Pea and lentil 7S globulin crystal structures with comparative immunoglobulin epitope mapping

Kelly A. Robinson [1], Antony D. St-Jacques [1], Isabella D. Bakestani, Benjamin A.G. Beavington, Michele C. Loewen [*]

Aquatic and Crop Resources Development Research Center, National Research Council of Canada, 100 Sussex Drive, Ottawa, Ontario K1A 0R6, Canada

A B S T R A C T

Legumes represent an affordable high protein, nutrient dense food source. However, the vast majority of legume crops contain proteins that are known allergens for susceptible individuals. These include proteins from the 7S globulin family, which comprise a vast majority of seed storage proteins. Here, the crystal structures of 7S globulins from *Pisum sativum* L. (pea) and *Lens culinaris* Medicus (lentil) are presented for the first time, including pea vicillin and convicilin, and lentil vicilin. All three structures maintain the expected 7S globulin fold, with trimeric quaternary structure and monomers comprised of β-barrel N- and C-modules. The potential impact of sequence differences on structure and packing in the different crystal space groups is noted, with potential relevance to packing upon seed deposition. Mapping on the obtained crystal structures highlights significant Ig epitope overlap between pea, lentil, peanut and soya bean and significant coverage of the entire seed storage protein, emphasizing the challenge in addressing food allergies. How recently developed biologicals might be refined to be more effective, or how these seed storage proteins might be modified *in planta* to be less immuno-reactive remain challenges for the future. With legumes representing an affordable, high protein, nutrient dense food source, this work will enable important research in the context of global food security and human health on an ongoing basis.

1. Introduction

Legumes represent an affordable high protein, nutrient dense feed and food source (Ferreira et al., 2021). More than 81 million ha of legumes are grown worldwide each year, yielding more than 92 million tonnes of grain globally per annum (FAOSTAT, 2022). Increased production of legumes is anticipated to be an important target to ensure food and nutritional security for 9 billion people by 2050 (Dutta et al., 2022). However, the vast majority of widely farmed legume crops, including peanut (*Arachis hypogaea* L.), pea (*Pisum sativum* L.), lentil (*Lens culinaris* Medicus), chickpea (*Cicer arietinum* L.), common bean (*Phaseolus vulgaris* L.), mung bean (*Vigna radiate* (L.) Wilczek) and soya bean (*Glycine max* (L.) Merrill), have been shown to contain food proteins with potential allergenicity for susceptible individuals (reviewed in (Riascos et al., 2010)). These potential allergenic proteins span four different protein superfamilies including i) cupins (globulins) ii) prolamins (e.g. non-specific lipid transfer proteins), iii) profilins and iv)

pathogenesis related proteins (e.g. pollen Betv1). As consumption of legumes increases, it will be essential to understand the molecular basis of legume allergenicity and its cross-reactivity, to enable development of hypo-allergenic varieties and best pulse grain processing practices.

To date, one of the most highly characterized legume allergen families is the 7S-8S globulins (in the cupin superfamily), also referred to as vicilins. They form a major component of seed storage proteins and are classically known as a source of nutrition during the development and germination of seeds, although a variety of additional roles at different developmental time have also been characterized (Kesari et al., 2017; Kriz, 1999). Crystal structures are available for an array of legume vicilins including *A. hypogaea* (peanut; (Chruszcz et al., 2011)), and several beans: *P. vulgaris* (common; (Lawrence et al., 1994; Lawrence et al., 1990), *Canavalia ensiformis* (L.) de Candolle (jack; (Ko et al., 2000; Ko et al., 2001; McPherson, 2020)), *G. max* (soya; (Maruyama et al., 2001; Maruyama et al., 2003)), *V. radiate* (mung: (Itoh et al., 2006)) and *Vigna angularis* (Willdenow) Ohwi & Ohashi (adzuki; (Fukuda et al.,

2008)). The structures are highly conserved displaying a homo-trimeric structure, with each monomer divided into very similar N- and C- terminal modules (reviewed in (Kesari et al., 2017)). Each module is comprised of a conserved β-barrel core (jelly roll-like) domain and an extended loop region with several helices.

Considerable epitope mapping has also been conducted across many of these 7S vicilin targets. For example, in the case of peanut globulin Ara h 1, extensive linear immunoglobulin (Ig) peptide epitopes were mapped on an obtained crystal structure of the peanut globulin (Cabanos et al., 2011). This analysis highlighted a complex array of dissociated epitopes, in extended loops, in coils between the N- and C- terminal modules, and even in the less accessible beta-barrel domain core of each monomer. These epitopes generally become (more) buried upon trimer formation. Epitope mapping for other legume globulins is also available, including for pea Pis s 1 and lentil Len c 1, despite the lack of available crystal structures. In the case of Pis s 1, eleven distinct peptides were identified as high IgE binders, where mapping of the peptides on a homology model indicated distribution over large areas of the Pis s 1 molecular surface with no predominant localization (Popp et al., 2020). In the case of Len c 1, four epitopes were proposed along a continuous span of amino acids (residues 314–400; epitope id 107–135), where mapping again highlighted discontinuous distribution over the surface of a derived homology model (Vereda et al., 2010).

Here, the three-dimensional crystal structures of recombinantly produced core (mature) pea and lentil vicilins as well as pea convicilin are reported for the first time revealing a common trimeric quaternary structure, with each monomer comprised of N- and C- terminal modules, each in turn made up of conserved β-barrel cores. Crystal packing is considered in the context of potential relevance to seed deposition. Mapping of known immunoglobulin epitopes revealed overlap between pea, lentil, soya and peanut. The potential for predicting missing epitopes, or other novel epitopes that may contribute to the diversity of allergenicities across the broader family of 7S-8S globulin containing pulse crops, is discussed.

2. Materials and methods

2.1. Materials

All chemicals and other materials were purchased from MilliporeSigma unless otherwise indicated.

2.2. Gene synthesis

Pulse allergen DNA sequences for Pis s 1 (GenBank: AJ626897), Pis s 2 (GenBank: AJ276875.1) and Len c 1 (GenBank: AJ551424) were selected for gene synthesis. The sequences were trimmed of N- and C- terminal residues that made up signal peptides or unstructured ends as determined by homology modelling (PHYRE 2), including 6 C-terminal amino acids for both Len c 1 and Pis s 1 and 194 N-terminal and 24 C- terminal amino acids for Pis s 2. The final sequences were optimized, synthesized and cloned into the pET28b+ expression vector with *Nde*I and *Xho*I restriction sites to include N-terminal 6×His-tag, by Bio Basic Inc.

2.3. Recombinant protein expression and purification

Plasmids were transformed into BL21DE3 cells for expression in *Escherichia coli*. Cells were grown in 100 mL LB media supplemented with 50 μg/mL kanamycin at 37 °C until an OD_{600} of 0.6 was reached, where cultures were induced with 1 mM IPTG and grown at 25 °C for 17 to 20 h. Cells were harvested by centrifugation at 3750 rpm, 4 °C for 30 min. Cell pellets were resuspended in 1× lysis buffer (50 mM Na_2HPO_4, 500 mM NaCl, 10 mM imidazole, pH 8.0) supplemented with 3 U/mL benzonase, 0.1 mM PMSF and 1 mg/mL lysozyme and rotated end-over-end for 1 h at 4 °C followed by 10–15 sonication cycles at 30 % duty, 40

output power with 25 s on and 30 s off. The lysate was clarified by centrifugation at 17,000×*g* and 4 °C for 30 min. Proteins were purified by two-step purification consisting of nickel affinity (Ni-NTA, QIAGEN) and size exclusion chromatography. Clarified lysate was loaded on resin pre-equilibrated with 1× lysis buffer, followed by 3 washes of 4 column volumes (CV) with wash buffer (lysis buffer, with 20 mM imidazole). The protein was eluted in three 1 CV fractions with elution buffer (lysis buffer, with 250 mM imidazole) then pooled and concentrated using an Amicon ultrafiltration device (10 kDa cut-off, EMD Millipore) and filtered using 0.22 μm Millex filter (EMD Millipore) in preparation for size exclusion chromatography. A 16/60 HiLoad Superdex 200 column (Cytiva) was pre-equilibrated with 25 mM Tris base, 500 mM NaCl, 1 mM EDTA, pH 7.5 using an AKTA Purifier FPLC (Cytiva) and the protein was loaded then eluted using a flow rate of 1.0 mL/min. Fractions were pooled and concentrated as previously described until the desired protein concentration for crystallography was obtained. Protein concentration was determined using the Bradford Assay (Bio-Rad) and purity was assessed by 12 % SDS-PAGE stained with Coomassie Blue (Bio-Rad).

2.4. Protein crystallization

The protein samples were centrifuged at 12000 rpm for 1 min followed by filter sterilization (0.22 μm) and stored on ice. Crystallography screens conducted at room temperature, testing 192 conditions (JCSG MD 1–37 and PACT premier MD 1–29, Molecular Dimensions), were used to determine appropriate buffer conditions for the formation of crystals. 2 μL of protein samples were mixed with equal volume of reservoir solution and equilibrated against 750 μL of reservoir solution, using the hanging-drop vapour-diffusion method. Crystals were observed using a stereomicroscope (ZEISS Axiocam ERc 5 s) for up to 5 weeks. Crystallography buffers that resulted in crystals were expanded to find ideal conditions to produce diffraction quality crystals. Supplemental Table T1 summarizes the final crystallization condition for each pulse allergen. In order to obtain diffraction quality crystals for Len c 1, streak microseeding was performed, seeding from the crystal-containing donor drop into a fresh drop containing 50 % or 100 % of the original protein concentration of the donor drop. Crystals were either crushed then streaked, or only streaked from the donor drop, depending on the types of crystals that formed. Crystals were harvested and stored in liquid nitrogen with no cryoprotectant for Pis s 1 and Pis s 2, and 20 % glycerol for Len c 1.

2.5. Crystal diffraction, structure elucidation and analysis

Diffraction data was collected on the crystals using the synchrotron 08B1-1 beamline at the Canadian Light Source, Canadian Macromolecular Crystallography Facility. The data was reduced and scaled within CCP4i2 software (Potterton et al., 2003; Winn et al., 2011), using DIALS (Diffraction Integration for Advanced Light Sources; (Winter et al., 2022)), followed by molecular replacement using MOLREP (Vagin and Teplyakov, 1997) with a molecular model of β-conglycinin from soya bean (PDBID 1UIK, (Maruyama et al., 2003)) with 55–57 % identity to the lentil and pea 7S globulins. The protein structures were refined using REFMAC5 (Murshudov et al., 2011) and manual modelling performed using COOT (Emsley et al., 2010); refinement statistics can be found in Table 1, for each 7S globulin. Primary sequence analyses were carried out using Clustal Omega (Madeira et al., 2022). Structural images were produced using PyMol (Schrödinger), with structural alignments carried out using the PyMol 'align' function.

2.6. Comparative structural mapping of previously identified pea and lentil Ig binding epitopes

All figures depicting protein structures and the solvent-accessible surface area (SASA) calculations were made using PyMol (Schrödinger).

Table 1

Data collection and refinement statistics. Values in parentheses are for the highest resolution shell.

	Pis s 1 (PDB ID: 7U1I)	Pis s 2 (PDB ID: 7U1J)	Len c 1 (PDB ID: 7U1H)
Data collection			
Space group	P 1 21 1	F 2 3	P 21 21 21
Unit-cell parameters (Å, °)	a = 98.32b = 52.69c = 127.84 α = 90 β = 95.19 γ = 90	a = 269.15b = 269.15c = 269.15 α = 90 β = 90 γ = 90	a = 86.44b = 92.50c = 143.74 α = 90 β = 90 γ = 90
Resolution (Å)	97.91–3.1 (3.31–3.10)	134.58–2.70 (2.81–2.70)	143.74–2.5 (2.60–2.50)
No. of unique reflections	20,422 (3738)	44,137 (4997)	39,945 (4426)
Completeness (%)	84.4 (82.8)	99.8 (100.0)	98.2 (97.8)
R_{merge}	0.102 (0.288)	0.13 (0.554)	0.14 (0.782)
$<I/\sigma(I)>$	6.0 (2.4)	8.9 (2.7)	9.6 (2.0)
Multiplicity	3.0 (2.1)	5.5 (5.7)	6.2 (6.2)
Refinement Statistics			
No. of reflections	20,413	44,134	39,887
R factor (%)	18.4–27.8	21.5	19.9
R_{free} (%)	28.5	28.1	28.4
No. of non-H atoms	8488	8329	8688
No. of waters	20	120	40
R.m.s.d., bonds (Å)	0.013	0.0095	0.0081
R.m.s.d., angles (°)	1.636	1.768	1.677
Average B factor (Å²)	49.0	46.0	38.0
Average B factor, waters (Å²)	14.5	28.7	31.2

3. Results

3.1. Recombinant protein production and crystallization.

Coding regions for the 'core' of each 7S globulin including His-tags, were recombinantly expressed in *E. coli*. Cell lysates containing the 3 target proteins were initially enriched by nickel-NTA affinity chromatography followed by size exclusion chromatography (Fig. 1A-C). This process yielded approximately 8.5, 13.5 and 25 mg of Pis s 1, Pis s 2 and Len c 1, respectively, all with greater than 95 % purity. Each protein was subjected to crystallization screening against a library of 192 conditions, with conditions yielding preliminary hits expanded to obtain diffraction quality crystals (Fig. 1D-F). In the case of Len c 1, streak microseeding was required. Crystal dimensions were in the range of 100–200 μm for Len c 1, and 200–300 μm for both Pis s 1 and Pis s 2.

3.2. Model generation and quality

A comparison of the primary amino acid sequences of the cores regions of Pis s 1, Pis s 2 and Len c 1 highlight their high degree of similarity to each other (Fig. 2). Pis s 1 and Len c 1 share 91 % amino acid identity. These vicilins respectively have 69 and 71 % identities to the Pis s 2 convicilin. More broadly (Supplemental Fig. S1), these pea and lentil globulins are 53–54 % identical to Ara h 1 (of known structure; PDBs: 3S7I, 3S7E), and 55–57 % identical to soya bean 7S vicilin (whose structure was used for molecular replacement; PDBs: 1UIK, 1UIJ, 1IPJ, 1IPK). Amino acid identities to other legume vicilins of known structure include 50–51 % with common bean (PDBIDs 1PHS; 2PHL), up to 52 % with jack bean (PDBIDs 2CAV, 2CAU; 6V7G, 6V7J, 6V7L, 1DGW, 1DGR) and up to 57 % with both mung (PDBID 2CV6) and adzuki (PDBIDs 2EA7, 2EAA) beans.

The data collection and refinement statistics for the three obtained structures are presented in Table 1. Crystal structures for Pis s 1, Pis s 2 and Len c 1 core regions were obtained at 3.1, 2.7 and 2.5 Å,
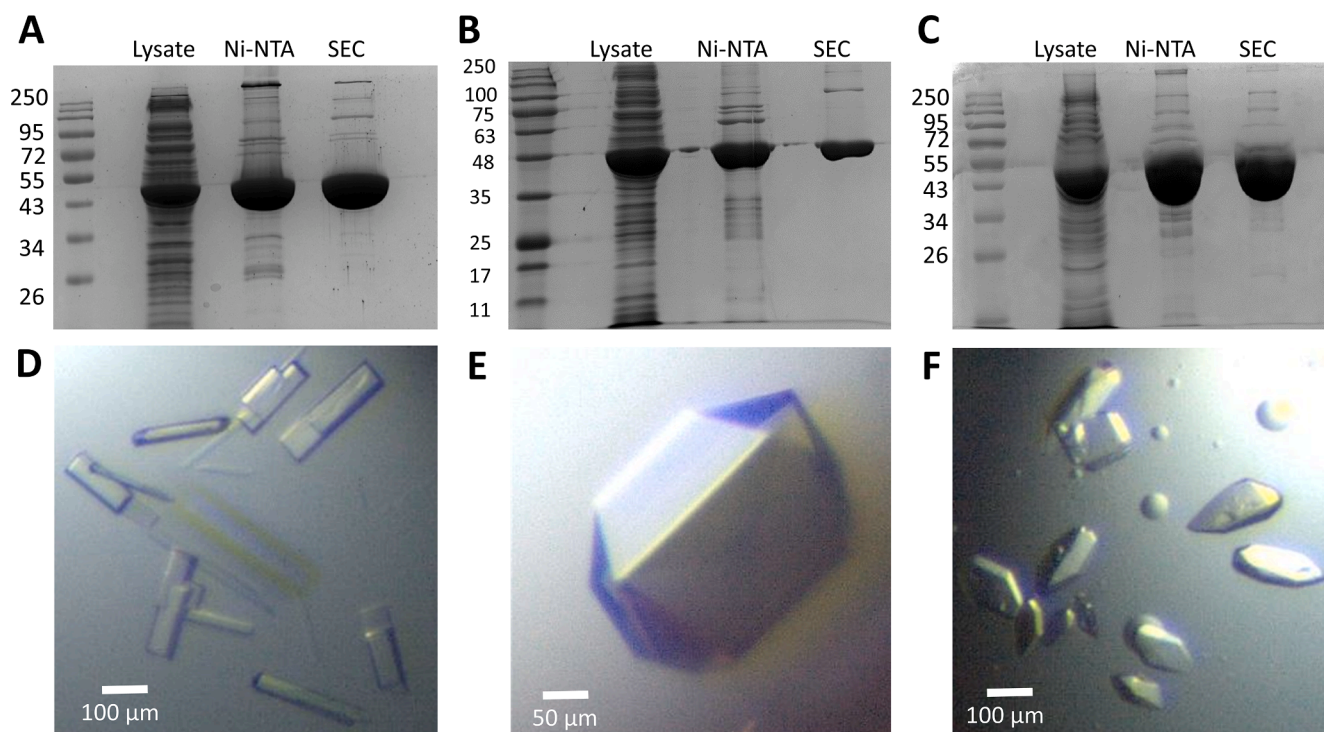
Fig. 1. Preparation of diffracting crystals of Pis s 1, Pis s 2, and Len c 1. Purification of recombinantly produced A) pea vicilin, Pis s 1; B) pea convicilin, Pis s 2; and C) lentil vicilin, Len c 1. SDS–PAGE showing total *E.coli* cell lysate fractions (Lysate), pooled fractions eluted from a Ni-NTA column (Ni-NTA) and the final purified product pooled from the size exclusion chromatography column (SEC). The SEC fractions were used directly for crystallization after concentration. Representative obtained crystals D) pea vicilin, Pis s 1, E) pea convicilin, Pis s 2, and F) lentil vicilin, Len c 1.

```
Pis_s_2    -SQERRNPFLFKSNKFLTLFENENGHIRLLQRFDKRSDLFENLQNYRLVEYRAKPHTIFL     253
Pis_s_1    SRSDQENPFIFKSNRFQTLYENENGHIRLLQKFDKRSKIFENLQNYRLLEYKSKPHTLFL     60
Len_c_1    SRSDQENPFIFKSNRFQTIYENENGHIRLLQKFDKRSKIFENLQNYRLLEYKSKPHTIFL     60
Ara_h_1    ---SRNNPFYFPSRRFSTRYGNQNGRIRVLQRFDQRSRQFQNLQNHRIVQIEAKPNTLVL     57
Gly_m_5.01 RRHKNKNPFLFGSNRFETLFKNQYGRIRVLQRFNQRSPQLQNLRDYRILEFNSKPNTLLL     247
           ...*** * *.* *  : *: *:**:**:*..**  ::**::*:: .:**:*:.*
Pis_s_2    PQHIDADLILVVLSGKAILTVLSPNDRNSYNLERGDTIKLPAGTTSYLVNQDDEEDLRLV     313
Pis_s_1    PQYTDADFILVVLSGKATLTVLKSNDRNSFNLERGDAIKLPAGTIAYLANRDDNEDLRVL     120
Len_c_1    PQFTDADFILVVLSGKAILTVLNSNDRNSFNLERGDTIKLPAGTIAYLANRDDNEDLRVL     120
Ara_h_1    PKHADADNILVIQQGQATVTVANGNNRKSFNLDEGHALRIPSGFISYILNRHDNQNLRVA     117
Gly_m_5.01 PNHADADYLIVILNGTAILSLVNNDDRDSYRLQSGDALRVPSGTTYYVVNPDNNENLRLI     307
           *:. *** ::*: . * ::: . .::*.*:.*: *.:::*:.*  *: * .:::.:**:
Pis_s_2    DLVIPVNGPGKFEAFDLAKN--KNQYLRGFSKNILEASYNTRYETIEKVLLEEQEKDRK-     370
Pis_s_1    DLAIPVNKPGGLQSFLLSGTQNQPSLLSGFSKNILEAAFNTNYEEIEKVLLEQQQEQEPQH     180
Len_c_1    DLAIPVNRPGGLQSFLLSGTQNQPSFLSGFSKNILEAAFNTEYEEIEKVLLEEQEQKSQH     180
Ara_h_1    KISMPVNTPGQFEDFFPASSRDQSSYLQGFSRNTLEAAFNAEFNEIRRVLLEENAGGEQE     177
Gly_m_5.01 TLAIPVNKPGRFESFFLSSTEAQQSYLQGFSRNILEASYDTKFEEINKVLFSREEGQQQG     367
           : .:*** **::: *   . .  : * ***.* ***.:::.:: *..**::.:
Pis_s_2    ---------RRQQGEETDAIVKVSREQIEELKKLAKSSSKKSLPSE---FEPINLRSHKPE     419
Pis_s_1    RRSLK--DRRQEINEENVIVKVSREQIEELSKNAKSSSKKSVSSE---SGPFNLRSRNPI     235
Len_c_1    RRSLR--DKRQEITNEDVIVKVSREQIEELSKNAKSSSKKSVSSE---SEPFNLRSRNPI     235
Ara_h_1    ERGQRRWSTRSSENNEGVIVKVSKEHVEELTKHAKSVSKKGSEEEGDITNPINLREGEPD     237
Gly_m_5.01 E----------ORLQESVIVEISKEQIRALSKRAKSSSRKTISSE---DKPFNLRSRDPI     427
           : ..**::*.*::  *.* *** *.*    *    *.***. .*
Pis_s_2    YSNKFGKLFEITPEKKYPQLQDLDLFVSCVEINEGALMLPHYNSRAIVVLLVNEGKGNLE     479
Pis_s_1    YSNKFGKFFEITPEK-NQQLQDLDIFVNSVDIKEGSLLLPNYNSRAIVIVTVTEGKGDFE     294
Len_c_1    YSNKFGKFFEITPEK-NPQLQDLDIFVNSVEIKEGSLLLPNYNSRAIVIVTVNEGKGDFE     294
Ara_h_1    LSNNFGKLFEVKPDKKNPQLQDLDMMLTCVEIKEGALMLPHFNSKAMVIVVVNKGTGNLE     297
Gly_m_5.01 YSNKLGKFFEITPEK-NPQLRDLDIFLSIVDMNEGALLLPHFNSKAIVILVINEGDANIE     486
           **.:**:**: .* *   **.***::. *:::.**.*:**:.**:*:*:* :**  . :*
Pis_s_2    LLGLKNEQQERE----------DRKERNNEVQRYEARLSPGDVVIIPAGHPVAITASSNL     529
Pis_s_1    LVGQRNENQGK---ENDK-EEEQEEETSKQVQLYRAKLSPGDVFVIPAGHPVAINASSDL     350
Len_c_1    LVGQRNENQQEQREENDE-EEGQEEETTKQVQRYRARLSPGDVLVIPAGHPVAINASSDL     353
Ara_h_1    LVAVRKEQQQRGRREEEEDEDEEEGSNREVRRYTARLKEGDVFIMPAAHPVAINASSEL     357
Gly_m_5.01 LVGLKEQQQEQ----------QQEEQPLEVRKYRAELSEQDIFVIPAGYPVVVNATSNL     535
           *:. .:::* .         .::  :*: * *.*.   *:::**..:**..::*.:*:*
Pis_s_2    NLLGFGINAENNERNFLSGSDDNVISQIENPVKELTFPGSVQEINRLIKNQKQSHFANA-     588
Pis_s_1    NLIGFGINAENNERNFLAGEEDNVISQVERPVKELAFPGSSHEVDRLLKNQKQSYFANA-     409
Len_c_1    NLIGFGINAKNNQRNFLAGEEDNVISQIQRPVKELAFPGSSREVDRLLTNQKQSHFANA-     412
Ara_h_1    HLLGFGINAENNHRIFLAGDKDNVIDQIEKQAKDLAFPGSGEQVEKLIKNQKESHFVSAR     417
Gly_m_5.01 NFFAIGINAENNQRNFLAGSQDNVISQIPSQVQELAFPGSAQAVEKLLKNQRESYFVDAQ     595
           :::.:*****.**.* **.*  ****.*:   .:**:***.   :*::**:.*:*.::*. ..
```

Fig. 2. Amino acid sequence alignment of the core crystallized regions of pea and lentil 7S globulins. Sequences of the crystallized regions of Pis s 1 (7U1I; CAF25232), Pis s 2 (7U1J; CAB82855) and Len c 1 (7U1H; CAD87730) and including peanut (Ara h 1; 3S7E) and the sequence of soya bean (Gly m 5.01 core; P0DO16). The unique Cys residue in Pis s 2 is highlighted in red font. Previously identified linear Ig epitopes are underlined with colors according to previous reports and Fig. 5 in this report, as follows: For Pis s 1 (Popp et al., 2020) - 3 purple; 14 green; 20 cyan; 28 yellow; 44 blue; 53 red; 64 magenta; 74 orange; 78 brown; 93 black; 100 sand; For Len c 1 (Vereda et al., 2010) - 1 red; 2 green; 3 blue; 4 yellow. The Ara h 1 (Cabanos et al., 2011) and Gly m 5.01 linear epitopes listed in Pi et al (2021) are underlined in black. Un-modelled regions in the crystal structures and loop regions in the alphafold model of Gly m 5.01 with a per residue confidence score < 90 are shown in light grey font. Amino acid numbering follows numbering of residues in the crystal structures or model. Numbers associated with colors in this legend refer to the epitope number as defined in the indicated prior reports and in Fig. 5 of this report. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

respectively. In the case of Pis s 1, 914 (86 %), of the amino acids were within the favoured core, and 127 (12 %) of the amino acids were within the allowed region, while 84 (2 %) of the amino acids were found in the outlier region of the Ramachandran plot as calculated by MolProbity (Chen et al., 2010). In contrast, Pis s 2 and Len c 1 were modelled with 91–92 % of amino acids in the favoured region, 6–8 % in the allowed regions and 1–2 % in the outlier regions. The R and R_{free} values ranged from 0.18 to 0.28, with the differences between R and R_{free} just broaching the recommended limit of 0.07 in all cases (Wlodawer et al., 2008). The asymmetric units all included three monomers

(Supplemental Fig. S2). However, only the Pis s 1 asymmetric unit actually contained the biological trimer, consistent with all three structures arising from crystals with different space groups (Table 1).

The numbers of waters in each structure varied from 20 in Pis s 1 and 120 in Pis s 2, to 40 in Len c 1, generally reflecting the unit cell sizes. The numbers of amino acid residues in each monomer include 362, 360 and 359 for Pis s 1, 349, 344 and 346 for Pis s 2, and 365, 365 and 363 for Len c 1. The *N*-terminal His-tag regions were not included in the models. Two additional regions were not included in any of the monomers due to weak $2|Fo| - |Fc|$ maps, making it impossible to trace even the Cα backbone in these regions. These un-modelled regions range from residues 164–172 and residues 295–319 respectively (Pis s 1 numbering). As well Pis s 2 had a third un-modelled region ranging from residues 394–407 (Pis s 1 numbering). All of these gaps are consistent with gaps previously observed in other 7S globulin structures (reviewed in (Cabanos et al., 2011)).

3.3. Overall structures

The core monomeric regions of the Pis s 1, Pis s 2 and Len c 1 structures oligomerize into homo-trimeric complexes observed at the unit cell level for all three proteins (Fig. 3A; Supplemental Fig. S3), representing the biological assembly as demonstrated by others (reviewed in (Kesari et al., 2017)). Each monomer can be divided into an N- and C- terminal module that share high structural similarity and are separated by a pseudo-dyad axis, as has been noted previously in other 7S globulins (Fig. 3B-D; (Cabanos et al., 2011)). Each module contains a conserved β-barrel core (jelly roll-like) domain comprised of two 7-stranded β-sheets and an extended loop domain comprised of six short α-helices. Superpositioning of the A chain monomers of both pea structures and the lentil structure with the A chain monomer of the peanut 7S globulin structure highlights a high degree of similarity across these four structures (Fig. 4). A comparison of A chain Cα atoms for Pis s 1 and Len c 1 yielded a root mean square deviations (RMSD) of 0.370 Å. Pis s 2 Cα atoms yielded RMSDs of 0.814 Å and 0.719 Å compared to these respectively.

Visual inspection of the aligned A chain Cα structures highlights minor differences among the 7S globulin, with Pis s 2 showing the greatest variability throughout, consistent with its lower sequence identity and higher RMSD values (Fig. 4). Pis s 2 variability is most notable in the *N*-terminal module including a loop (Pis s 2 residues 230–233), in a couple beta strands (Pis s 2 residues 283–288 and
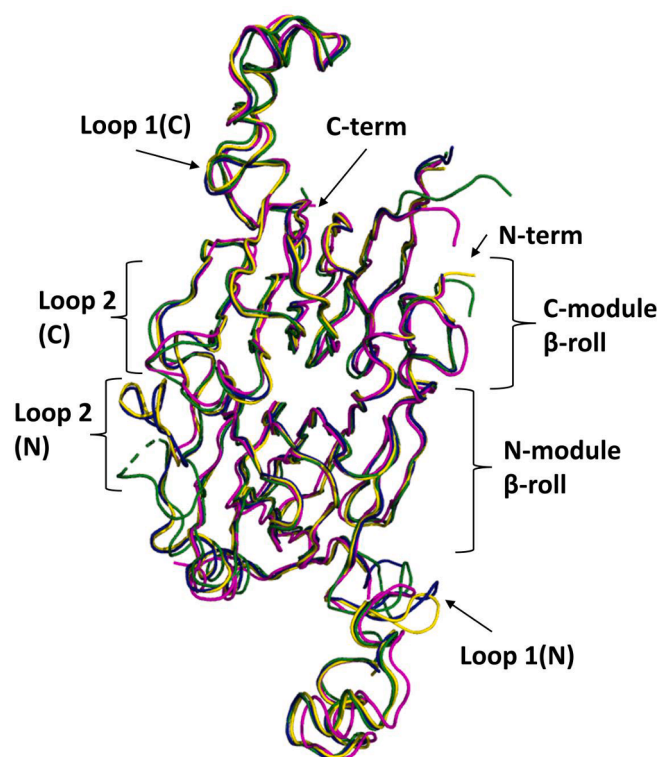
Fig. 4. Cα backbone superimposition of the A chain monomers of Pis s 1, Pis s 2 and Len c 1 with Ara h 1 7S globulin. Pis s 1, Pis s 2, Len c 1 and Ara h 1 are shown in yellow, pink, dark blue and green respectively. The two most highly structurally variable regions are noted, including Loop 1 (N) and Loop 2 (N). Comparable regions in the C-module are highlighted for comparison (Loop 1 (C), Loop 2 (C)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

302–309) and in the extended helical region (Pis s 2 residues 339–352). Pis s 2 also differs significantly from Pis s 1 and Len c 1 in a C-terminal module loop region (Pis s 2 residues 433–443). More broadly, residues immediately ahead and following un-modelled regions showed some variability across the 3 structures, consistent with the lack of observed density in the un-modelled regions arising from high flexibility of these loop regions (Pis s 2 residues 364–380 and 486–488 in Fig. 4). Finally,
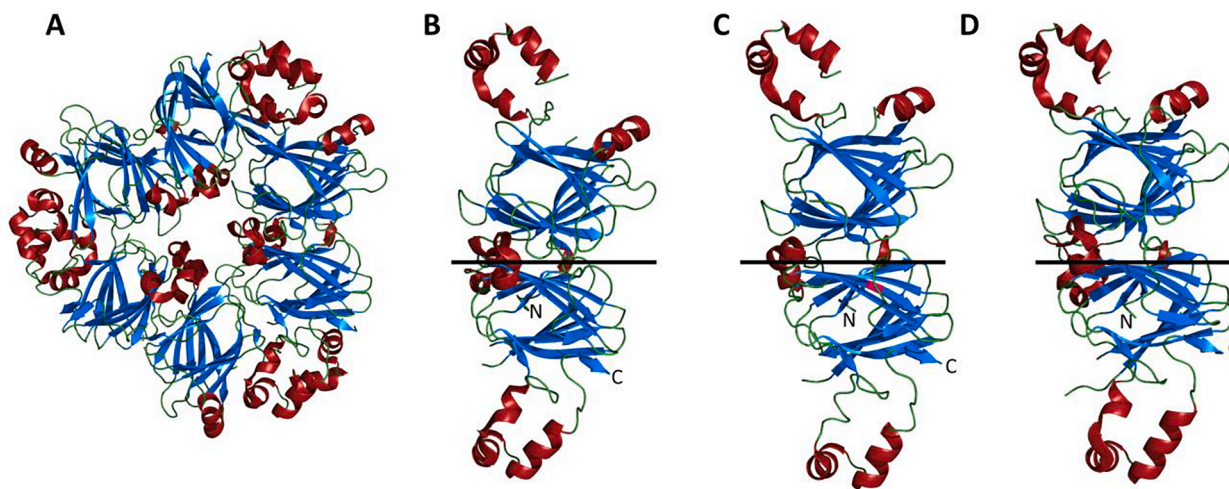

Fig. 3. Overall structure of the pea and lentil 7S globulins. A) Trimeric quaternary structure of Pis s 1 observed in the asymmetric unit. Monomer structures for B) Pis s 1 C) Pis s 2 D) Len c 1. The N-terminal modules are on the bottom, and the C-terminal modules are on the top of each image. The single unique Cys at position 448 of Pis s 2 is highlighted in pink. Strands of β-sheets and stretches of α-helices are shown in blue and red, respectively. The line represents the pseudo-dyad axis in between the N- and C-terminal modules. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the *N*-terminal core loop region connecting the extended helical region to the core β-roll (See Loop 2 in Fig. 4, equivalent to Pis s 2 residues 331–337) was highly variable across all 3 structures, an effect not observed in the equivalent loop segment in the C-terminal-module.

3.4. Comparative structural epitope mapping on pea and lentil crystal structures

Initially, the eleven candidate epitopes (Fig. 1 & Supplemental Table S2) of Pis s 1 identified by Popp *et al.* were mapped onto the crystal structure obtained in this current study (Popp et al., 2020). The

candidate epitopes were distributed throughout the surface of Pis s 1, both on the monomer and the trimer (Fig. 5A & Supplemental Fig. S4A), validating Popp *et al.*'s prior homology model that predicted a similar epitope distribution. Upon trimer formation, candidate epitopes 14, 20, 74, 78, 93, and 100 together lose at least 30 % of their accessible surface area. With the exception of one amino acid in chain A, there was no electron density for the residues of epitope 44. Additionally, there was a gap (residues 299–319) in the electronic density between peptides 74 and 78, including residues from both peptides.

In contrast, only four candidate epitopes (Fig. 2 & Supplemental Table T2) have been identified for Len c 1 (Vereda et al., 2010). Since
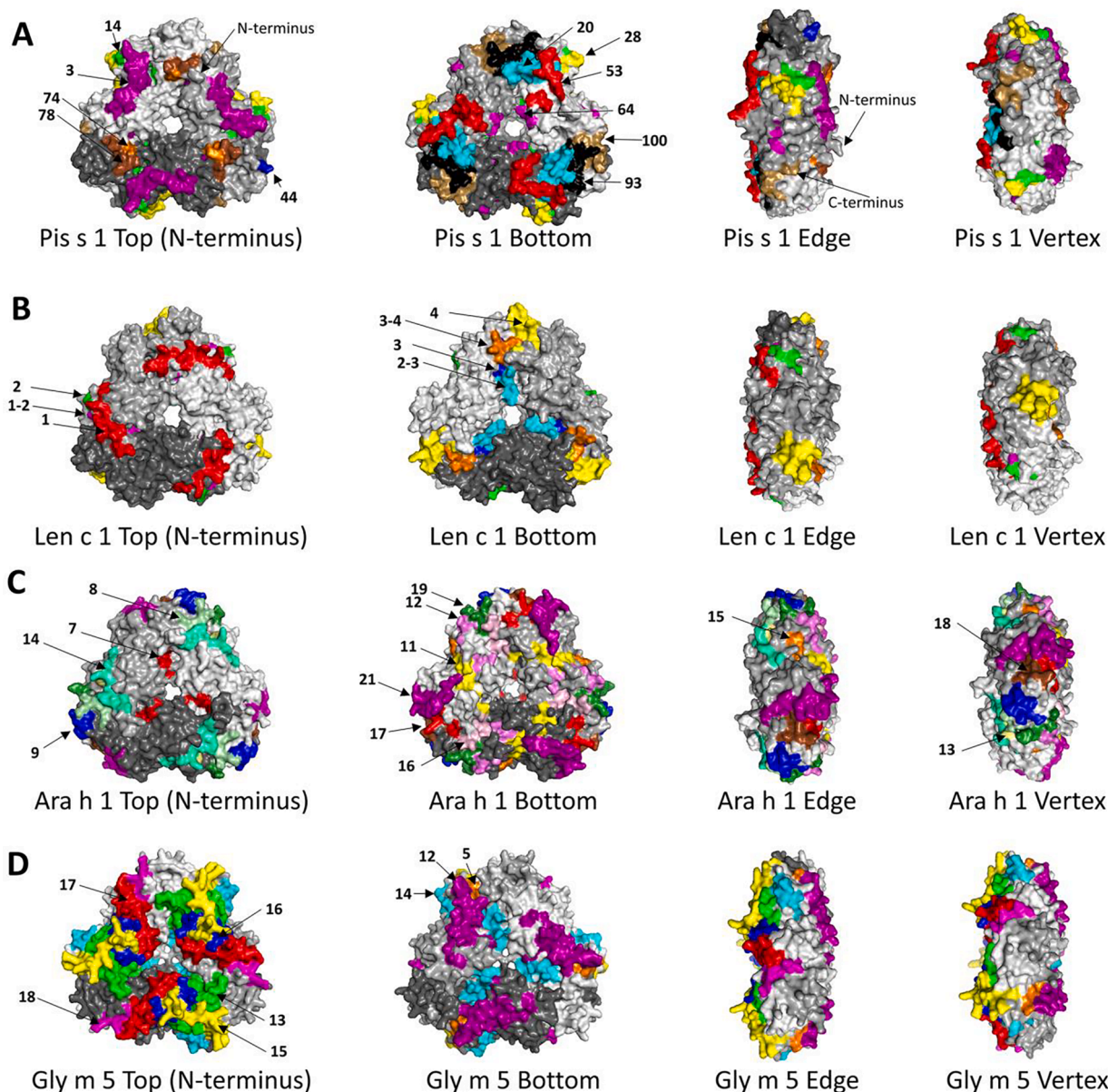
Fig. 5. Surface character and mapping of previously identified linear Ig epitopes on the molecular surface of the Pis s 1 and Len c 1 crystal trimeric unit structures. Epitopes were mapped onto the trimeric structures of (A) Pis s 1 (PDBID: 7U1I) and (B) Len c 1 (PDBID: 7U1H). (C) The Ara h 1 epitopes listed in Cabanos et al (2011) are shown mapped onto the trimeric structure of Ara h 1 (PDBIB: 3SMH). (D) The Gly m 5.01 epitopes listed in Pi et al (2021) are shown mapped onto an AlphaFold model of Gly m 5.01 core (AF-P0DO16) aligned to all three chains of the trimeric structure of Gly m 5.02 (PDBID: 1UIK). Pis s 1 epitopes are colored as per a previous report (Popp et al., 2020): 3 purple; 14 green; 20 cyan; 28 yellow; 44 blue; 53 red; 64 magenta; 74 orange; 78 brown; 93 black; 100 sand. For Len c 1 epitopes (Vereda et al., 2010): 1 red and magenta; 2 magenta, green, and cyan; 3 cyan, blue, and orange; 4 orange and yellow. The numbers on the trimeric structures correspond to the epitope numbers as defined in the indicated prior reports. The Ara h 1 epitopes are largely colored according to a previous report (Cabanos et al., 2011): 7 red; 8 pale green; 9 blue; 11 yellow; 12 magenta; 13 pale yellow; 14 cyan; 15 orange; 16 light pink; 17 red; 18 brown; 19 green; and 21 purple. For the Gly m 5.01 epitopes (Pi et al. (2021)): 5 orange; 12 purple; 13 green; 14 cyan; 15 yellow; 16 blue; 17 red; and 18 magenta. Due to their overlap with other epitopes, epitopes 3, 4, 6, 7, and 11 are excluded for clarity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

these epitopes have overlap between each other, residues that are shared between two epitopes are coloured differently when mapped on the new Len c 1 structure (Fig. 5B). Unlike Pis s 1, the linear candidate epitopes of Len c 1 are clustered at the C-terminus of the protein mostly distributed on the side of the monomeric structures that contains that N- and C-termini. Upon formation of the trimer, epitopes 1–2 and 3 are buried with a > 5-fold loss in surface accessible surface area. Nonetheless, all four epitopes are exposed to the solvent in the trimeric structure (Supplemental Fig. S4B): epitope 1 is on the *N*-terminus top side, epitope 3 is on the bottom side, epitope 4 is located on the corners of the triangular shape of the structure, and epitope 2 is solvent accessible along the edge of the triangular shape structure.

To the best of our knowledge, epitopes for the convicilin Pis s 2 have not been experimentally identified yet.

4. Discussion

In this report, the first obtained X-ray crystallographic structures of pea and lentil 7S globulins were solved and found to be similar to each other and to other legume vicilins. In particular they maintain the expected 7S globulin fold, with trimeric quaternary structure and monomers comprised of β-barrel N- and C-modules. Nonetheless a number of differences between the structures and mapped epitopes are observed upon closer inspection.

The higher RMSD for the Pis s 2 convicilin Cα atoms compared to Len c 1 and Pis s 1 is consistent with the observed 20 % lower amino acid sequence identity. Pea convicilin is also distinguished from vicilin by a significantly lower gene copy number (2:18), a lack of processing of the signal peptide and no glycosylation (Casey et al., 1986; Newbigin et al., 1990). As well, vicilin genes are expressed during early seed deposition (Casey et al., 1986). In contrast, the pea convicilin genes have been shown to be transcribed only in the late stages of seed deposition (Bown et al., 1988; Newbigin et al., 1990). Relative accumulation of convicilin and vicilin in the seed are consistent with this and also reflect the gene copy number, with significantly larger amounts of vicilin being stored.

Broader inspection of unit and super cell packing highlights both the amenability of all three proteins to form trimeric biological assemblies, as well as differential packing of these trimeric complexes in different crystal space groups (Supplemental Fig. S3). Interestingly, one of the more notable differences between pea vicilin and convicilin is the presence of a single Cys residue, at position 448 in Pis s 2 (also present in Ara h 1; Fig. 2), whereas the pea and lentil vicilins contain no Cys, presenting a Ser residue in this position. In Pis s 1, this Ser 263 is the foundation of a robust hydrogen bonding network with Asn 354 and Asp 265 (stabilizing two strands in the β-sheet fold in this C-module region) and more notably Arg 230 in Loop 2 adjacent to Arg 231 and Asn233. These latter residues make crystallographic packing contacts with *N*-module residues Asn 114 and Asp 166 in a neighbouring trimeric symmetry mate in this P 1 2₁ 1 space group. In contrast Cys 448 in Pis s 2 forms hydrogen bonds to Asn 530 and the main chain carbonyl of Leu 531, again linking two strands in the β-sheet, but not connecting to Loop 2. This alters the hydrogen bonding network associated with Loop 2, potentially contributing to stabilization of the alternative Loop 2 – Loop 2 inter-trimeric symmetry mate interaction in the F 2 3 crystallographic space group. The potential relevance of these packing differences to deposition during seed development cannot be ignored. Although whether they have relevance to early and late deposition, stabilization of the deposited protein or other traits remains for future consideration.

More broadly Pis s 1, Len c 1 and Pis s 2 yielded RMSDs of 0.699, 0.685 and 1.114 Å compared to the Cα atoms in peanut Ara h 1, indicating a high degree of structural similarity particularly with the vicilins. Ara h 1 was previously shown to have RMSDs of 1.00 to 1.15 Å compared to an array of other legume 7S structures including soya, adzuki and mung beans (Cabanos et al., 2011). Both pea and lentil vicilin structures vary from Ara h 1 in the regions of residues 216–226 (Pis s 1 numbering; see Loop 1 in Fig. 4) and in a C-terminal module loop

region (Pis s 2 residues 433–443). Interestingly, Pis s 2 remains virtually identical to Ara h 1 in the latter C-terminal module loop. The Loop 1 region is not modelled in Pis s 2. Further inspection of other missing regions reveals that these gaps in electron density were consistently observed across an array of representative legume 7S structures. An amino acid sequence alignment emphasizes the presence of non-conserved strings of Asp, Glu, Asn and Gln amino acids in these missing regions (Fig. 2 and Supplemental Fig. S1), proposed previously to be surface exposed based on their hydrophilic nature. Notably, the epitope mapping exercise emphasizes that these gap regions do have reactivity to immunoglobulins in some species, consistent with their proposed surface exposure.

With respect to the mapping of previously predicted immunoglobulin binding epitopes on Pis s 1, the degree to which epitopes are buried at interfaces in the trimeric structure (compare Fig. 5A to Supplemental Fig. S4A) is notable. The potential for increased allergenicity upon breakdown of this quaternary structure during food processing or digestion cannot be ignored. Further comparison of pea and lentil vicilin epitopes highlights significantly fewer epitopes for Len c 1 than Pis s 1. However, several regions of Len c 1 (not experimentally identified as epitopes), contain high sequence and structural similarity to Pis s 1 and in fact correspond exactly to Pis s 1 epitopes. For instance, Pis s 1 epitopes 28 and 53 are identical to the corresponding sequence in Len c 1. At the same time, Pis s 1 epitopes 3 (L19I), 14 (Y63F), 20 (T78I), and 64 (D265E) differ by one amino acid from the corresponding peptides in Len c 1, which represent relatively conservative changes (e.g. Leu to Ile). Additionally, two of the differing residues (residues 63 and 78) are buried (SASA < 2.5 Å²) and thus unlikely to be directly binding immunoglobulin. Therefore based on this comparison we predict that these regions may also be potential epitopes for Len c 1 that were not identified in the original linear epitope search. Validation of these computational predictions remains for the future.

At the same time, the degree to which epitopes cover a significant portion of the surface of trimers and individual monomers is consistent with challenges associated with addressing allergenicity of these proteins in food (de Silva et al., 2022). To gain more insight into epitope coverage across 7S globulins, the Ara h 1 epitopes identified by Cabanos et al. and the Gly m 5.01 epitopes listed by Pi et al., were mapped onto the trimeric crystal structure of Ara h 1 (Fig. 5C; PDB ID: 5SMH) and a model of Gly m 5.01 predicted by AlphaFold (Fig. 5D; AF-P0DO16), respectively (Cabanos et al., 2011; Pi et al., 2021). By aligning the sequences and structures of Pis s 1 and Len c 1 to Ara h 1 and Gly m 5, we were able to identify homologous epitopes at the same relative positions on the structures, with details compiled in Fig. 2 & Supplemental Table T2. All epitopes of Len c 1 had corresponding epitopes in Ara h 1, but no corresponding epitopes in Gly m 5.01. Also, several Pis s 1 epitopes corresponded to those identified in Ara h 1 (Cabanos et al., 2011), Gly m 5.01 (Pi et al. 2021), and those identified in Len c 1. Gly m 5.01 epitopes that did not correspond to those identified in Pis s 1 had homologous epitopes in Ara h 1. Additionally, Pis s 1 and Ara h 1 had epitopes in positions that were unique to each protein, which may account for different allergenic responses to different pulse vicilin proteins.

In terms of the allergenicity of these crops, it is notable that both peanut and soya bean are considered major allergens, while peas and lentils are not (FALCPA, 2004). And even in comparing peanut and soya bean allergenic responses, peanut allergies are highly prevalent and severe, while soya bean allergies tend to be less intense (Gupta et al., 2019). Pea and lentil allergenic responses are similar to soya bean in places where they are consumed as a staple (Verma et al., 2013). Certainly a comparison of the extent of epitopes mapped on Len c 1, compared to Ara h 1 might suggest a correlation (more epitopes/more potent allergic responses). However, a further more detailed comparison of Pis s 1, Gly m 5.01 and Ara h 1 epitopes suggests that the relationship is not quite so straightforward. Gly m 5.01 has just as much epitope coverage (or more) as Ara h 1 (Figs. 2 and 5). At the same time, one must

also consider that 7S globulins are not the only allergenic proteins in these crops (Riascos et al., 2010). Indeed, it is proposed that Ara h 2 and Ara h 6 allergens in peanut are responsible for the heightened severity of peanut allergies (Zhuang and Dreskin, 2013). Thus contributions of other epitopes on other proteins must be considered in any future treatments, breeding programs or processing strategies developed to reduce allergenicity of these species.

5. Conclusions

In conclusion, the first X-ray crystallographic structures of 7S-glob-ulin pea and lentil vicilin as well as pea convicilin are presented. All three structures are similar to those of other pulse 7S-8S globulins, maintaining the common trimeric quaternary structure and with each monomer comprised of N- and C- terminal modules, each comprised in turn of conserved β-barrel cores. The potential impact of sequence dif-ferences on structure and packing in the different crystal space groups is noted, with potential relevance to packing upon seed deposition. Map-ping on the obtained crystal structures highlights significant Ig epitope overlap between pea, lentil, peanut and soya bean, and significant coverage of the entire seed storage protein, emphasizing the challenge in addressing food allergies. How recently developed biologicals might be refined to be more effective, or how these seed storage proteins might be modified *in planta* to be less immuno-reactive remain challenges for the future.

Declaration of Competing Interest

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

Data availability

Data submitted to the PDB

Database

PDB: 7U1H, 7U1I, 7U1J.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochms.2022.100146.

References

Bown, D., Ellis, T. H. N., & Gatehouse, J. A. (1988). The sequence of a gene encoding convicilin from pea (*Pisum sativum L.*) shows that convicilin differs from vicilin by an insertion near the N-terminus. *The Biochemical Journal, 251*, 717–726.

Cabanos, C., Urabe, H., Tandang-Silvas, M. R., Utsumi, S., Mikami, B., & Maruyama, N. (2011). Crystal structure of the major peanut allergen Ara h 1. *Molecular Immunology, 49*, 115–123.

Casey, R., Domoney, C., & Ellis, T. H. N. (1986). Legume storage protein and their genes. In Oxford Surv Plant Molec Bell Biol pp. 1-95.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Biological Crystallography, 66*, 12–21.

Chruszcz, M., Maleki, S. J., Majorek, K. A., Demas, M., Bublin, M., Solberg, R., et al. (2011). Structural and immunologic characterization of Ara h 1, a major peanut allergen. *Journal of Biological Chemistry, 286*, 39318–39327.

de Silva, D., Singh, C., Arasi, S., Muraro, A., Zuberbier, T., Ebisawa, M., et al. (2022). Systematic review of monotherapy with biologicals for children and adults with IgE-mediated food allergy. *Clinical and Translational Allergy, 12*, e12123.

Dutta, A., Trivedi, A., Nath, C. P., Gupta, D. S., & Hazra, K. K. (2022). A comprehensive review on grain legumes as climate-smart crops: Challenges and prospects. *Environmental Challenges, 7*.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D, Biological Crystallography, 66*, 486–501.

FALCPA. (2004). Food Allergen Labeling and Consumer Protection Act of 2004. In II, U. F.A.D. Administration, ed.

FAOSTAT. (2022). Food and Agriculture Organization of the United Nations Statistics Database.

Fukuda, T. M. N., Salleh, M.R.M., Mikami, B., & Utsumi, S. (2008). Characterization and crystallography of recombinant 7S globulins of Adzuki bean and structure-function relationships with 7S globulins of various crops. Journal of Agricultural Food Chemistry *56*, 4145-4153.

Ferreira, H., Pinto, E., & Vasconcelos, M. W. (2021). Legumes as a Cornerstone of the Transition Toward More Sustainable Agri-Food Systems and Diets in Europe. *Frontiers in Sustainable Food Systems, 5*, 694121. https://doi.org/10.3389/fsufs.2021.694121

Gupta, R. S., Warren, C. M., Smith, B. M., Jiang, J., Blumenstock, J. A., Davis, M. M., et al. (2019). Prevalence and severity of food allergies among US adults. *JAMA Network Open, 2*, e185630.

Itoh, T., Garcia, R. N., Adachi, M., Maruyama, Y., Tecson-Mendoza, E. M., Mikami, B., et al. (2006). Structure of 8Salpha globulin, the major seed storage protein of mung bean. *Acta Crystallographica. Section D, Biological Crystallography, 62*, 824–832.

Kesari, P. N., Sharma A., Katiki, M., Kumar, P., Gurjar, B. R., Tomar, A, Sharma, A. K., & Kumar, P. (2017). Structure, functional and evolutionary aspects of seed globulins. Protein and Peptide Letters *24*, 267-277.

Ko, T.-P., Day, J., & McPherson, A. (2000). The refined structure of canavalin from jack bean in two crystal forms at 2.1 and 2.0 A resolution. *Acta Crystallographica Section D, 56*, 411–420.

Ko, T.-P., Kuznetsov, Y. G., Malkin, A. J., Day, J., & McPherson, A. (2001). X-ray diffraction and atomic force microscopy analysis of twinned crystals: Rhombohedral canavalin. *Acta Crystallographica Section D, 57*, 829–839.

Kriz, A. L. (1999). 7S globulins of cereals. In Seed proteins, P. R. Shewry, R. Casey, (Eds.) (Dordrecht: Springer Netherlands), pp. 477-498.

Lawrence, M. C., Izard, T., Beuchat, M., Blagrove, R. J., & Colman, P. M. (1994). Implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. *Journal of Molecular Biology, 238*, 748–776.

Lawrence, M. C., Suzuki, E., Varghese, J. N., Davis, P. C., Van Donkelaar, A., Tulloch, P. A., et al. (1990). The three-dimensional structure of the seed storage protein phaseolin at 3 A resolution. *EMBO Journal, 9*, 9–15.

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., et al. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*.

Maruyama, N., Adachi, M., Takahashi, K., Yagasaki, K., Kohno, M., Takenaka, Y., et al. (2001). Crystal structures of recombinant and native soybean beta-conglycinin beta homotrimers. *European Journal of Biochemistry, 268*, 3595–3604.

Maruyama, N., Maruyama, Y., Tsuruki, T., Okuda, E., Yoshikawa, M., & Utsumi, S. (2003). Creation of soybean β-conglycinin β with strong phagocytosis-stimulating activity. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1648*, 99–104.

McPherson, A. (2020). Binding of benzoic acid and anions within the cupin domains of the vicilin protein canavalin from jack bean (Canavalia ensiformis): Crystal structures. *Biochemical and Biophysical Research Communications, 524*, 268–271.

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., et al. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica. Section D, Biological Crystallography, 67*, 355–367.

Newbigin, E. J., deLumen, V. O., Chandler, P. M., Gould, A., Blagrove, R. J., March, J. F., Kortt, A. A., & Higgins, T. J. V. (1990). Pea convicilin: structure and primary sequence of the protein and expression of a gene in the seeds of transgenic tobacco. Planta *180*, 461-470.

Pi, X., Sun, Y., Fu, G., Wu, Z., & Cheng, J. (2021). Effect of processing on soybean allergens and their allergenicity. *Trends in Food Science & Technology, 118*, 316–327.

Popp, J., Trendelenburg, V., Niggemann, B., Randow, S., Volker, E., Vogel, L., et al. (2020). Pea (Pisum sativum) allergy in children: Pis s 1 is an immunodominant major pea allergen and presents IgE binding sites with potential diagnostic value. *Clinical and Experimental Allergy, 50*, 625–635.

Potterton, E., Briggs, P., Turkenburg, M., & Dodson, E. (2003). A graphical user interface to the CCP4 program suite. *Acta Crystallographica Section D, 59*, 1131–1137.

Riascos, J. J., Weissinger, A. K., Weissinger, S. M., & Burks, A. W. (2010). Hypoallergenic legume crops and food allergy: Factors affecting feasibility and risk. *Journal of Agriculture and Food Chemistry, 58*, 20–27.

Vagin, A., & Teplyakov, A. (1997). MOLREP: An automated program for molecular replacement. *Journal of Applied Crystallography, 30*, 1022–1025.

Vereda, A., Andreae, D. A., Lin, J., Shreffler, W. G., Ibanez, M. D., Cuesta-Herranz, J., et al. (2010). Identification of IgE sequential epitopes of lentil (Len c 1) by means of peptide microarray immunoassay. *The Journal of Allergy and Clinical Immunology, 126*, 596 e591–601.

Verma, A. K., Kumar, S., Das, M., & Dwivedi, P. D. (2013). A comprehensive review of legume allergy. *Clinical Reviews in Allergy and Immunology, 45*, 30–46.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D, 67*, 235–242.

Winter, G., Beilsten-Edmands, J., Devenish, N., Gerstel, M., Gildea, R. J., McDonagh, D., et al. (2022). DIALS as a toolkit. *Protein Science, 31*, 232–250.

Wlodawer, A., Minor, W., Dauter, Z., & Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal, 275*, 1–21.

Zhuang, Y., & Dreskin, S. C. (2013). Redefining the major peanut allergens. *Immunologic Research, 55*, 125–134.