ANNUAL REVIEWS

Annual Review of Food Science and Technology

Bioinformatic Approaches for Characterizing Molecular Structure and Function of Food Proteins

Harrison Helmick,[1] Anika Jain,[2] Genki Terashi,[2] Andrea Liceaga,[1] Arun K. Bhunia,[1] Daisuke Kihara,[2,3] and Jozef L. Kokini[1]

[1]Department of Food Science, Purdue University, West Lafayette, Indiana, USA; email: jkokini@purdue.edu

[2]Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA

[3]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA

Annu. Rev. Food Sci. Technol. 2023. 14:203–24

First published as a Review in Advance on January 9, 2023

The *Annual Review of Food Science and Technology* is online at food.annualreviews.org

https://doi.org/10.1146/annurev-food-060721-022222

Keywords

bioinformatics, protein modeling, computational biology, materials characterization, data science, protein functionality

Abstract

Structural bioinformatics analyzes protein structural models with the goal of uncovering molecular drivers of food functionality. This field aims to develop tools that can rapidly extract relevant information from protein databases as well as organize this information for researchers interested in studying protein functionality. Food bioinformaticians take advantage of millions of protein amino acid sequences and structures contained within these databases, extracting features such as surface hydrophobicity that are then used to model functionality, including solubility, thermostability, and emulsification. This work is aided by a protein structure–function relationship framework, in which bioinformatic properties are linked to physicochemical experimentation. Strong bioinformatic correlations exist for protein secondary structure, electrostatic potential, and surface hydrophobicity. Modeling changes in protein structures through molecular mechanics is an increasingly accessible field that will continue to propel food science research.

1. INTRODUCTION

Bioinformatics is a diverse field with applications in biological sciences, including identifying new vaccines and drugs, improving food protein functionality, and understanding protein interactions (Aguilar-Toalá et al. 2019, Gauthier et al. 2018, Goodman et al. 2016, Lin et al. 2017). Bioinformatics can be broadly defined as the development and use of computer algorithms to analyze biological data, including genetic information, protein amino acid sequences, and protein structures (Patel et al. 2019). With such a broad definition, it is useful to divide bioinformatics into categories, and structural bioinformatics is one discipline that shows significant promise in food science.

Structural bioinformatics analyzes experimental data and models protein molecular structures, enhancing understanding of protein structure–function relationships. This field grew from the work of Nobel Prize Laureate Christian Anfinsen, who demonstrated that a protein's structure dictates its function, which later became known as the thermodynamic hypothesis (Anfinsen 1973, Hirata et al. 2018). Since that seminal work, countless biochemists, polymer physicists, and computer scientists have contributed to the understanding of how protein molecules fold and its implications in biological processes. In food science, bioinformatics is used to analyze protein structures, providing insights into the effects of molecular structure on emulsification, foaming, gelation, solubility, denaturation, and other important functionalities (Garcia-Moreno et al. 2020, Gupta et al. 2016, Hou et al. 2019, Pucci et al. 2017, Tang 2017). This review focuses on the tools available in the field of structural bioinformatics and how they are used to develop an enhanced understanding of food functionalities.

2. PROTEIN SEQUENCE AND STRUCTURE OVERVIEW

Proteins comprise twenty unique amino acids, and the primary structure of the protein is the linear sequence of amino acids (**Figure 1**) (Ouellette & Rawn 2015). Secondary structure consists of specific regular arrangements within the polypeptide chain formed due to hydrogen bonding between the backbone atoms of the amino acid chains, namely oxygen atoms on the carbonyl groups and the hydrogen atoms on the amino groups. These structures include α-helices or β-pleated sheets that may be organized in parallel or antiparallel directions. Proteins can further

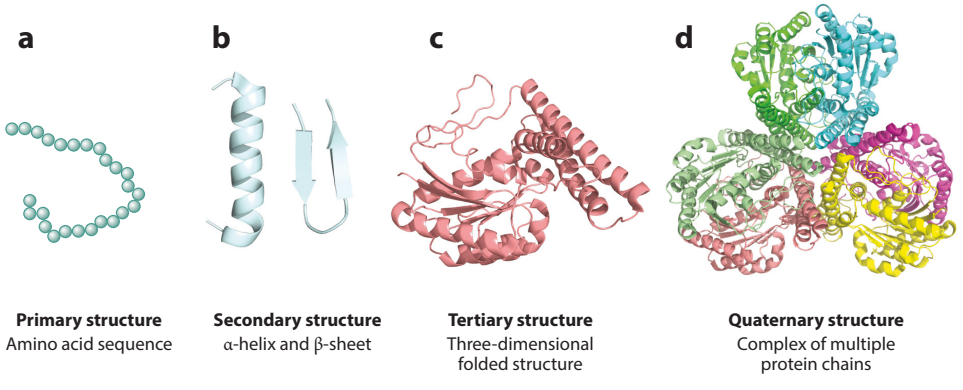

Figure 1

Protein structure overview depicting, from left to right, the primary structure (amino acid sequence); secondary structures (β-strands, α-helices, random coils); tertiary structure (folding of secondary structures into a three-dimensional conformation); and quaternary structures (the organization of protein monomer chains). Tertiary and quaternary structure PDB (Protein Data Bank) ID:1U6J (Berman et al. 2000).

adopt three-dimensional (3D) tertiary structures, resulting from interactions between groups of amino acids. Interactions can be covalent disulfide bonds between cysteine groups and noncovalent interactions, including electrostatic, hydrophobic, hydrogen bonds, and van der Waals interactions. In nature, many proteins exist as complexes consisting of two or more different polypeptide chains, which is known as quaternary structure. The quaternary structure is primarily stabilized by physical interactions, such as van der Waals, hydrophobic, hydrogen bonding, and/or electrostatic interactions between the different polypeptide chains (Ouellette & Rawn 2015).

3. PROTEIN SEQUENCE AND STRUCTURE DATABASES

Protein databases are foundational in structural bioinformatics, and these include amino acid sequence databases, structural databases, and databases that organize proteins based on structural features. Using databases allows food science researchers to develop models of protein functionality like solubility and denaturation (Hou et al. 2019, Pucci et al. 2017) as well as compare protein structures for relevant features (Tandang-Silvas et al. 2010). **Table 1** provides links to common databases as well as all the tools discussed in this manuscript.

3.1. Sequence Databases, Multiple Sequence Alignment, and Sequence Clustering Tools

Sequence databases are biological databases containing many nucleic acid or amino acid sequences deposited by researchers internationally. The large scale of these repositories led to the development of sequence retrieval and clustering programs to search for homologous sequences and create multiple sequence alignments (MSAs) because homologous sequences may share structural similarities. There are two sequence databases, UniProt (Universal Protein Resource) (**http://www.uniprot.org**) and NCBI (National Center for Biotechnology Information), that are used by most researchers.

UniProt is a comprehensive resource for protein sequences and their functional annotations (known binding substrates, biological functions, etc.). Each protein entry in the UniProt database compiles information from four different UniProt services. UniProtKB (UniProt Knowledgebase) cross-references and integrates information about proteins from various sources. UniRef (UniProt Reference Clusters) uses sequence identity to combine sequences that are closely related to aid in increasing search speeds. UniParc (UniProt Archive) is a repository that consists of the history of stored protein sequences. UniMES (UniProt Metagenomic and Environmental Sequences) primarily consists of environmental and metagenomic data (UniProt Consort. 2008). Information from these sources is aggregated into a single report about a protein when searched. The other major database is the NCBI Protein database, which contains amino acid sequences that have been translated from coding regions deposited in the GenBank and RefSeq databases. It also incorporates data from PDB (Protein Data Bank), UniProtKB/Swiss-Prot, PIR (Protein Information Resource), and PRF (Protein Research Foundation) databases (Acland et al. 2013).

Utilizing these databases is aided by sequence retrieval and alignment tools. The Basic Local Alignment Search Tool (BLAST) (**http://www.ncbi.nlm.nih.gov/blast**) is one of the most popular programs to search for sequences similar to the query (user input) sequence. BLAST can be run online through the NCBI website, or software can be downloaded and run on a user's computer. For amino acid sequence alignment, the algorithm finds similar regions between a query sequence and sequences in the database by splitting the query sequence into smaller groupings of amino acids. The segments that are similar to those in the database serve as hot spots where sequence alignment begins. This approach is relatively fast compared to other sequence alignment methods (Johnson et al. 2008).

Table 1 Software used in various bioinformatics analyses

Types of software and databases	Name	Function/brief description	Link for the software/database
Sequence database	UniProt	Functionally annotated protein sequences	**http://www.uniprot.org**
	NCBI protein database	Protein sequences translated from GenBank and RefSeq databases	**https://www.ncbi.nlm.nih.gov/protein/**
Structure database	PDB	Experimentally determined three-dimensional protein structures	**http://www.rcsb.org/**
	CATH	Classification of protein domain structures	**http://www.cathdb.info**
	SCOP	Classification of protein domain structures	**https://scop.mrc-lmb.cam.ac.uk/**
Model database	SWISS-MODEL repository	Annotated 3D models of protein structures generated by SWISS-MODEL	**http://swissmodel.expasy.org/repository/**
	AlphaFold database	3D models of protein structures generated by AlphaFold	**https://alphafold.ebi.ac.uk**
Sequence search/ clustering	BLAST	Searches for homologous sequences	**http://www.ncbi.nlm.nih.gov/blast**
	HMMER	Profile Hidden Markov Models are used to search for homologous sequences	**http://www.ebi.ac.uk/Tools/hmmer/**
	MMseqs2	Sequence similarity search tool using k-mer similarity algorithm	**https://github.com/soedinglab/mmseqs2**
Homology modeling	SWISS-MODEL	Homology modeling using rigid-fragment assembly technique	**http://swissmodel.expasy.org**
	MODELLER	Homology modeling using multiple template structures	**https://salilab.org/modeller/**
Machine learning–based modeling	AlphaFold2	Neural-network-based three-dimensional protein structure prediction	**https://github.com/deepmind/alphafold**
	RoseTTaFold	Multitrack network for protein structure prediction	**https://github.com/RosettaCommons/RoseTTAFold**
Structural assessment tools	pLDDT by AlphaFold	Assigns confidence values to the predicted structure	**https://github.com/deepmind/alphafold**
	MolProbity	Validates predicted protein models	**http://molprobity.biochem.duke.edu**
	Pcons	Consensus based method for model quality assessment	**http://pcons.net/**
	PrQ and ProQres	Structure-based method for model quality assessment	**https://proq.bioinfo.se/ProQ/index.html.iso8859-1**
	CASP	Double-blind structure assessments	**https://predictioncenter.org/**
	CAMEO		**https://www.cameo3d.org/**
Visualization	PyMol	Molecular visualization tool that supports custom Python and pml scripts	**https://pymol.org/2/**
	Chimera	Supports multiple extensions for additional visualization support	**https://www.cgl.ucsf.edu/chimera/**
	VMD	Supports Tcl or Python scripts for additional custom visualization tools and can be used to prepare molecular dynamics systems	**https://www.ks.uiuc.edu/Research/vmd/**
	Swiss-PDB Viewer	View and alter structural alignments for SWISS-MODEL inputs; provides structural assessment using Ramachandran Plots	**https://spdbv.unil.ch/**

(*Continued*)

Table 1 (*Continued*)

Types of software and databases	Name	Function/brief description	Link for the software/database
Protein–protein docking	LzerD and Multi-LzerD	LzerD can dock two proteins with each other; multi-LzerD can dock two or more proteins together	**https://lzerd.kiharalab.org**
	Rosetta Dock	Monte-Carlo-based algorithm for protein docking	**http://rosettadock.graylab.jhu.edu**
	HADDOCK	Utilizes interface information from experimental data; can also be used for protein-oligosaccharide docking	**https://wenmr.science.uu.nl/haddock2.4/**
	ClusPro	Uses PIPER docking program for rigid body docking of proteins	**https://cluspro.org**
Protein–ligand docking	AutoDock4	Predict protein–ligand interactions	**http://autodock.scripps.edu**
	Glide	Part of the Schrodinger suite of biological tools; has three docking modes that can be used for screening huge databases and accurate and high-precision docking	**https://www.schrodinger.com/products/glide**
	SwissDock	Uses EADock DSS docking program	**http://www.swissdock.ch/**
Molecular mechanics	CHARMM	Molecular simulation program to assess dynamic properties of macromolecules	**www.charmm.org**
	CHARMM-GUI	Generates input files and molecular systems for molecular dynamics simulations	**http://www.charmm-gui.org**
	AMBER	Can be used to build carbohydrate models for molecular dynamics simulations in addition to other macromolecules	**http://ambermd.org/**
	GROMACS	Fast molecular dynamics simulation program	**https://www.gromacs.org/**
	NAMD	Used for simulations of large biological systems	**https://www.ks.uiuc.edu/Research/namd/**

Abbreviations: AMBER, Associated Model Building with Energy Refinement; BLAST, Basic Local Alignment Search Tool; CAMEO, Continuous Automated Model EvaluatiOn; CASP, Critical Assessment of Structure Prediction; CATH, Classification, Architecture, Topology, Homology; CHARMM, Chemistry at HARvard Macromolecular Mechanics; CHARMM-GUI, Chemistry at HARvard Macromolecular Mechanics Graphical User Interface; GROMACS, Groningen Machine for Chemical Simulations; HADDOCK, High Ambiguity Driven protein-protein DOCKing); HMMER, name of software based on Hidden Markov Models; LzerD, Local 3D Zernike descriptor-based protein docking; MMseqs, many-against-many sequence searching; NAMD, nanoscale molecular dynamics; NCBI, National Center for Biotechnology Information; PDB, Protein Data Bank; pLDDT, predicted local distance difference test score; PrQ, protein quality predictors; SCOP, structural classification of proteins; VMD, virtual molecular dynamics.

A second approach to finding similar sequences is based on profile hidden Markov models (HMMs). HMMs are used across multiple disciplines to identify whether a sample is similar to a training data set. In sequence alignment, HMMs convert the MSAs into a position-specific scoring matrix that is used to search databases for similar sequences, reducing the perplexity (the number of computations needed). This encodes evolutionary changes that have occurred in a set of closely related sequences in a position-specific manner and captures information about gaps due to deletions or insertions in the sequences (Eddy 1998). One online tool using HMMs is HMMER (**http://www.ebi.ac.uk/Tools/hmmer/**), which identifies similar sequences by using profile HMMs. HMMER incorporates four programs that iteratively search databases and recommend similar sequences. The similarity of resultant sequences is reported using E-values or bit-scores, which are metrics that show sequence similarity, and low E-values indicate likely similar sequences (Finn et al. 2011).

Hhblits (HMM-HMM–based lightning-fast iterative sequence search) is a method that uses HMM–HMM alignment for sequence searching. It extends the Hhsearch HMM–HMM alignment method for faster iterative sequence searching by using two prefilters before HMM–HMM alignments are conducted. The method employed by Hhblits also improves the quality of alignment in the profile by adding pseudo counts, which prevent amino acid position probabilities from becoming 0. It is possible that a particular amino acid at a given position is not observed in a data set, but it is not necessarily true that it has a probability of 0, thus a pseudo count is added so that the probability is always greater than 0. The profile generated is used to search against a separate HMM database. Sequences that are below the predefined standard score (E-value) are added to the initial query MSA. This appended query MSA is then used to build the profile HMM for the next iteration of the search (Remmert et al. 2012).

3.2. Protein Structural Databases and Their Use in Developing an Understanding of Protein Structure–Function Relationships

The PDB (**http://www.rcsb.org**) is the primary database for experimentally determined 3D protein structures. It includes structures resolved using X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy methods. The database can be searched using the PDB ID of a deposited structure (e.g., 2PHL is the PDB ID for phaseolin, a plant seed storage protein), the amino acid sequence, the name of the protein (e.g., legumin), or the simplified molecular input line-entry system (SMILES) code of the docked ligand, where the SMILES annotation is a one-line method of writing molecular structure [e.g., c1nc(c2c(n1)n(cn2)[C@H]3[C@@H]([C@@H]([C@H](O3)CO[P@@](=O)(O)O[P@](=O)(O)OP(=O)(O)O)O)O)N) is the SMILES code for ADENOSINE-5′-TRIPHOSPHATE] (Weininger 1988). The Protein Feature View links with other databases to quickly compare sequences (UniProt) and structural similarity (Pfam) (Berman et al. 2000, Rose et al. 2017).

In addition to PDB's experimental structures, there are two primary databases that compile models of protein structures. Protein model databases contain predictions of protein structures based on machine learning methods or automated homology modeling and may not have experimental data to validate the accuracy of the model.

The SWISS-MODEL Repository (**http://swissmodel.expasy.org/repository/**) contains annotated 3D models of protein structures generated by automated homology modeling through the SWISS-MODEL pipeline (Waterhouse et al. 2018). As of May 26, 2022, the database has 2,217,761 models for sequences in UniProtKB generated by SWISS-MODEL. The database uses the QMEAN quality metric for model accuracy. QMEAN is a composite score that relies on calculating the mean force between atoms in the protein, which can be compared to other proteins, and a z-score is reported to determine the model's relative accuracy (Bienert et al. 2016, Studer et al. 2020).

The AlphaFold protein structure database (**https://alphafold.ebi.ac.uk**) contains 3D models for protein sequences predicted by DeepMind's AlphaFold2 (Jumper et al. 2021). It contains a large repertoire of predicted structures for the human proteome and other model organisms, including soybeans and maize. The number of structures continues to grow, but in June of 2022, it contained 995,411 predicted structures. It covers most sequences from the UniProt reference proteome as well as Swiss-Prot, although it excludes proteins that are shorter than 16 or longer than 2,700 amino acids. Proteins with nonstandard amino acids, such as hydroxyproline or hydroxylysine found in collagen or N-formylmethionine found in prokaryotic proteins, as well as proteins found in viruses, are also excluded (Varadi et al. 2022).

With millions of available structures, organizing them is a daunting task. The Class, Architecture, Topology, Homology (CATH) database (**http://www.cathdb.info**) classifies protein

structures from PDB hierarchically based on their domains. Protein domains are smaller regions of the amino acid chain that are capable of folding independently from the rest of the protein. The Class component groups proteins based on secondary structures; Architecture groups them according to orientation of secondary structures; Topology groups them based on the order of the secondary structures, and at the Homologous superfamily level, the domains are grouped based on their sequence and structural similarity (Knudsen & Wiuf 2010). CATH obtains domain information from data deposited in PDB and applies a combination of manual and computational methods to identify and classify the domains based on structural similarity. Proteins that share similar structures are also likely to have similar functionalities, allowing for the selection of proteins for targeted applications.

The Structural Classification of Proteins (SCOP) database (**https://scop.mrc-lmb.cam.ac.uk/**) is similar to the CATH database in that it also uses a hierarchical classification of protein domains. The classification is based on the family, superfamily, classes, and folds of the protein (Hubbard et al. 1997). Protein families are characterized by proteins that share similarities in their sequence, structure, or function owing to a shared ancestor. Superfamilies are the largest possible group of proteins that share a common ancestor and are characterized by structural similarity, even in the absence of strong sequence similarity. Classes correlate to the tertiary structure of the protein such as the globular organization of proteins. Fold-based classification groups proteins based on the topology of the tertiary structure, which considers similar secondary structure features such as all α-helix and all β-sheet regions (Csaba et al. 2009). The classifications between SCOP and CATH show some overlap, but the differences in grouping methods lead to some difference in how proteins are clustered. Using the databases in tandem is a useful strategy to gain insight into which proteins share the greatest similarity with a structure of interest.

By grouping proteins based on sequence and structural similarity, insight into functional attributes can be obtained by comparing conserved and divergent regions, where a conserved region is a similar structure or amino acid sequence motif across time and species and divergent areas show differences (Brown & Babbitt 2014; Tandang-Silvas et al. 2010, 2011, 2012; Uberto & Moomaw 2013). This is based on the hypothesis that conserved regions are of particular functional interest (Brown & Babbitt 2014, Rahman et al. 2020, Tandang-Silvas et al. 2010). Once conserved regions are identified using sequence or structure alignment tools, they are analyzed to understand functionality and its physical characteristics. This method has been applied to plant-based storage proteins, showing sequence conservation within the 11S proteins of pumpkin, soy, pea, and others ranged between 36% and 63% identical amino acids, and the maximum root mean square difference (RMSD) between structures was ∼9 Å after structure alignment (Tandang-Silvas et al. 2010). This was similar to the sequence conservation of pea protein 11S subunits when classified into legumin A, J, and S families, where sequence identity was 38.9% across all families (Helmick et al. 2021b). Further analysis revealed that internal protein cavities, the number of intramolecular hydrogen bonds, length of looping regions, and proline are important factors in distinguishing the high thermal stability of plant-based storage proteins, which is a common trait of cupin proteins, and was supported through differential scanning calorimetry in their work (Tandang-Silvas et al. 2010, Uberto & Moomaw 2013). Other research used homology models of napin and cruciferin to compare their structural and sequence similarity to known antimicrobial peptides, finding a conserved region with bacteriostatic properties (Rahman et al. 2020). Experimental validation showed that the napin proteins could effectively inhibit the growth of *Staphylococcus saprophyticus*, although the cruciferin did not show the same properties (Rahman et al. 2020). The conserved domain approach can also point to functional regions in enzymes and allergens, and there is a conserved domain between a variety of allergenic tree nut 2S proteins (Brown & Babbitt 2014, Goodman et al. 2016). This similarity-based approach can be

used in the food industry to identify which proteins have health-promoting peptides, allowing for the selection of ingredients that could facilitate health claims of newly developed products, and it can also be used to select proteins that would enhance shelf life through antimicrobial properties.

Some protein databases integrate structural information with experimental information, including solubility (Hou et al. 2019), thermostability (Bava et al. 2004), isoelectric points (Kozlowski 2016), and enzyme activity (carbohydrates) (Cantarel et al. 2008). When using large data sets that compile experimental information, it is important to consider that data are reported by researchers using different methods to obtain experimental information. For example, the temperature and enthalpy of denaturation can be obtained by measuring tryptophan fluorescence in ultraviolet visible spectroscopy, circular dichroism, and differential scanning calorimetry (Bava et al. 2004). Furthermore, thermodynamic parameters depend on solvent conditions (ionic strength and pH), rate of temperature increase during the experiment, and protein structure (Bava et al. 2004, Sun & Arntfield 2010). When generalizing models derived from data sets, experimental differences must be addressed for robust model generation. Additionally, databases frequently have redundant structures, over-representing a particular type of protein, reducing model generalizability, and leading to faulty conclusions (Walsh et al. 2015). Web servers like PISCES minimize this bias by removing similar proteins based on user-defined sequence similarity cutoffs (Walsh et al. 2015, Wang et al. 2003). Once a diverse and representative data set is developed, protein structures can be analyzed to produce models of functionality.

One approach to developing models from data sets is by comparing proteins from a data set with a known functionality (e.g., thermal stability) to a second independent data set that is a random collection of proteins. This allows for the development of statistical potentials (Miyazawa & Jernigan 1996, Shen & Sali 2006). If an attribute is significantly more common in the protein data set with known functionality than in the random data set, this may be a variable that imparts functionality. This concept was pioneered in assessing the interatomic distance of amino acids in protein structures, finding that particular amino acid pairs, such as hydrophobic amino acids, tended to be located physically close together (Miyazawa & Jernigan 1996, Shen & Sali 2006). These distances followed a Boltzmann distribution, which allowed for estimations of protein free energy based on interatomic distances and enhancing models of protein folding (Shen & Sali 2006). Statistical potential is now used to improve model accuracy in several applications, such as protein–protein docking, protein–ligand docking, and protein solubility or thermostability enhancement (Alford et al. 2017, Hou et al. 2019, Pucci et al. 2017).

One example of using database information to model protein functionality is in making predictions of protein solubility (Hou et al. 2019). Solubility is a property dependent on many factors (protein hydrophobicity, electrical charge, molecular weight, etc.), but through the use of a random forest model, coupled with statistical potentials based on π electron interactions, Hou et al. (2019) developed a model (named SOLart) that had R values of 0.78 when using the Esol database. Protein thermostability was investigated using the ProTherm database to predict protein stability using the Gibbs-Helmholtz model through the SCooP algorithm (Pucci et al. 2017). The algorithm uses a linear combination of statistical potentials to predict the enthalpy and temperature of denaturation from the database with R values of 0.80 and 0.72, respectively (Pucci et al. 2017). The thermodynamic parameters obtained are used to solve the Gibbs-Helmholtz model of protein denaturation, providing the range of protein stability (Becktel & Schellman 1987, Pucci et al. 2017).

Working with data sets that have thousands, if not millions, of proteins requires knowledge of data science programming languages, including R or Python, which can be a barrier to entry for some researchers. However, bioinformatics tools like SOLart and SCooP are also implemented through web servers or software that can help users apply the models to their own data, and entire

issues of leading bioinformatics journals aggregate these data into annual publications of web-based services (Brazas et al. 2010, Hou et al. 2019, Pucci et al. 2017). This can aid in selecting proteins for product developers when solubility and thermal stability are important variables, such as in nutritional beverages or egg replacement in the baking industry.

4. CONSTRUCTING NEW AND ACCURATE PROTEIN STRUCTURAL MODELS

Structural databases contain many protein structures, but researchers often need to generate protein models based on amino acid sequences generated in their research or amino acid sequences stored in databases. Model generation can be broadly grouped into homology modeling or ab initio methods. Homology modeling uses an experimentally solved structure of a homologous sequence as a template to generate a 3D model of the query sequence. Recent ab initio methods use machine learning to generate protein structures without the use of a single reference template. Both methods model secondary and tertiary structures of proteins by using either homologous sequences and structures or the incorporation of advanced neural network architectures (Mirdita et al. 2022).

4.1. Platforms for Generating Accurate Models of Protein Structure

SWISS-MODEL (**http://swissmodel.expasy.org**) is used for homology/comparative modeling of protein structures. It employs a rigid fragment assembly technique, matching all similar regions to the template and recreating similar structures (Waterhouse et al. 2018). To generate the model, a user's input sequence is compared against sequences in the ExPDB library, which is a structural database derived from PDB. Structural alignment is generated using the template structures after calculating the local pair-wise alignment of the query sequence to the template structure. The model is built by averaging the positions of the backbone atoms in the template structure. In regions with indels (insertions or deletions), the template coordinates cannot be used for model generation, so a set of compatible fragments with the neighboring stems in the template structure is generated using constraint space programming (CSP). A scoring scheme that considers force field energy, favorable interactions, and steric hindrance selects the best-formed loop for these regions. An experimentally determined loop library is used for fragments if CSP cannot model the loops well. Steepest descent energy minimization, using the GROMOS96 force field, normalizes deviations in the modeled structure. The output model file contains a C-score that estimates the variability of all template structures for every position. A C-score of 99 indicates regions where the model could not use template coordinates or information owing to indels (Schwede et al. 2003). These steps are completed in minutes using an online server, and SWISS-MODEL can produce quaternary structure, which is not always the case in homology modeling software.

MODELLER is another homology modeling software that follows the main steps of template search, alignment of the query and template, and building the model with evaluation metrics. Unlike SWISS-MODEL, MODELLER combines information derived from multiple template structures in two ways. Different template structures may be aligned to different domains of the query sequence without much overlap, aiding in constructing a homology-based model on the query sequence. Additionally, multiple template structures may align with the same domain of the query sequence in which case the homology model would be constructed by selecting the best template that fits the local context. Hence, MODELLER includes multiple template structures that may differ from each other but still have an overall similarity to the query sequence. These templates all contribute to modeling different sections of the target sequence. Query–template alignment is conducted by aligning multiple sequences from a database of

both the query and template. MODELLER then generates a homologous structure by using distance geometry or other techniques that satisfy information of spatial restraints derived from the alignment. For model accuracy, MODELLER extracts distance and dihedral angle restraints from the query–template alignment, the statistical information for the dihedral angles, the nonbonded related atomic distance from a representative subset of all known proteins, and stereochemical restraints (bond length and angle) from the CHARMM (Chemistry at Harvard Macromolecular Mechanics) force field. It then refines the model structure using methods that aid in spatial restraint violation minimization, relying on molecular dynamics (Fiser & Šali 2003). MODELLER does not build a quaternary structure by default, although this can be developed through modeling each subunit of a larger protein structure individually.

Homology models depend on the presence of an experimentally solved template with high sequence similarity (>30%) to model the structures of the query sequence (Patel et al. 2019). This makes it difficult for homology modeling algorithms to generate accurate models without template information from databases. Machine learning methods can be used to circumvent this drawback.

AlphaFold2 is a neural-network-based protein structure modeling tool that predicts the 3D structures of query sequences with high accuracy, even without close homolog templates (Jumper et al. 2021). Structures predicted by AlphaFold2 proved to be highly accurate in CASP14 (Critical Assessment of Structure Prediction), which is a biannual competition of more than 100 protein modeling researcher groups to develop better prediction algorithms (Kryshtafovych et al. 2021). AlphaFold2 structures had a median backbone accuracy of 0.96 Å Cα RMSD when considering 95% of amino acids and an all-atom accuracy of 1.5 Å when compared to X-ray crystallography models. The next best prediction method's median backbone accuracy was 2.8 Å RMSD at 95% residue coverage and all-atom accuracy of 3.5 Å RMSD (width of the carbon atom is 1.4 Å). This makes AlphaFold2 ab initio structures competitive with X-ray crystallography structures in terms of model accuracy.

In AlphaFold2, a query sequence is compared against sequence databases to build an MSA of homologous sequences that is used as one of several inputs into a neural network architecture. The first two layers exchange information from an MSA that aids in spatial arrangements of the final protein structure. A structural layer then incorporates information from structures used in training data. After an initial model is built, the program refines amino acid side-chain placements iteratively. This iterative process is referred to as recycling, and it was shown to greatly improve the accuracy of predictions for side chains (Jumper et al. 2021). AlphaFold2 is highly accurate, although it is limited to proteins of fewer than 2,700 amino acids, and takes several hours to construct protein models when run using the Google Colab notebook that serves as the AlphaFold2 web server (**https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb**).

AlphaFold2 assigns confidence scores to its predicted models through predicted local distance difference test scores (**Figure 2**). Every residue in the structure receives a score ranging from 0 to 100 wherein a score of 100 shows the highest confidence in the position and atomic coordinates of that residue. These are calculated at the end of the network using small-scale per-residue networks (Jumper et al. 2021).

RoseTTaFold is an alternative structure prediction algorithm that also incorporates a multi-track neural network to effectively design accurate protein models. It consists of a one-dimensional (1D) sequence alignment track, a two-dimensional (2D) track of the distance matrix (map), and a 3D backbone coordinate track. This parallel flow of information ensures the understanding of residue distances, sequence, atomic coordinates, and residue orientation information in modeling the final structure. This is in comparison to the AlphaFold2 network architecture wherein the 1D

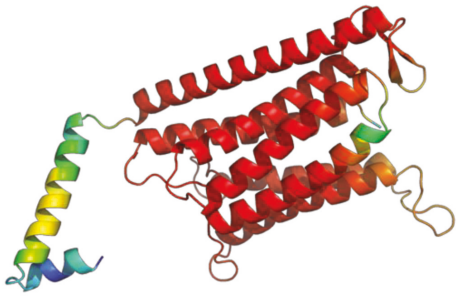

Figure 2

pLDDT (predicted local distance difference test) scores. Red regions indicate high confidence and blue regions indicate low confidence in the protein structure predicted using AlphaFold2. AlphaFold Database ID: AF-A0A1D6PQ72.

and 2D information of sequences and distance maps gets processed first, followed by the structure module considering the 3D coordinates. The relationship between the sequence and structure can be extracted more effectively using this three-track flow of information as compared to a two-track network. The final predicted 3D structures are constructed by averaging the 1D and 2D features and feeding them to a special Euclidian 3D [SE(3)]-equivariant layer, and RoseTTaFold also available as an online tool (**https://www.rosettacommons.org/docs/latest/Rosetta-Servers**) (Baek et al. 2021).

4.2. Tools Available for Independent Validation of Structure Model Accuracy

Although the above programs provide model quality estimators, several tools exist to independently validate models. Assessment methods provide a range of confidence levels in prediction accuracy over the entire structure. Regions of the structure with low prediction confidence can further be refined using alternative modeling strategies before using them to model functionality or interactions with other macromolecules.

MolProbity (**http://molprobity.biochem.duke.edu**) is one tool to validate models. It builds on software like PROCHECK that has primarily been used for the assessment of homology models. MolProbity has multiple categories of structure check parameters, such as Ramachandran plots for the ϕ, ψ backbone angles, all-atom contact assessment, sidechain rotamer χ angles, and torsion angle criteria. MolProbity incorporates the REDUCE software to add hydrogen bonds to structures, which is necessary for the assessment of all-atom contact clash analysis and optimizes their positions. The PROBE program then analyzes the all-atom contacts by measuring the overlap between nonbonded atom pairs. An overlap of 0.4 Å or more is considered a clash (Chen et al. 2010, Williams et al. 2018).

A second suite of programs is available from the Wallner & Elofsson (2007) research group. These methods use three categories of programs developed for the assessment of model quality: consensus-based, structure-based, and evolutionary-based. Consensus-based methods, such as Pcons, assess multiple predicted models for structural consensus by comparing the similarity of the selected models to all other predicted models in the consensus. It uses a structural superposition algorithm called dLGscore to obtain structural consensus. It depends on the idea that if a pattern is recurring in multiple models, it is likely to be correct. The result is a global and local quality score indicating correctness.

Structure-based methods (ProQ and ProQres) use structural features of proteins, including residue or atom contacts and agreement of tertiary structure to the secondary structure. These

features are used as inputs to a trained neural network that predicts model quality (Wallner & Elofsson 2007). Although ProQ predicts the global quality by extracting these features from a whole-protein model, ProQres predicts the local quality of the structure.

Consensus methods have been shown to perform best in most cases, whereas ProQprof performs best when there is a long evolutionary distance between the predicted model's query sequence and the template used for alignment. ProQres depends on high-quality predicted models to assess the local quality of the structural features of the overall model to match the quality of the training set of the network (Wallner & Elofsson 2007).

In addition to servers, competitions such as CASP, a biannual double-blinded experiment, assess protein monomer and oligomer structure prediction methodologies. Participants are provided with the amino acid sequence of a protein and then generate a 3D structure. These submissions are compared against the experimentally determined (X-ray crystallography) structures of the target and assessed for the accuracy of the modeled structure, including protein features and residue placement. CASP provides a platform for assessing the performance of breakthrough tools, such as AlphaFold2, in predicting structures and encourages the development of tools for improved predictions (Kryshtafovych et al. 2021).

CAMEO (Continuous Automated Model EvaluatiOn) assessments are a more frequent blind assessment of structure prediction methodologies. CAMEO is based on the model prediction of sequences that are prereleased weekly before the structures are published in PDB. It collects models over a four-day window and releases the benchmarking results of the models weekly. This provides developers with a tool to constantly validate their methods (Haas et al. 2018).

5. SOFTWARE FOR ANALYZING PROTEIN STRUCTURAL MODELS AND APPLICATIONS IN FOOD SCIENCE

Once an accurate protein model is available, whether from a database or modeling platform, researchers extract structural information of interest. There are several programs that aid in this research, but the most common are PyMol, Chimera, and Swiss-PDB Viewer.

PyMol is a molecular visualization tool (based on Python) that is used to display and assess atomic- and residue-level distances in biomolecular structures such as protein, DNA, and RNAs. It is an effective tool to visualize structures with different representations, including molecular surfaces, cartoons of the secondary structures, and atomic structures, aiding in the interpretation of interactions with ligands and protein folding. It further supports custom Python or PyMol Macro Language (pml) scripts that can be used for extracting protein characteristics. Molecular dynamics trajectories can be visualized as a movie in PyMol to understand the changes in interactions and folds throughout the simulation. It has additional plugins that can be used to show clashes and allow virtual screening of ligands and analysis of the binding sites. It also provides a mutagenesis tool for introducing point mutations in proteins or nucleic acids (Mooers 2020, Seeliger & de Groot 2010).

Chimera is another molecular visualization system primarily programmed in Python. It can incorporate multiple extensions akin to the plugins that have many of the same functionalities as PyMol and has the same visualization capacity as well. UCSF ChimeraX is a recent visualization tool that builds on the original Chimera by incorporating capabilities to handle a wider range of biological analyses such as medical imaging data as well as larger structures and has a user-friendly interface (Pettersen et al. 2004, 2021).

Swiss-PDB Viewer has many of the same molecular display and analysis tools as the other mentioned software programs (molecular representations, distances, etc.). It can be used to view and alter structural alignments before using them as a direct input in SWISS-MODEL. Like

PyMol and Chimera, it provides amino acid point mutation tools (Guex & Peitsch 1997). Unlike PyMol and Chimera, the scripting language is YACC (Yet Another Compiler Compiler), which is similar to languages such as Perl.

Using analytical software, proteins can be scanned for features, including surface hydrophobicity, secondary structure, and the electrostatic environment of the protein, that are related to physical properties (Helmick et al. 2021a,b; Hou et al. 2019; Pripp et al. 2005; Pucci et al. 2017). This analysis is aided by a quantitative structure–property relationship framework [also referred to as quantitative structure–activity relationships], and this leads to an enhanced understanding of emulsification abilities, solubility, gelation, and thermal stability (Garcia-Moreno et al. 2020, Helmick et. al. 2023, Hou et al. 2019, Katritzky et al. 2010, Pucci et al. 2017, Tang 2017). A difficulty in modeling food functionality is that protein isolates used in food are often a mixture of globulins, albumins, glutelins, and prolamins (Tandang-Silvas et al. 2011). As such, much of the work that models the functionality of food protein takes averages of protein categories that predominate the composition of extracted proteins, such as the 7S and 11S portions of many plant-based proteins, glutenin and gliadin in wheat, and α/β lactoglobulin in whey (Ainis et al. 2019, Helmick et al. 2021b, Keller 2018, Tandang-Silvas et al. 2011, Wu et al. 2021).

One approach to model food functionality is to first relate bioinformatic structural properties to physicochemical data from materials characterization techniques, including Fourier transform infrared (FTIR) spectroscopy, circular dichroism, differential scanning calorimetry, zeta potential, and surface hydrophobicity measurements (Helmick et al. 2021b, Katritzky et al. 2010, Pucci et al. 2017, Robertson & Murphy 1997). These techniques are often sensitive to specific bonding interactions that can be readily extracted from bioinformatic models, such as hydrogen bonding (FTIR) or electrostatic interactions (zeta potential). Information from these physicochemical techniques correlates to attributes in food, such as texture and mouthfeel, and rheological characterizations, and these techniques make up much of the base structure–function relationships in food protein (Dickie & Kokini 1983, Helmick et. al. 2023, Rasheed et al. 2020, Turasan et al. 2017). Utilizing these approaches, food bioinformaticians aid in providing molecular understanding and drivers of food functionality, developing in silico models of protein functionality.

Secondary structure is a useful attribute in describing the functionality of food proteins and has been shown to correlate with thermal stability, gelation, and emulsification (Barth 2007, Patel et al. 2019, Tang 2017). Experimentally, secondary structure is assigned using circular dichroism, Raman, or FTIR spectroscopy, and the quantification of secondary structure is based on curve fitting the Amide I region or observing ellipticity at specific wavelengths (Barth 2007, Turasan & Kokini 2016). Curve fits are based on comparing the results of spectroscopic techniques with X-ray crystallography models of proteins that predominantly have a single secondary structure, and these become representative spectroscopic readings for that secondary structure (Goormaghtigh et al. 2009, Turasan & Kokini 2016). Standard deviations for these techniques are between 8% and 24% when compared to X-ray crystallography models (Goormaghtigh et al. 2009). Homology models with sequence identities greater than 30% typically deviate from the X-ray crystallography models by ∼2 Å (Patel et al. 2019). Thus, the information gained from bioinformatic analysis, even of homology models, is comparable to the experimental results in observing native protein structures from X-ray crystallography. An example comparing an X-ray crystallography structure and homology model is shown in **Figure 3**. These models represent the native structure of the protein, and starting from the native structure can be useful in deriving relationships, even though the proteins have not been subjected to extraction and processing.

In work with kafirin films, solvent polarity was correlated with the formation of α-helical structures, confirmed by circular dichroism (Dianda et al. 2019). This was bioinformatically

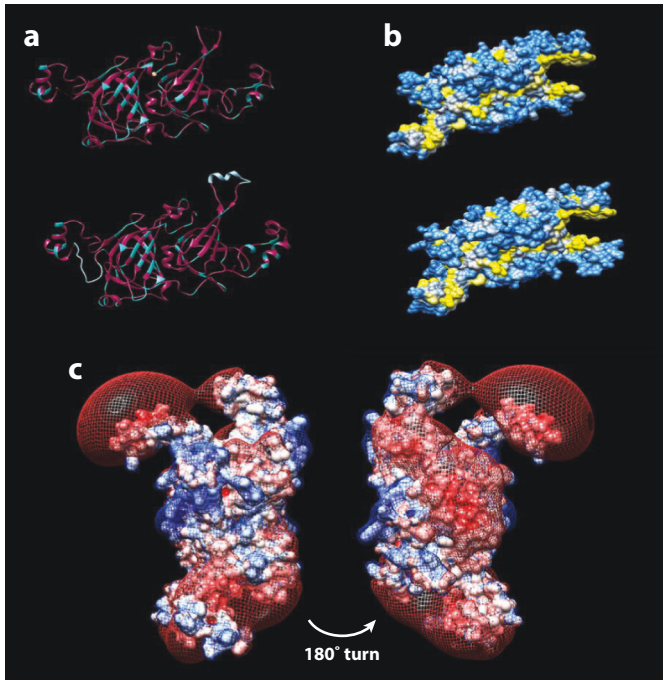

Figure 3

(*a*) Representative image of a homology model generated in SWISS-MODEL from a soybean amino acid sequence (PI11827.2) and visualized in Chimera. Coloring between the X-ray crystallography structure (1IPK) and homology model shows sequence conservation. (*b*) Display of solvent-accessible surface area colored using the Kyte-Doolittle hydrophobicity scale where yellow is the most hydrophobic (+4.5), blue is the least hydrophobic (−4.5), and gray is neutral (~0). (*c*) Representative image on solving the Poisson-Boltzmann model for a protein's structure. Homology model of pea 7S protein, accession CBK38923.1, solved at pH 7, 25°C, and no mobile ion presence. This is two sides of the same protein molecule showing the anisotropy of charge in the pea protein molecule.

investigated by using the I-TASSER homology modeling platform and analysis of hydrophobic amino acids in α-helices (Dianda et al. 2019). The burial and exposure of hydrophobic amino acids were the cause of the observed structural change and correlated with the hydrophobicity of films made with kafirin, measured through water contact angles (Dianda et al. 2019). Other work focused on the alignment of hydrophobic and hydrophilic amino acids within protein peptides and found that when hydrophobic residues occur at every N and N+4 position in α-helices, or every other residue in β-sheets, hydrophobic and hydrophilic faces are formed, allowing for suitable emulsifiers (Aguilar-Toalá et al. 2019). These findings were validated through tensiometer experiments, finding a reduction in surface tension of emulsions made using bioinformatically identified peptides (Garcia-Moreno et al. 2020). Less-ordered structures, indicated by high proportions of random coils, are also favorably correlated with emulsification and foaming abilities due to their molecular flexibility (Li et al. 2019, Tang 2017). Flexibility allows for rapid protein denaturation at the oil–water or air–water interface, producing superior emulsions and foams with strongly adsorbed protein (Li et al. 2019, Tang 2017). By using bioinformatics analysis, secondary structures of food materials can be estimated before processing, which can be used as one criterion in selecting proteins for targeted applications, such as emulsification or producing biodegradable films.

Surface hydrophobicity is another important feature that influences solubility, emulsification, gelation, foaming, and interactions with other molecules (Helmick et al. 2021b, Moro et al. 2001, Nakai 1983, Tong et al. 2021). Experimental measures of surface hydrophobicity include fluorescent probes that bind to hydrophobic surface regions on a protein such as ANS [1-(anilino)naphthalene-8-sulfonate] or PRODAN [6-propionyl-2-(N,N-dimethylamino) naphthalene] and hydrophobic chromatography (Alizadeh-Pasdar & Li-Chan 2000, Heldt et al. 2017). Bioinformatically, the surface hydrophobicity is calculated by measuring the solvent-accessible surface area of the protein and finding the contribution of the surface area imparted by hydrophobic amino acids using one of more than forty scales of amino acid hydrophobicity (Heldt et al. 2017, Lienqueo et al. 2002, Salgado et al. 2005). This value is the average surface hydrophobicity (ASH) (Heldt et al. 2017, Lienqueo et al. 2002). An example of a protein colored based on surface hydrophobicity according to the Kyte-Doolittle scale is shown in **Figure 3**. Two studies using small data sets (<7 proteins) found the ASH values from protein homology models and X-ray crystallography models, obtained using the Eisenberg or Kyte-Doolittle scale of hydrophobicity, correlated well with hydrophobic probes and chromatography (Heldt et al. 2017, Helmick et al. 2021b, Salgado et al. 2005). Calculating protein ASH requires 3D models of the protein; however, attempts at measuring hydrophobicity based strictly on the amino acid sequences have also been made, and R values between 0.769 and 0.803 were obtained using a sample of 1982 proteins and a machine learning approach (Salgado et al. 2005).

The electrostatic environment around a protein is another characteristic used by food scientists that can be derived bioinformatically (Helmick et al. 2021a, Honig & Nichols 1992, Klemmer et al. 2010). Experimentally, the pH-dependent electrostatic environment of protein is measured as the zeta potential, which is defined by the charged layer around a protein at the slipping plane that exists at some distance from the molecular surface of the protein (Chakravorty et al. 2017). Owing to the anisotropy of protein shape, determining where the slip plane lies bioinformatically is a challenging proposition, and the anisotropy of shape and charge leads to a difference in functionality between proteins; such is the case for β-lactoglobulin, lysozyme, and rapeseed napin protein (Ainis et al. 2019, Chakravorty et al. 2017). An example of this anisotropy in pea protein's 7S component is shown in **Figure 3**. By applying Coulombic or Poisson-Boltzmann models to solve the surface charge of the protein, estimates of the electrostatic environment of protein can be obtained bioinformatically (Ainis et al. 2019, Jurrus et al. 2018). The most robust model commonly used for these calculations is the Poisson-Boltzmann model (Honig & Nichols 1992, Jurrus et al. 2018).

This model allows for the presence of mobile ion species (e.g., NaCl), temperature, and pH of the calculation, and the results show the nature of the electrostatic interactions that surround the protein (**Figure 3*c***) (Jurrus et al. 2018). Protein protonation and deprotonation based on user-defined pH can be conducted using web servers, such as the adaptive Poisson-Boltzmann solver and PDB2PQR (Dolinsky et al. 2004, Jurrus et al. 2018). The Poisson-Boltzmann model correlates well with the zeta potential in pea legumin, vicilin, and commercial isolate when using protein homology models (Helmick et al. 2021a), and the net charge on soy protein calculated through the Henderson-Hasselbach equation considering Asp, Glu, Lys, and Arg correlated with the measured zeta potential in soy protein (Malhotra & Coupland 2004). Models like those for hydrophobicity and zeta potential show that bioinformatics can be a useful tool in describing processes like gelation that are partially driven by hydrophobic and electrostatic interactions (Guldiken et al. 2021, Liang & Tang 2013). This allows for the selection of proteins that minimize and maximize these interactions, aiding product developers in engineering specific textures of foods by minimizing or maximizing specific interactions.

6. PREDICTING AND SIMULATING PROTEIN MOLECULAR INTERACTIONS THROUGH MOLECULAR MECHANICS

Proteins are seldom present in a pure and isolated form, and understanding interactions with other biomacromolecules has important implications for food functionalities. This can be done through protein docking simulations or molecular mechanics simulations. Other reviews show how software like LzerD uses geometric hashing to find how proteins may interact with one another or how proteins interact with ligands (Aguilar-Toalá et al. 2019, Christoffer et al. 2021, Tao et al. 2020, Vidal-Limon et al. 2022). Although computationally expensive, molecular mechanics is another approach that shows strong potential for advancing food science research.

Molecular mechanics methods simulate the behavior of macromolecules in conditions similar to their biological environment or under experimental conditions of interest. These simulations can be run using protein homology models or X-ray crystallography models. Molecular mechanic/dynamic simulations then simulate atomic movement by solving equations for Brownian dynamics and incorporating restraints from predefined force fields as to how particles are allowed to move based on chemical interactions, such as van der Waals and electrostatic forces (Van Der Spoel et al. 2005). Although not by default, molecular systems are often built into programs such as CHARMM and AMBER (Associated Model Building with Energy Refinement), and the simulation is conducted in a second software such as GROMACS (Groningen Machine for Chemical Simulations) or NAMD (Nanoscale Molecular Dynamics) and finally visualized in programs like Chimera, PyMol, or Virtual Molecular Dynamics (VMD) (Case et al. 2005, Hsin et al. 2008, Humphrey et al. 1996, Phillips et al. 2005, Vanommeslaeghe et al. 2010). Many of these software programs must be run through a command line interface, which limits their usability for some researchers.

CHARMM-GUI helps make these tools more accessible (**http://www.charmm-gui.org**) by offering a graphical user interface to generate input files and molecular systems for simulations in other programs. This aids in creating accurate molecular systems with appropriate files for conducting simulations, which is complicated, even for those experienced in molecular mechanics (Jo et al. 2008). The platform walks users through several steps, starting with the PDB Reader, which prepares structures for analysis by converting a PDB file to CHARMM compatible files. The solvator tool adds water molecules to the system, solvating the molecule in a water box of a user-defined shape, such as cubic or octahedral. Ions, including KCl, NaCl, and $CaCl_2$, can be added at this point to obtain solution electrostatic neutrality. The quick molecular dynamics simulator step generates the input files for molecular dynamics simulations and instructions for running the simulation. The membrane builder tool simulates a realistic complex of proteins and membranes. It can be used to generate lipid bilayers with different types of lipid molecules, ions, pores, and bulk water, and can orient macromolecules through the use of the Orientations of Proteins in Membranes web server (Jo et al. 2008, Lomize et al. 2011).

In food science, molecular mechanics simulations have shown ovalbumin denaturation at high temperatures and resultant interactions in wheat dough (Sang et al. 2018), how β-lactoglobulin changes as the result of pressure treatment (Reznikov et al. 2011), and the nature of adsorption of β-lactoglobulin to oil interface during emulsification (Zare et al. 2016). It has also been shown that relatively long α-helices are converted to β-sheet structures under an applied strain (Qin & Buehler 2010), and this is consistent with experimental evidence from capillary rheology, which showed an increase in β-sheet structures with increasing levels of shear in pea protein (Beck et al. 2016). Lastly, gelation mechanisms of bovine serum albumin were investigated with the aid of molecular mechanics, showing a stabilizing impact of low concentrations of sodium dodecyl sulfate on protected helices as well as a hydrophobically driven aggregation that occurred below pH 3.5

(Baler et al. 2014, Nnyigide & Hyun 2020). Full utilization of these tools has yet to be realized in food science research, but as they become more accessible through programs like CHARMM-GUI, their use will likely expand into uncovering the nature of a variety of food processes.

7. CONCLUSION

Structural bioinformatics is a rich field that shows promise for food product developers, ingredient designers, and researchers. By using tools to analyze amino acid sequences and 3D models of protein structure, it is possible to draw strong relationships between models and key factors in food functionality, such as surface hydrophobicity, secondary structure, and zeta potential. These physicochemical relationships help create strong links between bioinformatics models and functional properties of food, including emulsification, foaming, gelation, and solubility. Future directions of research may be aided by developing databases of food-relevant proteins and their experimental characteristics. This would allow for more specific data science research for food science and would show greater applicability in the proteins of most interest to food developers, such as plant-based proteins, dairy proteins, and other proteins commonly used in the food industry. Lastly, as computing power increases to access longer timescales, the use of molecular mechanic tools to simulate food matrices in a great level of detail can be developed. The development of such programs could allow food developers to better model phenomena such as product shelf life and interactions during food processing like extrusion or baking and make inferences about the textures of food.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work is part of the USDA-NIFA grant 1024316. We thank the William R. Scholle Endowment for support to the Scholle Chair in food processing to cover research expenditures. We also thank Purdue University's Ross Fellowship for providing financial support to Harrison Helmick for graduate student study. D.K. acknowledges support from the National Institutes of Health (R01GM133840, 3R01GM133840-02S1) and the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, DBI2003635).

LITERATURE CITED

Acland A, Agarwala R, Barrett T, Beck J, Benson DA, et al. 2013. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41(D1):D8–20

Aguilar-Toalá JE, Hall FG, Urbizo-Reyes UC, Garcia HS, Vallejo-Cordoba B, et al. 2019. In silico prediction and in vitro assessment of multifunctional properties of postbiotics obtained from two probiotic bacteria. *Probiotics Antimicrob. Proteins* 12(2):608–22

Ainis WN, Boire A, Solé-Jamault V, Nicolas A, Bouhallab S, Ipsen R. 2019. Contrasting assemblies of oppositely charged proteins. *Langmuir* 35(30):9923–33

Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, et al. 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13(6):3031–48

Alizadeh-Pasdar N, Li-Chan ECY. 2000. Comparison of protein surface hydrophobicity measured at various pH values using three different fluorescent probes. *J. Agric. Food Chem.* 48(2):328–34

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science* 181(4096):223–30

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):871–76

Baler K, Martin OA, Carignano MA, Ameer GA, Vila JA, Szleifer I. 2014. Electrostatic unfolding and interactions of albumin driven by pH changes: a molecular dynamics study. *J. Phys. Chem. B* 118(4):921–30

Barth A. 2007. Infrared spectroscopy of proteins. *Biochim. Biophys. Acta* 1767(9):1073–101

Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 32(Suppl. 1):D120–21

Beck S, Knoerzer K, Sellahewa J, Emin A, Arcot J. 2016. Effect of different heat-treatment times and applied shear on secondary structure, molecular weight distribution, solubility and rheological properties of pea protein isolate as investigated by capillary rheometry. *J. Food Eng.* 208:66–76

Becktel WJ, Schellman JA. 1987. Protein stability curves. *Biopolymers* 26(11):1859–77

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28(1):235–42

Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, et al. 2016. The SWISS-MODEL repository—new features and functionality. *Nucleic Acids Res.* 45(D1):D313–19

Brazas MD, Yamada JT, Ouellette BFF. 2010. Providing web servers and training in bioinformatics: 2010 update on the bioinformatics links directory. *Nucleic Acids Res.* 38(Suppl. 2):W3–6

Brown SD, Babbitt PC. 2014. New insights about enzyme evolution from large scale studies of sequence and structure relationships. *J. Biol. Chem.* 289(44):30221–28

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2008. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 37(Suppl. 1):D233–38

Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, et al. 2005. The Amber biomolecular simulation programs. *J. Comput. Chem.* 26(16):1668–88

Chakravorty A, Jia Z, Li L, Alexov E. 2017. A new DelPhi feature for modeling electrostatic potential around proteins: role of bound ions and implications for zeta-potential. *Langmuir* 33(9):2283–95

Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, et al. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D* 66(1):12–21

Christoffer C, Chen S, Bharadwaj V, Aderinwale T, Kumar V, et al. 2021. LZerD webserver for pairwise and multiple protein-protein docking. *Nucleic Acids Res.* 49(W1):W359–65

Csaba G, Birzele F, Zimmer R. 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.* 9:23

Dianda N, Rouf TB, Bonilla J, Hendrick V, Kokini J. 2019. Effect of solvent polarity on the secondary structure, surface and mechanical properties of biodegradable kafirin films. *J. Cereal Sci.* 90:102856

Dickie AM, Kokini JL. 1983. An improved model for food thickness from non-Newtonian fluid mechanics in the mouth. *J. Food Sci.* 48(1):57–61

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32(Suppl. 2):W665–67

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755–63

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39(Suppl. 2):W29–37

Fiser A, Šali A. 2003. Modeller: generation and refinement of homology-based protein structure models. In *Methods in Enzymology*, Vol. 374, ed. CW Carter Jr., RM Sweet, pp. 461–91. Cambridge, MA: Academic

Garcia-Moreno P, Gregersen S, Nedamani E, Olsen T, Maracatili P, et al. 2020. Identification of emulsifier potato peptides by bioinformatics: application to omega-3 delivery emulsions and release from potato industry side streams. *Sci. Rep.* 10:690

Gauthier J, Vincent AT, Charette SJ, Derome N. 2018. A brief history of bioinformatics. *Brief. Bioinform.* 20(6):1981–96

Goodman RE, Ebisawa M, Ferreira F, Sampson HA, van Ree R, et al. 2016. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Mol. Nutr. Food Res.* 60(5):1183–98

Goormaghtigh E, Gasper R, Bénard A, Goldsztein A, Raussens V. 2009. Protein secondary structure content in solution, films and tissues: redundancy and complementarity of the information content in circular dichroism, transmission and ATR FTIR spectra. *Biochim. Biophys. Acta* 1794(9):1332–43

Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18(15):2714–23

Guldiken B, Stobbs J, Nickerson M. 2021. Heat induced gelation of pulse protein networks. *Food Chem.* 350:129158

Gupta JK, Adams DJ, Berry NG. 2016. Will it gel? Successful computational prediction of peptide gelators using physicochemical properties and molecular fingerprints. *Chem. Sci.* 7(7):4713–19

Haas J, Barbato A, Behringer D, Studer G, Roth S, et al. 2018. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins Struct. Funct. Bioinform.* 86(S1):387–98

Heldt CL, Zahid A, Vijayaragavan KS, Mi X. 2017. Experimental and computational surface hydrophobicity analysis of a non-enveloped virus and proteins. *Colloids Surfaces B* 153:77–84

Helmick H, Hartanto C, Bhunia A, Liceaga A, Kokini JL. 2021a. Validation of bioinformatic modeling for the zeta potential of vicilin, legumin, and commercial pea protein isolate. *Food Biophys.* 16:474–83

Helmick H, Hartanto C, Ettestad S, Liceaga A, Bhunia AK, Kokini JL. 2023. Quantitative structure-property relationships of thermoset pea protein gels with ethanol, shear, and sub-zero temperature pretreatments. *Food Hydrocoll.* 135:108066

Helmick H, Turasan H, Yildirim M, Bhunia A, Liceaga A, Kokini J. 2021b. Cold denaturation of proteins: where bioinformatics meets thermodynamics to offer a mechanistic understanding: pea protein as a case study. *J. Agric. Food Chem.* 26(22):6339–50

Hirata F, Sugita M, Yoshida M, Akasaka K. 2018. Perspective: structural fluctuation of protein and Anfinsen's thermodynamic hypothesis. *J. Chem. Phys.* 148(2):020901

Honig B, Nichols A. 1992. Classical electrostatics in biology and chemistry. *Science* 268(5214):1144–49

Hou Q, Kwasigroch JM, Rooman M, Pucci F. 2019. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* 36(5):1445–52

Hsin J, Arkhipov A, Yin Y, Stone JE, Schulten K. 2008. Using VMD: an introductory tutorial. *Curr. Protoc. Bioinform.* 24(1):5.7.1–5.7.48

Hubbard TJP, Murzin AG, Brenner SE, Chothia C. 1997. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.* 25(1):236–39

Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14(1):33–38

Jo S, Kim T, Iyer VG, Im W. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29(11):1859–65

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36(Suppl. 2):W5–9

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–89

Jurrus E, Engel D, Star K, Monson K, Brandi J, et al. 2018. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* 27(1):112–28

Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, et al. 2010. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* 110(10):5714–89

Keller R. 2018. Identification of potential lipid binding regions in cereal proteins and peptides with the use of bioinformatics. *J. Cereal Sci.* 80:128–34

Klemmer K, Stone W, Nickerson M. 2010. Complex coacervation of pea protein isolate and alginate polysaccharides. *Food Chem.* 130(3):710–15

Knudsen M, Wiuf C. 2010. The CATH database. *Hum. Genom.* 4:207–12

Kozlowski LP. 2016. Isoelectric point for PDB proteins Dec 2015. *RepOD*. **https://doi.org/10.18150/repod.1549954**

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. 2021. Critical assessment of methods of protein structure prediction (CASP)—round XIV. *Proteins Struct. Funct. Bioinform.* 89(12):1607–17

Li R, Wang X, Liu J, Cui Q, Wang X, et al. 2019. Relationship between molecular flexibility and emulsifying properties of soy protein isolate-glucose conjugates. *J. Agric. Food Chem.* 67(14):4089–97

Liang H-N, Tang C-H. 2013. Emulsifying and interfacial properties of vicilins: role of conformational flexibility at quaternary and/or tertiary levels. *J. Agric. Food Chem.* 61(46):11140–50

Lienqueo ME, Mahn A, Asenjo JA. 2002. Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. *J. Chromatogr. A* 978(1):71–79

Lin D, Lu W, Kelly AL, Zhang L, Zheng B, Miao S. 2017. Interactions of vegetable proteins with other polymers: structure-function relationships and applications in the food industry. *Trends Food Sci. Technol.* 68:130–44

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. 2011. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 40(D1):D370–76

Malhotra A, Coupland JN. 2004. The effect of surfactants on the solubility, zeta potential, and viscosity of soy protein isolates. *Food Hydrocoll.* 18(1):101–8

Mirdita M, Konstantin S, Yoshitaka M, Lim H, Sergey O, Martin S. 2022. ColabFold: making protein folding accessible to all. *Nat. Methods* 19(6):679–82

Miyazawa S, Jernigan RL. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256(3):623–44

Mooers BHM. 2020. Shortcuts for faster image creation in PyMOL. *Protein Sci.* 29(1):268–76

Moro A, Gatti C, Delorenzi N. 2001. Hydrophobicity of whey protein concentrates measured by fluorescence quenching and its relation with surface functional properties. *J. Agric. Food Chem.* 49(10):4784–89

Nakai S. 1983. Structure-function relationships of food proteins: with an emphasis on the importance of protein hydrophobicity. *J. Agric. Food Chem.* 31(4):676–83

Nnyigide OS, Hyun K. 2020. The protection of bovine serum albumin against thermal denaturation and gelation by sodium dodecyl sulfate studied by rheology and molecular dynamics simulation. *Food Hydrocoll.* 103:105656

Ouellette RJ, Rawn JD. 2015. *Principles of Organic Chemistry*. Cambridge, MA: Academic

Patel B, Singh V, Patel S. 2019. Structural bioinformatics. In *Essentials of Bioinformatics*, Vol. 1, ed. NA Shaik, KR Hakeem, B Banaganapalli, R Elango, pp. 169–99. Cham: Springer

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25(13):1605–12

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, et al. 2021. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* 30(1):70–82

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26(16):1781–802

Pripp AH, Isaksson T, Stepaniak L, Sørhaug T, Ardö Y. 2005. Quantitative structure activity relationship modelling of peptides and proteins as a tool in food science. *Trends Food Sci. Technol.* 16(11):484–94

Pucci F, Kwasigroch JM, Rooman M. 2017. SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics* 33(21):3415–22

Qin Z, Buehler MJ. 2010. Molecular dynamics simulation of the α-helix to β-sheet transition in coiled protein filaments: evidence for a critical filament length scale. *Phys. Rev. Lett.* 104(19):198304

Rahman M, Browne JJ, Van Crugten J, Hasan MdF, Liu L, Barkla BJ. 2020. In silico, molecular docking and in vitro antimicrobial activity of the major rapeseed seed storage proteins. *Front. Pharmacol.* 11:1340

Rasheed F, Markgren J, Hedenqvist M, Johansson E. 2020. Modeling to understand plant protein structure-function relationships—implications for seed storage proteins. *Molecules* 25(4):873

Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9(2):173–75

Reznikov G, Baars A, Delgado A. 2011. The initial stage of high-pressure induced β-lactoglobulin aggregation: the long-run simulation. *Int. J. Food Sci. Technol.* 46(12):2603–10

Robertson AD, Murphy KP. 1997. Protein structure and the energetics of protein stability. *Chem. Rev.* 97(5):1251–68

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, et al. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* 45(D1):D271–81

Salgado JC, Rapaport I, Asenjo JA. 2005. Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition? *J. Chromatogr. A* 1075(1):133–43

Sang S, Zhang H, Xu L, Chen Y, Xu X, et al. 2018. Functionality of ovalbumin during Chinese steamed bread-making processing. *Food Chem.* 253:203–10

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* 31(13):3381–85

Seeliger D, de Groot BL. 2010. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J. Comput.-Aided Mol. Des.* 24(5):417–22

Shen M, Sali A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15(11):2507–24

Studer G, Rempfer C, Waterhouse A, Gumienny R, Haas J, Schwede T. 2020. QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics* 36(6):1765–71

Sun X, Arntfield S. 2010. Dynamic oscillatory rheological measurement and thermal properties of pea protein extracted by salt method: effect of pH and NaCl. *J. Food Eng.* 105(3):577–82

Tandang-Silvas MR, Cabanos CS, Peña LDC, De la Rosa APB, Osuna-Castro JA, et al. 2012. Crystal structure of a major seed storage protein, 11S proglobulin, from *Amaranthus hypochondriacus*: insight into its physico-chemical properties. *Food Chem.* 135(2):819–26

Tandang-Silvas MRG, Fukuda T, Fukuda C, Prak K, Cabanos C, et al. 2010. Conservation and divergence on plant seed 11S globulins based on crystal structures. *Biochim. Biophys. Acta* 1804(7):1432–42

Tandang-Silvas MRG, Tecson-Mendoza EM, Mikami B, Utsumi S, Maruyama N. 2011. Molecular design of seed storage proteins for enhanced food physicochemical properties. *Annu. Rev. Food Sci. Technol.* 2:59–73

Tang C-H. 2017. Emulsifying properties of soy proteins: a critical review with emphasis on the role of conformational flexibility. *Crit. Rev. Food Sci. Nutr.* 57(12):2636–79

Tao X, Huang Y, Wang C, Chen F, Yang L, et al. 2020. Recent developments in molecular docking technology applied in food science: a review. *Int. J. Food Sci. Technol.* 55(1):33–45

Tong X, Cao J, Sun M, Liao P, Dai S, et al. 2021. Physical and oxidative stability of oil-in-water (O/W) emulsions in the presence of protein (peptide): characteristics analysis and bioinformatics prediction. *LWT* 149:111782

Turasan H, Barber E, Malm M, Kokini J. 2017. Mechanical and spectroscopic characterization of crosslinked zein films cast from solutions of acetic acid leading to a new mechanism for the crosslinking of oleic acid plasticized zein films. *Food Res. Int.* 108:357–67

Turasan H, Kokini J. 2016. Advances in understanding the molecular structures and functionalities of biodegradable zein-based materials using spectroscopic techniques: a review. *Biomacromolecules* 18:331–54

Uberto R, Moomaw EW. 2013. Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. *PLOS ONE* 8(9):e74477

UniProt Consort. 2008. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 36(Suppl. 1):D190–95

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26(16):1701–18

Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, et al. 2010. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* 31(4):671–90

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50(D1):D439–44

Vidal-Limon A, Aguilar-Toalá JE, Liceaga AM. 2022. Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides. *J. Agric. Food Chem.* 70(4):934–43

Wallner B, Elofsson A. 2007. Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins Struct. Funct. Bioinform.* 69(S8):184–93

Walsh I, Pollastri G, Tosatto SCE. 2015. Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinform.* 17(5):831–40

Wang G, Dunbrack J, Roland L. 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589–91

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46(W1,2):W296–303

Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.* 28(1):31–36

Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, et al. 2018. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci*. 27(1):293–315

Wu C, Wang T, Ren C, Ma W, Wu D, et al. 2021. Advancement of food-derived mixed protein systems: interactions, aggregations, and functional properties. *Compr. Rev. Food Sci. Food Saf.* 20(1):627–51

Zare D, Allison JR, McGrath KM. 2016. Molecular dynamics simulation of β-lactoglobulin at different oil/water interfaces. *Biomacromolecules* 17(5):1572–81

Annual Review of Food Science and Technology

Volume 14, 2023

Contents

Errata

An online log of corrections to *Annual Review of Food Science and Technology* articles may be found at http://www.annualreviews.org/errata/food