CHAP: A versatile tool for the structural and functional annotation of ion channel pores

Gianni Klesse[1,2], Shanlin Rao[2], Mark S.P. Sansom[2,3,] & Stephen J. Tucker[1,3,*]*

[1]Clarendon Laboratory, Department of Physics, University of Oxford, UK
[2]Department of Biochemistry, University of Oxford, UK
[3]OXION Initiative in Ion Channels and Disease, University of Oxford, UK

* to whom correspondence should be addressed: *stephen.tucker@physics.ox.ac.uk* or *mark.sansom@bioch.ox.ac.uk*

Abstract:

The regulation of ion channel and transporter function requires the modulation of energetic barriers or 'gates' within their transmembrane pathways. However, despite the ever-increasing number of available structures, our understanding of these barriers is often simply determined from calculating the physical dimensions of the pore. Such approaches (e.g. the HOLE program) have worked very well in the past, but there is now considerable evidence that the unusual behaviour of water within the narrow hydrophobic spaces found within many ion channel pores can also produce energetic barriers to ion conduction without requiring physical occlusion of the permeation pathway. Several different classes of ion channels have now been shown to exploit this principle of 'hydrophobic gating' to regulate ion flow. However, measurement of pore radius alone is unable to identify such barriers and new tools are required for more accurate functional annotation of an exponentially increasing number of ion channel structures. We have previously shown how molecular dynamics simulations of water behaviour can be used as a proxy to accurately predict hydrophobic gates. Here we now present a new and highly versatile computational tool, the Channel Annotation Package (CHAP) that implements this methodology to predict the conductive status of new ion channel structures.

Introduction

Ion channels form pores that facilitate the selective movement of ions across lipid bilayers. They therefore control a wide range of vital physiological functions and represent an important class of therapeutic targets. However, these pores are also highly dynamic structures that switch between functionally closed and open states in response to different physiological stimuli, a process referred to as gating (Hille, 1992).

Elucidating the molecular mechanisms that underlie these gating processes remains an important objective for ion channel structural biology and recent advances in techniques such as cryo-electron microscopy are now generating an increasing wealth of new ion channel structures, including many captured in different conformational states (**Figure 1**). However, before this structural information can be compared to the detailed electrophysiological data that can often be obtained for many ion channels, the conductive status of any new structure needs to be known. Conventionally, this has been achieved by determining the physical dimensions of the channel pore using software such as HOLE (Smart et al., 1996). This calculates a radius profile along the conduction pathway and the structure is typically considered to be open and conductive if its narrowest constriction exceeds the radius of the hydrated ion species it conducts (**Figure 1A**).

This approach has served the ion channel community well for several decades, but an increasing body of evidence has demonstrated this approach is unable to detect gates that are not based on steric occlusion of the permeation pathway. For example, 'hydrophobic gates' can occur in channel pores to prevent ion flow and rely on the capillary evaporation of water that occurs when the lining of the pore is particularly hydrophobic (Aryal et al., 2015). In these situations, ion movement through these 'de-wetted' sections of the pore is prevented by the free energy barrier associated with stripping the ions of their hydration shell (**Figure 1B**) rather than by van der Waals clashes with tightly packed pore residues (Trick et al., 2015). It is important to note that this definition of hydrophobic gating excludes

cases where de-wetting precedes collapse of hydrophobic pore segments involved in steric closure of a gate (Jensen et al., 2010).

The phenomenon of hydrophobic gating was first observed in simulations of simple model pores (Beckstein et al., 2001; Beckstein and Sansom, 2003; Allen et al., 2002) and has been found to occur within pores up to 1.2 nm in diameter (Beckstein and Sansom, 2004). Thus, an ion channel can be functionally closed even if the pore appears physically wide enough to accommodate the relevant hydrated ion.

The first example of a hydrophobic gate in an ion channel was found in an early cryo-EM structure of the nicotinic acetylcholine receptor, which represents a functionally closed state without being sterically occluded (Beckstein and Sansom, 2006). Further examples of hydrophobic gates and barriers include the MscL mechanosensitive channel (Spronk et al., 2006), the cation-selective bacterial channel GLIC (Zhu and Hummer, 2012), the BK and TWIK-1 potassium channels (Jia et al., 2018; Aryal et al., 2014), the BEST1 chloride channel (Rao et al., 2017), and the CorA magnesium channel (Neale et al., 2015). Thus hydrophobic gating clearly represents an important mechanism for regulating ion channel permeation and needs to be taken into consideration when assessing the functional status of any new ion channel structure.

In a previous study we have shown how molecular dynamics (MD) simulations of water behaviour within these nanoscale structures can be used to aid this process (Trick et al., 2016). Here we now present a new and highly versatile computational tool that implements this methodology to predict the conductive status of new ion channel structures.

Workflow for 'Channotation' of Pore Structures

The workflow underlying this Channel Annotation Package (CHAP) is shown in **Figure 2** and the required software is freely available (www.channotation.org). As an initial analysis step, CHAP can be run directly on the PDB structure of any channel or nanopore to generate both radius and hydrophobicity profiles of the putative ion conduction pathway. These initial steps are easily performed on an average desktop computer and do not require MD simulation of the structure. However, if these profiles suggest the channel may contain a hydrophobic gate or any other interesting features then a more in-depth analysis based on equilibrium MD simulation of the structure can be performed.

The example illustrated in **Figure 2** is a crystal structure of a mammalian 5-HT_3 receptor (PDB ID: 4PIR) in which the existence of a hydrophobic gate has been demonstrated by previous studies (Trick et al., 2016; Yuan et al., 2016). Initial structural analysis with CHAP (i.e. without simulation) yields plots of both radius and hydrophobicity for the entire channel pore. The simple radius plot shows the size of the transmembrane section of the pore to be roughly constant at ~0.3 nm radius, however, the accompanying hydrophobicity plot also reveals this section to be somewhat hydrophobic, especially in the region of the 9' ring of side chains.

To test for the presence of a hydrophobic gate within this region, the structure was inserted in a lipid bilayer system ready for equilibrium MD simulations using GROMACS 2016. Briefly, the starting structure was embedded in a POPC bilayer using a serial multiscale procedure (Stansfeld and Sansom, 2011) and the resulting protein-bilayer system was then simulated for 30 ns using the OPLS united atom force field (Jorgensen et

al., 1996) with united-atom lipids (Tieleman et al., 2006) and the TIP4P water model (Jorgensen et al., 1983). The $C\alpha$ atoms of all protein residues were also restrained to ensure the simulated channel closely resembles the experimentally determined structure (see **Appendix A** for details of the simulation parameters).

As shown in **Figure 2**, this allows calculation of the water density within the pore together with a free energy profile for hydration of the pore. The behaviour of water during this particular simulation is displayed in **Figure 3**. This shows that although the pore is fully hydrated in the initial configuration of the simulation system, water is expelled almost immediately (within < 0.5 ns), leaving a 1 nm long region of the conduction pathway fully dehydrated. From the profiles calculated by CHAP it therefore becomes immediately apparent that the transmembrane section of this conformation of the 5-HT_3 receptor pore contains a hydrophobic gate. This would not have been revealed if only the radius profile shown in **Figure 1A** had been considered.

Implementation of CHAP

Thus although tools like HOLE have proven exceptionally useful it is clear that additional information is required if more accurate predictions are to be made. However, a simple redesign of the HOLE program with added functionality was not practical for several reasons. In particular, HOLE was originally developed for the analysis of static protein structures and is unable to read common trajectory data formats or process time-dependent structures similar to those generated by MD simulation. We therefore decided to implement this new channel annotation package as an independent tool: CHAP is written in C++ and is based on the trajectory analysis framework of the popular MD simulation software GROMACS (Abraham et al., 2015).

The algorithm used by CHAP to calculate these different channel profiles is outlined in **Figure 4**. Starting from an input structure (either an experimentally derived PDB file or a single frame from a MD trajectory), the first step is to define the physical dimensions of the pore. A probe-based method is employed, similar to that used by HOLE, resulting in a sequence of discrete probe positions and associated pore radii. A continuous spatial curve is then determined from these probe positions by means of B-spline interpolation. A continuous radius profile is then calculated by interpolating the probe radii along this curve (**Figure 5**).

The interpolating spline curve represents the centre line of the pore and defines a coordinate system that forms the basis for all subsequent calculations. In this channel coordinate system, the position of a particle (e.g. a water molecule or a protein side chain) is represented by its distance to the closest point on the centre line and the distance along the centre line between this point and a reference point (**Figure 6A**). By convention, the reference point is the projection of the centre of geometry of the overall channel protein onto the centre line.

Next, a hydrophobicity profile is determined for the residues that line the pore. By converting residue positions to channel coordinates, CHAP identifies all residues within a cutoff distance from the pore surface and then selects only those whose side chains point inwards towards the pore. The resulting discrete values for hydrophobicity at the positions of the individual residues are then converted into a continuous hydrophobicity profile along the centre line by means of kernel smoothing (Hastie et al., 2013).

If the input structure contains water molecules, e.g. from an MD simulation, then CHAP will also estimate the solvent density along the centre line. To achieve this, water molecule positions are first transformed into channel coordinates, and then a one-dimensional probability density is calculated by kernel density estimation (Sheather and Jones, 1991). Together with the radius profile, this allows us to calculate the number density of water molecules throughout the channel structure. It is from this value that the related water free energy profile can then be estimated using Boltzmann inversion.

When the input to CHAP is an MD trajectory, the above procedure is repeated independently for each frame. CHAP will then calculate summary statistics (mean, standard deviation, minimum, and maximum) over time for each point in the channel profiles. The following sections now describe these different analysis steps in more detail.

Pore Geometry and Centre Line

The path-finding algorithm used by CHAP is based on a hard sphere representation of the protein, in which each atom is associated with a sphere of radius r_i centred at the position of the atom, $\mathbf{Q}_i$. **Figure 5A** illustrates how a pathway is found by squeezing a spherical probe through this collection of van der Waals spheres. Starting from an initial position inside the channel pore, $\mathbf{P}_0$, the probe is moved in steps of Δv along the direction of a unit vector $\mathbf{e}_v$ pointing in the overall direction of the permeation pathway:

$$\tilde{\mathbf{P}}_i = \mathbf{P}_{i-1} + \Delta v \mathbf{e}_v$$

The resulting point is then used to initialise a numerical optimisation procedure seeking to maximise the radius of the probe,

$$R(\mathbf{P}) = \min_{j \in [1,N]} ||\mathbf{Q}_j - \mathbf{P}|| - r_j$$

without overlapping the van der Waals sphere of any protein atom. This optimisation procedure is carried out in the plane supported by $\tilde{\mathbf{P}}_i$ and spanned by the unit vectors $\mathbf{e}_u$ and $\mathbf{e}_w$ which are chosen to be perpendicular to both $\mathbf{e}_v$ and one another. The optimal probe position

$$\mathbf{P}_i = \arg\max_{u,w} R(\tilde{\mathbf{P}} + u\mathbf{e}_u + v\mathbf{e}_w)$$

is accepted as a point on the pore centre line and a further step in the direction of $\mathbf{e}_v$ is carried out. This procedure is repeated until the probe radius exceeds a threshold, at which point it is assumed that the probe has left the pore. Subsequently, the iteration is carried out a second time moving in the direction of $-\mathbf{e}_v$ and starting from point $\mathbf{P}_1$.

In the original algorithm used by HOLE the optimisation of the probe position was performed by means of simulated annealing (Smart et al., 1996). However, being a global optimisation procedure, simulated annealing is prone to allowing the probe to jump out of the protein pore. The downhill simplex algorithm (Nelder and Mead, 1965) was therefore implemented in CHAP alongside the simulated annealing procedure and this latter method was found to be generally more robust.

This procedure has three free parameters: The probe step size, Δv, the number of Nelder-Mead iterations, N_{simplex}, and the initial probe position, P_0. For a sufficiently symmetric protein with a central pore, P_0 can simply be set to the centre of geometry of the overall channel protein. The step size needs to be small enough to resolve all structural features of interest. As illustrated in **Figure 5F**, a probe step of 0.1 nm (the van der Waals radius of a hydrogen atom) is sufficient and the radius profile does not change significantly when the probe step is decreased further. **Figure 5E** shows the dependence of the error in the radius profile on the number of simplex iterations and reveals that 100 steps are sufficient to converge the profile to within a tolerance of 0.001 nm even for a very small probe step. This is the case not just for this particular 5-HT_3 receptor structure, but also for a wide range of other ion channel proteins (**Figure S1**).

The points $\mathbf{P}_i$ on the centre line are then connected into a continuous spatial curve $\mathbf{S}(s)$ by means of B-spline interpolation as detailed in **Appendix B** so that the curve passes exactly through all centre line points:

$$\mathbf{S}(s_i) = \mathbf{P}_o$$

A corresponding continuous radius profile is then found by interpolating the probe radii along the centre line curve. The main advantage of a continuous centre line curve over individual points is that the curve can be used to define a simple one-dimensional reaction coordinate for the motion of particles through the channel pore, as is detailed in the following section. Additionally, it can be used to generate a pore surface representation through the path extrusion technique as shown in **Figure 5D**.

Water Density and Free Energy

The centre line spline can now be used to define a curvilinear coordinate system, in which the position of each particle is described by its distance from the spline curve, ρ, the arc length parameter value corresponding to the closest point on the spline curve, s, and the angle, ϕ, enclosed between the vector connecting this point and the particle position and the normal vector of the spline curve (**Figure 6A**). The resulting coordinate system can be viewed as a polar version of the Frenet-Serret frame, providing a generalisation of the canonical cylindrical coordinate system, where the z-coordinate is replaced by the length along the arc of the centre line.

Unfortunately, there is no analytical expression relating the Cartesian coordinates of a particle to the curvilinear centre line coordinates and this mapping has to be accomplished numerically by finding the minimal distance between the position of a particle and the spline curve. Given the position of the j-th particle in Cartesian coordinates, $\mathbf{Q}_j = (x_j, y_j, z_j)^\intercal$, the corresponding spline coordinates are found from

$$\rho_j = \min_s ||\mathbf{S}(s) - \mathbf{Q}_j||$$

$$s_j = \arg\min_s ||\mathbf{S}(s) - \mathbf{Q}_j|| - \arg\min_s ||\mathbf{S}(s) - \mathbf{P}_0||$$

$$\phi_j = \frac{\left(\mathbf{Q}_j - \mathbf{S}(s_j)\right) \cdot \mathbf{N}}{||\mathbf{Q}_j - \mathbf{S}(s_j)||\ ||\mathbf{N}||}$$

where the minimisation problem is solved using the method of (Brent, 1971). Note that in the definition of the s-coordinate the projection of the initial probe position is subtracted. This ensures that the offset of the s-coordinate is consistent across multiple trajectory frames by setting it to zero for the initial probe position (which typically corresponds to the centre of geometry of the protein).

The s-coordinate can now be used as a one-dimensional collective variable along which to determine the density of water and the corresponding free energy profile. The probability density $P(s, \rho, \phi)$ of finding a water molecule at the given centre line coordinates is given by the Boltzmann relation,

$$P(s, \rho, \phi) = \frac{1}{Z} \exp\big(- \beta G(s, \rho, \phi)\big)$$

where $G(s, \rho, \phi)$ is the Gibbs free energy, Z is the partition function, and $\beta^{-1} = k_{\mathrm{B}} T$ with k_{B} being the Boltzmann constant and T the simulation temperature. Assuming that inside a relatively narrow channel pore the free energy of a water particle is largely independent of its lateral position, i.e. $G(s, \rho,) = G(s)$, it follows that

$$P(s) = \int_0^{2\pi} \int_0^{R(s)} P(s, \rho, \phi) \rho d\rho d\phi = \frac{2\pi R(s)^2}{Z} \exp\big(- \beta G(s)\big)$$

where it is assumed that the cross-sectional area of the pore is approximately circular with a local radius $R(s)$. Inverting the above relationship yields an expression for the free energy of the water molecule,

$$G(s) = -\frac{1}{\beta} \left(\ln \frac{P(s)}{\pi R(s)^2} + \ln Z \right)$$

into which the partition function enters only as an additive constant. This constant is determined by the convention that the free energy in the bulk outside the protein should be zero, i.e. $G(s) \to 0$ as $s \to \pm\infty$.

The probability density appearing in the above expression can be estimated directly from the MD trajectory by means of kernel density estimation:

$$P_t(s) = \frac{1}{N} \sum_{i=1}^{N} K_h\big(s, s_i(t)\big)$$

Here $s_i(t)$ is the position of the i-th water molecule at time t, N is the overall number of water molecules, and $P_t(s)$ denotes the probability density estimated at time t. A Gaussian kernel of the form

$$K_h(s, s_i) = \frac{1}{\sqrt{2\pi} h} \exp\left(-\frac{s - s_i}{2h^2}\right)$$

is employed here. A conceptual representation of kernel density estimation is shown in **Figure 6B**.

The bandwidth h is a free parameter in kernel density estimation that fulfills a function similar to the bin width in histograms and care must be taken when selecting a value for this parameter as the resulting density profile is sensitive to it. A value for this parameter of 0.24 nm has been suggested for identifying liquid interfaces in three dimensions (Willard

and Chandler, 2010), whereas a value of 0.1 nm has been used in a study of pore de-wetting in the bacterial GLIC channel (Zhu and Hummer, 2012). It is also possible to estimate an appropriate bandwidth directly from the data (Sheather and Jones, 1991). This method minimises the asymptotic mean integrated square error (AMISE) of the density estimate.

Figure 7 uses the example of the 5-HT_3 receptor to illustrate how the water density and free energy profiles depend on the bandwidth parameter. For very small values of h, the water density estimate is very noisy and in low density regions it drops down to zero, so that the corresponding free energy at these points is singular. Conversely, if a very large bandwidth is chosen, the estimate of $P(s)$ is near uniform throughout the pore and the water density estimate exhibits over-densities at points where the pore radius is particularly small. In this case the free energy profile exhibits no clear barriers or wells.

The AMISE-optimal bandwidth, which in this case is 0.34 nm, finds a balance between these two extremes. The overall density profile is free of noise and the lowered water density in the region of the hydrophobic gate is readily apparent. However, there are still some regions where the water density is unphysically large. By comparison, a profile estimated using a bandwidth of 0.14 nm (the approximate radius of a water molecule) yields a water density that never much exceeds the experimental bulk density of water while still showing little sign of noise. Using a bandwidth fixed at the size of a water molecule also permits a more direct comparison between different channel structures, while the AMISE-optimal bandwidth may differ across channel structures of different overall pore volume. It is therefore recommended to use a default value of 0.14 nm.

Hydrophobicity Profiles

In addition to calculating the channel radius and water density profiles inside the pore, it is also necessary to quantify the hydrophobicity along the conduction pathway. For cylindrical nanopores, it has been found that the free energy difference between hydrated and dehydrated pore states is given by

$$\Delta\Omega = \Omega_{\text{vapour}} - \Omega_{\text{liquid}} = 2\pi R\gamma_{\text{lv}}\big(R + L\cos(\theta_e)\big)$$

where R and L are the radius and length of the nanopore, γ_{lv} is the liquid-vapour surface tension of water, and θ_e is the contact angle from the Young-Dupré equation (Beckstein and Sansom, 2004). Determining a hydrophobicity profile for a channel therefore requires knowledge about the amino acid residues facing a given section of the pore and their associated contact angle. The individual hydrophobicity values at the positions of the amino acids can then be combined into a continuous hydrophobicity profile by means of kernel smoothing.

CHAP can automatically identify pore-facing residues, i.e. residues exposing their amino acid side chains to the channel pore, via a two-step procedure. First, all residues whose centre of geometry lies within a given cutoff, R_{thres}, from the pore surface are determined. To accomplish this, the position of the centre of geometry of each residue is mapped from Cartesian coordinates to centre line spline coordinates so that each residue is associated with a distance from the centre line, ρ_i^{cog}, and a distance along the centre line, s_i^{cog}, where the index runs over all protein residues. If

$$\rho_i^{\mathrm{cog}} > R(s_i^{\mathrm{cog}}) + R_{\mathrm{thres}}$$

the i-th residue will be considered pore-lining, where $R(s_i^{\mathrm{cog}})$ is the pore radius at the position of the residue. A typical threshold value for an α-helical pore is 0.75 nm.

Second, for each pore-lining residue, the position of the α-carbon is also mapped to pathway coordinates ρ_i^{α} and s_i^{α}. If the distance from the α-carbon to the pore centre line is larger than the distance between the centre of geometry of the overall residue and the pore centre line, i.e. if

$$\rho_i^{\alpha} > \rho_i^{\mathrm{cog}}$$

then the i-th residue will be considered pore-facing. The property of being pore-lining and pore-facing is encoded in two binary indicator variables, whose values are written to the temperature factor and occupancy fields of a PDB file of the input structure to allow easy visualisation of these residues. If CHAP is run with a trajectory as input, the above analysis is performed repeatedly (once for each frame) and the values are written to the PDB file correspond to the time average of the indicator variables. **Figure 8A** illustrates the distinction between pore-lining and pore-facing residues in the 5-HT_3 receptor.

To the best of our knowledge no experimental data quantifying the contact angle at the interface between water and specific amino acids are available. However, various amino acid hydrophobicity scales have been proposed in the literature. An early hydropathy scale based on a collection of experimental observations has been suggested by (Kyte and Doolittle, 1982), whilst another has been based on measurements of the stability of α -helices (Monera et al., 1995). Other scales have used the free energy of partitioning of oligopeptides between water and a lipid environment (Wimley et al., 1996; Wimley and White, 1996), while some report free energies of transferring amino acids from water into a lipid bilayer (Moon and Fleming, 2011). A biological scale has also been proposed based on the recognition of α-helices by the endoplasmic reticulum translocon (Hessa et al., 2005), whilst another is derived from MD simulations of a water nanodroplet on a planar layer of amino acids (Zhu et al., 2016).

For comparability, the hydrophobicity scales listed above have been shifted and scaled to the interval ranging from -1 for the most hydrophilic to +1 for the most hydrophobic residues while preserving their natural zero (**Figure S4**). Note this contrasts with the physical definition of hydrophobicity in terms of $\cos\theta_e$ and instead follows the biochemical convention of associating a larger value with a more hydrophobic residue. CHAP uses the normalised hydrophobicity scales to associate a hydrophobicity value, H_i, with each residue position s_i^{cog}. A continuous hydrophobicity profile is then calculated from these discrete hydrophobicity values using a Nadaraya-Watson kernel-weighted average. Mathematically this is expressed as

$$H(s) = \frac{\sum_i K_\lambda(s,s_i) H_i}{\sum_i K_\lambda(s,s_i)}$$

where

$$K_\lambda(s, s_i) = \exp\left(-\frac{|s-s_i|^2}{2\lambda^2}\right)$$

is a Gaussian kernel. The smoothing span λ is chosen to reflect the finite extension of an amino acid residue. Since the pitch of an α-helix is approximately 0.15 nm per residue and on average every third residue is pore-facing, a value of ~0.225 nm (for a width of the Gaussian kernel of 0.45 nm) can be considered reasonable. Note that far away from any residues (e.g. in the bulk water region outside the channel pore) $H(s)$ will decay exponentially to zero.

In **Figure 8D** this is illustrated on the 5-HT_3 receptor using the Wimley and White hydrophobicity scale. As can be seen, there is a considerable increase in hydrophobicity in the region of the transmembrane domain containing the hydrophobic gate. This section of the profile is shown in **Figure 8C** mapped onto a 3D representation of the pore together with the M2 helices of the 5-HT_3 receptor, illustrating how the hydrophobicity of the pore-facing residues influences the hydrophobicity profile.

Conclusion

In this study we present a versatile computational tool for the functional annotation of ion channel structures: the Channel Annotation Package CHAP. This tool implements an MD-based methodology for predicting the conductive state of ion channel structures that goes beyond a measurement of the pore radius to include profiles of pore hydrophobicity as well as water density and free energy. Crucially, this enables CHAP to identify hydrophobic gates which cannot be detected using existing methodologies.

The analysis performed by CHAP is computationally inexpensive compared to more complex simulation techniques such as computational electrophysiology (Kutzner et al., 2011) and is therefore highly suitable for the high-throughput analysis of the rapidly growing number of ion channel structures. The utility of this approach is demonstrated in a recent study (Rao et al., 2018) employing CHAP to investigate the prevalence of hydrophobic gates in ~200 different ion channel structures.

CHAP is made available as free and open-source software. Copies of the source code can be obtained from www.channotation.org and include scripts for visualising results in the molecular graphics programmes VMD (Humphrey et al., 1996) and PyMOL (DeLano, 2002).

Acknowledgements

This work was funded by grants from the BBSRC, the EPSRC, and the Wellcome Trust.
GK holds a studentship funded by the EPSRC.

Appendix A: Molecular Dynamics Simulation Protocol

Equilibrium molecular dynamics simulations of the 5-HT_3 receptor were performed using GROMACS 2016. Starting structures were generated from the original crystal structure (PDB ID: 4PIR) with missing atoms being added by the WHAT IF tool (Vriend, 1990). This starting structure was embedded in a 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) bilayer using an established serial multiscale protocol (Stansfeld and Sansom, 2011). The OPLS all-atom protein force field was used with united-atom lipids (Jorgensen et al., 1996; Tieleman et al., 2006) and the TIP4P model of water (Jorgensen et al., 1983). Long-range electrostatic interactions are treated using the particle mesh Ewald (PME) method (Vriend, 1990; Darden et al., 1993) employing a short range cutoff of 1 nm and a Fourier spacing of 0.12 nm.

To permit subsequent determination of the Gibbs free energy from the water molecule distribution, the system is simulated in the isothermal-isobaric (NPT) ensemble at a temperature of 310 K and a pressure of 1 bar. A v-rescale thermostat (Bussi et al., 2007) with a coupling constant of 0.1 ps is used for temperature control and a pressure is maintained semi-isotropically using the method of (Parrinello and Rahman, 1981) with a coupling constant of 1 ps. The canonical leapfrog method with a time step of 2 fs is used to integrate the system for 30 ns with bonds constrained through the linear constraints solver algorithm LINCS (Hess et al., 1997). Additionally, the Cα atoms of all protein residues are placed under a harmonic restraint with a force constant of 1000 kJ/mol/nm^2 to ensure that the configuration of the simulated channel does not deviate from the experimentally determined structure.

Appendix B: B-Spline Curves

Formally, a spline curve of degree p along some parameter s can be written as the linear combination

$$\tilde{\mathbf{S}}_p(\tilde{s}) = \sum_{k=1}^{N} \mathbf{c}_k B_{k,p}(\tilde{s})$$

where

$$B_{k,p}(\tilde{s}) = \frac{\tilde{s}-t_k}{t_{k+p}-t_k} B_{k,p-1}(\tilde{s}) + \frac{t_{k+p+1}-\tilde{s}}{t_{k+p+1}-t_{k+1}} B_{k+1,p-1}(\tilde{s})$$

with $B_{k,0}(\tilde{s}) = 1$ if $t_k \leq \tilde{s} \leq t_{k+1}$ and $B_{k,0}(\tilde{s}) = 0$ otherwise are the recursively defined B-spline basis functions. Here $\{\mathbf{c}_k\}_{k=1}^{N}$ denotes the spline curve control points and $\{t_k\}_{k=1}^{N+p+1}$ is referred to as knot vector. Note that a zero division in the above equations is treated as $0/0 = 0$.

In order to find the spline curve that interpolates the centre line points $\mathbf{P}_i$ from the pore finding algorithm, one demands

$$\mathbf{S}_p(\tilde{s}_i) = \mathbf{P}_i \quad (*)$$

which requires a parameterisation of the discrete point set. Ideally, this should be the distance between the points along the arc of the spline curve, but unfortunately this distance is not known *a priori*. Instead, the commonly applied chord length approximation is used here, in which points are parameterised by their Euclidean distance according to:

$$\tilde{s}_{i+1} = \tilde{s}_i + ||\mathbf{P}_{i+1} - \mathbf{P}_i||$$

Importantly, it has been also shown that this choice of parameterisation guarantees full approximation order for spline curves up to third degree (Floater and Surazhsky, 2006).

For a cubic spline curve $\mathbf{S}(\tilde{s})$, where the subscript $p = 3$ has been dropped for clarity, Equation (*) together with the knot vector

$$\{t_k\}_{k=1}^{N+p+1} = \{\tilde{s}_1, \tilde{s}_1, \tilde{s}_1, \tilde{s}_1, \tilde{s}_2, \tilde{s}_3, \ldots, \tilde{s}_{N-1}, \tilde{s}_N, \tilde{s}_N, \tilde{s}_N, \tilde{s}_N\}$$

and hermite boundary conditions then leads to a tridiagonal linear system, from which the control points $\mathbf{c}_k$ can be determined efficiently through the Thomas algorithm. The choice of cubic splines is motivated by the fact that cubic splines are twice continuously differentiable so that the curvature of the resulting spline curve is guaranteed to be continuous. At the same time, cubic splines are known to minimise overall curvature, preventing undue variation of the interpolant between the given centre line points.

The resulting spline curve is subsequently re-parameterised by arc length using a three-step procedure proposed by (Wang et al., 2002). In a first step, the arc length in each interval is computed according to

$$l_k = \int_{\tilde{s}_k}^{\tilde{s}_{k+1}} ||\tilde{\mathbf{S}}(\tilde{s})|| d\tilde{s} = \int_{\tilde{s}_k}^{\tilde{s}_{k+1}} \sqrt{\left(\tilde{\mathbf{S}}'(\tilde{s})\Big|_x\right)^2 + \left(\tilde{\mathbf{S}}'(\tilde{s})\Big|_y\right)^2 + \left(\tilde{\mathbf{S}}'(\tilde{s})\Big|_z\right)^2} d\tilde{s}$$

from which the overall length of the spline curve between the two openings of the pore can be calculated as:

$$L = \sum_{k=1}^{N-1} l_k$$

The integral in the expression for l_k is solved numerically through sixth-order accurate Newton-Cotes quadrature (Boole's rule).

The second step comprises calculating the chord length parameter values $\tilde{s}_m$ corresponding to the equidistant arc length parameter values $s_m = m\Delta s = mL/M$ with $m \in [0, M]$. This is equivalent to finding the root of

$$m\Delta s - \left(\sum_{k=1}^{q} l_k + \int_{\tilde{s}_q}^{\tilde{s}_m} ||\tilde{\mathbf{S}}'(\tilde{s})||\ d\tilde{s}\right) = 0$$

where $q \in [1, N]$ is the index of the spline interval for which

$$\sum_{k=0}^{q} l_k \leq m\Delta s < \sum_{k=0}^{q+1} l_k.$$

The root-finding problem is solved through the algorithm of (Alefeld et al., 1995), which is known to be the asymptotically most efficient root-finding method.

In the final step, the interpolation problem

$$\mathbf{S}(s_m) = \tilde{\mathbf{S}}(\tilde{s}_m)$$

is solved to obtain a new spline curve, $\mathbf{S}(\mathrm{s})$, that approximates the original curve, but is parameterised by arc length rather than chord length. This curve is then used to describe the centre line of the channel pore.

Bibliography

Abraham, M.J., T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, and E. Lindahl. 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 1-2:19–25.

Alefeld, G.E., F.A. Potra, and Y. Shi. 1995. Algorithm 748; enclosing zeros of continuous functions. *ACM Trans. Math. Softw.* 21:327–344.

Allen, R., S. Melchionna, and J.-P. Hansen. 2002. Intermittent permeation of cylindrical nanopores by water. *Phys. Rev. Lett.* 89. doi:10.1103/physrevlett.89.175502.

Aryal, P., F. Abd-Wahab, G. Bucci, M.S.P. Sansom, and S.J. Tucker. 2014. A hydrophobic barrier deep within the inner pore of the TWIK-1 K2P potassium channel. *Nat. Commun.* 5:4377.

Aryal, P., M.S.P. Sansom, and S.J. Tucker. 2015. Hydrophobic gating in ion channels. *J. Mol. Biol.* 427:121–130.

Beckstein, O., P.C. Biggin, and M.S.P. Sansom. 2001. A hydrophobic gating mechanism for nanopores. *J. Phys. Chem. B*. 105:12902–12905.

Beckstein, O., and M.S.P. Sansom. 2003. Liquid-vapor oscillations of water in hydrophobic nanopores. *Proc. Natl. Acad. Sci. U. S. A.* 100:7063–7068.

Beckstein, O., and M.S.P. Sansom. 2004. The influence of geometry, surface character, and flexibility on the permeation of ions and water through biological pores. *Phys. Biol.* 1:42–52.

Beckstein, O., and M.S.P. Sansom. 2006. A hydrophobic gate in an ion channel: the closed state of the nicotinic acetylcholine receptor. *Phys. Biol.* 3:147–159.

Brent, R.P. 1971. An algorithm with guaranteed convergence for finding a zero of a function. *Comput. J.* 14:422–425.

Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.

Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

DeLano, W.L. 2002. PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*. 40:82–92.

Floater, M.S., and T. Surazhsky. 2006. Parameterization for curve interpolation. *In* Studies in Computational Mathematics. 39–54.

Hastie, T., R. Tibshirani, and J. Friedman. 2013. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media. 536 pp.

Hessa, T., H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, and G. von Heijne. 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*. 433:377–381.

Hess, B., H. Bekker, H.J.C. Berendsen, and Johannes G E. 1997. LINCS: A linear constraint solver

for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

Hille, B. 1992. Ionic Channels of Excitable Membranes. Sinauer Associates Incorporated. 607 pp.

Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

Jensen, M.O., D.W. Borhani, K. Lindorff-Larsen, P. Maragakis, V. Jogini, M.P. Eastwood, R.O. Dror, and D.E. Shaw. 2010. Principles of conduction and hydrophobic gating in K channels. *Proceedings of the National Academy of Sciences*. 107:5833–5838.

Jia, Z., M. Yazdani, G. Zhang, J. Cui, and J. Chen. 2018. Hydrophobic gating in BK channels. *Nat. Commun.* 9:3408.

Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225–11236.

Kutzner, C., H. Grubmüller, B.L. de Groot, and U. Zachariae. 2011. Computational electrophysiology: the molecular dynamics of ion channel permeation and selectivity in atomistic detail. *Biophys. J.* 101:809–817.

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.

Monera, O.D., T.J. Sereda, N.E. Zhou, C.M. Kay, and R.S. Hodges. 1995. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *J. Pept. Sci.* 1:319–329.

Moon, C.P., and K.G. Fleming. 2011. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc. Natl. Acad. Sci. U. S. A.* 108:10174–10177.

Neale, C., N. Chakrabarti, P. Pomorski, E.F. Pai, and R. Pomès. 2015. Hydrophobic gating of ion permeation in magnesium channel CorA. *PLoS Comput. Biol.* 11:e1004303.

Nelder, J.A., and R. Mead. 1965. A Simplex Method for Function Minimization. *Comput. J.* 7:308–313.

Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.

Rao, S., G. Klesse, P.J. Stansfeld, S.J. Tucker, and M.S.P. Sansom. 2017. A BEST example of channel structure annotation by molecular simulation. *Channels* . 11:347–353.

Rao, S., G. Klesse, P.J. Stansfeld, S.J. Tucker, and M.S.P. Sansom. 2018. Identification of hydrophobic gates by simulation-enabled analysis of the ion channel structural proteome. doi:10.1101/498386.

Sheather, S.J., and M.C. Jones. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Series B Stat. Methodol.* 53:683–690.

Smart, O.S., J.G. Neduvelil, X. Wang, B.A. Wallace, and M.S.P. Sansom. 1996. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.*

14:354–360.

Spronk, S.A., D.E. Elmore, and D.A. Dougherty. 2006. Voltage-dependent hydration and conduction properties of the hydrophobic pore of the mechanosensitive channel of small conductance. *Biophys. J.* 90:3555–3569.

Stansfeld, P.J., and M.S.P. Sansom. 2011. From coarse grained to atomistic: a serial multiscale approach to membrane protein simulations. *J. Chem. Theory Comput.* 7:1157–1166.

Tieleman, D.P., J.L. Maccallum, W.L. Ash, C. Kandt, Z. Xu, and L. Monticelli. 2006. Membrane protein simulations with a united-atom lipid and all-atom protein model: lipid-protein interactions, side chain transfer free energies and model proteins. *J. Phys. Condens. Matter*. 18:S1221–34.

Trick, J.L., P. Aryal, S.J. Tucker, and M.S.P. Sansom. 2015. Molecular simulation studies of hydrophobic gating in nanopores and ion channels. *Biochem. Soc. Trans.* 43:146–150.

Trick, J.L., S. Chelvaniththilan, G. Klesse, P. Aryal, E.J. Wallace, S.J. Tucker, and M.S.P. Sansom. 2016. Functional annotation of ion channel structures by molecular simulation. *Structure*. 24:2207–2216.

Vriend, G. 1990. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* 8:52–56.

Wang, H., J. Kearney, and K. Atkinson. 2002. Wang, H., Kearney, J., & Atkinson, K. (2002, June). Arc-length parameterized spline curves for real-time simulation. *In* Proc. 5th International Conference on Curves and Surfaces. 387–396.

Willard, A.P., and D. Chandler. 2010. Instantaneous liquid interfaces. *J. Phys. Chem. B*. 114:1954–1958.

Wimley, W.C., T.P. Creamer, and S.H. White. 1996. Solvation energies of amino acid side chains and backbone in a family of host−guest pentapeptides. *Biochemistry*. 35:5109–5124.

Wimley, W.C., and S.H. White. 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* 3:842–848.

Yuan, S., S. Filipek, and H. Vogel. 2016. A gating mechanism of the serotonin 5-HT_3 receptor. *Structure*. 24:816–825.

Zhu, C., Y. Gao, H. Li, S. Meng, L. Li, J.S. Francisco, and X.C. Zeng. 2016. Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network. *Proc. Natl. Acad. Sci. U. S. A.* 113:12946–12951.

Zhu, F., and G. Hummer. 2012. Drying transition in the hydrophobic gate of the GLIC channel blocks ion conduction. *Biophys. J.* 103:219–227.

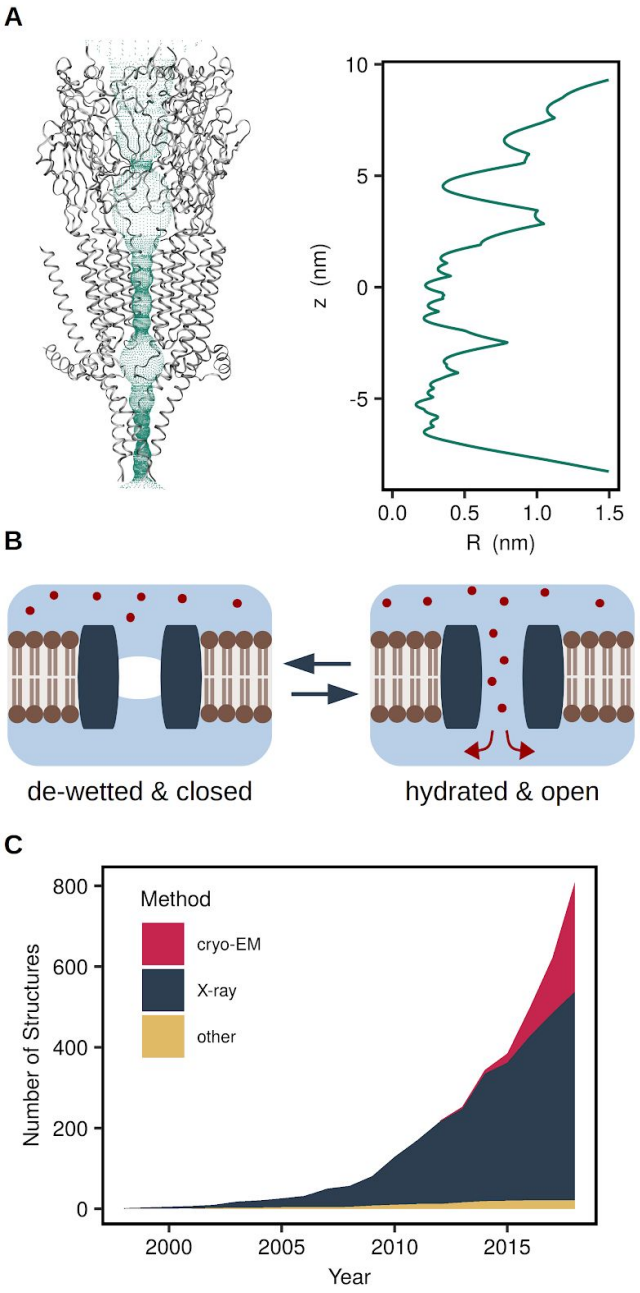

Figure 1: Hydrophobic gates pose a challenge to the functional annotation of ion channel structures. (A) HOLE profile of the entire PDB structure of a 5-HT_3 receptor (PDB ID: 4PIR). (B) Hydrophobic gating: A non-conducting state is characterised by the de-wetting of a part of the permeation pathway lined by hydrophobic residues. Ions are prevented from crossing this region by the free energy barrier associated with their desolvation even if the pore is sufficiently large to accommodate them. (C) The number of ion channel structures deposited in the protein data bank is growing exponentially (data retrieved on 5 December 2018).

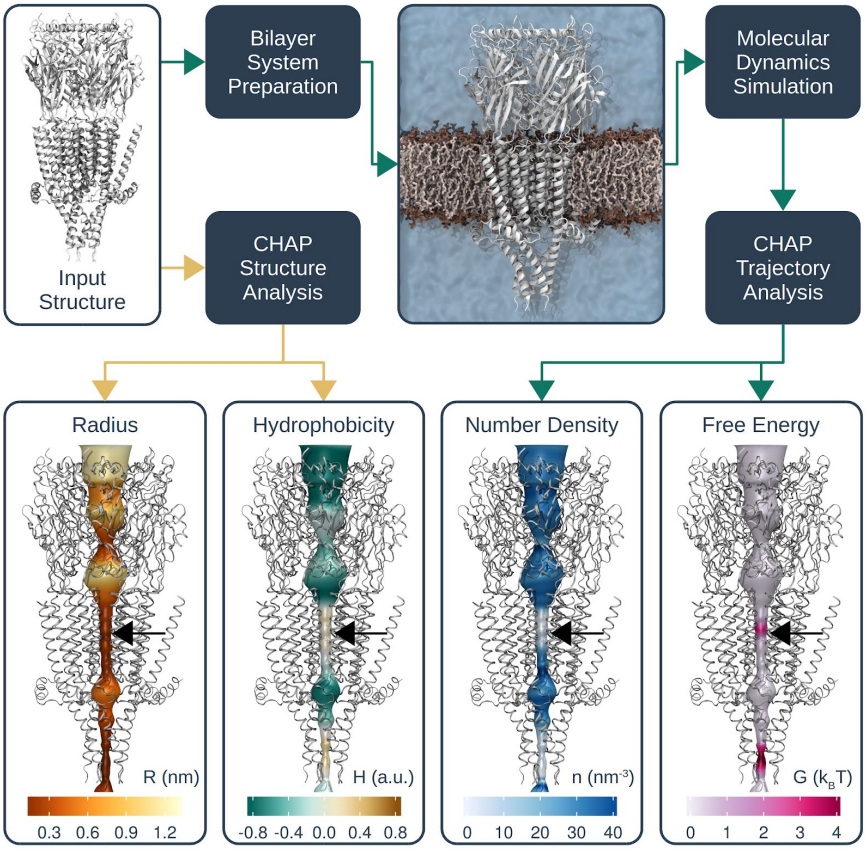

Figure 2: Overview of workflow for functional annotation of ion channels with CHAP. Given a simple protein structure as input, CHAP can calculate pore radius and hydrophobicity profiles. Additionally, the protein can be simulated in its native lipid bilayer environment in order to assess the behaviour of water inside the channel pore. From the resulting simulation trajectory data, CHAP can derive water density and free energy profiles that aid the identification of hydrophobic gates. Yellow arrows indicate steps of the pipeline that can be carried out on structural data alone while cyan arrows indicate those steps that are based on subsequent molecular dynamics simulation. The black arrows indicate the location of the 9' ring of leucine residues.

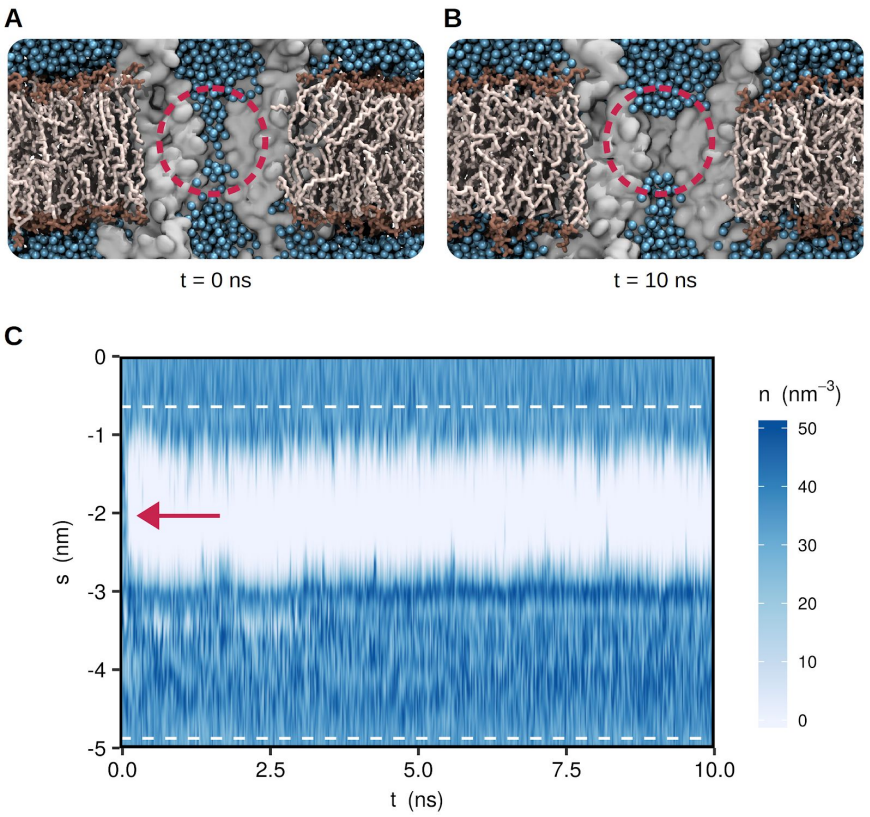

Figure 3: Hydrophobic gates can be identified through equilibrium MD simulation. (A) Cross-section of the starting structure for MD simulation of a 5-HT_3 receptor (PDB ID: 4PIR) embedded in a POPC lipid bilayer. The protein is represented by a grey surface with two of five subunits omitted for visual clarity. Only the transmembrane (TM) domain is shown. Water oxygens are represented by blue spheres and lipid molecules are shown in liquorice representation. Note that the entire channel pore is initially filled with water. (B) Snapshot from the MD trajectory clearly exhibiting a de-wetted region (highlighted with red circle) in the TM domain of the channel. Note that the channel pore is not physically occluded. (C) Water density profile in the receptor pore over time. An approximately 2 nm long region of the TM domain indicated by a red arrow de-wets at the beginning of the simulation and remains devoid of water throughout.

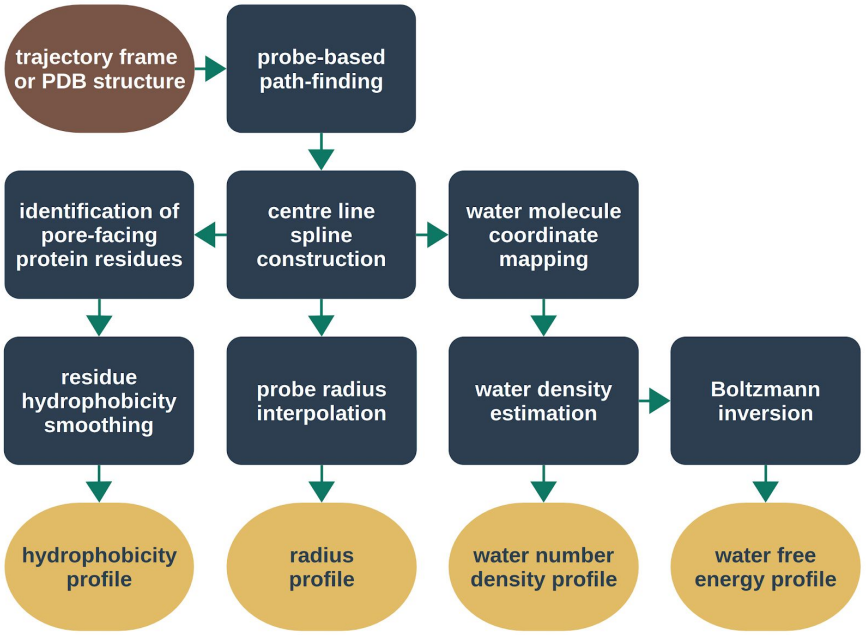

Figure 4: Workflow diagram of data processing in CHAP. For an individual PDB structure this procedure is carried out once. For trajectory data, the above procedure is carried out individually for each frame, i.e. for each structural configuration at a sampled time point. The resulting hydrophobicity, radius, density, and free energy profiles are averaged over all time steps and confidence bands are estimated from the standard deviation over time.

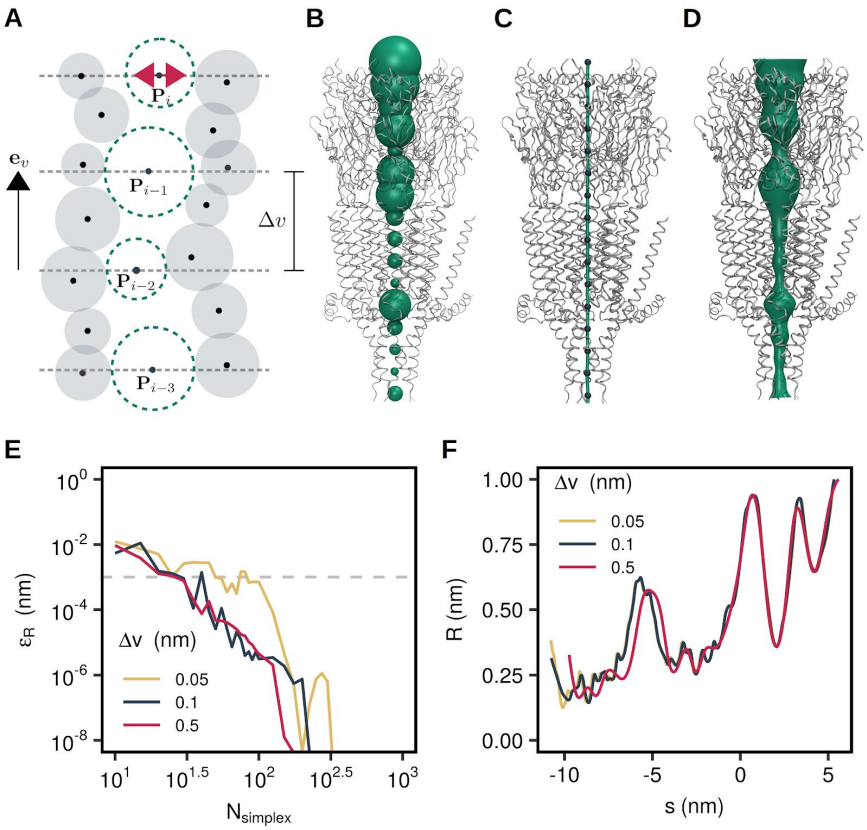

Figure 5: Pore geometry calculations in CHAP. (A) Schematic representation of probe-based pathway finding. The position of a spherical probe is optimised in subsequent parallel planes to maximise its radius without overlapping the van der Waals sphere of any ion channel atom. (B) Configuration space view of probe positions (green spheres) in the permeation pathway of a 5-HT_3 receptor structure (PDB ID: 4PIR). For visual clarity a large probe step of $\Delta v = 1$ nm was used here. (C) The probe centre positions (blue spheres) are interpolated with cubic B-splines to yield a continuous spatial curve (green tube) representing the pore centre line. (D) A surface representation of the channel pore is generated by extruding a circular cross-section along the centre line spline curve. (E) Convergence of the Nelder-Mead algorithm carried out to optimise the probe position in each parallel plane. The radius error falls below a threshold of 0.001 nm within 100 optimisation steps irrespective of the value of the probe step. The reference radius profile is calculated with a very large number of 2000 Nelder-Mead iterations. (F) Influence of probe step on radius profile resolution. While a probe step of 0.5 nm only captures the general feature of the radius profile, decreasing the probe step from 0.1 nm to 0.05 nm leads only to minimal changes.

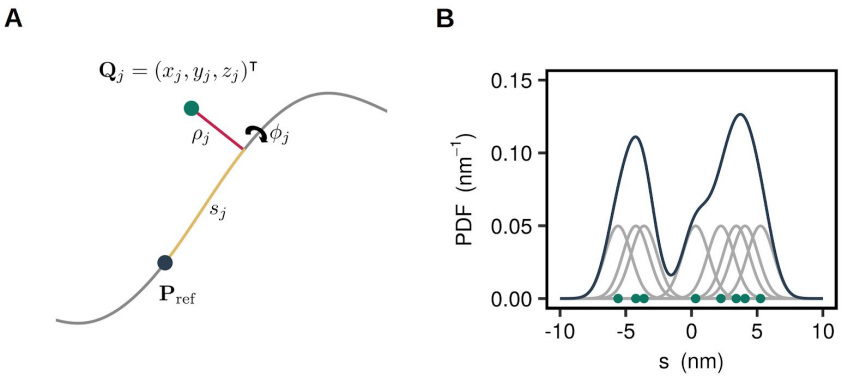

Figure 6: Determination of water density. (A) Cartesian coordinates of a water molecule (green) are converted into a curvilinear coordinate system by projecting its position onto the channel centre line (grey). The new coordinates are then expressed in terms of the distance along the centre line (yellow) between the projection base point and a reference point (dark blue), the distance from the water molecule to the centre line (red), and an angle around the centre line. (B) Conceptual representation of kernel density estimation: Each discrete water molecule position is associated with a Gaussian probability distribution and the sum of all Gaussians yields the probability density function of water along the channel centre line.

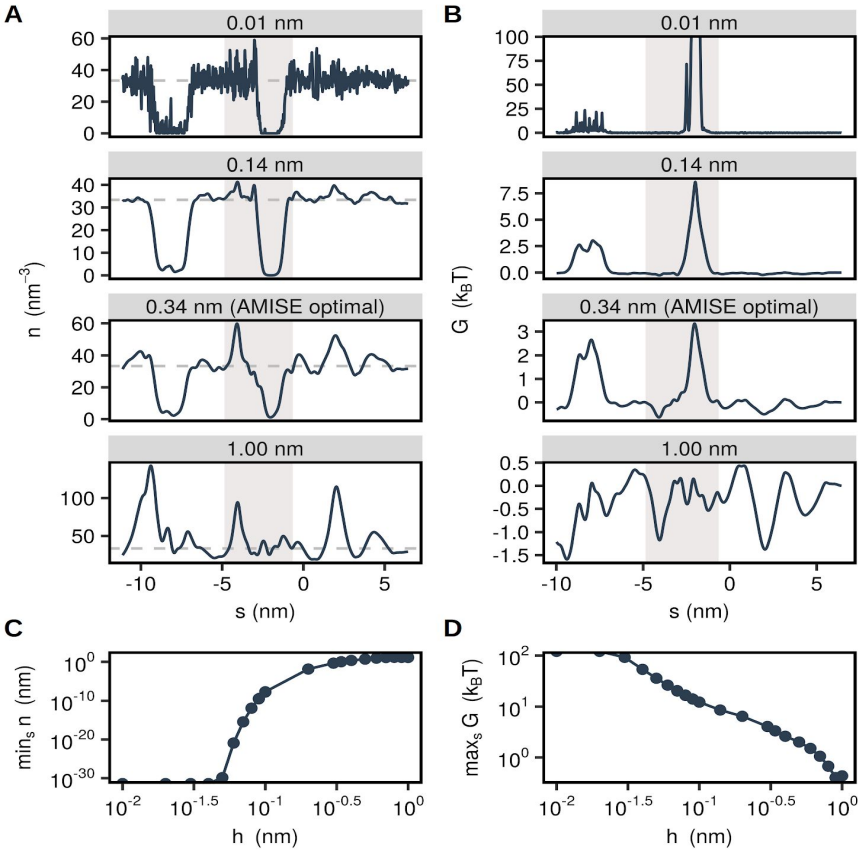

Figure 7: Influence of bandwidth parameter. (A) Water number density along the permeation pathway of the 5-HT_3 receptor for various values of the kernel bandwidth. A value of 0.14 nm corresponds to the approximate size of a water molecule and a value of 0.34 nm corresponds to the statistically optimal bandwidth. (B) Corresponding free energy profiles calculated via Boltzmann inversion of the density profiles. The grey shaded area in (A) and (B) represents the transmembrane domain. (C) Minimum of number density profile in dependence of bandwidth. (D) Maximum of free energy profile in dependence of bandwidth.

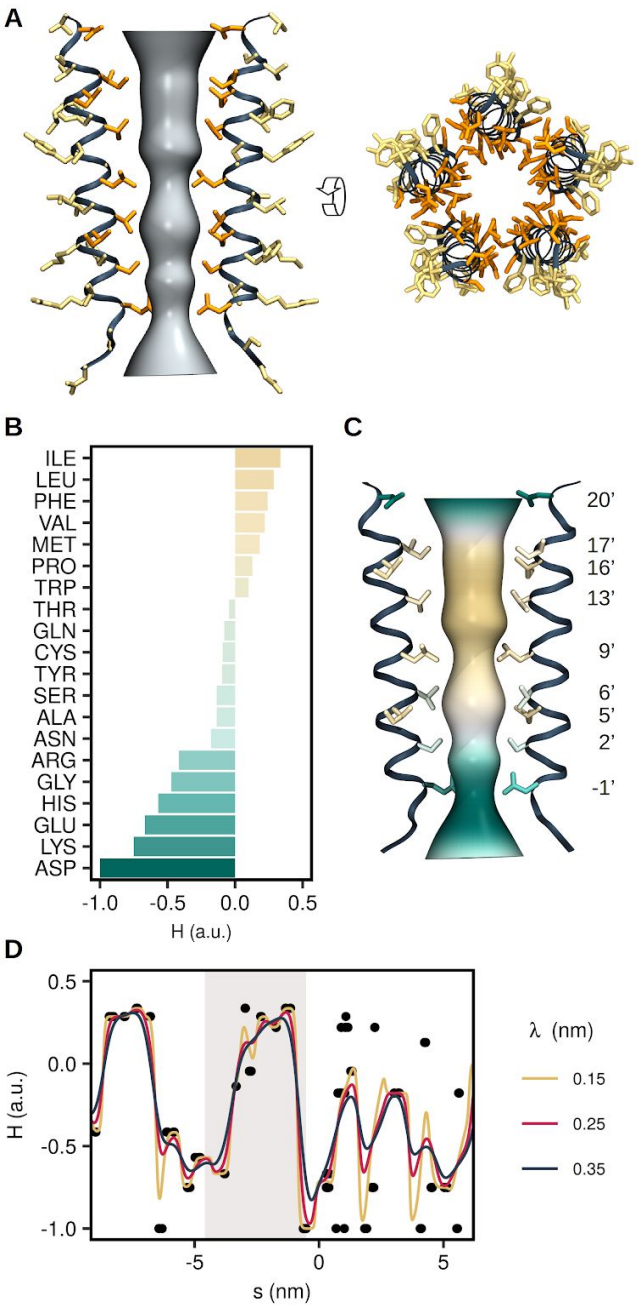

Figure 8: Hydrophobicity profile of pore-lining residues. (A) Classification of residues in the M2 helix bundle of the 5-HT_3 receptor into pore-lining (orange) and non-pore-lining (yellow). (B) Amino acid hydrophobicities according to the Wimley-White scale. In CHAP the original hydrophobicities are rescaled from -1 (very hydrophilic) to +1 (very hydrophobic) to facilitate comparison between different scales. (C) Permeation pathway coloured according to the Wimley-White hydrophobicity scale. (D) Hydrophobicity profile of the entire pore. Black circles represent the position and hydrophobicity of the pore-facing residues (averaged over all five subunits). Lines represent the kernel-smoothed continuous hydrophobicity profile underlying the surface colouring in panel C. Different values of the smoothing span, λ, influence the magnitude of fluctuations, but the overall profile remains unaffected. The shaded area corresponds to the TM region shown in panels A and C.

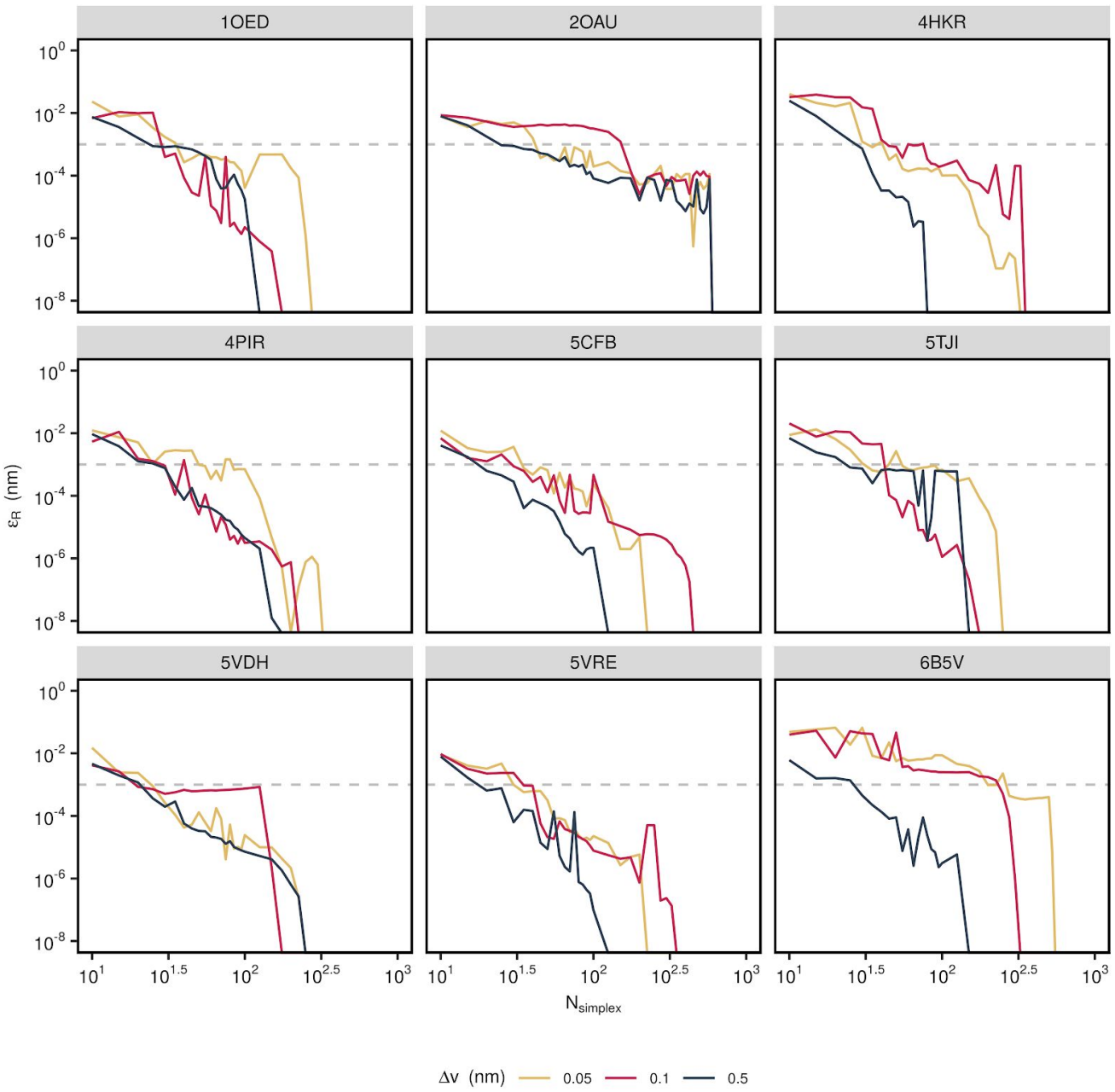

Figure S1: Convergence of the Nelder-Mead algorithm. The error of the radius profile calculated for nine different channel structures using three different values for the probe step is shown. Convergence to within 0.001 nm is typically reached within 100 iterations.

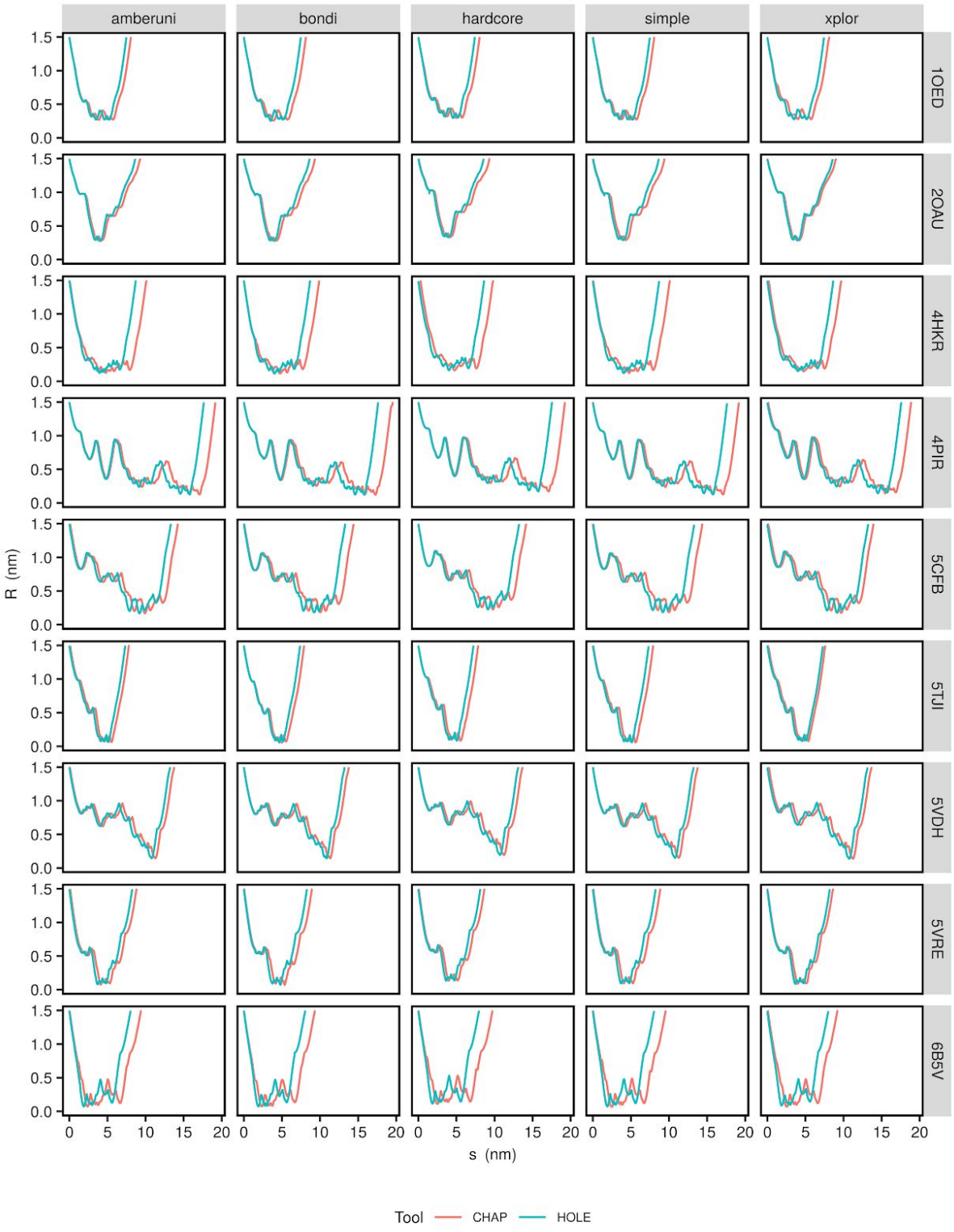

Figure S2: Comparison of the radius profiles calculated by CHAP and HOLE. Radius profiles were calculated for nine different ion channel structures using the five van der Waals radius datasets distributed with HOLE. Note that the CHAP profile is slightly stretched with respect to the HOLE profile because CHAP uses the position along the (potentially curved) centre line as its reaction coordinate whereas HOLE uses the position along the transmembrane axis.

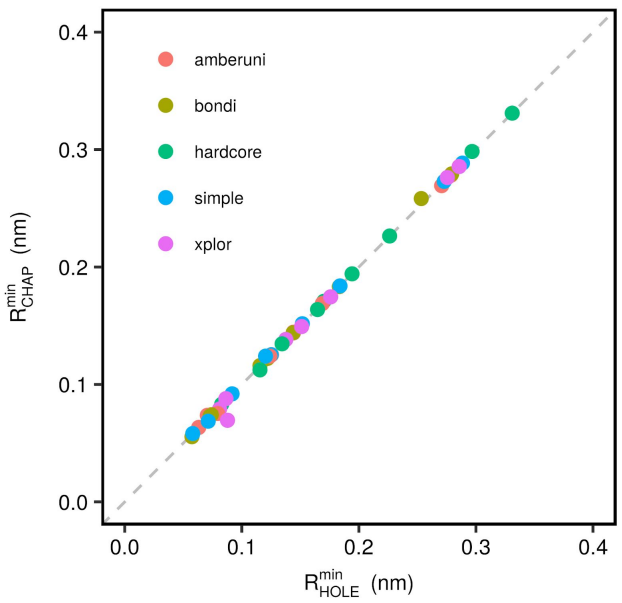

Figure S3: Comparison of the radius at the narrowest constriction between CHAP and HOLE. Calculated for nine representative ion channel structures using the five different van der Waals radius datasets available in HOLE. In all cases the radii agree to within 0.01 nm.

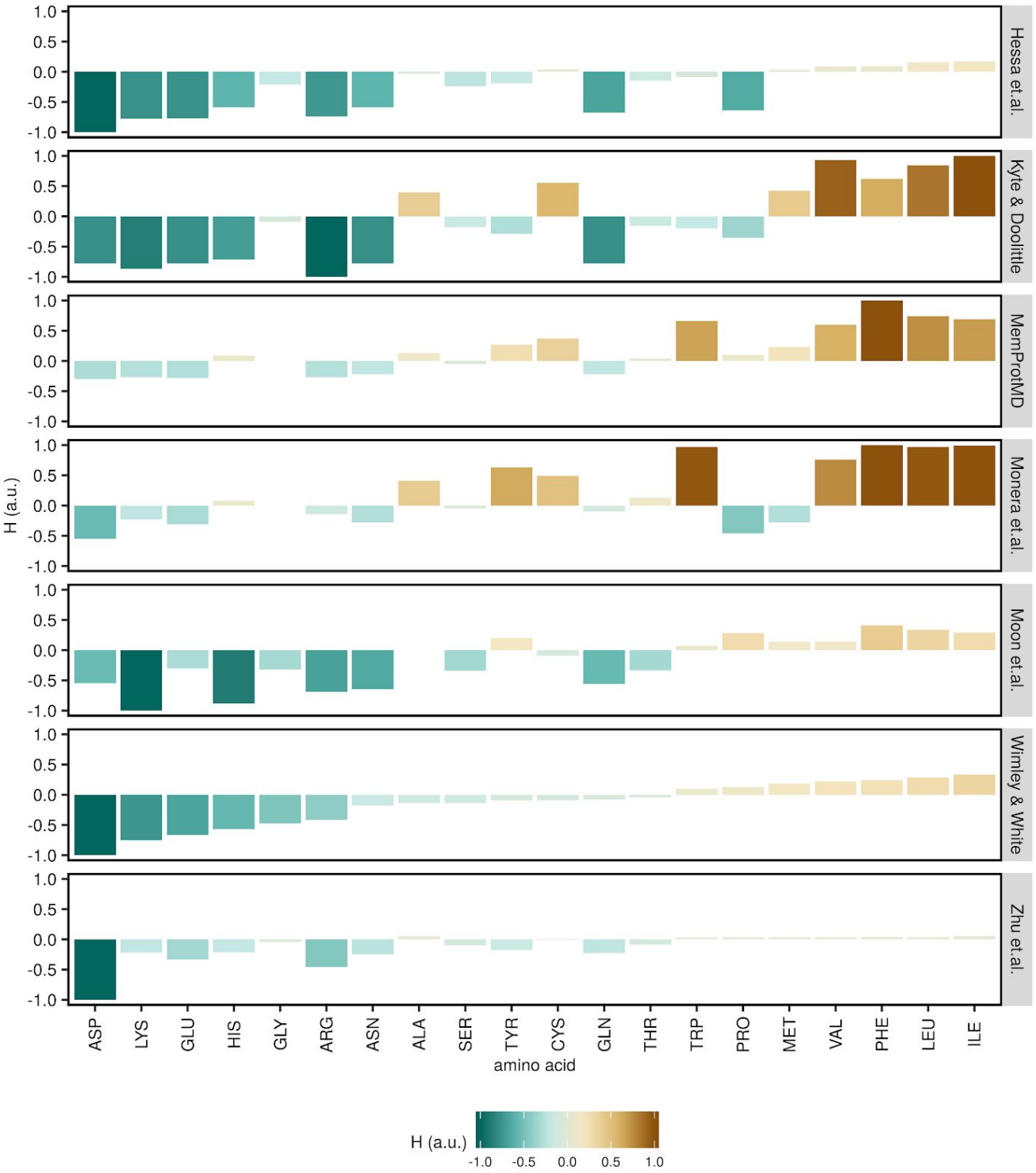

Figure S4: Comparison of amino acid hydrophobicity scales available in CHAP. The amino acids are listed in order of increasing hydrophobicity according to the Wimley-White scale, which is the default scale used by CHAP.