



# The structure of the *Gemella haemolysans* M26 IgA1 protease trypsin-like domain

Norman Tran,<sup>a</sup> Jasmina S. Redzic,<sup>b</sup> Elan Z. Eisenmesser<sup>b</sup> and Todd Holyoak<sup>a\*</sup>

<sup>a</sup>Department of Biology, University of Waterloo, Waterloo, ON N2L 3G1, Canada, and <sup>b</sup>Department of Biochemistry and Molecular Genetics, School of Medicine, University of Colorado Denver, Aurora, CO 80045, USA. \*Correspondence e-mail: tholyoak@uwaterloo.ca

Received 31 December 2024

Accepted 10 February 2025

Edited by N. Sträter, University of Leipzig, Germany

**Keywords:** IgA proteases; serine proteases; trypsin; chymotrypsin; M26 proteases.

**PDB reference:** *Gemella haemolysans* M26 IgA1 protease trypsin-like domain, 9ect

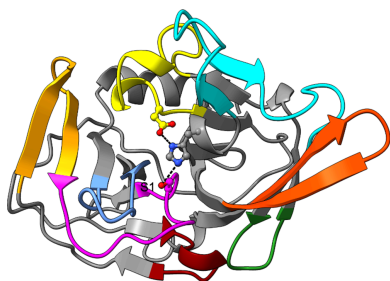
**Supporting information:** this article has supporting information at journals.iucr.org/f

Immunoglobulin A (IgA) proteases are a group of bacterial-derived enzymes that selectively hydrolyze human IgA in the hinge region that is unique to this immunoglobulin. Several IgA protease (IgAP) families have evolved this ability using both metalloprotease and serine protease chemical mechanisms. One family of metal-dependent IgAPs is the M26 family. This family can be grouped into two subfamilies based upon the presence or absence of a trypsin-like domain found N-terminal to the IgAP domain. The role of this domain in IgAP structure and function is poorly understood. Here, we present the first structural characterization of an M26 IgAP trypsin-like domain from *Gemella haemolysans* (*GhTrp*). These structural data demonstrate that the *GhTrp* domain possesses a trypsin-like fold but contains significant deviations in the surface-loop structure that is known to be coupled to protease selectivity. The lack of observable catalytic function coupled with the structural data suggest that this domain may exist in a pro-enzyme-like state that can potentially be activated when the domain is N-terminally proteolytically excised from the larger M26 IgAP structure.

## 1. Introduction

Immunoglobulin A proteases (IgAPs) represent an interesting group of proteolytic enzymes that have convergently evolved to specifically cleave the unique hinge region present in IgA1 from humans and great apes via several different chemical mechanisms. Representative members of all three known IgAP families have been biochemically and structurally characterized. These include an S6 serine IgAP, two M26 metal-dependent IgAPs and most recently an M64 metal-dependent IgAP (Johnson *et al.*, 2009; Wang *et al.*, 2020; Redzic *et al.*, 2022; Tran *et al.*, 2024).

The M26 IgAP family can be split into two subfamilies with distinct domain architectures. The subfamily represented by the *Gemella haemolysans* IgAP (*GhIgAP*) contains an additional trypsin-like domain (*GhTrp*) found N-terminal to the IgAP domain (Supplementary Fig. S1a; residues 684–896<sup>1</sup>; Redzic *et al.*, 2022). This trypsin-like domain is missing from the other subfamily represented by the *Streptococcus pneumoniae* IgAP (Redzic *et al.*, 2022). Prior studies that compared *GhIgAP* constructs with and without this domain concluded that *GhTrp* had no effect on IgA1 proteolysis (Redzic *et al.*, 2022). This left the role of *GhTrp* in the context of the larger M26 IgAP structure open to further investigation (Redzic



<sup>1</sup>To be consistent with the numbering found in the deposited full-length structure of *GhIgAP* (PDB entry 7uvk), in the manuscript we have used the sequence numbering corresponding to NCBI entry WP\_040464465.1, which is offset from the sequence found in the PDB deposition corresponding to UniProt entry C5NYF3 by 23 residues.

**Table 1**Data-collection and refinement statistics for *GhTrp* (PDB entry 9ect).

Values in parentheses are for the highest resolution shell.

Wavelength (Å)	1.521
Resolution range (Å)	74.21–1.75 (1.79–1.75)
Space group	$P2_1$
$a, b, c$ (Å)	47.27, 58.55, 76.37
$\alpha, \beta, \gamma$ (°)	90, 103.66, 90
Total reflections	233882 (21006)
Unique reflections	40706 (4006)
Multiplicity	5.7 (5.2)
Completeness (%)	99.30 (98.86)
Mean $I/\sigma(I)$	2.60 (0.28)
Wilson $B$ factor (Å <sup>2</sup> )	13.03
$R_{\text{merge}}$	0.205 (0.621)
$R_{\text{meas}}$	0.224 (0.691)
$R_{\text{p.i.m.}}$	0.089 (0.296)
$CC_{1/2}$	0.984 (0.835)
No. of reflections used in refinement	40700 (4005)
No. of reflections used for $R_{\text{free}}$	2031 (181)
$R_{\text{work}}$	0.1936 (0.2865)
$R_{\text{free}}$	0.2421 (0.3443)
No. of atoms	
Total	3855
Protein	3313
Water	542
$B$ factors (Å <sup>2</sup> )	
Overall	20.94
Protein	19.67
Water	28.67
Root-mean-square deviations	
Bond lengths (Å)	0.004
Angles (°)	0.735
Rotamer outliers (%)	0
Clashscore	3.65
Ramachandran statistics (%)	
Favoured	98.32
Allowed	1.68
Outliers	0

*et al.*, 2022). To gain insight into the structure and potential functional role of this domain, we solved the crystal structure of *GhTrp* and demonstrated that the domain does indeed possess a trypsin-like protease fold. This fold, however, contains many unique changes in the well characterized surface loops that are known to contribute to trypsin-like protease specificity. The crystal structure suggests that *GhTrp*, as it exists in the full-length M26 *GhIgAP*, may be an inactive pro-enzyme. We propose a mechanism of pro-enzyme activation through the proteolytic removal of the N-terminal region of the full-length enzyme from the *GhTrp* domain.

## 2. Materials and methods

### 2.1. Protein expression and purification

The trypsin-like domain of *G. haemolysans* IgAP (WP\_040464465.1; residues 684–896; *GhTrp*) was cloned into pET-21b with an N-terminal His-tag and thrombin cleavage site as described previously (Redzic *et al.*, 2022). *Escherichia coli* BL21(DE3) cells were transformed with this vector and used for recombinant protein expression. An overnight culture was inoculated into ZYP-5052 autoinduction medium (Studier, 2005) at a ratio of 50 ml overnight culture to 1 l final medium volume with a minimum headspace:medium ratio of

1:1. ZYP-5052 medium was supplemented with 50  $\mu\text{g ml}^{-1}$  kanamycin and the cells were grown at 20°C at 150 rev  $\text{min}^{-1}$  for 40–48 h, harvested at 6000g and the cell pellets were stored at  $-80^\circ\text{C}$ .

All purification steps were carried out at 4°C. Cell pellets were thawed in buffer *A* (25 mM HEPES pH 7.5, 0.5 M NaCl, 10 mM imidazole), passed twice through a French pressure cell (Thermo Fisher Scientific, Waltham, Massachusetts, USA) at 7.6 MPa for cell lysis and debris was removed via high-speed centrifugation at 17 000g. The clarified cell lysate was then incubated with Ni-NTA resin (Qiagen) pre-equilibrated in buffer *A* for 1 h. The resin was first washed with ten column volumes (CV) of buffer *B* [25 mM HEPES pH 7.5, 0.1% (v/v) IGEPAL CA-630, 10 mM imidazole] to remove nonspecific hydrophobically bound contaminants, followed by a wash with 15 CV buffer *A*. The protein was eluted with buffer *C* (25 mM HEPES pH 7.5, 0.5 M NaCl, 300 mM imidazole). The Ni-NTA flowthrough was concentrated to less than 1 ml and loaded onto a pre-packed HiLoad Superdex 75 pg 16/600 column pre-equilibrated in crystallization buffer (25 mM HEPES pH 7.5) and run at 0.5 ml  $\text{min}^{-1}$ . The purity of the protein in the non-aggregate absorbance peak was qualitatively analysed using SDS-PAGE. Pure fractions were concentrated, frozen in pellets by direct immersion in liquid nitrogen and stored at  $-80^\circ\text{C}$ . Protein concentration was measured using a 1% mass extinction coefficient of 10.95, theoretically determined from the primary sequence of the protein (Gasteiger *et al.*, 2005).

### 2.2. Protein crystallization

2.0  $\mu\text{l}$  10 mg  $\text{ml}^{-1}$  *GhTrp* was mixed with 2.0  $\mu\text{l}$  reservoir solution [0.2 M  $\text{KNO}_3$ , 22% (w/v) PEG 3350] in a hanging-drop crystallization tray. Thin plate clusters appeared after several days and were manually manipulated to acquire single crystals suitable for diffraction. Crystals were cryoprotected in 0.2 M  $\text{KNO}_3$ , 25% (w/v) PEG 3350 supplemented with 20% (v/v) PEG 400 before being plunged into liquid nitrogen for data collection.

### 2.3. Data collection and processing

Diffraction data were collected on the CMCF-BM beamline at the Canadian Light Source (CLS) using a Dectris PILATUS3 S 6M. Data were indexed, integrated and scaled with *DIALS* (Winter *et al.*, 2018) and imported into the *CCP4* suite (Agirre *et al.*, 2023) with *AIMLESS* (Evans & Murshudov, 2013). The structure was solved with *phenix.mr\_rosetta* through a combination of *ab initio* modelling and molecular replacement (DiMaio *et al.*, 2011; Terwilliger *et al.*, 2012). Refinement was performed using *phenix.refine* (Afonine *et al.*, 2012) in conjunction with manual model building in *Coot* (Emsley *et al.*, 2010). Translation–libration–screw parameters were automatically determined and used by *phenix.refine*. Model geometry was analysed and optimized based on suggestions by *MolProbity* (Williams *et al.*, 2018). Data-collection and model statistics are summarized in Table 1.

**Table 2**

Top-ranking structures from *DALI* analysis of *GhTrp* against the PDB50 data set.

Protein	PDB code	<i>DALI</i> Z-score	$C^\alpha$ r.m.s.d. (Å)	Sequence identity (%)
<i>Bacillus intermedius</i> glutamyl-endopeptidase	1p3c	21.8	2.0	19
<i>Arthrobacter nicotinovorans</i> protease	3wy8	19.6	2.4	12
Protease DO	4ynn	19.3	2.8	21
Exfoliative toxin D2	5c2z	19.1	2.3	21
Exfoliative toxin C	8r3i	18.9	2.3	21
Epidermolytic toxin A	1agj	18.6	2.2	21

### 3. Results and discussion

#### 3.1. Activity analysis of *GhTrp*

Several attempts at identifying potential substrates using small chromogenic peptide-based substrates as well as proteomic identification of protease cleavage sites (PICS) analysis against a bacterial (*E. coli*) peptide library (Eckhard *et al.*, 2016) failed to demonstrate any measurable catalytic activity for *GhTrp* (data not shown).

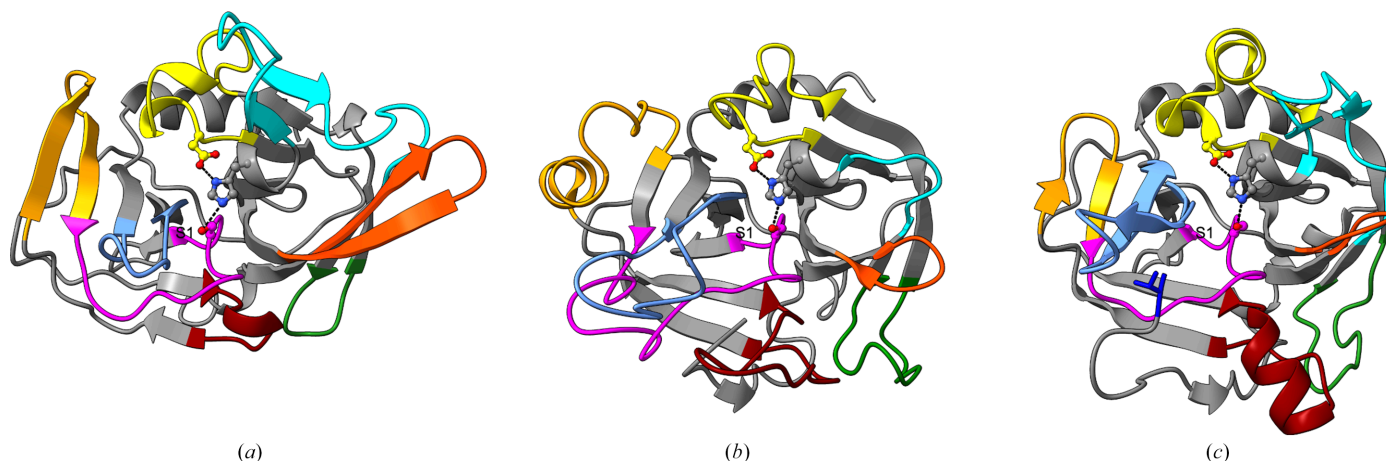
#### 3.2. Structure solution

A crystallographic property present in the *GhTrp* crystal structure prevented initial structure solution. Due to the presence of translational noncrystallographic symmetry (tNCS) in the crystal, the structure was unable to be solved using simple molecular-replacement strategies. The tNCS was identified by *phenix.xtriage* (Zwart *et al.*, 2005), which showed a strong off-origin Patterson peak at  $(u, v, w) = (0.00, 0.06, -0.50)$  with a height of 28% of the Patterson origin peak. This structure was solved at a time (early 2021) when structural modelling techniques had yet to reach the more accurate predictive capabilities of *AlphaFold* (Jumper *et al.*, 2021) and *RosettaFold* (Baek *et al.*, 2021). The best search model identified through sequence alone only had ~30% sequence identity and a  $C^\alpha$  r.m.s.d. of ~2.5 Å (PDB entry 1dt2), which may have been sufficient for determining phases if not for the

artefacts associated with tNCS interfering with molecular-replacement techniques (Read *et al.*, 2013). This was nevertheless a better search model than the ~3.0 Å  $C^\alpha$  r.m.s.d. homology model predicted by *I-TASSER* at that time (Supplementary Fig. S2; Roy *et al.*, 2010). The structure was ultimately solved using *phenix.mr\_rosetta* as this was one of the first programs that incorporated *ab initio* model building as part of the phasing process (DiMaio *et al.*, 2011; Terwilliger *et al.*, 2012). As expected, the *GhTrp* crystal structure depicts two molecules in the asymmetric unit, related to each other along the *c* axis by a tNCS vector of approximately half the *c*-axis length (Supplementary Fig. S3).

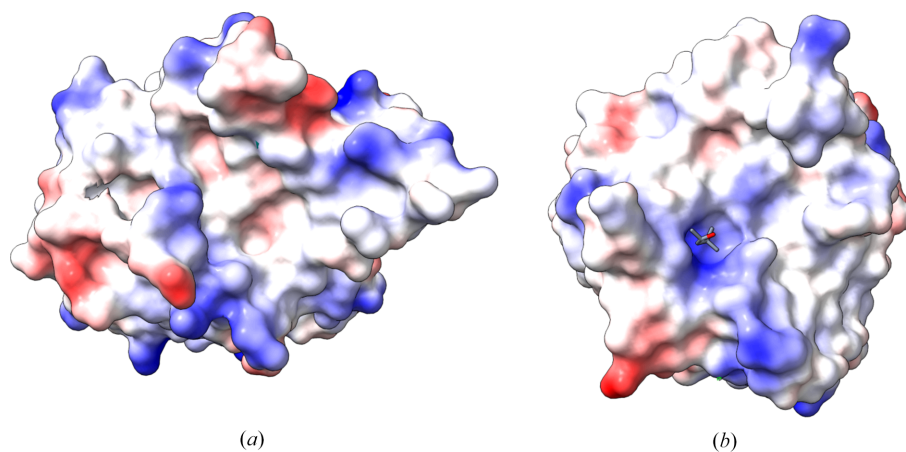
#### 3.3. The general fold shows modifications to trypsin-like specificity loops

Despite having low sequence identity (<20%) to most known chymotrypsin-like and trypsin-like proteases, *DALI* analysis (Holm, 2022) demonstrates that the fold of *GhTrp* is consistent with other members of the S1 family of glutamyl endopeptidases, as categorized by the MEROPS database (Rawlings *et al.*, 2018). *GhTrp* exhibits reasonable overall structural homology with this family of glutamyl endopeptidases, with the best-aligning structures having *DALI* scores of >18.5 and overall  $C^\alpha$  r.m.s.d. values of between 2 and 2.8 Å despite sequence identities of 21% or less (Table 2).



**Figure 1**

A comparison of the loop structures of (a) *GhTrp* (PDB entry 9ect), (b) bovine trypsin (PDB entry 1hj9) and (c) *Bacillus intermedius* glutamyl-endopeptidase (PDB entry 1p3c). The known specificity loops, loop A (37 loop; dark orange), loop B (60 loop; cyan), loop C (99 loop; yellow), loop D (148 loop; maroon), loop E (75 loop; green), loop 1 (189 loop; magenta), loop 2 (220 loop; light blue) and loop 3 (175 loop; orange), are illustrated in each structure with the remaining protein rendered in grey. In (c), the location of the N-terminus is indicated by the N-terminal leucine residue rendered as a blue stick model. Potential interactions between members of the catalytic triad are rendered as dashed lines and the location of the S1 pocket is annotated. All molecules are presented in an identical orientation.



**Figure 2**

A comparison of the electrostatic surfaces of (a) *GhTrp* (PDB entry 9ect) and (b) *B. intermedius* glutamyl-endopeptidase (PDB entry 1p3c). In (b) the position of the S1 binding pocket is indicated by the MPD molecule that was co-crystallized (grey sticks coloured by atom type). In the *GhTrp* structure (a), the S1 pocket site is occluded by the structure of loop 2. Both molecules are presented in the same orientation as in Fig. 1.

The trypsin-like fold has been well characterized and the involvement of the many surface loops as determinants of subsite selectivity for peptide and protein substrates has been well documented (reviewed in Goettig *et al.*, 2019). An analysis of these surface loops in the structure of *GhTrp* demonstrates that there are considerable differences in the structures of the specificity loops between the classic trypsin structure and *GhTrp*, with the exception of loop C. Comparisons between the structures of *GhTrp*, bovine trypsin and *Bacillus intermedius* glutamyl peptidase (BGP; Fig. 1) demonstrate that loops A and B are considerably larger in bovine trypsin and loops D and E are shorter in *GhTrp* than either of the other two enzymes. *GhTrp* therefore lacks the calcium-binding residues that stabilize the more elongated loop structure in bovine trypsin and thus no ions are observed in the structure of *GhTrp* (Leiros *et al.*, 2001).

In *GhTrp* and BGP, loop 3 forms additional  $\beta$ -strands that extend the core  $\beta$ -sheet, which is quite different from the helical structure found in bovine trypsin. Loop 1, which contains the serine nucleophile (Ser167) and the oxyanion-hole residues (amides of Ser167/Gly165), is similar in structure between *GhTrp* and BGP but is truncated when compared with bovine trypsin. This may be a consequence of their correspondingly truncated loops 2, which act as a supporting structure for the placement of loop 1. As both loops 1 and 2 are truncated in *GhTrp* relative to bovine trypsin, loop 2 is still able to function as a backing structure for loop 1 in the fold.

Most notably, the conformation of loop 2 of *GhTrp* places it in the middle of the putative S1 pocket, bifurcating the substrate-binding groove (Fig. 2). This malformed S1 pocket is consistent with the functional data that demonstrate a lack of proteolytic function for this enzyme construct. In contrast, the the prime-side subsites are well structured. Taken together, these data suggest that the *GhTrp* structure could represent a pro-enzyme-like form of the putative zymogen in which some activation event is required to properly stabilize loop 2 in an active conformation to generate a viable S1 pocket. One could argue that the bifurcation of the S1 pocket may be the result

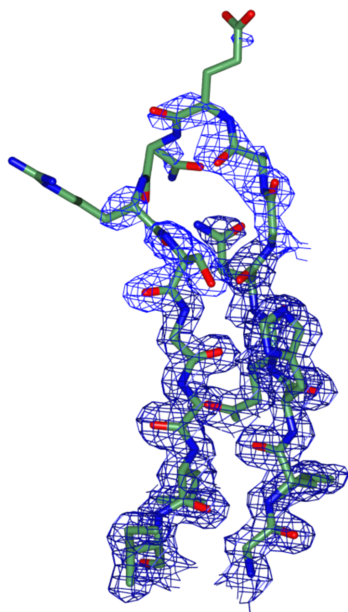
of characterizing *GhTrp* outside the context of *GhIgAP*. However, the *AlphaFold3* model of full-length *GhIgAP* shows a similar conformation of loop 2 in which it still bifurcates the S1 pocket, consistent with the persistence of the pro-enzyme-like conformation in the full-length enzyme (Supplementary Fig. S1b).

The N-termini of many trypsin-like serine proteases have been shown to regulate protease activation and activity. For example, the N-terminal helix of the *Staphylococcus aureus* exfoliative toxin A stabilizes the S1 pocket and deletions in the N-terminal region abolish activity (Cavarelli *et al.*, 1997). An alternative explanation for this substrate-binding-groove bifurcation comes from examining the structure of BGP, where zymogen activation liberating the N-terminal leucine residue stabilizes a correct loop 2 conformation and formation of the S1 pocket (Fig. 1c; Meijers *et al.*, 2004). In the crystallized construct, the N-terminus of *GhTrp* is too short to interact with loop 1 to stabilize an open, active conformation. Even if the N-terminus is extended by  $\sim 30$  amino acids, *GhTrp* remained inactive and this extra N-terminal tail was shown to lack a defined structure (Redzic *et al.*, 2022). If the *GhTrp* structure truly depicts a pro-enzyme, the activation mechanism for BGP suggests that the N-terminus of *GhTrp* must be cleaved at a specific site to properly activate *GhTrp*. Further support for this activation mechanism comes from the electron-density maps corresponding to loop 2, in which the distal end of this loop (residues 868–872) is poorly ordered in the crystal structure in its modelled conformation (Fig. 3). However, based upon the data, we cannot rule out the possibility that this domain of *GhIgAP* evolved from the trypsin protease fold, lost its ability to function as a protease and acquired a different, but as of yet unknown, function.

#### 4. Conclusions

The crystal structure of *GhTrp* was solved to gain insight into the potential functions of this domain despite difficulties in finding a substrate for the putative enzyme. These structural





**Figure 3**  
Apparent disorder in loop 2 (863–877) of *GhTrp*. The backbone and side chains are represented as stick models and coloured by atom type with C atoms in green.  $2F_o - F_c$  density at  $1\sigma$  is rendered as a blue mesh

data showed that the lack of activity observed is unsurprising due to the aberrant position of loop 2 occluding the S1 pocket in the crystal structure and *AlphaFold* model. Based upon this result, we hypothesize that the current structure of *GhTrp* represents the pro-enzyme structure of the enzyme that is present in the full-length M26 IgAP. We hypothesize that this putative pro-enzyme form must undergo a specific cleavage event to generate an N-terminal segment that interacts with loop 2 to stabilize a more open and active conformation of the S1 pocket.

### Acknowledgements

Part of the research described in this paper was performed using beamline CMCF-BM at the Canadian Light Source, a national research facility of the University of Saskatchewan, which is supported by the Canada Foundation for Innovation (CFI), the Natural Sciences and Engineering Research Council (NSERC), the National Research Council (NRC), the Canadian Institutes of Health Research (CIHR), the Government of Saskatchewan and the University of Saskatchewan.

### Conflict of interest

The authors declare no conflicts of interest.

### Data availability

The model coordinates and structure factors for *GhTrp* have been deposited in the PDB (<https://www.rcsb.org/pdb>) under accession code 9ect. The PDB deposition is cross-referenced with residues 661–873 of UniProt entry C5NYF3. It should be noted that this entry contains a sequence that lacks 23

N-terminal residues relative to the *GhIgAP* sequence used in this and previous literature (GenBank WP\_040464465.1; residues 684–896).

### Funding information

This work was supported in part by funds provided through a Discovery Grant issued to TH by the Natural Sciences and Engineering Research Council of Canada (NSERC).

### References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst. D* **68**, 352–367.

Agirre, J., Atanasova, M., Bagdonas, H., Ballard, C. B., Baslé, A., Beilsten-Edmands, J., Borges, R. J., Brown, D. G., Burgos-Mármol, J. J., Berrisford, J. M., Bond, P. S., Caballero, I., Catapano, L., Chojnowski, G., Cook, A. G., Cowtan, K. D., Croll, T. I., Debreczeni, J. É., Devenish, N. E., Dodson, E. J., Drevon, T. R., Emsley, P., Evans, G., Evans, P. R., Fando, M., Foadi, J., Fuentes-Montero, L., Garman, E. F., Gerstel, M., Gildea, R. J., Hatti, K., Hekkelman, M. L., Heuser, P., Hoh, S. W., Hough, M. A., Jenkins, H. T., Jiménez, E., Joosten, R. P., Keegan, R. M., Keep, N., Krissinel, E. B., Kolenko, P., Kovalevskiy, O., Lamzin, V. S., Lawson, D. M., Lebedev, A. A., Leslie, A. G. W., Lohkamp, B., Long, F., Malý, M., McCoy, A. J., McNicholas, S. J., Medina, A., Millán, C., Murray, J. W., Murshudov, G. N., Nicholls, R. A., Noble, M. E. M., Oeffner, R., Pannu, N. S., Parkhurst, J. M., Pearce, N., Pereira, J., Perrakis, A., Powell, H. R., Read, R. J., Rigden, D. J., Rochira, W., Sammito, M., Sánchez Rodríguez, F., Sheldrick, G. M., Shelley, K. L., Simkovic, F., Simpkin, A. J., Skubak, P., Sobolev, E., Steiner, R. A., Stevens, K., Tews, I., Thomas, J. M. H., Thorn, A., Valls, J. T., Uski, V., Usón, I., Vagin, A., Velankar, S., Vollmar, M., Walden, H., Waterman, D., Wilson, K. S., Winn, M. D., Winter, G., Wojdyr, M. & Yamashita, K. (2023). *Acta Cryst. D* **79**, 449–461.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). *Science*, **373**, 871–876.

Cavarelli, J., Prévost, G., Bourguet, W., Moulinier, L., Chevrier, B., Delagoutte, B., Bilwes, A., Mourey, L., Rifai, S., Piémont, Y. & Moras, D. (1997). *Structure*, **5**, 813–824.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwai, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature*, **473**, 540–543.

Eckhard, U., Huesgen, P. F., Schilling, O., Bellac, C. L., Butler, G. S., Cox, J. H., Dufour, A., Goebeler, V., Kappelhoff, R., Keller, U., Klein, T., Lange, P. F., Marino, G., Morrison, C. J., Prudova, A., Rodriguez, D., Starr, A. E., Wang, Y. & Overall, C. M. (2016). *Matrix Biol.* **49**, 37–60.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst. D* **66**, 486–501.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst. D* **69**, 1204–1214.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. (2005). *The Proteomics Protocols Handbook*, edited by J. M. Walker, pp. 571–607. Totowa: Humana Press.

Goettig, P., Brandstetter, H. & Magdolen, V. (2019). *Biochimie*, **166**, 52–76.

Holm, L. (2022). *Nucleic Acids Res.* **50**, W210–W215.

Johnson, T. A., Qiu, J., Plaut, A. G. & Holyoak, T. (2009). *J. Mol. Biol.* **389**, 559–574.

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.
- Leiros, H.-K. S., McSweeney, S. M. & Smalås, A. O. (2001). *Acta Cryst. D* **57**, 488–497.
- Meijers, R., Blagova, E. V., Levnikov, V. M., Rudenskaya, G. N., Chestukhina, G. G., Akimkina, T. V., Kostrov, S. V., Lamzin, V. S. & Kuranova, I. P. (2004). *Biochemistry*, **43**, 2784–2791.
- Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A. & Finn, R. D. (2018). *Nucleic Acids Res.* **46**, D624–D632.
- Read, R. J., Adams, P. D. & McCoy, A. J. (2013). *Acta Cryst. D* **69**, 176–183.
- Redzic, J. S., Rahkola, J., Tran, N., Holyoak, T., Lee, E., Martín-Galiano, A. J., Meyer, N., Zheng, H. & Eisenmesser, E. (2022). *Commun. Biol.* **5**, 1190.
- Roy, A., Kucukural, A. & Zhang, Y. (2010). *Nat. Protoc.* **5**, 725–738.
- Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
- Terwilliger, T. C., DiMaio, F., Read, R. J., Baker, D., Bunkóczi, G., Adams, P. D., Grosse-Kunstleve, R. W., Afonine, P. V. & Echols, N. (2012). *J. Struct. Funct. Genomics*, **13**, 81–90.
- Tran, N., Frenette, A. & Holyoak, T. (2024). *bioRxiv*, 2024.12.31.630911.
- Wang, Z., Rahkola, J., Redzic, J. S., Chi, Y.-C., Tran, N., Holyoak, T., Zheng, H., Janoff, E. & Eisenmesser, E. (2020). *Nat. Commun.* **11**, 6063.
- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (2018). *Protein Sci.* **27**, 293–315.
- Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst. D* **74**, 85–97.
- Zwart, P., Grosse-Kunstleve, R. W. & Adams, P. (2005). *CCP4 Newsl.* **43**, 7.