predicted solubility ranging from 0.6 to 1.0. In the case of the stacking model, it was evident that the predictions in the regions where each base model failed were successfully alleviated. This improvement suggested the stacking model's ability to leverage the strength of its base models. While the base models MLP and GCN were both constructed from predicted structures of the same proteins, they utilized distinct-level information of the proteins. Specifically, the MLP model leveraged information from the protein's structure and sequence, and GCN utilized a recreated graph structure based on connectivity and node features. MLP approach might overlook the intricate interactions and spatial relationships between amino acids. Conversely, GCN could underestimate long-range interactions but lacked a comprehensive viewpoint on protein structure (Rampásek & Wolf,

2021). Hence, the stacking model seemed to be able to achieve a more holistic understanding of proteins compared to its base models.

Table 2.3. Prediction performance of the stacking models and existing prediction models on *S. cerevisiae* dataset.

Solubility models	R ²
ESM-2 (GCN)	0.280 / 0.314 [*]
MLP	0.303 / 0.433 [*]
GCN	0.421 / 0.459 [*]
Stacking (LR)	0.452 / 0.485 [*]
Stacking (SVR)	0.468 / 0.494[*]
GATSol	0.424 ^a
HybridGCN (ensemble)	0.390 ^b
HybridGCN	0.378 ^b
SOLart	0.423 ^{*c}
Graphsol (ensemble)	0.372 ^c
Graphsol	0.358 ^c
ccSOL	0.303 ^{*d}
Protein-Sol	0.281 ^{*d}
DeepSol	0.090 ^{*d}
proGAN	0.084 ^d

^{*}Square of Pearson's correlation

^aGenerated by Li et al. (2024)

^bGenerated by Chen et al. (2023)

^cGenerated by Chen et al. (2021)

^dGenerated by Hou et al. (2020)

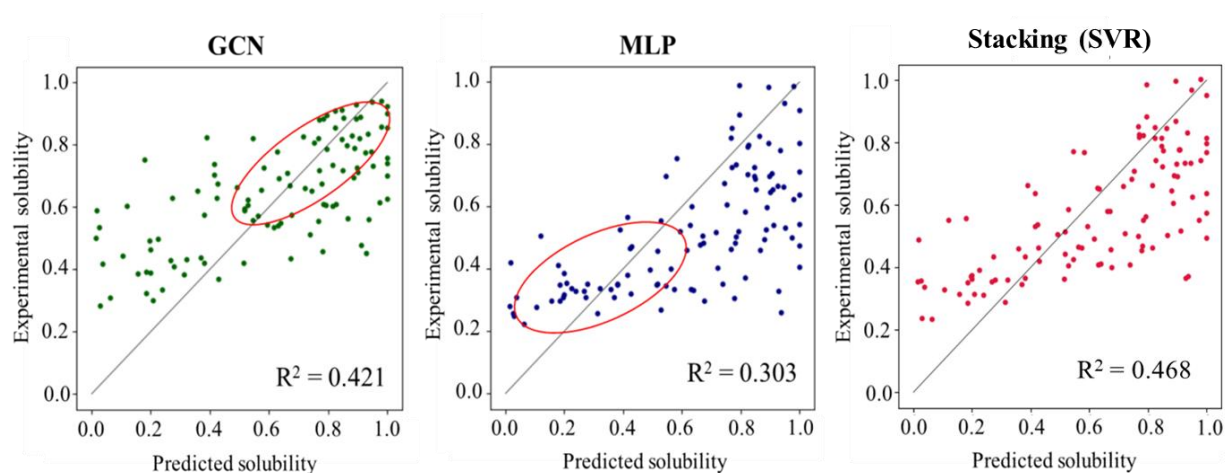


Figure 2.5. The scatter plots of predicted and experimental solubility on *S. cerevisiae* validation dataset.

2.3.8. Transferability of the stacking model to seed storage proteins

While the stacking model demonstrated a promising performance on microbial proteins, its applicability to food-related proteins remained unknown. Validating the transferability of the model would ideally involve comparing predicted values with experimental data. However, in the field of food and agricultural science, the solubility of individual protein is not commonly reported, making it nearly impossible to construct a solubility dataset. To address this challenge, the Osborne classification of plant storage proteins was employed. This classification system categorizes proteins based on their solubility in various solvents, dividing them into albumin (water soluble), globulin (salt soluble), prolamin (alcohol soluble), and glutelin (alkaline soluble) fractions (Osborne, 1924). These four fractions are relevant for predicting aqueous solubility, as each displays distinct solubility behaviors in water. Specifically, albumins are highly soluble in water, while globulins are known to be partially soluble in water (Yang et al., 2024). In contrast, glutelin and prolamin, which require more harsh conditions to solubilize, are barely soluble in water (Sathe et al., 2018).

The process of obtaining Osborne fractions first involves solubilizing a sample in water, centrifuging it, and collecting the supernatant as albumin. The remaining residue is then subjected to saline water, and the supernatant is collected as globulin. This process continues with prolamins using an alcohol solution and with glutelins using an alkaline solution (Taylor et al., 1984). The major proteins in each supernatant are then classified into the respective Osborne classes. Therefore, although numerical quantification does not follow, the process of obtaining the albumin fraction is analogous to the relative solubility definition used in *E.coli* dataset. In this sense, by comparing the predicted solubilities of proteins falling under each classification, it would be possible to indirectly demonstrate the applicability of the model to food-related protein solubility. Based on the documented solubility of each Osborne fraction, it was anticipated that albumin would exhibit the highest solubility, followed by globulin with lower solubility, and finally, prolamins and glutelins would be insoluble.

Thus, the stacking model was applied to a collected dataset of 200 seed storage proteins, and their predicted solubility was presented in Fig. 2.6. The mean predicted solubility for each classification was 0.621 ± 0.147 for albumin, 0.407 ± 0.116 for globulin, 0.310 ± 0.116 for prolamins, and 0.301 ± 0.079 for glutelins. Notably, the model predicted some level of solubility for glutelins and prolamins, which differed from the expected behavior. This discrepancy likely stemmed from the inherent uncertainty in the model (R^2 of 0.502 in the *E. coli* dataset). Nevertheless, the class-level prediction pattern suggested that the model retained some degree of transferability to plant proteins outside its original training domain.

Moreover, the existing models were tested on the seed storage protein dataset (Table 2.4). Evaluations of Graphsol and HybridGCN were excluded due to the heavy computational demands of generating PSSM and HMM. As shown in the table, while albumin was generally

predicted to be the most soluble across all models, discrepancies were observed for other Osborne fractions. For example, Protein-Sol, SoLart, and GATSol, on average, predicted prolamins to be more soluble in water than globulin, which did not align with the reported behaviors of globulin (partially soluble) and prolamins (insoluble). Although the stacking model did not perfectly reproduce the expected insolubility of prolamins and glutelins, it produced a more conservative class-level pattern, with albumins showing the highest predicted solubility, globulins showing intermediate values, and prolamins and glutelins showing the lowest values. These results suggest that the stacking model may retain limited but meaningful transferability to plant seed storage proteins at the level of relative Osborne-class trends, rather than as an absolute predictor of individual protein solubility.

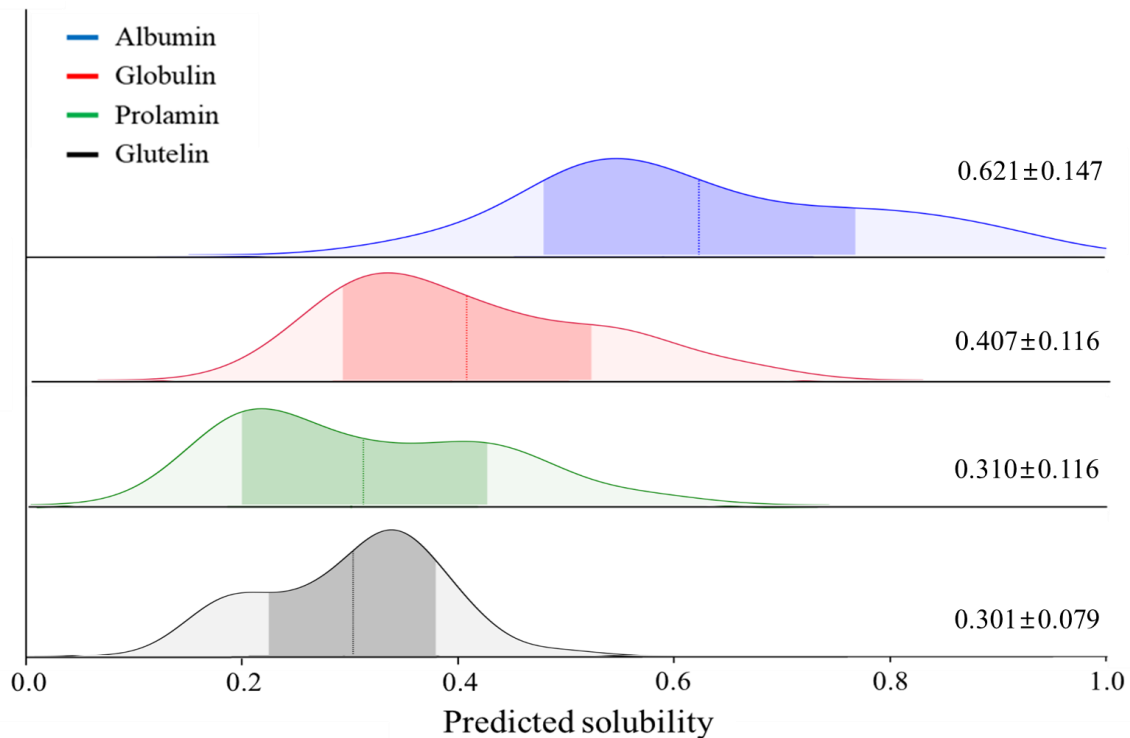


Figure 2.6. Stacking model (SVR) predictions for the solubility of seed storage proteins by their Osborne classifications.

Table 2.4. Predictions of existing models on seed storage protein dataset.

	Albumin (soluble)	Globulin (partially soluble)	Prolamin (insoluble)	Glutelin (insoluble)
DeepSol	0.492±0.137	0.464±0.078	0.315±0.167	0.456±0.096
Protein-Sol	0.599±0.120	0.373±0.101	0.528±0.070	0.325±0.091
SoLart	0.659±0.067	0.500±0.065	0.565±0.032	0.521±0.109
GATSol	0.578±0.119	0.356±0.066	0.496±0.105	0.336±0.117
Stacking (SVR)	0.621±0.147	0.407±0.116	0.310±0.116	0.301±0.079

2.3.9. Case study on seed storage proteins: identification of important residues

To determine the significance of specific residues in predicting solubility using the GCN model, the saliency values generated for 2S albumin from soybean and 22 kDa zein (prolamin) from maize were analyzed. These proteins were selected as they displayed the highest and lowest predicted solubility: 0.898 for albumin and 0.142 for prolamin in the stacking model, with GCN model predictions of 0.825 and 0.169, respectively. For 2S albumin, residues 8-31 (region 1), 65-83 (region 2), and 139-143 (region 3) were found to have the greatest impact. For 22 kDa prolamin, residues 31-135 (region 1), 160-168 (region 2), and 171-175 (region 3) were identified as the most influential (Fig. 2.7). Notably, these regions exhibited fewer charged residues per length compared to the overall protein. The selected regions of albumin had charged residue densities of 0.0869, 0.278, and 0.000, respectively, compared to 0.367 for the entire protein. For prolamin, the regions had charged group densities of 0.019, 0.000, and 0.000, respectively, compared to 0.023 for the whole protein. This charge disparity was further highlighted by the electrostatic potential data from the APBS web server.

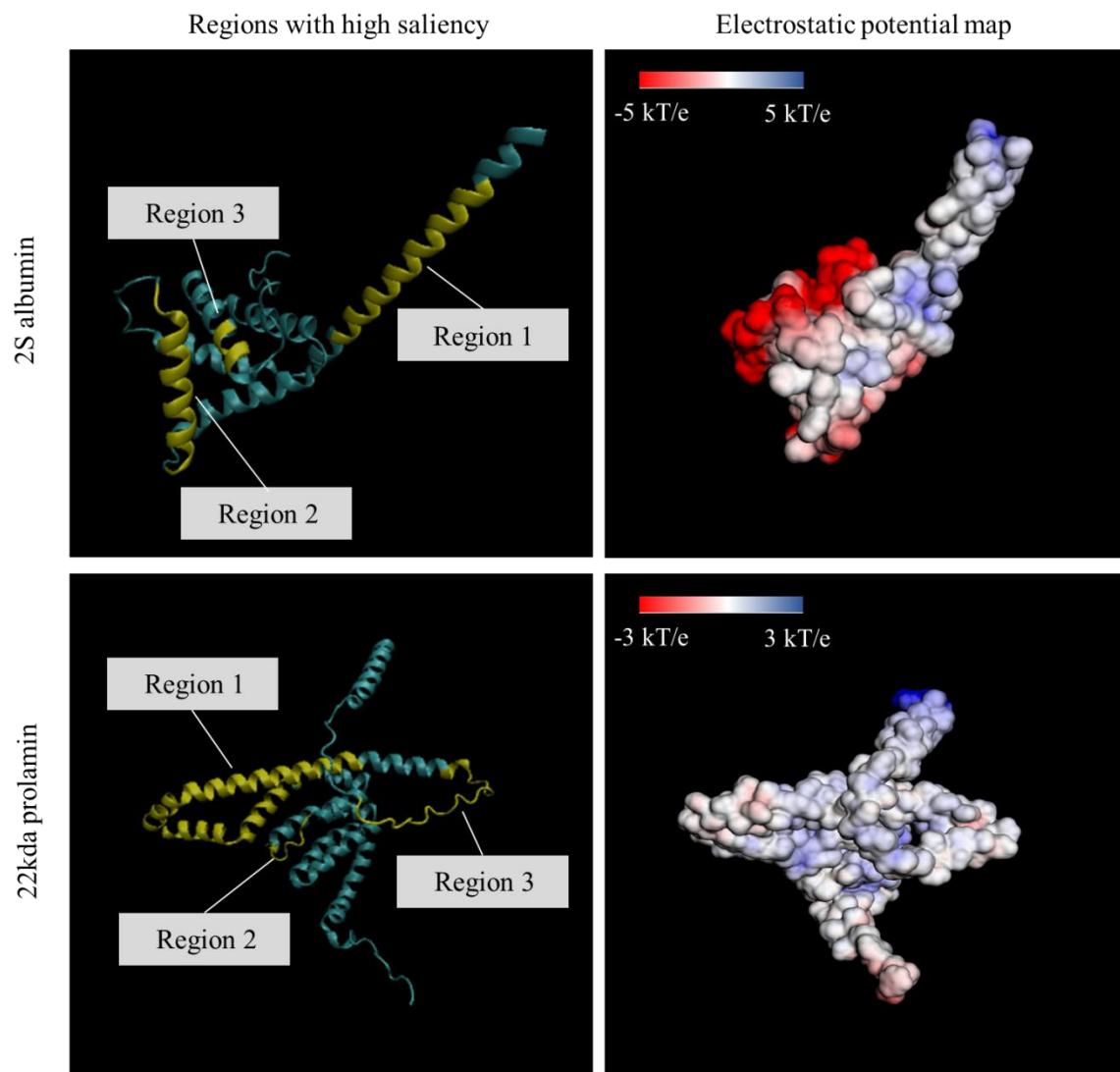


Figure 2.7. Selected regions with high saliency and electrostatic potential maps of soybean 2S albumin and maize 22 kDa prolamin.

2.3.10. Case study on seed storage proteins: CG-MD simulation

To further investigate the dissolution behaviors of storage proteins predicted by the stacking model, 4 μ s CGMD simulations were conducted with 2S albumin and 22 kDa prolamin (Fig. 2.8a). The important residues from the saliency map of the GCN models were colored in red. As shown in the figure, both 2S albumin and 22 kDa prolamin exhibited aggregating

behaviors over the simulation time. However, while 2S albumin tended to aggregate into small clusters of two or three monomers, larger aggregates involving multiple 22 kDa prolamins were observed. The more aggregation-prone behavior of prolamin compared to albumin was also reflected in their mean square displacement (MSD) over time (Appendix Fig.1). MSD measures the average squared distance that particles move over a certain time interval. Compared to 2S albumin, 22 kDa prolamin displayed faster and more severe aggregation, reducing overall mobility and MSD. The diffusion coefficient, roughly estimated from the slope from the later 2 μs , was $17.91 \pm 16.20 \text{ nm}^2 \mu\text{s}^{-1}$ for 2S albumin and $5.75 \pm 7.35 \text{ nm}^2 \mu\text{s}^{-1}$ for 22 kDa prolamin. Given that the formation of large aggregates could directly lead to increased amount of protein precipitation (Golovanov et al., 2004), the simulation results seemed to support the solubility predictions made by the stacking model.

While Fig. 2.8a demonstrated the predicted behaviors of the two proteins, the role of the important regions determined in Section 2.3.8 in protein aggregation remained unclear. Therefore, the trajectories of the simulations during the early stage of aggregation were investigated. From the snapshots of the trajectories, it was observed that initial protein aggregation began with the contact between proteins and the influential regions (Fig. 2.6b), whether in the large aggregate of prolamin or the smaller ones of albumin. This was more prominent in the prolamin simulation at 0.15 μs , where the binding of multiple selected regions to a prolamin monomer was observed. The binding of these regions further promoted overall aggregation, as both selected and unselected regions came into contact and aggregated (0.50 μs in Fig. 2.8b and 1.5 μs in Fig. 2.8a). The snapshots of the early stage of aggregation suggested that the selected regions were more aggregation-prone compared to the unselected regions, potentially inducing further aggregation. This point was further supported by the fewer charged

groups in the identified regions (Fig. 2.4), which would result in less favorable interaction of these regions with water compared to highly charged ones. The importance of charge-related features, the deficiency of charged groups within the selected residues, and the more aggregation-prone nature of the selected regions from the simulation therefore collectively suggested that the GCN model predicts solubility by identifying residues likely to aggregate, emphasizing the charged states of residues.

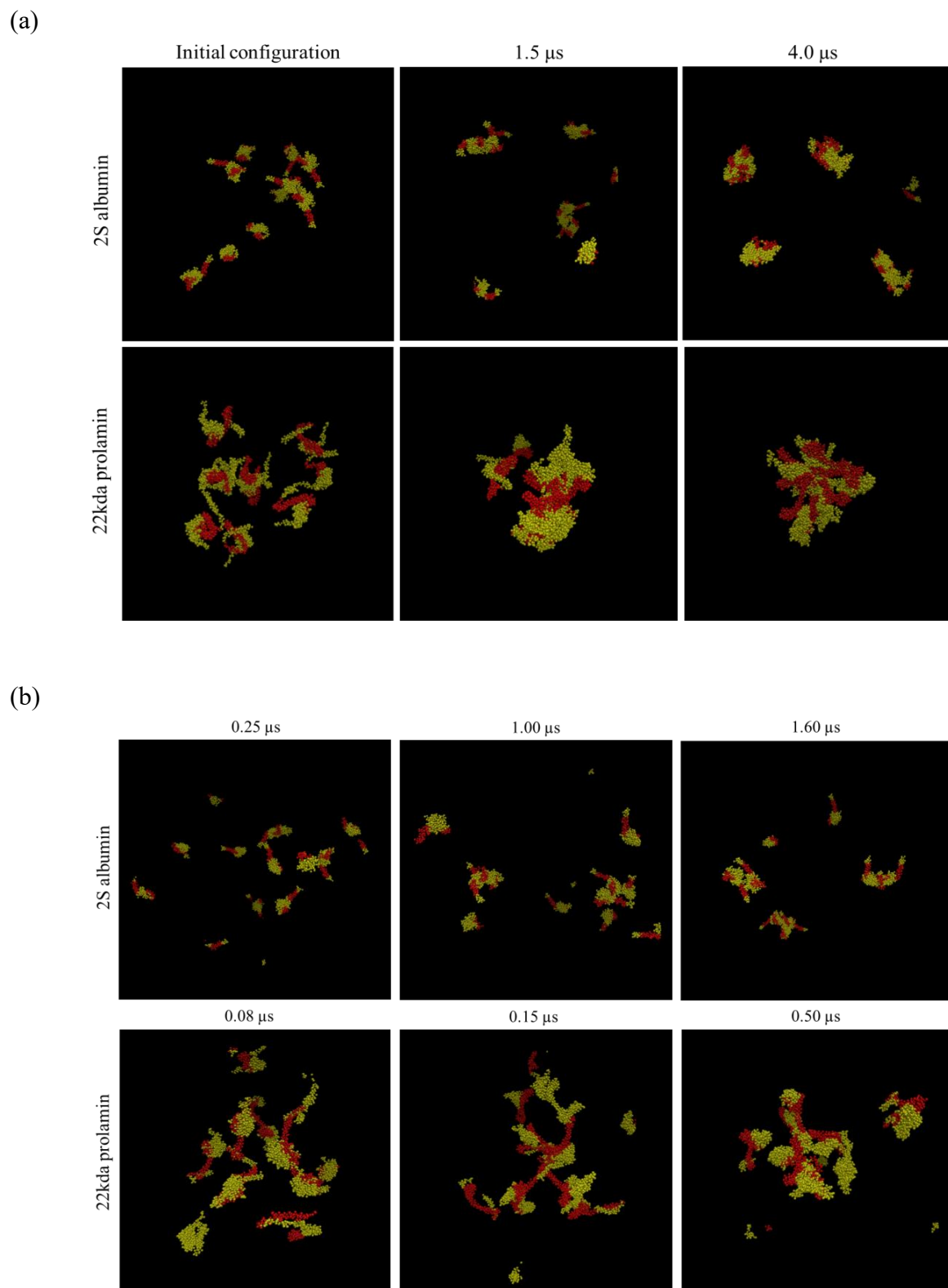


Figure 2.8. Snapshots of CGMD simulations for soybean 2S albumin and maize 22 kDa prolamin at different intervals (a) and initial aggregation stage (b).

2.4. Conclusion

In this study, a novel solubility prediction model was developed using the protein structures generated from AlphaFold 2. Based on the predicted structures, two distinctive feature sets on their sequence+structural information and on residual feature+contact map were subjected to MLP and GCN model, respectively. In the case of GCN model, the use of evolutionary features related to *E.coli* proteins was avoided during training, presumably providing more robust prediction applicable to food related proteins. The resulting out-of-fold predictions from the two base models were combined to create stacking models using various meta-regressors. The stacking model with SVR achieved the best performance among existing models in an external validation dataset. The higher performance of the stacking model compared to its base models suggested that stacking approach effectively leveraged the strengths of base models and the information from the distinct protein features. The stacking model (SVR) was further validated using a dataset consisting of seed storage proteins, generating expected solubility trend for seed storage proteins and exhibiting bulk transferability from microbial proteins to food and agricultural proteins.

The limitation of this study is the lack of wet experimental validation data, due to the difficulty in obtaining isolated seed storage proteins. Additionally, although the extracted physicochemical features for training the model were fairly consistent across each dataset, the adequacy of using *E.coli* proteins to model seed storage proteins needs further validation. Moreover, the lack of multimeric structure data might impede the model performance, which could be improved with multimer prediction tools. Lastly, a further comparison with existing regressors on different dataset outside the one used in the research would make the performance evaluation more rationale. Yet, this research marks the first study to develop a solubility model for food proteins, and the framework adopted here could serve as a pioneering approach for future research in this area.

Acknowledgements

This is contribution No. 24-083-J from the Kansas Agricultural Experimental Station. This research was supported by the USDA National Institute of Food and Agriculture Hatch project (7003330).

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*, 1–2, 19–25.
<https://doi.org/10.1016/j.softx.2015.06.001>
- Agostini, F., Cirillo, D., Livi, C. M., Ponti, R. D., & Tartaglia, G. G. (2014). ccSOL omics: A webserver for large-scale prediction of endogenous and heterologous solubility in *E. coli*. *Bioinformatics*, 30(20), 2975–2977. <https://doi.org/10.1093/bioinformatics/btu420>
- Arican, O. C., & Gumus, O. (2023). PredDRBP-MLP: Prediction of DNA-binding proteins and RNA-binding proteins by multilayer perceptron. *Computers in Biology and Medicine*, 164, Article 107317. <https://doi.org/10.1016/j.combiomed.2023.107317>
- Baranwal, M., Magner, A., Saldinger, J., Turali-Emre, E. S., Elvati, P., Kozarekar, S., VanEpps, J. S., Kotov, N. A., Violi, A., & Hero, A. O. (2022). Struct2Graph: A graph attention network for structure based predictions of protein–protein interactions. *BMC Bioinformatics*, 23(1), Article 370. <https://doi.org/10.1186/s12859-022-04910-9>
- Bauer, P., Hess, B., & Lindahl, E. (2023). *GROMACS 2022.5 source code* (Version 2022.5) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7586780>

- Berman, H. M., Gabanyi, M. J., Kouranov, A., Micallef, D., & Westbrook, J. (2017). *Protein Structure Initiative–TargetTrack 2000–2017: All data files* [Data set]. Zenodo.
<https://doi.org/10.5281/zenodo.1234567>
- Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8), 2102–2110.
<https://doi.org/10.1093/bioinformatics/btac020>
- Chen, J., Zheng, S., Zhao, H., & Yang, Y. (2021). Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *Journal of Cheminformatics*, 13(1), Article 7. <https://doi.org/10.1186/s13321-021-00488-1>
- Chen, L., Wu, R., Zhou, F., Zhang, H., & Liu, J. K. (2023). HybridGCN for protein solubility prediction with adaptive weighting of multiple features. *Journal of Cheminformatics*, 15(1), Article 118. <https://doi.org/10.1186/s13321-023-00788-8>
- Cheng, Z., Yan, C., Wu, F. X., & Wang, J. (2022). Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4), 2208–2218.
<https://doi.org/10.1109/TCBB.2021.3077905>
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & 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>
- Deng, R., Mars, M., Van Der Sman, R. G. M., Smeets, P. A. M., & Janssen, A. E. M. (2020). The importance of swelling for in vitro gastric digestion of whey protein gels. *Food Chemistry*, 330, Article 127182. <https://doi.org/10.1016/j.foodchem.2020.127182>

- 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>
- Du, Z., Xu, Y., Liu, C., & Li, Y. (2024). pLM4Alg: Protein language model-based predictors for allergenic proteins and peptides. *Journal of Agricultural and Food Chemistry*, 72(1), 752–760. <https://doi.org/10.1021/acs.jafc.3c07143>
- Evans, P., Wyatt, K., Wistow, G. J., Bateman, O. A., Wallace, B. A., & Slingsby, C. (2004). The P23T cataract mutation causes loss of solubility of folded γ D-crystallin. *Journal of Molecular Biology*, 343(2), 435–444. <https://doi.org/10.1016/j.jmb.2004.08.050>
- Gligorijević, V., Renfrew, P. D., Kosciolk, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K., & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1), Article 3168. <https://doi.org/10.1038/s41467-021-23303-9>
- Golovanov, A. P., Hautbergue, G. M., Wilson, S. A., & Lian, L.-Y. (2004). A simple method for improving protein solubility and long-term stability. *Journal of the American Chemical Society*, 126(29), 8933–8939. <https://doi.org/10.1021/ja049297h>
- Grossmann, L., & McClements, D. J. (2023). Current insights into protein solubility: A review of its importance for alternative proteins. *Food Hydrocolloids*, 137, Article 108416. <https://doi.org/10.1016/j.foodhyd.2022.108416>
- Han, X., Ning, W., Ma, X., Wang, X., & Zhou, K. (2020). Improving protein solubility and activity by introducing small peptide tags designed with machine learning models.

- Metabolic Engineering Communications*, 11, Article e00138.
<https://doi.org/10.1016/j.mec.2020.e00138>
- Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein–Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19), 3098–3100. <https://doi.org/10.1093/bioinformatics/btx345>
- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22), 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>
- Hou, Q., Kwasigroch, J. M., Rooman, M., & Pucci, F. (2020). SOLart: A structure-based method to predict protein solubility and aggregation. *Bioinformatics*, 36(5), 1445–1452. <https://doi.org/10.1093/bioinformatics/btz773>
- 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](https://doi.org/10.1016/0263-7855(96)00018-5)
- Ismi, D. P., & Pulungan, R. (2022). Deep learning for protein secondary structure prediction: Pre and post-AlphaFold. *Computational and Structural Biotechnology Journal*, 20, 6271–6286. <https://doi.org/10.1016/j.csbj.2022.11.012>
- Jia, J., & Benson, A. R. (2020). Residual correlation in graph neural network regression. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 588–598). Association for Computing Machinery. <https://doi.org/10.1145/3394486.3403101>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ...

- Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., ... Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1), 112–128. <https://doi.org/10.1002/pro.3280>
- Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv. <https://doi.org/10.48550/arXiv.1412.6980>
- Kipf, T. N., & Welling, M. (2016). *Semi-supervised classification with graph convolutional networks*. arXiv. <https://doi.org/10.48550/arXiv.1609.02907>
- Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1), 105–132. [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
- Li, B., & Ming, D. (2024). GATSol, an enhanced predictor of protein solubility through the synergy of 3D structure graph and large language modeling. *BMC Bioinformatics*, 25(1), Article 204. <https://doi.org/10.1186/s12859-024-05820-8>
- Li, Y., Zhang, Z., Teng, Z., & Liu, X. (2020). PredAmyl-MLP: Prediction of amyloid proteins using multilayer perceptron. *Computational and Mathematical Methods in Medicine*, 2020, Article 8845133. <https://doi.org/10.1155/2020/8845133>
- Lijnzaad, P., Berendsen, H. J. C., & Argos, P. (1996). A method for detecting hydrophobic patches on protein surfaces. *Proteins: Structure, Function, and Genetics*, 26(2), 192–203. [https://doi.org/10.1002/\(SICI\)1097-0134\(199610\)26:2<192::AID-PROT9>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0134(199610)26:2<192::AID-PROT9>3.0.CO;2-I)

- Lijnzaad, P., Berendsen, H. J. C., & Argos, P. (1996). Hydrophobic patches on the surfaces of protein structures. *Proteins: Structure, Function, and Genetics*, 25(3), 389–397.
[https://doi.org/10.1002/\(SICI\)1097-0134\(199607\)25:3<389::AID-PROT10>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0134(199607)25:3<389::AID-PROT10>3.0.CO;2-E)
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>
- Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A structured self-attentive sentence embedding*. arXiv.
<https://doi.org/10.48550/arXiv.1703.03130>
- Magnan, C. N., Randall, A., & Baldi, P. (2009). SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics*, 25(17), 2200–2207.
<https://doi.org/10.1093/bioinformatics/btp386>
- 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>
- Mills, J. E., & Dean, P. M. (1996). Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *Journal of Computer-Aided Molecular Design*, 10(6), 607–622. <https://doi.org/10.1007/BF00134183>
- Niwa, T., Ying, B. W., Saito, K., Jin, W., Takada, S., Ueda, T., & Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11), 4201–4206. <https://doi.org/10.1073/pnas.0811922106>

- Osborne, T. B. (1924). The vegetable proteins. Longmans, Green and Company.
- Pollastri, G., & Baldi, P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18(suppl_1), S62–S70. https://doi.org/10.1093/bioinformatics/18.suppl_1.S62
- Radhika, V., & Rao, V. S. H. (2015). Computational approaches for the classification of seed storage proteins. *Journal of Food Science and Technology*, 52(7), 4246–4255. <https://doi.org/10.1007/s13197-014-1500-x>
- Rampásek, L., & Wolf, G. (2021). Hierarchical graph neural nets can capture long-range interactions. arXiv. <https://doi.org/10.48550/arXiv.2107.07432>
- Rashid, M., Saha, S., & Raghava, G. P. S. (2007). Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics*, 8, Article 337. <https://doi.org/10.1186/1471-2105-8-337>
- Rawi, R., Mall, R., Kunji, K., Shen, C. H., Kwong, P. D., & Chuang, G. Y. (2018). PaRSnIP: Sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, 34(7), 1092–1098. <https://doi.org/10.1093/bioinformatics/btx662>
- Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173–175. <https://doi.org/10.1038/nmeth.1818>
- Ros-Lucas, A., Martínez-Peinado, N., Bastida, J., Gascón, J., & Alonso-Padilla, J. (2022). The use of AlphaFold for in silico exploration of drug targets in the parasite *Trypanosoma cruzi*. *Frontiers in Cellular and Infection Microbiology*, 12, Article 944748. <https://doi.org/10.3389/fcimb.2022.944748>

- Rudnicki, W. R., Mroczek, T., & Cudek, P. (2014). Amino acid properties conserved in molecular evolution. *PLOS ONE*, 9(6), Article e98983.
<https://doi.org/10.1371/journal.pone.0098983>
- Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), Article e1249.
<https://doi.org/10.1002/widm.1249>
- Sathe, S. K., Zaffran, V. D., Gupta, S., & Li, T. (2018). Protein solubilization. *Journal of the American Oil Chemists' Society*, 95(8), 883–901. <https://doi.org/10.1002/aocs.12058>
- Schnepf, M. I. (1992). Protein-water interactions. In B. J. F. Hudson (Ed.), *Biochemistry of food proteins* (pp. 1–33). Springer. https://doi.org/10.1007/978-1-4684-9895-0_1
- Shapovalov, M., Dunbrack, R. L., Jr., & Vucetic, S. (2020). Multifaceted analysis of training and testing convolutional neural networks for protein secondary structure prediction. *PLOS ONE*, 15(5), Article e0232528. <https://doi.org/10.1371/journal.pone.0232528>
- Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., & Frishman, D. (2012). PROSO II: A new method for protein solubility prediction. *The FEBS Journal*, 279(12), 2192–2200.
<https://doi.org/10.1111/j.1742-4658.2012.08603.x>
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C., Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., ... 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/s41592-021-01098-3>

- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958. <https://doi.org/10.5555/2627435.2670313>
- Susanty, M., Mursalim, M. K. N., Hertadi, R., Purwarianti, A., & Rajab, T. L. (2024). Classifying alkaliphilic proteins using embeddings from protein language model. *Computers in Biology and Medicine*, 173, Article 108385. <https://doi.org/10.1016/j.combiomed.2024.108385>
- Taylor, J. R. N., Schüssler, L., & van der Walt, W. H. (1984). Fractionation of proteins from low-tannin sorghum grain. *Journal of Agricultural and Food Chemistry*, 32(1), 149–154. <https://doi.org/10.1021/jf00121a036>
- Tran, C., Khadkikar, S., & Porollo, A. (2023). Survey of protein sequence embedding models. *International Journal of Molecular Sciences*, 24(4), Article 3775. <https://doi.org/10.3390/ijms24043775>
- 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>
- Uemura, E., Niwa, T., Minami, S., Takemoto, K., Fukuchi, S., Machida, K., Imataka, H., Ueda, T., Ota, M., & Taguchi, H. (2018). Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Scientific Reports*, 8(1), Article 678. <https://doi.org/10.1038/s41598-017-18977-5>
- Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

- JMLR Workshop and Conference Proceedings.
<http://proceedings.mlr.press/v9/glorot10a.html>
- 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](https://doi.org/10.1016/S1359-0294(97)80099-4)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762>
- Weinert, W. R., & Lopes, H. S. (2004). Neural networks for protein classification. *Applied Bioinformatics*, 3(1), 41–48. <https://doi.org/10.2165/00822942-200403010-00006>
- Westerhout, J., Krone, T., Snippe, A., Babé, L., McClain, S., Ladics, G. S., Houben, G. F., & Verhoeckx, K. C. M. (2019). Allergenicity prediction of novel and modified proteins: Not a mission impossible! Development of a random forest allergenicity prediction model. *Regulatory Toxicology and Pharmacology*, 107, 104422. <https://doi.org/10.1016/j.yrtph.2019.104422>
- Whitnall, T., & Pitts, N. (2019). Global trends in meat consumption. *Agricultural Commodities*, 9(1), 96–99.
- Wood, P., & Tavan, M. (2022). A review of the alternative protein industry. *Current Opinion in Food Science*, 47, Article 100869. <https://doi.org/10.1016/j.cofs.2022.100869>
- Xu, B., Wang, N., Chen, T., & Li, M. (2015). *Empirical evaluation of rectified activations in convolutional network*. arXiv. <https://doi.org/10.48550/arXiv.1505.00853>
- Yang, J., Kornet, R., Ntone, E., Meijers, M. G. J., van den Hoek, I. A. F., Sagis, L. M. C., Venema, P., Meinders, M. B. J., Berton-Carabin, C. C., Nikiforidis, C. V., & Hinderink,

- E. B. A. (2024). Plant protein aggregates induced by extraction and fractionation processes: Impact on techno-functional properties. *Food Hydrocolloids*, 155, 110223. <https://doi.org/10.1016/j.foodhyd.2024.110223>
- Yang, Z. R., & Thomson, R. (2005). Bio-basis function neural network for prediction of protease cleavage sites in proteins. *IEEE Transactions on Neural Networks*, 16(1), 263–274. <https://doi.org/10.1109/TNN.2004.836196>
- Yuan, H., Tang, J., Hu, X., & Ji, S. (2020). XGNN: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 430–438). Association for Computing Machinery. <https://doi.org/10.1145/3394486.3403085>
- Zhang, F., Li, J., Wen, Z., & Fang, C. (2024). FusPB-ESM2: Fusion model of ProtBERT and ESM-2 for cell-penetrating peptide prediction. *Computational Biology and Chemistry*, 111, Article 108098. <https://doi.org/10.1016/j.compbiolchem.2024.108098>
- Zhang, X., Hu, X., Zhang, T., Yang, L., Liu, C., Xu, N., Wang, H., & Sun, W. (2024). PLM_Sol: Predicting protein solubility by benchmarking multiple protein language models with the updated Escherichia coli protein solubility dataset. *Briefings in Bioinformatics*, 25(5), Article bbae404. <https://doi.org/10.1093/bib/bbae404>

Chapter 3 - Unraveling the physicochemical differences among Osborne protein classes via bioinformatics and AI²

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 1,039 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

² This chapter was adapted from: Kwon, H., Xu, Y., Xu, X., & Li, Y. (2025). Unraveling the physicochemical differences among Osborne protein classes via bioinformatics and AI. Food Research International, 117322.

<https://doi.org/10.1016/j.foodres.2025.117322>

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-atom 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 were 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.

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

3.1. 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 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 prolamins) (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 prolamins 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 1,039 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-atom (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.

3.2. Materials and methods

3.2.1. Dataset

3.2.1.1. 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 1,039 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>.

3.2.1.2. 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 3.1.

Table 3.1. 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
	Garden pea (42), Rapeseed (24), Field mustard		
	(17), Soybean (14), False flax (12), Wild soy (10), Peanut (10), Lupine (10)...		
Globulin	Maize (106), Wheat (46), Sorghum (30), Hall's panicgrass (24), Job's tear (23), Foxtail millet (7), Goatgrass (7), Sugarcane (6)...	112	336
Prolamin		27	294

	Japonica rice (35), Red rice (11), Malo sina (10),		
Glutelin	African rice (10), Indica rice (10), Oryza glumaepatula (6), Oryza barthii (6)...	32	159

3.2.2. 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).

3.2.3. 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.

3.2.4. 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 2.3.2.2). 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 optimization.

3.2.4.1. 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 (interprotein). 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).

3.2.4.2. 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. 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.

3.2.4.3. 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. 3.1). 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).

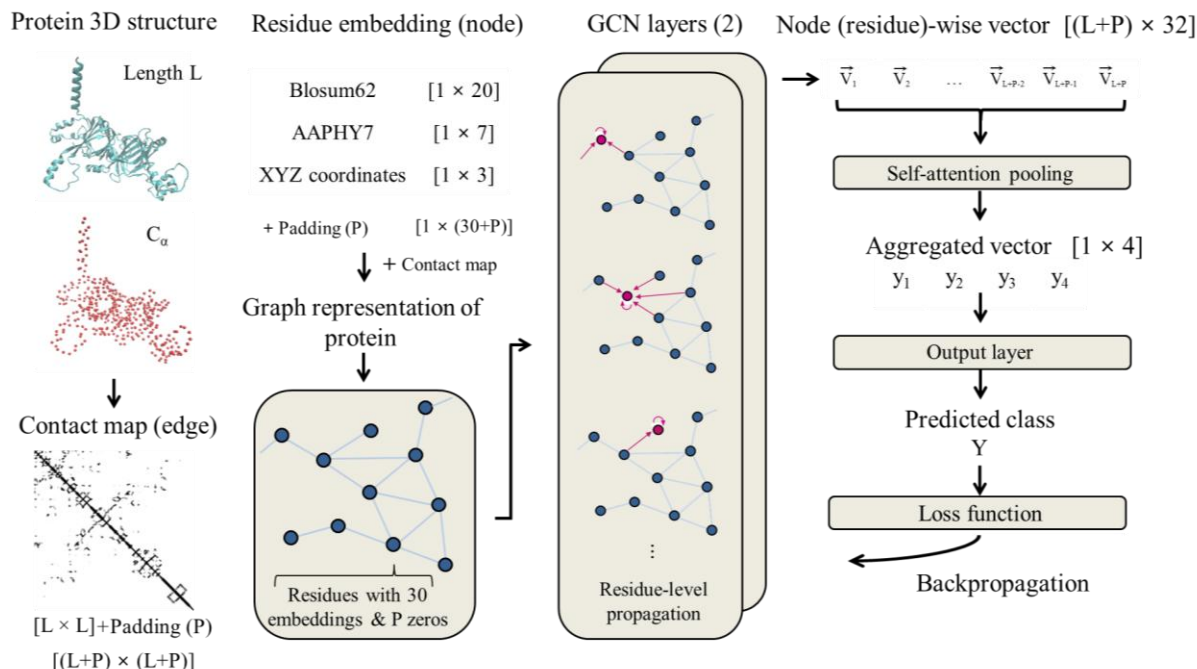


Figure 3.1. Graphical representation of protein structure embedding and architecture of the GCN model.

3.2.5. 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.

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. pdb2gm module in GROMACS was used to assign protonation states.

Molecular dynamics simulations were performed using GROMACS 2022.5 (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).

3.2.5.1. All-atom 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, 2013) using the steepest descent algorithm for 5,000 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 gm module, and counterions were added to neutralize the net charge using the gm 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 300K 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 $4.5 \times 10^{-5} \text{ bar}^{-1}$ and a 5.0 ps pressure coupling constant. The protein structure was position-restrained with restraint strength of $1000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ 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.

3.2.5.2. 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 $700.0 \text{ kJ mol}^{-1} \text{ nm}^{-2}$, 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 ($\epsilon_r = 15$, $r_{\text{cutoff}} =$

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 = $-5.23 \text{ kcal mol}^{-1}$) 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 $-1.95 \text{ kcal mol}^{-1}$ (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 ($\Delta t = 20.0 \text{ fs}$). This was followed by NPT equilibration using the Parrinello–Rahman barostat ($\tau_p = 5.0 \text{ ps}$, compressibility = $4.5 \times 10^{-5} \text{ bar}^{-1}$) for another 50.0 ns. During equilibration, position restraints were applied to the protein heavy atoms with a force constant of $4000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$. 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.

3.3. Results and discussion

3.3.1. Protein characterization

3.3.1.1. Comparative analysis of 38 physicochemical properties

To highlight differences among the four Osborne classes, 38 physicochemical properties were compared (Appendix Table A.9). 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 3.2 (a). For instance, cysteine content (rC) in albumins was classified as a meaningful feature, as its distribution was entirely separated from that of other classes. In contrast, leucine content, although statistically different, had overlapping IQRs and was therefore not considered biologically distinctive (Appendix Table A.9).

Table 3.2 (a) 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 nm²). 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 (Grumezescu & Holban, 2018), as well as high contents of hydrophobic residues (Xing et al., 2023) (Appendix Table A.9), 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 nm²), 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 isoelectric point (6.067 vs 9.090) relative to glutelin (Table 3.2 (b))

Table 3.2. Key physicochemical properties distinguishing Osborne protein classes (a), including features with notable differences between globulin and glutelin (b). Values are expressed as mean $\pm$ 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	4.196	5.918	0.232	7.719	4.689	7.913	4.192
	$\pm 1.236/$	$\pm 1.510/$	$\pm 1.599/$	$\pm 0.173/$	$\pm 3.383/$	$\pm 1.531/$	$\pm 2.836/$	$\pm 2.251/$
	(5.594, 1.627a)	(4.000, 0.000a)	(5.689, 1.654b)	(0.203, 0.253a)	(6.711, 4.464a)	(4.576, 1.341a)	(7.000, 3.735a)	(4.505, 3.388a)
Globulin	0.945	1.642	16.941	0.224	8.031	4.327	8.987	3.864
	$\pm 0.608/$	$\pm 1.280/$	$\pm 3.478/$	$\pm 0.223/$	$\pm 2.529/$	$\pm 0.866/$	$\pm 3.290/$	$\pm 2.030/$
	(1.040, 0.558c)	(2.000, 1.000b)	(16.043, 3.658a)	(0.227, 0.350a)	(7.551, 3.983a)	(4.329, 1.091a)	(8.924, 4.356a)	(3.178, 2.713a)
Prolamin	1.408	1.044	15.591	-0.556	1.460	0.398	0.735	0.246
	$\pm 1.213/$	$\pm 1.575/$	$\pm 3.576/$	$\pm 0.162/$	$\pm 0.836/$	$\pm 0.533/$	$\pm 0.677/$	$\pm 0.473/$
	(0.939, 2.084c)	(0.000, 2.000c)	(16.684, 4.878a)	(-0.635, 0.318c)	(1.361, 0.808b)	(0.357, 0.649c)	(0.670, 0.410c)	(0.000, 0.404b)
Glutelin	1.297	1.956	16.760	-0.004	7.504	3.243	5.945	3.078
	$\pm 0.337/$	$\pm 0.235/$	$\pm 2.073/$	$\pm 0.077/$	$\pm 1.293/$	$\pm 0.825/$	$\pm 1.051/$	$\pm 0.598/$
	(1.261, 0.602b)	(2.000, 0.000b)	(16.453, 0.897a)	(-0.015, 0.048b)	(7.361, 1.089a)	(3.158, 1.151b)	(5.732, 0.684b)	(3.104, 0.728a)

(b)

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

3.3.1.2. 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. 3.2).

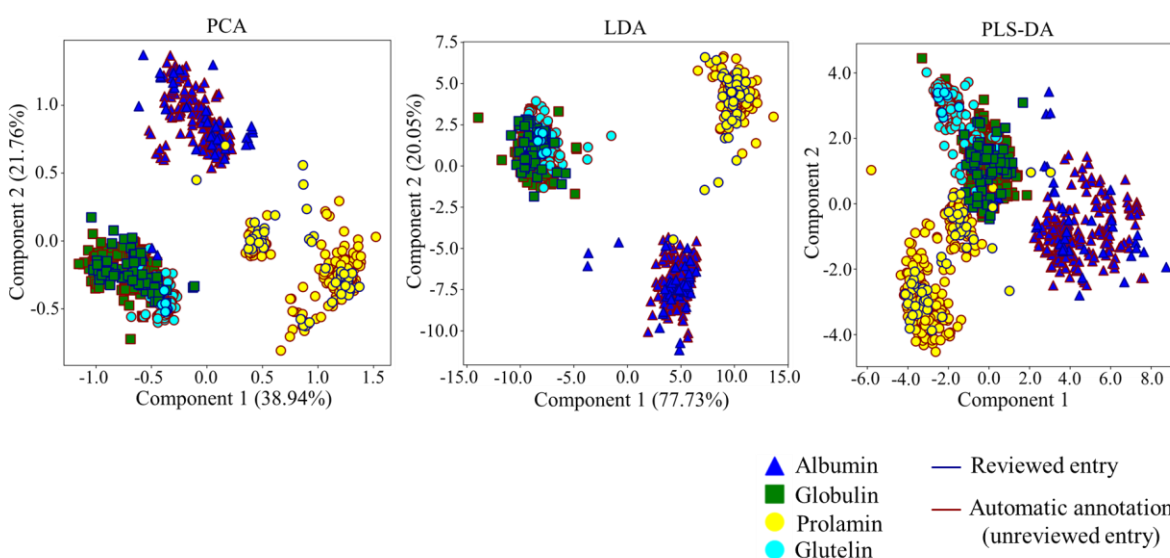


Figure 3.2. Projection of seed storage proteins by Osborne classification using principal component analysis (PCA), linear discriminant analysis (LDA), and partial least squares discriminant analysis (PLS-DA).

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 general similarity of their extracted features also aligned with this idea (Appendix Table A.9). 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 (Appendix Fig.2). The clear boundaries of LDA suggested that simple and linear techniques could be sufficient for broad classification across albumin, prolamin, and globulin/glutelin.

3.3.2. Machine learning classification and interpretation of the model

3.3.2.1. Classification using global features: SHAP analysis & interpretability

As demonstrated in Fig. 3.2, 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 Appendix Table A.10.

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 Appendix Table A.11 and 12, 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. 3.3). 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).

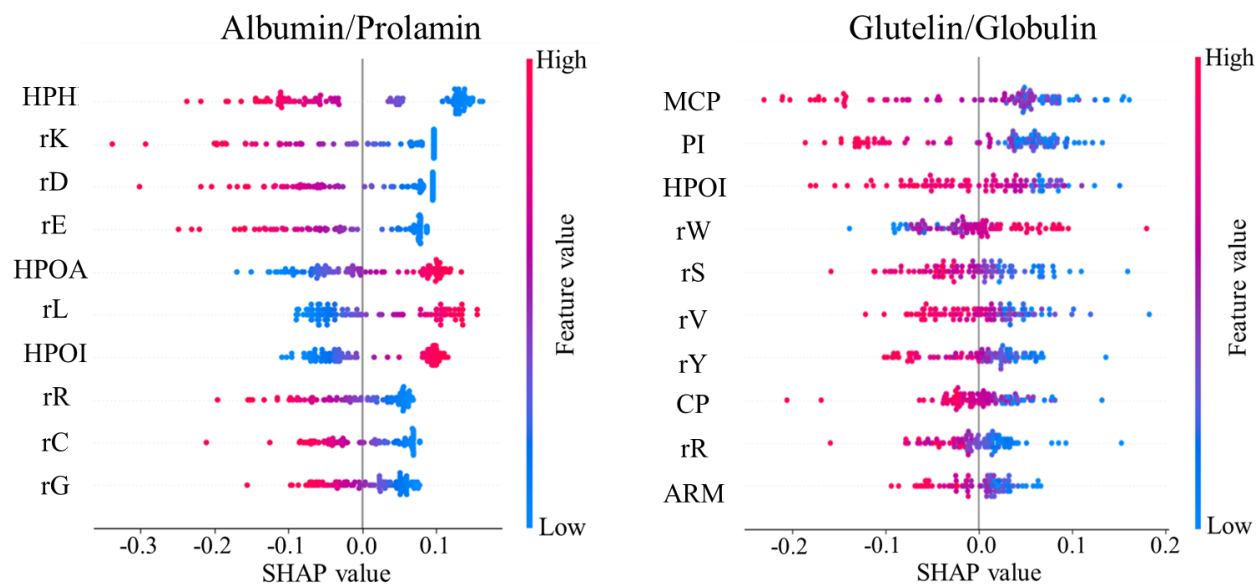


Figure 3.3. SHapley Additive exPlanations (SHAP) scatter plots showing feature importance in seed storage protein classification. Feature values are color-coded from low (blue) to high (red), and SHAP values indicate each feature's impact on the model output for individual predictions.

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 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.

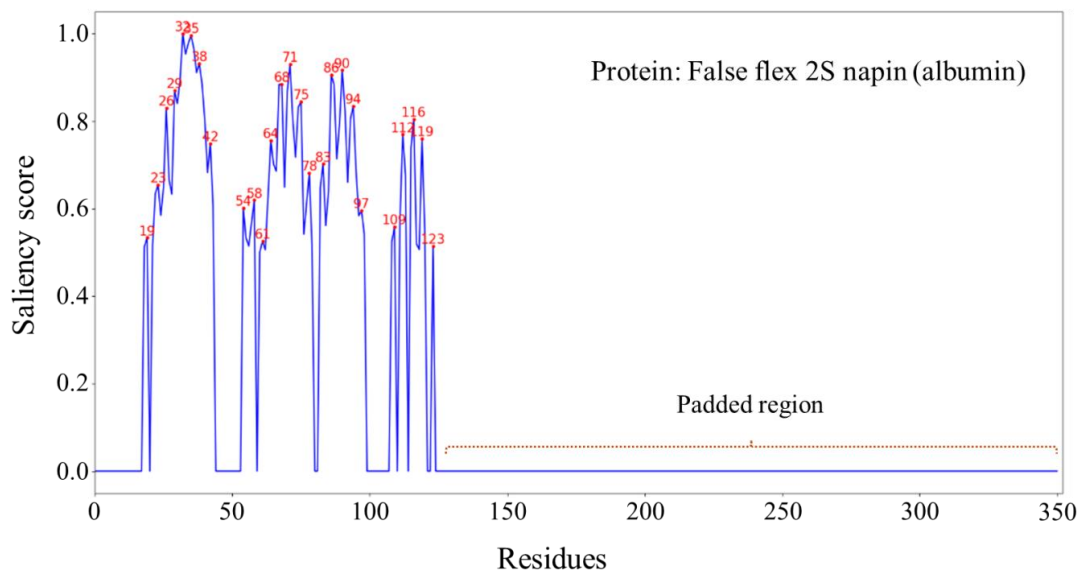
3.3.2.2. Residue-level classification via GCN: Saliency mapping & interpretability

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 train_test split seeds (Appendix Table A.11 and 12).

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. 3.4a.

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. 3.4a), 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. 3.4b).

(a)



(b)

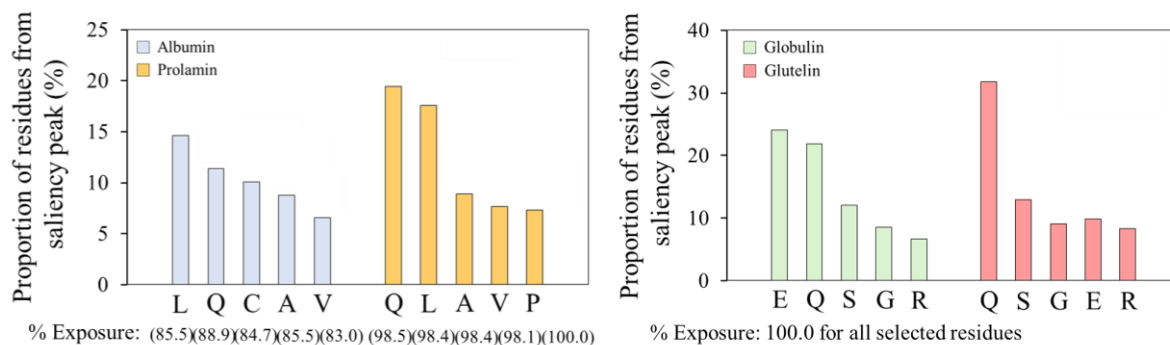


Figure 3.4. Representative saliency score peaks illustrating the identification of key residues for seed storage protein classification (a) and top 5 residue types identified from the score peaks and their relative exposures (b).

3.3.2.3. 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. 3.5a and 3.5b, respectively. No clear association between pattern type and protein origin or functional subclass was observed.

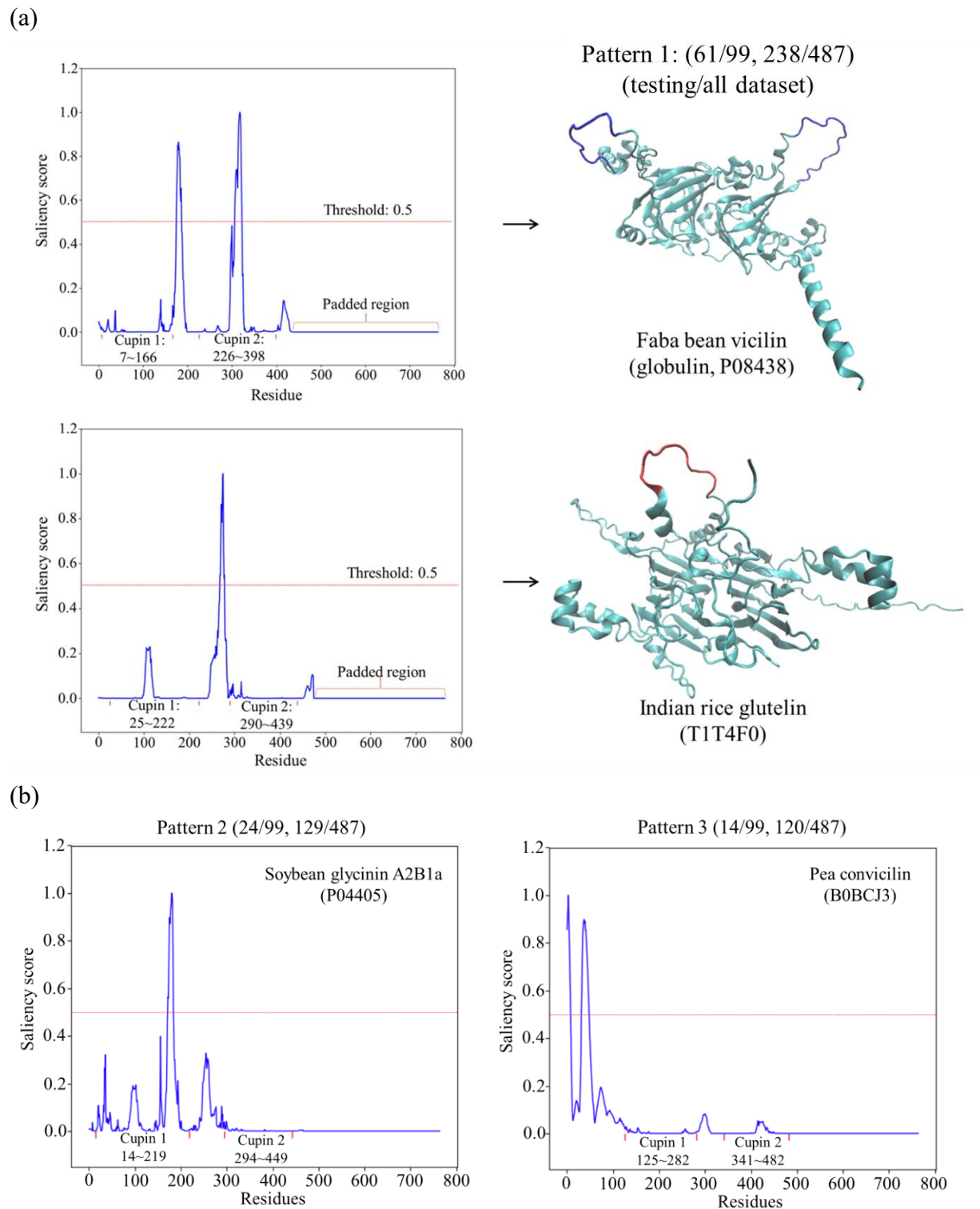
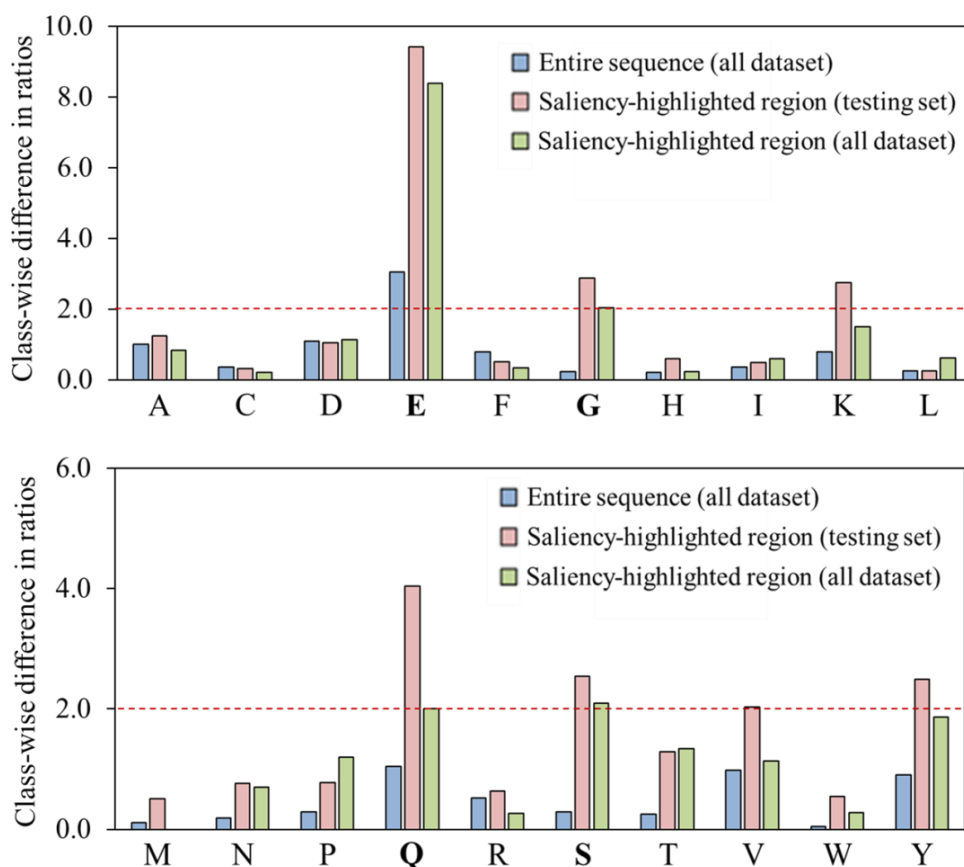


Figure 3.5. Representative saliency scores highlighting high-saliency regions (>0.5), shown with 3D protein structure mapping for pattern 1 (a), and saliency profiles for patterns 2 and 3 (b).

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. 3.6a). 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. 3.6b). 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 and glutelins.

(a)



(b)

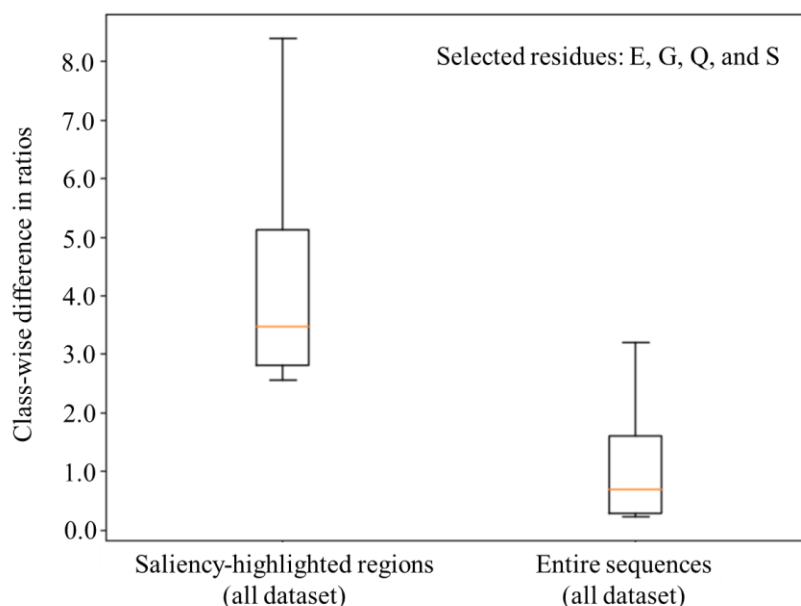


Figure 3.6. Comparison of amino acid composition between globulins and glutelins, focusing on regions highlighted by saliency scores (>0.5) (a). Class-wise differences in the relative abundance of selected residues (E, G, Q, S) based on saliency-highlighted regions versus entire sequences (b).

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 likely 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.

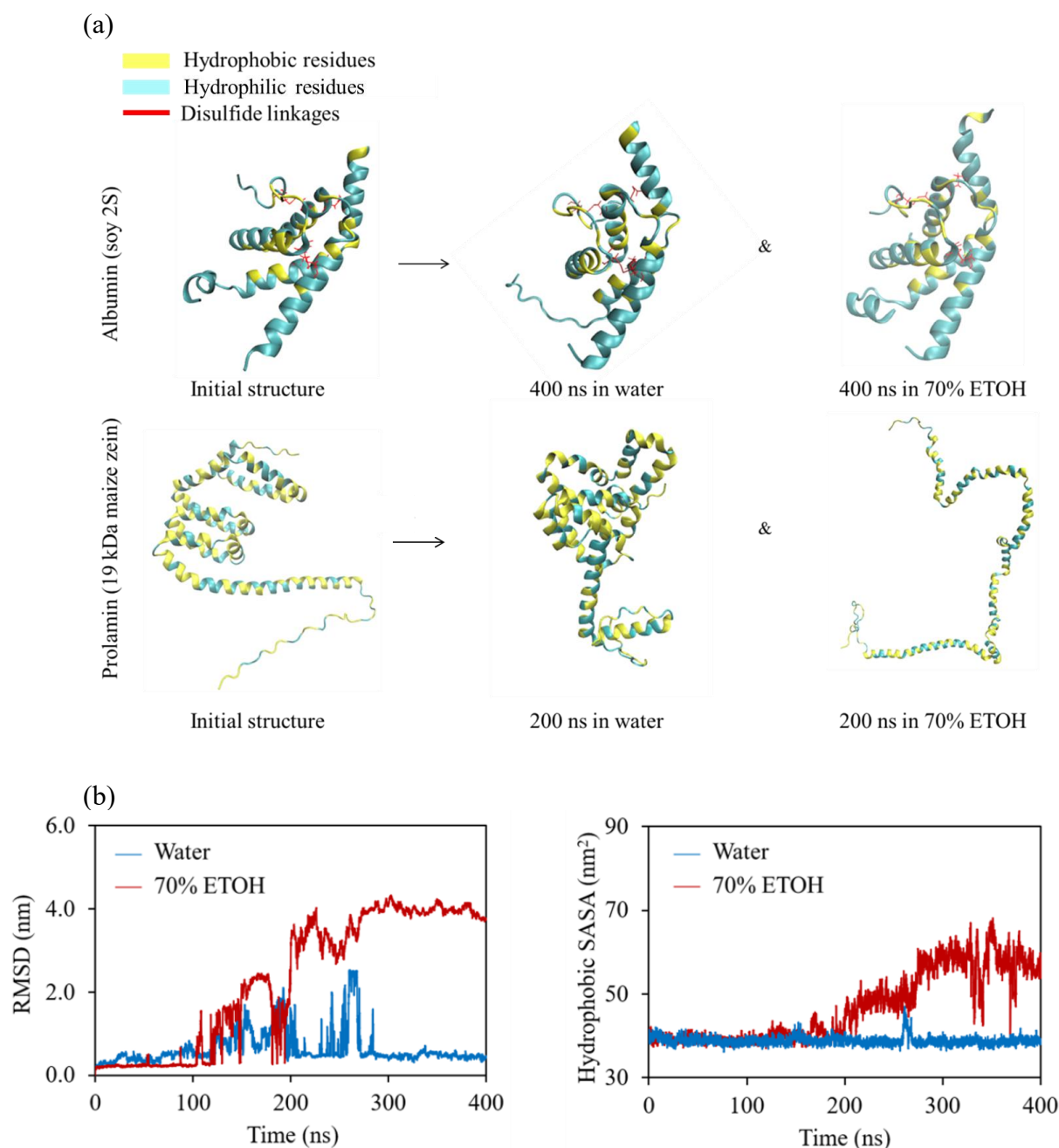
It should be noted that model outputs were interpreted in two distinct ways depending on the protein classes. For albumins and prolamins, the top-ranked SHAP features (e.g., charged residues, hydrophilic residue content, hydrophobic patch area) were biophysically consistent with known mechanisms of solubility 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.

3.3.3. All-atom molecular dynamics simulation

3.3.3.1. 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 subjected to water, only a minor change in the tertiary structure from the initial structure was observed (Fig. 3.7a). 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. 3.7b). In the case of 19 kDa maize zein, more pronounced 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. 3.7c). This expanded structure of 19kDa zein was consistent with 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).



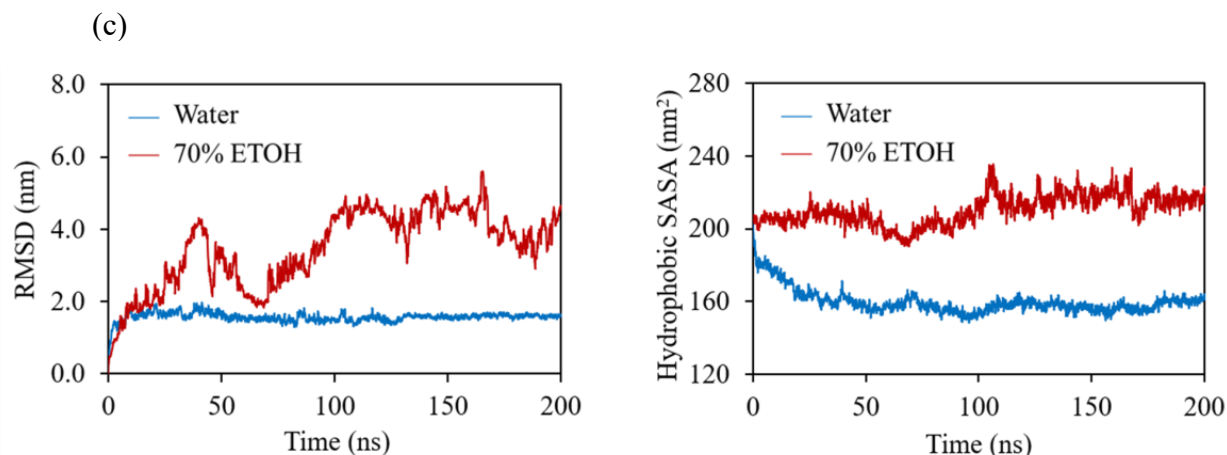


Figure 3.7. Conformational changes of soy 2S albumin and 19 kDa zein in water and 70% ethanol (a) and evolution of root-mean-square deviation (RMSD) and hydrophobic solvent-accessible surface area (SASA) of the albumin (b) and prolamin (c).

To identify the factors driving the observed exposures of hydrophobic residue in 70% ETOH, energy decomposition and hydrogen bonding analysis were conducted. Energy calculations were performed using the `gmx_energy` function, which isolate 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. 3.8a). On the other hand, when the number of intramolecular 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. 3.8).

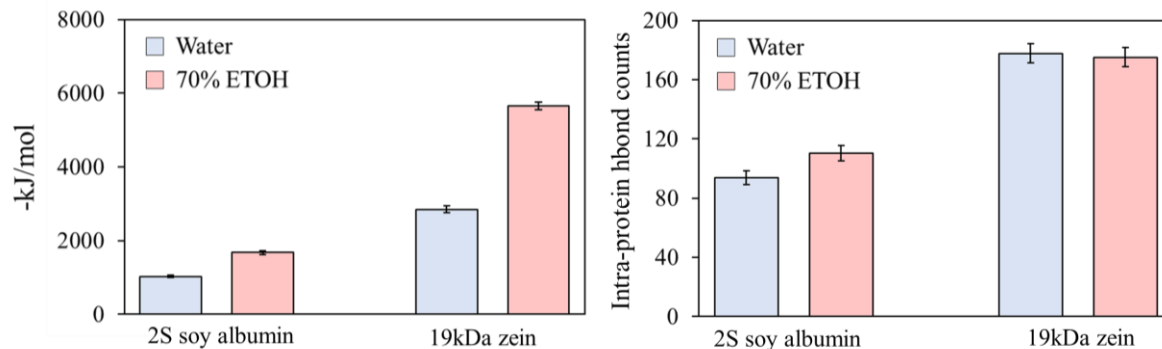


Figure 3.8. Average Lennard-Jones potential energy between protein and system in water and 70% ethanol (left) and intra-protein hydrogen bond counts (right).

As the observed changes in the number of hydrogen bonds (nohb) were counterintuitive to the idea that alcohol would disrupt intramolecular hydrogen bonding (Thomas & Dill, 1993), the evolution of nohb between the solvent and proteins was computed (Fig. 3.9a-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 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. 3.9b,c). Moreover, in 70% EtOH, the decrease in the solvent-protein nohb accompanied an increase in the intra-protein nohb (Fig. 3.9e), suggesting that hydrophilic albumin rearranged to form more intra-molecular hydrogen bonds to maintain structural stability. This observed increase in internal hydrogen bonding in albumin under 70% EtOH aligned with recent MD and CD studies on

biomacromolecules such as spidroins, where ethanol exposure led to enhanced intramolecular hydrogen bonding and helix/turn stabilization (Tolmachev et al., 2023). Furthermore, CD spectroscopy on human serum albumin revealed a recovery of α -helical content at ethanol concentrations above 55% (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. 3.9b, d). However, unlike the case of albumin, this decrease did not accompany gain in the intra-protein hydrogen bonds (Fig. 3.9f). These observations highlighted the differences in solvent adaptation between 2S albumin and 19kDa 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. 3.7 b, c), the increased magnitude of Lennard-Jones potential (Fig. 3.8a), and the stability of the internal hydrogen bonding network (Fig. 3.8b), collectively suggested that hydrophobic interactions, rather than the disruption of internal hydrogen bonds, were the dominant factor driving the observed denaturation in 70% EtOH for the two tested SSPs.

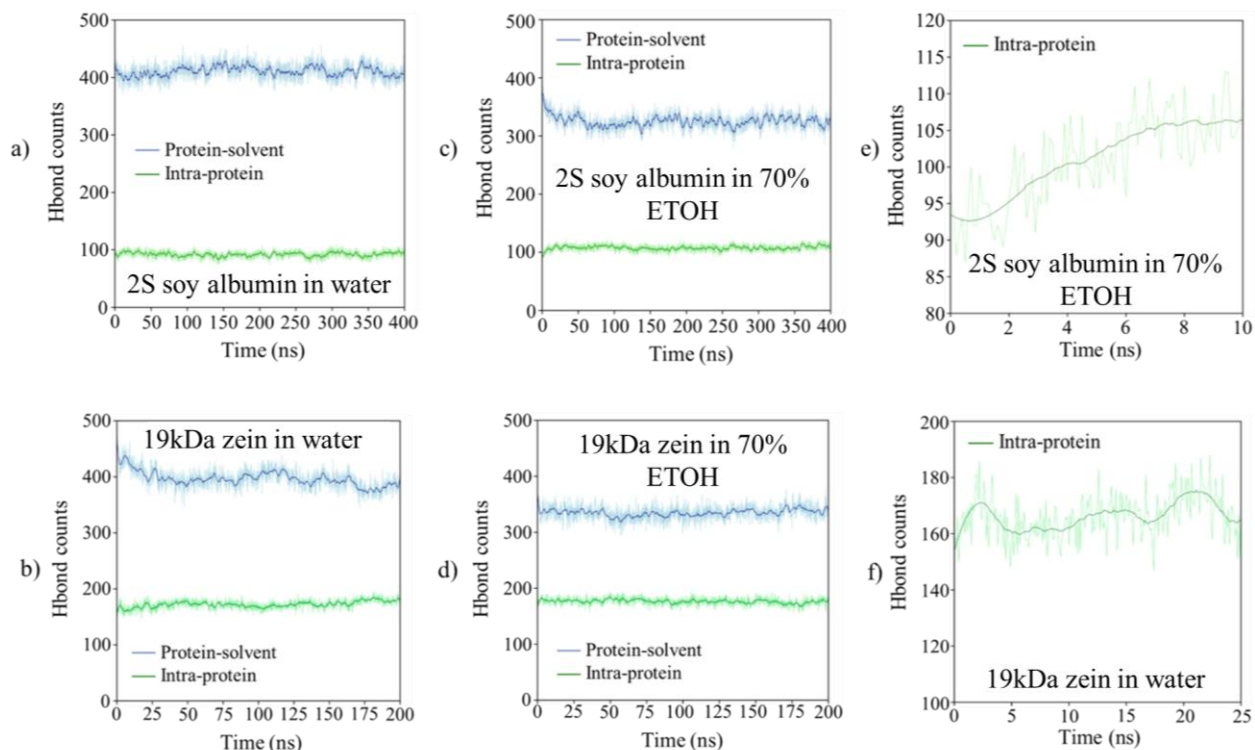


Figure 3.9. Time-evolution of protein-solvent and intra-protein hydrogen bonding in water and 70% ethanol (a–d). Initial changes in the number of hydrogen bonds (e–f) for 2S soy albumin and 19 kDa zein.

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. 3.7b and 3.9, 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 equilibration, and considering the substantial differences observed across solvent conditions, it was regarded that the current trajectories were sufficient to support our conclusions.

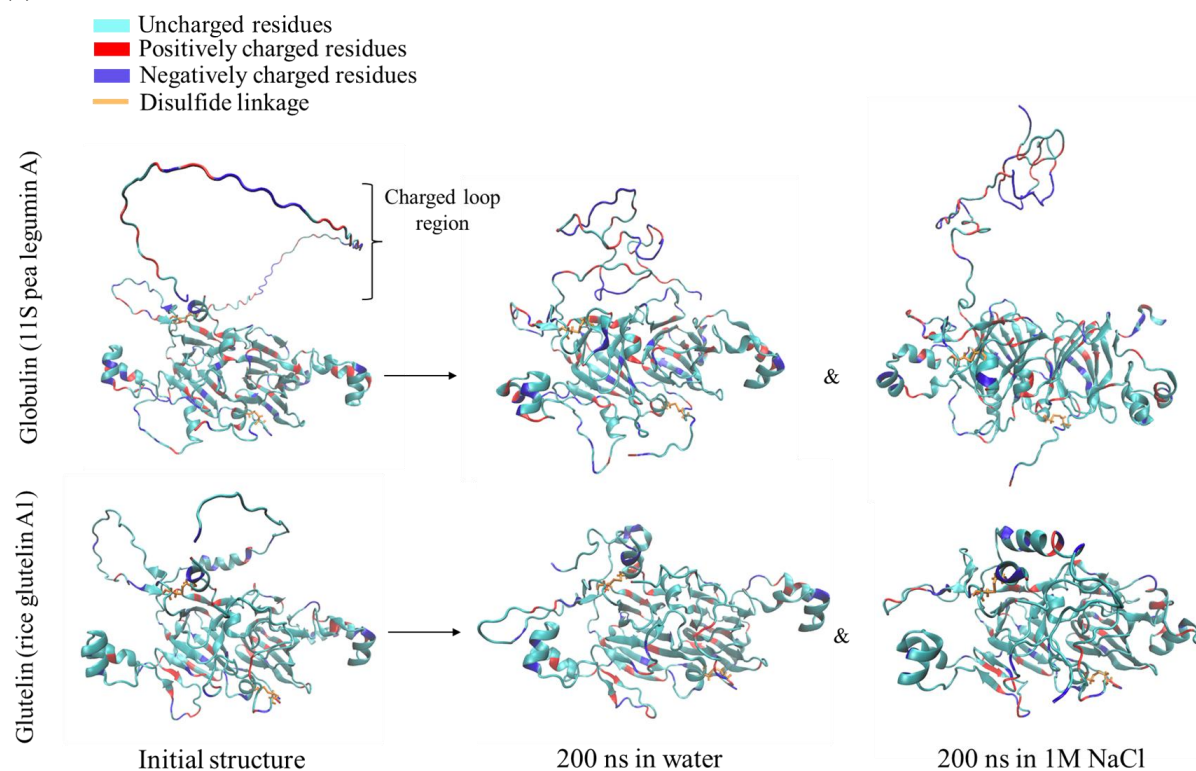
3.3.3.2. Simulation of 11S pea globulin & rice glutelin in water & 1M NaCl

When the structures of pea legumin A (globulin) and rice glutelin A1 were simulated in water and 1M NaCl, no dramatic changes in the tertiary structures were observed for both proteins (Fig. 3.10a). Globulin displayed an increased RMSD in 1M NaCl compared to in water (Fig. 3.10b), 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. 3.10a), 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. 3.10b, c).

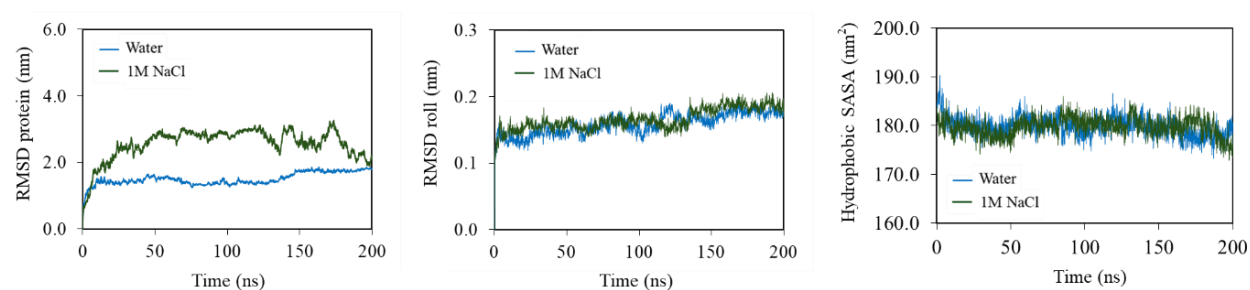
Unlike the addition of hydrophobic solvents, the most direct effect of adding salt would be the disruption of electrostatic interactions between charged residues. To quantify the degree of disrupted electrostatic interactions, the Coulombic potential between the charged residues, as well as the number of salt bridges were computed (Fig. 3.10b,c). As can be seen from the figures, both globulin and glutelin displayed less stabilized coulombic potential in 1M NaCl compared to water. This was also reflected in the decreased number of salt bridges in the saline solution. On the other hand, while 1M NaCl effectively decreased the magnitude of the interaction between charged residues, no significant changes in the nohb of the barrel roll was observed (Fig. 3.10d). Therefore, considering the minimal changes in hydrophobic SASA (Fig. 3.10b,c) and the 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.

(a)



(b)



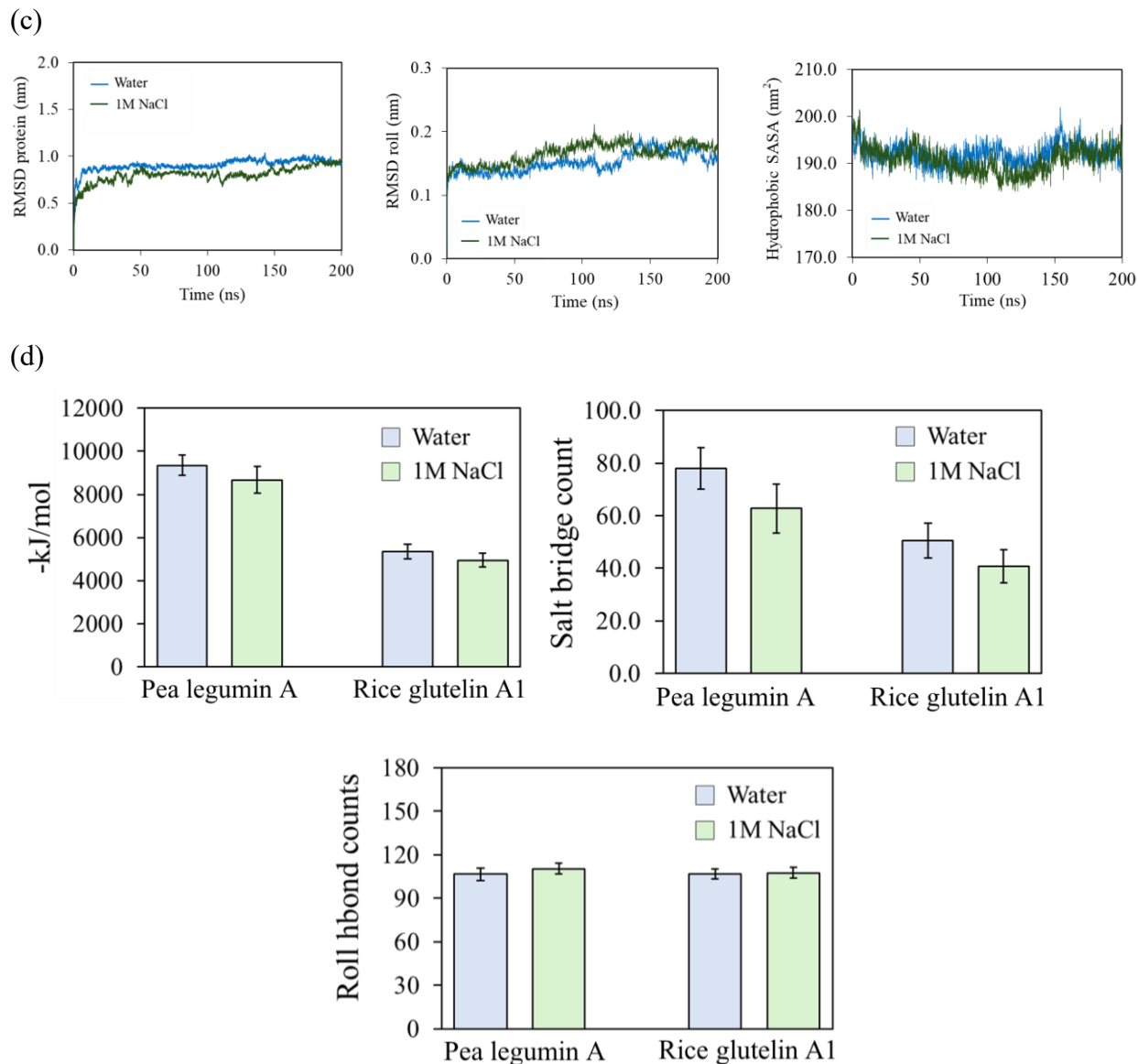


Figure 3.10. Conformational changes of pea legumin A and glutelin A1 in water and 1M NaCl (a), and the evolution of root-mean-square deviation (RMSD) for the protein and barrel roll, along with hydrophobic solvent-accessible surface area (SASA), for globulin (b) and glutelin (c) under the same conditions. The average Coulombic potential between protein and system, salt bridge count, and hydrogen bonding count at barrel rolls in water and 1M NaCl (d).

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. 3.11 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 1M NaCl for both globulin and glutelin. The exception was at the region 3 of pea legumin, where notably higher movements in 1M NaCl were found.

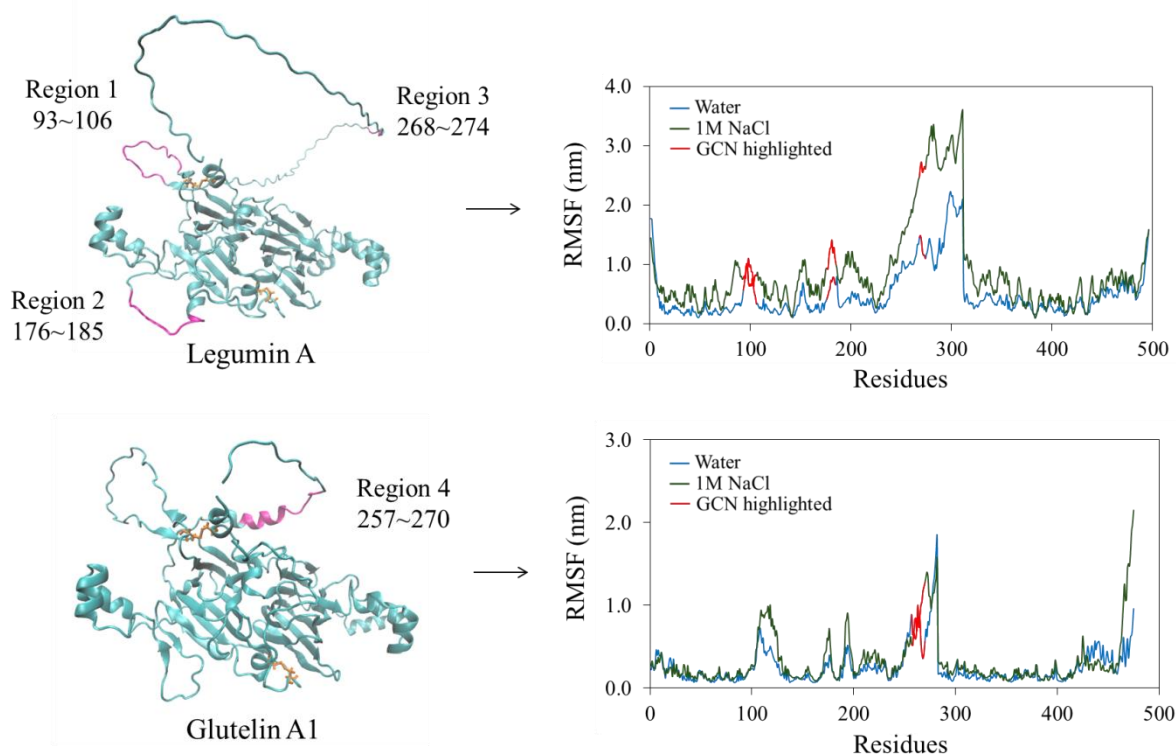


Figure 3.11. Saliency-highlighted regions of pea legumin A and rice glutelin A1 with corresponding root-mean-square fluctuation (RMSF) profiles. The highlighted saliency region is shown in magenta, and disulfide bonds are shown as orange sticks.

This high fluctuation likely stemmed from the highly charged nature of region 3 and the residues nearby (Fig. 3.10a). 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. 3.10a), it was more plausible that the GCN model primarily identified regions with distinct

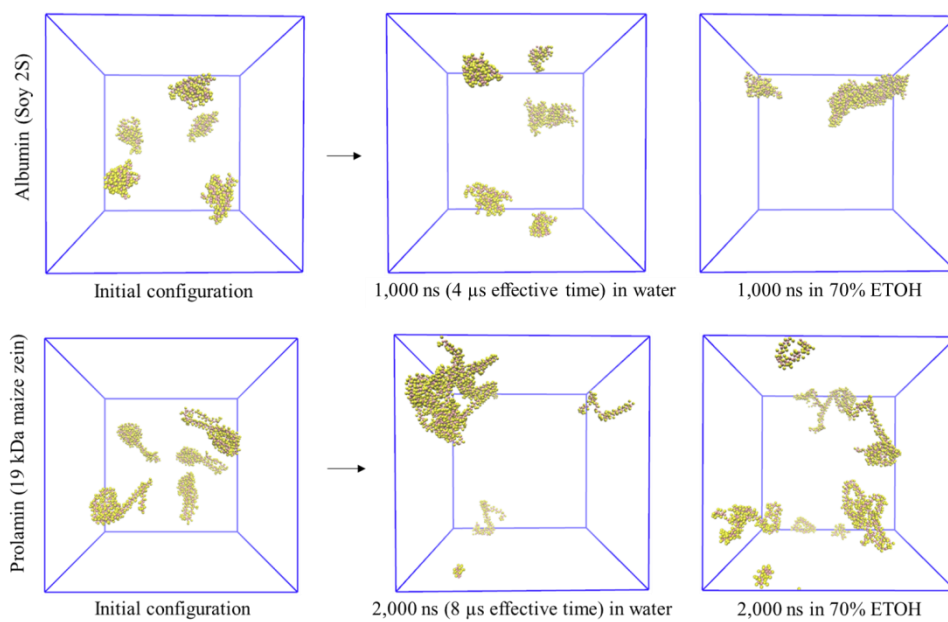
residue compositions rather than those directly contributing to solubility differences. This result suggested that while the GCN model effectively captured sequence variation, additional factors such as intermolecular interactions must be considered to fully understand the solubility differences between globulin and glutelin.

3.3.4. Coarse-grained molecular dynamics simulation

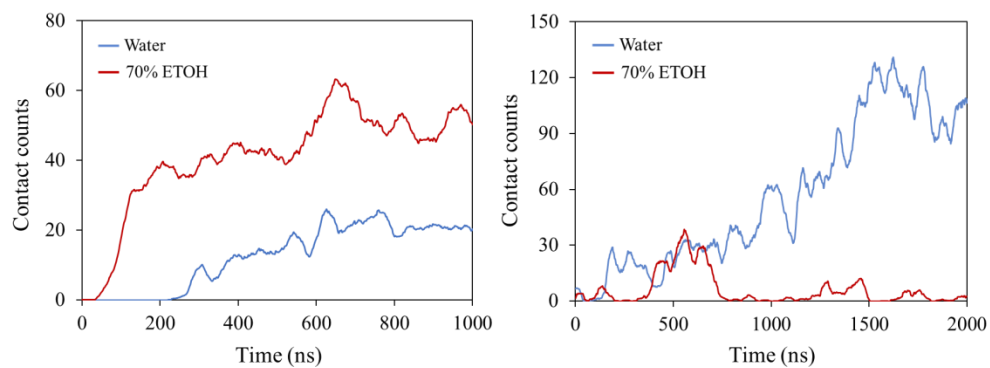
3.3.4.1. Simulation of 2S soy albumin & 19 kDa maize zein in water & 70% EtOH

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. 3.12a, 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. 3.12b). 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 prolamins aggregation.

(a)



(b)



(c)

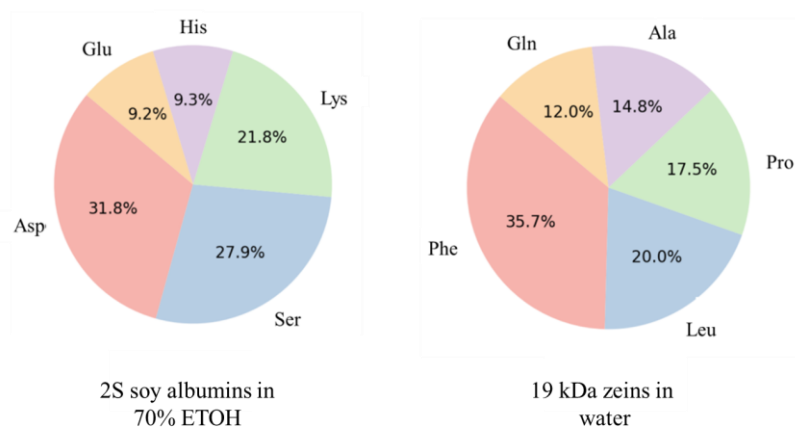


Figure 3.12. Coarse-grained molecular dynamics simulation of multiple soy 2S albumin and 19 kDa zein (a), evolution of contacts (b), and top 5 residues involved in aggregation (c).

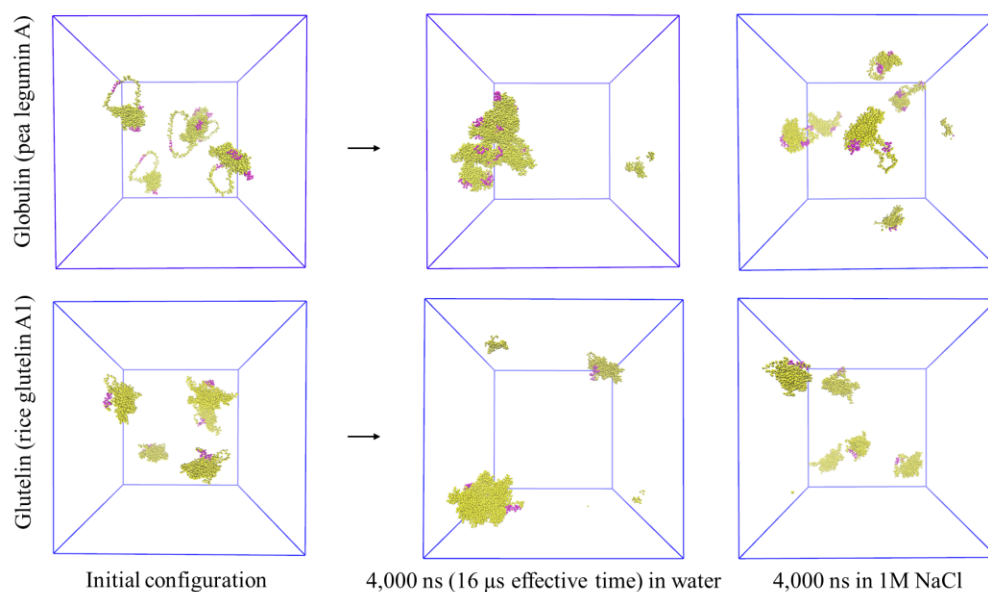
To determine the residues contributing the most to the observed aggregation, residue-specific contacts were computed throughout the trajectory, and the top five residues with the highest degree of contacts were identified (Fig. 3.12c). For albumins in 70% ethanol, charged (Asp, Glu, Lys) and polar (Ser, 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.

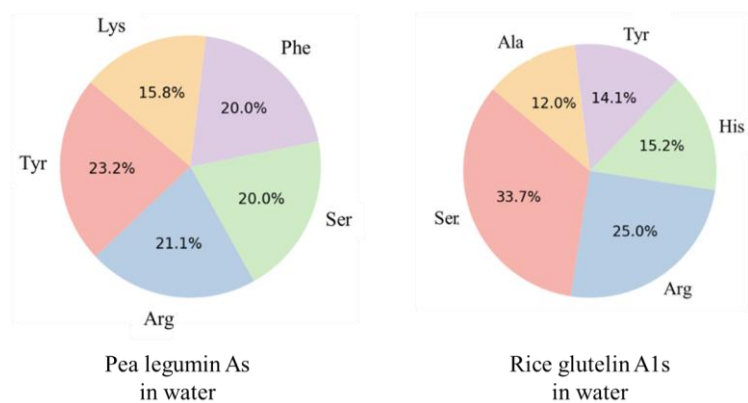
3.3.4.2. Simulation of 11S pea legumin A & rice glutelin A1 in water & 1M NaCl

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 1M NaCl (Fig. 3.13a), suggesting that electrostatic interactions played a crucial role in their clustering. On the other hand, residue-specific contact analysis in water revealed a more diverse interaction profile for globulins and glutelins (Fig. 3.13b). These interactions included electrostatic attractions (Lys, Arg), π - π stacking (Phe, Tyr), hydrogen bonding (Ser, His), and hydrophobic effects (Phe, Ala).

(a)



(b)



(c)

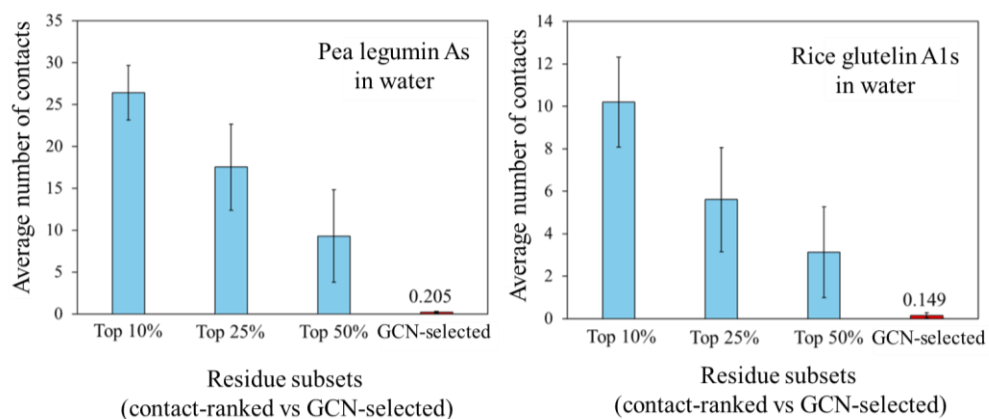


Figure 3.13. Coarse-grained molecular dynamics simulation of multiple 11S legumin A and glutelin A1 with saliency-highlighted regions (purple) (a), top 5 residues involved in aggregation (b), and comparison of contact frequencies between residues with high contact ranks and GCN-selected regions (c).

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. 3.13c, 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 1M 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 interprotein 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.

3.4. 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 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-atom 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.

Acknowledgement

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.

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., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., ... Jumper, 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), 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](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., & Gavligi, H. A. (2025). Exploring the effect of natural deep eutectic solvents on zein: Structural and functional properties. *Current Research in Food Science*, 10, 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., Pinto, B. L., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G. A., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I., Llinares-López, F., ... Bateman, 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., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., ... Karplus, 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), 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., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & 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. *Sustainable 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., Domanski, J., ... & Beckstein, O. (2019). *MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations* (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). *Role of materials science in food bioengineering* (Vol. 19). 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](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](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. DOI: [10.5169/seals-266450](https://doi.org/10.5169/seals-266450)

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... 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 (BBA) - 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), 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, 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](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. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. arXiv. <https://doi.org/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), 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>
- Nakata, N., Okamoto, R., Sumi, T., Koga, K., Morita, T., & Imamura, H. (2023). Molecular mechanism of the common and opposing cosolvent effects of fluorinated alcohol and urea on a coiled coil protein. *Protein Science*, 32(10), e4763. <https://doi.org/10.1002/pro.4763>
- 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.), *Recent advances in computer aided drug designing* (Chapter 10, 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](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, 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., & Casey, R. (1999). Seed proteins. In *Seed proteins* (pp. 1–10). Springer Netherlands. https://doi.org/10.1007/978-94-011-4431-5_1
- 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., 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](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., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C., Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., ... 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., & 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](https://doi.org/10.1016/S0733-5210(05)80002-5)
- 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](https://doi.org/10.1016/S0021-9258(19)74308-7)
- 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](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, 0599. <https://doi.org/10.34133/research.0599>
- Wilson, C. M., Shewry, P. R., Faulks, A. J., & Mifflin, 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>

Chapter 4 - Ethanol-dependent multiscale transitions in zein–lutein encapsulation: From monomer collapse to path-dependent thermal stability

Abstract

This study examined whether ethanol-dependent structural transitions in zein–lutein systems correspond to lutein protection across hierarchical levels. To address this question, ethanol composition was used as a common reference across four levels: monomer-level behavior based on an AlphaFold 3-derived α -zein model, dilute-solution behavior, suspension-level encapsulation, and powder-level thermal stability. At the monomer level, all-atom molecular dynamics simulations predicted α -zein compaction and increased zein–lutein contacts near 40% ethanol. In dilute solution, UV–Vis and particle-size analyses showed loss of solution-like behavior and particle formation in the same range. However, suspension experiments showed that these lower-scale transitions were not directly reflected in lutein protection. Hexane-extraction resistance remained negligible at 70–40% ethanol and emerged only at lower ethanol fractions, where turbidity and particle-size measurements indicated extensive aggregation. Powder-level thermal stability followed a different trend: rapid quenching from high-ethanol conditions resulted in the highest lutein retention (43.84% after 72 h at 60 °C), whereas samples prepared under low-ethanol aggregation-promoting conditions showed lower stability (15.82–18.71%) and the physical mixture showed the lowest retention (6.77%). These findings demonstrate that structure–function relationships in zein–lutein systems do not necessarily translate across hierarchical scales, as lower-scale transitions and suspension-level retention did not directly correspond to stability in the powder state. Together, these results provide a

multiscale perspective on ethanol-dependent zein–lutein assembly and offer insight into scale-aware design of zein-based encapsulation systems.

Keywords: Antisolvent precipitation, Protein-based encapsulation, Molecular dynamics simulation, Solvent-protein interaction, Ethanol-induced transition, Thermal stability

4.1. Introduction

Zein is the predominant storage protein in maize and belongs to a class of alcohol-soluble seed storage proteins (prolamin). Based on their solubility difference, zein proteins can be classified into four major fractions: α -, β -, γ -, and δ -zeins (Argos et al., 1982). Among these fractions, α -zein represents the largest proportion, accounting for approximately 75–80% of total zein (Esen, 1986), whereas β -, γ -, and δ -zeins account for approximately 1–5%, 10–20%, and 1–5%, respectively (Wilson, 1991).

Although zein has limited nutritional value due to poor aqueous solubility and an unbalanced essential amino acid profile (Shukla & Cheryan, 2001), its pronounced hydrophobicity and ability to self-assemble into films, fibers, and nanoparticles provide distinct advantages for encapsulating lipophilic nutraceuticals (Tortorella et al., 2021). Several strategies have been employed for zein-based encapsulation, including chemical crosslinking (Chang et al., 2017), pH-shift methods (Pan & Zhong, 2016), supercritical CO₂ techniques (Hu et al., 2012), and antisolvent precipitation (Wang et al., 2018). Among these, the antisolvent approach has been widely applied due to its simplicity and mild processing conditions (Li et al., 2017). During this process, zein and hydrophobic nutraceuticals are co-dissolved in ethanol solution, followed by the addition of water as an antisolvent to induce self-assembly into colloidal matrices.

During antisolvent processing, zein-containing systems are not exposed to a single fixed solvent condition, but to a changing ethanol–water environment as the ethanol fraction is progressively reduced. This ethanol-reduction pathway can influence not only zein solubility, but also zein structural organization and interactions with co-existing molecules. For instance, ethanol concentration has been reported to modulate the secondary-structure distribution and intermolecular interaction balance of zein (Yu et al., 2020). Changes in ethanol fraction have also been associated with altered glass transition behavior of zein powders (Zhang et al., 2025) and viscosity changes in concentrated zein solutions (Fu & Weller, 1999). In related mixed protein–zein nanoparticle systems, zein has been dissolved in 60–80% aqueous ethanol and then introduced into aqueous protein phases, thereby reducing the ethanol fraction and inducing zein-containing complex formation (Huang et al., 2022; Wang et al., 2025). Beyond zein-based systems, ethanol composition has also been reported to affect other food-protein behavior, including the rheology of whey protein systems (Feng et al., 2021) and the surface hydrophobicity of bovine lactoferrin (Liao et al., 2022) and wheat gliadin (Zhang et al., 2021). These observations suggest that, during antisolvent processing, zein-containing systems may pass through distinct physical states depending on the ethanol fraction.

One of the most extensively investigated hydrophobic ligands in zein-based encapsulation systems is lutein. Despite its well-documented health benefits, particularly in retinal protection (Yang et al., 2020), its conjugated polyene structure renders it highly susceptible to thermal and oxidative degradation, while its pronounced hydrophobicity contributes to poor bioavailability (Mitri et al., 2011). To enhance storage stability or bioavailability, zein–lutein assemblies formed via antisolvent precipitation have been developed

in diverse physical formats, including nanoparticles (Ma et al., 2020), hydrogels (Fang et al., 2025), and powders (Zhang et al., 2023).

Like many other hydrophobic ligands (Liu et al., 2020; Xue et al., 2018), lutein is commonly described as being physically entrapped within a self-assembled zein matrix during antisolvent precipitation (Yuan et al., 2019). In this framework, increasing solvent polarity induces zein self-assembly, during which lutein may become incorporated into the resulting protein-rich matrix. Lutein protection is generally associated with the formation of a barrier between the carotenoid and the surrounding environment (Han et al., 2024; Steiner et al., 2018).

Mechanistic interpretations of protein-based encapsulation, however, are not limited to matrix-level entrapment. Several studies have used computationally derived affinity scores or predicted binding modes from monomer-level structures to interpret encapsulation behavior. For example, molecular docking between pulse proteins and apigenin has been used to relate predicted affinity scores to nanoparticle encapsulation performance (Li et al., 2025). Similarly, a favorable predicted affinity between zein and phloretin has been used to support the interpretation of enhanced ligand stability in zein–phloretin particles (Chen et al., 2026). Whereas these approaches provide useful molecular-level insight, antisolvent precipitation involves self-assembly at a larger structural scale. At this scale, ligand retention may also be influenced by multimeric assembly, rather than by local binding interactions alone.

This distinction is important because it remains uncertain whether a transition observed at one structural level is associated with lutein protection at larger structural scales. In the context of zein–lutein powder formation, the system may involve ethanol-rich solution, suspension-like intermediate, and dried matrix states (Zhang et al., 2023). When examined at different structural levels, ethanol-dependent behavior may appear in different forms, such as the loss of solution-

like behavior in dilute systems or large-cluster formation in suspensions. A relevant question is therefore whether ethanol-dependent transitions observed in monomer-level models or dilute solutions are predictive of lutein protection in more functionally relevant suspension and powder states. Addressing this question would support not only a more integrative understanding of ethanol-mediated zein–lutein assembly but also its rational design.

This study therefore used a common ethanol-fraction axis to examine zein–lutein behavior across different structural scales, from a computationally isolated monomer model to dilute solutions, suspensions, and dried powders. Specifically, an AlphaFold 3-predicted α -zein structure was used as the initial model for all-atom molecular dynamics simulations to examine the ethanol-dependent emergence of zein–lutein contacts. In parallel, dilute solutions were used to locate the range where the system begins to transition toward particle formation. More concentrated suspensions were then examined to determine whether these ethanol-dependent transition ranges coincided with lutein retention. Finally, powders prepared under distinct antisolvent histories were evaluated for thermal stability to determine how transitions from lower scales relate to powder-level protective performance. By evaluating these systems using ethanol composition as a common reference, this study aims to offer a more scale-aware understanding of ethanol-mediated zein–lutein assembly and its functional implications for lutein protection.

4.2. Materials and methods

4.2.1. Materials

Zein was purchased from Sigma-Aldrich (St. Louis, MO, USA). Lutein (90% purity) was obtained from Thermo Fisher Scientific (Waltham, MA, USA). Ethanol (analytical grade) was purchased from Decon Laboratories, Inc. (King of Prussia, PA, USA). Hexane, β -

mercaptoethanol, and other reagents were obtained from Thermo Fisher Scientific (Waltham, MA, USA). All ethanol concentrations reported in this study are expressed on a w/w basis. Zein and lutein concentrations were also defined on a w/w basis across the systems prepared. Ethanol fraction was used as the common reference variable for comparing solvent conditions across the different experimental and simulation systems.

4.2.2. Pretreatment of zein

Zein (10 g) was first washed with 100 mL ethanol to remove innate pigments including lutein (De Boer et al., 2018). The sediment was then solubilized in 200 mL of 70% ethanol and stirred for two hours with 0.5% β -mercaptoethanol (v/v) to reduce possible disulfide linkages (Christensen, 2024). The solution was then dialyzed against deionized water (MW cutoff: 3.5 kDa) at 4 °C for 3 days to remove the reducing agent as well as ethanol. The dialyzed solution was subsequently centrifuged and freeze-dried.

As an additional check for residual β -mercaptoethanol after dialysis, DTNB assay (Ellman, 1959) was performed. Specifically, β -Mercaptoethanol standards were prepared over the range of 0–100 μ M in 0.1 M sodium phosphate buffer (pH 8.0), which exhibited a linear absorbance response at 412 nm ($R^2 = 0.9997$). Untreated commercial zein and β -mercaptoethanol-treated/dialyzed zein powders were extracted by suspending 100 mg of powder in 10 mL of the same buffer, followed by shaking for 30 min, centrifugation, and filtration through a 0.22 μ m membrane filter. After correction for sample-only absorbance and DTNB reagent background, the extractable DTNB-reactive thiol concentrations were $1.07 \pm 0.46 \mu$ M for commercial zein and $1.17 \pm 0.54 \mu$ M for pretreated zein. The pretreated zein extract therefore

did not show a noticeable increase in extractable thiols, supporting that residual β -mercaptoethanol was sufficiently removed.

4.2.3. Preparation of dilute zein–lutein solutions

For dilute-solution experiments, samples were prepared from a common ethanol-rich zein stock solution rather than by independently dissolving zein and lutein at each ethanol fraction. This design allowed the ethanol fraction to be adjusted from a shared starting condition while keeping the zein and lutein amounts constant across samples.

Briefly, a zein stock solution was prepared by dissolving 30 mg of pretreated zein in 30.0 g of 70% ethanol. For zein–lutein samples, 1.00 g of this zein stock solution was mixed with 224 μ L of a 1.0 mM lutein stock solution prepared in absolute ethanol, corresponding to approximately 0.18 g. The ethanol contribution from the lutein stock solution was included in the mass-balance calculation. For each target ethanol fraction, 8.82 g of an ethanol–water diluent with a calculated composition was slowly added to obtain the designated ethanol fractions of 70, 60, 50, 40, and 35% at a final total mass of 10.0 g. This preparation fixed the zein concentration at 0.10 mg/g and the lutein amount at 0.224 μ mol per 10.0 g sample. Zein-only controls were prepared using the same procedure, with the lutein stock solution replaced with absolute ethanol. Lutein-only controls were prepared by replacing the zein stock solution with 70% ethanol. After thorough mixing, all samples were allowed to equilibrate at room temperature for 1 h prior to measurements.

4.2.4. UV–vis absorption spectra

UV–vis absorption spectra of zein–lutein samples, zein-only controls, lutein-only controls, and ethanol–water blanks at the corresponding ethanol fractions were recorded using a double-beam spectrophotometer (UV-6300PC, VWR International, Radnor, PA, USA) over a wavelength range of 300–600 nm. For comparison of lutein-associated absorption, lutein-only spectra were corrected using the corresponding ethanol–water blank, whereas zein–lutein spectra were corrected by subtracting the corresponding zein-only spectrum at the same ethanol fraction.

4.2.5. Particle size measurement

Particle-size distributions were measured by dynamic light scattering (DLS) using a Zetasizer Nano ZS instrument (Malvern Panalytical, Malvern, UK) at 25 °C. Solvent refractive index and viscosity parameters were adjusted according to the ethanol fraction using literature-based values for ethanol–water mixtures on a w/w basis (National Research Council, 1926–1930; Nowakowska, 1939). Z-average diameter, polydispersity index (PDI), and particle-size distributions were obtained from the Zetasizer Nano software (version 3.30).

4.2.6. Preparation of zein-lutein suspension

Zein and lutein were co-dissolved in 70% ethanol at concentrations of 2.5 mg/g (zein) and 0.1 mg/g (lutein) to obtain a homogeneous precursor solution. For each condition, 25.0 g of the precursor solution was transferred to a round-bottom flask and subjected to rotary evaporation (30 °C, 150 mbar, 120 rpm). The remaining mass was tracked at 30 s intervals to monitor solvent loss. Within the 70–20% ethanol range, the total mass decreased linearly (Appendix Fig. 3, $R^2=0.9995$), allowing estimation of intermediate solvent compositions based

on gravimetric tracking under the assumption of preferential ethanol evaporation (O'Hare et al., 1993). Accordingly, evaporation was continued with gravimetric monitoring until the remaining mass reached the calculated target value corresponding to each ethanol fraction: 18.75, 15.00, 12.50, 10.71, and 9.38 g for 60, 50, 40, 30, and 20% ethanol, respectively. This endpoint mass was controlled to within 2% of the target mass for each sample.

After evaporation, the flask walls were rinsed with 2.0 g of 70% ethanol to recover any possible residual material. The rinse was combined with the suspension, and its ethanol contribution was included in the final mass-balance calculation. Each sample was then brought to 25.0 g using an ethanol–water diluent of calculated composition to restore the designated ethanol fraction. The resulting suspensions were allowed to equilibrate at room temperature for an hour prior to measurements.

4.2.7. Measurement of solvent extraction resistance and turbidity

The prepared suspensions were divided into 12.5 g aliquots and extracted with an equal mass of hexane (1:1, w/w). The mixtures were vortexed vigorously for 10 minutes, centrifuged for 10 minutes under 4500 rpm, and the supernatant was collected. This extraction step was repeated three times. Lutein concentration in the hexane phase was determined using a calibration curve constructed from lutein standards in hexane. To account for solvent-dependent variations in lutein partitioning between ethanol–water mixtures and hexane, control samples without zein were prepared at each ethanol composition and subjected to the same extraction protocol (Appendix Fig. 4). The hexane-extractable lutein from zein-containing suspensions was normalized to the total extractable lutein measured from the corresponding ethanol-only controls at the same solvent composition (Equation (1)).

$$\text{Extraction resistance (\%)} = 100 \times \left(1 - \frac{M_{zein}}{M_{control}}\right) \quad (1)$$

Where M_{zein} and $M_{control}$ denote the amount of lutein in the hexane extracts obtained from zein-containing suspensions and lutein-only controls at the same solvent composition, respectively.

Turbidity of the suspensions was assessed by measuring absorbance at 650 nm using a UV–vis spectrophotometer (UV-6300PC, VWR International, Radnor, PA, USA), following a previously established method (Hernandez et al., 1991).

4.2.8. Confocal laser scanning microscopy

Confocal laser scanning microscopy (CLSM) images of the suspension were acquired using a Zeiss LSM 5 PASCAL confocal microscope (Carl Zeiss, Jena, Germany) equipped with a Plan-Apochromat 100×/1.4 NA oil immersion objective. Images were recorded at 1024×1024 pixel resolution (pixel size: 0.09 μm) under identical acquisition settings for all samples. Although excitation was initially applied to detect lutein fluorescence, signals were not resolved. Therefore, structural morphology was evaluated using the refractive detection mode. Single optical sections were collected with a pinhole diameter of 150 μm .

4.2.9. Molecular dynamics simulation

Because α -zein constitutes the largest fraction of whole zein, it was selected as the representative zein type for molecular simulations. As no experimentally resolved structure of α -zein is available, the predicted structure of 22 kDa α -zein (UniProt accession ID: O48966) generated by AlphaFold 3 (Abramson et al., 2024) was used as the initial model, consistent with a previous simulation study on zein (Devnani et al., 2024). Signal peptide residues (residues 1–

22) were removed prior to structure generation to represent the mature protein. Among the five models generated by the AlphaFold 3 server (<https://alphafoldserver.com>), the structure with the highest overall predicted local distance difference test (pLDDT) score was selected. The uncertainty of the predicted model and the structural equilibration process are discussed in detail in Section 4.3.1.1.

Protonation states of ionizable residues were assigned assuming neutral pH conditions using standard protonation states, while maintaining charged N- and C-termini. All molecular dynamics simulations were carried out using GROMACS 2022.5 (GROMACS development team, 2022). Subsequent analysis was conducted using gmx modules and MDAnalysis python package (Michaud-Agrawal et al., 2011). Visualization was performed using VMD (version 1.9.4) (Humphrey et al., 1996).

4.2.9.1. All-atom molecular dynamics simulation

Prior to introducing lutein, a 1 μ s simulation of α -zein only in 70% ethanol was performed using the parameters described below. The 3D structure of lutein was obtained from a resolved crystal structure that includes lutein (PDB ID: 3PL9) and parametrized using CGenFF (Vanommeslaeghe et al., 2010). The protein structures were embedded in a cubic simulation box, and four lutein molecules were placed randomly inside the simulation box, maintaining a minimum distance of 1.2 nm from the box edges. The AA-MD system contained one α -zein molecule and four lutein molecules, providing multiple possible zein–lutein contact events within a finite simulation box.

Initial energy minimization was performed in vacuum under the CHARMM36 force field (Huang & MacKerell, 2013) using the steepest descent algorithm for 5000 steps. A Verlet cutoff scheme was applied with 1.2 nm cutoffs for both electrostatic and van der Waals

interactions. The system was then solvated with TIP3P_{mod} water molecules and counterions were introduced via gmx genion to neutralize the net charge. A second minimization step was carried out in the solvated state to eliminate unfavorable contacts among protein, solvent, and lutein. For all solvated simulations, hydrogen-containing bond lengths were constrained using the LINCS algorithm (Hess et al., 1997), and long-range electrostatic interactions were treated with the Particle Mesh Ewald (PME) method (Darden et al., 1993). Van der Waals interactions were smoothly truncated using a force-switching scheme.

Equilibration was performed sequentially under constant volume (NVT) and constant pressure (NPT) conditions at 300 K. The Nose–Hoover thermostat was employed with separate temperature coupling groups for protein and solvent ($\tau_T = 1.0$ ps). Each equilibration phase was conducted for 125 ps using a 1.0 fs integration time step. During equilibration, positional restraints ($1000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$) were applied to the protein backbone. Subsequent NPT equilibration was conducted at 1 bar using the Parrinello–Rahman barostat with a compressibility of $4.5 \times 10^{-5} \text{ bar}^{-1}$ and a pressure coupling constant of 5.0 ps. Ethanol was modeled using CHARMM parameters (toppar distribution), consistent with the CHARMM36 force field. Ethanol–water mixtures corresponding to the desired weight fractions were constructed by adjusting the number of ethanol and water molecules prior to minimization. The final configuration obtained at one composition was subsequently re-solvated to the next ethanol condition, followed by the same minimization, equilibration, and production protocol. Production simulations were performed for 200 ns at a given ethanol composition.

For the initial α -zein-only relaxation simulation, the 500–1000 ns interval was used for geometry-based clustering to select a representative solvent-relaxed α -zein conformation from the relaxed ensemble. This broad window was used because the purpose of this step was to select

a reference conformation from the solvent-relaxed trajectory rather than to analyze a condition-specific endpoint response. In contrast, for the subsequent α -zein–lutein simulations, the final 50 ns of each production trajectory was used for structural and contact analyses. This narrower window was selected to compare ethanol-dependent behavior after each system had evolved under the corresponding solvent condition, without averaging the final condition-specific response together with the earlier relaxation period.

4.2.9.2. Coarse-grained molecular dynamics simulation

Coarse-grained (CG)-MD simulations were employed to extend the solvent-dependent interaction regimes identified in all-atom simulations to multi-protein system. Rather than resolving atomistic binding events, the CG framework was used as a mesoscale model to compare relative assembly behavior and morphology across ethanol compositions.

All CG-MD simulations were performed using the Martini 3 force field (Souza et al., 2021). The coarse-grained topology and parameters for lutein were adopted from the work of Martini 3 group (Chiariello et al., 2024). CG-mapping of the protein was conducted using Martinize2.py (Kroon et al., 2025). Prior to multi-protein simulation, α -zein monomer was equilibrated in 70% ethanol for 500 ns. Five of the resulting conformation were then randomly positioned within the simulation box using `gmx_insert_molecules`. Twenty lutein molecules were added to the simulation box, following the same molar ratio with the AA simulation. The system was first energy-minimized in vacuum using the steepest descent algorithm for 10,000 steps. Electrostatic interactions were treated using the reaction-field method ($\epsilon_r = 15$) with a cutoff distance of 1.1 nm. Van der Waals interactions were smoothly shifted to zero at 1.1 nm under the Verlet cutoff scheme.

Following energy minimization, the system was solvated using standard Martini water beads (W). The molar ratio between water and ethanol was determined under the Martini 3 mapping scheme, where four water molecules are represented by one W bead. The N6 bead was used to represent ethanol, following our previous simulations of 19 kDa zein in ethanol solution (Kwon et al., 2025). Following solvation, a second energy minimization was performed. Bond lengths were constrained using the LINCS algorithm. NVT equilibration was carried out at 300 K for 50 ns using the velocity-rescaling thermostat ($\tau_T = 1.0$ ps) with a time step of 20 fs. This was followed by 50 ns of NPT equilibration using the Parrinello–Rahman barostat ($\tau_p = 5.0$ ps; compressibility = 4.5×10^{-5} bar⁻¹). During equilibration, positional restraints were applied to protein heavy atoms with a force constant of 4000 kJ mol⁻¹ nm⁻². Production simulations were subsequently performed without restraints using the identical parameters. Production runs were conducted for 1 μ s, and extended to 5 μ s for systems with ethanol concentrations of 40% and below. This decision was guided by exploratory 5 μ s simulations at 50% ethanol, where protein–lutein contacts remained transient and fluctuating and did not develop sustained high-contact states. In contrast, longer-lived high-contact configurations were observed at 40% ethanol (Appendix Fig. 5).

To preserve solvent-responsive conformational flexibility observed in AA simulations, no elastic network was applied (EN = 0) for CG simulations. Preliminary tests indicated that increasing elastic network strength artificially constrained structural expansion and suppressed solvent-dependent conformational transitions (Appendix Fig. 6). To minimize bias from ligand self-association in the CG model, lutein molecules were re-dispersed prior to each simulation to standardize initial configurations. Upon reaching the solvent regime corresponding to the AA-identified interaction onset (~40% ethanol), lutein configurations obtained after equilibration

were propagated to subsequent lower-ethanol conditions to preserve solvent-history continuity. Excessive aggregation arising from imbalanced non-bonded interactions has been reported in Martini-based coarse-grained simulations (Wang & Zhou, 2023). Furthermore, calibration of CG parameters against atomistic references is a commonly adopted strategy to preserve system-specific conformational shifts (Thomasen et al., 2024). Accordingly, elastic network constraints were not applied, and ligand initialization procedures were adjusted to maintain consistency with solvent-dependent transitions observed in AA simulations. Solvent accessible surface area was computed using the standard van der Waals radii of the Martini 3 beads (i.e. 0.264, 0.230, and 0.191 nm for regular, small, and tiny coarse-grained beads, respectively) (Souza et al., 2021).

4.2.10. Preparation of zein-lutein powder

Zein (3000 mg) and lutein (30 mg) were co-dissolved in 300 g of 70% ethanol to obtain a homogeneous stock solution. For each condition, 50 g of the stock solution was transferred to a beaker and subjected to antisolvent addition to reduce the ethanol fraction to 10%. Two distinct addition routes were employed to evaluate the effect of solvent history. In the fast-quenching condition, 300 g of deionized water was added rapidly (~1 s) under continuous stirring, immediately lowering the ethanol fraction to 10%. In the stepwise condition, water was added dropwise using a burette to gradually decrease the ethanol fraction from 70% to 50%, 40%, 30%, and 20%, with equilibration for 10 min at each designated composition. After reaching the desired ethanol fraction, the remaining water was added rapidly to achieve the final ethanol fraction of 10%. The total mass of each preparation was 350.0 g, ensuring identical final composition and sample mass across conditions. Following antisolvent addition, the suspensions were centrifuged at 6000 rpm for 15 min to collect the precipitated zein–lutein complexes. The

concentrated samples were subsequently freeze-dried, gently pulverized, and passed through a 150 μm sieve to obtain uniformly sized powder fractions for further analyses.

4.2.11. Measurement of thermal stability

Powder samples (15 mg) were placed into sealed centrifuge tubes and incubated at 60 °C in a convection oven for 0, 2, 4, 8, 24, 48, and 72 h. Three independent samples were prepared for each time point to avoid repeated thermal exposure. The amount of lutein was quantified using sequential extraction (Ding et al., 2020) with hexane and 70% ethanol to recover hexane-extractable lutein and ethanol-released lutein, respectively. More specifically, following incubation, each sample was extracted three times with 20 mL of hexane to remove surface-bound lutein. The residual solid was subsequently dissolved in 20 mL of 70% ethanol and subjected to sonication to ensure complete disruption of the zein matrix. The lutein recovered from the ethanol phase was quantified as deeply bound lutein.

4.2.12. Scanning electron microscopic analysis

Powder morphology was observed using scanning electron microscopy (SEM; Hitachi S-3500N, Hitachi High-Technologies, Japan). Powder samples were mounted on SEM stubs using conductive carbon tape and sputter-coated with platinum prior to imaging. SEM images were acquired under high-vacuum conditions using secondary electron detection at an accelerating voltage of 5 kV. Images were collected at 100 $\times$ and 1000 $\times$ magnifications.

4.2.13. Statistical analysis

All wet-lab experiments were conducted in triplicate using different batches. Statistical differences were evaluated by analysis of variance (ANOVA) using the R statistical package at a

95% confidence level, followed by Duncan's multiple range test. Results are presented as mean values $\pm$ standard deviation. For AA-MD simulation, two independent replicas were carried out for 60% and subsequent ethanol concentrations, as zein–lutein interactions were observed from 60%. Replicas were initiated with distinct velocity seeds and lutein placements to enhance sampling robustness. For CG-MD simulation, all simulations were conducted in duplicate with distinct initial orientations and placements of protein and lutein. Because the simulation designs involved sequential re-solvation of the final structure from one ethanol composition to the next to mimic antisolvent processing, the ethanol fractions were not treated as independent condition-wise groups for formal ANOVA. Instead, MD-derived quantities were interpreted descriptively, focusing on whether comparable solvent-dependent trends were reproduced across replicas.

4.3. Results and discussion

4.3.1. Lower-scale structural transition of zein-lutein across ethanol concentration

4.3.1.1. All-atom molecular dynamics simulation of α -zein

In molecular simulations of protein–ligand systems, apparent binding events are often highly sensitive to the initial protein conformation, particularly when the protein exhibits substantial intrinsic flexibility (Dror et al., 2011). In contrast to lutein, a highly conjugated and structurally rigid carotenoid whose conformational degrees of freedom are limited, prolamins such as zeins or hordeins are known to undergo pronounced solvent-dependent structural rearrangements (Devnani et al., 2024). Moreover, the selected model from AlphaFold 3 showed relatively low confidence, with a mean atom-level pLDDT of 48.01 and a mean off-diagonal PAE (predicted aligned error) of 25.25 Å (Appendix Fig. 7), indicating limited local structural

confidence across the model and uncertainty in the relative positioning of residue pairs, respectively (Abramson et al., 2024).

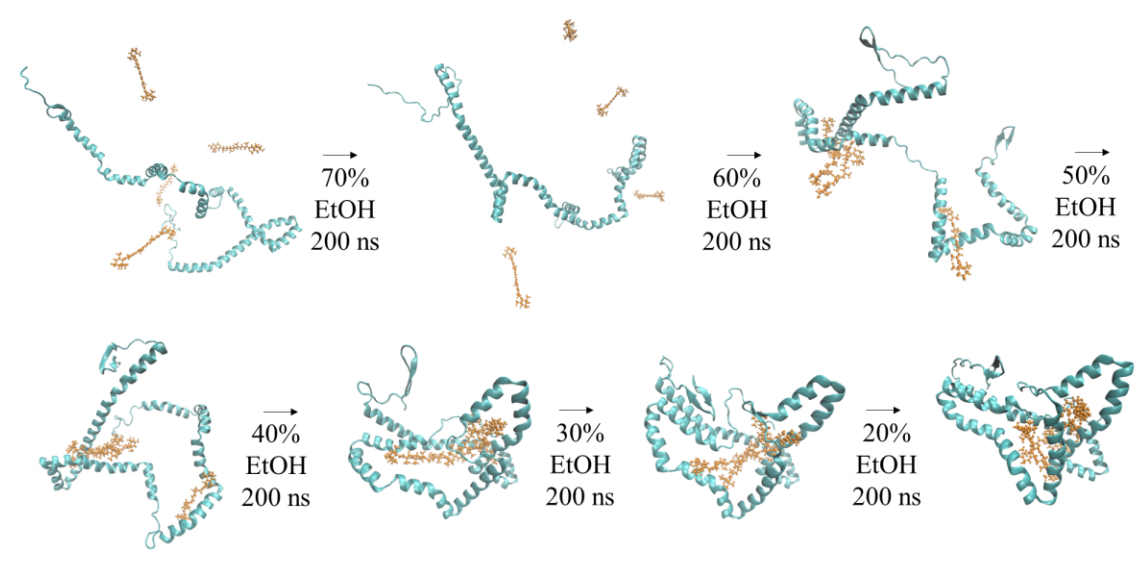
Therefore, before introducing lutein, the derived α -zein structure was first relaxed in 70% ethanol to obtain a solvent-relaxed reference conformation, reducing the likelihood that subsequent zein–lutein interactions reflected bias from non-equilibrated protein conformations. In 70% ethanol, α -zein adopted and maintained an expanded conformational ensemble throughout the 1 μ s AA-MD simulation. No spontaneous compaction was observed, and the expanded state persisted during the latter half of the trajectory (500–1000 ns) (Appendix Fig. 8 a). This behavior was consistent with previous computational studies on 19 kDa α -zein (Kwon et al., 2025), other prolamin (Kumar and Li, 2025), and experimental reports of ethanol-induced swelling (Li et al., 2012). RMSD-based clustering of equilibrated frames identified the most populated conformation between 500–1000 ns, which was selected as the starting structure for subsequent zein–lutein simulations (Appendix Fig. 8b).

4.3.1.2. All-atom molecular dynamics simulation of ethanol-dependent association between α -zein and lutein

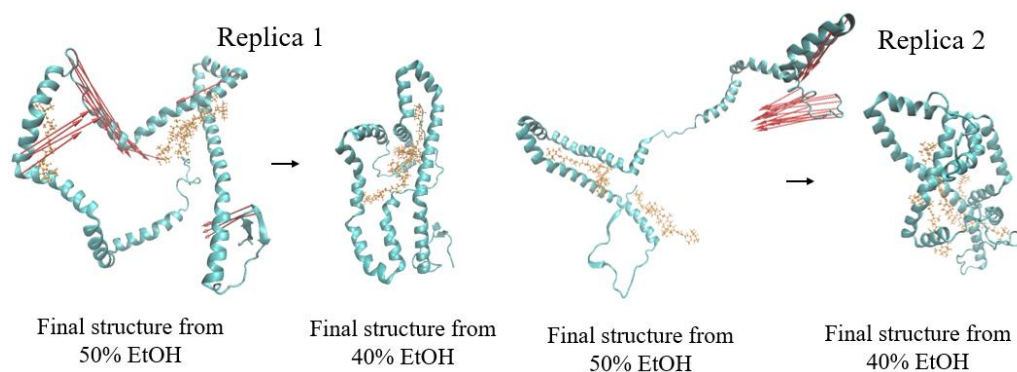
Simulations across ethanol concentrations (70–20%) predicted distinct conformational and interaction regimes in the α -zein–lutein system (Fig. 4.1a). At 70% ethanol, α -zein remained expanded and lutein exhibited only transient and nonpersistent contacts. At 60–50% ethanol, intermittent associations were observed while the protein retained an overall or partially expanded conformation. In contrast, a pronounced structural transition was observed at 40% ethanol, where α -zein adopted a compact collapsed state. This transition was further highlighted by the final snapshots from the two independent replicas (Fig. 4.1b). The arrows indicate the 20 largest C α displacement vectors between the 50% and 40% ethanol structures. Concomitant with

this transition, lutein became more persistently associated with the compact α -zein structure. This collapsed configuration was maintained at lower ethanol fractions (e.g., 30% and 20%).

(a)



(b)



(c)

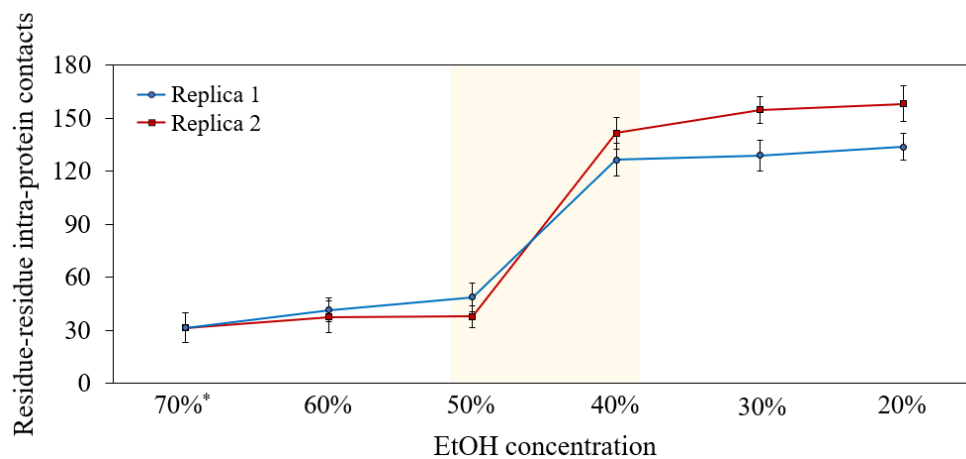


Figure 4.1. Structural evolution of zein–lutein complexes across ethanol concentrations from all-atom molecular dynamics simulations (a). Each system was initialized from the final configuration of the preceding ethanol condition to reflect the gradual solvent transition encountered during antisolvent processing. Endpoint comparison of α -zein–lutein conformations at 50% and 40% EtOH (b). Red arrows indicate the top 20 largest C α displacement vectors between the two endpoint conformations. Nonlocal residue–residue intra-protein contacts calculated over the final 50 ns of all-atom MD simulations at different EtOH concentrations (c). Contacts between residues within six sequence positions were excluded. * The 70% EtOH condition (panel c) is based on a single simulation trajectory, whereas all other conditions are shown for two independent replicas.

Quantitative structural analysis was consistent with this transition, as residue–residue intra-protein contacts increased sharply between 50% and 40% ethanol in both replicas and remained high at 30% and 20% ethanol (Fig. 4.1c). Radius of gyration (R_g) and α -zein–lutein contact analyses further supported this ethanol-dependent compaction, while also showing replica-dependent differences in the intermediate range (Appendix Fig. 9). At 70–50% ethanol, R_g decreased from 4.92 to 3.29 nm in Replica 1, whereas Replica 2 remained more expanded at 4.92–5.14 nm (Appendix Fig. 9a). However, both replicas shifted to lower R_g values at 40–20% ethanol, ranging from 2.53 to 2.45 nm in Replica 1 and from 2.40 to 2.28 nm in Replica 2. Similarly, α -Zein–lutein contacts increased from 3.2–46.6 contacts at 70–50% ethanol to 59.9–76.1 contacts at 40–20% ethanol in Replica 1 and from 3.2–29.8 to 53.3–70.7 contacts in Replica 2 (Appendix Fig. 9b).

The noticeable differences at 60% and 50% ethanol between the replicas were expected because the two replicas were initiated from distinct microscopic initial states, including initial velocity assignments and lutein placements. This design allowed simulations to sample trajectory-to-trajectory variability rather than reproducing near-duplicate trajectories (Pérez et al., 2016). The effect of such variance could be particularly strong for the present system because lutein was not constrained to a predefined binding pocket. Instead, lutein could interact with

different α -zein surface residues depending on its initial placement, as further examined below. The resulting differences in early local contacts could influence subsequent intra-protein rearrangement and contribute to replica-specific R_g and α -zein–lutein contact values. Nevertheless, the key ethanol-dependent transition was reproduced in both replicas: α -zein adopted a compact morphology at 40% ethanol, and this compact morphology was maintained at 30% and 20% ethanol.

To further examine whether the distinction between the 50% and 40% ethanol states persisted, additional 200 ns simulations were performed from the corresponding endpoint conformations. In both replicas, the structures in 50% ethanol remained relatively expanded during the extended simulations, whereas the structures in 40% maintained a collapsed morphology (Appendix Fig. 10). Residue–residue intra-protein contact distributions from the final 50 ns and the additional 200 ns extension simulations were also largely comparable (Appendix Fig. 11). Together with the persistent collapsed morphology at and below 40% ethanol in both replicas, these results showed that the 50% and 40% ethanol states remained structurally distinguishable over the simulated timescale. Therefore, within the two-replica and extended AA-MD dataset, the reproducible trend was a shift from a relatively expanded α -zein morphology at 50% ethanol to a more collapsed monomer-level morphology at 40% ethanol, accompanied by increased α -zein–lutein contacts.

To further characterize the spatial arrangement of lutein in α -zein-lutein complex, residue-level contact frequencies were calculated over the final 50 ns of each trajectory (Appendix Fig. 12). In the expanded regime (60% ethanol), contact-occupancy profiles showed no agreement between independent simulations, indicating an absence of reproducible contact hotspots under expanded conformation. In contrast, upon decreasing ethanol concentration to

$\leq 40\%$, recurrent contact-enriched regions consistently emerged within two helical segments (residues 57–75 and 105–120). Within each replica, these contact-prone regions were largely preserved between 40% and 20% ethanol, suggesting that the collapsed-state interaction pattern was stable across lower ethanol fractions. However, the limited number of simulations and compact conformation prevented clear discrimination between specific binding and nonspecific packing. Instead, what seemed evident from the contact analysis was that a specific and reproducible binding site was not apparent under high-ethanol (expanded) conditions. Therefore, these results suggested that lutein association does not appear to involve a pre-organized binding pocket in the expanded state but instead could be associated with solvent-induced protein collapse.

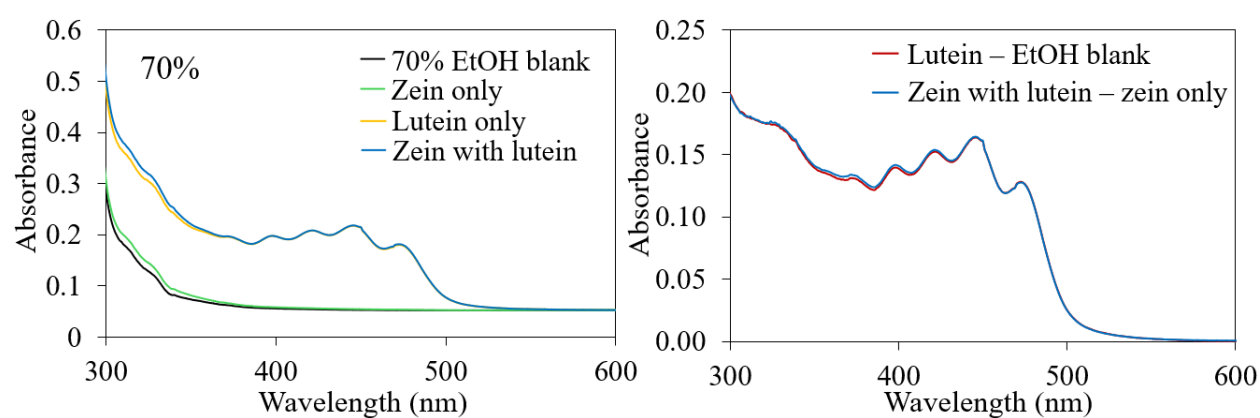
4.3.1.3. UV–Vis spectral transitions in dilute zein–lutein systems

The AA-MD simulations suggested an ethanol-dependent transition of α -zein–lutein systems using monomer model, with surface-associated lutein contacts at 60–50% ethanol and a more collapsed, contact-rich zein state emerging between 50% and 40% ethanol. However, experimental systems are not equivalent to an isolated single-protein model, because ethanol reduction can also promote particle formation and aggregation of zein. Therefore, UV–Vis spectra were collected to examine how ethanol reduction affected the spectral behavior of the dilute zein–lutein solution. To provide a dilute experimental comparison to the isolated-protein conditions used in the AA-MD simulations, zein solutions were prepared at a low final concentration of 0.10 mg/g, as described in Section 4.2.3.

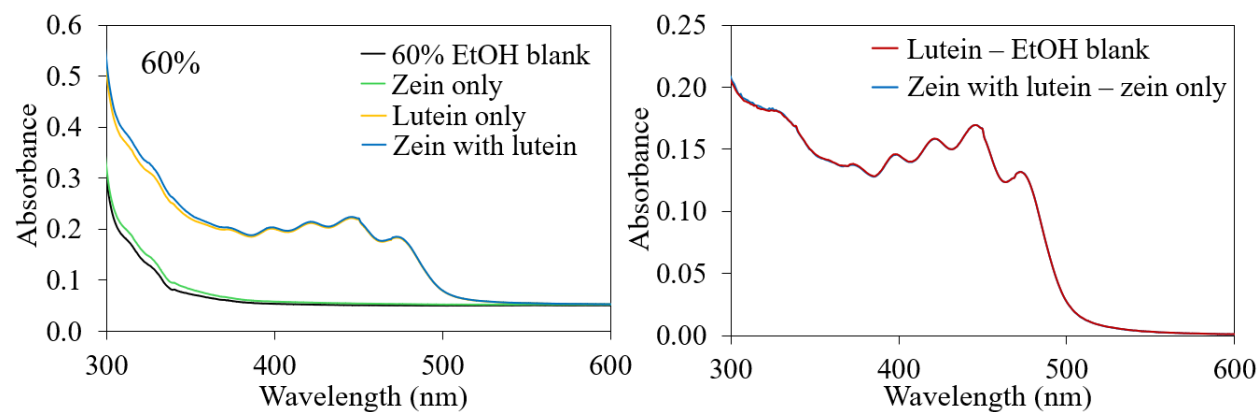
Across the ethanol-rich range of 70–50% ethanol, the UV–Vis spectra showed largely additive spectral behavior between zein and lutein (Fig. 4.2). Across the spectrum collected between 70% and 50% ethanol range, the lutein-only spectrum after subtracting the

corresponding ethanol blank closely overlapped with the zein–lutein spectrum after subtracting the zein-only contribution. These results indicate that, within this ethanol-rich range, the presence of zein did not strongly distort the lutein absorption spectrum, consistent with a relatively solution-like dilute system.

(a)



(b)



(c)

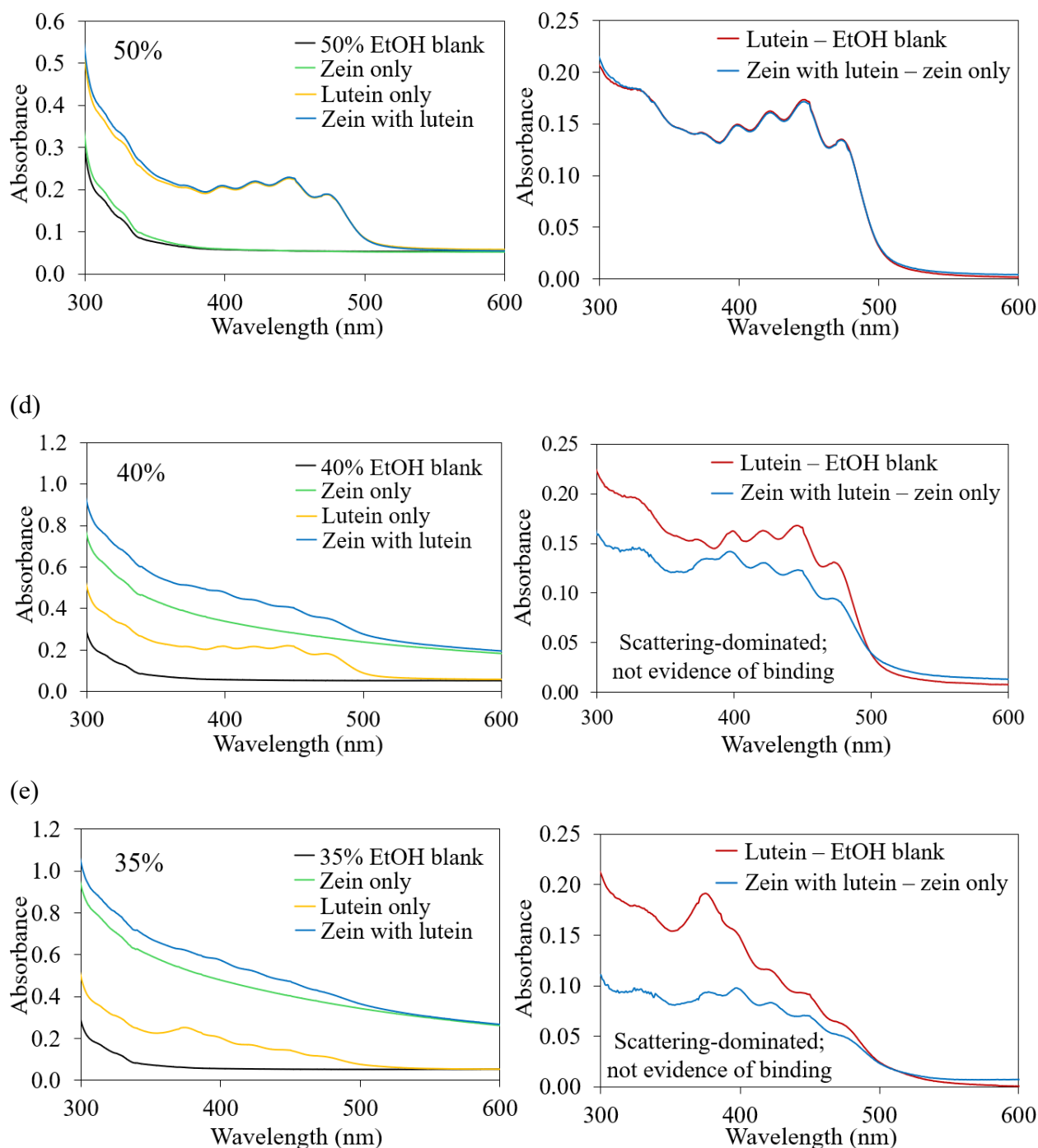


Figure 4.2. UV-Vis absorbance spectra of dilute lutein, zein, and zein-lutein systems at 70–35% EtOH. Left panels show raw spectra of lutein-only, zein-only, zein-lutein, and EtOH blank samples, while right panels show blank-subtracted lutein spectra and zein-subtracted zein-lutein spectra. (a) 70%, (b) 60%, (c) 50%, (d) 40%, and (e) 35% EtOH.

By contrast, at 40% ethanol, the zein-only and zein–lutein spectra showed pronounced baseline elevation and broad spectral distortion, suggesting the emergence of scattering-dominated optical behavior. A further decrease to 35% ethanol intensified this baseline distortion, supporting the ethanol-dependent nature of the spectral changes observed near 40% ethanol. Although the background-corrected zein–lutein spectrum at 40% and 35% deviated from the lutein-only spectrum at corresponding ethanol fractions, such deviations could not be directly assigned to molecular zein–lutein interaction because the raw spectra were strongly affected by baseline scattering (Basalla & Kendrick, 2023). Therefore, the non-additive UV–Vis behavior observed near 40% ethanol was interpreted as evidence for the loss of dilute-solution behavior rather than direct spectroscopic evidence of zein–lutein binding.

4.3.1.4. Particle size analysis of low-ethanol ranges in dilute zein–lutein solution

To assess whether the baseline elevation observed in the low-ethanol UV–Vis spectra was associated with particle or aggregate formation, particle size measurements were conducted on the dilute samples. Between 70% and 50% ethanol, stable DLS particle-size distributions could not be reliably obtained for the lutein-containing dilute samples, preventing direct comparison across all sample types in this ethanol-rich range. Therefore, only the 40% and 35% dilute DLS results are reported in this section. Ethanol-dependent particle-size changes in zein-containing systems across a broader ethanol range are discussed in Section 4.3.2 using suspension systems prepared at a 25-fold higher zein concentration.

At 40% ethanol, stable particle-size distribution measurements were obtained for all three sample types, with Z-average values of 190.37 ± 3.25 nm for lutein-only, 621.57 ± 99.71 nm for zein-only, and 677.97 ± 41.14 nm for zein–lutein samples (Appendix Fig. 13). The corresponding PDI values were 0.20 ± 0.05 , 0.21 ± 0.05 , and 0.26 ± 0.06 , respectively. At 35%

ethanol, the lutein-only particles remained near 200 nm, with a Z-average of 200.10 ± 6.46 nm and a PDI of 0.13 ± 0.01 . In contrast, the zein-only and zein–lutein samples increased to 1137.33 ± 225.39 nm and 1204.77 ± 168.82 nm, with PDI values of 0.27 ± 0.08 and 0.12 ± 0.08 , respectively. These low-ethanol conditions also corresponded to the scattering-dominated baseline distortion observed in the UV–Vis spectra. Thus, together with the UV–Vis results, the DLS data showed that aggregation-associated loss of solution-like behavior became apparent near 40% ethanol in the dilute zein–lutein system.

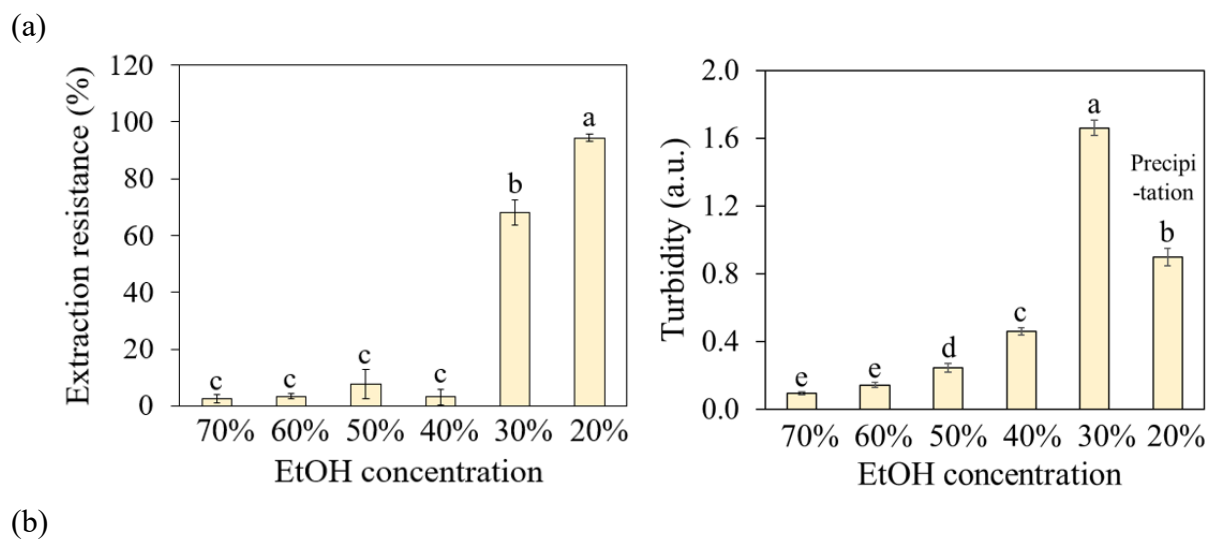
Importantly, this transition toward aggregation in dilute solution should be distinguished from the monomer-level transition predicted by AA-MD, which describes conformational changes within an isolated single-protein model. Thus, the main finding from Section 3.1 was not that the two transitions were physically identical, but that the MD-predicted α -zein-lutein collapse/contact enrichment and the experimental loss of dilute-solution behavior occurred over an overlapping ethanol range near 40% ethanol. This overlap raised the question of whether these lower-scale transitions near 40% ethanol would translate into functional lutein protection in higher-concentration zein suspensions.

4.3.2. Collective assembly behavior of zein–lutein systems across ethanol concentrations

Section 4.3.1 showed that ethanol reduction produced lower-scale transitions near 40% ethanol, including predicted monomer-level compaction and particle formation in dilute solution. Because it remained unclear whether these apparent transitions would translate into lutein retention, suspension-level experiments were conducted across the same ethanol range at a higher protein concentration (2.5 mg/g).

4.3.2.1. Lutein retention under solvent extraction

To assess encapsulation behavior at the suspension level, lutein in zein–lutein suspensions was extracted using hexane. As shown in Fig. 4.3a, hexane resistance remained negligible between 70% and 40% ethanol. In contrast, a marked increase in extraction resistance was observed at 30% and 20% ethanol. This increase was strongly correlated with suspension turbidity ($R^2 = 0.980$, using turbidity value of 2.0 for 20% ethanol suspension to reflect precipitation). Representative CLSM images acquired (Fig. 4.3b) showed that particles remained mostly dispersed at 70% and 50% ethanol. Cluster formation became visible at 40% ethanol, and larger aggregates were observed at 30% ethanol. This trend was consistent with the turbidity increase and the emergence of extraction resistance at lower ethanol fractions.



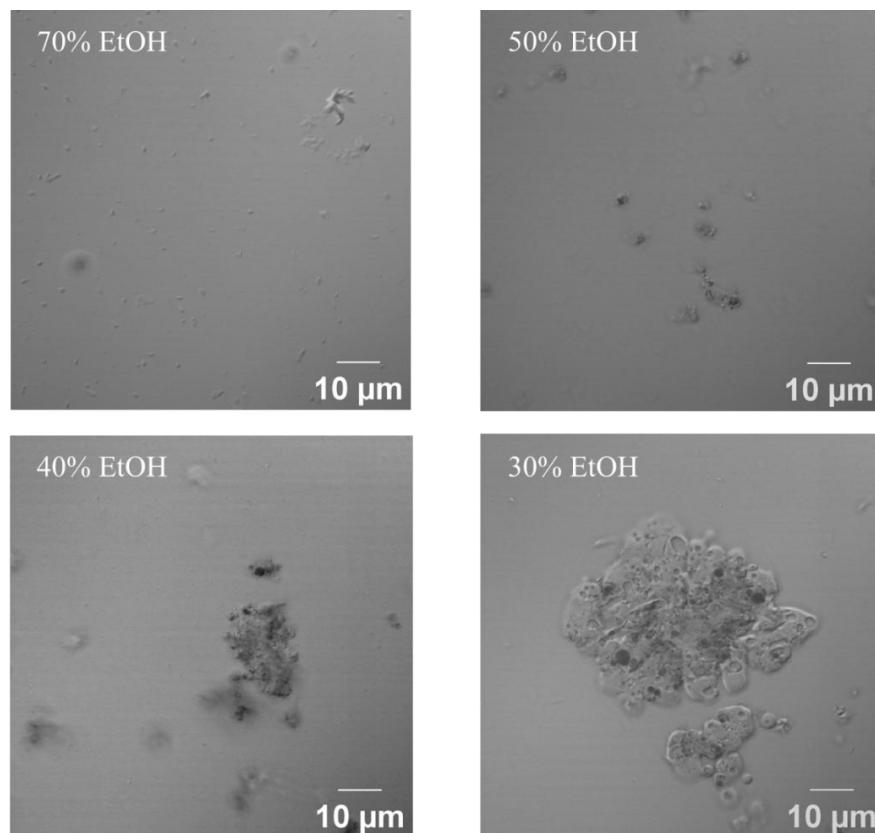


Figure 4.3. Hexane extraction resistance (left) and turbidity (right) of zein–lutein suspension across ethanol concentrations (a). CLSM images of the suspensions across ethanol concentrations (b).

Because lutein fluorescence was not resolved under the CLSM conditions used, as described in Section 2.8, the CLSM images should be interpreted only as refractive/morphological observations of the aggregates, not as direct evidence of zein–lutein co-localization.

Notably, no increase in hexane resistance was detected at 40% ethanol, despite the early lower-scale transition near 40% ethanol identified in Section 3.1. Instead, resistance emerged only at lower ethanol fractions (30–20%). This distinction was further discussed with particle-size distribution analysis in Section 4.3.2.2.

4.3.2.2. Ethanol-dependent particle size distribution in zein–lutein suspensions

To examine whether the ethanol-dependent increase in suspension turbidity was associated with changes in particle-size distribution, DLS measurements were performed on zein–lutein suspensions. As displayed in Fig. 4.4, particle-size distributions demonstrated the ethanol-dependent development of zein-driven aggregation in the suspension. At 70% and 60% ethanol, zein–lutein suspensions were dominated by small particles, with Z-average values of 10.65 ± 0.39 nm and 25.74 ± 2.77 nm, respectively (Fig. 4.4 and Table 4.1). At 50% ethanol, the Z-average increased to 254.70 ± 45.79 nm, accompanied by a higher PDI of 0.39 ± 0.02 . However, the volume distribution indicated that small particles remained dominant and that the increased Z-average was likely influenced by the appearance of a minor larger-aggregate population.

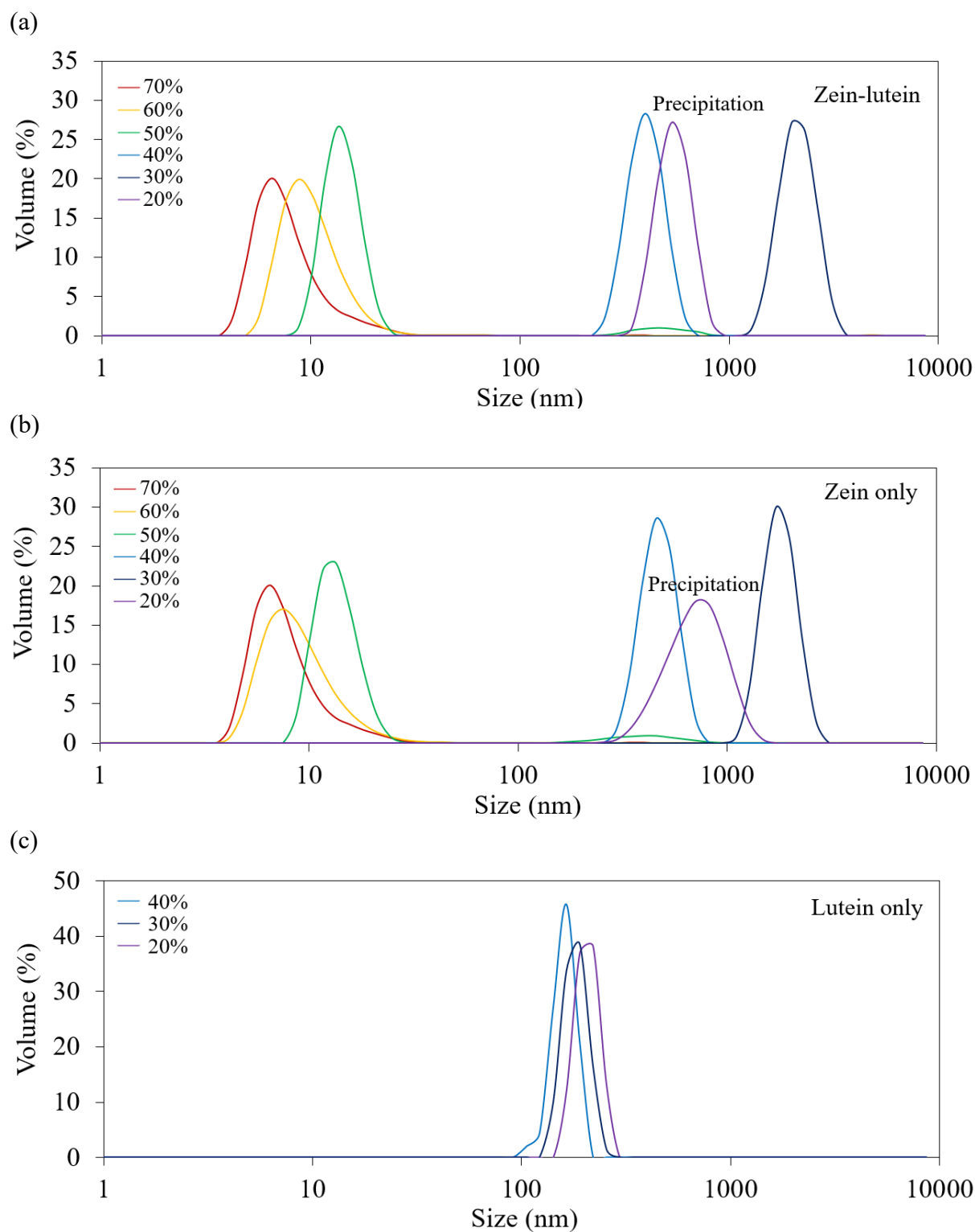


Figure 4.4. Particle size distributions of zein–lutein suspensions and control systems at different EtOH concentrations. Distributions are shown for zein–lutein (a), zein-only (b), and lutein-only (c) systems.

Table 4.1. EtOH-dependent particle size and polydispersity of zein–lutein, zein-only, and lutein-only suspensions at different ethanol fractions. Values are presented as mean $\pm$ SD. Different lowercase letters within the same column indicate significant differences among ethanol fractions for each sample type ($p < 0.05$). Asterisks indicate no significant difference. Statistical analysis of Z-average values was performed after log transformation because particle-size values differed by orders of magnitude across ethanol fractions. PDI, polydispersity index; ND, not determined.

	Z_{avg} (nm)			PDI		
	Zein-lutein	Zein only	Lutein only	Zein-lutein	Zein only	Lutein only
70% EtOH	10.65 $\pm 0.39\text{e}$	11.60 $\pm 0.72\text{e}$	ND	0.22 $\pm 0.04\text{bc}$	0.18 $\pm 0.05\text{bc}$	ND
60% EtOH	25.74 $\pm 2.77\text{d}$	23.97 $\pm 1.73\text{d}$	ND	0.23 $\pm 0.03\text{b}$	0.23 $\pm 0.02\text{bc}$	ND
50% EtOH	254.70 $\pm 45.79\text{c}$	273.27 $\pm 84.85\text{c}$	ND	0.39 $\pm 0.02\text{a}$	0.44 $\pm 0.11\text{a}$	ND
40% EtOH	451.93 $\pm 59.29\text{b}$	490.83 $\pm 33.57\text{b}$	176.13 $\pm 6.67\text{b}$	0.20 $\pm 0.05\text{bc}$	0.17 $\pm 0.03\text{bc}$	0.10 $\pm 0.02^*$
30% EtOH	1920.00 $\pm 256.07\text{a}$	1892.67 $\pm 363.6\text{a}$	203.2 $\pm 12.44\text{a}$	0.38 $\pm 0.09\text{a}$	0.35 $\pm 0.10\text{ab}$	0.09 $\pm 0.02^*$
20% EtOH	444.40 $\pm 53.57\text{b}$	562.73 $\pm 58.25\text{b}$	215.3 $\pm 5.01\text{a}$	0.09 $\pm 0.01\text{c}$	0.12 $\pm 0.04\text{c}$	0.10 $\pm 0.02^*$

A more pronounced size transition was observed beginning at 40% ethanol, where the dominant zein–lutein particle population shifted to the submicron range, with a Z-average of 451.93 ± 59.29 nm. Upon further ethanol reduction to 30%, the particle population shifted to micrometer-scale aggregates, with a Z-average of 1920.00 ± 256.07 nm, indicating extensive suspension-level aggregation. At 20% ethanol, visible precipitation was observed, and the measured Z-average decreased to 444.40 ± 53.57 nm. This decrease was not interpreted as reduced aggregation, but as the remaining dispersed fraction after precipitation. The zein-only suspension showed a similar ethanol-dependent size-shift pattern to the zein–lutein suspension,

with no major difference in particle-size development across the tested ethanol fractions. In contrast to zein-added samples, lutein-only samples remained in a smaller size range at 40–20% ethanol, with Z-average values ranging from 176.13 ± 6.67 nm to 215.3 ± 5.01 nm.

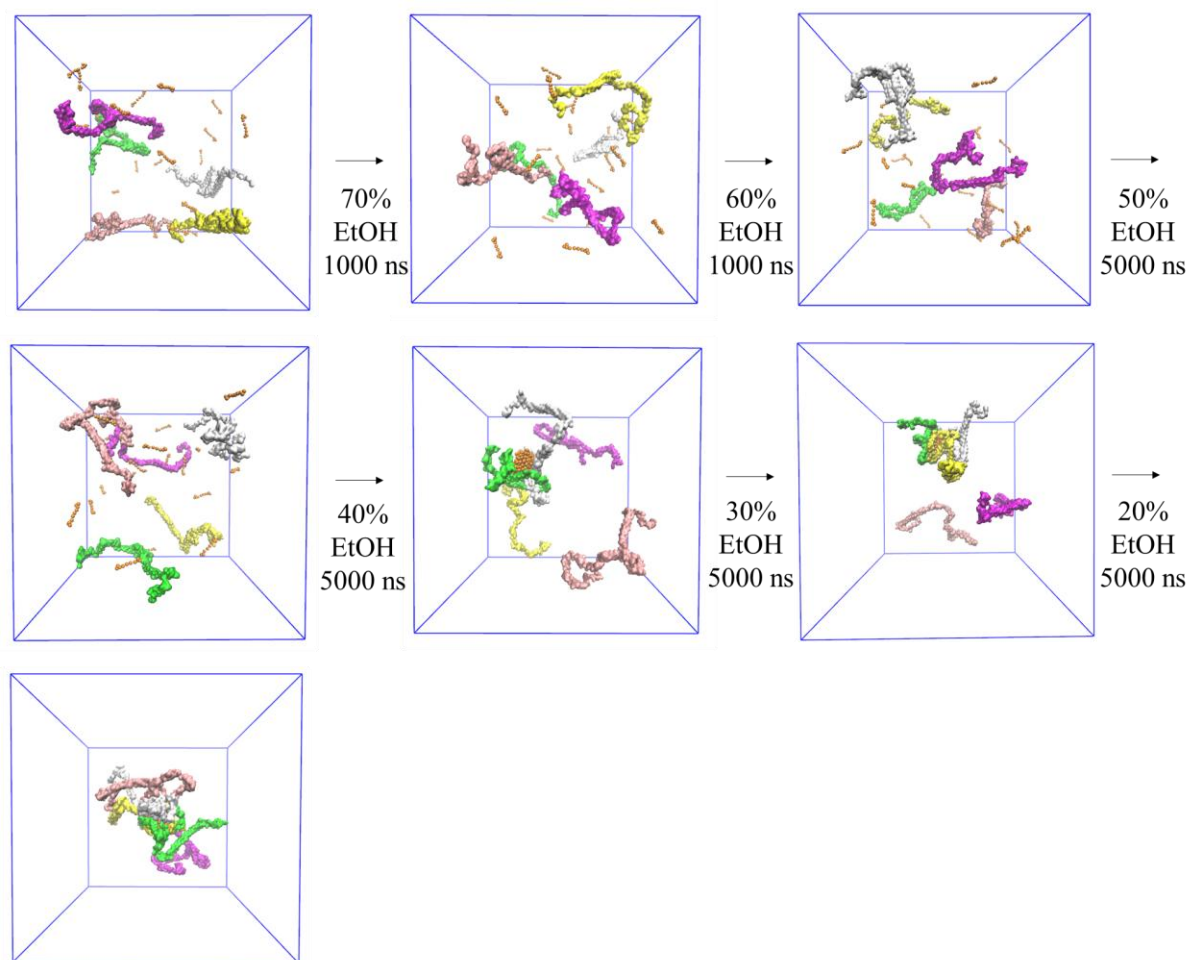
When considered together with the lower-scale observations in Section 4.3.1, the suspension-level results showed that the monomer/dilute-system transition near 40% ethanol did not directly translate into solvent-extraction resistance. Instead, the pronounced increases in turbidity and particle size between 40% and 30% ethanol suggested that reduced hexane extractability was more closely associated with the development of extensive zein aggregation.

4.3.2.3. Coarse-grained molecular dynamics simulation

The suspension experiments showed that solvent-extraction resistance emerged after zein-containing systems developed into extensive aggregates at lower ethanol fractions. To examine whether such collective assembly could reduce lutein accessibility to the surrounding solvent, CG-MD simulations were performed using multi-protein α -zein–lutein systems. Unlike the AA-MD simulations, which were used to evaluate conformational and contact changes of one α -zein molecule, the CG-MD simulations were used as a simplified aggregate-level model to examine ethanol-dependent zein assembly and lutein solvent exposure.

As shown in Fig. 4.5a, the CG-MD simulations showed limited α -zein–lutein association under ethanol-rich conditions (70–50% ethanol). At 40% ethanol, collective assembly began to emerge, with zein molecules associating around clustered lutein molecules. As ethanol concentration decreased further, the simulated systems formed more compact zein-rich aggregates, in which lutein molecules became more centrally located within the aggregate phase.

(a)



(b)

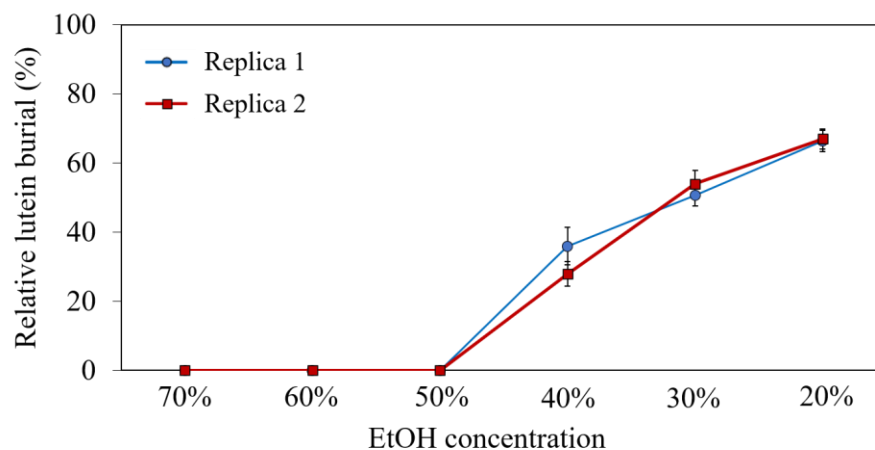


Figure 4.5. Structural evolution of zein–lutein assemblies across ethanol concentrations from coarse-grained molecular dynamics simulations (a). Each system was initialized from the final configuration of the preceding ethanol condition to reflect the gradual solvent transition encountered during antisolvent processing. Individual zein monomers are shown in distinct

colors. Relative burial of lutein averaged over the final 500 ns of simulation, defined as $100 \times [1 - (\text{exposed solvent-accessible surface area (SASA) of lutein} / \text{SASA of lutein in the absence of zein})]$ (b). Values were not calculated for ethanol concentrations $\geq 50\%$ due to the absence of stable complexes.

Consistent with this structural reorganization, the relative burial of lutein progressively increased with decreasing ethanol fraction (Fig. 4.5b). Although the CG-MD results should not be interpreted as quantitative validation of the extraction data, the ethanol-dependent increase in lutein burial was consistent with the experimental observation that hexane resistance emerged only under more strongly aggregated suspension conditions. These results provide model-based support for the interpretation that extensive zein-rich aggregation can reduce lutein accessibility to the extracting solvent, thereby contributing to solvent-extraction resistance in suspension.

Because CG simulations employ a reduced representation in which groups of atoms are mapped into beads and interactions are described by effective potentials, a comparison with AA simulations was necessary. As shown in Fig. 4.1a, AA-MD simulations at 40% ethanol exhibited collapsed monomer morphology, whereas proteins in CG-MD remained relatively more expanded under comparable conditions. The difference in protein conformation can be attributed to the different molecular resolutions and force-field parameterizations used in the AA and CG simulations. The atomistic CHARMM36 protein force field has been extensively parameterized to accurately describe structural details of protein conformation, such as backbone dihedral angles and side-chain interactions (Huang & MacKerell, 2013). In contrast, as a coarse-grained model, Martini 3 targets the reproduction of mesoscale phenomena, such as partition behavior or self-assembly (Souza et al., 2021). More importantly, CG models map groups of atoms into coarse-grained beads and describe interactions using effective bead-level potentials. As a result, fine interaction details such as directional hydrogen bonding and residue-specific interactions are

averaged out in CG simulation. Moreover, the absence of elastic network to reproduce the expanded monomeric structure can also contribute to the conformational differences observed between the AA-MD and CG-MD simulations.

Given the scope of the CG force field, the focus of interpretation in CG-MD should not be the exact replication of the protein conformation in AA-MD, but the aggregate-level organization of zein and lutein, where the CG models can still provide useful model-based information. At the qualitative level, both AA-MD and CG-MD simulations predicted increased zein–lutein association near 40% ethanol, with the emergence of lutein-rich domain. Moreover, the degree of zein-lutein aggregation became more pronounced as the ethanol fraction further decreased. In this sense, the conformational difference between AA-MD and CG-MD was considered a resolution-dependent limitation of the simulation framework, rather than evidence against the aggregate-level interpretation of the CG simulations.

4.3.3. Powder-level thermal stability of zein-lutein system

The suspension-level results demonstrated that lutein retention emerged at lower ethanol fractions. Such functionality observed below 30% ethanol may translate into enhanced functional stability in more product-relevant solid forms. To evaluate this, zein–lutein powders prepared under different processing histories were tested for thermal stability of lutein.

4.3.3.1. Thermal stability of zein–lutein powders prepared under different ethanol conditions

To assess how ethanol-dependent preparation history influence lutein retention and distribution in the resulting powders, three distinct process routes were compared: a physical mixture; fast quenching, defined as rapid addition of water to directly reduce ethanol from 70%

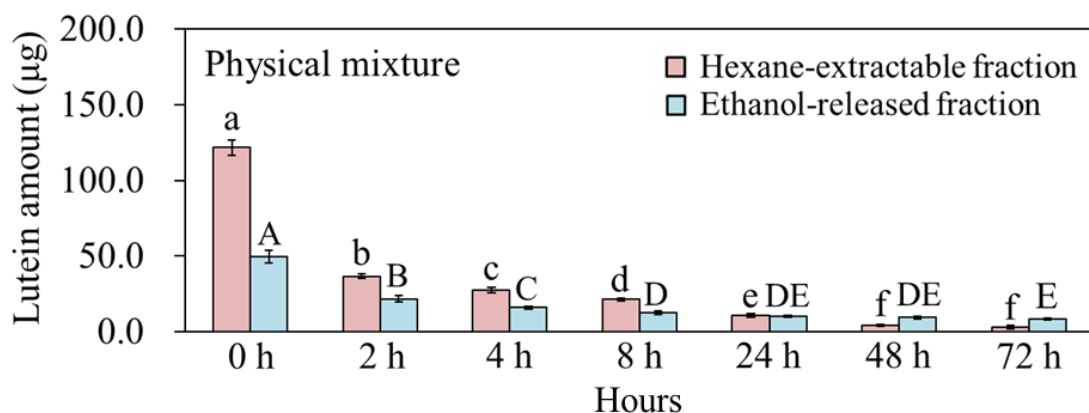
to 10%; and gradual quenching, defined as stepwise reduction from 70% to intermediate ethanol fractions (50%, 40%, 30%, and 20%) with equilibration at each step (10 min) prior to final quenching to 10% ethanol. The thermal stability of lutein fractions across all ethanol conditions is summarized in Table 4.2, while selected representative conditions—physical mixture, fast quenching from 70% ethanol (70% quench), and gradual quenching with equilibration down to 20% ethanol prior to final quenching (20% quench)—are highlighted in Fig. 4.6 for clarity and comparison.

When the distribution of lutein fractions was compared across preparation routes, only the fast-quenched powder exhibited a markedly higher proportion of ethanol-released fraction relative to the hexane-extractable fraction (Table 4.2 and Fig. 4.6). Specifically, the ratio between the two fractions in the fast-quenched powder was 4.55, which was substantially greater than those observed in other samples (0.41–0.83), indicating a distinct distribution of lutein within the protein matrix. Upon thermal exposure at 60°C, all samples exhibited a time-dependent reduction in recoverable lutein fractions. The physical mixture showed a pronounced loss of both fractions within the first 2 hr of heating, demonstrating minimal protective capacity. In contrast, encapsulated powders displayed a more gradual decrease between 0 and 8 hr. Between 8 and 24 hr, a significant reduction in the hexane-extractable fraction was observed across encapsulated samples, whereas the ethanol-released fraction remained comparatively sustained during the same interval. This fraction-dependent depletion pattern suggested preferential loss of surface-exposed lutein under prolonged thermal stress.

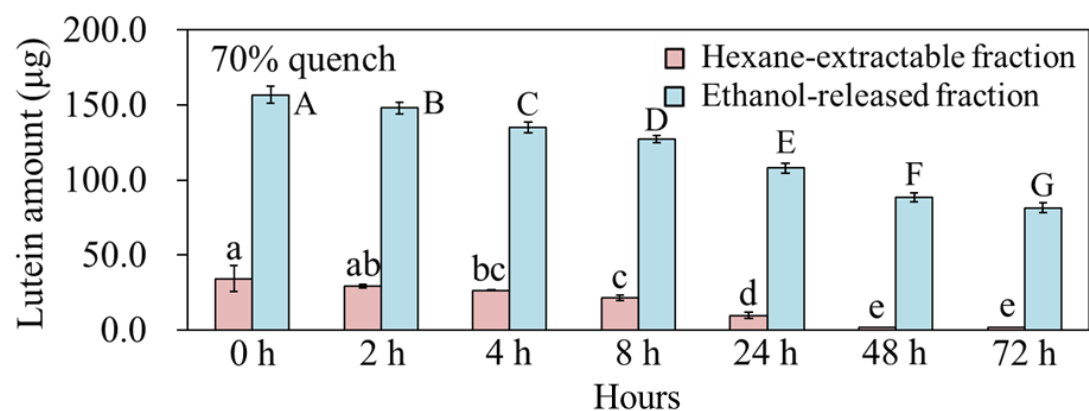
Table 4.2. Thermal stability and final lutein retention of zein–lutein powders prepared under different ethanol-dependent processing routes. Thermal stability was evaluated at 60 °C for 72 h. PM: physical mixture; FQ: fast quenching from 70% to 10% ethanol by rapid water addition; GQ: gradual quenching with equilibration at the indicated ethanol concentrations prior to final quenching to 10% ethanol (e.g., GQ 50% indicates equilibration at 50% ethanol for 10 min before final quenching). Values are presented as mean $\pm$ standard deviation. The upper and lower values represent hexane-extractable and ethanol-released lutein fractions, respectively. Different lowercase (hexane-extractable fraction) and uppercase (ethanol-released fraction) letters indicate significant differences over time within each sample ($p < 0.05$). All values are expressed as μg of lutein per 15 mg of zein–lutein powder. The final lutein retention column indicates differences among processing routes ($p < 0.05$). Final retention (%) was calculated as the amount of lutein at 72 h relative to 0 h.

	0 hr	2 hr	4 hr	8 hr	24 hr	48 hr	72 hr	Final lutein retention (%)
PM	121.86 $\pm$ 5.00a 49.42 $\pm$ 4.23A	36.79 $\pm$ 1.14b 21.86 $\pm$ 2.07B	27.34 $\pm$ 1.75c 16.02 $\pm$ 0.78C	21.27 $\pm$ 1.01d 12.68 $\pm$ 1.31D	10.80 $\pm$ 1.17e 10.07 $\pm$ 0.57D E	4.39 $\pm$ 0.58f 9.32 $\pm$ 0.78DE	3.04 $\pm$ 0.51f 8.58 $\pm$ 0.22E	6.77 $\pm$ 0.53e
FQ (70%)	34.43 $\pm$ 8.65a 156.81 $\pm$ 5.60A	29.37 $\pm$ 1.05ab 148.24 $\pm$ 3.91B	26.67 $\pm$ 0.58bc 135.33 $\pm$ 3.67C	21.60 $\pm$ 1.55c 127.39 $\pm$ 2.64D	9.80 $\pm$ 2.41d 108.14 $\pm$ 3.38E	2.03 $\pm$ 0.00e 88.41 $\pm$ 2.82F	2.03 $\pm$ 0.00e 81.58 $\pm$ 2.84G	43.84 $\pm$ 3.15a
GQ (50%)	134.00 $\pm$ 8.85a 87.16 $\pm$ 6.07A	116.12 $\pm$ 4.57b 83.44 $\pm$ 5.58A B	102.62 $\pm$ 7.17c 77.73 $\pm$ 1.49B C	96.88 $\pm$ 8.43c 72.76 $\pm$ 3.17C	49.96 $\pm$ 6.51d 56.37 $\pm$ 2.26D	13.16 $\pm$ 2.76e 50.91 $\pm$ 3.18D	10.13 $\pm$ 6.16e 40.48 $\pm$ 4.03E	22.91 $\pm$ 4.72b
GQ (40%)	172.15 $\pm$ 13.92 a 75.25 $\pm$ 4.80A	151.22 $\pm$ 13.92 b 73.01 $\pm$ 3.98A	139.75 $\pm$ 10.52 bc 71.89 $\pm$ 0.78A	128.95 $\pm$ 20.51 c 73.76 $\pm$ 10.44A	87.09 $\pm$ 5.06d 62.95 $\pm$ 2.28B	35.78 $\pm$ 4.22e 48.68 $\pm$ 3.35C	26.33 $\pm$ 4.05e 39.62 $\pm$ 4.60D	26.83 $\pm$ 4.33bc
GQ (30%)	111.73 $\pm$ 5.10a 92.38 $\pm$ 14.22A	99.92 $\pm$ 7.18b 93.00 $\pm$ 17.55A	90.46 $\pm$ 2.29bc 87.04 $\pm$ 5.04A	81.35 $\pm$ 12.66c 75.62 $\pm$ 6.20A	19.58 $\pm$ 1.55d 57.74 $\pm$ 4.75B	4.73 $\pm$ 0.59e 42.10 $\pm$ 1.31B C	2.36 $\pm$ 0.19e 35.52 $\pm$ 2.28C	18.71 $\pm$ 2.66cd
GQ (20%)	137.72 $\pm$ 7.91a 79.22 $\pm$ 3.55A	123.88 $\pm$ 2.26b 75.00 $\pm$ 2.81A	112.41 $\pm$ 6.62c 74.62 $\pm$ 5.22A	93.84 $\pm$ 7.14d 71.27 $\pm$ 0.57A	5.06 $\pm$ 0.41e 46.32 $\pm$ 9.29B	2.37 $\pm$ 0.59e 38.50 $\pm$ 2.26B C	2.03 $\pm$ 0.00e 32.29 $\pm$ 7.34C	15.82 $\pm$ 0.37d

(a)



(b)



(c)

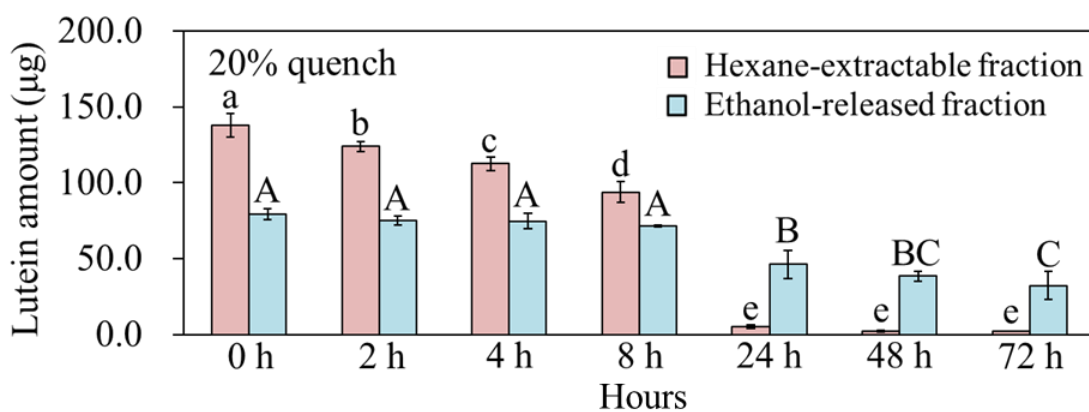


Figure 4.6. Thermal stability of zein–lutein powders under different processing histories: physical mixture (a), fast antisolvent addition from 70% ethanol (b), and stepwise ethanol reduction to 20% with equilibration followed by rapid dilution (c). All values are expressed as µg of lutein per 15 mg of zein–lutein powder.

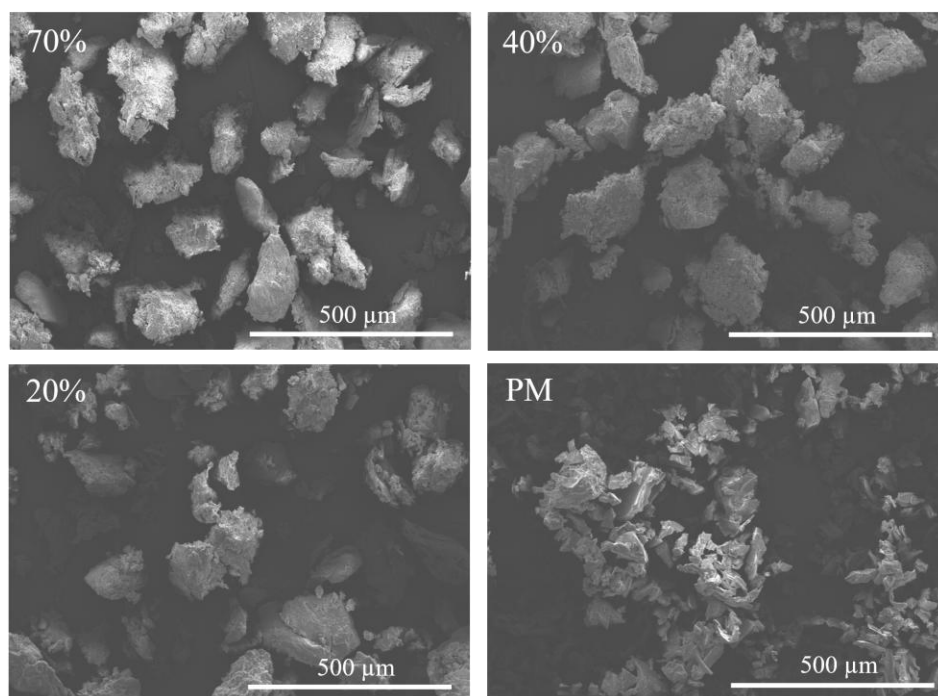
Although all encapsulated powders exhibited improved final retention of lutein (15.82-43.84%) compared to that of physical mixture (6.77%), substantial differences were observed across processing histories. Powders prepared via rapid quenching from 70% ethanol demonstrated the highest final retention after 72 hr (43.84%), exceeding those prepared under lower ethanol pre-assembly conditions. Among the gradual-quench powders, samples equilibrated at 50%, 40%, 30%, and 20% ethanol exhibited final retention values of 22.91%, 26.83%, 18.71%, and 15.82%, respectively. Because all encapsulated powders were ultimately quenched to the same final ethanol fraction of 10%, these differences indicate that lutein protection in the dried powders was not determined by the final solvent composition alone, but was strongly influenced by the processing pathway and quenching history. The noticeable differences across samples, particularly in the amount of extractable lutein and final lutein retention, may be attributed to processing-history effects before final quenching, including ethanol-dependent suspension states and changes in the solute-to-solvent ratio caused by intermediate water addition. However, thermal stability was compared based on retention relative to each sample's own 0 h value, so the analysis primarily reflected lutein loss during heating rather than initial recovery differences.

Interestingly, this ranking of thermal retention did not reflect the suspension-level hexane resistance described in Section 4.3.2. In the suspension state, lower ethanol conditions (30–20%) promoted extensive aggregation and enhanced resistance to hexane extraction. Instead, the highest retention (43.84%) was observed in powders prepared via rapid quenching from 70% ethanol, where suspension-level extraction resistance was negligible (Fig. 4.3a). This discrepancy indicated that the structural features governing solvent resistance in suspension were not identical to those determining thermal stability in the dried matrix.

4.3.3.2. Powder morphology observed by SEM

To examine whether the variance in thermal stabilities were associated with apparent morphological differences, representative zein–lutein powders were observed by SEM (Fig. 4.7). At low magnification, powders prepared through the different quenching routes all consisted of irregular micron-scale aggregates (Fig. 4.7a). Some particles in the 20% quench sample appeared to contain relatively smoother and more compact regions. However, this feature was not consistently reflected at higher magnification, where rough and heterogeneous surface textures were observed across the encapsulated powders (Fig. 4.7b). The physical mixture also consisted of irregular particles, but it appeared more fragmented and angular, with plate-like features and smoother fractured surfaces in some regions. This morphology was distinct from the rough, heterogeneous aggregate surfaces observed in the encapsulated powders. Overall, the SEM observations did not reveal a clear ethanol-dependent morphological trend among the encapsulated powders, suggesting that the differences in thermal stability were not primarily attributable to gross powder morphology.

(a)



(b)

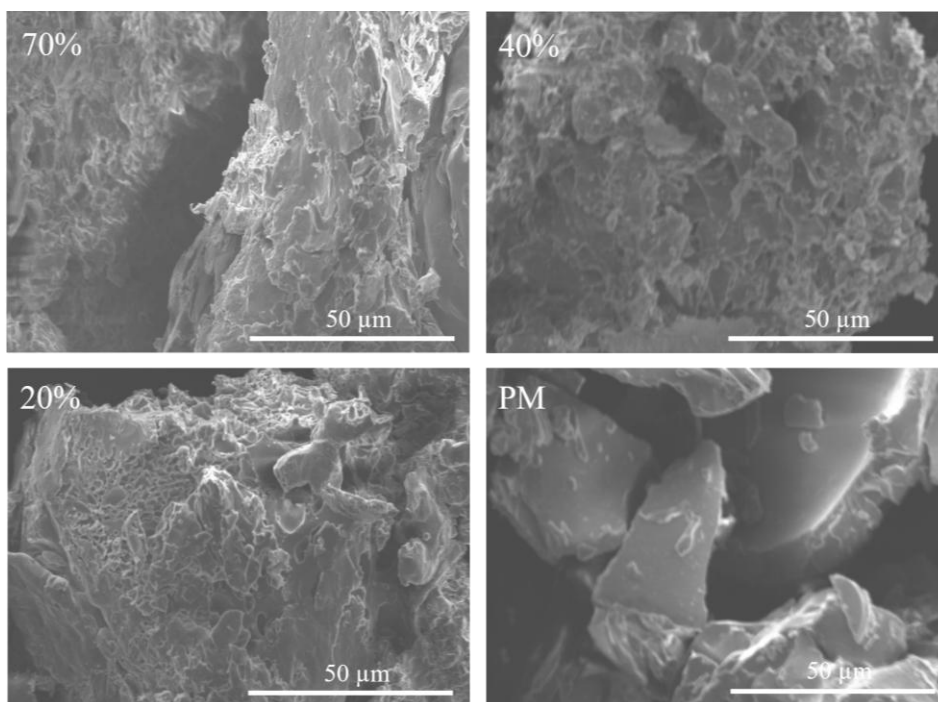


Figure 4.7. SEM morphology of zein–lutein powders prepared under different EtOH processing histories and the physical mixture (PM). Images are shown at 100 $\times$ magnification (a) and 1000 $\times$ magnification (b).

4.3.3.3. Possible explanation for differences in powder thermal stability

Because no clear ethanol-dependent morphological trend was observed among the encapsulated powders, a possible model-based explanation was explored to help interpret the different thermal stability profiles. For this purpose, representative α -zein–lutein configurations obtained from the AA-MD and CG-MD simulations in Sections 4.3.1 and 4.3.2 were subjected to water quenching. Specifically, structures corresponding to two processing histories were selected: (i) a system equilibrated at 70% ethanol, representing the fast-quenching pathway, and (ii) a system obtained after stepwise ethanol reduction (70%→20%), approximating the gradual-quenching pathway.

Following water quenching, the two systems exhibited distinct spatial distribution of lutein (Fig. 4.8). In the fast-quenching model, lutein distribution was more dispersed throughout the protein network. On the other hand, in the gradual-quenching model, lutein molecules remained concentrated within localized domains associated with aggregated protein assemblies. Specifically, the number of lutein–lutein contacts over the last 50 ns of the AA quenching simulations was 16.41 ± 0.79 for fast quenching and 83.75 ± 2.37 for gradual quenching. Similarly, in CG-MD simulations, the corresponding values for the last 500 ns were 89.37 ± 1.55 and 312.20 ± 1.20 for fast and gradual quenching, respectively. This observation was consistent with the notion that processing kinetics influence the distribution of encapsulated molecules by shifting the window for structural reorganization (Chandrasekaran et al., 2024; Johnson & Prud'homme, 2003). During solvent evaporation and mechanical processing, such heterogeneous clusters of lutein may become exposed to powder surfaces or fracture interfaces. This increased surface exposure could thereby render lutein more susceptible to thermal degradation. Furthermore, the local concentration of lutein may further increase its vulnerability toward

degradation. On the other hand, a more homogeneous distribution of lutein established in fast quenching may reduce the extent of the exposure of lutein-rich domain. This interpretation was consistent with the low proportion of hexane-extractable fraction and superior thermal retention observed in fast-quenched powders (Fig. 4.6 and Table 4.2). However, this model- and process-based interpretation does not directly demonstrate lutein localization in the dried powders and should therefore be considered a possible explanation for the observed thermal stability differences.

Taking the results across Sections 4.3.1–4.3.3 together, ethanol-dependent structural transitions in zein–lutein systems did not translate uniformly across hierarchical scales. The lower-scale transitions near 40% ethanol, observed as predicted monomer-level reorganization and the loss of solution-like behavior in dilute solutions, were not sufficient to produce solvent-extraction resistance in suspension. Instead, extraction resistance emerged only at lower ethanol fractions below 30%, where extensive zein-rich aggregation was observed. However, even this suspension-level resistance did not consistently translate into improved thermal stability in powders, where the most stable retention was obtained through rapid quenching from 70% ethanol. Thus, the protective performance of zein–lutein systems from antisolvent processing was better described as a multiscale and path-dependent outcome, rather than as the direct consequence of a single ethanol-dependent structural transition.

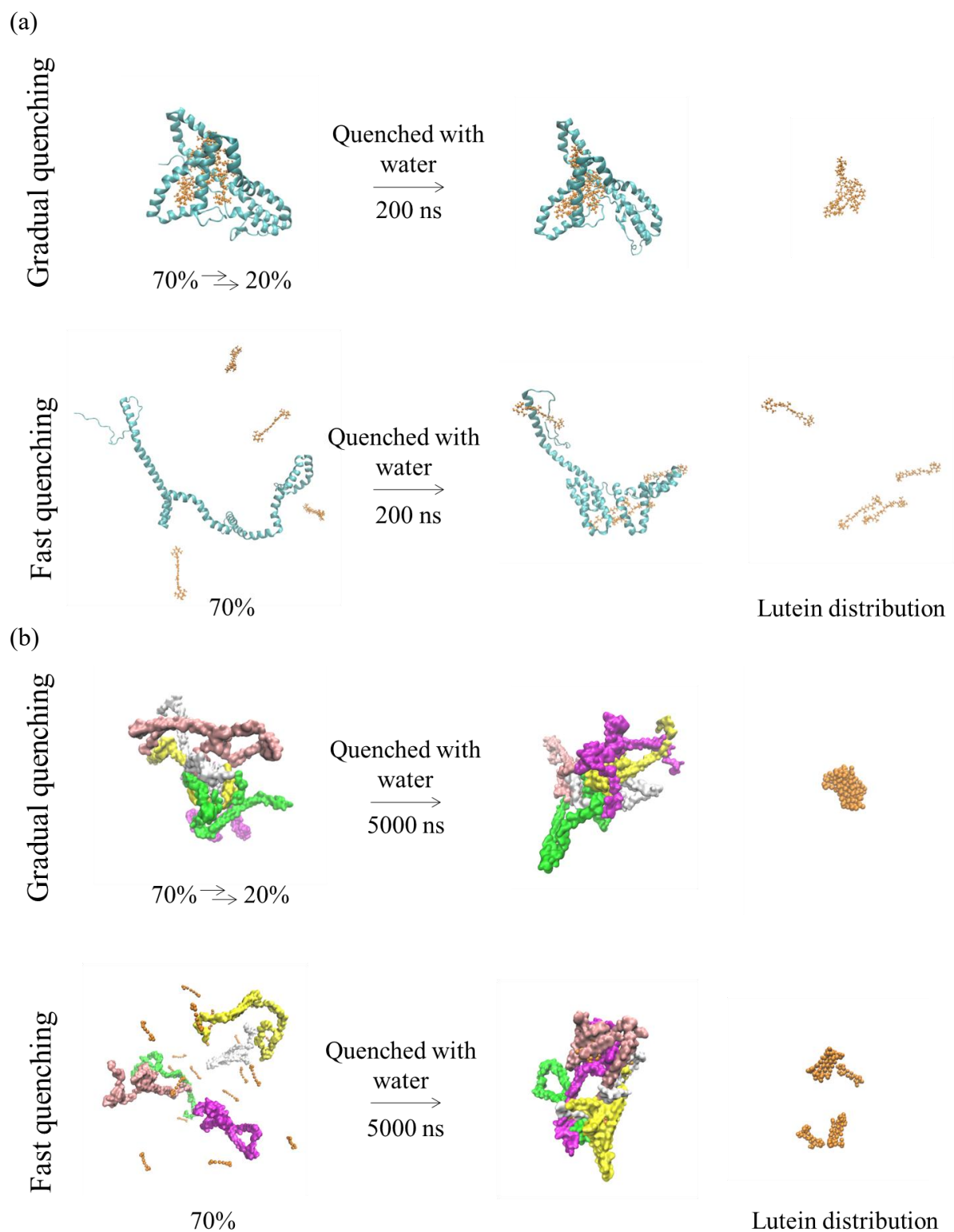


Figure 4.8. Configurations of zein–lutein complex and spatial distribution of lutein after water quenching under different processing conditions, predicted from all-atom (a) and coarse-grained molecular dynamics simulations (b).

4.4. Conclusions

This study examined ethanol-dependent zein–lutein assembly across monomer-level simulations, dilute solutions, suspensions, and dried powders to determine whether structural transitions observed at lower scales translate into functional lutein protection at higher scales. Our results demonstrated that the ethanol range associated with the monomer/dilute-solution transition, the condition that produced solvent-resistant suspensions, and the corresponding pathway that generated the most thermally stable powder were distinct.

Specifically, the lower-scale transition near 40% ethanol, observed as predicted monomer-level α -zein collapse and loss of additive solution-like behavior in dilute solution, did not directly translate into suspension-level solvent-extraction resistance. Instead, extraction resistance emerged only at lower ethanol fractions, where extensive zein-rich aggregation was observed. However, this suspension-level resistance did not lead to superior thermal stability in dried powders, for which rapid quenching from 70% ethanol produced the highest lutein retention. Thus, the lower-scale transitions near 40% ethanol were not sufficient predictors of lutein protection in either suspensions or powders. Moreover, solvent-extraction resistance in suspension did not directly translate into superior thermal stability in the dried powder state.

Several limitations should be considered when interpreting these findings. First, the AA-MD simulations were conducted using an isolated 22 kDa α -zein monomer and therefore do not directly represent whole-zein assembly. Although a consistent compaction trend between 50% and 40% ethanol was observed in two independent replicas, additional replicas and longer simulations would be useful for further characterization. In addition, the suspension experiments did not resolve the specific mass-transfer mechanism underlying solvent-extraction resistance or directly visualize lutein localization. Similarly, the proposed explanation for the superior stability

of fast-quenched powders remains inferential, because sequential extraction and quenching simulations provide indirect support rather than direct evidence of lutein distribution in dried powders. Direct visualization using Raman mapping or hyperspectral imaging would be valuable for confirming lutein localization within the dried matrix. Finally, powder stability was evaluated under accelerated thermal storage at 60 °C for 72 h, and whether the same ranking is maintained under other storage conditions remains to be examined.

Despite these limitations, the present study highlights that structural changes observed at earlier structural scales do not necessarily indicate functional protection at later stages. From a formulation-design perspective, these findings suggest that ethanol-dependent changes observed at one structural scale may not directly predict lutein protection at another scale. This point may be particularly important for monomer-level computational analyses of zein antisolvent assembly, where lutein protection appeared to depend strongly on multi-protein assembly. Beyond its implications for zein-based encapsulation, this multiscale, path-dependent framework may also inform data-driven modeling approaches. In particular, functional properties that emerge from collective assembly and processing history may not be adequately captured by conventional static descriptor-based models, suggesting the need for modeling strategies that incorporate structural hierarchy and process-dependent states.

Acknowledgement

This is contribution No. 26-177-J from the Kansas Agricultural Experimental Station. This study was supported in part by the USDA National Institute of Food and Agriculture project (2026-68017-45854).

References

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., ... Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
- Argos, P., Pedersen, K., Marks, M. D., & Larkins, B. A. (1982). A structural model for maize zein proteins. *Journal of Biological Chemistry*, 257(17), 9984–9990. [https://doi.org/10.1016/S0021-9258\(18\)33974-7](https://doi.org/10.1016/S0021-9258(18)33974-7)
- Basalla, A. J., & Kendrick, B. S. (2023). Correcting ultraviolet-visible spectra for baseline artifacts. *Journal of Pharmaceutical Sciences*, 112(12), 3240–3247. <https://doi.org/10.1016/j.xphs.2023.08.015>
- Bouman, J., Belton, P., Venema, P., van der Linden, E., de Vries, R., & Qi, S. (2016). Controlled release from zein matrices: Interplay of drug hydrophobicity and pH. *Pharmaceutical Research*, 33(3), 673–685. <https://doi.org/10.1007/s11095-015-1818-8>
- Chandrasekaran, A., Graham, K., Stachowiak, J. C., & Rangamani, P. (2024). Kinetic trapping organizes actin filaments within liquid-like protein droplets. *Nature Communications*, 15(1), 3139. <https://doi.org/10.1038/s41467-024-46726-6>
- Chang, C., Wang, T., Hu, Q., & Luo, Y. (2017). Caseinate–zein–polysaccharide complex nanoparticles as potential oral delivery vehicles for curcumin: Effect of polysaccharide type and chemical cross-linking. *Food Hydrocolloids*, 72, 254–262. <https://doi.org/10.1016/j.foodhyd.2017.05.039>
- Chen, S., Jia, J., Du, J., Qin, N., & Xia, X. (2026). Fabrication of a zein-phloretin complex and nanoparticles: Interaction mechanism and characterization. *RSC Advances*, 16, 7717–7725. <https://doi.org/10.1039/D5RA08898D>
- Chiariello, M. G., Zarmiento-Garcia, R., & Marrink, S. J. (2024). Martini 3 coarse-grained model for the cofactors involved in photosynthesis. *International Journal of Molecular Sciences*, 25(14), 7947. <https://doi.org/10.3390/ijms25147947>

- Christensen, N. J. (2024). Conformations of a highly expressed Z19 α -zein studied with AlphaFold2 and MD simulations. *PLOS ONE*, 19(5), e0293786. <https://doi.org/10.1371/journal.pone.0293786>
- Dai, Z., Song, J., Chen, Y., Feng, L., Xu, Y., Li, D., Wu, C., Zhang, Z., & Liu, J. (2021). Study on the bioavailability of stevioside-encapsulized lutein and its mechanism. *Food Chemistry*, 354, 129528. <https://doi.org/10.1016/j.foodchem.2021.129528>
- 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(12), 10089–10092. <https://doi.org/10.1063/1.464397>
- De Boer, F. Y., Kok, R. N. U., Imhof, A., & Velikov, K. P. (2018). White zein colloidal particles: Synthesis and characterization of their optical properties on the single particle level and in concentrated suspensions. *Soft Matter*, 14(15), 2870–2878. <https://doi.org/10.1039/c7sm02415k>
- Devnani, B., Tan, N. K., Sanders, J. M., Shelley, J. C., Moraca, F., & Grossmann, L. (2024). Molecular insights into the structure forming properties of zein and a rheological comparison with hordein. *Future Foods*, 10, Article 100503. <https://doi.org/10.1016/j.fufo.2024.100503>
- Ding, Z., Tao, T., Wang, X., Prakash, S., Zhao, Y., Han, J., & Wang, Z. (2020). Influences of different carbohydrates as wall material on powder characteristics, encapsulation efficiency, stability and degradation kinetics of microencapsulated lutein by spray drying. *International Journal of Food Science and Technology*, 55(7), 2872–2882. <https://doi.org/10.1111/ijfs.14544>
- Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., Xu, H., & Shaw, D. E. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32), 13118–13123. <https://doi.org/10.1073/pnas.1104614108>
- Ellman, G. L. (1959). Tissue sulphydryl groups. *Archives of Biochemistry and Biophysics*, 82(1), 70–77. [https://doi.org/10.1016/0003-9861\(59\)90090-6](https://doi.org/10.1016/0003-9861(59)90090-6)

- 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.80.3.62>
- Fang, S., Wang, Y., Zhu, L., Zhang, Y., & Yu, L. L. (2025). Effect of zein nanoparticles addition on anthocyanin and lutein dual-loaded nanocomposite hydrogels: Structure, physico-chemical and delivery properties. *International Journal of Biological Macromolecules*, 309, 142967. <https://doi.org/10.1016/j.ijbiomac.2025.142967>
- Feng, Y., Ma, X., Kong, B., Chen, Q., & Liu, Q. (2021). Ethanol induced changes in structural, morphological, and functional properties of whey proteins isolates: Influence of ethanol concentration. *Food Hydrocolloids*, 111, 106379.
<https://doi.org/10.1016/j.foodhyd.2020.106379>
- Fu, D., & Weller, C. L. (1999). Rheology of zein solutions in aqueous ethanol. *Journal of Agricultural and Food Chemistry*, 47(5), 2103–2108. <https://doi.org/10.1021/jf9811121>
- GROMACS development team. (2022). GROMACS 2022 source code (Version 2022) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.6103835>
- Han, H., Chang, Y., & Jiao, Y. (2024). Recent advances in efficient lutein-loaded zein-based solid nano-delivery systems: Establishment, structural characterization, and functional properties. *Foods*, 13(14), 2304. <https://doi.org/10.3390/foods13142304>
- Hernandez, E., Baker, R. A., & Crandall, P. G. (1991). Model for evaluating turbidity in cloudy beverages. *Journal of Food Science*, 56(3), 747–750. <https://doi.org/10.1111/j.1365-2621.1991.tb05373.x>
- Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. (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](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H)
- Hu, D., Lin, C., Liu, L., Li, S., & Zhao, Y. (2012). Preparation, characterization, and in vitro release investigation of lutein/zein nanoparticles via solution enhanced dispersion by supercritical fluids. *Journal of Food Engineering*, 109(3), 545–552.
<https://doi.org/10.1016/j.jfoodeng.2011.10.015>

- 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>
- Huang, M., Song, Y., Lv, N., Liu, C., Ren, G., Shen, Q., Wang, B., Cao, Z., & Xie, H. (2022). Theoretical and experimental perspectives of interaction mechanism between zein and lysozyme. *Food Hydrocolloids*, 132, 107876. <https://doi.org/10.1016/j.foodhyd.2022.107876>
- 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](https://doi.org/10.1016/0263-7855(96)00018-5)
- Johnson, B. K., & Prud'homme, R. K. (2003). Flash nanoprecipitation of organic actives and block copolymers using a confined impinging jets mixer. *Australian Journal of Chemistry*, 56(10), 1021–1024. <https://doi.org/10.1071/CH03115>
- Joye, I. J., Davidov-Pardo, G., Ludescher, R. D., & McClements, D. J. (2015). Fluorescence quenching study of resveratrol binding to zein and gliadin: Towards a more rational approach to resveratrol encapsulation using water-insoluble proteins. *Food Chemistry*, 185, 261–267. <https://doi.org/10.1016/j.foodchem.2015.03.128>
- Kroon, P. C., Grünewald, F., Barnoud, J., van Tilburg, M., Brasnett, C., Souza, P. C. T., Wassenaar, T. A., & Marrink, S. J. (2025). Martinize2 and Vermouth provide a unified framework for molecular topology generation. *eLife*, 12, RP90627. <https://doi.org/10.7554/eLife.90627>
- Kumar, N., & Li, Y. (2025). Solvent-dependent structural and dynamic behavior of α -gliadin in aqueous ethanol systems: Insights from molecular dynamics simulations. *Food Chemistry*, 147044. <https://doi.org/10.1016/j.foodchem.2025.147044>
- Kwon, H., Xu, Y., Xu, X., & Li, Y. (2025). Unraveling the physicochemical differences among Osborne protein classes via bioinformatics and AI. *Food Research International*, 117322. <https://doi.org/10.1016/j.foodres.2025.117322>
- Li, F., Chen, Y., Liu, S., Qi, J., Wang, W., Wang, C., Zhong, R., Chen, Z., Li, X., Guan, Y., Kong, W., & Zhang, Y. (2017). Size-controlled fabrication of zein nano/microparticles

- by modified anti-solvent precipitation with/without sodium caseinate. *International Journal of Nanomedicine*, 12, 8197–8209. <https://doi.org/10.2147/IJN.S143733>
- Li, H., Yuan, Y., Zhu, J., Wang, T., Wang, D., & Xu, Y. (2020). Zein/soluble soybean polysaccharide composite nanoparticles for encapsulation and oral delivery of lutein. *Food Hydrocolloids*, 103, 105715. <https://doi.org/10.1016/j.foodhyd.2020.105715>
- Li, Y., Li, J., Xia, Q., Zhang, B., Wang, Q., & Huang, Q. (2012). Understanding the dissolution of α -zein in aqueous ethanol and acetic acid solutions. *The Journal of Physical Chemistry B*, 116(39), 12057–12064. <https://doi.org/10.1021/jp305709y>
- Li, Z., Liu, Y., Lu, S., Gao, S., Chen, L., Deng, Q., Xu, Z., & Xiang, X. (2025). pH-driven plant protein nanoparticles modulated by flaxseed gum oligosaccharides for apigenin delivery. *Food Chemistry*, 146835. <https://doi.org/10.1016/j.foodchem.2025.146835>
- Li, Z., Xiao, Z., Jiang, M., & Zhang, Y. (2023). A comparison on the binding mechanisms of zein and gliadin with curcumin to guide the self-assembly of nanoparticles for delivery purpose. *Journal of Agriculture and Food Research*, 13, 100660. <https://doi.org/10.1016/j.jafr.2023.100660>
- Liao, W., Zhang, L., Xie, J., Wang, Y., Tong, Z., Wang, Y., Abliz, A., Li, Q., Shu, X., & Gao, Y. (2022). Impact of ethanol shock on the structural change and emulsifying capacity of bovine lactoferrin. *Food Hydrocolloids*, 133, 107894. <https://doi.org/10.1016/j.foodhyd.2022.107894>
- Liu, C., Yuan, Y., Ma, M., Zhang, S., Wang, S., Li, H., Chen, L., Zhao, Y., & Wang, D. (2020). Self-assembled composite nanoparticles based on zein as delivery vehicles of curcumin: Role of chondroitin sulfate. *Food & Function*, 11(6), 5377–5388. <https://doi.org/10.1039/D0FO00964D>
- Ma, M., Yuan, Y., Yang, S., Wang, Y., & Lv, Z. (2020). Fabrication and characterization of zein/tea saponin composite nanoparticles as delivery vehicles of lutein. *LWT*, 125, 109270. <https://doi.org/10.1016/j.lwt.2020.109270>
- Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10), 2319–2327. <https://doi.org/10.1002/jcc.21787>

- Mitri, K., Shegokar, R., Gohla, S., Anselmi, C., & Müller, R. H. (2011). Lipid nanocarriers for dermal delivery of lutein: Preparation, characterization, stability and performance. *International Journal of Pharmaceutics*, 414(1–2), 267–275. <https://doi.org/10.1016/j.ijpharm.2011.05.008>
- Mobley, D. L., & Dill, K. A. (2009). Binding of small-molecule ligands to proteins: “What you see” is not always “what you get.” *Structure*, 17(4), 489–498. <https://doi.org/10.1016/j.str.2009.02.010>
- Momany, F. A., Sessa, D. J., Lawton, J. W., Selling, G. W., Hamaker, S. A., & Willett, J. L. (2006). Structural characterization of α -zein. *Journal of Agricultural and Food Chemistry*, 54(2), 543–547. <https://doi.org/10.1021/jf052417e>
- National Research Council. (1926–1930). International critical tables of numerical data, physics, chemistry and technology (E. W. Washburn, Ed.-in-Chief; Vols. 1–7). McGraw-Hill.
- Nowakowska, J. (1939). The refractive indices of ethyl alcohol and water mixtures [Master’s thesis, Loyola University Chicago]. Loyola eCommons.
- O’Hare, K. D., Spedding, P. L., & Grimshaw, J. (1993). Evaporation of the ethanol and water components comprising a binary liquid mixture. *Developments in Chemical Engineering and Mineral Processing*, 1(2–3), 118–128. <https://doi.org/10.1002/apj.5500010203>
- Pan, K., & Zhong, Q. (2016). Low energy, organic solvent-free co-assembly of zein and caseinate to prepare stable dispersions. *Food Hydrocolloids*, 52, 600–606. <https://doi.org/10.1016/j.foodhyd.2015.08.014>
- Pérez, J. J., Tomás, M. S., & Rubio-Martínez, J. (2016). Assessment of the sampling performance of multiple-copy dynamics versus a unique trajectory. *Journal of Chemical Information and Modeling*, 56(10), 1950–1962. <https://doi.org/10.1021/acs.jcim.6b00347>
- Shukla, R., & Cheryan, M. (2001). Zein: The industrial protein from corn. *Industrial Crops and Products*, 13(3), 171–192. [https://doi.org/10.1016/S0926-6690\(00\)00064-9](https://doi.org/10.1016/S0926-6690(00)00064-9)
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C.,

- Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., ... 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/s41592-021-01098-3>
- Steiner, B. M., McClements, D. J., & Davidov-Pardo, G. (2018). Encapsulation systems for lutein: A review. *Trends in Food Science & Technology*, 82, 71–81.
<https://doi.org/10.1016/j.tifs.2018.10.003>
- Thomasen, F. E., Skaalum, T., Kumar, A., Srinivasan, S., Vanni, S., & Lindorff-Larsen, K. (2024). Rescaling protein–protein interactions improves Martini 3 for flexible proteins in solution. *Nature Communications*, 15(1), 6645. <https://doi.org/10.1038/s41467-024-50647-9>
- Tortorella, S., Maturi, M., Buratti, V. V., Vozzolo, G., Locatelli, E., Sambri, L., & Franchini, M. C. (2021). Zein as a versatile biopolymer: Different shapes for different biomedical applications. *RSC Advances*, 11(62), 39004–39026. <https://doi.org/10.1039/d1ra07424e>
- Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & Mackerell, A. D., Jr. (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4), 671–690. <https://doi.org/10.1002/jcc.21367>
- Wang, M., Fu, Y., Chen, G., Shi, Y., Li, X., Zhang, H., & Shen, Y. (2018). Fabrication and characterization of carboxymethyl chitosan and tea polyphenols coating on zein nanoparticles to encapsulate β -carotene by anti-solvent precipitation method. *Food Hydrocolloids*, 77, 577–587. <https://doi.org/10.1016/j.foodhyd.2017.10.029>
- Wang, Y., & Zhou, R. (Eds.). (2023). *A practical guide to recent advances in multiscale modeling and simulation of biomolecules*. AIP Publishing.
<https://doi.org/10.1063/9780735425279>
- Wang, Y., Xie, J., Wang, Y., Liao, W., & Gao, Y. (2025). Ethanol-tolerant dual-protein nanoparticles: Inhibiting lactoferrin aggregation and stabilizing Pickering emulsions in ethanol-water system. *Food Hydrocolloids*, 168, 111533.
<https://doi.org/10.1016/j.foodhyd.2025.111533>

- Wilson, C. M. (1991). Multiple zeins from maize endosperms characterized by reversed- phase high-performance liquid chromatography. *Plant Physiology*, 95(3), 777–786. <https://doi.org/10.1104/pp.95.3.777>
- Xue, J., Zhang, Y., Huang, G., Liu, J., Slavin, M., & Yu, L. L. (2018). Zein–caseinate composite nanoparticles for bioactive delivery using curcumin as a probe compound. *Food Hydrocolloids*, 83, 25–35. <https://doi.org/10.1016/j.foodhyd.2018.04.037>
- Yang, J., Li, D., Zhang, Y., Zhang, L., Liao, Z., Aihemaitijiang, S., Hou, Y., Zhan, Z., Xie, K., & Zhang, Z. (2020). Lutein protected the retina from light-induced retinal damage by inhibiting increasing oxidative stress and inflammation. *Journal of Functional Foods*, 73, 104107. <https://doi.org/10.1016/j.jff.2020.104107>
- Yu, X., Afreen, S., Kong, Q., & Wang, J. (2020). Study on self-assembled morphology and structure regulation of α -zein in ethanol–water mixtures. *Langmuir*, 36(40), 11975–11984. <https://doi.org/10.1021/acs.langmuir.0c02143>
- Yuan, Y., Li, H., Liu, C., Zhang, S., Xu, Y., & Wang, D. (2019). Fabrication and characterization of lutein-loaded nanoparticles based on zein and sophorolipid: Enhancement of water solubility, stability, and bioaccessibility. *Journal of Agricultural and Food Chemistry*, 67(43), 11977–11985. <https://doi.org/10.1021/acs.jafc.9b05175>
- Zhang, G., Hu, S., Wang, X., Zhao, Y., Liu, M., Han, J., Prakash, S., Wang, Z., & Ding, Z. (2023). Development and performance evaluation of amorphous microencapsules containing lutein nanoparticles via antisolvent precipitation followed by spray/freeze-drying. *International Journal of Food Science and Technology*, 58(1), 62–73. <https://doi.org/10.1111/ijfs.16161>
- Zhang, X., Zhang, Z., Liang, H., Li, J., Wen, L., Geng, F., & Li, B. (2021). Influence of solvent polarity of ethonal/water binary solvent on the structural, emulsifying, interfacial rheology properties of gliadin nanoparticles. *Journal of Molecular Liquids*, 344, 117976. <https://doi.org/10.1016/j.molliq.2021.117976>
- Zhang, Y., Wu, C., Fang, X., Xu, L., Shen, H., Song, M., Zhang, H., & Luan, G. (2025). Structural transitions and molecular interactions of zein induced by ethanol and extrusion. *Food Chemistry: X*, 103131. <https://doi.org/10.1016/j.fochx.2025.103131>

Chapter 5 - Feasibility of direct lutein encapsulation using corn gluten meal ethanol extract for application in cookies

Abstract

Corn gluten meal (CGM) is a protein-rich coproduct of corn wet milling and a primary industrial source of zein. This study evaluated the feasibility of using the ethanol-soluble fraction of CGM directly for the preparation of lutein-containing powders, examined the effects of the solvent-reduction pathway on their thermal stability, and assessed their performance following incorporation into cookies. Zein-rich materials were collected at successive stages of ethanol extraction and fractionation and compared with commercial zein. Their compositions, protein profiles, and solubilities in 70% ethanol were characterized. The CGM-derived fractions contained 89.68–95.80% protein and were predominantly composed of α -zeins, although their recovery yields, and proximate composition differed according to the fractionation procedure. Molecular dynamics simulations were conducted to examine the ethanol-dependent interactions of zein with lutein and linoleic acid. Simulations showed that lutein could associate with zein at relatively high ethanol concentrations of 60%, whereas persistent zein–linoleic acid contacts emerged near 40% ethanol. Lutein-containing powders were prepared through either rapid or stepwise reduction of ethanol concentration. During heating at 60°C for 72 h, physical mixtures retained only 2.66–7.05% of their initial apparent lutein contents. In contrast, fast-quenched powders retained 35.52–43.86%, whereas stepwise-prepared powders retained 12.84–21.19%. The powder prepared directly from the CGM ethanol extract retained 42.43%, comparable to the 39.94% obtained with commercial zein. After baking, cookies containing encapsulated powders exhibited apparent lutein retentions of 85.65–93.01%, higher than 71.74% for the physical-mixture treatment. Overall, these findings support the feasibility of directly converting a zein-

rich CGM ethanol extract into lutein-containing powders without intermediate protein recovery, while retaining protective functionality following incorporation and baking in a cookie system.

Keywords: Zein; antisolvent precipitation; solvent-reduction pathway; thermal stability; baking stability

5.1. Introduction

Zein is the major prolamin protein of maize and is characterized by limited water solubility and appreciable solubility in aqueous alcohols. These solvent-dependent properties allow zein to be processed through antisolvent precipitation. In this process, zein and a hydrophobic compound are initially dissolved in aqueous ethanol, followed by the addition of water to reduce the solvent quality. The resulting decrease in ethanol concentration promotes zein association and precipitation, leading to the formation of protein–ligand assemblies containing the hydrophobic compound. Owing to its operational simplicity and relatively mild processing conditions, antisolvent precipitation has been widely applied to zein-based encapsulation (Yan et al., 2022).

Many studies of antisolvent-based zein encapsulation have used commercial zein isolate as the starting material (Kim et al., 2025; Wang et al., 2022; Zhao et al., 2022). Commercial zein is generally recovered from corn gluten meal (CGM), a protein-rich coproduct of corn wet milling, by extracting the alcohol-soluble protein fraction and subsequently precipitating and drying the recovered protein (Shukla & Cheryan, 2001). A widely cited process for zein recovery is based on the method of Carter and Reck (Carter & Reck, 1970), in which zein is extracted

using 2-propanol, cold-precipitated at approximately -20°C , and recovered after drying. However, this process has been reported to provide limited recovery of the major α -zein fraction (Wu et al., 1997). To improve α -zein recovery and purity, Anderson and Lamsal developed a modified fractionation process (Anderson & Lamsal, 2011). In this method, zein is first extracted using an aqueous alcohol solution, such as 70% ethanol or 88% 2-propanol, followed by precipitation at a higher alcohol concentration. The remaining soluble fraction is subsequently subjected to cold precipitation to recover an α -zein-rich fraction. This approach was reported to improve both the recovery and purity of α -zein.

Because both zein recovery from CGM and antisolvent encapsulation depend on the solubility of zein in aqueous alcohol, the two operations could potentially be integrated. Specifically, the ethanol-soluble fraction extracted from CGM could be used directly for encapsulation, thereby avoiding intermediate protein precipitation, drying, and redissolution. In addition, alcohol fractionation and cold precipitation may generate zein-rich materials with different protein compositions and relative abundances of zein fractions. It therefore remains unclear whether direct CGM extracts and progressively fractionated CGM-derived materials provide comparable encapsulation and protective functionality.

A further question is whether the pathway-dependent lutein stability observed in the previous chapter is retained in CGM-derived materials. Chapter 4 showed that rapid and stepwise reductions in ethanol concentration produced different lutein distributions and thermal stabilities in a pretreated commercial zein system. Because CGM-derived materials may contain different zein fractions, and other co-extracted components, it is uncertain whether the same pathway-dependent behavior persists in these less-refined systems. Lastly, the previous chapter evaluated

thermal stability only at the powder level and did not assess whether the protective effect would be maintained after incorporation into an actual food matrix. Therefore, it would be valuable to evaluate the performance of the lutein-containing powders during baking, which represents a practical food-processing operation involving substantial thermal stress.

To address these questions, this study compared commercial zein with CGM-derived materials collected at successive stages of ethanol extraction and fractionation. Lutein-containing powders were prepared using rapid and stepwise solvent-reduction pathways to determine whether (i) the ethanol-soluble CGM extract could be used directly for lutein incorporation, (ii) downstream fractionation and differences in protein composition affected encapsulation performance, (iii) the pathway-dependent thermal stability observed with commercial zein was also observed in CGM-derived systems, and (iv) the encapsulated powders could provide protective functionality in a cookie matrix during baking.

5.2. Materials and methods

5.2.1. Materials

Corn gluten meal was purchased from CZ Grain LLC (Russell, IA, USA). Commercial zein was obtained from Sigma-Aldrich (St. Louis, MO, USA), while lutein with a stated purity of 90% was supplied by Thermo Fisher Scientific (Waltham, MA, USA). Analytical-grade ethanol was sourced from Decon Laboratories, Inc. (King of Prussia, PA, USA). Hexane, sodium hydroxide and sodium bisulfite, and all remaining chemicals were purchased from Thermo Fisher Scientific. Ethanol concentrations are reported throughout this chapter as weight fractions (w/w).

5.2.2. Zein fractionation

Zein was extracted and fractionated from CGM using a modified method based on Anderson and Lamsal (Anderson & Lamsal, 2011). Briefly, CGM (23.33 g, dry basis) was suspended in 140 g of 70% ethanol containing 0.5% sodium bisulfite and 0.25% NaOH. The suspension was vigorously agitated at 60°C for 2 h and centrifuged at 12,000 rpm. The recovered supernatant was designated as corn gluten meal extract (CGME). Absolute ethanol was slowly added to CGME until the ethanol concentration reached 95%. After stirring for 30 min, the mixture was centrifuged at 12,000 rpm. The resulting supernatant was designated as the before-cold-precipitation fraction (BCP), and the pellet was collected as the insoluble fraction (IP). BCP was then maintained at −18°C for 2 h to induce cold precipitation. Following centrifugation, the precipitated solid was collected as the after-cold-precipitation fraction (ACP).

Independent extraction batches were used to recover CGME, BCP, IP, and ACP at their respective stages of the fractionation process. Each recovered fraction was extensively dialyzed, freeze-dried, and pulverized with coffee grinder to obtain a powdered sample. The overall fractionation procedure is illustrated in Fig. 5.1.

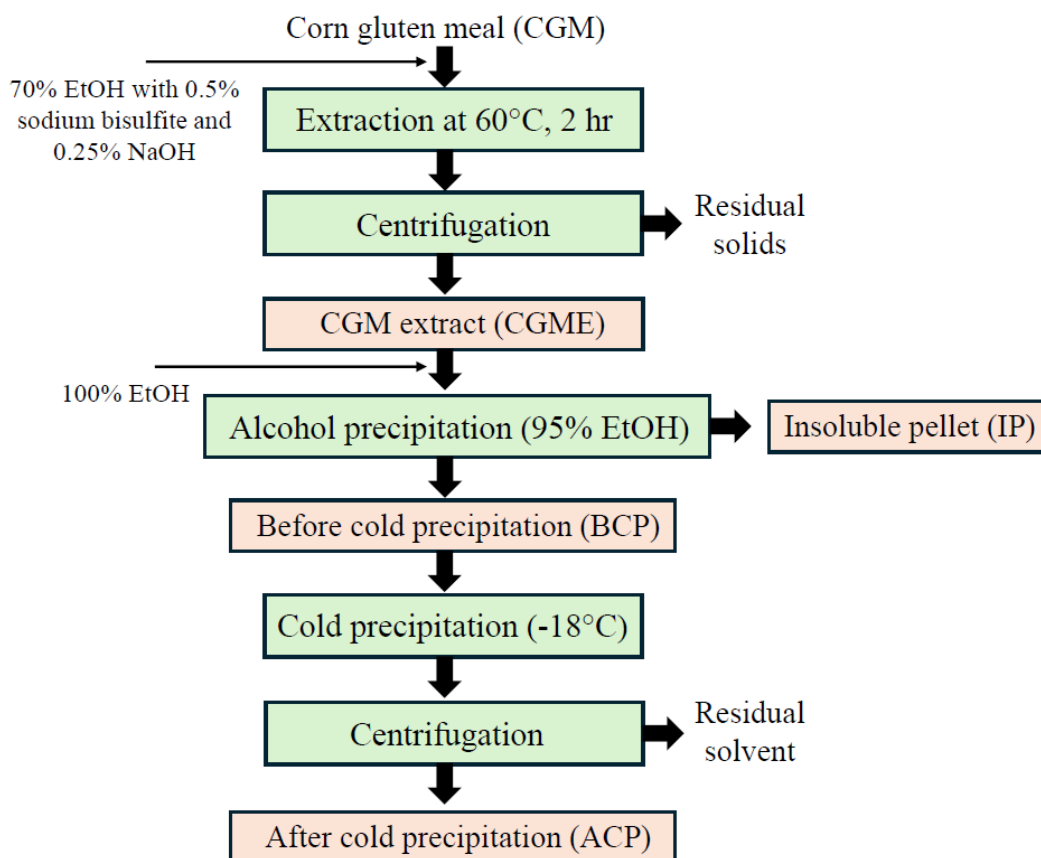


Figure 5.1. Schematic representation of zein extraction and sequential fractionation from corn gluten meal.

5.2.3. Proximate analysis

Proximate composition was analyzed by the University of Missouri Agricultural Experiment Station Chemical Laboratories (Columbia, MO, USA). Moisture, ash, crude protein, crude fat, and crude fiber contents were determined using standard AOAC procedures. Crude protein was measured by combustion using a LECO nitrogen analyzer and calculated as total nitrogen $\times 6.25$. Crude fat was determined by Soxhlet extraction, while moisture and ash were measured by oven drying and incineration, respectively. Analyses were performed in duplicate using independently prepared sample batches.

5.2.4. SDS-PAGE

Because zein exhibits limited solubility in conventional aqueous buffers, the CGM-derived fractions were initially dispersed in 70% ethanol at a concentration of 2% (w/w). After vortex mixing for 1 h, the samples were centrifuged at 12,000 rpm, and the supernatants were collected for analysis. An aliquot of the protein solution (20 μ L) was mixed with 20 μ L of Laemmli sample buffer containing 10% β -mercaptoethanol, 20 μ L of 2% SDS solution, and 20 μ L of glycerol. Glycerol was included to facilitate stable loading of the samples into the gel wells. The mixtures were heated at 100°C for 5 min and then loaded onto a 4% polyacrylamide gel. Following electrophoresis, the gel was stained with Coomassie Brilliant Blue to visualize the protein bands.

5.2.5. Solubility assay

The apparent solubility of the CGM-derived fractions in 70% ethanol was evaluated using the bicinchoninic acid (BCA) assay. A calibration curve was prepared using commercial zein dissolved in 70% ethanol at concentrations ranging from 0 to 1 mg/mL ($R^2 = 0.993$), following a previously reported method (Podaralla & Perumal, 2012). Briefly, each CGM-derived fraction (5 mg) was dispersed in 10 mL of 70% ethanol, vigorously vortexed for 1 h, and centrifuged at 12,000 rpm. The resulting supernatant was passed through a 0.22 μ m syringe filter and subjected to the BCA reaction. Absorbance was measured at 562 nm, and the soluble protein concentration was determined from the zein calibration curve.

5.2.6. Molecular dynamics simulation

Molecular dynamics simulations were conducted to examine the interactions among zein, lutein, and linoleic acid across different ethanol concentrations. Linoleic acid was selected as a representative lipid component because it has been reported as a major fatty acid in CGM (Yururdurmaz & Yildiz, 2022). Parameters for lutein and linoleic acid were generated using the CHARMM General Force Field (CGenFF) (Vanommeslaeghe et al., 2010). The protein model was obtained from the zein structure previously equilibrated in 70% ethanol in the preceding chapter.

For methodological consistency, the simulation procedure followed that described in the previous chapter. Each system was energy-minimized using the steepest-descent algorithm, followed by 125-ps NVT and NPT equilibration steps. Production simulations were performed for 200 ns in ethanol–water mixtures ranging from 70 to 20% ethanol at 10% intervals. All simulations were conducted using the CHARMM36 force field (Huang & MacKerell, 2013) and GROMACS version 2022.5 (Bauer et al., 2022).

5.2.7. Powder formulation

To prepare encapsulated zein–lutein powders, each CGM-derived fraction (10 g) and lutein (10 mg) were dispersed in 200 g of 70% (w/w) ethanol. Two solvent-reduction pathways were used to decrease the final ethanol concentration to 10%. For fast quenching, 1,200 g of deionized water was added rapidly (~1 s) under continuous stirring, directly reducing the ethanol concentration from 70 to 10%. For stepwise quenching, 500 g of deionized water was first added dropwise using a burette to gradually reduce the ethanol concentration from 70 to 20%. The

suspension was equilibrated for 10 min at 20% ethanol, after which the remaining 700 g of deionized water was added rapidly to reach the final ethanol concentration of 10%.

For direct encapsulation, the amount of CGME solution corresponding to 10 g of recovered CGME solids was estimated from the measured CGME recovery yield of 33.20% (Table 5.1). This corresponded to 180.7 g of 70% ethanol extract. An additional 19.3 g of 70% ethanol was added to provide the same nominal formulation basis of 10 g solids per 200 g of 70% ethanol used for the powder-derived samples. Lutein powder (10 mg) was then added directly to the adjusted CGME solution before quenching. Following quenching, the CGMD suspension was centrifuged at 12,000 rpm, and the recovered solids were lyophilized, pulverized, and passed through a 150 μm sieve in the same manner as the other formulations. The resulting powder was designated as CGMD (corn gluten meal direct encapsulation).

5.2.8. Powder thermal stability

To evaluate changes in lutein content under accelerated degradation conditions, powder samples (100 mg) were sealed in centrifuge tubes and incubated in a convection oven at 60°C. Samples were collected after 4, 8, 24, 48, and 72 h of heating. To assess differences in lutein extractability within the powders, sequential extraction with hexane followed by 70% ethanol was performed. The resulting hexane-extractable and ethanol-released fractions were quantified by UV–Vis spectrophotometry at 445 nm and reported as apparent lutein contents.

5.2.9. Cookie formulation

Cookies were prepared using a modified version of the method reported by Jeong et al. (Jeong et al., 2024). Each treatment consisted of soft wheat flour (76.40 g), lutein-containing powder (10.00 g), sucrose (53.80 g), shortening (30.00 g), distilled water (12.15 g), baking powder (1.90 g), and sodium chloride (0.27 g). The tested powders included the physical mixture and fast-quenched CZ, ACP, CGME, and CGMD powders, as defined previously. For the control formulation, the treatment powder was replaced with an equal mass of flour, resulting in a total flour content of 86.40 g.

Shortening and sucrose were mixed for 3 min. Sodium chloride was dissolved in distilled water, added to the shortening mixture, and mixed for an additional 1 min. The flour, baking powder, and corresponding lutein-containing powder were premixed and incorporated into the wet ingredients, followed by mixing for 1 min. A portion of each dough (5 g) was collected before baking to determine the initial lutein content. The remaining dough was rolled to a thickness of 7 mm and cut into circular pieces using a 6-cm-diameter cutter. The cookies were baked at 205°C for 10 min and cooled at room temperature for 1 h. The collected dough samples and baked cookies were freeze-dried, ground, and thoroughly homogenized before lutein extraction.

5.2.10. Apparent lutein retention in cookie

Freeze-dried dough and cookie samples were finely ground before analysis. Each powdered sample (2.00 g) was successively extracted three times with 10 mL of 70% ethanol per

extraction, based on a previously reported method for lutein-fortified bread (Kwon et al., 2023). After each extraction, the suspension was centrifuged, and the supernatant was collected. The supernatants from the three extraction cycles were combined for analysis.

Apparent lutein concentration was determined by UV–Vis spectrophotometry at 445 nm using a calibration curve prepared with lutein standards in 70% ethanol. To account for absorbance contributed by the cookie matrix and baking-derived compounds, the absorbance of an identically extracted lutein-free control sample was subtracted from that of each lutein-containing sample. The corrected values were reported as apparent extractable lutein contents. Apparent lutein retention after baking was calculated as the apparent extractable lutein content of the baked cookie divided by that of the corresponding unbaked dough and expressed as a percentage (equation 1).

$$\text{Apparent lutein retention (\%)} = \left(\frac{C_{\text{cookie}}}{C_{\text{dough}}} \right) \times 100 \quad (1)$$

Where C_{cookie} and C_{dough} represents the apparent extractable lutein contents of the freeze-dried baked cookie and dough, respectively.

5.2.11. Statistical analysis

Unless otherwise specified, wet-laboratory experiments and fractionations were independently performed in triplicate using separately prepared batches. Statistical significance was determined using analysis of variance (ANOVA) and Duncan's test under R package at 95% confidence level. All values are presented as mean $\pm$ standard deviation.

5.3. Results and discussion

5.3.1. Proximate analysis and yield

The proximate composition and recovery yields of the CGM-derived fractions prepared using the process described in Section 5.2.2 are presented in Table 5.1. The protein content of the original CGM was 64.22%, whereas the protein contents of the derived fractions ranged from 89.68 to 95.80%. Among the recovered fractions, ACP exhibited the highest protein content at 95.80%, followed by IP (91.05%), BCP (89.98%), and CGME (89.68%). These values were comparable to or higher than that of commercial zein (89.06%), indicating substantial protein enrichment through ethanol-based fractionation.

Table 5.1. Proximate composition and recovery yields of commercial zein, corn gluten meal, and corn gluten meal-derived fractions. Different lowercase letters within the same column indicate significant differences among samples ($p < 0.05$). Yield was calculated relative to the initial total dry mass of CGM. CZ, commercial zein; CGM, corn gluten meal; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction.

	Protein (%)	Moisture (%)	Fat (%)	Fiber (%)	Ash (%)	Yield (%)
CZ	89.06 ±0.44c	3.66 ±0.00e	3.13 ±0.01c	0.00 ±0.00b	2.15 ±0.04a	—
CGM	64.22 ±0.32d	6.05 ±0.01a	1.30 ±0.03d	0.84 ±0.00a	2.17 ±0.02a	—
CGME	89.68 ±0.23c	5.38 ±0.02b	3.48 ±0.04b	0.00 ±0.00c	0.15 ±0.03c	33.20 ±1.49a
IP	91.05 ±0.06b	4.85 ±0.22c	0.31 ±0.08e	0.08 ±0.01b	0.64 ±0.02b	11.17 ±0.50c
BCP	89.98 ±0.14c	4.71 ±0.04c	4.95 ±0.14a	0.04 ±0.06b	0.19 ±0.01c	9.96 ±0.86d

ACP	95.80 ±0.06a	4.32 ±0.07d	0.00 ±0.00f	0.00 ±0.00c	0.22 ±0.00c	15.12 ±1.70b

Fat content differed depending on the fraction recovery procedure. BCP and CGME contained 4.95 and 3.48% fat, respectively, whereas IP contained only 0.31% fat and no detectable fat was observed in ACP. CGME and BCP were recovered through solvent removal from the entire ethanol-soluble fraction without an additional selective protein precipitation step. Consequently, ethanol-soluble non-protein components, including lipids, were likely co-recovered with the proteins. In contrast, cold precipitation selectively recovered a protein-rich ACP fraction, while a substantial proportion of the soluble lipids likely remained in the supernatant. The visual appearance of the powders was consistent with these compositional differences (Fig. 5.2). CGME and BCP exhibited intense yellow coloration similar to that of commercial zein, whereas IP and ACP were considerably lighter in color. This coloration was consistent with the co-recovery of endogenous corn carotenoids in the ethanol-soluble fractions.



Figure 5.2. Visual appearance of CGM, CGME, IP, BCP, ACP, and CZ powders. CGM, corn gluten meal; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction; and CZ, commercial zein

CGME showed the highest recovery yield at 33.20%, followed by ACP at 15.12%, IP at 11.17%, and BCP at 9.96%. Thus, recovery of the entire ethanol-soluble fraction maximized the mass yield but also retained more non-protein components, whereas cold precipitation generated a fraction with higher protein content and undetectable fat at a lower overall yield.

5.3.2. SDS-PAGE profiles of CGM-derived fractions

The protein profiles of commercial zein and the CGM-derived fractions were examined using SDS-PAGE (Fig. 5.3). The bands were assigned based on the work of Hamaker et al., (Hamaker et al., 1995). All samples exhibited dominant banding in the approximately 19–22 kDa region, corresponding to the 19- and 22-kDa α -zeins, which constitute the major zein protein classes in maize. Commercial zein showed a particularly intense band in this region, and similar α -zein-dominant profiles were observed in CGME, BCP, and ACP. These results indicate that the ethanol-based extraction and recovery procedures retained α -zein as the predominant protein component of the CGM-derived fractions.

In contrast, IP exhibited a broad and heterogeneous banding pattern, with several additional bands above and below the major α -zein region. This profile suggests that the ethanol-insoluble fraction retained a wider range of CGM proteins. In contrast, BCP and ACP were strongly dominated by bands in the approximately 19–22 kDa region, corresponding to the major α -zeins, with relatively few additional bands. Their similar profiles indicate that both fractions

were enriched in α -zeins and that cold precipitation did not substantially alter the dominant protein composition.

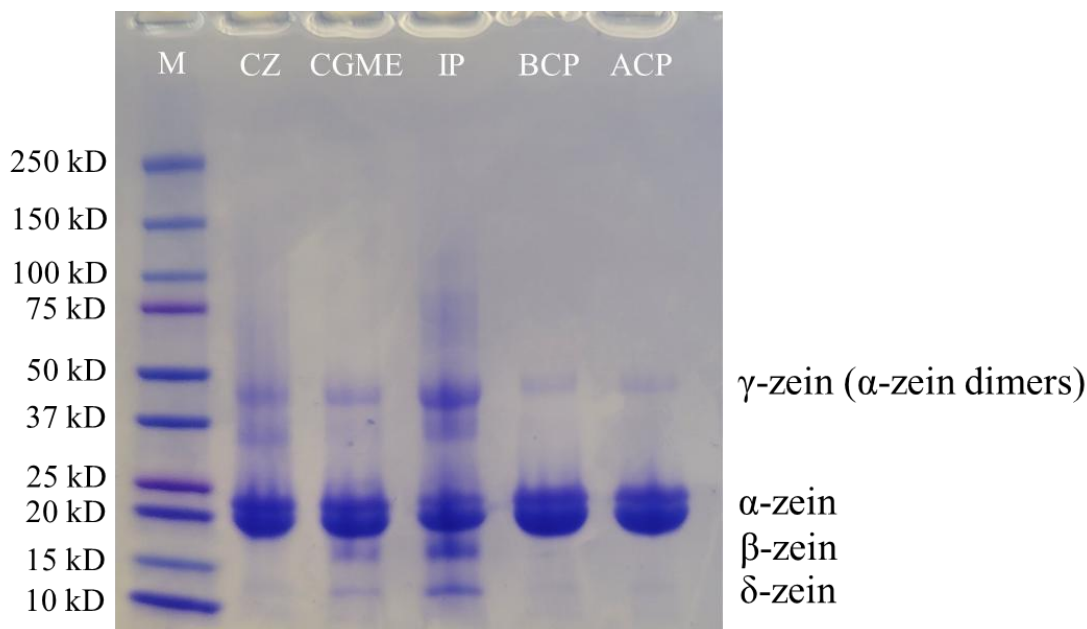


Figure 5.3. SDS-PAGE profiles of commercial zein and CGM-derived fractions. Lanes: M, molecular-weight marker; CZ, commercial zein; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; and ACP, after-cold-precipitation fraction.

5.3.3. Protein solubility in 70% EtOH

The apparent soluble protein concentrations of commercial zein and the CGM-derived fractions in 70% ethanol are shown in Fig. 5.4. Commercial zein, CGME, BCP, and ACP exhibited similar soluble protein concentrations, ranging from approximately 0.50 to 0.54 mg/mL, with no significant differences among these samples. ACP showed the numerically highest soluble protein concentration, which likely reflect its higher protein composition (Table 5.1). In contrast, IP exhibited a soluble protein concentration of approximately 0.15 mg/mL, which was significantly lower than those of all other samples. This value was approximately

70% lower than those observed for commercial zein and the ethanol-soluble CGM-derived fractions.

The low ethanol solubility of IP was notable because this fraction contained 91.05% protein and exhibited prominent zein bands in the SDS-PAGE analysis. Given that the IP fraction was recovered as a pellet after increasing the ethanol concentration of CGME to 95%, it could be possible that precipitation followed by drying produced a poorly redispersible protein structure. Protein dehydration and strong protein–protein association during precipitation may have contributed to the formation of aggregates that were not readily dissociated upon re-exposure to 70% EtOH. Because the subsequent antisolvent encapsulation process required dissolution of the protein fraction in 70% EtOH for preparation and quantification, IP was excluded from the encapsulated powder preparation experiments.

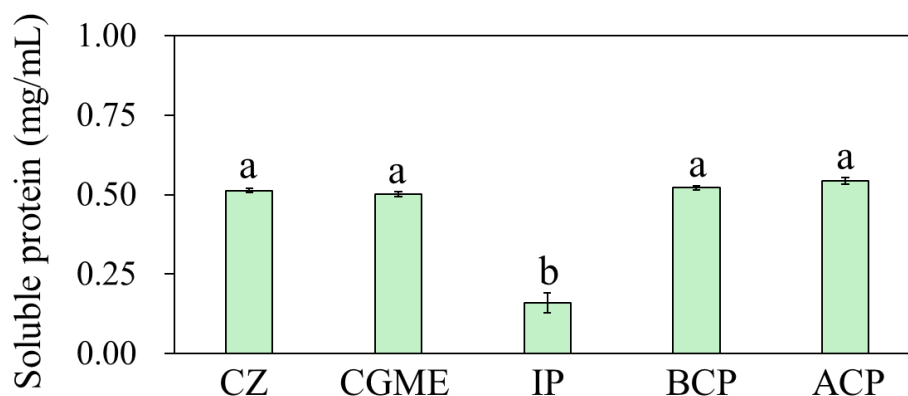


Figure 5.4. Apparent soluble protein concentrations of commercial zein and CGM-derived fractions in 70% EtOH. CZ, commercial zein; CGME, corn gluten meal extract; IP, insoluble fraction; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction

5.3.4. Molecular dynamics simulation

In the previous chapter, the ethanol-dependent association between zein and lutein was examined in the absence of other hydrophobic components. Because the CGM-derived fractions contained residual fat, linoleic acid was included in the present simulation as a representative fatty acid to determine whether its presence affected the molecular behavior of the zein–lutein system during ethanol reduction.

Representative simulation configurations and protein–ligand contact numbers are shown in Fig. 5.5a and b, respectively. Lutein began to form contacts with zein at 60% ethanol, and the number of zein–lutein contacts progressively increased as the ethanol concentration decreased. In contrast, linoleic acid showed only limited association with zein between 70 and 50% ethanol. Substantial and persistent zein–linoleic acid contacts emerged at 40% ethanol and were maintained as the ethanol concentration was further reduced to 30 and 20%.

For both ligands, a pronounced increase in protein–ligand contacts occurred at 40% ethanol. This increase coincided with the conformational collapse of zein and the formation of a more compact protein–ligand assembly. The observed transition was consistent with the ethanol-dependent collapse previously observed for the zein–lutein system in Chapter 4. Thus, the inclusion of linoleic acid did not appear to substantially alter either the ethanol-dependent association of zein with lutein or the overall conformational response of zein at the monomer level. Overall, although lutein and linoleic acid exhibited different initial association behaviors at relatively high ethanol concentrations, the simulations suggest that both hydrophobic compounds can become incorporated into mixed zein–ligand assemblies under low-ethanol conditions.

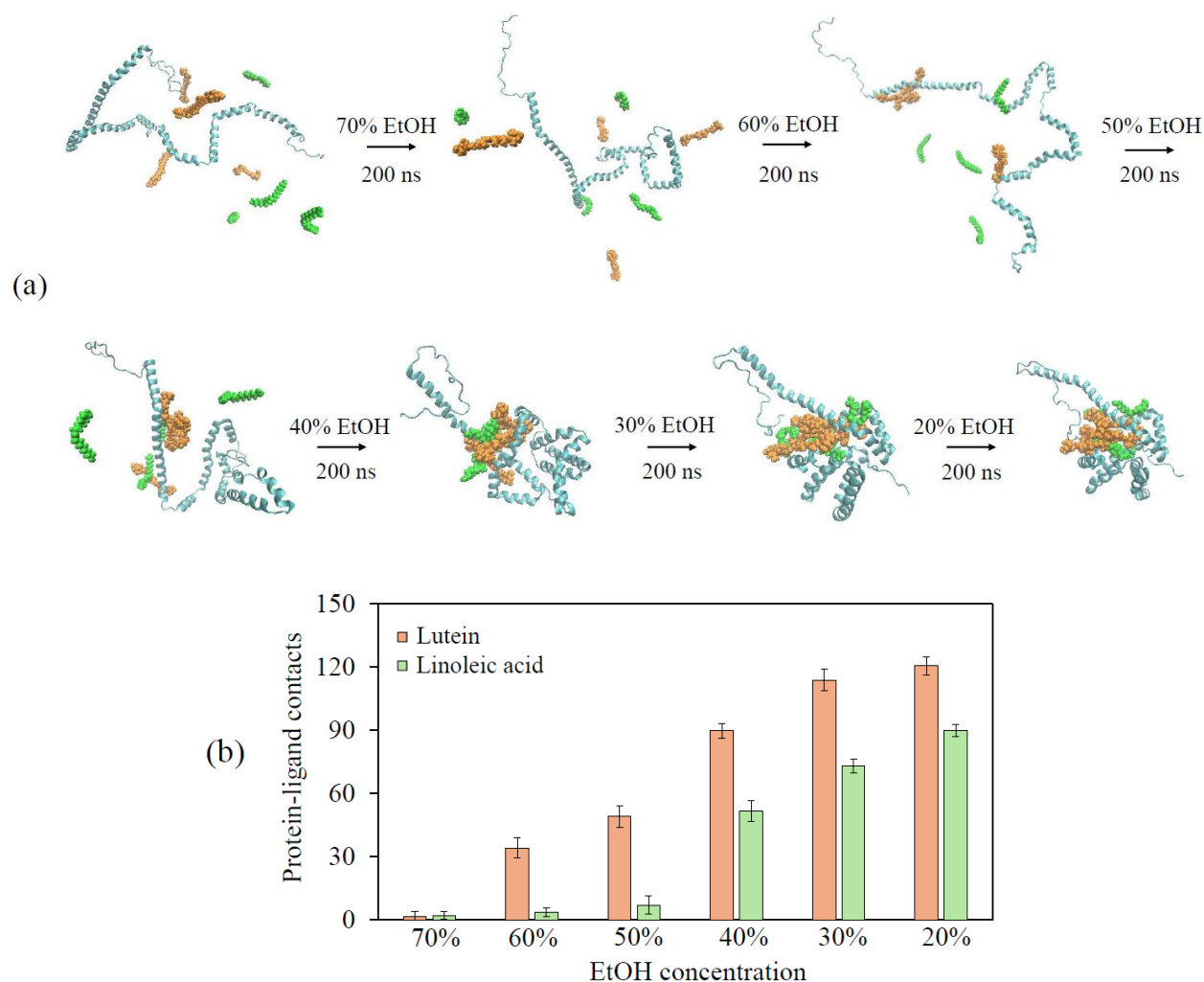


Figure 5.5. Ethanol-dependent association of zein with lutein and linoleic acid. (a) Representative configurations obtained after 200 ns of simulation at each ethanol concentration. Zein is shown in cyan, lutein in orange, and linoleic acid in green. (b) Average numbers of zein–lutein and zein–linoleic acid contacts calculated over the final 50 ns of each 200 ns simulation.

5.3.5. Powder thermal stability

The thermal stability of lutein in physical mixtures containing commercial zein or CGM-derived protein fractions was first evaluated during heating at 60°C for 72 h (Table 5.2).

Apparent lutein content, determined as the sum of the hexane-extractable and ethanol-released

fractions, decreased substantially with increasing heating time in all physical mixtures. After 72 h, total lutein retention ranged from only 2.66 to 7.05%. BCP exhibited the highest retention at 7.05%, followed by commercial zein and CGME at 5.39 and 5.25%, respectively, whereas ACP showed the lowest retention at 2.66%. Although some differences were observed among the protein sources, the overall low retention demonstrated that simple physical mixing provided limited protection of lutein against thermal degradation.

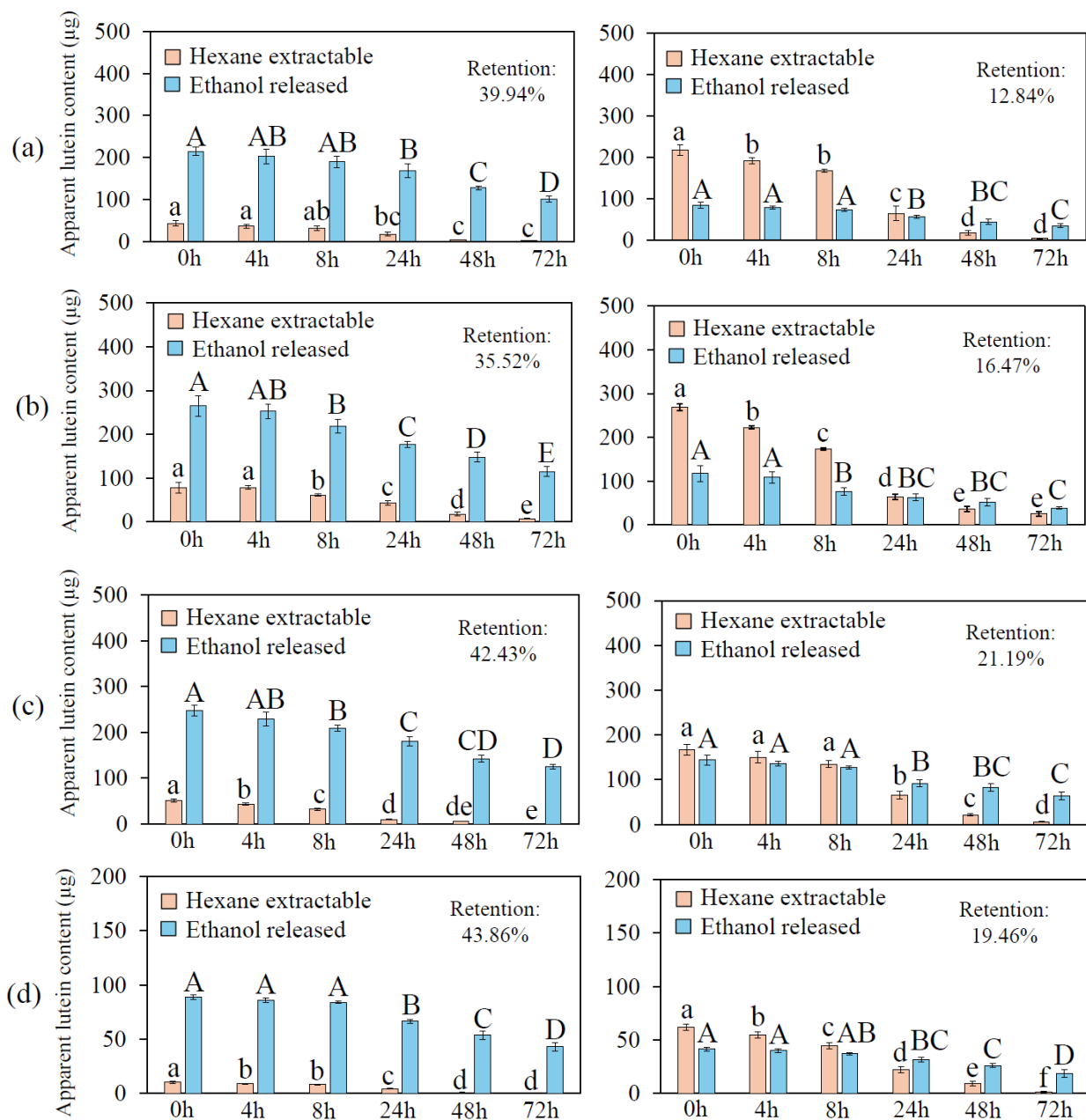
Encapsulation markedly improved lutein retention compared with the corresponding physical mixtures (Fig. 5.6). In powders prepared by fast quenching, the 72h retention values ranged from 35.52 to 43.86%. ACP exhibited the numerically highest retention at 43.86%, followed by CGMD at 42.43%, commercial zein at 39.94%, BCP at 37.20%, and CGME at 35.52%. In contrast, powders prepared by stepwise ethanol reduction showed lower retention values ranging from 12.84 to 21.19%. CGMD exhibited the numerically highest retention among the stepwise-prepared powders at 21.19%, followed by ACP at 19.46%, BCP at 19.38%, CGME at 16.47%, and commercial zein at 12.84%.

The distribution of lutein between the two operationally defined extraction fractions also differed according to the powder preparation pathway. In the fast-quenched powders, most of the initial lutein was resistant to direct hexane extraction and was subsequently released by ethanol. In contrast, the stepwise-prepared powders generally contained a larger initial proportion of hexane-extractable lutein. Moreover, fast quenching consistently resulted in higher final retention than stepwise ethanol reduction for every tested protein source. Rapid reduction of ethanol concentration may have promoted the immediate co-precipitation of zein and lutein, thereby limiting lutein redistribution and producing a protein-rich matrix with reduced solvent

accessibility. During stepwise ethanol reduction, lutein and protein molecules had more time to reorganize or form separate ligand-rich domains before complete precipitation, potentially leaving a greater proportion of lutein at exposed or weakly protected locations. Thus, the pathway-dependent trends in lutein distribution and thermal stability observed in the previous chapter appeared to be retained across both commercial zein and CGM-derived protein fractions.

Table 5.2. Changes in apparent lutein content and final retention of physical mixtures during heating at 60°C. Within each cell, the upper and lower values represent the hexane-extractable and ethanol-released lutein contents, respectively. Different lowercase and uppercase letters within the same row indicate significant differences across heating times in the hexane-extractable and ethanol-released fractions, respectively. Different letters in the retention column indicate significant differences among physical mixtures. Retention (%) was calculated as the total apparent lutein content at 72 h divided by that at 0 h $\times$ 100. The amount of lutein is expressed as μg per 100 mg of powder. CZ, commercial zein; CGME, corn gluten meal extract; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction

	0 h	4 h	8 h	24 h	48 h	72 h	Retention (%)
CZ	261.52 $\pm$ 3.21a 48.02 $\pm$ 3.41A	53.49 $\pm$ 5.88b 35.48 $\pm$ 1.27B	29.04 $\pm$ 2.45c 16.69 $\pm$ 0.49C	9.07 $\pm$ 1.84d 12.30 $\pm$ 1.41D	5.05 $\pm$ 1.99e 10.80 $\pm$ 1.59D	4.37 $\pm$ 1.56e 12.32 $\pm$ 1.35D	5.39 $\pm$ 0.67B
CGME	282.64 $\pm$ 8.57a 78.39 $\pm$ 5.97A	84.94 $\pm$ 9.41b 44.44 $\pm$ 8.92B	43.17 $\pm$ 5.29c 37.94 $\pm$ 3.59B	16.41 $\pm$ 2.43d 21.40 $\pm$ 3.31C	6.67 $\pm$ 2.17e 19.38 $\pm$ 4.80C	4.94 $\pm$ 0.86e 14.02 $\pm$ 0.72D	5.25 $\pm$ 0.35B
BCP	229.16 $\pm$ 10.20a 134.05 $\pm$ 4.53A	93.40 $\pm$ 2.27b 109.96 $\pm$ 7.67B	42.30 $\pm$ 4.84c 75.25 $\pm$ 5.16C	21.39 $\pm$ 3.53d 30.39 $\pm$ 3.33D	9.27 $\pm$ 1.66e 20.85 $\pm$ 4.56DE	8.31 $\pm$ 0.88e 17.28 $\pm$ 3.04E	7.05 $\pm$ 0.90A
ACP	92.39 $\pm$ 1.66a 11.08 $\pm$ 1.27A	26.28 $\pm$ 1.24b 4.89 $\pm$ 1.00B	11.30 $\pm$ 0.78c 4.23 $\pm$ 0.55B	3.56 $\pm$ 0.65d 2.64 $\pm$ 0.48C	1.81 $\pm$ 0.78de 2.07 $\pm$ 0.51C	1.07 $\pm$ 0.05e 1.63 $\pm$ 0.68C	2.66 $\pm$ 0.66C



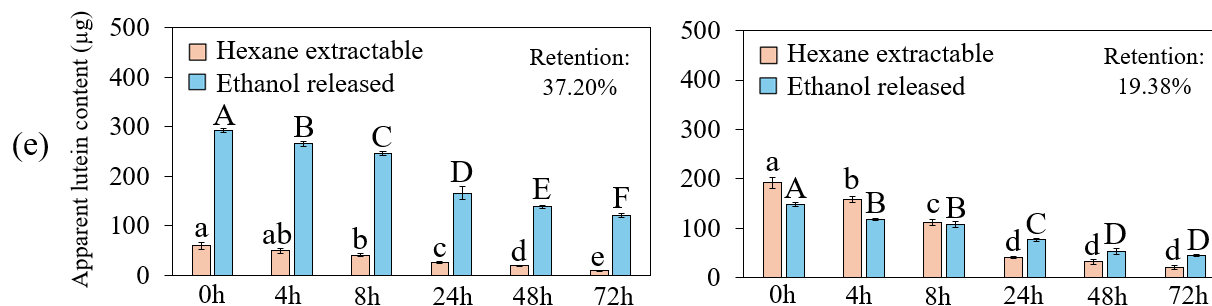
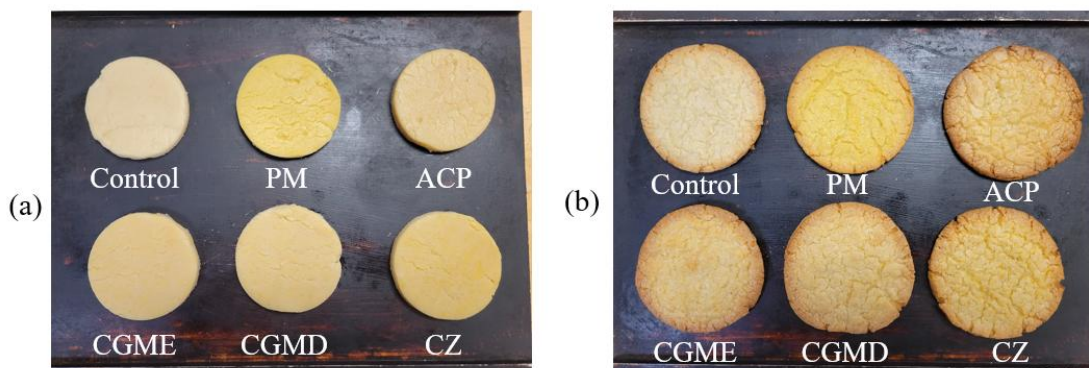


Figure 5.6. Thermal stability and sequential extractability of lutein in encapsulated powders prepared by fast and stepwise quenching during heating at 60°C: (a) CZ, (b) CGME, (c) CGMD, (d) ACP, and (e) BCP. The left and right panels represent powders prepared by fast and stepwise quenching, respectively. Values are presented as means $\pm$ standard deviations. Different lowercase and uppercase letters within each panel indicate significant differences across heating times in the hexane-extractable and ethanol-released fractions, respectively. The amount of lutein is expressed as μg per 100 mg of powder. Retention was calculated by dividing the total apparent lutein content at 72 h by that at 0 h $\times$ 100. CZ, commercial zein; CGME, corn gluten meal extract; CGMD, corn gluten meal direct powder; BCP, before-cold-precipitation fraction; ACP, after-cold-precipitation fraction

5.3.6. Lutein retention before and after baking

The visual appearance of the cookie doughs before baking and the corresponding baked cookies is shown in Fig. 5.7a and b, respectively. The control formulation exhibited the light cream-to-golden coloration expected from the base cookie formulation, whereas the lutein-containing treatments showed varying degrees of yellow coloration. The physical-mixture formulation displayed a particularly pronounced yellow appearance both before and after baking. This may have resulted from the greater accessibility of lutein in the physical mixture, which could have allowed it to disperse more readily within the shortening phase. However, this interpretation remains qualitative because instrumental color measurements and direct analyses of lutein distribution within the cookie matrix were not performed.

Apparent lutein retention after baking differed among the formulations (Fig. 5.7c). The physical-mixture treatment showed the lowest retention at 71.74%, whereas cookies containing encapsulated powders exhibited numerically higher retention values ranging from 85.65 to 93.01%. CZ showed the highest numerical retention at 93.01%, followed by ACP at 91.97%, CGMD at 88.15%, and CGME at 85.65%. The retention values of CZ, ACP, and CGMD were significantly higher than that of the physical mixture, whereas CGME showed an intermediate value. The improved retention observed for most encapsulated formulations indicates that incorporating lutein into a zein-rich matrix before its addition to the cookie formulation reduced lutein degradation during baking. Notably, CGMD provided apparent lutein retention comparable to that obtained with commercial zein and the other CGM-derived fractions. These results suggest that direct encapsulation from the CGM ethanol extract can produce lutein-containing powders that retain their protective function after incorporation into a baked food system.



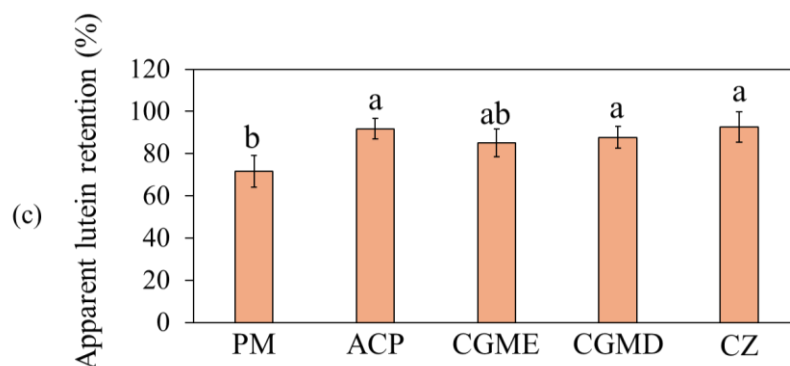


Figure 5.7. Visual appearance and apparent lutein retention of cookie formulations before and after baking. Representative images of (a) unbaked doughs and (b) baked cookies, and (c) apparent lutein retention after baking. PM, physical mixture; ACP, after-cold-precipitation fraction; CGME, corn gluten meal extract; CGMD, corn gluten meal direct powder; and CZ, commercial zein.

5.4. Conclusion

This study evaluated whether zein-rich fractions recovered from corn gluten meal could be directly incorporated into an ethanol-based antisolvent process for the preparation of lutein-containing powders. Ethanol extraction and subsequent fractionation substantially increased the protein content of the CGM-derived fractions, although the recovery method affected their yields and the extent to which lipids and other ethanol-soluble components were co-recovered. The major ethanol-soluble fractions exhibited α -zein-dominant SDS-PAGE profiles, proximate compositions broadly comparable to those of commercial zein, although high-ethanol precipitation further enriched the fraction of α -zein and protein composition.

The molecular simulations indicated that lutein and linoleic acid differed in their initial association with zein at relatively high ethanol concentrations. Nevertheless, both compounds formed persistent contacts with zein following the conformational transition observed at approximately 40% ethanol. Therefore, simulation suggested that at monomer-level, the presence

of a representative fatty acid did not substantially alter the ethanol-dependent collapse and ligand-association behavior previously observed for the zein–lutein system.

At the powder level, encapsulation substantially improved thermal lutein retention compared with physical mixing. More importantly, fast quenching consistently produced higher retention than stepwise ethanol reduction across both commercial zein and CGM-derived fractions. These results indicate that the pathway-dependent functional performance observed for commercial zein in the previous chapter was also maintained in powders prepared from CGM-derived zein-rich fractions. The CGM-derived powders also retained their protective function following incorporation into a cookie system, with several fractions providing baking retention comparable to that of commercial zein.

Several limitations should be considered when interpreting these findings. First, the analytical method quantified apparent lutein but did not distinguish lutein from zeaxanthin or potentially other co-extracted carotenoids originating from CGM. The reported values may therefore include contributions from structurally related endogenous carotenoids. Second, the proximate composition and SDS-PAGE profiles of the ethanol-soluble CGM fractions were generally similar to those of commercial zein. Consequently, the present results do not demonstrate that residual CGM components provided a unique protective effect or that the CGM-derived fractions were functionally superior to purified zein.

A further limitation is that the practical and economic advantages of the proposed process is not demonstrated. While the process described in this chapter removes the drying and possible purification step in the production of commercial zein, it is questionable whether antisolvent precipitation step would have practical implication, since it requires the addition of a substantial

quantity of water, which dilutes the ethanol stream and may increase solvent recovery and energy costs. Instead, the main implication is the potential to integrate an ethanol-based value-adding operation with existing zein extraction from CGM. Within such an integrated process, the zein-rich ethanol extract could potentially be directed toward encapsulated powder production before complete zein purification and solvent removal.

Overall, this study provides proof of concept that the antisolvent processing pathway and its associated functional consequences can be transferred from commercial zein to CGM-derived zein-rich fractions, while demonstrating food-level functionality in baked goods.

References

- Anderson, T. J., & Lamsal, B. P. (2011). Development of new method for extraction of α -zein from corn gluten meal using different solvents. *Cereal Chemistry*, 88(4), 356–362. <https://doi.org/10.1094/CCHEM-08-10-0117>
- Bauer, P., Hess, B., & Lindahl, E. (2022). *GROMACS 2022 source code* (Version 2022) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.6103835>
- Carter, R., & Reck, D. R. (1970). *Low-temperature solvent extraction process for producing high-purity zein* (U.S. Patent No. 3,535,305). U.S. Patent and Trademark Office. <https://patents.google.com/patent/US3535305A/en>
- Hamaker, B. R., Mohamed, A. A., Habben, J. E., Huang, C. P., & Larkins, B. A. (1995). Efficient procedure for extracting maize and sorghum kernel proteins reveals higher prolamin contents than the conventional method. *Cereal Chemistry*, 72(6), 583–588.
- 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>
- Jeong, S., Kim, G., Ryu, K., Park, J., & Lee, S. (2024). Effect of different sweeteners on the thermal, rheological, and water mobility properties of soft wheat flour and their application to cookies as an alternative to sugar. *Food Chemistry*, 432, 137193. <https://doi.org/10.1016/j.foodchem.2023.137193>
- Kim, S., Cheng, H. N., & Biswas, A. (2025). Use of zein for nanoencapsulation of eucalyptol. *Discover Polymers*, 2(1), 23. <https://doi.org/10.1007/s44347-025-00035-3>
- Kwon, H., Lee, D. U., & Lee, S. (2023). Lutein fortification of wheat bread with marigold powder: Impact on rheology, water dynamics, and structure. *Journal of the Science of Food and Agriculture*, 103(11), 5462–5471. <https://doi.org/10.1002/jsfa.12621>
- Podaralla, S., & Perumal, O. (2012). Influence of formulation factors on the preparation of zein nanoparticles. *AAPS PharmSciTech*, 13(3), 919–927. <https://doi.org/10.1208/s12249-012-9816-1>
- Shukla, R., & Cheryan, M. (2001). Zein: The industrial protein from corn. *Industrial Crops and Products*, 13(3), 171–192. [https://doi.org/10.1016/S0926-6690\(00\)00064-9](https://doi.org/10.1016/S0926-6690(00)00064-9)
- Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & MacKerell, A. D., Jr. (2010). CHARMM

- general force field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4), 671–690. <https://doi.org/10.1002/jcc.21367>
- Wang, C., Wang, J., Sun, Y., Freeman, K., McHenry, M. A., Wang, C., & Guo, M. (2022). Enhanced stability and oral bioavailability of cannabidiol in zein and whey protein composite nanoparticles by a modified anti-solvent approach. *Foods*, 11(3), 376. <https://doi.org/10.3390/foods11030376>
- Wu, S., Myers, D. J., & Johnson, L. A. (1997). Factors affecting yield and composition of zein extracted from commercial corn gluten meal. *Cereal Chemistry*, 74(3), 258–263. <https://doi.org/10.1094/CCHEM.1997.74.3.258>
- Yan, X., Li, M., Xu, X., Liu, X., & Liu, F. (2022). Zein-based nano-delivery systems for encapsulation and protection of hydrophobic bioactives: A review. *Frontiers in Nutrition*, 9, 999373. <https://doi.org/10.3389/fnut.2022.999373>
- Yururdurmaz, C., & Yildiz, H. (2022). Fatty acid compositions of *Zea mays* L. varieties in Turkey. *Progress in Nutrition*, 24, e2022112. <https://doi.org/10.23751/pn.v24i3.13268>
- Zhao, Y., Zhi, J., Huang, S., Zhang, X., Kim, Y.-R., Xu, Y., Wang, D., & Luo, K. (2022). Fabrication of starch/zein-based microcapsules for encapsulation and delivery of fucoxanthin. *Food Chemistry*, 392, 133282. <https://doi.org/10.1016/j.foodchem.2022.133282>

Chapter 6 - Conclusion and future perspectives

6.1. Conclusion

The findings of this dissertation highlight both the potential and the limitations of computational approaches in plant protein research. Plant protein functionality is often shaped by features and relationships that are difficult to isolate experimentally, such as patterns embedded in large protein datasets or monomer-level molecular behavior. Computational methods can make these aspects more accessible by converting complex protein information into forms that can be compared and interpreted.

The results showed that sequence- and structure-level patterns can provide useful information for solubility prediction and protein classification, but their relevance depends on protein source and structural organization. They also showed that monomer-level information can explain solvent-dependent protein responses, but does not necessarily predict functional outcomes when higher-order assembly and processing history become important. Therefore, a central conclusion of this work is that prediction in plant protein systems must be treated as context-dependent. In this sense, the value of computation lies not only in generating predictions, but also in clarifying the conditions under which those predictions remain meaningful.

At the same time, this context dependence also defines an important limitation of the approaches used in this dissertation. Data-driven models are constrained by the quality and domain of the datasets on which they are trained, and molecular simulations are constrained by sampling time and model resolution. Experimental systems introduce additional complexity through source variability and processing kinetics. These limitations do not reduce the value of

computational approaches, but they emphasize that their outputs must be interpreted as conditional information rather than direct substitutes for experimental validation.

Future progress in plant protein research will therefore require closer integration between computational analysis and experimental design. Larger and better-curated protein datasets can improve the reliability of data-driven models, while targeted experiments can help test whether predicted patterns or simulated molecular behaviors remain relevant under food-processing conditions. Similarly, molecular simulations can be most useful when they are designed around specific functional questions and interpreted alongside measurements at the appropriate experimental level.

Overall, this dissertation supports the use of computational approaches as an additional layer of analysis for studying plant protein behavior in food systems. By combining data-driven pattern recognition, molecular-level interpretation, and experimental observation, these approaches can extend what is learned from empirical characterization and support more systematic development of plant-protein-based food systems.

6.2. Future perspectives

The findings of this dissertation suggest several directions through which computational approaches could have broader impact in plant protein research. First, future studies would benefit from larger and better-curated plant protein datasets. Current data-driven models are often limited by datasets generated from specific organisms, expression systems, or narrowly defined functional measurements. Building more standardized datasets for plant and food

proteins would improve model transferability and allow computational predictions to be linked more directly to food-relevant properties.

Second, future computational work should continue to bridge molecular-scale information with higher-order protein organization. The results of this dissertation show that monomer-level properties can provide useful insight, but they are not always sufficient to explain functionality in assembled or processed systems. More effective use of multiscale and process-informed models could help clarify how molecular interactions develop into functional properties during food processing.

Finally, although the dissertation touches on the application in corn gluten meal system, broader impact will require extending these approaches from purified model proteins to ingredient-level and food-product systems. Plant protein ingredients often contain mixtures of proteins, lipids, pigments, and other components that influence functional behavior in ways that are difficult to predict from purified proteins alone. Applying and calibrating computational and experimental frameworks to systems such as protein-rich cereal byproducts would help extend plant protein research toward more industrially and economically relevant applications.

Appendix A - Supplementary data

Supplementary data for Chapter 1.

Appendix Table A.1. Representative applications of all-atom and coarse-grained molecular dynamics simulations in food protein systems.

Topics	System	Timescale	Tools (engine/ force field)	Key findings	Reference
Solvent- and co-solute-induced protein responses	Wheat α -gliadin with varying ethanol concentration	1 μ s	GROMACS 2023.3/ Amber99sb-ILDN (AA)	Increased ethanol concentration led to global expansion of the protein	Kumar and Li, 2025
	Five soy 2S albumin and maize zein in water and ethanol solution	1 – 2 μ s	GROMACS 2021.5/ Martini 3 (CG)	The two proteins displayed distinct aggregation pattern depending on solvent	Kwon et al., 2025
	β -lactoglobulin dimer docked with three different co-solutes	500 ns	AMBER 22/ Amber14sb (AA)	The chemical nature of the ligands led to different interaction mechanics	Song et al., 2025

Process-induced structural evolution	Multiple zein monomers with PEG and ECGC	1.5 μ s	GROMACS 2018/ Martini (CG, version unspecified)	Zein, PEG, ECGG undergo spontaneous aggregation. The aggregate remained stable at oil-water interface	Wang et al., 2025
	Pork myosin docked with κ -carrageenan monomer under 0.3 and 0.6M NaCl	100 ns	GROMACS 5.1.4/ GROMOS96 (AA)	Myosin at 0.6M showed higher binding energy, higher NaCl concentration restricted the structural flexibility	Cao et al., 2023
	Egg ovalbumin under varying high hydrostatic pressure	500 ns	GROMACS 2023.3/ CHARMM (AA, version unspecified)	Increased pressure led to global compression of the protein, pressure-sensitive regions identified	Yang et al., 2024
	Walnut protein under ultrasound treatment	200 ns (AA), 1 μ s (CG)	GROMACS 2022.5/ Amber14sb (AA), Martini 3 (CG)	The frequency with maximal structural perturbation identified. More compact aggregate observed under the frequency	Zhang et al., 2025

	96 22kda zein in water with under 300K and 500K	5 μ s	DESMOND/ Martini 2 (CG)	Heating promotes chain extension and alignment, which can withstand higher pulling stress than unheated aggregates	Devnani et al, 2024
	Ten solvated proteins with different charges and their binary mixtures under heating	1 μ s	GROMACS 2018.7/ MA RTINI 1 (CG)	Oppositely charged proteins are more prone to aggregation compared to similarly charged solution	Nnyigide and Hyun, 2023
	Solvated transglutaminase under different oscillating electric fields	100 ns	GROMACS 2018.8/ GROMOS96 (AA)	Significantly more unstable structure was observed for the protein under 10^9 V/m, lacking orderly structures	Tao et al., 2024
	Peanut arachin docked with resveratrol	300 ns	GROMACS 2023.2/ Amber14sb-parmbc1 (AA)	Resveratrol binds within a hydrophobic cavity of arachin primarily through van der Waals interactions	Yang et al., 2025
Bioactive ligand interactions and redistribution	β -casein docked with ferulic acid and epicatechin	310 – 380 ns	GROMACS 2018.4/ CHARMM (AA, version unspecified)	Umbrella sampling revealed stronger binding for epicatechin than for ferulic acid	Condict et al., 2024

Bovine serum albumin with black rice anthocyanin	100 ns (AA), 1 μ s (CG)	GROMACS 2022.5/ Amber14sb (AA), Martini 3 (CG)	Anthocyanins associate on the protein surface. Ethanol addition can promote encapsulation	Zhao et al., 2025
κ -and β -casein aggregates with curcumin	1 μ s	NAMD (version unspecified)/ Martini (CG, version unspecified)	Spontaneous retention of curcumin was not observed but redistribution on the surface of the aggregates	Gao et al., 2025
Pea legumin dimer docked with 12 EC GC	250 ns	GROMACS 2019.2/ CH ARMM36m	ECGC interacts primarily with nonpolar side chains reducing the accessible surface area of legumin	Chen et al., 2021

Supplementary data for Chapter 2.

Appendix Table A.2. Source organisms of seed storage dataset and their numbers by each Osborne class.

Class	Source organism	Number of organisms	Number of data
Albumin	Mouse-ear cress (9), Garden pea (7), Rapeseed (4), Buckwheat (2), Soybean (2), Mabinlang (2), Winter squash (2), Brazil nut (1), Castor bean (1) ..etc	28	50
Globulin	Rapeseed (7), Mouse-ear cress (5), Buckwheat (3), English walnut (3), Almond (2), Oat (2), Pecan (2), Pistachio (2), Cashew (1) ..etc	29	50
Prolamin	Wheat (20), Maize (6), Rice (<i>Japonica</i>) (6), Barley (5), Oat (3), Sorghum (3), Rice (<i>Indica</i>) (1), Wild oat (1), Wild rye (1), Goatgrass (1) ...etc	13	50
Glutelin	Rice (<i>Japonica</i>) (10), Rice (<i>Indica</i>) (5), Sedge (3), Coyote tobacco (2), Wild rice (2), Oil palm (2), Pineapple (2), Tree cotton (2) ...etc	28	50

Appendix Table A.3. Effect of feature selection methods on the performance of MLP and benchmark models.

	(R ² / RMSE)					
	LR	SGD	SVR	RF	GB	MLP
Pearson correlation*	0.358 / 0.262	0.357 / 0.262	0.407 / 0.251	0.421 / 0.249	0.417 / 0.249	0.426 / 0.247
F-regression**	0.331 / 0.267	0.322 / 0.269	0.375 / 0.258	0.357 / 0.262	0.377 / 0.258	0.389 / 0.255
Recursive feature elimination	0.358 / 0.262	0.347 / 0.264	0.418 / 0.249	0.322 / 0.269	0.409 / 0.251	0.436 / 0.245
Without feature selection	0.368 / 0.260	0.365 / 0.260	0.443 / 0.244	0.436 / 0.245	0.441 / 0.244	0.469 / 0.238

LR: linear regression, SGD: Stochastic gradient descent, SVR: Support vector regressor,

RF: Random forest, GB: Gradient boosting, MLP: multi-layer perceptron

*Threshold of $r > 0.8$ was used to detect colinearity

**Top 10 descriptors were selected

Appendix Table A.4. Selected hyperparameters for the MLP (a), GCN (b), and stacking (SVR) meta-regressor (c).

(a)

MLP	
Epochs	25
Number of hidden layers	3
Number of nodes	128, 64, 32
Dropout rate	0.2
Activation function	LeakyReLU, sigmoid
Learning rate	0.001

(b)

GCN	
Epochs	10
Number of GCN layers	2
Number of nodes	256, 64
Activation function	LeakyReLU, sigmoid
Attention dimension	16
Attention heads	4
Learning rate	0.0001

(c)

Meta-regressor (SVR)	
Kernal	Rbf
Gamma	0.10
C	0.10

Appendix Table A.5. Comparison of extracted MLP features across the datasets used.

Features	<i>E.coli</i> dataset	<i>S. cerevisiae</i> dataset	Seed storage dataset
Molecular weight	34.727 ±20.571	37.528 ±15.414	37.357 ±17.589
Aromaticity	0.078 ±0.028	0.091 ±0.022	0.078 ±0.022
Instability index	38.011 ±10.269	36.929 ±7.648	61.186 ±24.673
Hydrophobic index	-0.146 ±0.330	-0.339 ±0.199	-0.466 ±0.395
Alipathic index	289.696 ±170.708	286.942 ±119.826	251.741 ±119.952
Absolute charge per residue	0.047 ±0.028	0.051 ±0.028	0.031 ±0.017
Hydrophilic index	-10.114 ±54.526	19.798 ±38.635	-5.198 ±89.961
hydrophobic patch area	9965.75 ±5187.194	10076.415 ±4054.545	14183.421 ±5478.721
hydrophilic area	6534.736 ±3556.231	7079.267 ±2983.201	10049.422 ±4915.376
hydrophobic area ratio	0.606 ±0.047	0.588 ±0.021	0.594 ±0.053
mean lipophilicity potential	-2.562 ±1.411	-3.161 ±0.548	-2.606 ±0.985
mean coulombic potential	-0.656 ±1.671	-1.078 ±1.341	0.045 ±0.972
surface area	13533.002 ±7229.448	14095.648 ±5678.104	18981.785 ±7974.796

Appendix Table A.5. (continued)

Features	<i>E.coli</i> dataset	<i>S. cerevisiae</i> dataset	Seed storage dataset
Solvent accessible surface area	16151.322 ±8318.966	16763.76 ±6839.81	23930.924 ±10064.226
volume	45215.954 ±26977.267	48733.426 ±20163.582	48416.050 ±22996.647
hbonds	277.481 ±175.499	292.676 ±129.661	239.500 ±134.517
helix propensity	0.432 ±0.181	0.378 ±0.128	0.347 ±0.195
strand propensity	0.177 ±0.117	0.183 ±0.097	0.160 ±0.147
coil propensity	0.392 ±0.103	0.439 ±0.074	0.493 ±0.101
favorable contacts	11834.805 ±7483.192	12999.639 ±5695.359	9646.635 ±5565.607

Appendix Table A.6. Comparison of amino acids composition across the datasets used.

Amino acid	<i>E.coli</i> dataset	<i>S. cerevisiae</i> dataset	Seed storage dataset
Ala	9.389±2.723	6.095±2.232	6.810±2.447
Cys	1.302±1.200	1.458±0.936	3.067±2.568
Asp	5.395±1.756	6.541±1.746	3.373±2.153
Glu	6.232±2.249	6.953±1.711	5.551±3.484
Phe	3.638±1.605	4.391±1.473	4.445±1.520
Gly	7.053±2.293	6.460±2.133	6.491±2.741
His	2.398±1.290	2.299±0.976	1.854±1.101
Ile	5.882±2.006	6.649±1.634	4.569±1.343
Lys	4.787±2.283	7.437±2.039	3.160±2.440
Leu	10.353±2.850	9.034±1.965	8.249±2.076
Met	2.830±1.177	2.019±0.872	2.317±2.486
Asn	3.920±1.630	4.947±1.625	4.172±2.073
Pro	4.365±1.647	4.661±1.476	6.641±3.815
Gln	4.625±1.897	3.824±1.685	12.984±9.997
Arg	5.714±2.163	3.923±1.462	5.932±4.109
Ser	5.601±1.845	6.738±1.716	6.802±2.165
Thr	5.325±1.774	5.241±1.395	3.942±1.461
Val	7.012±2.062	6.621±1.589	6.279±1.945
Trp	1.415±1.109	1.133±0.668	0.828±0.508
Tyr	2.764±1.420	3.575±1.293	2.533±1.261

Appendix Table A.7. Comparison of selected seed proteins to reference human proteins. Means with different letters indicate significant differences within each row ($p < 0.05$).

	Albumin	Globulin	Prolamin	Glutelin	Seed protein average	Keratin	Hemo globin	Amylase
Hydrophobic index	-0.519 ± 0.042 b	-0.933 ± 0.156 c	0.344 ± 0.035 a	-0.498 ± 0.039 b	-0.402 ± 0.482 b	-0.409 ± 0.095 b	0.029 ± 0.054 a	-0.424 ± 0.017 b
Hydrophobic area ratio	0.618 ± 0.010 bcd	0.544 ± 0.012 e	0.707 ± 0.008 a	0.572 ± 0.008 de	0.610 ± 0.064 bcd	0.651 ± 0.010 bc	0.662 ± 0.012 ab	0.604 ± 0.004 cd
Absolute charge per residue	0.029 ± 0.012 bc	0.059 ± 0.021 a	0.023 ± 0.006 bc	0.011 ± 0.004 c	0.031 ± 0.021 bc	0.043 ± 0.011 ab	0.023 ± 0.008 bc	0.035 ± 0.001 b
Hydrophilic index	0.055 ± 0.060 b	0.394 ± 0.135 a	-0.656 ± 0.021 c	-0.056 ± 0.025 b	-0.066 ± 0.395 b	0.204 ± 0.041 ab	-0.142 ± 0.009 b	-0.110 ± 0.010 b
MW	19035.239 ± 237.739 d	50751.307 ± 10409.652 ab	28213.106 ± 1127.976 cd	56055.149 ± 725.901 a	38513.700 ± 16472.844 bc	54748.592 ± 2508.758 a	15830.058 ± 381.813 d	57743.355 ± 32.604 a
mean lipophilic potential	-2.141 ± 0.219 b	-3.444 ± 0.349 c	-0.035 ± 0.272 a	-2.783 ± 0.173 bc	-2.101 ± 1.334 b	-1.914 ± 0.281 b	-1.898 ± 0.329 b	-3.760 ± 0.012 c
mean coulombic potential	0.276 ± 0.553 b	-1.380 ± 0.605 c	0.262 ± 0.194 b	1.392 ± 0.202 a	0.138 ± 1.090 b	-0.634 ± 0.391 bc	0.396 ± 0.366 b	-0.110 ± 0.120 b
accessible surface area	13710.800 ± 474.593 ef	32043.600 ± 2827.388 b	22959.00 ± 1016.841 d	30356.600 ± 659.254 bc	24767.500 ± 7564.657 cd	55369.8 ± 4126.348 a	7725.980 ± 102.691 f	19243.200 ± 124.570 de
helix propensity	0.591 ± 0.012 bc	0.245 ± 0.079 de	0.739 ± 0.041 ab	0.193 ± 0.009 e	0.442 ± 0.239 cd	0.609 ± 0.027 abc	0.804 ± 0.031 a	0.289 ± 0.008 de
strand propensity	0.000 ± 0.000 c	0.229 ± 0.046 ab	0.000 ± 0.000 c	0.275 ± 0.008 a	0.126 ± 0.132 b	0.010 ± 0.022 c	0.000 ± 0.000 c	0.127 ± 0.000 b
coil propensity	0.409 ± 0.012 c	0.526 ± 0.049 ab	0.261 ± 0.041 d	0.532 ± 0.017 a	0.432 ± 0.117 bc	0.381 ± 0.039 c	0.196 ± 0.031 d	0.584 ± 0.008 a

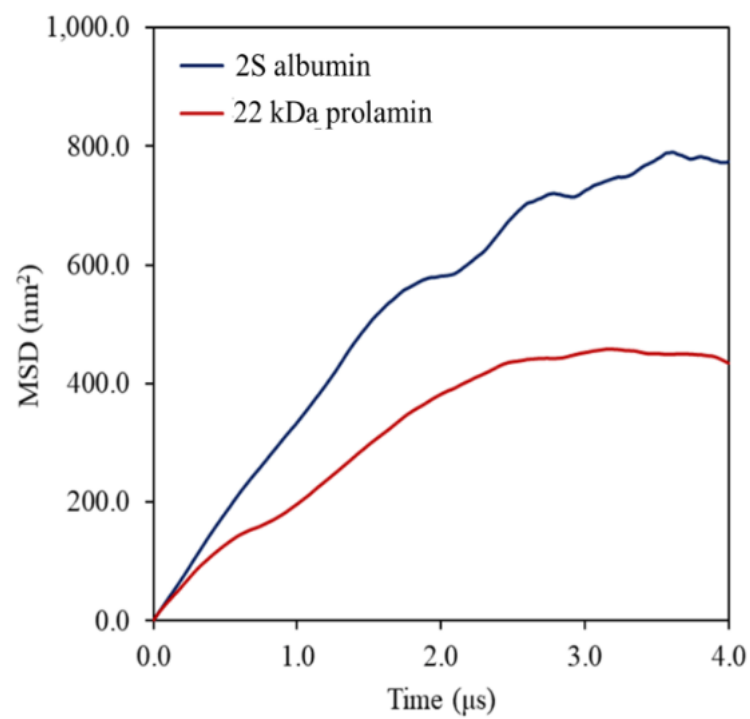
Appendix Table A.7. (continued)

	Albumin	Globulin	Prolamin	Glutelin	Seed protein average	Keratin	Hemo globin	Amylase
Ala	6.639 ±0.647cd	6.288 ±0.952cd	14.721 ±0.567a	7.068 ±0.484cd	8.679 ±3.645bc	7.475 ±0.728cd	11.606 ±1.883b	5.362 ±0.107d
Arg	5.789 ±0.667b	8.937 ±0.515a	1.081 ±0.319c	7.348 ±0.614ab	5.789 ±3.055b	6.822 ±0.655ab	2.492 ±1.082c	5.479 ±0.000b
Asn	3.381 ±0.568f	6.543 ±0.884bc	5.003 ±0.276de	7.021 ±0.597ab	5.487 ±1.57cd	4.323 ±0.211def	3.561 ±1.869ef	8.141 ±0.175a
Asp	4.945 ±1.164b	3.826 ±0.750bc	0.160 ±0.219d	2.789 ±0.650c	2.930 ±1.951c	3.929 ±0.753bc	4.839 ±0.914b	6.771 ±0.107a
Cys	5.790 ±0.534a	0.895 ±0.555c	0.553 ±0.390c	1.694 ±0.292c	2.233 ±2.190bc	3.828 ±2.562b	0.962 ±0.364c	2.309 ±0.088bc
Gln	13.272 ±0.498b	6.645 ±0.391c	19.102 ±0.847a	10.451 ±0.563b	12.368 ±4.694b	5.216 ±0.925cd	2.060 ±0.954d	2.505 ±0.214d
Glu	6.155 ±0.696bc	12.514 ±2.935a	0.538 ±0.202d	5.569 ±0.504bc	6.194 ±4.584bc	9.443 ±1.282ab	4.674 ±1.260c	3.836 ±0.107cd
Gly	4.697 ±1.476c	6.956 ±0.457bc	1.229 ±0.488d	6.984 ±0.338bc	4.966 ±2.527c	9.152 ±1.847ab	6.593 ±2.164c	10.098 ±0.107a
His	1.567 ±0.325cde	2.901 ±0.837b	0.915 ±0.320e	2.183 ±0.354bc	1.892 ±0.887cd	1.070 ±0.276de	5.524 ±1.022a	2.309 ±0.088bc
Ile	3.855 ±0.984b	4.163 ±0.235b	4.244 ±0.599b	5.134 ±0.64ab	4.349 ±0.785ab	4.169 ±0.558b	1.666 ±2.345c	5.519 ±0.088a
Leu	8.204 ±0.559bc	8.010 ±0.803bc	18.838 ±0.870a	7.874 ±0.701bc	10.732 ±4.851b	9.594 ±0.923b	12.004 ±0.675b	5.362 ±0.107c
Lys	5.431 ±1.148b	4.479 ±1.202bcd	0.309 ±0.174e	2.911 ±0.496d	3.283 ±2.143cd	5.229 ±0.51bc	7.715 ±1.149a	4.618 ±0.107bcd
Met	2.773 ±0.684a	0.425 ±0.499d	1.685 ±0.468bc	0.967 ±0.355cd	1.463 ±1.018bc	1.852 ±0.803abc	1.929 ±0.563abc	2.114 ±0.088ab
Phe	5.429 ±0.852ab	3.399 ±0.940cd	3.960 ±0.839c	5.209 ±0.523ab	4.499 ±1.140bc	2.593 ±0.262d	5.373 ±0.491ab	5.753 ±0.107a
Pro	6.508 ±1.336b	4.374 ±0.884c	8.494 ±0.383a	4.400 ±0.237c	5.944 ±1.912bc	2.229 ±1.009d	4.275 ±0.572c	4.384 ±0.107c
Ser	3.737 ±0.888c	7.168 ±1.380b	7.107 ±0.528b	7.140 ±1.360b	6.288 ±1.816b	9.060 ±1.211a	6.101 ±2.419b	6.380 ±0.107b
Thr	3.254 ±0.893c	3.323 ±0.624c	3.466 ±0.311c	3.835 ±0.396bc	3.470 ±0.597c	4.693 ±1.312ab	5.407 ±2.020a	4.423 ±0.107abc
Trp	0.484 ±0.505cd	0.583 ±0.010c	0.077 ±0.171d	0.727 ±0.116c	0.468 ±0.353cd	0.319 ±0.106cd	1.375 ±0.473b	3.640 ±0.107a
Tyr	2.534 ±0.796bc	2.484 ±0.402bc	2.939 ±0.307b	3.713 ±0.147a	2.918 ±0.668b	2.725 ±0.200b	1.934 ±0.322c	4.070 ±0.088a
Val	5.556 ±1.194b	6.084 ±1.313b	5.580 ±1.53b	6.983 ±0.434b	6.051 ±1.244b	6.280 ±0.972b	9.911 ±1.909a	6.928 ±0.107b

Appendix Table A.8. Linear correlation coefficient of MLP features with experimental solubility

Feature	Correlation coefficient (r)
Molecular weight	-0.327
Aromaticity	-0.284
Instability index	-0.042
Hydrophobic index	-0.202
Aliphatic index	-0.335
Absolute charge per residue	0.357
Hydrophilic index	0.361
Hydrophobic patch area	-0.321
Hydrophilic surface area	-0.248
Hydrophobic area ratio	-0.199
Mean lipophilicity potential	-0.230
Mean coulombic potential	-0.205
Surface area	-0.306
Solvent-accessible surface area	-0.291
Volume	-0.328
Hydrogen bonding	-0.327
Helix propensity	-0.121
Strand propensity	0.097
Coil propensity	0.103
Favorable contacts	-0.326

Appendix Figure A.1. Mean square displacement of the selected proteins over the simulation time



Supplementary data for Chapter 3.

Appendix Table A.9. Comparison of 38 extracted physicochemical features across Osborne classes. Values are presented as mean $\pm$ std and median, interquartile range.

	Aromaticity (a.u.)	Instability Index (a.u.)	Hydrophobic index (a.u.)	Turn forming residue fraction (%)	Absolute charge per residue	Hydrophilic index (a.u.)	Predicted PI	MW (kDa)	rA (%)	rR (%)
Albumin	6.447 ± 2.241 / (6.708, 2.319c)	63.828 ± 13.422 / (63.887, 17.153b)	-0.933 ± 0.214 / (-0.901, 0.264d)	20.589 ± 2.698 c/ (20.779, 4.007)	1.680 ± 2.324 / (0.817, 1.789b)	0.232 ± 0.173 / (0.203, 0.253a)	6.480 ± 1.100 / (6.391, 1.116c)	17.337 ± 4.602 / (16.783, 3.176c)	4.292 ± 1.615 / (4.452, 1.799c)	7.719 ± 3.383 / (6.711, 4.464a)
Globulin	7.433 ± 0.979 / (7.425, 1.126b)	56.556 ± 9.883 / (55.140, 14.991c)	-0.802 ± 0.278 / (-0.761, 0.394c)	25.941 ± 1.749 a/ (25.986, 2.395)	1.592 ± 1.337 / (1.266, 2.133b)	0.224 ± 0.223 / (0.227, 0.350a)	6.346 ± 1.020 / (6.067, 1.402c)	54.585 ± 6.223 / (53.382, 6.880a)	5.760 ± 1.392 / (5.943, 2.180b)	8.031 ± 2.529 / (7.551, 3.983a)
Prolamin	8.026 ± 2.030 / (7.377, 2.369b)	81.297 ± 19.892 / (74.156, 33.058a)	-0.091 ± 0.502 / (0.186, 0.993a)	23.581 ± 2.687 b/ (23.113, 4.413)	0.656 ± 0.490 / (0.635, 0.427c)	-0.556 ± 0.162 / (-0.635, 0.318c)	7.933 ± 0.846 / (8.089, 0.825b)	25.023 ± 6.015 / (26.586, 7.889b)	11.024 ± 5.567 / (13.609, 11.720a)	1.460 ± 0.836 / (1.361, 0.808b)
Glutelin	9.098 ± 0.916 / (8.956, 1.263a)	50.787 ± 7.102 / (50.931, 5.831d)	-0.561 ± 0.110 / (-0.587, 0.094b)	25.674 ± 1.317 a/ (25.942, 1.470)	1.678 ± 0.823 / (1.684, 0.834a)	-0.004 ± 0.077 / (-0.015, 0.048b)	8.674 ± 1.040 / (9.090, 0.546a)	53.376 ± 3.583 / (53.688, 1.582a)	6.746 ± 0.969 / (6.737, 1.041a)	7.504 ± 1.293 / (7.361, 1.089a)

Appendix Table A.9. (continued)

	rN (%)	rD (%)	rC (%)	rQ (%)	rE (%)	rG (%)	rH (%)	rI (%)	rL (%)	rK (%)
Albumin	2.981 ±1.087/ (2.667, 1.301c)	4.689 ±1.531/ (4.576, 1.341a)	5.838 ±1.236/ (5.594, 1.627a)	15.036 ±4.412/ (15.543, 4.054b)	7.913 ±2.836/ (7.000, 3.735a)	5.787 ±1.547/ (5.863, 1.779b)	2.166 ±0.953/ (2.182, 1.138b)	3.818 ±1.318/ (3.497, 1.609c)	6.000 ±1.609/ (5.667, 1.386c)	4.192 ±2.251/ (4.505, 3.388a)
Globulin	6.591 ±1.122/ (6.533, 1.538a)	4.327 ±0.866/ (4.329, 1.091a)	0.945 ±0.608/ (1.040, 0.558c)	9.139 ±2.970/ (8.571, 4.032d)	8.987 ±3.290/ (8.924, 4.356a)	7.563 ±1.601/ (7.460, 2.119a)	2.163 ±0.684/ (2.016, 0.603b)	4.834 ±0.890/ (4.762, 1.228b)	7.686 ±1.385/ (7.516, 1.862b)	3.864 ±2.030/ (3.178, 2.713a)
Prolamin	4.271 ±1.736/ (4.695, 2.924b)	0.398 ±0.533/ (0.357, 0.649c)	1.408 ±1.213/ (0.939, 2.084c)	23.871 ±6.821/ (21.361, 12.133a)	0.735 ±0.677/ (0.670, 0.410c)	1.886 ±1.019/ (1.653, 1.145c)	1.354 ±0.816/ (1.224, 0.813c)	4.746 ±0.933/ (4.694, 1.038b)	13.911 ±5.106/ (16.000, 10.036a)	0.246 ±0.473/ (0.000, 0.404b)
Glutelin	6.774 ±1.399/ (6.947, 1.499a)	3.243 ±0.825/ (3.158, 1.151b)	1.297 ±0.337/ (1.261, 0.602b)	10.179 ±1.724/ (10.774, 1.602c)	5.945 ±1.051/ (5.732, 0.684b)	7.332 ±0.735/ (7.158, 0.704a)	2.376 ±0.746/ (2.311, 0.815a)	5.189 ±0.863/ (4.873, 0.882a)	7.431 ±0.729/ (7.392, 0.953b)	3.078 ±0.598/ (3.104, 0.728a)

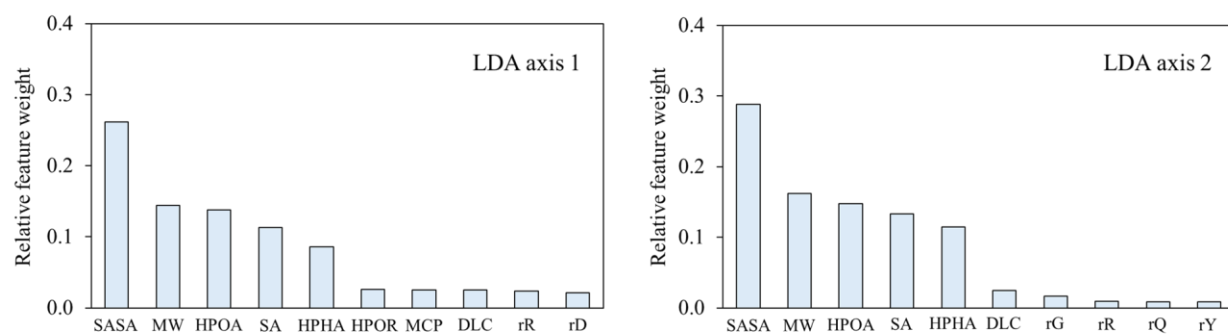
Appendix Table A.9. (continued)

	rM (%)	rF (%)	rP (%)	rS (%)	rT (%)	rW (%)	rY (%)	rV (%)	Mean Coulombic potential (kcal mol ⁻¹ e ⁻¹)	Solvent- accessible surface area (nm ²)
Albumin	2.830 ±2.015/ (2.400, 1.344a)	2.913 ±1.798/ (3.145, 2.250c)	6.271 ±3.067/ (7.641, 5.403b)	5.549 ±2.342/ (5.053, 2.869c)	3.776 ±1.434/ (4.013, 1.860a)	1.012 ±0.768/ (0.840, 0.612a)	2.523 ±1.193/ (2.210, 1.235c)	4.693 ±1.864/ (5.096, 2.939c)	-0.143 ±1.292/ (-0.165, 1.148c)	10.779 ±2.981/ (10.283, 2.403c)
Globulin	1.033 ±0.911/ (0.820, 1.615b)	4.034 ±0.970/ (4.107, 1.493b)	5.067 ±0.841/ (4.928, 1.360c)	6.721 ±1.236/ (6.652, 1.941a)	3.713 ±0.929/ (3.744, 1.233a)	0.847 ±0.431/ (0.862, 0.488a)	2.552 ±0.656/ (2.434, 0.797c)	6.141 ±1.075/ (6.118, 1.646b)	-0.776 ±0.874/ (-0.740, 1.325d)	31.755 ±7.302/ (29.917, 8.133a)
Prolamin	1.216 ±1.371/ (1.036, 1.183b)	4.737 ±1.371/ (4.761, 1.906a)	11.198 ±3.809/ (9.678, 6.512a)	6.226 ±1.381/ (6.127, 1.676b)	2.830 ±1.447/ (2.559, 1.345b)	0.446 ±0.543/ (0.376, 0.562b)	2.844 ±1.365/ (3.077, 1.632b)	5.193 ±1.647/ (5.263, 2.057c)	0.162 ±0.227/ (0.150, 0.240b)	23.374 ±6.420/ (23.739, ,9.329b)
Glutelin	1.148 ±0.775/ (0.848, 0.764b)	4.836 ±0.684/ (4.904, 1.109a)	4.553 ±0.450/ (4.627, 0.620d)	7.015 ±1.223/ (7.368, 1.800a)	3.965 ±0.596/ (3.967, 0.689a)	0.800 ±0.222/ (0.732, 0.219a)	3.461 ±0.556/ (3.619, 0.662a)	7.126 ±0.783/ (7.219, 0.823a)	1.167 ±0.813/ (1.350, 1.015a)	29.679 ±3.698/ (29.281, 1.642a)

Appendix Table A.9. (continued)

	Helix propensity (%)	Strand propensity (%)	Coil propensity (%)	Quilt surface area (nm ²)	Hydrophobic patch area (nm ²)	Hydrophilic area (nm ²)	Hydrophobic surface area ratio (%)	Disulfide Linkage Count
Albumin	57.164 ±8.740/ (57.634, 9.578a)	0.545 ±4.022/ (0.000, 0.000b)	42.291 ±7.144/ (42.235, 9.420c)	10.961 ±3.036/ (10.498, 2.571c)	5.918 ±1.599/ (5.689, 1.654b)	5.042 ±1.547/ (4.710, 0.932c)	0.540 ±0.035/ (0.543, 0.049c)	4.196 ±1.510/ (4.000, 0.000a)
Globulin	18.116 ±5.533/ (16.913, 7.867b)	28.591 ±3.687/ (28.571, 4.233a)	53.293 ±5.586/ (54.188, 7.486b)	32.229 ±7.152/ (30.604, 7.767a)	16.941 ±3.478/ (16.043, 3.658a)	15.288 ±3.801/ (14.327, 4.492a)	0.528 ±0.022/ (0.528, 0.028d)	1.642 ±1.280/ (2.000, 1.000b)
Prolamin	57.958 ±20.466/ (68.058, 38.798a)	2.378 ±8.770/ (0.000, 0.000b)	39.664 ±17.530/ (31.942, 35.057c)	23.444 ±6.310/ (23.934, 9.111b)	15.591 ±3.576/ (16.684, 4.878a)	7.854 ±3.071/ (7.199, 5.417b)	0.674 ±0.049/ (0.695, 0.096a)	1.044 ±1.575/ (0.000, 2.000c)
Glutelin	15.521 ±2.050/ (15.543, 1.953b)	28.493 ±1.990/ (29.053, 1.706a)	55.986 ±2.808/ (55.484, 2.624a)	30.324 ±3.596/ (30.009, 1.428a)	16.760 ±2.073/ (16.453, 0.897a)	13.564 ±1.613/ (13.541, 0.669a)	0.553 ±0.014/ (0.549, 0.015b)	1.956 ±0.235/ (2.000, 0.000b)

Appendix Figure A.2. Top 10 relative contributions of input features to class separation in linear discriminant analysis (LDA).



SASA: Solvent-accessible surface area, MW: Molecular weight, HPOA: Hydrophobic patch area, SA: surface area, HPHA: Hydrophilic area, HPOR: Hydrophobic area ratio, MCP: Mean Coulombic potential, DLC: disulfide linkage count, rX: ratio of residue X in entire sequence

Appendix Table A.10. Performances of the models trained (training/cross-validation/testing).

	Accuracy	Precision	Recall	F1	AUC
SVM (Albumin/ Prolamin)	0.995/ 0.995/ 1.000	0.995/ 0.995/ 1.000	0.996/ 0.996/ 1.000	0.995/ 0.996/ 1.000	0.999/ 0.999/ 1.000
SVM (Globulin/ Glutelin)	0.987/ 0.972/ 0.960	0.986/ 0.972/ 0.954	0.984/ 0.967/ 0.954	0.985/ 0.968/ 0.954	0.999/ 0.988/ 0.993
GCN (Albumin/ Prolamin)	1.000/ 1.000/ 1.000	1.000/ 1.000/ 1.000	1.000/ 1.000/ 1.000	1.000/ 1.000/ 1.000	1.000/ 1.000/ 1.000
GCN (Globulin/ Glutelin)	0.974/ 0.987/ 0.968	0.976/ 0.984/ 0.972	0.956/ 0.970/ 0.964	0.966/ 0.977/ 0.964	0.987/ 0.992/ 0.984

Appendix Table A.11. Generalizability performances of the models tested on globulin/glutelin set and respective top 5 SHAP features selected using different train_test split seeds. The first column indicates the random seeds tested. All hyperparameters were selected via cross-validation at random seed = 98. The hyperparameters were fixed for the other 5 train_test split seeds. SHAP was not applied to GCN model. Model abbreviations: SVM (support vector machine), DT (decision tree), RF (random forest), KNN (K-nearest neighbor), MLP (multilayer perceptron).

SVM	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.960	0.954	0.984	0.954	0.993
Top features	Mean coulombic potential, predicted PI, hydrophobic index, rS, rW				
Split seed: 14	0.990	0.945	0.993	0.989	0.998
Top features	Mean coulombic potential, predicted PI, hydrophobic index, rW, rY				
Split seed: 31	0.990	0.985	0.993	0.989	0.997
Top features	Mean coulombic potential, predicted PI, hydrophobic index, rS, rW				
Split seed: 48	0.970	0.962	0.969	0.966	0.999
Top features	Mean coulombic potential, hydrophobic index, predicted PI, rS, rV				
Split seed: 65	0.980	0.977	0.977	0.977	0.994
Top features	Mean coulombic potential, predicted PI, hydrophobic index, rV, rW				
Split seed: 82	0.970	0.962	0.970	0.966	0.997
Top features	Predicted PI, mean coulombic potential, hydrophobic index, rV, rS				

Appendix Table A.11. (continued)

DT	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.940	0.947	0.914	0.928	0.914
Top features	Mean coulombic potential, rE, coil propensity, aromaticity, rN				
Split seed: 14	0.950	0.954	0.930	0.941	0.930
Top features	Mean coulombic potential, aromaticity, instability index, rP, hydrophobic area				
Split seed: 31	0.940	0.926	0.940	0.932	0.939
Top features	Mean coulombic potential, aromaticity, rP, hydrophobic area, coil propensity				
Split seed: 48	0.889	0.871	0.877	0.874	0.877
Top features	Mean coulombic potential, aromaticity, rP, instability index, rN				
Split seed: 65	0.909	0.926	0.898	0.889	0.868
Top features	Mean coulombic potential, aromaticity, rN, rE, instability index				
Split seed: 82	0.967	0.957	0.978	0.966	0.978
Top features	Mean coulombic potential, aromaticity, rG, rP, instability index				

Appendix Table A.11. (continued)

RF	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.940	0.959	0.906	0.927	0.996
Top features	Mean coulombic potential, predicted PI, aromaticity, rE, rY				
Split seed: 14	0.980	0.977	0.977	0.977	0.999
Top features	Mean coulombic potential, predicted PI, aromaticity, hydrophobic area, rY				
Split seed: 31	0.960	0.962	0.946	0.953	0.997
Top features	Mean coulombic potential, predicted PI, aromaticity, rY, hydrophobic area				
Split seed: 48	0.940	0.947	0.914	0.928	0.990
Top features	Mean coulombic potential, predicted PI, aromaticity, hydrophilic index, hydrophobic area				
Split seed: 65	0.970	0.979	0.953	0.965	0.996
Top features	Mean coulombic potential, predicted PI, aromaticity, rE, hydrophobic area				
Split seed: 82	0.990	0.985	0.993	0.989	0.999
Top features	Mean coulombic potential, predicted PI, aromaticity, hydrophobic area, hydrophilic index				

Appendix Table A.11. (continued)

MLP	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.940	0.931	0.931	0.931	0.987
Top features	Mean coulombic potential, rW, rP, Hydrophobic index, rV				
Split seed: 14	0.970	0.962	0.969	0.966	0.998
Top features	Mean coulombic potential, rW, rE, hydrophilic index, hydrophobic index				
Split seed: 31	0.980	0.977	0.977	0.977	0.997
Top features	Mean coulombic potential, rW, rE, hydrophobic area, hydrophilic index				
Split seed: 48	0.970	0.957	0.978	0.966	0.999
Top features	Mean coulombic potential, rW, rE, hydrophilic index, hydrophobic index				
Split seed: 65	0.970	0.970	0.961	0.965	0.994
Top features	Mean coulombic potential, rW, rE, hydrophobic index, predicted PI				
Split seed: 82	0.980	0.971	0.985	0.977	0.998
Top features	Mean coulombic potential, rW, rE, hydrophilic index, hydrophobic index				

Appendix Table A.11. (continued)

KNN	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.950	0.939	0.946	0.943	0.987
Top features	Predicted PI, rE, rN, mean coulombic potential, rV				
Split seed: 14	0.970	0.962	0.970	0.966	0.983
Top features	Predicted PI, rE, mean coulombic potential, hydrophilic index, rY				
Split seed: 31	0.980	0.977	0.977	0.977	0.991
Top features	Predicted PI, rE, mean coulombic potential, hydrophilic index, rY				
Split seed: 48	0.960	0.962	0.950	0.953	0.990
Top features	Predicted PI, rE, mean coulombic potential, hydrophilic index, rY				
Split seed: 65	0.970	0.970	0.961	0.965	0.990
Top features	Predicted PI, rE, mean coulombic potential, rN, hydrophilic index				
Split seed: 82	0.980	0.971	0.985	0.977	0.998
Top features	Predicted PI, rE, mean coulombic potential, rV, hydrophilic index				

Appendix Table A.11. (continued)

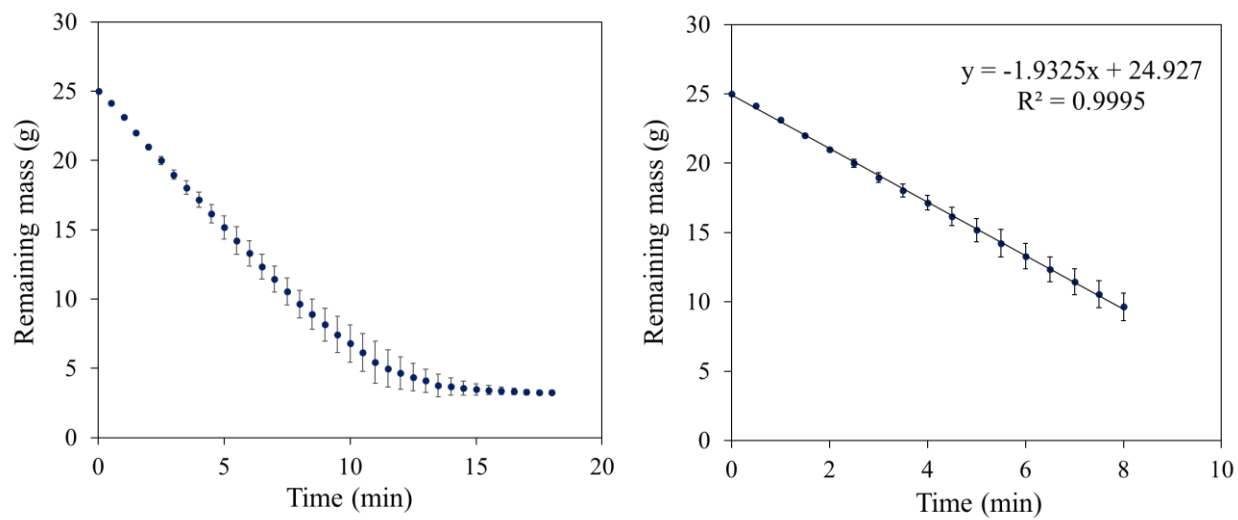
GCN	Accuracy	Precision	Recall	F1	AUC
Split seed: 98	0.968	0.972	0.964	0.964	0.984
Split seed: 14	0.970	0.970	0.970	0.970	0.990
Split seed: 31	0.980	0.982	0.980	0.981	0.992
Split seed: 48	0.956	0.964	0.960	0.962	0.979
Split seed: 65	0.980	0.971	0.985	0.977	0.998
Split seed: 82	0.970	0.962	0.969	0.966	0.998

Appendix Table A.12. Average generalizability performances of the models tested on globulin/glutelin set across 6 different train_test split random seeds.

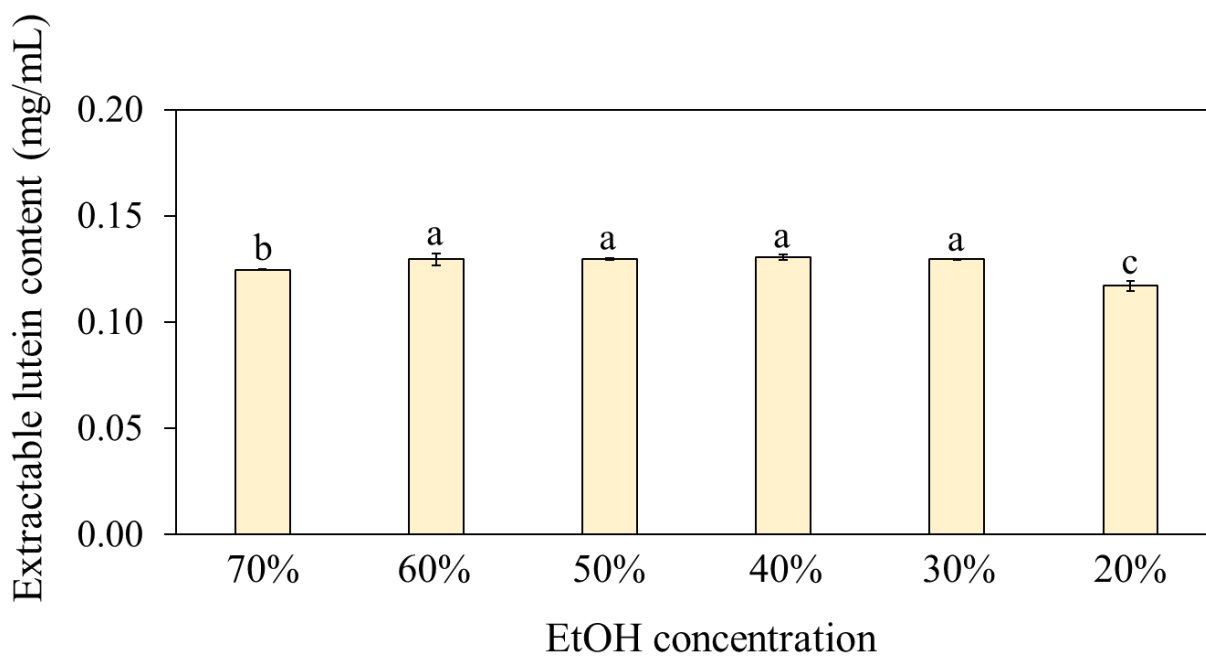
Model	Accuracy	Precision	Recall	F1	AUC	Jaccard similarity
SVM	0.977±0.012	0.964±0.015	0.981±0.011	0.974±0.014	0.996±0.002	0.733
DT	0.933±0.028	0.93±0.032	0.923±0.035	0.922±0.034	0.918±0.041	0.488
RF	0.963±0.021	0.968±0.014	0.948±0.034	0.957±0.025	0.996±0.003	0.679
KNN	0.968±0.012	0.964±0.013	0.965±0.015	0.964±0.013	0.990±0.005	0.686
MLP	0.968±0.015	0.961±0.016	0.967±0.019	0.964±0.017	0.996±0.005	0.626
GCN	0.971±0.009	0.970±0.007	0.971±0.010	0.970±0.008	0.990±0.008	-

Supplementary data for Chapter 4.

Appendix Figure A.3. Gravimetric monitoring of solvent evaporation during zein–lutein suspension preparation. Total mass change over time during rotary evaporation (left). Linear relationship between mass loss and time within the 70–20% ethanol range (right).

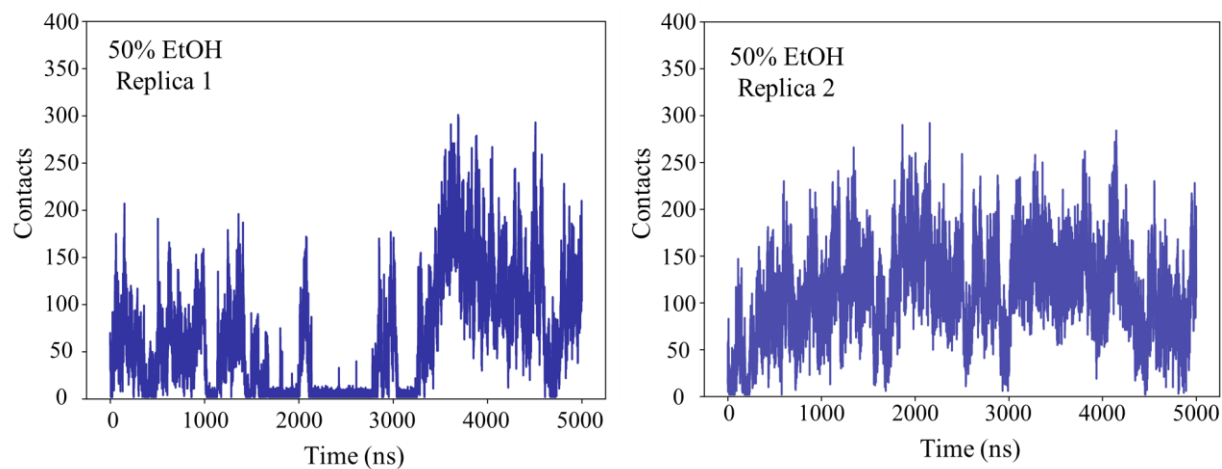


Appendix Figure A.4. Hexane-extractable lutein in ethanol solutions without zein across different ethanol concentrations.

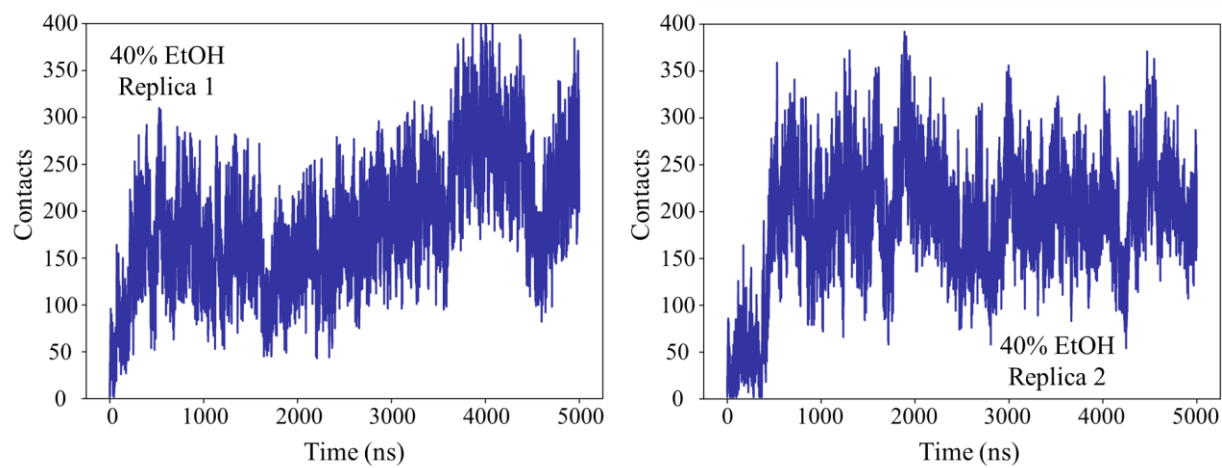


Appendix Figure A.5. Time evolution of zein–lutein contacts in coarse-grained simulations at 50% (a) and 40% ethanol (b).

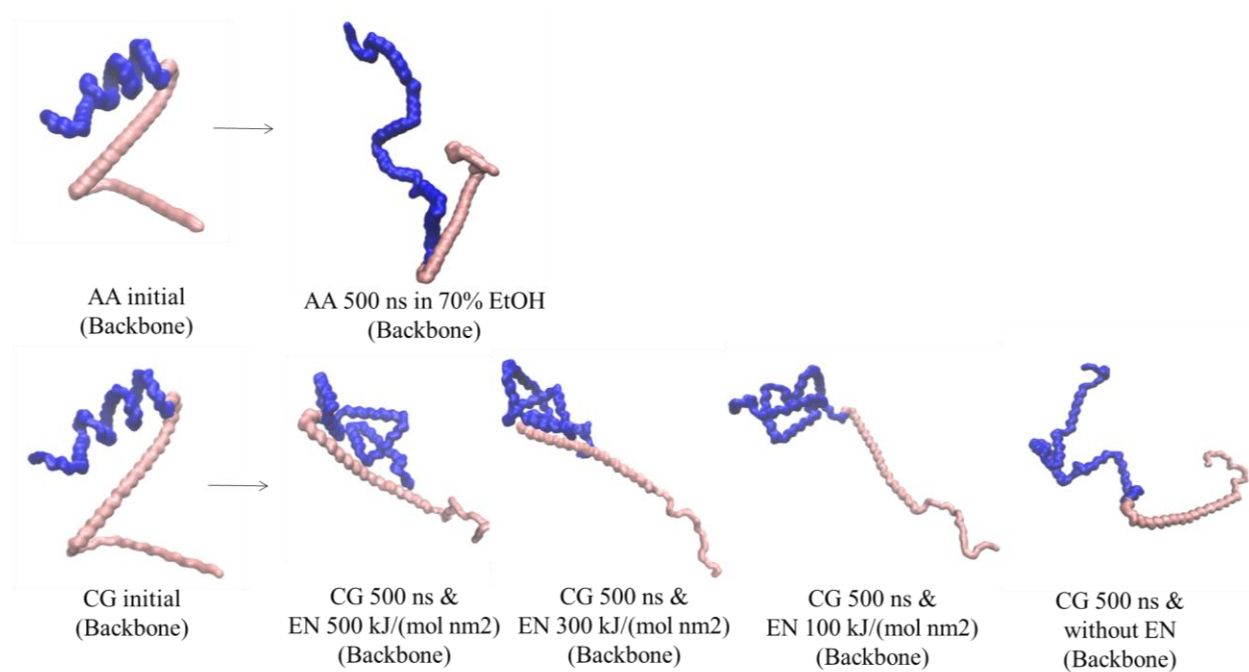
(a)



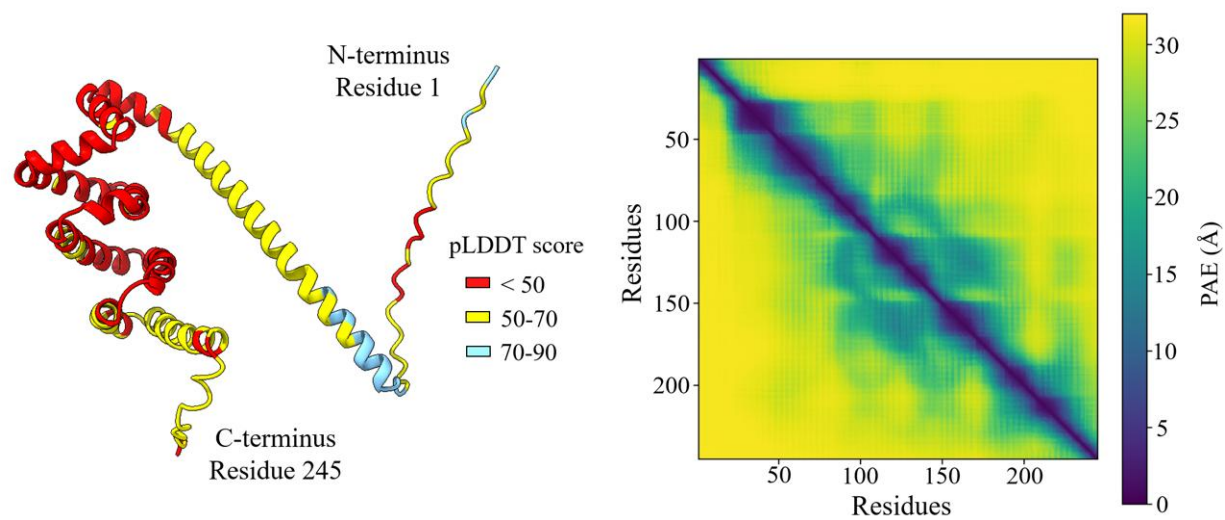
(b)



Appendix Figure A.6. Comparison of zein backbone conformations between all-atom and coarse-grained simulations with varying elastic network (EN) strength. The helical region where pronounced expansion was observed (residues 100–245) is colored in blue.

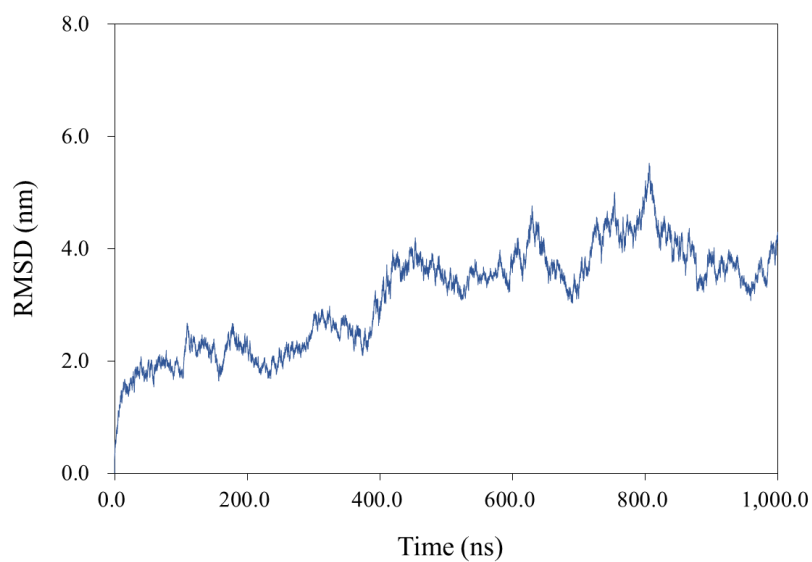


Appendix Figure A.7. AlphaFold 3-predicted α -zein monomer structure colored by pLDDT score (left) and the corresponding PAE matrix (right).

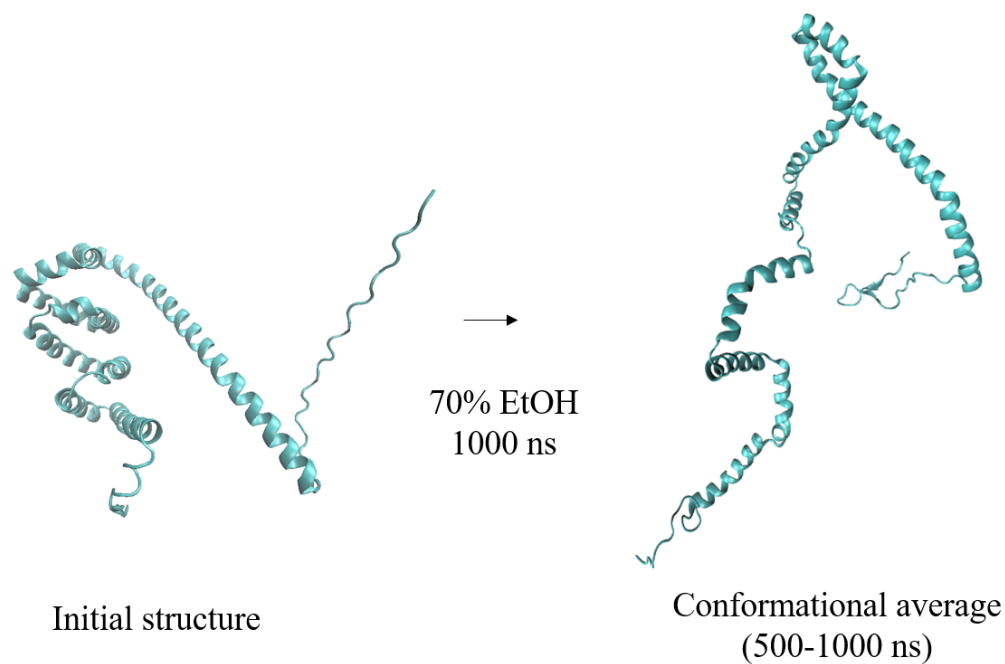


Appendix Figure A.8. Conformational behavior of zein in 70% ethanol from all-atom molecular dynamics simulations. RMSD of zein over the 1 μ s simulation (a). Representative structure obtained from RMSD-based clustering of equilibrated frames (500–1000 ns) (b).

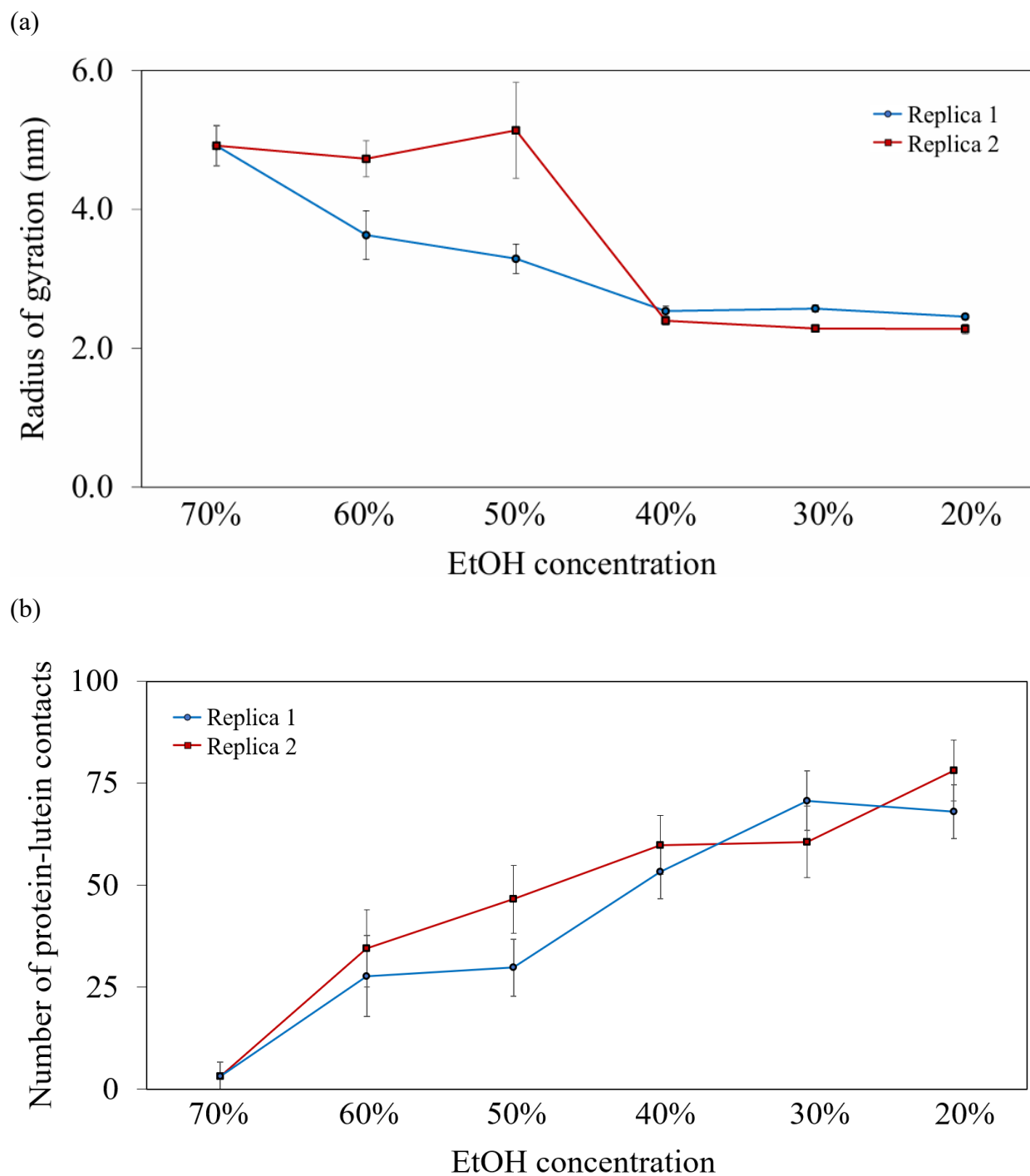
(a)



(b)

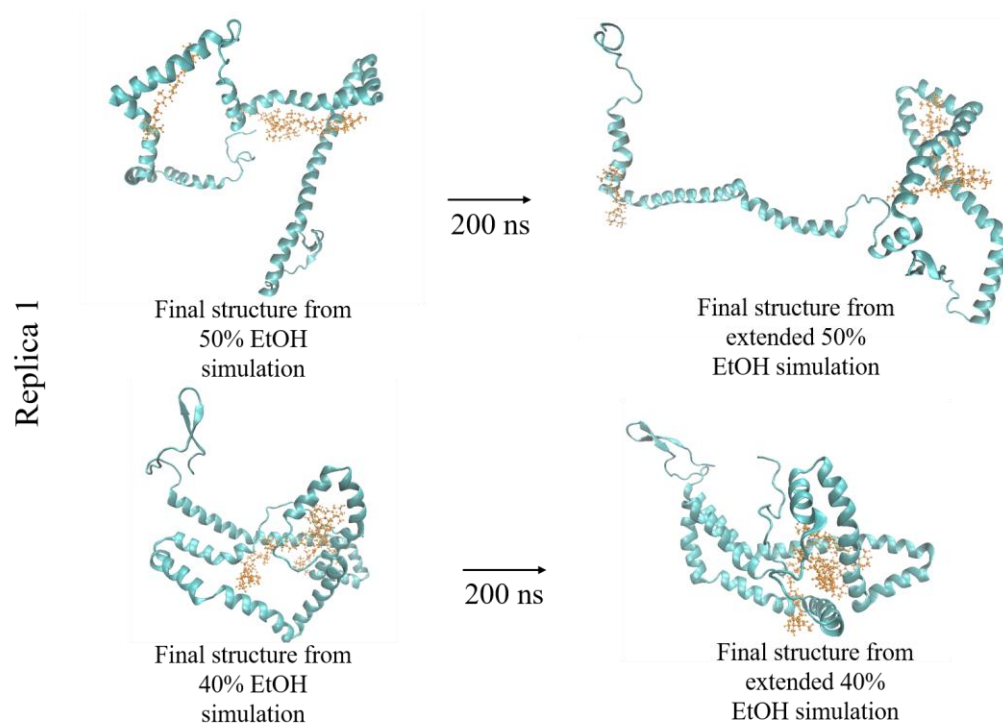


Appendix Figure A.9. Average radius of gyration (a) and protein–lutein contacts (b) during the last 50 ns of all-atom MD simulations at different ethanol concentrations, shown separately for each replica.

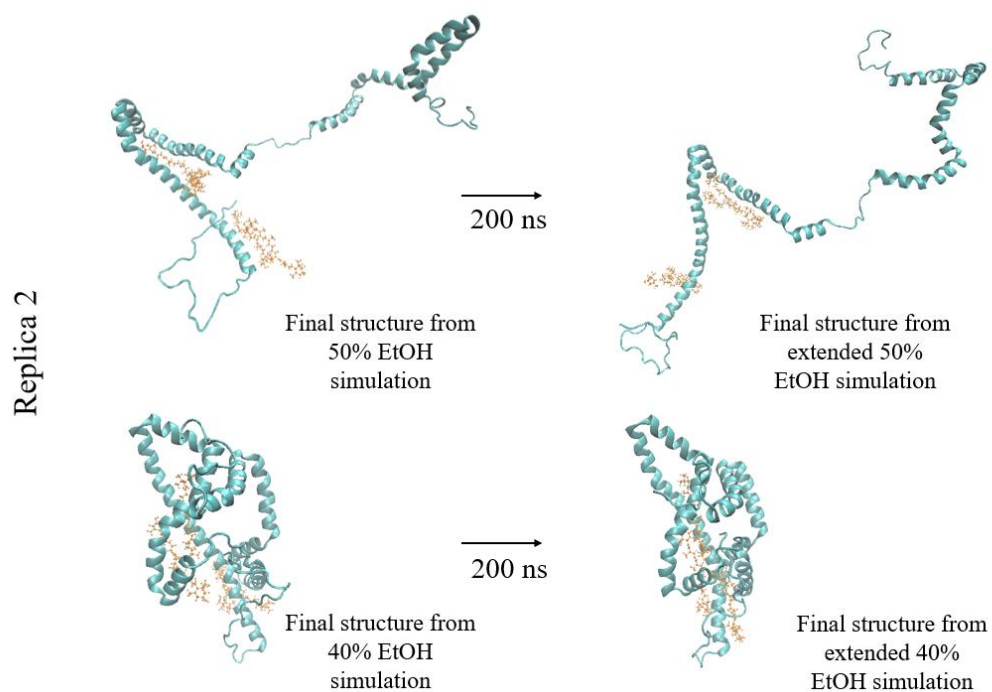


Appendix Figure A.10. Final conformations of the α -zein–lutein system at 50% and 40% EtOH from the original simulation and after the 200 ns extension. (a) Replica 1; (b) replica 2.

(a)

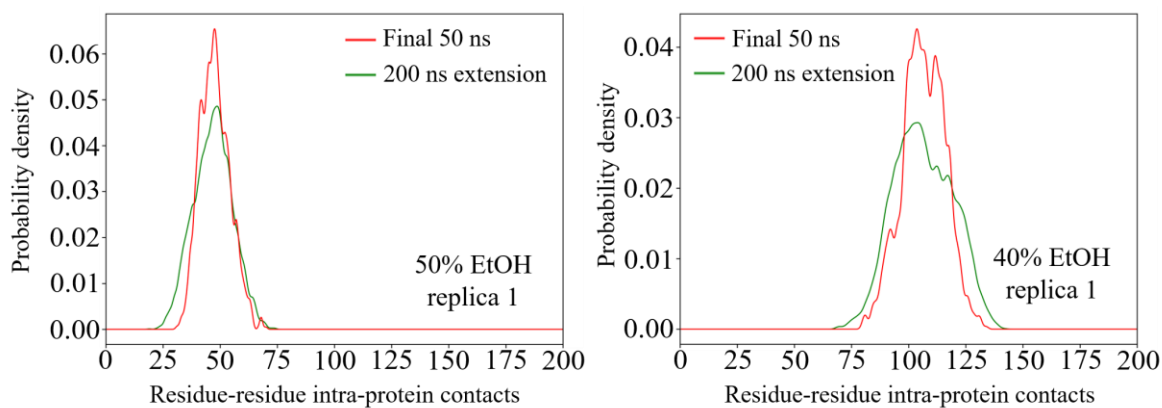


(b)

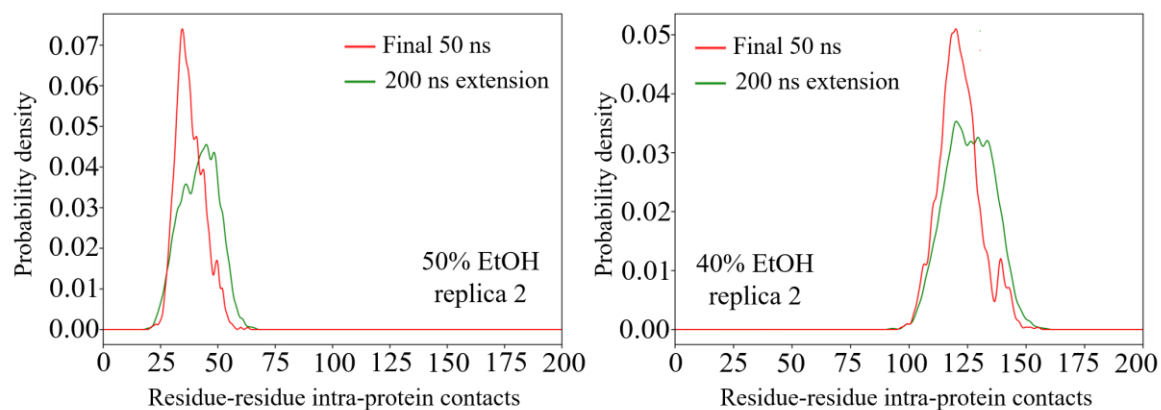


Appendix Figure A.11. Residue–residue intra-protein contact distributions from the final 50 ns and subsequent 200 ns extension of all-atom MD simulations at 50% and 40% EtOH. (a) Replica 1; (b) replica 2.

(a)

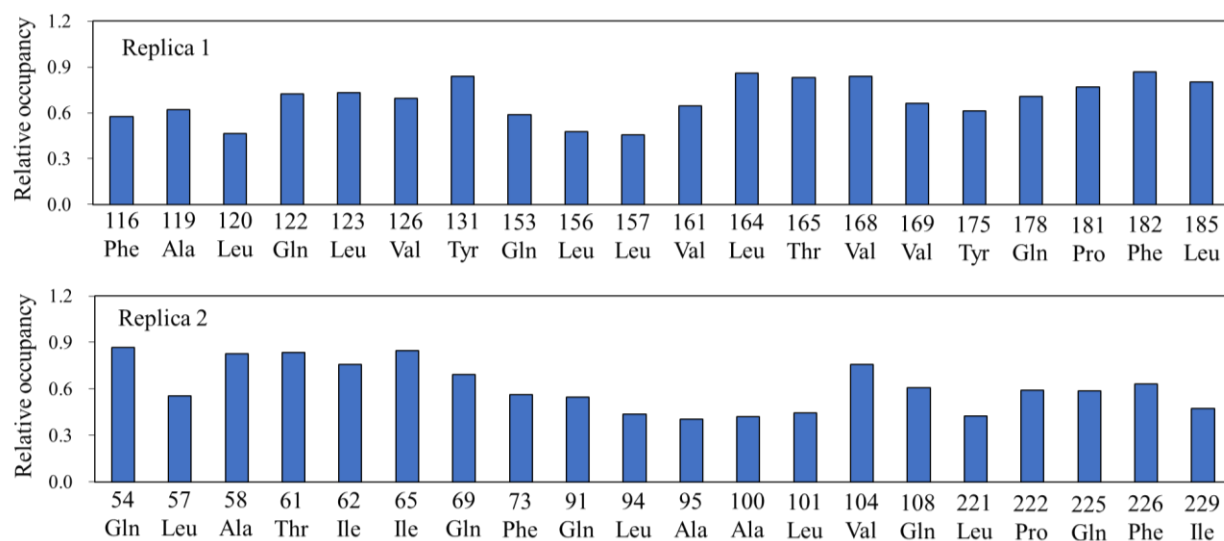


(b)

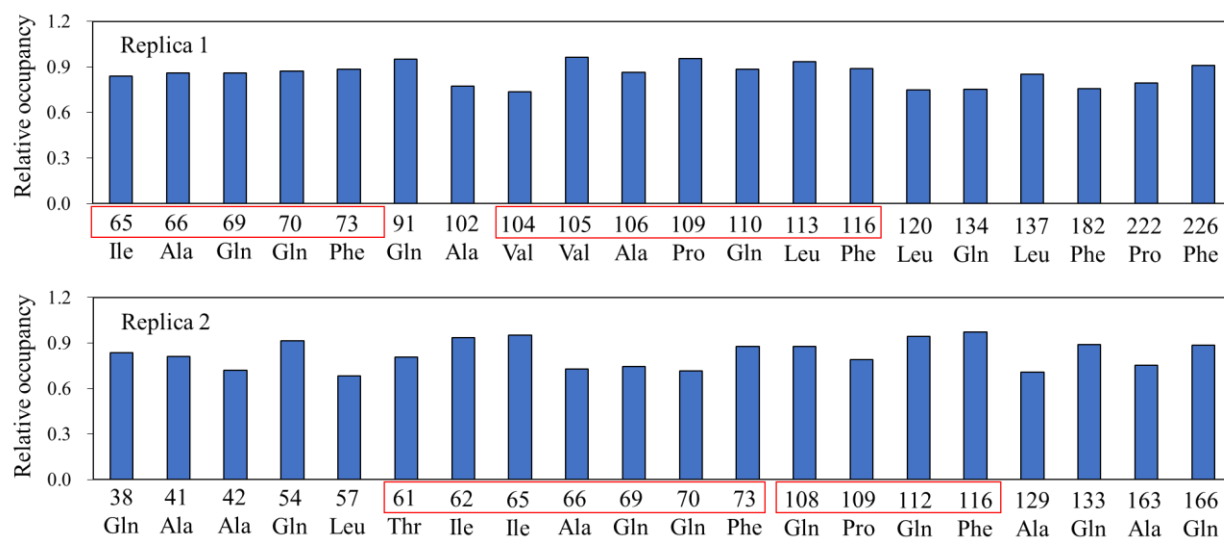


Appendix Figure A.12. Residue-level contact frequency of zein–lutein interactions at different ethanol concentrations of 60% (a), 40% (b), and 20% (c). Contact frequencies were calculated over the final 50 ns of each trajectory. Results from independent replicas are shown.

(a)

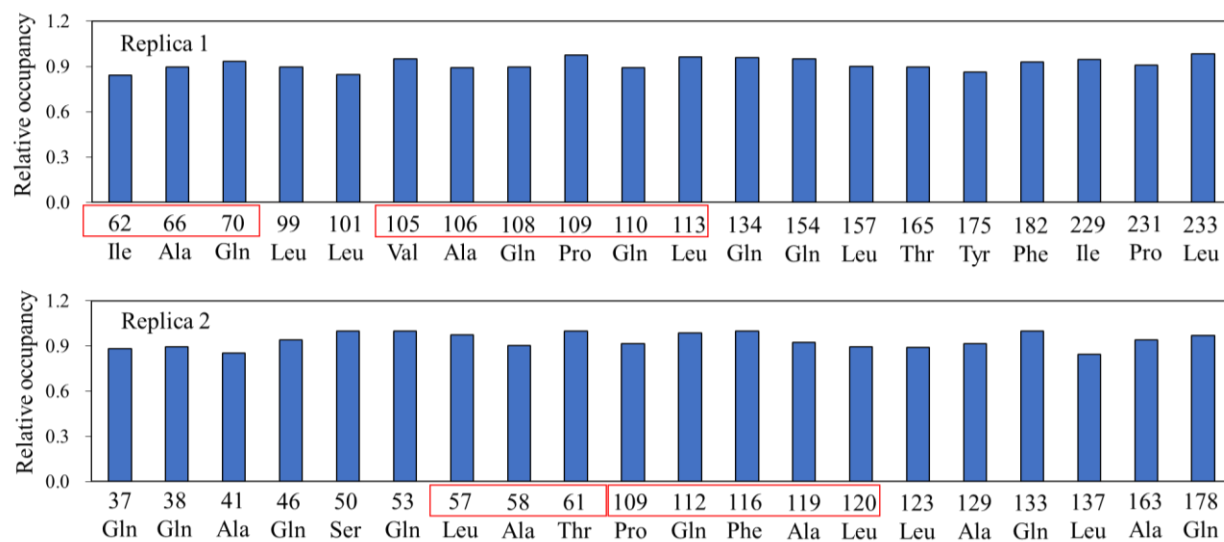


(b)



Appendix Figure A.12. (Continued)

(c)



Appendix Figure A.13. Particle size distributions of dilute solution samples of zein–lutein, zein-only, and lutein-only systems at 40% and 35% EtOH.

