



STRUCTURAL  
BIOLOGY

**Volume 79 (2023)**

**Supporting information for article:**

**Crystal structure of a polyglycine hydrolase determined using a *RoseTTAFold* model**

**Nicole V. Dowling, Todd A. Naumann, Neil P. J. Price and David R. Rose**

```

Bz-cmp      SPTTNQNLPSSKR-DATVEQN-----VFFDINGKEVETRIAKLKAEGYRPTSLNIHGTP
Es-cmp      FPANQQDLTFAKR-NGTFEQS-----VFYGLTGPEVEAKLAKLKADGYRPTSLNIHGST
Fvan-cmp    -----LPNQRRSNVTSHVE-----TYYSVDGATHAEKSKALKADGYRIVSLSSYGSP
Gm_CMP      -SPTDLSLKRTVPSNNTLSSRSIPAWQFYGVGTGAQHQTSTFNQWSVSGYRMISLSAYGQP
           *   : *           : : *           *** * * : *

Bz-cmp      TDAKYAGIWIQKQDGNPYETIVGANKTTYDAWLQWKASGYVSTQVSATGPASNAVFAGVM
Es-cmp      SDAKYAGIWTQKQGDDFETILGANKTVYDAWLDSHKAQGYVSTHVSATGGSSDALFAGVM
Fvan-cmp    DSANYAAIWVQEEGPSFEIHD ADEATYNSWLQTWKSRGYVSTQVSATGPAENAVFAGVM
Gm-CMP      PNHRYAAVWVQRSGPAYFAIHEASGSQYQTFSDSHTPNGVSTVITVTPANAAVYAGVM
           ** : * : * : * : * : * : * : * : * : * : * : * : * : * : * : *

Bz-cmp      EKIPSIRNWTQICGTSSP-YAYLNATGDAP-IVITGVSlyGTPTERQYCVLGHEDTDNHQ
Es-cmp      EKVPSVANWIQVCGLDNP-YAYANATIDEP-MYIKGVSMYGAPNERQYCILGHENLVNYQ
Fvan-cmp    ENIN-VANWFQSCLENP-WAFSNTTGNVD-VVVKGFRMFGTPEERRYCILGHENVGNEQ
Gm-CMP      EQNG-VTNWYQKCGLTNAQYATELNNGFANRYFLKSFTYGTSSDRRRCGIWYFNDQYDK
           * : : * * * * : * : : * : * : * : * : :

Bz-cmp      QTVWFPTS-SSQDYKTIEPAETTKRFWRPVIDSSEDKLLASIFDDTSVGNWTVRDLTA
Es-cmp      QTVFYQTDYFKKDYAKLLQSETSKRHWRPVFDLSEDLLPTPIFDDTSVGQWVARTLSA
Fvan-cmp    TTIQYSTPSFTVNFASFETTKRFWRPRLFLSEDHITPSFADTSVGKWSHAVDLTK
Gm-CMP      YTAWYDQS--YSGYQTTFAETTKPFWRPSYLSVSEDHLISSAFVDTDVGSWVARHAMTA
           * : : : * * * * * * : * * * : * * * * * * : :

Bz-cmp      SQLQSEITAQKANNMHPHISGAGSTET-QYVIFTEQATPLERNWHAVGDITGFPDNAR
Es-cmp      SELEAEIAAQKAKNLYAVHIA GAGSKGS-KYAVLFAEHLSPLEKRWTVTGEVTGFKTNDV
Fvan-cmp    AELKEKTIETRAKGLYPIDIQGGGSGSSERFTVFAERTSPKPRQWNVVRGEITGFEDNKA
Gm-CMP      SDLQAEYQTQLSAGRYIIHLQGGGTGANANYAALWATQDIPTRFRATGSITGFQNNQA
           : * : : : : * * : : : : * * : * : * * *

Bz-cmp      VKRDLDEILQKFMKRNSVRQAQV SASINGTVIASRAYTWAESNRAIVEPSDKFLLGSVSK
Es-cmp      VAKDMDAVMEEFMKNNSVRQAQVAASVNGTVAERSYTWAEEDRAVVKPTDKFGLGSVSK
Fvan-cmp    AEEVVDSIMRRFMEKNGVRQAQFAVALEGKTIERSYTWAEEDRAIVEPDDIFLLASVSK
Gm-CMP      ASAQADALMQAWMKTNVGRQAQFTVGKNGNILLQKGSWSEATRHHTTAPNDVFLASNSK
           : * : : * : * * * * : * : : * : * * * * * * * * * * * *

Bz-cmp      AFTYAAVDHLVSTGVVNLTDHIYPLLGYTKPADPRSLNITIQHLLDHTGGFDRGMS-PDI
Es-cmp      MFTYAAATNLLNEGLLNHTTRVYPFLGMNPNADNRSLDITVDHLLQHTAGYNRDIK-PDI
Fvan-cmp    MFLHASIDWLVSMDLNFSTPVYDLLGY-KPADSRANDINVQHLLDHSAGYDRSMS-GDP
Gm-CMP      MFVEAAVQTLYNSNKLTAATKVYPLLGYTKTTPRLQKITVDQLLTHYGGNLDTASGFDP
           * * : * : : * * * * : * * : * : * * * * * * :

Bz-cmp      GFIFTTVAQSL--NQSTPATLRQLIEYVLAKPLDFEPGTRSVYSNYGTLGLYLIANKTG
Es-cmp      GFIFRNIALER--NQTPVSLRELIEYVVEQPLDFTPGTDVYSNYGTMLLSYLIANITG
Fvan-cmp    SFMFRERIAQSLPTKGAKAATLRDVIEYVVAKPLDFTPGDYSAYSNYCPMLLSYVVTNITG
Gm-CMP      TYRMRDIAIAQG-TGASPASVKNIVDYSRYTLDYNPGGGYAYSNYGYLLSYVVEHVTG
           : : * * : : : : * : * * * * * * * * * * * *

Bz-cmp      ETYMSYLEKNVLK--GLDVEMYTPGADHTNDRIVQETKFTSISALTPLEKRVANVYGG
Es-cmp      ESFNYSYIHKVNLN--GLDVELYPTSPELNANNPQVETKYTFYPAQDPASTKQVSNANGG
Fvan-cmp    VPYLDLEKNILD--GLNVRLYETAASKHTEDRIVQESKNTGQDPVHPQSAKLVPGPHGG
Gm-CMP      MAYDYDYLSSALLTPGGYDVRKWVTPSAHASDPITQESAYTLNAAALPQSQTQVAAVFGG
           : : : * * : * : * : : * * * : * * * * * *

Bz-cmp      DGAVKEAAVSSFALKASADTVSRFIGKHAAYGLGPRQLFM-----YRDGTVAGARALAYS
Es-cmp      DGSIREEAIGAFGLRASASTISQFLANHAAYDIGPRQAYT-----YRDGTIVGSRAFAQS
Fvan-cmp    DGAVKEECAGTFAMAASASSLAKFIGSHAVWGTGGR-VSS-----NRDGSLSGARAYVES
Gm-CMP      DGMYKEDAFGPASLAASATTLANFIHTHAVWGIGPRPTGTGQWTWARSGSTPGTRTWAQS
           ** : * : * * * : * : * * : * * : * * : * * : * * : *

Bz-cmp      MA-KLDWAITLNTREYLDENA-----WNTLVFTDLYQLWGALDSALPL
Es-cmp      QD-LIDWSLILNNTREYSEK-----WEQLVFGPISQWY--KYALAE
Fvan-cmp    RG-TIDWALTNTREYISETE-----FDELRWYSLPDLFS--AFPIAG
Gm_CMP      QWNGLDWAVTVNTRDFPPNSGQDPFDDVLCESLPGFLN--AYPTV-
           : * : : * * : : : * : :

```

**Figure S1** Multiple sequence alignment. The mature protease sequences were compared by multiple sequence alignment. Conserved amino acids are in grey. Stars represent conserved amino acids; colons represent similar amino acids. The two catalytic motifs are indicated with black boxes.

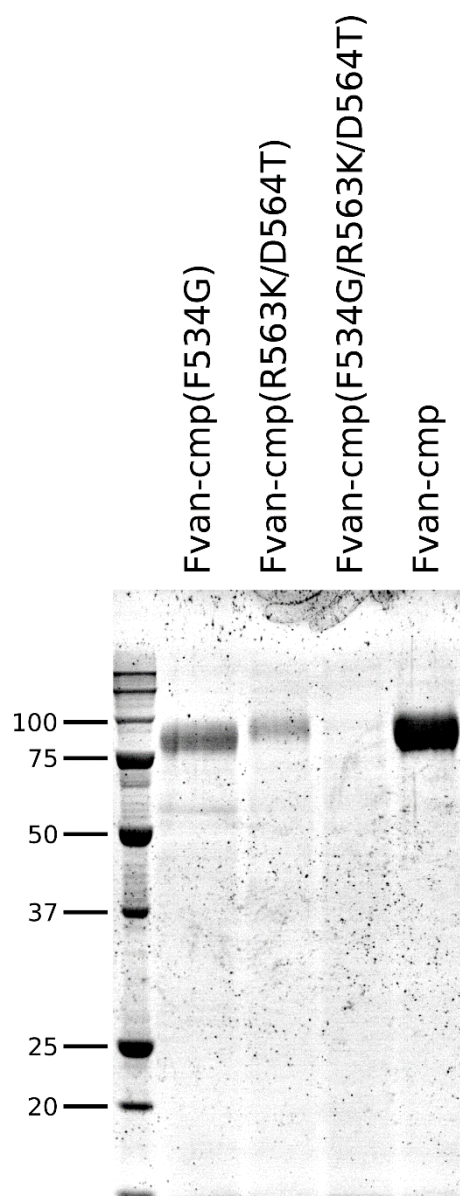
**A**

EFLPNQRRSNVTSHVETYYSVGDGATHAEKSKALKADGYRIVSLSSYGSPDSANYAAIWVQEEGPSFE  
 IIHDADEATYNSWLQTWKSRGYVSTQVSATGPAENAVFAGVMENINVANWFQSCLENPWAFSNT  
 TGNVDVVVKGFRMFGTPEERRYCILGHENVGNEQTTIQYSTPSFTVNFASFTEAETTKRFRWRPRLF  
 LSEDHITPSFADTSVGKWSHAVDLTKAELKEKIETERAKGLYPIDIQGGGSGSSERFTVVFA**EERTSP**  
**KPRQWNVRGEITGFEDNKAAEEEVDSIMRRFMEKNGVRQAQFAVALEGKTI**AERSYTWAEEDRAI  
 VEPDDIFLLASVSKMFLHASIDWLVSHDMLNFSTPVYDLLGYKPADSRANDINVQHLLDHSAGYDR  
 SMSGDPSFMFREIAQSLPTKGAKAATLRDVIEYVVAKPLDFTPGDYSAYSNYCPMLLSYVVTNITGV  
 PYLDFLEKNILDGLNVRLYETAASKHTEDRIVQESKNTGQDPVHPQSAKLVPGPHGGDGAVKEECA  
 GTFAMAASASSLAKFIGSHAVWGTGGRVSSNRDGSLSGARAYVESRGTIDWALTLNTREYISETEFD  
 ELRWYSLPDFLSAFPIAG

**B**

EFLPNQRRSNV<sub>15</sub>TSHVETYYSVGDGATHAEKSKALKADGYRIVSLSSYGSPDSANYAAIWVQ<sub>60</sub>EEGP  
**S<sub>66</sub>EIIHDADEATYNSWLQTWKSRGYVSTQVSATGPAENAVFAGVME<sub>110</sub>NINVAN<sub>116</sub>WFQSCLEN**  
**PWAFSNTTGNVDVVVKGFRMFGTPEERRYCILGHE<sub>160</sub>NVGNEQT<sub>167</sub>TIQYSTPSFTVNFASFTEAETT**  
**KRFRWRPRLFSEDHITPSFA<sub>211</sub>DTSVGKW<sub>218</sub>SHAVDLTKAELKEKIETERAKGLYPIDIQGGGSGSS**  
**ERFTVVFA<sub>262</sub>ERTSPKPRQWNVRGEITGFEDNKAAEEEVDSIMRRFMEKNGVRQAQFAVALEGKTI**  
 AERSYTWAEEDRAIVEPDDIFLLASVSKMFLHASIDWLVSHDMLNFSTPVYDLLGYKPADSRANDIN  
 VQHLLDHSAGYDRSMSGDPSFMFREIAQSLPTKGAKAATLRDVIEYVVAKPLDFTPGDYSAYSNYC  
 PMLLSYVVTNITGVPLYLDFLEKNILDGLNVRLYETAASKHTEDRIVQESKNGQDPVHPQSAKLVPGP  
 HGGDGAVKEECAGTFAMAASASSLAKFIGSHAVWGTGGRVSSNRDGSLSGARAYVESRGTIDWAL  
 TLNTREYISETEFDELRWYSLPDFLSAFPIAG

**Figure S2** (A) The full-length sequence of Fvan-cmp protease. The N-domain spans residues 1-282, C-domain spans 271-616 with the linker region spanning residues 263-270. The linker region is bolded for visualization within the sequence. (B) The N-domain consists of five structural repeats connected by a small loop. The EHEE repeats are indicated by colour **VAL<sub>15</sub> - GLN<sub>60</sub>** (1), **PHE<sub>66</sub> - GLU<sub>110</sub>** (2), **ASN<sub>116</sub> - GLU<sub>160</sub>** (3), **THR<sub>167</sub> - ALA<sub>211</sub>** (4) and **TRP<sub>218</sub> - ALA<sub>262</sub>** (5).



**Figure S3** SDS-PAGE and Coomassie staining analysis of cell-free media. After two days of expression, cell-free media from *K. phaffii* cultures expressing either Fvan-cmp or the single, double, or triple mutants was analyzed. The site-directed mutants all showed reduced amounts of protein, likely due to decreased stability of the proteins.

**Table S1** Top 50 FoldSeek hits against the AlphaFold Uniprot Database

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-A0A166M196	59.1	1480	1.83E-32	<i>Colletotrichum incanum</i>	Penicillin-binding protein
AF-A0A135SQZ1	54.6	1462	4.87E-32	<i>Colletotrichum salicis</i>	Pectate lyase
AF-A0A1B7YR40	55.3	1438	1.80E-31	<i>Colletotrichum higginsianum IMI 349063</i>	Penicillin-binding protein
AF-N4V2J0	56.1	1400	1.42E-30	<i>Colletotrichum orbiculare MAFF 240422</i>	Beta-lactamase containing domain
AF-A0A2V1DDC1	53.7	1361	1.18E-29	<i>Periconia macrospinosa</i>	Beta-lactamase/transpeptidase-like protein
AF-A0A4R8RCR4	48.6	1230	1.45E-26	<i>Colletotrichum trifolii</i>	flp gene product
AF-A0A6A5RWU7	44.9	1227	1.71E-26	<i>Didymella exigua CBS 183.55</i>	Beta-lactamase/transpeptidase-like protein
AF-A0A3M7MH02	43.3	1197	8.73E-26	<i>Pyrenophora seminiperda CCB06</i>	Penicillin-binding protein
AF-A0A4Q6A4M9	29.8	1092	2.62E-23	<i>Sphingobacteriales bacterium</i>	Class A beta-lactamase related serine hydrolase
AF-A0A067T494	36.7	1086	3.63E-23	<i>Galerina marginata CBS 339.88</i>	Uncharacterized
AF-A0A1J9QVD6	37.7	1081	4.77E-23	<i>Diplodia corticola</i>	Beta-lactamase containing domain
AF-A0A321LBL7	33.4	1067	1.02E-22	<i>Blastocatellia bacterium AA13</i>	Beta-lactamase containing domain
AF-A0A6A6B650	35.6	1067	1.02E-22	<i>Aplosporella prunicola CBS 121167</i>	Beta-lactamase containing domain
AF-K3V3V4	33.6	1065	1.14E-22	<i>Fusarium pseudograminearum CS3096</i>	Beta-lactamase containing domain
AF-A0A409Y2P4	37.7	1044	3.56E-22	<i>Gymnopilus dilepis</i>	Beta-lactamase containing domain
AF-A0A067T7P9	30.9	1026	9.46E-22	<i>Galerina marginata CBS 339.88</i>	Uncharacterized
AF-A0A1L7TAS3	41.9	972	1.78E-20	<i>Fusarium mangiferae</i>	Beta-lactamase containing domain

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-A0A3M9ZDX4	27.1	965	2.60E-20	<i>Leptolyngbya sp. IPPAS B-1204</i>	Class A beta-lactamase related serine hydrolase
AF-A0A849TPW4	23.7	926	2.17E-19	<i>Nitrospira sp.</i>	Uncharacterized
AF-A0A7W7CQE2	29.8	925	2.29E-19	<i>Actinoplanes abujensis</i>	Uncharacterized
AF-A0A532CUY9	24.9	912	4.64E-19	<i>Nitrospira sp.</i>	Uncharacterized
AF-A0A7W1TEK5	29.3	910	5.17E-19	<i>Planctomycetes bacterium</i>	Serine hydrolase
AF-A0A7W0KYF8	29.4	905	6.78E-19	<i>Acidimicrobiia bacterium</i>	Serine hydrolase
AF-A0A838DV87	25.6	877	3.10E-18	<i>Ktedonobacteraceae bacterium</i>	Serine hydrolase
AF-A0A6H9YRX4	24.6	873	3.86E-18	<i>Actinomadura rudentiformis</i>	Uncharacterized
AF-A0A7G5IK55	22.1	872	4.07E-18	<i>Sandaracinobacter sp. M6</i>	Serine hydrolase
AF-A0A5J6P455	25.8	869	4.80E-18	<i>Cellvibrio sp. KY-GH-1</i>	Class A beta-lactamase related serine hydrolase
AF-A0A3N1JPB1	23.1	843	1.97E-17	<i>Granulicella sp. GAS466</i>	Beta-lactamase
AF-A0A2V8HSC7	25.6	823	5.84E-17	<i>Acidobacteria bacterium</i>	Beta-lactamase containing domain
AF-A0A3A4B505	25.6	794	2.82E-16	<i>Bailinhaonella thermotolerans</i>	Non-specific serine/threonine protein kinase
AF-A0A2J6QYV7	18.9	784	4.86E-16	<i>Hyaloscypha variabilis F</i>	Beta-lactamase containing domain
AF-A0A2L2U0F8	29	752	2.76E-15	<i>Fusarium venenatum</i>	Beta-lactamase containing domain
AF-D2B7Z4	19.7	746	3.83E-15	<i>Streptosporangium roseum DSM 43021</i>	Beta-lactamase
AF-A0A3N7JU83	21.1	743	4.51E-15	<i>Albitalea terrae</i>	Class A beta-lactamase related serine hydrolase
AF-A0A7Y6IQ54	22.1	733	7.76E-15	<i>Nonomuraea rhodomycinica</i>	Uncharacterized
AF-A0A1H1BZ32	22.5	724	1.27E-14	<i>Thermostaphylospora chromogena</i>	Uncharacterized
AF-A0A7X0U1N1	17.8	723	1.34E-14	<i>Nonomuraea rubra</i>	Uncharacterized
AF-A0A2J6SIY4	33.3	719	1.66E-14	<i>Hyaloscypha bicolor E</i>	Beta-lactamase/transpeptidase-like protein

AlphaFold Entry	Percent Identity	FoldSeek Score	E-value	Organism	Function
AF-A0A367FK14	20.8	716	1.96E-14	<i>Sphaerisporangium album</i>	Uncharacterized
AF-A0A7W8ECQ3	22.5	696	5.80E-14	<i>Nonomuraea endophytica</i>	Non-specific serine/threonine protein kinase
AF-A0A5R8MSZ6	18.8	695	6.12E-14	<i>Nonomuraea sp. KC401</i>	Uncharacterized
AF-A0A5S4FIC2	20.1	694	6.46E-14	<i>Nonomuraea turkmeniaca</i>	Uncharacterized
AF-A0A848DM22	15	691	7.60E-14	<i>Pseudonocardia bannensis</i>	Beta-lactamase containing domain
AF-A0A0J9ECQ5	19.3	683	1.17E-13	<i>Candidatus Rhodobacter lobularis</i>	Beta-lactamase containing domain
AF-A0A553Y292	16.5	676	1.72E-13	<i>Streptomyces benahoarensis</i>	Beta-lactamase containing domain
AF-A0A2H3RIQ8	35.4	672	2.14E-13	<i>Fusarium fujikuroi</i>	Beta-lactamase containing domain
AF-A0A124DZT4	22.9	659	4.33E-13	<i>Mycolicibacterium brisbanense</i>	Beta-lactamase containing domain
AF-A0A239BKE8	20.3	646	8.77E-13	<i>Streptosporangium subroseum</i>	Uncharacterized
AF-A0A4Q5NQF3	19.2	626	2.60E-12	<i>bacterium</i>	Beta-lactamase containing domain
AF-A0A022VXE6	14	624	2.90E-12	<i>Trichophyton rubrum CBS 288.86</i>	Uncharacterized

The matches are sorted based on their FoldSeek score metric. All proteins that matched the Fvan-cmp N-domain, none characterized the purpose of the tertiary fold. Within the top 50 results, 20 proteins were from the fungal domain and 30 proteins were from the bacterial domain.

**Table S2** Conserved  $\beta$ -lactamase shell residues

Residue	Fvan-cmp residue*	Shell	Function
E37 (A, D, Q, S)	Q308	III	Proper folding
R65 (A, T, P, L, H, K, C)	P335	III	Loop interactions
T71 (S, A, V, L)	L341	II	Reduces mobility of active site serine
D131	D403	II	Stability and global positioning
A185 (S, T, V, E, Q, R, N, G)	S456	III	
W229 (S, A, Y, C, F)	S501	III	Hydrophobic and stacking interactions

\*The Fvan-cmp C-domain begins at residue 271, so the corresponding residue was scaled up to fit the full sequence length.

The residues conserved across Class A  $\beta$ -lactamases and are designated to shells based on proximity to the active site and their related function. The bracketed residues are the conserved alternates found in nature.

Conserved residues shared between polyglycine hydrolases and  $\beta$ -lactamases are listed.