The structure and dynamics of water molecule networks underlie catalytic efficiency in a glycoside exo-hydrolase

Check for updates

Sukanya Luang[1,5], Xavier Fernández-Luengo[2,5], Victor A. Streltsov[3], Jean-Didier Maréchal[2], Laura Masgrau [2,4] ✉ & Maria Hrmova [1] ✉

Glycoside hydrolases break glycosidic bonds by transferring a water molecule onto the glycosidic oxygen of carbohydrates, but on the nanoscale, the dynamics of water molecules remains unclear. We investigate the role of the non-nucleophilic E220 glutamate, essential for maintaining the water molecule network in a family 3 β-D-glucan glucohydrolase, but not involved directly in catalysis. Kinetic data disclose that the E220A mutant retains substrate poly-specificity but has drastically reduced catalytic efficiency compared to the wild-type. High-resolution structures in-complex with a hydrolytic product and a mechanism-based inhibitor reveal that in wild-type, the concatenated water molecules near acid/base E491 and neighbouring N219 and E220 form a harmonised network. In contrast, in the E220A mutant, this network is uncoordinated. Computational models of covalent complexes show that water flux through the wild-type protein correlates with high catalytic efficiency dissimilar to E220A, where this correlation is lost. Ancestral sequence reconstructions of family 3 enzymes divulge the evolutionary conservation of residues participating in water molecule networks, which underlie substrate-product-assisted processivity. Our findings provide a blueprint for the dynamics of catalysis mediated by hydrolytic enzymes, which could inspire bioengineering to create a sustainable bio-economy.

Water is central to life and fulfils multiple metabolic roles as solvent, substrate, intermediate, cofactor, and product during chemical transformations in living cells[1,2]. In the Mechanism and Catalytic Site Atlas Database[3], 57% of entries use water in chemical reactions, and approximately a third to a half of these reactions, classed in the Enzyme Catalytic-mechanism Database v3.0, utilise or generate water[4].

Water and its availability are critical to mechanical support, thermal coupling, dielectric screening, mass and charge transport, and ligand competition, where water thermodynamics appears to be a driving force for protein folding and macromolecular interactions[5,6] and regulating biomolecular condensation[7,8]. In these roles, structure-based water prediction algorithms envisage water molecule locations and how to incorporate them in a protein environment of buried pockets[9]. In some enzymes, such as dehalogenase, time-dependent fluorescence shifts and molecular dynamics (MD) simulations revealed hydration dynamics in active sites and how they differ between wild-type (WT) and re-designed enzymes[10,11]. In proteins, the collective motions of water molecules could be rotationally restricted. However, water molecules still show libration as it is a dominant motion for breaking or making hydrogen bonds (H-bonds) in water molecule networks, with an average lifetime in liquid water of ≈1 ps[12]. The entropy of water molecules in a protein interior could be similar to that of bulk water[13], although the loss of translational and rotational freedom of water molecules in confinement (compared to bulk water environment) means that the loss of their configurational entropy occurs, due to the narrowing of H-bond energies. This entropy loss reflects an increase in molecular order, where an unfavourable decrease in water molecules' entropy echoes favourable enthalpic interactions[5].

In glycoside hydrolases, e.g. agarases[14], a family-1 β-glycosidase[15] and sialidases[16], the networks of H-bonded water molecules are connected to catalytic residues, which control hydrolysis versus transglycosylation (synthetic) balance. MD simulations revealed that transglycosylation reactions in an agarase correlated with higher active site water molecule dynamics,

[1]School of Agriculture, Food and Wine, and Waite Research Institute, Faculty of Sciences, Engineering and Technology, University of Adelaide, Adelaide, SA, Australia. [2]Department de Química, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. [3]The Florey Institute, University of Melbourne, Melbourne, VIC, Australia. [4]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. [5]These authors contributed equally: Sukanya Luang, Xavier Fernández-Luengo. ✉e-mail: laura.masgrau@uab.cat; maria.hrmova@adelaide.edu.au

water survival time reduction, and a decrease in water purge times[14]. In an α-amylase, collective variables (nucleophile and proton donor availability, H-bond and hydration environments, active site water molecule sites) controlled substrate reactivity[17].

During evolution, hydrolytic reactions are deployed to carry out chemical breakdowns of compounds owing to reactions with water molecules, which assist in breaking glycosidic bonds by providing a hydroxyl group for nucleophilic attack and accepting/donating a proton from the leaving group[18]. The term hydrolysis is broadly used for substitution, elimination, and solvation/desolvation reactions in which water is a nucleophile. In reactions catalysed by glycoside hydrolase (GH) enzymes, the water molecule breaks a chemical bond of substrates, where retaining GHs (contrary to inverting GHs) follow a double displacement reaction mechanism in two steps with the formation of a covalent glycosyl-enzyme intermediate[15,18–20].

A wide range of glycoside substrates is used by retaining, e.g. GH3 enzymes, which enable hydrolytic and transglycosylation reactions through the substrate-product-assisted processive catalytic mechanism[21,22]. Molecular mechanisms underlying processivity in enzymes remain underexplored, although recently, these concepts received heightened attention, e.g. investigating α-glucanotransferases[23] and cellulases[24]. Based on crystallographic and multi-scale MD modelling, we have described substrate-product-assisted processive catalysis in a plant exo-hydrolytic GH3 β-D-glucan glucohydrolase (HvExoI)[21,22].

The structure of HvExoI folds into two domains [(α/β)$_8$ barrel and (α/β)$_6$ sandwich] that at the interface accommodate an enclosed 13 Å-deep pocket-shaped active site. This pocket accommodates D285-nucleophile and E491-acid/base catalysts in the −1 subsite, which provide an anionic environment for the evolution of an oxocarbenium ion-transition state upon proton transfer. A vital conclusion was made that the glucose (Glc) product released from the β-D-glucoside substrates remains entrapped in the enzyme active site until an incoming substrate binds, presumably lowering the energy barrier to facilitate Glc displacement[21]. Further structural analyses revealed how each hydrolytic event is coordinated with the incoming substrate binding and hydrolysis to enable substrate-product-assisted processivity, which is dictated by the chemical nature of a substrate, and where multiple hydrolytic events occur without HvExoI de-contacting saccharide substrates[22].

In HvExoI, the W286-W434 clamp at the +1 subsite, playing a role in substrate binding, is situated next to an enclosed pocket-shaped active site accommodating nucleophile D285, acid/base E491 and non-nucleophilic E220 in an anionic −1 subsite environment. The catalysts are in near rigid states stabilised by a dense H-bond network to mediate hydrolysis, which

proceeds in two steps (Supplementary Fig. 1). Firstly, substrate glycosidic oxygen is protonated by E491, the glycosidic bond at the reducing end of a substrate is cleaved, with the Glc moiety adopting a flattened, positively charged oxocarbenium ion-like transition state, which reacts with D285 forming a glycosyl-enzyme covalent complex. In the next step, a water molecule with the assistance of E491 (that now acts as the catalytic base receiving a proton from the water molecule) restores the reducing end of the Glc moiety with retention of anomeric configuration, where Glc stays attached in the −1 subsite until a new substrate binds[19,20]. Here, a critical role of a nucleophilic water molecule is providing the hydroxyl group for nucleophilic attack and accepting a proton from the leaving group of the β-D-glucoside substrates[18,19].

In one of our studies with HvExoI, we observed that mechanism-based inhibitors, conduritol B epoxide, 2',4'-dinitrophenyl 2-deoxy-2-fluoro-α-D-glucoside (2F-DNPGlc), a non-hydrolysable S-glycoside analogue 4^I, 4^{III}, 4^V-S-trithiocellohexaose[19], and transition state gluco-phenylimidazole mimics[25] dislodged the Glc product from the −1 subsite. We concluded that because HvExoI hydrolysed various β-D-glucoside positional isomers, the non-reducing Glc moiety was locked stably in the −1 subsite position. Conversely, the Glc reducing-end moiety could adopt multiple orientations in the +1 subsite, which explains substrate poly-specificity[20].

In this work, we leveraged a multidisciplinary approach using kinetic and X-ray crystallography measurements, and classical molecular dynamics (cMD) simulations, using the unliganded and complexed WT and non-nucleophilic E220 mutant (to Ala) forms of HvExoI, to examine the organisation and dynamics of water molecule networks. We observed that acid/base E491 and nearby N219 and E220 water-coordinating residues interlinked water molecules at the van der Waals distances in active site pocket confinement and connected them with those of bulk water. In the E220A mutant structure, we noted a dramatic change in the organisation of a water molecule network and water flux movements through the enzyme active site, which correlated with profound modifications of catalytic properties. We concluded that designated residues participated in the formation of water molecule networks and served as principal operators to regulate water flux, which directly affected hydrolytic reactions.

The structural data were further rationalised through bioinformatics analyses of the GH3 family entries using ancestral sequence reconstruction (ASR)[26,27], which shed light on the evolutionary trajectories of GH3 entries and revealed water molecule network signatures. We detected that during evolution, the plant GH3 sub-clades retained unique residue signatures for water coordination in catalytic pockets. Our analyses, identifying the structural and evolutionary determinants of water molecule networks, could be beneficial for a rational design of interconnected biotechnologies and offer values beyond enzymology and biochemistry.

This information could be exploited in biotechnologies to develop more efficient enzyme forms. In this approach, the impact of catalytic water molecules on protein structure and dynamics could be modified to generate enzymes with modified catalytic rates or controlled preference for hydrolytic versus transglycosylation (synthetic) reactions.

Results

The exo-acting HvExoI enzyme adopts a retaining hydrolytic mechanism proceeding in glycosylation and de-glycosylation steps (Supplementary Fig. 1). To define the structural determinants of water molecule networks and catalytic pocket hydration sites during hydrolysis, we built on our previous findings with HvExoI[19] (Supplementary Fig. 1), created the E220A mutant (Supplementary Table 1), generated stability and kinetic data (Supplementary Fig. 2, Tables 1–3 and Supplementary Tables 2, 3), elucidated high-resolution X-ray structures in unliganded and complexed states (Fig. 1, Table 4 and Supplementary Table 4), and compared these data with those of WT. Crystallographic and kinetic data were complemented by cMD simulations of the WT and E220A

Table 1 | Relative rates of hydrolysis of poly- and oligosaccharides, and an aryl-glycoside of recombinant HvExoI wild-type (WT) and mutant E220A forms

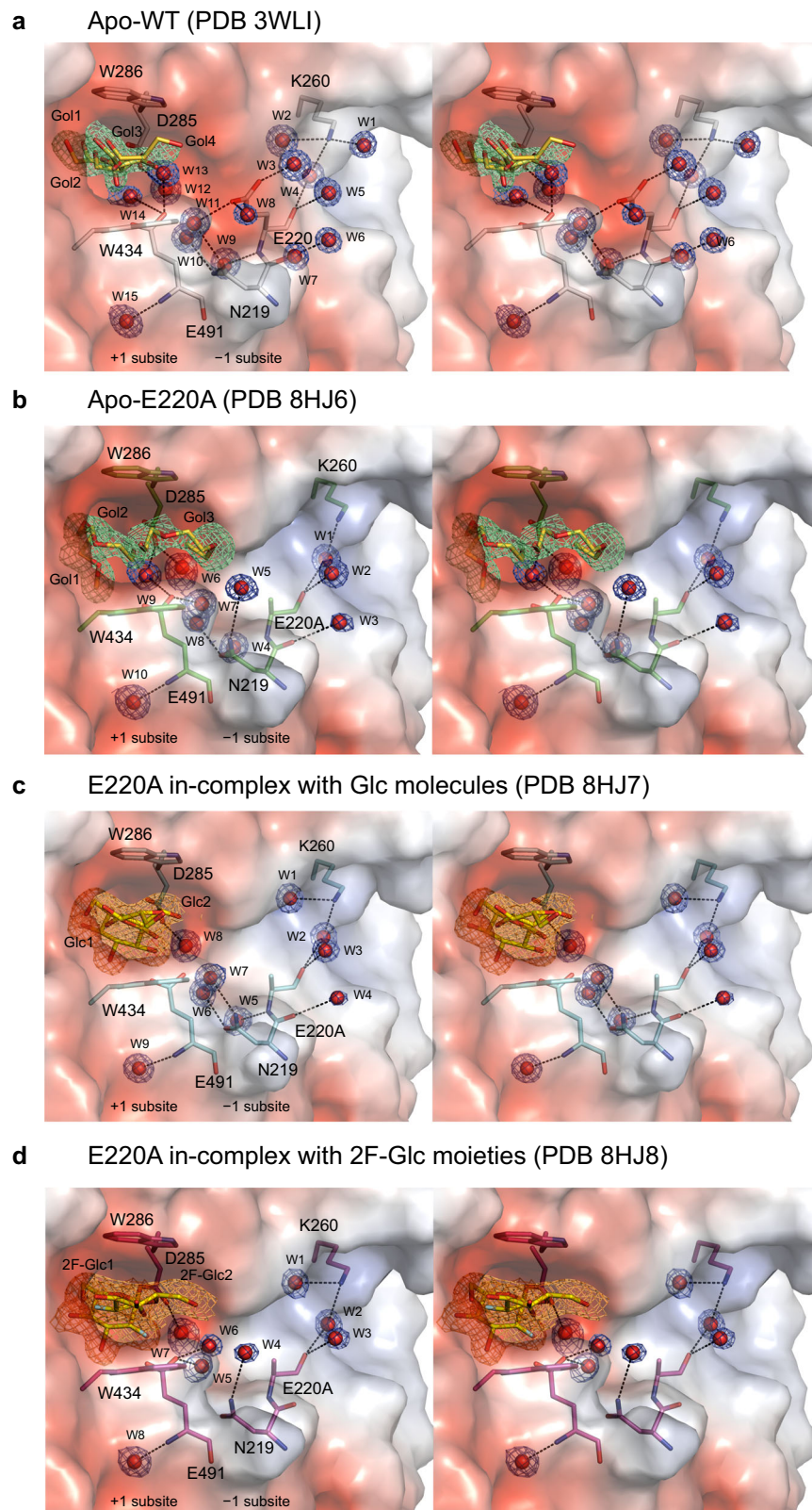| Substrate	Activity (µmole/min per mg protein)[a]	
	WT[b]	E220A
(1,3;1,6)-β-D-Glucan (laminarin; 3:2)[c] (1,4;1,3)-β-D-Glucan (barley glucan; 2.3-2.7:1)[c]	54.4 ± 0.5 5.4 ± 0.3	0.4 ± 0.01 nd[d]
Sophorose	6.0 ± 0.1	1.5 ± 0.10
Laminaribiose	1.3 ± 0.01	1.6 ± 0.06
Cellobiose	4.6 ± 0.2	0.3 ± 0.03
Gentiobiose	4.7 ± 0.4	0.6 ± 0.02
4NP-Glc	12.4 ± 0.8	1.6 ± 0.08

[a]Data, expressed in one or two significant digits, are given as mean values with standard error of measurements (n = 2) after ± signs.
[b]From ref. 22.
[c]Main types linkages and the ratio of linkage type[84].
[d]nd, (activity) not detected.

Fig. 1 | Cross-eyed stereo views of HvExoI wild-type (WT) and E220A in apo (unliganded) states, and in-complex with glucose (Glc) and 2,4-dinitrophenyl 2-deoxy-2-fluro-α-D-glucopyranoside (2F-DNPGlc). a Apo or unliganded-WT (PDB 3WLI; residues and Gol in cpk and cpk yellow sticks, and water molecules in red spheres, respectively); **b** apo-E220A (PDB 8HJ6; residues and Gol in cpk green and cpk yellow sticks, respectively); **c** E220A in-complex with Glc (PDB 8HJ7; residues and Glc in cpk cyan and cpk yellow sticks, respectively); and **d** E220A in-complex with two 2F-Glc moieties (PDB 8HJ8; residues and 2F-Glc in cpk magenta, and cpk yellow sticks, respectively), where 2F-Glc1 is covalently linked to D285. Structures shown on backgrounds of surface morphologies are coloured by electrostatic potentials (white, neutral; blue, +5 kT e^{-1}; red, −5 kT e^{-1}). **a–d** depict derived difference 2 m|F_o | -D | F_c| electron density maps for water molecules (blue mesh,) Gol (green mesh), Glc and 2F-Glc (orange mesh) contoured at 1.0σ levels, associated with N219, E220 or E220A, K260 and E491 in unliganded WT and E220A, Glc-bound E220A, and 2F-Glc-bound E220A structures, respectively. Identical rotation matrices were used in all images. Subsites −1/+1 and +2 are marked. Number of water molecules in structures varies between WT and E220A. Dashed lines indicate the separations at 2.6 to 3.5 Å between residues and water molecules. There are partial overlaps of water molecules in images.

a Apo-WT (PDB 3WLI)

b Apo-E220A (PDB 8HJ6)

c E220A in-complex with Glc molecules (PDB 8HJ7)

d E220A in-complex with 2F-Glc moieties (PDB 8HJ8)

covalent complexes using entropy water clustering[28], WaterKit[29] and DynaWatProt[30] (Figs. 2–5, Supplementary Figs. 3–7, Supplementary Tables 5–7, Supplementary Movies 1–2 and Supplementary Methods), which revealed geometrical parameters of selected residues and intricate organisation, motions and thermodynamics profiles of water molecule networks. Structural data were complemented by ASR analyses[26,27] (Fig. 6, Supplementary Figs. 8, 9 and Supplementary Data 1–3) to reconstruct evolutionary trajectories of GH3 enzymes and how structural signatures underlie the organisation of water molecule networks.
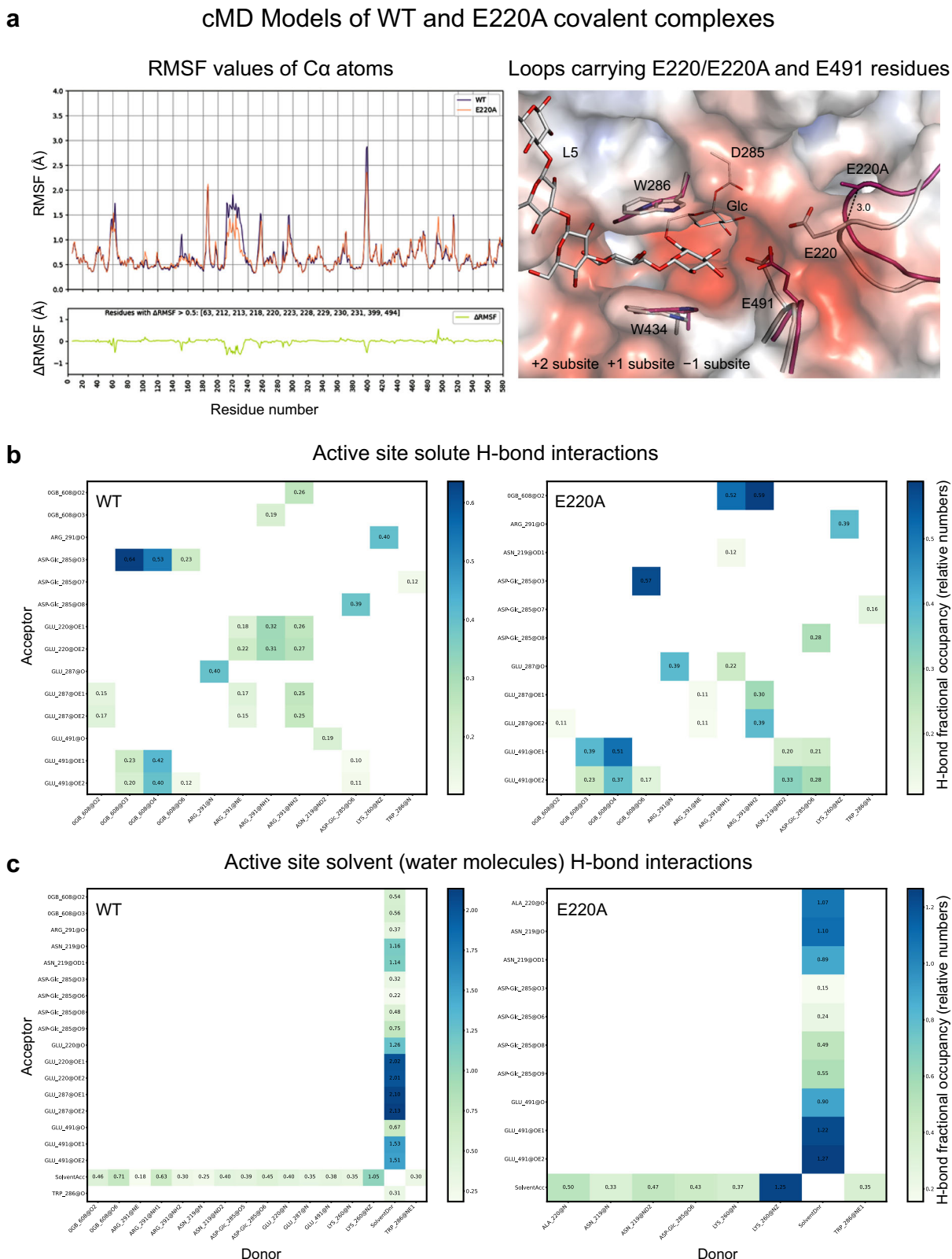
Fig. 2 | Root-mean-square fluctuation (RMSF) and H-bond interactions in computational models of HvExoI wild-type (WT) and E220A covalent complexes. a RMSF values (Å) are specified as the average of three replicates. Panels indicate RMSF values and their differences along protein chains, respectively. Loops carrying E220 (E220A) residues in 220–226 regions, and E491 residue in 494–497 regions of WT (cartoons, cpk sticks) and E220A (cartoons, magenta cpk sticks) covalent complexes. W286, W434, D285, L5 and D285-Glc adduct (cpk sticks) are shown for WT. Separations (Å) between Cα atoms of E220 (E220A) and E491 residues are indicated in dashed lines. Properties of surface morphologies coloured by electrostatic potentials are described in Fig. 1. Subsites −1, +1 and +2 are marked. Selected Glc moieties of L5 are displayed. **b** Active site residues-solute H-bond interactions. **c** Active site residues-solvent (water molecules) (SolventAcc) and donor residues-solvent (SolventDnr) H-bond interactions, calculated by cpptraj[80]. H-bond fractional occupancy heat term is given as the average of three replicates.

Physico-chemical properties, substrate specificity and kinetic constants of the HvExoI wild-type (WT) and mutant E220A forms signal differences in protein structures

Measurements of physico-chemical properties of the WT and E220A forms, such as hydrolytic activity dependences on pH (Supplementary Fig. 2a) and temperature (Supplementary Fig. 2b) indicated that the pH optimum of E220A was shifted toward an acidic region of 4.5, compared to that of 5.0 in WT. In contrast, the temperature sensitivity in both enzymes remained similar (Supplementary Fig. 2b).

To quantify the kinetic parameters of the WT and E220A with carbohydrate substrates, we measured relative rates of hydrolysis and calculated apparent first-order (K_M, k_{cat}) constants, and second-order (k_{cat}/K_M) reaction rate constants (catalytic efficiency), and first-order (K_i) constants of enzyme-inhibitor complexes (Tables 1–3 and Supplementary Tables 2, 3). We observed that the E220A mutation predominantly affected the k_{cat} parameters measured with laminaribiose, cellobiose and 4-nitrophenyl β-D-glucoside (4NP-Glc) (Supplementary Table 3). Consistent with our predictions, kinetic data revealed that the E220A mutant retained substrate poly-specificity (Table 1 and Supplementary Table 2) but was 10 to 30-fold less efficient with (1,3)-, and (1,4)-linked gluco-oligosaccharides and 4NP-Glc, and lost activity with (1,3;1,6)- and (1,3;1,4)-linked polysaccharides, compared to WT (Table 2 and Supplementary Tables 2, 3). To define the strength of binding with methyl 6-thio-β-gentiobioside (G6SG-OMe), and substrate analogue and mechanism-based inhibitors glucono-δ-lactone and 2F-DNPGlc, we determined K_i constants, using a competitive inhibition model (Table 3). These examinations revealed around 13-fold weaker binding of glucono-δ-lactone in E220A ($K_i = 103 \times 10^{-6}$ M) compared to WT ($K_i = 10 \times 10^{-6}$ M), while 2F-DNPGlc bound in WT and E220A with the same strength ($K_i = 0.5 \times 10^{-6}$ M). However, the differences in K_i values with G6SG-OMe between WT and E220A were less pronounced with around 3-fold tighter binding in WT (Table 3).

Substrate specificity and kinetic data signalled the decreased hydrolytic efficiency with β-D-glucosides in the E220A mutant compared to WT, and correlated with the weaker binding of G6SG-OMe and glucono-δ-lactone, while 2F-DNPGlc bound with the same strength.

Crystal structures of the HvExoI wild-type (WT) and mutant E220A forms describe water molecule networks in active sites

We elucidated three crystal structures of the E220A mutant [Protein Data Bank (PDB) 8HJ6, 8HJ7, 8HJ8] with a resolution of better than 1.95 Å at atomic levels when atoms of C–C bonds are resolved. We compared these structures with the solved WT enzyme forms (PDB 3WLI and 1IEW)[19,21,22] to establish binding modes of substrates and inhibitors and define structural differences (Fig. 1, Table 4 and Supplementary Table 4).

Apo-form of E220A and comparison with WT. Crystal structures of E220A (PDB 8HJ6; 1.78 Å resolution) and WT (PDB 3WLI; 1.45 Å resolution) indicated that both structures had similar backbone dispositions (the root-mean-square deviation—RMSD value 0.14 Å). However, the E220 side-chain replacement for alanine instigated a modified pattern of water molecule coordination by E491, N219, E220 (E220A), D285 and K260. The most prominent difference between the two structures was the number of water molecules associated with catalytic acid/base E491, and the N219, E220 (or E220A) and K260 water-coordinating residues creating characteristic networks, and where K260 connected these networks to bulk water milieu (Fig. 1a, b). We observed 10 and 15 respective water molecules in the E220A and WT structures at the separations between 2.5 to 3.5 Å (Fig. 1a, b). Explicitly, in WT and E220A, N219 and E491 coordinated four and three water molecules, E220 or E220A channelled four and one water molecules, while K260 coordinated three and one water molecules, respectively (Fig. 1a, b).

E220A in-complex with the Glc product and comparison with WT. Hydrolytic product Glc is a relatively weak binder in HvExoI with a K_i value of 0.16×10^{-3} M[22]. In HvExoI (PDB 3WLO; 1.55 Å resolution), two Glc molecules, one of each in the −1 and +1 subsites, were observed in $^4C_1$ chair and $^1S_3$ skew boat conformations with classic Cremer–Pople ring-puckering parameters[31], and forming a network of 12–13 mono- and cooperative bi- and tri-dentate H-bonds[22]. Correspondingly to the WT structure, the analysis of the E220A mutant form in-complex with Glc (PDB 8HJ7; 1.85 Å resolution) revealed the presence of two Glc molecules in $^4C_1$ chair conformation, each in the −1 and +1 subsites, also consistent with classical Cremer–Pople ring-puckering parameters, which were fitted to the calculated difference electron density map, observed in these subsites (Supplementary Fig. 1c, d). The most prominent interactions of the −1 subsite-bound Glc molecule were short H-bonds of 2.5 to 2.7 Å between Oδ2 atoms of D95 and C4-OH and C6-OH groups of Glc. The Glc molecule in the +1 subsite was constrained through stacking interactions at the 4.2 Å and 4.4 Å separations from the centres of phenyl ring moieties of W286 and W434, respectively (Fig. 1c). In the E220A structure, the most noticeable was the presence of nine water molecules coordinated by N219, E220A, D285 and K260, and where E491 channelled one water molecule, compared to 15 water molecules in WT with four channelled through E491[21] (Fig. 1c).

Table 2 | Apparent second-order reaction rate kinetic constants of recombinant HvExoI wild-type (WT) and mutant E220A forms

Substrate	k_{cat}/K_M (mM^{-1} s^{-1})[a]	
	WT[b]	E220A
(1,3;1,6)-β-D-Glucan (laminarin)[c]	234.0 ± 41.0	nd[d]
(1,4;1,3)-β-D-Glucan (barley glucan)[c]	365.5 ± 58.0	nd
Laminaribiose	41.2 ± 5.6	2.9 ± 0.1
Cellobiose	3.7 ± 0.2	0.2 ± 0.03
4NP-Glc	13.9 ± 0.6	1.3 ± 0.2

[a]Data, expressed in one or two significant digits, are given as mean values with standard error of measurements (n = 4–6) after ± signs.
[b]From ref. 22.
[c]Main types linkages and the ratio of linkage type[84].
[d]nd, (activity) not detected.

Table 3 | Apparent inhibition constants of recombinant HvExoI wild-type (WT) and mutant E220A forms

Inhibitor	K_i[a]		
	WT glycosylated	WT De-glycosylated	E220A De-glycosylated
G6SG-OMe K_i (M) ΔG (kJ mol^{-1})[b]	109×10^{-6} −23.0	71×10^{-6} −24.1	340×10^{-6} −20.1
Glucono-δ-lactone K_i (M) ΔG (kJ mol^{-1})	10×10^{-6} −29.0	8×10^{-6} −29.6	103×10^{-6} −23.1
2F-DNPGlc K_i (M) ΔG (kJ mol^{-1})	0.5×10^{-6} −36.6	0.5×10^{-6} −36.6	0.5×10^{-6} −36.6

[a]Data expressed in one significant digit.
[b]Calculated according to $\Delta G = -RT \ln [1/K_i]$[85].

Table 4 | Data collection and refinement statistics of HvExoI E220A structures in apo (unliganded)-form, and in-complex with glucose (Glc) and the mechanism-based inhibitor 2F-DNPGlc

	E220A Apo	E220A Glc bound	E220A 2F-DNPGlc bound
PDB accession	8HJ6	8HJ7	8HJ8
Data collection			
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$
Number of molecules/ASU	1	1	1
Cell dimensions			
a, b, c (Å)	100.23, 100.23, 183.17	100.63, 100.63, 182.32	100.50, 100.50, 182.69
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	87.93–1.79 (1.84–1.79)[a]	88.10–1.85 (1.90–1.85)	88.06–1.95 (2.00–1.95)
R_{merge} (%)[b]	9.6 (83.5)	9.4 (82.3)	9.2 (80.0)
$I/\sigma I$	48.6 (2.3)	50.8 (2.3)	38.7 (3.0)
Completeness (%)	95.3 (68.3)	98.0 (80.0)	100 (100)
Redundancy	24.8	26.0	28.2
CC1/2 (%)	99.8	99.8	99.7
Refinement			
Resolution (Å)	87.93–1.79	88.10–1.85	88.06–1.95
No. reflections	79,237	74,941	64,685
R_{work}/R_{free}	0.155/0.187	0.175/0.205	0.154/0.184
No. of atoms	5313	5300	5270
Protein	4618	4577	4618
Ligand/ion	0	24	23
Water	695	699	629
B-factors	27.86	38.03	32.58
Protein	70.03	35.48	30.19
Ligand	–	44.84	40.50
Water	37.94	49.79	41.72
RMSD values			
Bond lengths (Å)	0.021	0.009	0.024
Bond angles (°)	1.661	1.159	1.818

[a]Values in parentheses are for the highest resolution shell.
[b]$R_{merge} = 100 [\sum(I_i - <I>)^2/\sum I_i^2]$, summed over all (highest shell) independent reflections.

E220A in-complex with the 2-deoxy-2-fluoro-α-D-glucopyranosyl (2F-Glc) moiety covalently linked to D285 and comparison with WT. 2F-DNPGlc bound with a similar strength in WT and E220A (Table 3). The WT structure containing the 2F-Glc moiety linked to D285 (PDB 1IEW; 2.55 Å resolution; cf. Fig. 5c in ref. 19) was collected using an in-house X-ray source and solved with an electron density in the −1 subsite corresponding to the 2F-Glc-D285 adduct (Supplementary Fig. 1b). This electron density map contoured at 4.2σ was interpreted containing the covalent glycosyl-enzyme intermediate of a catalytically competent enzyme. The C1 separation between 2F-Glc and Oδ1 of D285 was 1.47 Å, indicating a covalent linkage formed through α-anomeric configuration. A similar adduct formed in the E220A structure (PDB 8HJ8; 1.95 Å resolution) revealed the presence of two 2F-Glc molecules in $^4C_1$ chair conformation in the −1 and +1 subsites, consistent with classical Cremer–Pople ring-puckering parameters (Supplementary Fig. 1d). Here 2F-Glc in the −1 subsite slightly tilted towards D285, compared to the Glc molecule in E220A (Fig. 1c) or the 2F-Glc moiety in WT (cf. Fig. 5c in ref. 19). The Oδ2 of D285 H-bond with the F-atom in the E220A/2F-Glc

complex (Fig. 1d) was within an overall coordinate error limit of H-bonds between Oδ1 of D285 and the C2-OH group of Glc in the E220A/Glc complex (Fig. 1c, d and Table 4). In E220A, the 2F-Glc moiety in the −1 subsite formed a covalent linkage at the separation of 1.39 Å between C1 of 2F-Glc and Oδ1 of D285 (Fig. 1d). Here, we also observed that E491 and G57 formed short H-bonds with the C1-OH and C3-OH groups of 2F-Glc, and D95, R158, K206, H207 and R291 established H-bond at 2.5 to 3.0 Å to 2F-Glc in the −1 subsite, alike the WT structure with 2F-Glc[19]. In the E220A/2F-Glc complex, we identified eight water molecules coordinated via H-bonds with N219, E220A, K260, D285 and E491 (Fig. 1d).

Crystal structures of an unliganded E220A mutant form and in-complex with Glc and a mechanism-based inhibitor 2F-DNPGlc revealed significant differences in distributions of water molecule networks in active sites compared to those in WT.

Computational models of the HvExoI wild-type (WT) and E220A covalent complexes through cMD simulations establish differences in water molecule networks

We built computational models of WT and E220A covalent complexes (Supplementary Fig. 3 and Supplementary Table 5) to investigate structural determinants of water molecule networks and how they are coordinated, through cMD in three replicates, each in one µs duration. These models contained the hydrolytic product (1,3)-linked β-D-glucoside laminaripentaose (L5) in the +1 and +2 subsites (and facing bulk solvent) originating from laminarihexaose (L6), and the 2F-Glc moiety linked covalently to catalytic nucleophile D285 in the −1 subsite. This configuration reflects processive catalysis in HvExoI, where the oligosaccharide product (which becomes a new substrate in the next hydrolysis round) is retained in the +1 and +2 subsites. Here, we kept the full L5 substrate chain, as the dynamics and interactions of the oligosaccharide chain are likely to affect its positioning and the influx of water molecules into the active site[21,22].

Three trajectory replicates revealed that the root-mean-square fluctuation (RMSF) values of Cα atoms were higher in the 220–230 residue region in WT, whereas the loop carrying these residues had surprisingly higher flexibility compared to E220A (Fig. 2a). In the E220A complex, we observed that the loop in the 224–227 residue region had re-positioned in the active site cavity compared to WT. This re-positioning caused the mutated E220A residue to adopt an altered orientation (Fig. 2a). Analyses of protein-solute H-bond interactions indicated that the E220A mutation was accompanied by differences in N219-E491 H-bonds frequencies, where the N219-E491 interaction values in WT and E220A were 0.19 and 0.53, respectively (Fig. 2b; calculations described in "Methods"). Notably, in WT this H-bond was formed between Nδ2 and the E491 backbone carbonyl atom, whereas in E220A, it switched towards the E491 carboxylate. This change resulted in a lower frequency of the E491-water H-bonds in E220A (2.44) than in WT (3.04) (Fig. 2c). Other residue interactions, e.g. between E220 and R291 were seen in WT only but were absent in the E220A complex. We also observed new H-bonds formed between R291-K260 and R291-E287 in the latter complex. These changes were mirrored by the adjustments of H-bond frequencies between residues and water molecules in both complexes. In WT, a total of E220-water molecules H-bonds reached the value of 3.92, while those of the N219-water molecules H-bonds were higher in WT (1.57), compared to E220A (1.36). These evaluations indicated that some prevalent and strong protein-water molecule H-bonds appearing in WT were lost upon mutation (Fig. 2c).

In summary, while in WT the loop carrying the 220–230 residues showed unexpectedly increased flexibility in the E220A complex, we observed the altered dispositions and H-bond interactions of residues that neighbour E220A, which resulted in modified water accessibility—this cooperatively led to losses of hydrophilic interactions of the loop carrying the E220A residue.

Distributions of side-chain dihedral angles and frequencies of active site residues in the computational models of wild-type (WT) and E220A covalent complexes

To investigate how the E220A mutation propagates into spatial dispositions of active site residues involved in water molecule network formations, potentially affecting residue conformations during catalysis, we analysed geometrical parameters such as side-chain dihedral angle distributions and frequencies of the E491 and N291 residues. These analyses were based on cMD frames with active site water molecules included (Fig. 3 and Supplementary Figs. 4, 5).

Analyses of distributions and frequencies of the side-chain dihedral χ1, χ2 and χ3 angles in E491 of the WT covalent complex revealed that the distribution of the χ1 angle along C-Cα-Cβ-Cγ atoms exhibited a narrow distribution of around 60°. Conversely, in the E220A complex, a multimodal and wider distribution was observed (within angular angles of 0 to 200° and 0 to −200°) with the maximum population shifted to 170°. Here, we detected a significantly lower occurrence of the dominant χ2 angle configuration observed in WT between 40° and 80° that was compensated by the increased populations at around −100° and 180° in the E220A mutant. In contrast, distributions of the χ3 dihedral angles were similar in both covalent complexes (Fig. 3a). These quantitative data reflected on overall geometries of active sites and divergent conformations adopted by the E491 side-chains (Fig. 3b).

Similar analyses of side-chain dihedral angle distributions and frequencies were conducted for the N219 residue of the WT and E220A covalent complexes (Supplementary Fig. 4a and Supplementary Table 6). We observed differences between the χ1 and χ2 angle conformations in both complexes, with preferential values of around 170° and 100°, respectively, for WT and 60°/−60° and −100° for E220A, and the χ1 angle of E220A featuring a more noticeable bimodal distribution due to the defined parameters of the N219 dihedral angles. These alterations may be related to the H-bond formed at the separation of 3.9 Å between Nδ2 of N219 and Oε1 or Oε2 of E491 in the E220A complex. This H-bond was absent in WT, where a much looser interaction was observed with the E491 backbone (Supplementary Fig. 4b). Although the presence of these H-bonds was not limited to the unique N219 and E491 dihedral values (Supplementary Fig. 5), the E220A mutation induced their differential orientations and differences in H-bonds with catalytic acid/base E491, and surrounding residues and water molecules. This observation suggested that water molecules near E491 could be misallocated in the E220A complex and would not exhibit ideal dispositions for the nucleophilic attack at the anomeric carbon compared to WT. Moreover, this new N219-E491 H-bond could affect the pK_a value of E491 and reduce its proton-binding affinity for catalytic water activation, thus reducing the hydrolytic activity of the E220A mutant[32].

Frequency profiles of water molecules in active sites of computational models of the wild-type (WT) and E220A covalent complexes

We examined the radial pair distribution function (RDF) parameters (or how density varies as a function of distance from a reference particle) of water molecules centred on Cα atoms of D285 as indicators of characteristics of water molecule networks in the computational models of the WT and E220A covalent complexes. These analyses indicated that the RDF profiles exhibited similar trends in both complexes, although the frequency profile of water molecules around the 6.5 Å radius was higher in the E220A complex (Fig. 4a).

The frequency histograms of the total numbers of water molecules located within the 6.5 and 8 Å spheres of Cα atoms of E220/E220A, E491, N219 and D285 of the WT and E220A complexes showed mixed characteristics (Fig. 4b and Supplementary Fig. 6). While the histogram of water molecules around E491 in WT exhibited higher numbers and a marginally higher frequency (attributable to the function of E491 as a catalytic acid/ base), in the E220A mutant, we observed higher frequencies of water molecule distributions within the same spheres of the E220A residue, as presumably water molecules fill-in the region vacated by a side-chain

substitution upon mutation (Fig. 4b). Water molecules linked to E220 and E491 in the WT complex formed an internal and coordinated network of H-bonds, while in E220A, this pattern of coordination was absent (Fig. 4c). Here, the role of the E491 residue in catalysis, as a catalytic base, is in the activation of water molecules for nucleophilic attack in both complexes. Consequently, in the E220A mutant, the E220A replacement would affect the environment and the function of E491 and influence hydrolytic rates and efficiency. In the frequency histograms of water molecules centred at Cα atoms of N219, we observed higher and lower water molecule numbers located within 6.5 and 8 Å spheres, respectively, however, with similar frequencies (Supplementary Fig. 6a). Conversely, higher numbers of water molecules were identified within 6.5 and 8 Å spheres of Cα atoms of D285 (Supplementary Fig. 6a), with short contacts at the separations of ≤3.5 Å between water molecules and the Oδ2 atoms of D285 in both enzymes (Supplementary Fig. 6b).

Distributions of catalytic water molecules in computational models of the HvExoI wild-type (WT) and E220A covalent complexes

In our analyses, we define catalytic water at a separation of ≤3.5 Å from the Oε1 or Oε2 atoms of E491 and at ≤4.5 Å from the C1 atom of the D285-Glc adduct, which corresponds to the centre of a glycosidic bond cleavage. Distributions of catalytic water molecules as a mean of three replicates, expressed in total numbers and percentages of all water molecules identified in three replicates of both proteins, indicated that the higher average occurrences of catalytic water molecules, 68% versus 57%, were identified in the WT complex compared to mutant E220A (Fig. 5a). Dispositions of selected catalytic water molecules in active sites (at the separations ≤2.9 Å from E491), based on evaluations of three computational frames of HvExoI wild-type (WT) and E220A along cMD simulations, indicated that four water molecules associated through H-bonds with E491 in both cases (Fig. 5a). These water molecules were placed discreetly in regular patterns in WT between E491 and E220 residues. Conversely, in the E220A complex, water molecules were clumped together and failed to exhibit even dispositions within the space of a catalytic pocket.

WaterKit describes catalytic pocket water hydration sites in the HvExoI wild-type (WT) and E220A crystal structures in-complex with the Glc product

The WaterKit simulation protocol[29], which performs the thermodynamic profiling of protein hydration sites, was used to predict dispositions of individual water molecules and their densities in active sites of the crystal structures of WT (PDB 3WLH) and E220A (PDB 8HJ7) in-complex with a Glc product. Here, we detected a differently organised water density network in the E220A catalytic site compared to the WT structure. Specifically, in the E220A complex we identified a deeper internal water density and a disrupted water network around the E220A residue and its surroundings (Fig. 5b). This altered water density was predominantly seen in the region that would otherwise be occupied by the E220 side-chain (Fig. 5b, c, Supplementary Fig. 7 and Supplementary Table 7).

DynaWatProt predicts global dynamics and density of water molecule networks in the computational models of HvExoI wild-type (WT) and E220A covalent complexes

Analyses of water movement predictions using the Python package DynaWatProt[30], based on cMD simulations, revealed dramatic and profound contrasts in global dynamics and density of water molecule networks in WT and E220A covalent complexes (Supplementary Movies 1, 2). Here, the extensive water molecule networks were formed in the vicinity of entries into the active sites of both complexes (Fig. 5d).

The evaluations of the entropy of these water clusters revealed a higher entropy in the mutant E220A (Supplementary Table 5). However, the disposition, coordination and penetration of water molecules near a protein surface were disrupted in the E220A complex, compared to those processes in WT, due to the mutated E220 (Fig. 5). As expected, this mutation led to
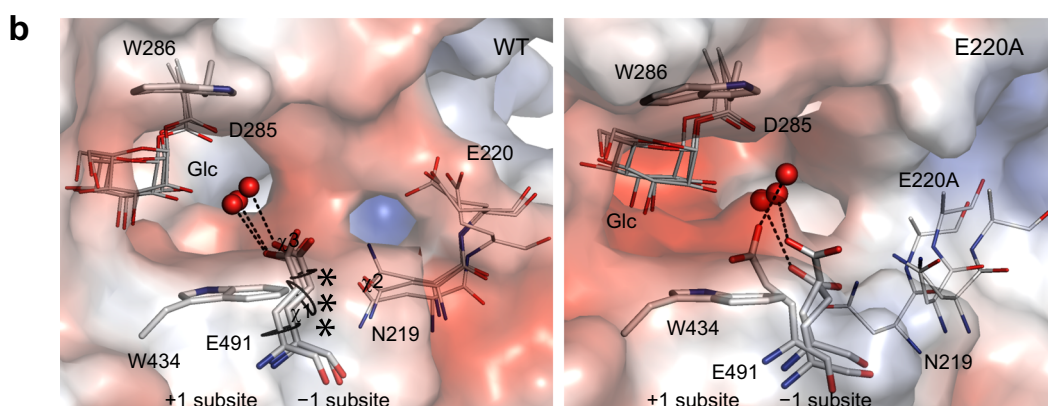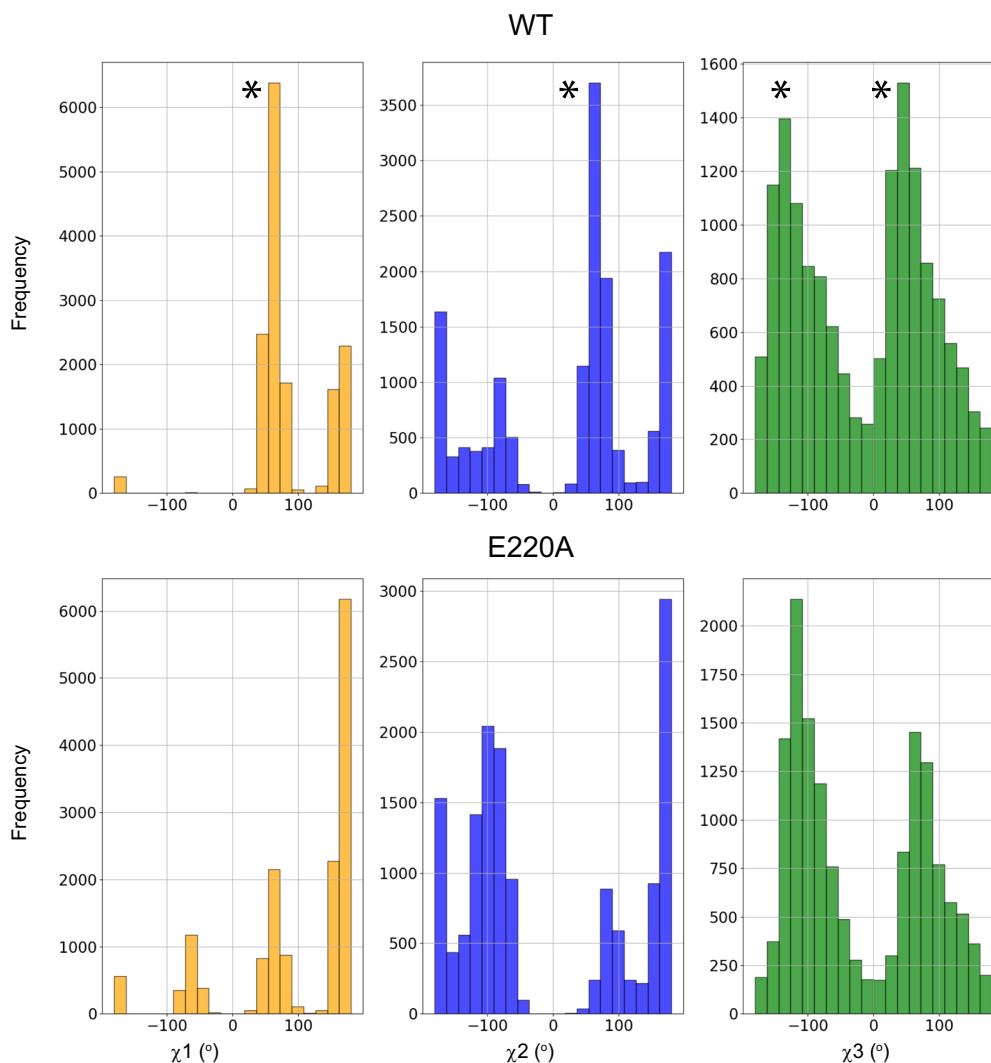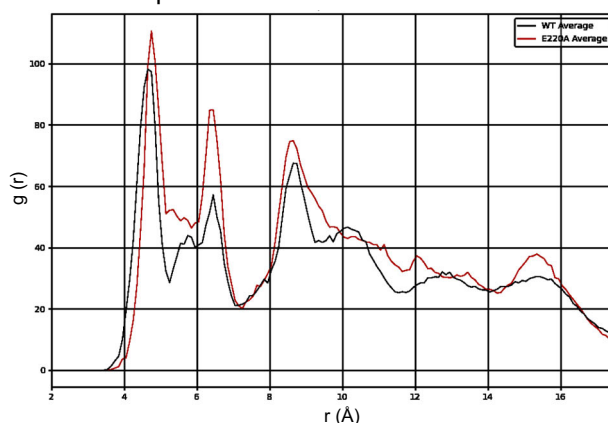
Fig. 3 | Frequencies of side-chain χ1, χ2 and χ3 dihedral angle distributions of E491 in HvExoI computational models of wild-type (WT) and E220A covalent complexes. a Frequencies of χ1, χ2 and χ3 dihedral angle (°) distributions of E491. **b** Dispositions of active site residues and D285-Glc adduct (cpk sticks and lines) in WT and E220A are shown for three selected frames (represented by D285-Glc adducts) along cMD simulations. Water molecules (red spheres) at separations of ≤3.0 Å between Oε1 or Oε2 atoms of E491 are indicated in dashed lines. Dihedral angles χ1, χ2 and χ3 of E491 in WT are marked by lassos in (**b**), and peaks of frequencies in asterisks in (**a**, **b**). Subsites −1 and +1 are marked. Properties of surface morphologies coloured by electrostatic potentials are described in Fig. 1.
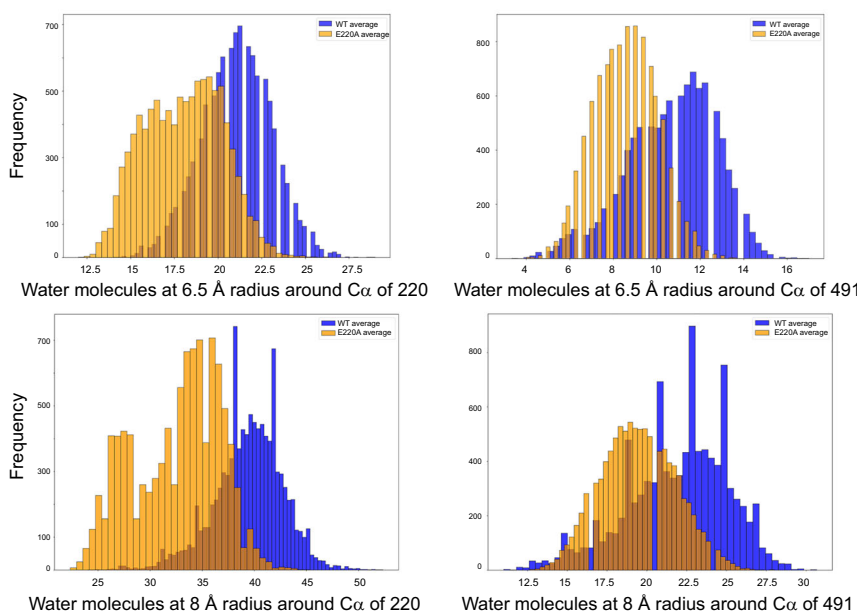
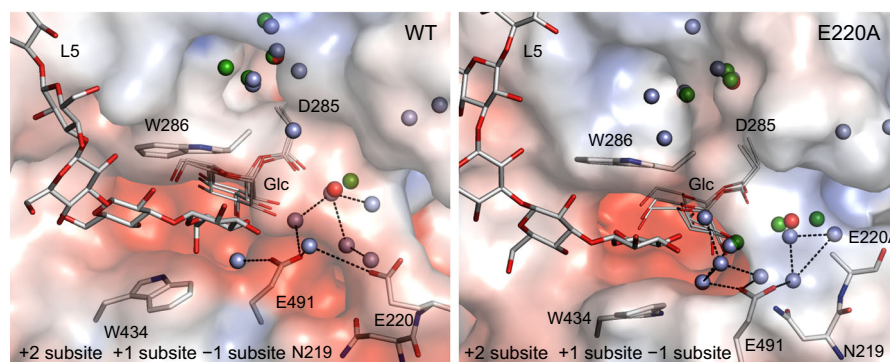## Water molecule networks in wild-type (WT) and E220A covalent complexes

a   Radial pair distribution function profiles of water molecules centered on Cα atoms of D285

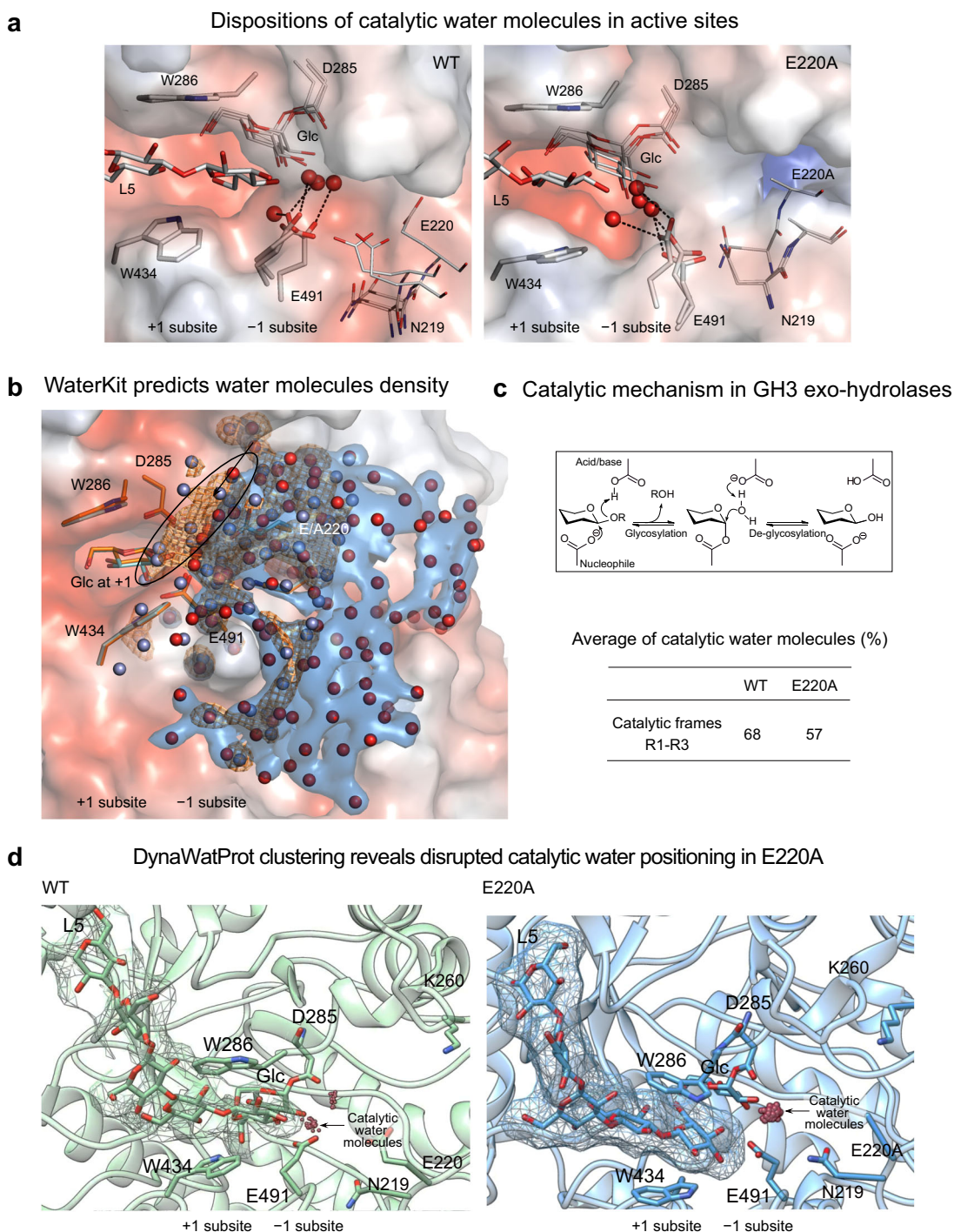**b**   Water molecules located within 6.5 Å and 8 Å radii of E220/E220A and E491

c   Separations of water molecules around E220/E220A and E491

Fig. 4 | Water molecule networks of active site pockets in HvExoI computational models of wild-type (WT) and E220A covalent complexes. a Radial pair distribution function profiles (quantified by density variation as a function of distance from a reference particle) of water molecules centred on Cα atoms of D285 (average of three replicates). **b** Frequencies of water molecules located within 6.5 Å and 8 Å spheres of E220/E220A and E491. **c** Dispositions of active site residues and D285-Glc adduct (cpk sticks and lines) in WT and E220A covalent complexes are shown for three selected frames (represented by D285-Glc adduct) along cMD simulations. Water molecules at separations of ≤4.7, 6.5 and 9.5 Å from E220/E220A, E491, N219 and D285-Glc adduct are specified in red, green and blue spheres (displayed in-depth cue mode), respectively. Subsites −1, +1 and +2 are marked. Selected Glc moieties of L5 are displayed. Separation between E220 and E491, and water molecules of ≤3.6 Å are shown in dashed lines. Properties of surface morphologies coloured by electrostatic potentials are described in Fig. 1.

Catalytic water molecules and water network densities in wild-type (WT) and E220A covalent complexes

a Dispositions of catalytic water molecules in active sites

b WaterKit predicts water molecules density

c Catalytic mechanism in GH3 exo-hydrolases

Average of catalytic water molecules (%)

	WT	E220A
Catalytic frames R1-R3	68	57

d DynaWatProt clustering reveals disrupted catalytic water positioning in E220A

Fig. 5 | Predictions of catalytic water molecules and water molecule network densities in HvExoI computational models of wild-type (WT) and E220A covalent complexes. a Prevalence of catalytic water molecules (cpk spheres) in computational models of WT and E220A are shown for three selected frames (represented by N219, E220/E220A, D285-Glc adduct) along cMD simulations. Separations of water molecules (cpk spheres) from Oε1 and Oε2 atoms of E491 at ≤2.9 Å are indicated in dashed lines. Properties of surface morphologies coloured by electrostatic potentials are described in Fig. 1. Selected Glc moieties of L5 are displayed. **b** Predicted water molecule network densities in superposed (RMSD value 0.13 Å) WT (PDB 3WLH) and E220A (PDB 8HJ7) structures in-complex with Glc using WaterKit[29]. Images of WT (blue surface contoured at −1 kT e^{-1} level) and E220A (orange mesh contoured at −1 kT e^{-1} level) depict clusters of water

molecules (red and blue spheres for WT and E220A, respectively). Alternate water molecule dispositions in E220A are marked in an ellipsoid. Key active site residues and Glc molecules (cpk sticks) are indicated for WT (cyan cpk sticks) and E220A (orange cpk sticks). Properties of a surface morphology of WT coloured by an electrostatic potential are described in Fig. 1. **c** Retaining catalytic mechanism of HvExoI and averages of catalytic water molecules (%) of all water molecules identified in replicates 1–3 in WT and E220A covalent complexes. **d** Snapshots from molecular animations illustrating the participation of catalytic water molecules (red spheres) in active site residues of WT and E220A covalent complexes, predicted by DynaWatProt[30]. Protein folds are shown in lime green (WT) and light blue (E220A) and correspondingly coloured residues in cpk sticks, with L5 substrate surfaces indicated in a mesh. Subsites −1 and +1 are marked in all images.

modifications of the overall intake of water molecules. Specifically, in the mutant E220A complex (Supplementary Movie 2), we observed a deeper penetration of water molecules into the active site pocket in smaller numbers, compared to those in WT (Supplementary Movie 1). Indeed, the visualisation of water clusters in both complexes illustrated the dispositions of putative catalytic water molecules (Fig. 5c, d). These water clusters shared close proximity to anomeric carbons, E491 and D285 side-chains in both complexes. However, only in WT, a potential catalytic water was observed within an optimal distance and orientation to the anomeric carbon at the β-face of the Glc molecule for a nucleophilic attack, as described above (Supplementary Movies 1, 2).

Ancestral sequence reconstruction (ASR) of the glycoside hydrolase family-3 (GH3) estimates the evolution of water molecule network-associated signatures

We employed FireProt[ASR26,27] to reconstruct ancestral (or progenitor) sequences and infer evolutionarily conserved motifs linked to water molecule networks in the GH3 family. Two analyses were undertaken, where we first used the HvExoI sequence and specified the essential D285 and E491 catalysts while searching databases at 20–90% sequence identity to the HvExoI sequence. Here, FireProt[ASR] generated a homologous group of bacterial, metagenome and plant sequences, where monocot and dicot entries formed a unique plant clade containing HvExoI (Fig. 6 and Supplementary Data 1–3). This GH3 clade with monocot and dicot entries and involving the N219, E220 and K260 residues (HvExoI numbering), engaged in water molecule networks, was disconnected from bacterial clades. The ancestral node 196 represented the most plausible node for the evolution of a plant clade. Conversely, the N219-E220 residues (or NE signature) were substituted in nodes 179, 180 and 195 by the DE or DQ signatures, whereas D285 and E491 were conserved in all entries (Supplementary Fig. 8). Overall, 39–96% sequence identity was observed in the plant group, with ancestral sequences 179 and 180 being the most identical at 96%, meaning that these nodes may have originated concurrently during evolution (Supplementary Fig. 8).

When we used the HvExoI sequence and as the essential residues D285, E491, N219 and E220A, and searched databases at 30–90% sequence identity, monocot and dicot entries allocated in separate clades (Supplementary Fig. 9a). For monocots, ancestral nodes 131, 132, 133, 136 and 164 appeared to be the most related containing conserved catalytic signatures participating in water molecule networks (Supplementary Fig. 9b). We observed that ancestral sequences 132 and 163 were identical at 75% to HvExoI and that in plant entries an overall 73–99% sequence identity was seen, with ancestral sequences 132 and 163 being the most identical at 99%, implying these ancestral nodes could share an evolutionary history.

In summary, ASR analyses indicated that the residues equivalent to D285, E491, N219, E220 and K260 in HvExoI were invariant in the examined plant GH3 and their ancestral nodes. This conservation implies an ancient evolutionary history for active site residues.

Discussion

An exohydrolytic β-D-glucan glycohydrolase HvExoI, classified in the GH3 family[33] exemplifies a class of enzymes with a pocket-shaped catalytic site, which employs substrate-product assisted processive catalysis[21,22,34]. This mechanism operates in GH3 hydrolases recognising oligo- and polymeric carbohydrate substrates[21,22,35], which reflects the architecture of a catalytic site, where structural determinants govern substrate binding, including directionality and specificity, to ensure optimal catalytic efficiency. The essence of substrate-product-assisted processivity resides in the role of a hydrolytic Glc product, which induces the formation of a lateral cavity to allow Glc egress and the next hydrolytic event to proceed using a retaining catalytic mechanism[18,19,36,37]. Processivity thus enables multiple catalytic cycles to proceed without an enzyme losing contact with oligo- or polysaccharides. This mechanism contrasts with distributive catalysis, where catalysts follow a conventional enzyme-reactant-association-transformation-dissociation pattern[34].

In HvExoI, we defined substrate-product-assisted processivity, reactant trajectories and mechanisms of breakdown of oligo- and polysaccharides based on nearly 40 crystal structures of WT and mutant ligand-free forms and in-complex with a product, substrate analogues, and positional isomers of O- and S-linked substrates[19-22,25,38]. One of the residues, which cooperatively coordinated water molecules with acid/base E491 was non-nucleophilic E220[19,21,22]. To precisely define the function of this residue, here, we resolved three high-resolution X-ray crystal structures of the E220A HvExoI mutant in an unliganded form and in-complex with the mechanism-based 2F-DNPGlc inhibitor and a Glc product. Based on these structures, we constructed computational models of covalent complexes of the WT and E220A forms to define dispositions of water molecule networks and the organisation of catalytic pocket hydration sites. The precise definition of placements of water molecules during catalysis is vital, as they are critical for ligand binding and enthalpic and entropic water molecule contributions to catalysis[29,39-41]. These firmly bound water molecules with slow dynamics are often considered an extension of a protein or its binding site[5].
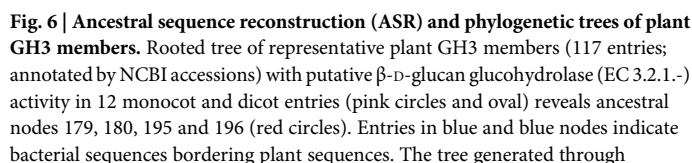
In the WT structure of HvExoI in-complex with 2F-DNPGlc (Supplementary Fig. 1b), we observed an interconnected network of water molecules near acid/base E491 and neighbouring non-nucleophilic E220[19]. We hypothesised that these water molecules are directed precisely to their target locations (rather than through arbitrary movements) and that the nearest water molecule to E491 could fulfil the role of a candidate nucleophile during hydrolysis. In our earlier studies, we noted that mostly charged residues, including D285 and E491 and neighbouring D219 and E220, in addition to the Glc product, stabilised water molecules[19]. In the current work, our interpretations of electron density maps at 1.95 Å resolution in the E220A mutant unambiguously placed one 2F-Glc moiety covalently linked to D285 at the separation of 1.39 Å between C1 of 2F-Glc and Oδ1 of D285, and constrained by H-bonds through neighbouring residues (Fig. 1d). The second 2F-Glc moiety was bound by the W286-W434 clamp, forming face-to-face stacking interactions, known as the London dispersion-driven C-H-π forces[42]. The comparison of $N_{water}/N_{residue}$ ratios 1.21 versus 0.47 (N = number of water molecules/residues) in the WT (1.95 Å; synchrotron data collection) and E220A (2.55 Å; in-house data collection) structures in-complex with 2F-DNPGlc, revealed a higher number of experimentally observable water molecules in WT. This disparity can be explained by the brilliance of X-rays used for WT data collection, allowing a better interpretation of protein conformation with associated ligands and water molecules[43,44].

Evaluation of side-chain dihedral angle distributions of residues is critical as they define discrete orientations of geometric and rotameric structural features. These parameters modulate entropy and spatial distributions of side-chains and rearrange H-bonds while adopting a finite set of statistically significant 3D positions[45,46]. In this framework, catalytic proteins represent special states as they traverse a multidimensional energy landscape during catalysis and adopt various rotamers. Notably, the typical free energy of an H-bond (around 15 to 20 kJ/mole) is in the same order as an entire protein folding free energy[47].

The analyses of dihedral angle distributions for side-chain χ2 and χ3 angles of E491, and χ1 and χ2 angles of N219 indicated multimodal patterns. Notably, in the E220A complex, the χ1 E491 angle distribution was broader and multifaceted, and had lower occurrences of the preferred χ1 and χ2 angle configurations seen in WT that had shifted towards a variety of values (Fig. 3a). These values reflected dispositions of E491 in the E220A complex, which are less likely to be compatible with optimal structural preferences for high-efficiency catalysis. We also observed that the dihedral χ1 and χ2 angles of N219 in WT and E220A exhibited reciprocal patterns in a side-chain frequency (Supplementary Fig. 4a), mainly due to the H-bond formed between Nδ2 of N219 and Oε1 or Oε2 of E491 in the E220A complex (Supplementary Fig. 4b). This 3.9 Å H-bond appears to "hijack" E491 and hinder the extraction of a proton evolving from the catalytic water required for hydrolysis.

We further compared the values of side-chain dihedral angle distributions of the WT and E220A crystal structures (Supplementary Table 6). We observed differences in dihedral angle distributions of N219 and E491

Fig. 6 | Ancestral sequence reconstruction (ASR) and phylogenetic trees of plant GH3 members. Rooted tree of representative plant GH3 members (117 entries; annotated by NCBI accessions) with putative β-ᴅ-glucan glucohydrolase (EC 3.2.1.-) activity in 12 monocot and dicot entries (pink circles and oval) reveals ancestral nodes 179, 180, 195 and 196 (red circles). Entries in blue and blue nodes indicate bacterial sequences bordering plant sequences. The tree generated through FireProt[ASR26,27] and visualised in FigTree v1.4.3[81] is based on the HvExoI sequence at 20–90% sequence identity to other entries as input. All plant sequences carry conserved catalytic D285 and E491, and N219, E220 and K260 residues involved in catalysis and water coordination, respectively. The percentage of replicate trees with taxa clustering is shown at node bipartitions. Substitutions per site scale are indicated.

between the WT and E220A structures and computational models, where those of the computational models showed much wider and/or multimodal distributions (Fig. 3b, Supplementary Fig. 4b and Supplementary Table 6). These differences amongst other factors, most likely reflect side-chain flexibilities of in-solution computational models compared to those of the crystal structures. They could largely be explained by the χ1 of N219 values, which underlie the disruption of an underlying H-bond network of a loop that carries E220A and is involved in water intake. This broader dihedral

angle distribution difference between the two complexes (around 80°) points to a more de-localised electronic density of N219 in the mutant complex, reflecting increased loop flexibility, which could upsurge the overall entropy of water molecules near the water-intaking loop and mitigate the correct E220A disposition, while forming the stable water molecule networks.

Differences between water molecule networks in the unliganded E220A and WT structures and their mechanism-based inhibitor complexes

warranted explorations by classical MD (cMD) simulations. These methods are powerful tools for studying the mechanisms of chemical reactions in complex environments. cMD simulations are considered the "gold standard" when using explicit water molecules[29]. The rationale for using cMD simulations is justified by a large inhomogeneity of water density, polarity and mobility, which is problematic to assess experimentally[5,39,48,49]. Water behaviour has also been studied by knowledge-based predictions, umbrella sampling and coarse-grained models, which are versatile as they sample configurations of water molecules and correlate thermodynamic properties of binding sites[40]. To gain insights into the dynamics of water molecule networks and catalytic pocket hydration sites contributing to long-range protein interactions, we defined how water molecules form hydration sites in complex networks and how designated water molecules function as nucleophiles during catalysis (Figs. 2–5, Supplementary Figs. 3–6, Supplementary Tables 5–7 and Supplementary Methods). These definitions could explain how water becomes a driving force for hydrolytic efficiency and how certain residues direct water molecules to target locations. More broadly, the quantification of thermodynamic contributions of a water solvent could impact the design of inhibitors of carbohydrate-processing enzymes[44].

Our computational models of covalent complexes contained the L5 substrate bound in the +1 and +2 subsites and facing bulk solvent, and the covalent D285-Glc adduct located in the −1 subsite[18,36,37]—hence, these reactants represented catalytic intermediates during a hydrolytic reaction. Through cMD simulations, we established the differences in water molecule network distributions and frequencies and how water molecule trajectories diverge in active site pockets. Their additional explorations of densities and movements, illustrated through molecular animation and based on DynaWatProt[30] using the WT and E220A crystal structures (in-complex with Glc and 2F-DNPGlc) revealed profound contrasts in global water dynamics and density (Supplementary Movies 1, 2). Similarly, through the data obtained with physics-based WaterKit[29] (Fig. 5b and Supplementary Fig. 7), we observed that water molecules in the E220A mutant were absorbed in lower numbers and were distributed deeper in the pocket in chaotic and disorganised manners. Higher configuration numbers of water molecules, adopted in the E220A active site, were reflected in higher entropy of water clusters. The changes in global dynamics and density of water molecule networks in both enzyme forms mirrored the availability of optimally positioned catalytic water molecules (Fig. 5d), which could explain 10- to 30-fold reduced catalytic efficiency by a mutant form. We concluded that the WaterKit predictions agreed with RDF parameters and frequency histograms of water molecule dispositions and predictions of dynamics of water molecule networks. Additionally, the cMD analyses of water molecule dispositions in the WT and E220A computational models matched those in crystal structures, whereby the dispositions of catalytic residues aligned with predicted water molecule networks (Figs. 1, 4c, 5a). To this end, additional analysis of water networks could include calculating their residence time. Here, the survival time of water molecules (or their average residence time) could be estimated from cMD data by obtaining a survival probability curve along the trajectory[14]. Such analysis was used to rationalise the inversion of a transglycosylation/hydrolysis (T/H) ratio of the engineered agarose AgaD carrying two mutations distant from the active site[14]. No mispositioning of catalytic waters nor changes in the conformation of a catalytic acid/base were identified that could explain the hydrolytic rate decrease. However, the purge time of the water channel (or residence time of a given water molecules cluster in a channel at a specific time to leave it) was much lower for the mutant than for WT. This agreed with a hydrolytic rate reduction, while the transglycosylation activity was almost unaffected. Although we did not perform this analysis, a similar effect may contribute to the reduced hydrolytic activity in the E220A mutant. These T/H relationships could be investigated in future work.

Substantial effort has been made to understand protein interactions. Conversely, little attention is directed to the principles of how they evolved. Evolution is a natural process through which we explore how it shapes structural determinants and molecular mechanisms of extant proteins. This knowledge could be used to improve catalytic function under rational "selection pressure" to engineer re-designed proteins. It is thought that ancestral sequences, defined through ASR, are more promiscuous than extant ones. Consequently, the enzyme-substrate specificity may have evolved through gene duplications or modifications, providing new utilities[50,51]. These could arise from an ancestor protein with a different function (neo-functionalisation)[52] or through the specialisation of a promiscuous ancestor (sub-functionalisation)[53]. In this sense, a vertical ASR comparative approach to analyse extant sequences, rather than a horizontal one, could distinguish which changes yield functional optimisation of a pre-existing activity or strict partitioning among paralogs[51,54].

A deeper understanding of the significance of ancestral sequences and evolutionarily conserved motifs in hydrolytic enzymes is needed, which explains the importance of active site residues, partaking in the coordination of water molecules during catalysis. To build evolutionary history in GH3 enzymes and to reconstruct ancestral sequences, generate rooted phylogenetic trees, and infer evolutionary conserved motifs with probable evolutionary relationships, we employed FireProt[ASR26,27] (Fig. 6, Supplementary Figs. 8, 9 and Supplementary Data 1–3). The emphasis was on the evolutionary conservation of E491 acid/base and the N219-E220-N221 (NEN) water-coordinating signature. Here, we observed that monocot and dicot plant entries, which carried the D285 and E491 catalysts and the N219-E220-K260 signature, were separated from bacterial clades. This analysis also identified node 196, as the most plausible ancestral node for plant entries—it would be inspirational to evaluate the structural traits of this node. We also observed that the NEN signature occurred in HvExoI, while NED or NEG signatures featured in other monocots (*Triticum turgidum* and *Hordeum vulgare*) or dicots (*Medicago truncatula*)—this observation indicated that the residue following E220 could vary, nonetheless, E220 was absolutely conserved. Notably, the K260 and M261 residues (KM motif), where K260 also participates in a water molecule network (Fig. 1), were conserved to respective 100% and 85% in monocot or dicot clades (Supplementary Figs. 8 and 9). We unexpectedly noted the KL signature instead of KM (Supplementary Fig. 8) in *Nymphaea colorata* (XP_031484720.1), which carries monocot and dicot traits. The extant Charophyte green alga *Klebsormidium nitens* (GAQ91658.1) unusually harboured a nearly identical catalytic signature as HvExoI (except DEG instead of NEN), signifying that residues around E220 could diversify from this point onwards (Supplementary Fig. 9). These data pointed to the evolution of unique conserved motifs controlling catalytic water organisation in GH3 exo-hydrolases.

We conclude that water molecules as one of the key reactants play a fundamental role in protein function, as they underline the properties of enzyme-substrate complexes and are essential for driving chemical reactions. We suggest that the water molecule-protein interactions harmonise solvent properties and control the entropy of water molecule networks that energise catalysis. Exploring these relationships in carbohydrate-active enzymes, such as HvExoI, is essential for understanding a global carbon cycle, which forms the basis of a multibillion-dollar biotechnological industry. In our view, the findings regarding water molecule networks and catalytic pocket hydration sites are central to a rational design of interrelated biotechnological applications and offer values beyond enzymology. These findings could provide a blueprint for the mechanism and dynamics of catalysis facilitated by hydrolytic enzymes, which may inspire future efforts to create a sustainable bio-economy through bioengineering.

Methods
Preparation of the HvExoI E220A mutant, protein expression, and purification
We used the pPICZαBNH8 vector containing an optimised cDNA fragment encoding mature HvExoI as a template, to prepare the E220A mutant by singly replacing glutamate (E220) for alanine (E220A). Mutagenesis was performed with primers (Supplementary Table 1) and the QuikChange® Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA, USA). The DNA construct with the E220A mutation was transformed in the *Pichia pastoris (Komagataella phaffii)* Yeast Strain SMD1168H (Invitrogen, Carlsbad, CA, USA) to express and purify the recombinant protein *via*

Immobilised Metal Affinity Chromatography[21,22,55,56] using the TALON metal affinity resin (Takara Bio, San Jose, CA, USA).

Enzyme stability, substrate specificity, kinetic analyses, and inhibition

Activities of the WT and E220A forms, dependent on pH (McIlvaine buffers in the pH range of 3.5–8.5 at 30 °C) and temperature (100 mM sodium acetate buffer, pH 5.25, temperature range 0–80 °C, 15 min incubations) were determined using 0.2% (w/v) 4NP-Glc. Relative rates of hydrolysis of β-linked poly-, oligosaccharides, and aryl-glucosides were measured using 45–50 pmoles of near-homogenous enzymes per reaction mixture. Apparent first-order rate parameters (K_M and k_{cat}) were determined with (1,3;1,6)-β-D-glucan (laminarin), barley (1,4;1,3)-β-D-glucan, laminaribiose, cellobiose, and 4NP-Glc at 0.3–3 times K_M value substrate concentrations in duplicate. Apparent inhibition constants of enzyme-inhibitor complexes (K_i) using 4NP-Glc were determined at six concentrations with G6SG-OMe, glucono-δ-lactone and 2F-DNPGlc at 0.3–2 times K_i values in duplicate[22]. Kinetic parameters were calculated from data based on analyses of Michaelis-Menten plots with Grafit 7.0.3[57] and processed by a proportional weighted fit, using nonlinear regression[58]. Relative rates of hydrolysis, and kinetic and inhibition data were rounded to one or two decimal points. Standard errors of measurements ($n = 2$–6) were calculated in Microsoft Excel.

Protein crystallisation and inhibitor perfusion

WT and mutant HvExoI were crystallised by macro- and cross-seeding with native HvExoI microcrystals[59]. Briefly, crystallisation was carried out at 4 °C in handing drops consisting of proteins and a precipitant solution A (100 mM HEPES-NaOH buffer, pH 7.0 containing 2.3% (w/v) polyethylene glycol 400 and 1.6 M ammonium sulfate). Drops were placed on siliconised glass coverslips, supplemented with several native microcrystals of about one μm length in the longest dimension and the E220A protein. Crystals of E220A grew within 7–24 days at 4 °C to 100–400 μm sizes in the longest dimension. Glc and 2F-DNPGlc were perfused at saturating concentrations dissolved in solution A for 5–60 min, cryo-protected with 15% (v/v) glycerol in solution A and mounted on a goniometer.

Crystal structure determination

The X-ray diffraction data were collected on the MX1 beamline of the Australian Synchrotron at 100 K (Oxford Instruments, UK) with a collimating mirror, double-crystal Si(111) monochromator 03BM1 dipole/bending magnet, and ADSC Quantum 210r Detector, as described[21,22]. All data were collected at 0.5° to 1° oscillations throughout 180–720°. Data were processed with the DENZO/SCALEPACK HKL 2000[60]. Autoindexing determined that space groups of crystals were consistent with a tetragonal space group P4₃2₁2. Structures were refined using REFMAC5[61] and Phenix[62] using the HvExoI structures (PDB accessions 1EX1, 1IEQ, 1IEV, 1IEW, and 1J8V) as starting models after removing ligands, ions, glycerol, and water molecules[19,20,38]. The iterative model building using REFMAC5[61], MIFIT[63], and Coot[64] traced residues in electron density maps. Following convergence, approximately 2% improvements in R_{work}/R_{free} factor ratios were achieved by refining domain 1 (residues 1–357) and domain 2 (residues 374–559) as independent anisotropic domains with translation-libration-screw motion[65]. The positions of inhibitors were defined in active sites at 3σ levels in $m|F_o| - |F_c|$ electron density maps (m is the figure of merit and D is an estimated coordinate error), and water molecules were located automatically with CCP4 ARP at levels higher than 2σ levels. Water molecules were assigned with satisfied H-bond criteria and if $2m|F_o| - D|F_c|$ maps were confirmed after refinement. During model building and refinement, 5% of the data were flagged for cross-validation using R_{free} statistics[66,67]. PROCHECK[68] was used to check stereo-chemical parameters, where Ramachandran plots indicated that 99.8% residues were found in the most favourable, additionally allowed, and generously allowed regions of the plot, with a well-defined density map for Ile432, which was the only residue located in a disallowed region. Structural images for stereoscopic convergent

and other viewing were prepared in PyMOL Molecular Graphics System v2.3.3 (Schrödinger LLC, Portland, OR, USA). The $2m|F_o| - D|F_c|$ maps were calculated using Phenix[62] with mtz and pdb files after the refinements and contoured at 1σ levels. The omit electron density maps of E220A in-complex with Glc and 2F-DNPGlc were calculated with REFMAC5[61] after removing Glc and 2F-Glc, respectively, from the −1 and +1 subsites.

Classical molecular dynamics (cMD) simulations of the (1,3)-linked β-D-glucoside laminarihexaose (L6) hydrolysis products

Computational models were built for the WT and E220A mutant forms containing (1,3)-linked β-D-glucoside laminaripentaose (L5), originating from L6 in the +1 and +2 subsites and the Glc-D285 adduct in the −1 subsite, based on the crystal structures of WT (in-complex with Glc; PDB 3WLH; 1.65 Å) and E220A (in-complex with 2F-DNPGlc; PDB 8HJ8; 1.85 Å) enzyme forms. Starting computational models were prepared using USFC (University of California San Francisco, USA) Chimera[69], where ion, water, glycerol (Gol), and polyethylene molecules in crystal structures were removed, and proteins were protonated at pH 7 with E491 catalytic acid/base in a neutral form[38]. The L5 substrate was generated using the carbohydrate builder GLYCAM (http://glycam.org/cb/legacy). To ensure that starting models contained the non-reducing end of L5 occupying the +1 and +2 subsites, L5 was spatially matched to substrates present in the crystal structures of the WT and E220A, applying a match command using USFC Chimera. The same protocol was followed to ensure the correct spatial arrangement of the non-standard covalent Glc-D285 adduct. System solvation, coordinate, and topology file acquisitions for standard protein residues and ligands (prmtop and inpcrd files) were performed in tleap of the AMBERTools21 suite of Amber21[70]. All systems were solvated in TIP3P water molecules and neutralised with Na⁺ ions, using a cubic box, with standard periodic boundary conditions set. The systems were parametrised using the Amber ff99SB-ILDN force fields for proteins[71], carbohydrates[72], and TIP3P water molecules[73]. The parametrisation of the covalent non-standard Glc-D285 adduct was performed with GAFF[74], and atomic charges were obtained using initial geometry optimisation followed by the calculation of restrained electrostatic potential (RESP) charges with Gaussian16[75]. In addition, the AmberTools21[76] package antechamber was used to process RESP charges and obtain the files required by tleap to generate the final set of topological parameters. Simulations were carried out in AMBER 22[76] with CUDA acceleration. The cMD simulation protocol consisted of (1) initial minimisation (10^5 steps); (2) heating (100 ps, with protein backbones restrained); (3) NVT equilibration (100 ps, with restrained backbones); (4) NPT equilibration (500 ps, unrestrained); and (5) NPT production at 300 K under the pressure of 1.0 atm. Each simulation was carried out with a time step of 2 ps using the SHAKE algorithm[77] to constrain bonds involving hydrogens. A cutoff value of 14.0 Å was used for non-bonding interactions and the Particle Mesh Ewald method for long-range electrostatics[78]. Production simulations were run for 1000 ns (1 μs) in all cases with three replicates of each system, amounting to a total of six simulations. The H-bond fractional occupancy heat term (average of three replicates) refers to the frequency of H-bonds, based on the analyses of solute or solvent H-bonds incidences over total frames of a given cMD simulation—e.g. the value 1.54 for N219 in WT stems from the contributions of the side-chain Oδ1 and Nδ2 atoms by adding 1.14 and 0.4, respectively (cf. Fig. 2c). The cluster entropy analyses of replicates, unrelated to thermodynamic entropy, were conducted as described in Supplementary Methods.

Thermodynamics profiling of catalytic pocket hydration sites using WaterKit

Predictions of dispositions of water molecules and hydration densities were performed using WaterKit[29] using WT (PDB 3WLH) and E220A (PDB 8HJ7) crystal structures in-complex with Glc. The resulting pdb files were transformed into pdbqt files by modules AmberTools21[70] (included in WaterKit), followed by the sampling of water molecule dispositions with the script run_waterkit.py. After individual trajectories were generated, the calculations based on the grid inhomogeneous solvation theory (GIST)[79]

were implemented using cpptraj[80], where catalytic pocket hydration sites were identified iteratively using oxygen density maps resulting from GIST calculations and applying python scripts[29].

Catalytic pocket water molecule network density profiling using DynaWatProt

Analyses and density profiling in computational models of the WT and E220A covalent complexes were performed using the Python package DynaWatProt (Benoît: Getwatchan), based on explicit MD simulations. DynaWatProt includes scripts and programmes to isolate and analyse the dynamics and positions of putative water molecule networks buried within a protein structure. For each system and individual replicate, trajectories with implicit water molecules were aligned onto an initial frame and saved in several pdb files with a step corresponding to a time step of 200 ps of simulation time. The analysis was run with a cutoff of 2.5 Å and putative water molecule networks were acquired as an ensemble of individual water clusters present during MD simulations. DynaWatProt structural profiles were visualised in PyMOL.

Ancestral sequence reconstruction (ASR) and phylogeny of the glycoside hydrolase family-3 (GH3)

The ASR analysis proceeded using FireProt[ASR26,27] in two steps using the default parameters for evolutionary models and bootstrap selection. We used the HvExoI sequence as an input. We also specified as essentials D285 and E491 (HvExoI numbering) to search for homologous sequences at 20-90% sequence identity to HvExoI to identify ancestral sequences of GH3 entries (Fig. 6, Supplementary Fig. 8 and Supplementary Data 1–3), and at 30-90% sequence identity to HvExoI to identify ancestral sequences in plant GH3 entries (Supplementary Fig. 9 and Supplementary Data 1–3). This step, after homology search and sequence filtering, aggregated a dataset of 233 bacterial, metagenome and plant sequences to identify GH3 ancestral sequences, and 130 sequences to identify plant GH3 ancestral sequences. This dataset, from which we removed incomplete sequences and entries containing variable catalytic residues, was used for the identification of ancestral sequences and the building of phylogenetic trees. Trees were visualised in FigTree v1.4.3[81] with individual sequences annotated by the National Centre for Biotechnology Information (NCBI) accession numbers and in selected instances by the species of origin. Trees were drawn to scale, with branch lengths specified in the same units as those of evolutionary distances, which were used to infer these trees. Evolutionary distances were computed using the p-distance method and are expressed in units of a residue number difference per site. The PROMALS3D multiple sequence alignments[82] were generated in selected cases to visualise catalytic and water-coordinating residues.

Statistics and reproducibility

Enzyme stability, substrate specificity, kinetic analyses, and inhibition, based on two to three independent replicates, are given as mean values in one to three significant digits with standard error of measurements. Protein crystallisation and inhibitor perfusion experiments were performed more than three times. cMD simulation data, based on three replicates, are expressed in one or two significant digits as mean values with standard error of measurements. Dihedral angles of N219 and E491, based on the average of 15,000 frames of three replicates, are given as mean values with standard error of measurements. Averages for $\chi 2$, $\chi 3$ of E491, and $\chi 1$, $\chi 2$ of N219 are expressed as squared values to account for dihedral angle distributions with standard error of measurements. ASR and phylogeny analyses of GH3 entries were performed twice. The method section and supplementary files contain comprehensive information regarding the experimental particulars and the data used.

Data availability

Atomic coordinates and structure factors of HvExoI in the apo (unliganded)-form and in-complex with D-glucose, and 2F-Glc were deposited in the Protein Data Bank (www.wwpdb.org) under the accession codes 8HJ6,

8HJ7, and 8HJ8, respectively. Primary data from molecular dynamics simulations are stored in the RDR CORA repository (https://cora.csuc.cat/) under https://doi.org/10.34810/data2160. Supplementary Movies 1, 2 are available in the Figshare repository under 28367753. Supplementary Data 1–5 are provided with this paper.

Code availability

Software applications described in "Methods" with associated references were used without code modifications. For analyses of geometrical parameters for water molecule network dynamics and profiling, and cluster entropy analysis, we used in-house scripts available from the Zenodo repository[83].

References

1. Shinobu, A. & Agmon, N. Proton wire dynamics in the green fluorescent protein. *J. Chem. Theory Comput.* **13**, 353–369 (2017).
2. Frenkel-Pinter, M., Rajaei, V., Glass, J. B., Hud, N. V. & Williams, L. D. Water and Life: the medium is the message. *J. Mol. Evol.* **89**, 2–11 (2021).
3. Ribeiro, A. J. M. et al. Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
4. Nagano, N. et al. EzCatDB: the enzyme reaction database, 2015 update. *Nucleic Acids Res.* **43**, D453–D458 (2015).
5. Maurer, M. & Oostenbrink, C. Water in protein hydration and recognition. *J. Mol. Recognit.* **32**, e2810 (2019).
6. de Souza, J. P. & Stone, H. A. Protein condensation regulates water availability in cells. *Nature* **623**, 698–699 (2023).
7. Watson, J. L. et al. Macromolecular condensation buffers intracellular water potential. *Nature* **623**, 842–852 (2023).
8. Joshi, A. et al. Hydrogen-bonded network of water in phase-separated biomolecular condensates. *J. Phys. Chem. Lett.* **15**, 7724–7734 (2024).
9. Smith, S. T., Shub, L. & Meiler, J. PlaceWaters: real-time, explicit interface water sampling during Rosetta ligand docking. *PLoS ONE* **17**, e0269072 (2022).
10. Pavlova, M. et al. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.* **5**, 727–733 (2009).
11. Sykora, J. et al. Dynamics and hydration explain failed functional transformation in dehalogenase design. *Nat. Chem. Biol.* **10**, 428–430 (2014).
12. Keutsch, F. N. & Saykally, R. J. Water clusters: untangling the mysteries of the liquid, one molecule at a time. *Proc. Natl Acad. Sci. USA* **98**, 10533–10540 (2001).
13. García, A. E. & Hummer, G. Water penetration and escape in proteins. *Proteins* **38**, 261–272 (2000).
14. David, B. et al. Internal water dynamics control the transglycosylation/hydrolysis balance in the agarase (AgaD) of *Zobellia galactanivorans*. *ACS Catal.* **7**, 3357–3367 (2017).
15. Romero-Téllez, S., Lluch, J. M., González-Lafont, A. & Masgrau, L. Comparing hydrolysis and transglycosylation reactions catalyzed by *Thermus thermophilus* β-glycosidase. A combined MD and QM/MM study. *Front. Chem.* **7**, 200 (2019).
16. Buschiazzo, A. et al. Structural basis of sialyltransferase activity in trypanosomal sialidases. *EMBO J.* **19**, 16–24 (2000).
17. Das, S., Raucci, U., Neves, R. P. P., Ramos, M. J. & Parrinello, M. How and when does an enzyme react? Unraveling α-amylase catalytic activity with enhanced sampling techniques. *ACS Catal.* **12**, 8092–8098 (2023).
18. Sinnott, M. L. Catalytic mechanisms of enzymic glycosyl transfer. *Chem. Rev.* **90**, 1171–1202 (1990).

19. Hrmova, M. et al. Catalytic mechanisms and reaction intermediates along the hydrolytic pathway of plant β-D-glucan glucohydrolase. *Structure* **9**, 1005–1016 (2001).

20. Hrmova, M. et al. Structural basis for a broad specificity in higher plant β-D-glucan glucohydrolases. *Plant Cell* **14**, 1–22 (2002).

21. Streltsov, V. A. et al. Discovery of processive catalysis by an exo-hydrolase with a pocket-shaped active site. *Nat. Commun.* **10**, 2222 (2019).

22. Luang, S. et al. The evolutionary advantage of an aromatic clamp in plant family 3 glycoside exo-hydrolases. *Nat. Commun.* **13**, 5577 (2022).

23. Dong, J. et al. Tailor-made α-glucans by engineering the processivity of α-glucanotransferases via tunnel-cleft active center interconversions. *J. Agric. Food Chem.* **72**, 11041–11050 (2024).

24. Nong, D. et al. Single-molecule tracking reveals dual front door/back door inhibition of Cel7A cellulase by its product cellobiose. *Proc. Natl Acad. Sci. USA* **121**, e2322567121 (2024).

25. Hrmova, M. et al. Structural rationale for low nanomolar binding of transition state mimics to a family GH3 β-D-glucan glucohydrolase from barley. *Biochemistry* **44**, 16529–16539 (2005).

26. Khan, R. T., Musil, M., Stourac, J., Damborsky, J. & Bednar, D. Fully automated ancestral sequence reconstruction using FireProt^ASR. *Curr. Protoc.* **1**, e30 (2021).

27. Musil, M. et al. FireProt^ASR: a web server for fully automated ancestral sequence reconstruction. *Brief. Bioinform.* **22**, 1–11 (2021).

28. Wheat, M. How to calculate clustering entropy? A working example or software code. https://stackoverflow.com/questions/35709562/how-to-calculate-clustering-entropy-a-working-example-or-softwarecode (2016).

29. Eberhardt, J. & Forli, S. WaterKit: thermodynamic profiling of protein hydration sites. *J. Chem. Theory Comput.* **19**, 2535–2556 (2023).

30. Benoît, D. Getwatchan: Python package for determination of water channels from explicit MD simulations. Université de Nantes. https://gitlab.univ-nantes.fr/Ben/DynaWatProt (2017).

31. Cremer, D. & Pople, J. A. A general definition of ring puckering coordinates. *J. Am. Chem. Soc.* **97**, 1354–1358 (1975).

32. Geronimo, I., Payne, C. M. & Sandgren, M. The role of catalytic residue pKa on the hydrolysis/transglycosylation partition in family 3 β-glucosidases. *Org. Biomol. Chem.* **16**, 316–324 (2018).

33. Drula, E. et al. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).

34. Hrmova, M. & Schwerdt, J. G. Molecular mechanisms of processive glycoside hydrolases underline catalytic pragmatism. *Biochem. Soc. Trans.* **51**, 1387–1403 (2023).

35. Hrmova, M., Zimmer, J., Bulone, V. & Fincher, G. B. Enzymes in 3D: synthesis, remodelling and hydrolysis of cell wall (1,3;1,4)-β-glucans. *Plant Phys.* **194**, 33–50 (2024).

36. Barrett, T., Suresh, S. G., Tolley, S. P., Dodson, E. J. & Hughes, M. A. The crystal structure of cyanogenic, β-glucosidase from white clover, a family 1 glycosyl hydrolase. *Structure* **3**, 951–960 (1995).

37. MacKenzie, L. F. et al. Crystal structure of the family 7 endoglucanase I (Cel7B) from *Humicola insolens* at 2.2 Å resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate. *Biochem J.* **335**, 409–416 (1998).

38. Varghese, J. N., Hrmova, M. & Fincher, G. B. Three-dimensional structure of a barley β-D-glucan exohydrolase, a family 3 glycosyl hydrolase. *Structure* **7**, 179–190 (1999).

39. Damjanović, A. B., García-Moreno, E., Lattman, E. E. & García, A. E. Molecular dynamics study of water penetration in staphylococcal nuclease. *Proteins Struct. Funct. Gen.* **60**, 433–449 (2005).

40. Levitt, M. & Park, B. H. Water: now you see it, now you don't. *Structure* **1**, 223–226 (1993).

41. Sobhia, M. E. et al. The role of water network chemistry in proteins: a structural bioinformatics perspective in drug discovery and development. *Curr. Top. Med. Chem.* **22**, 1636–1653 (2022).

42. Raju, R. K., Ramraj, A., Vincent, M. A., Hillier, I. H. & Burton, N. A. Carbohydrate-protein recognition probed by density functional theory and ab initio calculations including dispersive interactions. *Phys. Chem. Chem. Phys.* **10**, 6500–6508 (2008).

43. Gnesi, M. & Carugo, O. How many water molecules are detected in X-ray protein crystal structures. *J. Appl. Cryst.* **50**, 96–101 (2017).

44. Muench, S. P., Antonyuk, S. V. & Hasnain, S. S. The expanding toolkit for structural biology: synchrotrons, X-ray lasers and cryoEM. *IUCrJ* **6**, 167–177 (2019).

45. Dunbrack, R. L. Jr. & Karplus, M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574 (1993).

46. Gaudreault, F., Chartier, M. & Najmanovich, R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* **28**, i423–i430 (2012).

47. Scaletti, C. et al. Hydrogen bonding heterogeneity correlates with protein folding transition state passage time as revealed by data sonification. *Proc. Natl Acad. Sci. USA* **121**, e2319094121 (2024).

48. Shen, L. & Yang, W. Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks. *J. Chem. Theory Comput.* **14**, 1442–1455 (2018).

49. Matthews, B. W. & Liu, L. A review about nothing: are apolar cavities in proteins really empty?. *Protein Sci.* **18**, 494–502 (2009).

50. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).

51. Scossa, F. & Fernie, A. R. Ancestral sequence reconstruction—an underused approach to understand the evolution of gene function in plants?. *Comp. Struct. Biotech. J.* **19**, 1579–1594 (2021).

52. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).

53. Steindel, P. A., Chen, E. H., Wirth, J. D. & Theobald, D. L. Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. *Prot. Sci.* **25**, 1319–1331 (2016).

54. Attah, V. et al. Duplication and neofunctionalization of a horizontally transferred xyloglucanase as a facet of the Red Queen coevolutionary dynamic. *Proc. Natl Acad. Sci. USA* **121**, e2218927121 (2024).

55. Luang, S., Hrmova, M. & Ketudat Cairns, J. R. High-level expression of barley β-D-glucan exohydrolase HvExoI from a codon-optimized cDNA in *Pichia pastoris*. *Prot. Exp. Purif.* **73**, 90–98 (2010).

56. Luang, S., Ketudat Cairns, J. R., Streltsov, V. A. & Hrmova, M. Crystallisation of wild-type and variant forms of a recombinant β-D-glucan glucohydrolase from barley (*Hordeum vulgare* L.) by macroseeding with wild-type native microcrystals and preliminary X-ray analysis. *Int. J. Mol. Sci.* **11**, 2759–2769 (2010).

57. Perella, E. EZ-FIT: a practical curve-fitting microcomputer program for the analysis of the enzyme kinetic data on IBM-PC compatible computers. *Anal. Biochem.* **174**, 437–447 (1988).

58. Leatherbarrow, R. J. *Data Analysis and Graphics Program for the IBM PC* (Erithacus Software Ltd., 1998).

59. Hrmova, M., Varghese, J, N., Høj, P. B. & Fincher, G. B. Crystallization and preliminary X-ray analysis of β-glucan exohydrolase isoenzyme ExoI from barley (*Hordeum vulgare*). *Acta Crystallogr.* **D54**, 687–689 (1998).

60. Otwinowski, Z. & Minor, W. A. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1996).

61. Marshudov, G. N. P. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr.* **D67**, 355–367 (2011).

62. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **D66**, 213–221 (2010).

63. McRee, D. E. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1996).

64. Emsley, P., Lohkamp, B., Scott, G. W. & Cowtan, K. Features and development of coot. *Acta Crystallogr.* **D66**, 486–501 (2010).

65. Painter, J. & Merritt, E. A. Molecular viewer for the analysis of TLS rigid-body motion in macromolecules. *Acta Crystallogr.* **D61**, 465–471 (2005).

66. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

67. Brünger, A. T. *X-PLOR Version 3.851: A System for Crystallography and NMR* (Yale University Press, 1992).

68. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).

69. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

70. Case, D. A. et al. *Amber 2021* (University of California, 2021).

71. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

72. Kirschner, K. N. et al. GLYCAM06: a generalizable biomolecular force field. *Carbohydr. J. Comput. Chem.* **29**, 622–655 (2008).

73. Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R. & Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

74. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

75. Frisch, M. J. et al. Gaussian 16, Revision C.01 (2016).

76. Case, D. A. et al. *AMBER 2022* (University of California, 2022).

77. Ryckaert, J. P., Ciccotti, G. & Berendsen, J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (2017).

78. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089 (1993).

79. Nguyen, C. N., Young, T. K. & Gilson, M. K. Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **137**, 044101 (2012).

80. Roe, D. & Cheatham, T. Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

81. Rambaut, A. Figtree v1.4.3 (Institute of Evolutionary Biology, University of Edinburgh, 2018).

82. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–22300 (2008).

83. Luang, S. et al. The structure and dynamics of water molecule networks underlie catalytic efficiency in a glycoside exo-hydrolase. *Commun. Biol*. ZENODO repository. https://zenodo.org/records/13149550.

84. Hrmova, M. & Fincher, G. B. Purification and properties of three (1→3)-β-D-glucanase isoenzymes from young leaves of barley (*Hordeum vulgare*). *Biochem. J.* **289**, 453–461 (1993).

85. Fersht, A. *Structure and Mechanism in Protein Science* 1–631 (W. H. Freeman and Co, 1999).

Acknowledgements

Author contributions

M.H. and L.M. designed experiments. S.L. completed enzyme characterisation and kinetics; S.L. and M.H. purified and crystallised proteins. V.A.S. and M.H. collected X-ray data; S.L., V.A.S., and M.H. refined crystal structures and deposited in the Protein Data Bank; X.F.-L., J.-D.M., and L.M. implemented MD simulations. M.H. performed ancestral sequence reconstruction; X.F.-L. and M.H. generated animations and structural graphics. All authors analysed data. M.H. wrote the manuscript with contributions from X.F.-L. and L.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08113-9.

Correspondence and requests for materials should be addressed to Laura Masgrau or Maria Hrmova.

Peer review information *Communications Biology* thanks Rui Neves and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Laura Rodríguez Pérez. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.