OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Bioinformatics-Guided Pre-Formulation Screening of Plant-Protein Sauce Ingredients Using Nutritional Profiling, Allergen Sequence Similarity, and Control-Ligand Docking

Arjun Dugar and Nirupma Singh
arjun.a.dugar@gmail.com, nirupmajadaun21@gmail.com

ABSTRACT

The increasing global demand for protein-rich plant-based foods creates a need for early-stage methods that compare nutritional composition and identify sequence-similarity signals relevant to potential allergen cross-reactivity. This study developed a computational pre-formulation framework to prioritize candidate ingredients for subsequent formulation and experimental testing; no physical sauce prototype was produced or experimentally characterized. Twenty-three plant-based ingredients were evaluated using nutritional and amino-acid composition data from the USDA FoodData Central SR Legacy dataset. Eighteen ingredients produced evaluable AllergenOnline Sliding 80-mer outcomes, while five had insufficient sequence or comparable-match information. Fourteen evaluable ingredients met the RED criterion of at least 35% identity across an 80-amino-acid window, two were classified YELLOW, and mung beans and kidney beans were classified GREEN because no qualifying match was identified for the single representative sequence queried under this criterion. A GREEN classification does not establish absence of allergenicity or clinical safety. Separately, thirteen known digestive enzyme-control inhibitor pairs were docked with AutoDock Vina to benchmark execution of the docking module; Vina scores ranged from -8.785 to -5.115 kcal/mol. Because the proposed sauce proteins and phytochemicals were not docked, these control-ligand scores do not predict ingredient-enzyme binding, enzyme inhibition, digestibility, or biological effects of the proposed formulation. The framework therefore supports computational pre-formulation prioritization rather than demonstration of product safety, digestive performance, or a completed sauce design.

Keywords: Plant-protein ingredients; Pre-formulation screening; Amino-acid composition; Allergen sequence similarity; Control-ligand docking

INTRODUCTION

A protein-rich sauce was researched and developed as a medium to increase protein intake, while not compromising on texture and long-term stability. The present study did not manufacture or experimentally test a sauce; instead, published evidence on plant-protein solubility and aggregation was used to define pre-formulation considerations for candidate ingredients. Higher protein solubility promoted a smoother mouthfeel, whereas aggregation within the same system reduced structural uniformity and promoted phase separation during storage. (Grasberger et al. 2024) High-protein condiments and sauces allow for more practical protein enrichment systems, especially in our increasingly convenience-fuelled world. In India and globally, plant-based protein fortification has gained popularity due to sustainability constraints and allergen considerations connected to animal protein sources. Previous research also reported that pea protein isolate demonstrated lower solubility and high aggregation in acidic conditions, resulting in lower quality dispersion and reduced long term stability. (Chen, Hall, and Moraru 2023) However, completely soluble protein fractions do not correlate with the strongest mechanical strength of interfacial layers, which means that the optimal balance between solubility and aggregation is necessary for forming stable emulsions. Follow-up work has verified that both the delivery and structure in these applications are directly related to plant protein function, especially in an acidic, high-solids environment as found in sauces. The overall significance of this work is that successful formulation needs tight control of both interfaces and bulk protein. (Chen, Hall, and Moraru 2023; Pușcaș et al. 2025)

One strategy that had been extensively researched to improve emulsion stability using plant proteins was the formation of protein-polysaccharide complexes. These complexes enhanced continuous phase viscosity and interface elasticity, thus hindering droplet movement and creaming and phase separation rates. High interfacial elasticity and steric repulsion also resulted in smaller droplet sizes, which improved visual uniformity as well as stability during storage. These improvements in macroscopic stability were associated with changes in interfacial mechanics, alongside bulk rheological effects. (Farooq, Ahmad, and Abdullah 2022) Emulsion behaviour was also strongly influenced by protein identity; different plant proteins formed interfaces with different elasticities and resistance to oiling-off. Ionic strength and pH also shifted stability, with near isoelectric conditions resulting in maximum viscosity. Moreover, viscosity rankings in both potato and faba protein isolates remained consistent across pH conditions under high salt concentrations of around 5.5% NaCl. (Cao, Corstens, and Schroën 2024) Processing conditions also played a role in the stability of the product. High-shear homogenisation resulted in smaller droplets and higher physical stability in dressing-type emulsions. However, it did not significantly reduce oxidation, revealing that there is no clear link between mechanical and chemical stability. In addition, this decoupling illustrates the finding that upgrading the microstructure of emulsion alone fails to guarantee oxidative stability for protein-rich system emulsions. (Yang, Chen, and Mis Solval 2024)

A plethora of analytical techniques like rheology, particle-size analysis, FTIR, DSC, and microscopy have been used to study the relationship between molecular protein changes and macroscopic stability

outcomes. With these approaches, correlations were possible among transitions in secondary structure, interfacial elastic properties, and macroscopic stability measures like viscosity and phase separation. (Schneider, Bu, and Ismail 2023; Mandliya et al. 2022) Structural modification methods, such as biologically mediated processing, have been demonstrated to improve hydrationability, solubility, and inhibit aggregation by specific changes in protein structures. Storage studies quantified lipid oxidation using peroxide value (PV), iodine value (IV), and TBARS. Chemical degradation was quantified independently of physical separation. During higher temperature storage conditions, a dramatic increase in oxidation was noted which led to quality deterioration despite the limited physical separation. Changes in mixing intensity affected physical stability but did not proportionally reduce oxidation. Improvements in physical stability occurred independently of oxidative degradation. (Yang, Chen, and Mis Solval 2024) Protein safety and functionality were also analysed in terms of sequence- and structure-dependent biological interactions. Bioinformatics-based screening methods served as a tool to identify allergenicity and structural risk during the initial stages of protein selection and modification. Sequence homology analysis, motif searches, and in silico digestion models made possible early selection for proteins with increased propensity for allergenicity or off-target biological activity, which could be problematic in formulation. Computational screening was incorporated alongside experimental formulation during early-stage protein selection. (Ferrari et al. 2026)

Long-term storage studies highlighted that the stability of plant-protein products reduced over time due to ongoing biochemical changes. Proteolysis and the release of amino acids during storage were linked to pH drift over time. This pH drift resulted in increased aggregation and sedimentation, which worsened the consistency and uniformity of the sauce. (Roland et al. 2024) In a particular study, microbial growth and survival were modelled as functions of pH, water activity, temperature and time. While lower pH and water activity kept it under control, higher temperatures increased growth within permissive ranges. This directly influenced product shelf life. Predictive models enabled quantitative shelf-life planning, which was an improvement from prolonged empirical storage trials. Such models allow formulation decisions to be evaluated without reliance on prolonged empirical storage trials, which are impractical during early-stage product development. (Stavropoulou and Bezirtzoglou 2019) Existing literature has largely examined solubility control, emulsion stability, chemical degradation, microbial safety, and protein safety screening in isolation. Nevertheless, these conditions are normally assessed in relation to one another independently, and there is little research exploring the interactions between them in one complex food system. Accordingly, the present study evaluated whether nutritional profiling, allergen sequence-similarity screening, and a control-ligand docking benchmark could be integrated as an early-stage ingredient-screening framework. Physical and chemical stability of an actual sauce was not tested and remains a subject for subsequent formulation experiments.

METHODOLOGY

DATA COLLECTION

Data relating to the nutrition and composition of the ingredients has been obtained from FoodData Central, operated by the USDA, with greater emphasis given to the SR Legacy Foods data, owing to the standardized values for the composition of foods and nutrient depth, compared to the other data sets, such as Branded Foods and Foundation Foods. Ingredients have been selected to resemble the actual formulation of the protein-enriched sauce, such as legumes and pulses for the protein content, grain and pseudo-cereal-based staples for the bulk solid content, oilseed-based ingredients for protein delivery, and finally, the use of polysaccharide-based structuring agents, although at the formulation level, owing to their proven use for viscosity enhancement and protein-polysaccharide interactions, as proven in the literature, rather than relying on them as data sources within FoodData Central. Ingredients were selected in their raw or minimally processed form to obtain values for their composition rather than relying on formulation data commonly found within processed products.

For each raw ingredient and flour, the dataset recorded the focus areas, which included protein content (in grams per 100 grams) and amino acid composition, with comparisons being drawn on the basis of the composition on a per gram of protein basis. The emphasis on amino acid composition data stemmed from the direct relevance of such composition differences to the protein functionalities of interest in the context of the specific project requirements. Along with the raw ingredient and flour composition data, the composition data of protein isolate were also recorded, as they represent the industrial protein fractions, which are more applicable to high-protein sauce formulation compared to the raw ingredient or flour composition data. The protein isolate composition data were recorded as a separate category, as the elimination of non-protein components such as starch and fiber results in changes to the protein content, which affects the protein functional properties with respect to solubility and aggregation in an acidic environment. The simultaneous collection of raw ingredient/flour composition and protein isolate composition data will allow for the controlled comparison of the two data sets, thereby providing the basis for the selection and screening of the raw ingredient, which is in line with the emphasis on the balance between nutritional content and protein functional properties.

ALLERGEN PROFILING

For allergen profiling of each of the candidate ingredients, AllergenOnline was used to query for potential cross-reactivity with known food allergens. For each candidate ingredient, one representative protein sequence was selected from UniProt and obtained in FASTA format. The sequence was analysed using the AllergenOnline Sliding 80-mer search with the default criterion of at least 35% identity across an 80-amino-acid window. A qualifying match was treated as a conservative screening flag warranting further assessment; it was not interpreted as evidence that an ingredient was clinically allergenic. Conversely, absence of a qualifying match was not interpreted as evidence of non-allergenicity or safety.

(FAO/WHO 2001). This approach is accepted as the first-tier bioinformatics approach for determining potential allergenicity of novel and modified food proteins (Ferrari et al. 2026).

For each of the candidate ingredients, the parameters used were: query protein name (from UniProt), top hit allergen name and accession number (from AllergenOnline), maximum percentage identity over 80-mer window, alignment length in amino acids, and E-value or score. Operational categories were defined as RED when the search returned at least 35% identity across an 80-amino-acid window, YELLOW for reported sub-threshold or borderline similarity, and GREEN when no qualifying match was identified for the representative sequence. Ingredients for which no UniProt sequence was available were not included in the flag level summary.

MOLECULAR DOCKING

The docking module was benchmarked using thirteen human digestive or brush-border enzymes paired with known pharmacological control inhibitors. AutoDock Vina (Trott and Olson 2010) was executed through the Tamarind Bio interface. This analysis evaluated whether the computational workflow could process established enzyme-control pairs and return rankable docking poses and scores. It did not test the proposed sauce ingredients, food-derived proteins, peptides, or food-relevant phytochemicals. The enzyme panel comprised pepsin A (1PSO), gastricsin/pepsin C (1AVF), trypsin-1/PRSS1 (1TRN), trypsin-2/PRSS2 (7Z9F), mesotrypsin/PRSS3 (9BOT), chymotrypsin C/CTRC (4H4F), pancreatic alpha-amylase (1B2Y), pancreatic lipase (2PPL), a human elastase-family surrogate (6SMA), carboxypeptidase A1 (2V77), carboxypeptidase B/CPB1 (1ZLI), aminopeptidase N/CD13 (4FYR), and DPP4/CD26 (2P8S). Enzyme structures were obtained from the RCSB Protein Data Bank in PDB format, and control-ligand structures were obtained from PubChem in 3D SDF format.

For each enzyme-control pair, the enzyme PDB file and control-ligand SDF file were uploaded to the AutoDock Vina interface. A whole-protein search space was used to avoid prespecifying a binding site. The lowest, or most negative, Vina score was retained in kcal/mol to three decimal places. These scores are computational ranking estimates and should not be interpreted as experimentally measured binding free energies, inhibitor potency, enzyme inhibition, digestive efficiency, or biological activity.

RESULTS

DATA AGGREGATION

The nutritional composition data for the ingredients of interest were obtained from USDA FoodData Central, SR Legacy Foods, for the following twenty-three ingredients of interest: lentils, pink or red, raw (FDC 174284); chickpeas, mature seeds, raw (FDC 173756); peas, green, split, raw (FDC 172428); mung beans, mature seeds, raw (FDC 174256); mungo beans/black gram, mature seeds, raw (FDC 174259); beans, kidney, red, raw (FDC 173744); soy flour, full-fat, raw (FDC 174273); peanut flour, defatted (FDC 174267); quinoa, uncooked (FDC 168874); amaranth grain, uncooked (FDC 170682); buckwheat flour,

whole-groat (FDC 170687); oat flour, partially debranned (FDC 169741); rice flour, brown (FDC 168898); sorghum flour, whole-grain (FDC 168943); millet flour (FDC 172023); seeds, sunflower seed flour, partially defatted (FDC 170156); seeds, flaxseed (FDC 169414); soy protein isolate (FDC 174276); seaweed/spirulina, dried (FDC 170495); nuts, walnuts (FDC 170187); nuts, almonds (FDC 170567); seeds, chia seeds, dried (FDC 170554); and seeds, pumpkin and squash seed kernels, dried (FDC 170556). Xanthan gum, guar gum, and aquafaba powder were not SR Legacy Foods entries and thus excluded.

For each constituent, the following parameters were determined: total protein content (g/100g), and the content of individual amino acids normalized to mg/g protein: histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, valine, cystine, and tyrosine, along with the calculated sums methionine + cystine and phenylalanine + tyrosine. A resolution at the amino acid level was chosen because these values are used directly to calculate protein quality indices and predict protein structure behaviors relevant to sauce stability.

ALLERGEN PROFILING

Among the 23 ingredients considered, 18 produced evaluable Sliding 80-mer screening outcomes, while five ingredients (spirulina, walnuts, almonds, chia seeds, and pumpkin/squash seed kernels) were classified as having insufficient data because an adequate representative sequence or comparable AllergenOnline result was unavailable. Fourteen evaluable ingredients met the RED criterion of at least 35% identity across an 80-amino-acid window: lentils (Len c 1.0101, 100%); chickpeas (provicilin, 75.5%); peas (vicilin, 85.4%); mungo beans (vignain, 54.2%); soy flour (glycinin G1, 99.2%); peanut flour (Ara h 1.0101, 100%); quinoa (11S seed storage globulin, 48.6%); amaranth (11S globulin seed storage protein, 50.7%); buckwheat (major allergenic storage protein, 91.1%); oat flour (avenin-E, 54.0%); rice flour (glutelin type-B 1, 46.1%); sunflower seed flour (albumin-8, 100%); flaxseed (conlinin, 87.0%); and soy protein isolate (glycinin G1, 99.2%). Sorghum flour (26.4%) and millet flour (33.9%) were classified YELLOW. Mung beans and kidney beans were classified GREEN because no qualifying match was detected for the single representative sequence queried. These GREEN outcomes do not establish that the ingredients are non-allergenic or clinically safe, and the insufficient-data cases were not treated as negative results. Summary: 14 RED, 2 YELLOW, 2 GREEN, and 5 insufficient data.

MOLECULAR DOCKING ANALYSIS

Molecular docking was performed for thirteen human digestive enzyme-known control inhibitor pairs as a benchmark of the computational docking module. Table 1 reports the lowest AutoDock Vina score for each control pair. The tested ligands were pharmacological controls and were not ingredients or phytochemicals from the proposed sauce. Therefore, the scores describe only the modelled control enzyme-ligand pairs and cannot be used to infer how candidate food ingredients bind to digestive enzymes, inhibit enzyme activity, or affect digestion.

The five enzyme-control pairs with the most negative Vina scores were:

1. Chymotrypsin C/CTRC (PDB code: 4H4F), in combination with Chymostatin (Figure 1) (PubChem CID: 443119), had a Vina score of -8.785 kcal/mol. Chymostatin is known as a peptide aldehyde-based inhibitor of chymotrypsin-like serine proteases, which interfere with hydrolysis at aromatic and large hydrophobic amino acid sites, i.e., phenylalanine, tyrosine, and tryptophan. Chymostatin is a pharmacological control and not a candidate ingredient, so this score does not describe how any proposed ingredient would interact with CTRC. It records only that the docking module returned a rankable pose for this control pair. The FDC amino acid data, considered separately, show that the ingredients with the highest levels of phenylalanine and tyrosine per gram of protein are seeds, chia seeds, dried (95.9 mg/g protein; FDC 170554); seeds, pumpkin and squash seed kernels, dried (93.4 mg/g protein; FDC 170556); seaweed, spirulina, dried (93.2 mg/g protein; FDC 170495); and peanut flour, defatted (92.3 mg/g protein; FDC 174267). These ingredients are known to have high levels of storage proteins with high levels of aromatic amino acids, e.g., 2S albumins, 11S globulins, and phycobiliproteins, which is the compositional reason they would be carried forward. Whether that composition produces differential susceptibility to chymotrypsin-type proteolysis was not tested here, and establishing it would require docking or enzymatic assays run on the ingredient proteins themselves. (Kårlund et al. 2021).

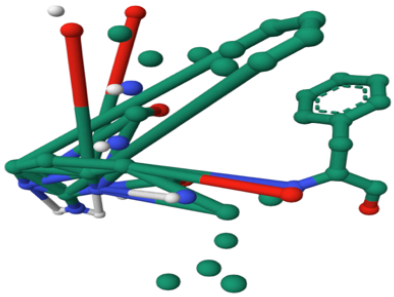

Figure 1. Three-dimensional structure of Chymostatin, the control ligand corresponding to the highest-affinity enzyme-ligand pair identified for Chymotrypsin C (CTRC; PDB: 4H4F). This molecule is a pharmacological control ligand used to benchmark the docking workflow and is not a proposed sauce ingredient. This figure shows the chemical structure of Chymostatin in ball-and-stick form. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, and hydrogen in white. Chymostatin was used as the control ligand in docking against Chymotrypsin C and showed a Vina score of -8.785 kcal/mol. The figure is included as one of the five highest-affinity enzyme-ligand pairs identified in the docking analysis.

2. DPP4/CD26 (PDB ID: 2P8S) in combination with Sitagliptin (Figure 2) (PubChem CID: 4369359) returned a Vina score of -8.368 kcal/mol. Sitagliptin is a clinically validated inhibitor of DPP4, having a fluorophenyl-triazolopyrazine pharmacophore, and its high affinity for DPP4 is consistent with the enzyme's role as a brush border exopeptidase, cleaving N-terminal dipeptides from proline or alanine-containing peptides. DPP4 activity, in the context of protein digestion, is particularly relevant to legume storage proteins, which are characterized by their high proline content in their vicilin and legumin subunits. Among the legume sources of the FDC ingredient, the major sources of protein in the formulation, lentils, pink or red, raw (FDC 174284; 23.9 g protein/100 g), chickpeas, mature seeds, raw (FDC 173756; 20.5 g protein/100 g), and peas, green, split, raw (FDC 172428; 23.1 g protein/100 g), are all rich in these proline-rich 7S and 11S storage proteins. Their behavior under DPP4-mediated peptide processing was not modeled, since Sitagliptin and not an ingredient protein was the docked ligand. (Kårlund et al. 2021).

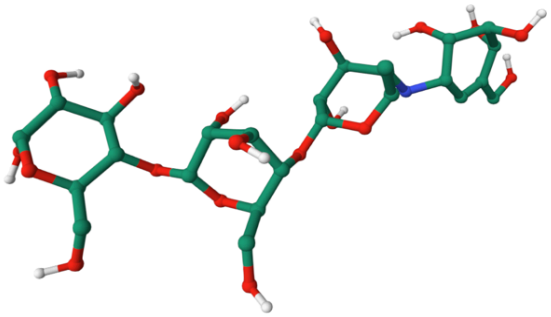

Figure 2. Three-dimensional structure of Sitagliptin, the control ligand corresponding to one of the five highest-affinity enzyme-ligand pairs identified for DPP4/CD26 (PDB: 2P8S). This molecule is a pharmacological control ligand used to benchmark the docking workflow and is not a proposed sauce ingredient. This figure shows the chemical structure of Sitagliptin in ball-and-stick form. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, hydrogen in white, and fluorine in light green. Sitagliptin was used as the control ligand in docking against DPP4/CD26 and showed a Vina score of -8.368 kcal/mol. The figure is included as one of the five highest-affinity enzyme-ligand pairs identified in the docking analysis.

3. Pancreatic α-Amylase (PDB: 1B2Y) in complex with Acarbose (Figure 3) (PubChem CID: 41774) has a Vina score of -8.278 kcal/mol. Acarbose is a pseudotetrasaccharide α-glucoside inhibitor of α-glucosidase. Acarbose is a pharmacological control ligand, so this score reflects how the docking workflow performed on the control pair. It says nothing about the cereal and pseudo-cereal components considered in this study. Starch-containing food components from the Food Data Central (FDC) database are: oat flour, partly debraned (ID: 169741; 16.9 g protein/100g); brown rice flour (ID: 168898; 7.23 g protein/100g); uncooked quinoa (ID: 168874; 14.1 g protein/100g); uncooked amaranth grain (ID: 170682; 13.6 g protein/100g); whole grain buckwheat flour (ID: 170687; 12.6 g protein/100g); whole grain sorghum flour (ID: 168943; 8.43 g protein/100g); and millet flour (ID: 172023; 10.8 g protein/100g). From the biochemical perspective, these components are characterized by the presence of

starch crystals surrounded by a protein matrix composed of prolamins and glutelins, whose gelatinization and enzymatic hydrolysis behavior would have to be established experimentally in any later formulation work. (Kårlund et al. 2021).

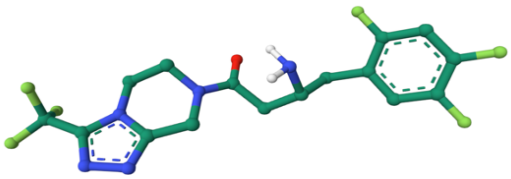

Figure 3. Three-dimensional structure of Acarbose, the control ligand corresponding to one of the five highest-affinity enzyme-ligand pairs identified for Pancreatic α-Amylase (PDB: 1B2Y). This molecule is a pharmacological control ligand used to benchmark the docking workflow and is not a proposed sauce ingredient. This figure shows the chemical structure of Acarbose in ball-and-stick form. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, and hydrogen in white. Acarbose was used as the control ligand in docking against Pancreatic α-Amylase and showed a Vina score of -8.278 kcal/mol. The figure is included among the five highest-affinity enzyme-ligand pairs identified in the docking analysis.

4. Human Elastase-family Surrogate (PDB 6SMA) + Sivelestat (Figure 4) (PubChem CID 107706) has a Vina score of -7.604 kcal/mol. Sivelestat is a selective benzenesulfonamide ester inhibitor of human neutrophil elastase. Serine elastases are known to cleave preferentially at aliphatic residue positions such as valine, alanine, leucine, and isoleucine. Based on FDC data, the amino acid residues with the highest content of valine in a gram of protein were present in seaweed, spirulina, dried (61.0 mg/g protein; FDC 170495); seeds, flaxseed (58.5 mg/g protein; FDC 169414); rice flour, brown (58.6 mg/g protein; FDC 168898); and seeds, sunflower seed flour, partially defatted (57.8 mg/g protein; FDC 170156). Biochemical classification of these proteins is that of albumins and globulins, storage proteins that are rich in exposed aliphatic residues. Their susceptibility to elastase cleavage during pancreatic digestion was not evaluated, since Sivelestat and not an ingredient protein was docked against the elastase-family surrogate. (Kårlund et al. 2021).

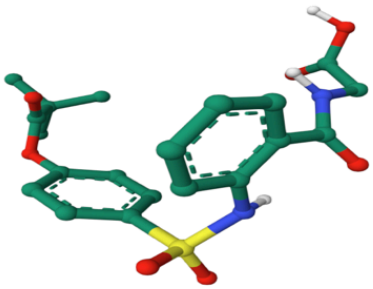

Figure 4. Three-dimensional structure of Sivelestat, the control ligand corresponding to one of the five highest-affinity enzyme-ligand pairs identified for the human elastase-family surrogate (PDB: 6SMA). This molecule is a pharmacological control ligand used to benchmark the docking workflow and is not a proposed sauce ingredient. This figure shows the chemical structure of Sivelestat in ball-and-stick form. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, hydrogen in white, and sulfur in yellow. Sivelestat was used as the control ligand in docking against the human elastase-family surrogate and showed a Vina score of -7.604 kcal/mol. The figure is included as one of the five highest-affinity enzyme-ligand pairs identified in the docking analysis.

5. The complex of Gastricsin/Pepsin C (PDB code 1AVF) and Pepstatin A (Figure 5) (PubChem CID 5478883) has a Vina score of -6.869 kcal/mol. The naturally occurring pentapeptide protease inhibitor is a product of actinomycetes and is a potent inhibitor of aspartic proteases, with two statine residues that mimic the peptide cleavage in the transition state. Gastricsin is a gastric aspartic protease with broad substrate specificity, initiating protein denaturation and hydrolysis in the acidic gastric medium after ingestion. Among the candidate ingredients, those with the highest overall protein density are soy protein isolate (FDC 174276; 88.3 g/100g), seaweed, spirulina, dried (FDC 170495; 57.5 g/100g), defatted peanut flour (FDC 174267; 52.2 g/100g), and sunflower seed flour, partially defatted (FDC 170156; 48.1 g/100g). Soy glycinin, phycobiliproteins, arachins, and sunflower albumin are reported in the literature as substrates for gastric aspartic proteases, which makes them candidates for later digestion work. No claim about digestive efficiency follows from the present docking data, which covered the Gastricsin-Pepstatin A control pair alone. (Ferrari et al. 2026; Kårlund et al. 2021).

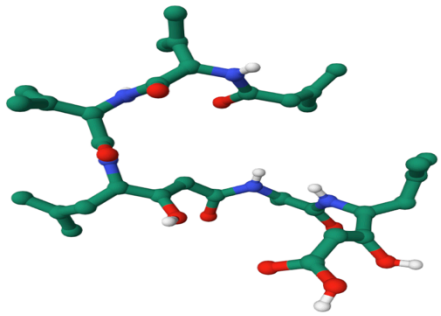

Figure 5. Three-dimensional structure of Pepstatin A, the control ligand corresponding to one of the five highest-affinity enzyme-ligand pairs identified for Gastricsin/Pepsin C (PDB: 1AVF). This molecule is a pharmacological control ligand used to benchmark the docking workflow and is not a proposed sauce ingredient. This figure shows the chemical structure of Pepstatin A in ball-and-stick form. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, and hydrogen in white. Pepstatin A was used as the control ligand in docking against Gastricsin/Pepsin C and showed a Vina score of -6.869 kcal/mol. The figure is included as one of the five highest-affinity enzyme-ligand pairs identified in the docking analysis.

Table 1. AutoDock Vina scores (kcal/mol) for thirteen digestive enzyme-pharmacological control ligand pairs. The listed ligands were used only to benchmark the docking workflow; none is a proposed sauce ingredient.

Enzyme (Target)	DB ID	Pharmacological Control Ligand	PubChem CID	Vina Score (kcal/mol)
Chymotrypsin C (CTRC)	4H4F	Chymostatin	443119	−8.785
DPP4 / CD26	2P8S	Sitagliptin	4369359	−8.368
Pancreatic α-Amylase	1B2Y	Acarbose	41774	−8.278
Human Elastase-family	6SMA	Sivelestat	107706	−7.604
Gastricsin / Pepsin C	1AVF	Pepstatin A	5478883	−6.869
Aminopeptidase N (CD13)	4FYR	Bestatin (Ubenimex)	72172	−6.717
Pepsin A	1PSO	Pepstatin A	5478883	−6.478
Trypsin-2 (PRSS2)	7Z9F	Benzamidine	2332	−6.167
Trypsin-1 (PRSS1)	1TRN	Benzamidine	2332	−6.045
Carboxypeptidase A1	2V77	Benzylsuccinic acid	3858	−5.622
Pancreatic Lipase	2PPL	Orlistat	3034010	−5.439
Carboxypeptidase B (CPB1)	1ZLI	GEMSA	3464	−5.213
Mesotrypsin (PRSS3)	9BOT	Benzamidine	2332	−5.115

DISCUSSION

The study integrated nutritional characterization, allergen sequence-similarity screening, and control-ligand docking as a computational pre-formulation framework for prioritizing candidate ingredients. The AllergenOnline results should be interpreted as conservative sequence-similarity flags rather than as direct evidence of allergenicity, cross-reactivity, or safety. Fourteen ingredients met the selected RED criterion and therefore warrant further evaluation. For the single representative sequences queried for mung beans and kidney beans, no qualifying match was identified; these GREEN classifications mean only that the selected Sliding 80-mer threshold was not met. They do not exclude allergenicity arising from other proteins, conformational or discontinuous epitopes, processing effects,

sensitization history, or mechanisms not captured by the search. Likewise, the five insufficient-data ingredients cannot be regarded as having negative allergenicity results.

The results obtained in the molecular docking studies further reinforce the importance of using both composition and enzyme specificity profiles. In this context, it is noted that the thirteen control enzyme-inhibitor pairs returned Vina scores in the order expected from the reported pharmacology of the control inhibitors. That ordering is the basis on which the docking module was judged to be executing correctly. Chymotrypsin C (CTRC) in complex with Chymostatin returned the most negative Vina score, at -8.785 kcal/mol, and Mesotrypsin with Benzamidine the least negative, at -5.115 kcal/mol. The composition data, treated as a separate line of evidence, identify chia seeds at 95.9 mg/g protein of combined Phe+Tyr content, pumpkin and squash seed kernels at 93.4 mg/g protein, and spirulina at 93.2 mg/g protein as the ingredients richest in the aromatic residues that chymotrypsin-type proteases act on. Lentils, chickpeas and split peas are the richest in the proline-containing legumin and vicilin subunits relevant to DPP4/CD26, an enzyme whose established role in the intestinal brush-border membrane is to cleave N-terminal dipeptides from proline-containing substrates. These observations mark where ingredient-level docking and digestion assays should be directed next. None of these ingredient proteins was docked, so the control-ligand scores cannot be read as predictions of their proteolytic susceptibility. (Kårlund et al. 2021)

The methodological strategies employed in this investigation are placed within the context of existing literature to both affirm their individual validity and articulate the specific contribution of the combined framework in relation to existing literature on protein sauces and protein emulsions. In terms of affirming the individual validity of the various methodological strategies employed in this investigation, the technique of allergen profiling, conducted via sequence homology analysis against existing allergen databases, is an exemplification of the bioinformatics-first approach espoused by a research group as the appropriate first-tier technique for assessing the allergenicity of novel and modified food proteins, owing to the limitations of traditional chemical toxicology frameworks in capturing protein-specific immunological hazards. In terms of affirming the individual validity of the various methodological strategies employed in this investigation, the computational docking technique used in this investigation is a validated technique for assessing the Vina score between small molecule ligands and macromolecular targets, with its scoring function being benchmarked against existing pharmacological datasets. (Trott and Olson 2010)

In contrast, past research on protein-sauce and emulsion systems only utilized physicochemical characterization techniques with no prior screening based on any bioinformatics approach. A research group studied Maillard-glycated pea protein isolate for acidic beverage development using a range of physicochemical characterization techniques, including differential scanning calorimetry, Fourier transform infrared spectroscopy, surface hydrophobicity, and solubility. However, they did not study allergenic potential or computer modeling of protein-enzyme interactions. Similarly, a researcher studied the interfacial characteristics of potato and faba bean protein isolates in high oil content emulsions using dilatational interfacial rheology and laser diffraction techniques. However, they did not study allergenic potential or computer modeling of protein-enzyme interactions. The present study differs in the addition

of a prior screening approach based on bioinformatics techniques, which was not done in the above-mentioned studies, so that the selection criteria draw on a sequence-similarity screen for potential allergenicity as well as on functional food chemistry.

Several limitations influence the results and suggest avenues for further investigation. Foremost, the molecular docking analysis was limited to pharmacological control inhibitors and was used only as a validation tool for the experimental platform. That is, the eleven food-relevant ligands identified in the experimental design, namely, phytic acid, tannic acid, gallic acid, quercetin, catechin, epicatechin, chlorogenic acid, caffeic acid, oleic acid, linoleic acid, and glycerophosphocholine, were not used in the molecular docking analysis against the thirteen-enzyme panel. Accordingly, the major goal of the study, namely, the investigation of the interactions between food-relevant phytochemicals and digestive enzymes, has not been addressed and constitutes the most important future direction in the field. Indeed, a full analysis would provide up to 143 enzyme food ligand Vina score values, thereby allowing a quantitative assessment of the potential of sauce-relevant food phytochemicals to modulate the activity of human digestive enzymes.

In the context of the five aforementioned ingredients, the lack of sufficient UniProt sequence information or the absence of similar matches in the AllergenOnline database meant that there was a lack of information in the safety assessment profile for the five ingredients of interest. This was particularly pertinent for the spirulina ingredient, which had the second highest protein content in the panel (57.5g/100g) and was thus a priority for further investigation in the context of the allergen profiling. However, in the future, the allergenicity assessment of the five ingredients would be conducted by utilizing alternative databases, such as the WHO/IUIS allergen nomenclature database, or alternative software tools, such as AllerHunter. One further limitation of the present study was the linear sequence-based assessment of the five ingredients, which did not take into consideration the presence of conformational or discontinuous epitopes that may be present in the proteins of interest. This would be significantly addressed in future studies by the inclusion of the AlphaFold tool, which would provide a much more detailed assessment of the allergenicity risks of the five ingredients of interest. Furthermore, the present study used a single UniProt accession per ingredient, which did not take into consideration the heterogeneity of the food proteins that may have different immunological profiles to the primary sequence used in the present study.

With respect to the experimental domain, the investigation was performed in silico only, without the development and characterization of the sauce prototype. Empirical verification of the docking predictions using enzyme inhibition kinetics, as well as the allergenicity indicators using IgE immunoblotting, would need to be performed before the in silico predictions could be used to inform the formulation decisions. Additionally, the SR Legacy Foods database did not contain the protein isolate data for most of the candidate legumes and cereals, only including the data for the soy protein isolate. Future studies should aim to use the protein isolate composition data, as it has functional differences compared to the whole food matrices due to the removal of the starch and fiber fractions, which affect the solubility and aggregation characteristics in acidic environments. With respect to the structuring agents identified in the formulation design using the in silico predictions, the xanthan gum, guar gum, and aquafaba powder

were not included in the nutritional and allergenicity analysis as they were not present in the SR Legacy Foods database. Although their functional characteristics in enhancing the viscosity as well as the interactions between the proteins and the polysaccharides are well established in the literature, their absence in the database limits the overall bioinformatics analysis. Future studies involving the prototyping of the sauce using the GREEN-flagged ingredients, namely the mung beans and the kidney beans, as the primary protein sources, in combination with the protein-polysaccharide structuring approach as well as the glycation approach using the Maillard reaction, as described in the literature, would provide a clear pathway for the future investigation as it pertains to the development of the food product.

CONCLUSION

This study established a bioinformatics-guided pre-formulation framework integrating USDA nutrient and amino-acid composition data, AllergenOnline sequence-similarity screening, and a control-ligand docking benchmark. Under the selected Sliding 80-mer criterion, the representative mung-bean and kidney-bean sequences produced no qualifying matches. This finding supports their provisional prioritization for further study but does not demonstrate that either ingredient is non-allergenic or clinically safe. Five ingredients remained unclassified because of insufficient data. Docking of thirteen known pharmacological enzyme-inhibitor pairs demonstrated execution of the computational module for those control pairs only. Because the proposed food-derived proteins and phytochemicals were not docked, the docking scores do not establish ingredient-enzyme binding, enzyme inhibition, digestibility, bioavailability, or biological effects. No physical sauce was produced or experimentally characterized. The framework therefore supports hypothesis generation and ingredient prioritization for subsequent computational analyses, formulation trials, physicochemical characterization, digestion assays, and allergenicity assessment rather than constituting a completed sauce design.

REFERENCES

1. Cao, J., M. Corstens, and K. Schroën. 2024. 'Interfacial and Bulk Properties of Potato and Faba Protein in Connection with Physical Emulsion Stability at Various pH Values and High Salt Concentrations', *Foods*, 13.
2. Chen, S., A. E. Hall, and C. I. Moraru. 2023. 'Functionality of pea protein isolate solutions is affected by reconstitution conditions', *J Food Sci*, 88: 4630-38.
3. FAO/WHO. 2001. Evaluation of Allergenicity of Genetically Modified Foods: Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, 22-25 January 2001 (Food and Agriculture Organization of the United Nations: Rome).
4. Farooq, S., M. I. Ahmad, and Abdullah. 2022. 'Interfacial rheology of sodium caseinate/high acyl gellan gum complexes: Stabilizing oil-in-water emulsions', *Curr Res Food Sci*, 5: 234-42.
5. Ferrari, A., E. Sánchez-Brunete, F. J. Moreno, M. Ardizzone, I. C. Dewhurst, T. Goumperis, and A. Fernández Dumont. 2026. 'What makes a food protein unsafe?', *Food Chem Toxicol*, 207: 115823.

6. Grasberger, K. F., F. W. Lund, A. C. Simonsen, M. Hammershøj, P. Fischer, and M. Corredig. 2024. 'Role of the pea protein aggregation state on their interfacial properties', *J Colloid Interface Sci*, 658: 156-66.
7. Kårlund, A., I. Paukkonen, C. Gómez-Gallego, and M. Kolehmainen. 2021. 'Intestinal Exposure to Food-Derived Protease Inhibitors: Digestion Physiology- and Gut Health-Related Effects', *Healthcare (Basel)*, 9.
8. Mandliya, Shubham, Anubhav Pratap-Singh, Siddharth Vishwakarma, Chandrakant G. Dalbhagat, and Hari N. Mishra. 2022. "Incorporation of Mycelium (Pleurotus eryngii) in Pea Protein Based Low Moisture Meat Analogue: Effect on Its Physicochemical, Rehydration and Structural Properties." In *Foods*, 2476.
9. Pușcaș, Andreea, Anda E. Tanislav, Andruța E. Mureșan, and Vlad Mureșan. 2025. "Emulgels as Fat-Replacing Systems in Biscuits Developed with Ternary Mixtures of Pea and Soy Protein Isolates and Gums." In *Gels*, 478.
10. Roland, Ida S., Thao T. Le, Tony Chen, Miguel Aguilera-Toro, Søren D. Nielsen, Lotte B. Larsen, and Nina A. Poulsen. 2024. "Storage Stability of Plant-Based Drinks Related to Proteolysis and Generation of Free Amino Acids." In *Foods*, 367.
11. Schneider, Alissa A., Fan Bu, and Baraem P. Ismail. 2023. 'Enhancement of pea protein solubility and thermal stability for acidic beverage applications via endogenous Maillard-induced glycation and chromatography purification', *Current Research in Food Science*, 6: 100452.
12. Stavropoulou, E., and E. Bezirtzoglou. 2019. 'Predictive Modeling of Microbial Behavior in Food', *Foods*, 8.
13. Trott, O., and A. J. Olson. 2010. 'AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading', *Journal of computational chemistry*, 31: 455-61.
14. Yang, B., J. Chen, and K. Mis Solval. 2024. 'Microencapsulated polyphenol extracts from Georgia-grown pomegranate peels delay lipid oxidation in salad dressing during accelerated and ambient storage conditions', *Food Sci Nutr*, 12: 370-84.

ABOUT THE AUTHOR

Arjun Dugar is a high school senior at The Cathedral and John Connon School, Mumbai in Mumbai, India. He enjoys learning biology and medical Bio-sciences, hoping to explore these subjects in a STEM-related degree in college too.