CaviTracer: robust automation of channel identification

0. Overview: the test corpus, before and after

21 structures from the MOLE2 and CAVER reference sets plus a porin and two protonated dehalogenases, all seeded from a hand-placed start_point.

case	chl main	chl head	bn main	bn head	t main	t head	speedup
1bl8	15	10	1.72	1.73	2.62	2.75	0.95x
1cqw	3	1	1.42	1.51	1.91	1.64	1.16x
1grm	4	2	1.43	1.34	0.37	0.47	0.79x
1m56	47	9	0.88	2.11	62.41	16.89	3.70x
1mj5_prot	6	4	1.30	1.31	5.11	5.39	0.95x
1mqf	13	4	1.95	1.99	3.47	3.36	1.03x
1mxt	9	4	1.59	1.59	8.09	7.93	1.02x
1s5l	28	19	1.17	1.18	46.18	43.26	1.07x
1su4	3	1	1.23	1.24	9.01	6.10	1.48x
1thg	15	1	1.19	1.20	3.14	2.65	1.18x
1tqn	21	3	2.16	2.22	4.19	3.42	1.23x
2ace	6	1	2.00	1.99	2.86	2.81	1.02x
2bg9	1	46	1.98	7.64	14.65	14.92	0.98x
2cqt	8	2	2.20	2.48	12.26	8.20	1.50x
2fgq	1	2	0.83	2.20	2.04	1.90	1.07x
2jbv	13	5	1.14	1.34	7.94	6.43	1.23x
2oar	40	27	1.82	1.85	5.41	6.55	0.83x
2oj3	6	2	1.22	1.23	1.78	1.96	0.91x
3cap	27	1	1.52	1.55	5.79	4.53	1.28x
3eyx	11	2	1.75	1.78	2.48	2.32	1.07x
dbja_prot	1	1	1.43	1.79	4.76	5.06	0.94x
total					206.47	148.54	1.39x

- 18/21 cases report fewer channels
- 2fgq (porin) and 2bg9 (nicotinic receptor) are find more channels as the recalibrated surf_radius spearates wide lumens from the surface, enabling channel identification
- overall 1.39x faster which is a side effect: less work is spent on routes that are then merged away
- 66 channels were removed by larger inner radius and 65 by improved merging of dupplicates

1. One channel per way out

The single largest source of redundant output was that a surface opening was identified with the tetrahedron at its centre. Mouths form a layer one tetrahedron thick, so a route could walk around a mouth through the shell behind it and surface a few Angstrom further on, and the two results were then compared as if they were different exits. Three changes follow from replacing that identification:

- *an opening is a mouth's inscribed ball*, not a tetrahedron. A search stops at the first opening it enters; the survivors leave by the cheapest mouth of that same opening, and the reported route is the seed's own path to it rather than two searches spliced at the arrival (38d804bc);
- *two mouths are compared symmetrically* - The old test asked whether one exit fell inside the other's mouth sphere; sibling exits sit about a mouth radius apart, so nearly coincident openings were called distinct and the corridor test never ran (fd7f98fb);

- *duplicates are merged on how far their surfaces diverge*, in Angstrom, scaled by the local clearance (surface distnaces), - which collapsed the similarity and route_tolerance pair into a single route_divergence (098a80a1, c34b94a5);
- *when two routes do merge, the better ending is reported* - a wider mouth or a wider bottleneck, with mouth not narrowing the bottleneck (3d235565, 95dabcd6).

mouths as spheres (fd7f98fb)			end at first opening (38d804bc)			route_divergence (c34b94a5)	
case	bef	aft	case	bef	aft	case	bef aft
1bl8	14	10	1mqf	6	4	1su4	2 1
1mqf	8	6	1tqn	4	3	2ace	2 1
1mxt	6	4	2ace	3	2		
1su4	3	1	2fgq	3	2		
2ace	4	3	1su4	1	2		
2cqt	4	3					
2fgq	4	3					
2oj3	3	2					
1tqn	5	4					
16-case total	64	48	16-case total	48	44	16-case total	43 41

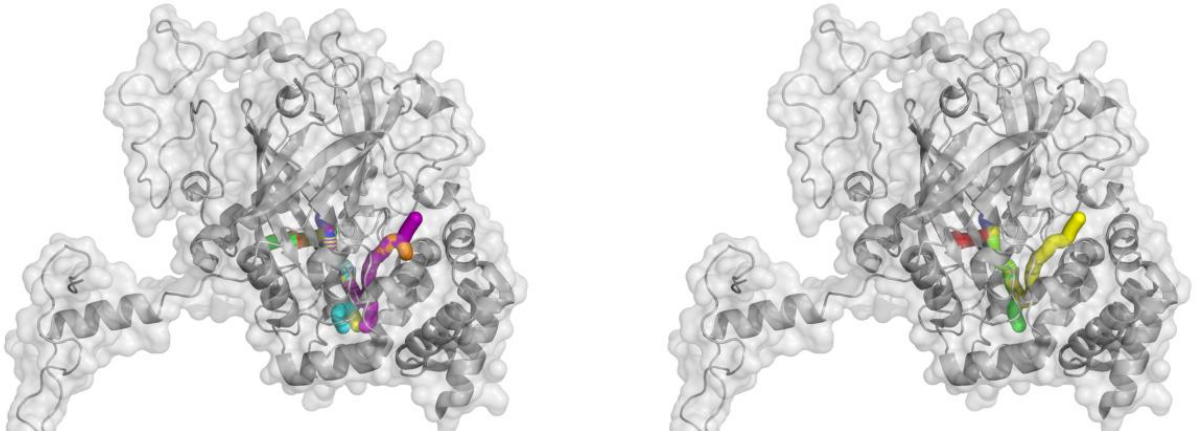


Figure 1. 1mqf, left main's algorithm at the branch's defaults (8 channels), right this branch at the same defaults (4).

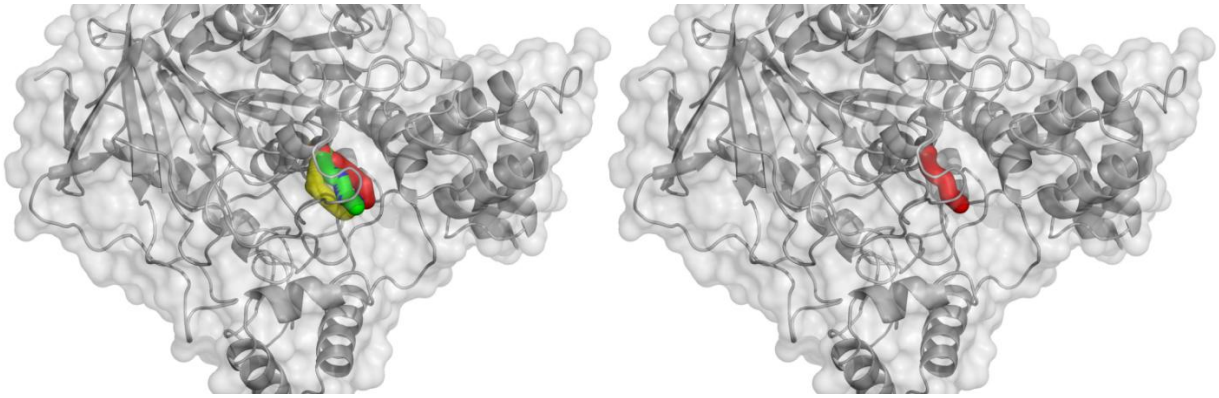


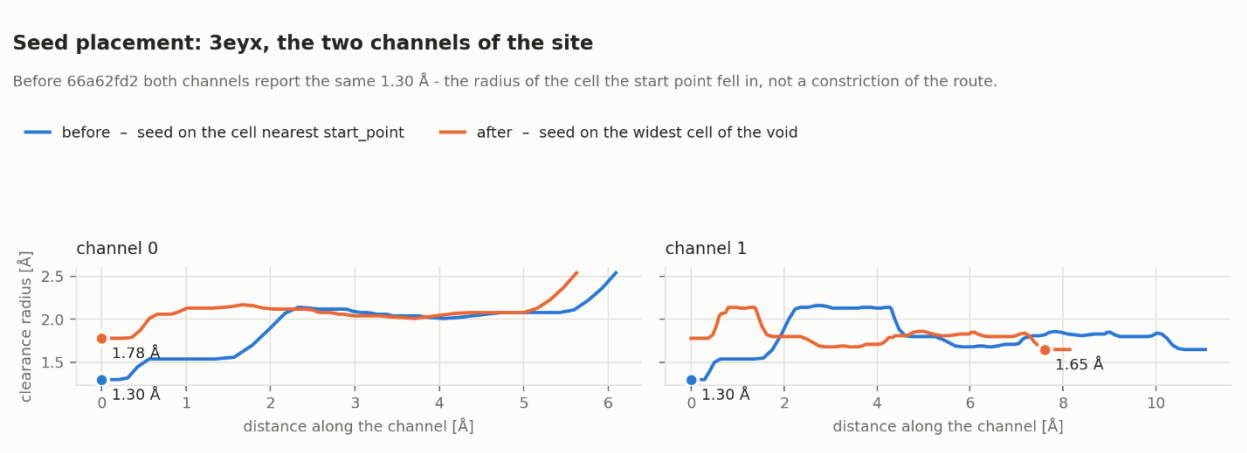
Figure 2. 2ace (acetylcholinesterase), left main's algorithm at the branch's defaults (4 channels), right this branch (1). The four run the same gorge and part only near the surface.

2. Seeding from a user start point

The seed was capped by the point the user gave. Every channel of a site starts at the seed tetrahedron, so its inscribed radius is an upper bound on all of their bottlenecks. The floor for widening the seed was the anchor's own depth - but a start_point is usually placed on a ligand or a catalytic residue, deeper than the widest part of the pocket around it, which left nothing eligible. The seed then stayed on the narrow cell nearest the point and silently capped the whole site. The floor is now min_depth raised to the cavity's widest mouth, as the chamber pass already did for automatic seeds.

seed placement (66a62fd2), bottlenecks in A

case	before	after
2cqt	2.20, 2.20, 1.52	2.48, 2.48, 1.51
3eyx	1.30, 1.30	1.78, 1.65
1tqn	1.33, 1.90, 2.17	1.33, 1.90, 2.22



3. Centerline spline

The centerline spline was parameterized by tetrahedron index. Tetrahedra are not evenly spaced, so the curve spent equal parameter on unequal distances. It is now parameterized centripetally by distance (√step), with the radius profile carried by a PCHIP interpolant so it cannot overshoot between samples.

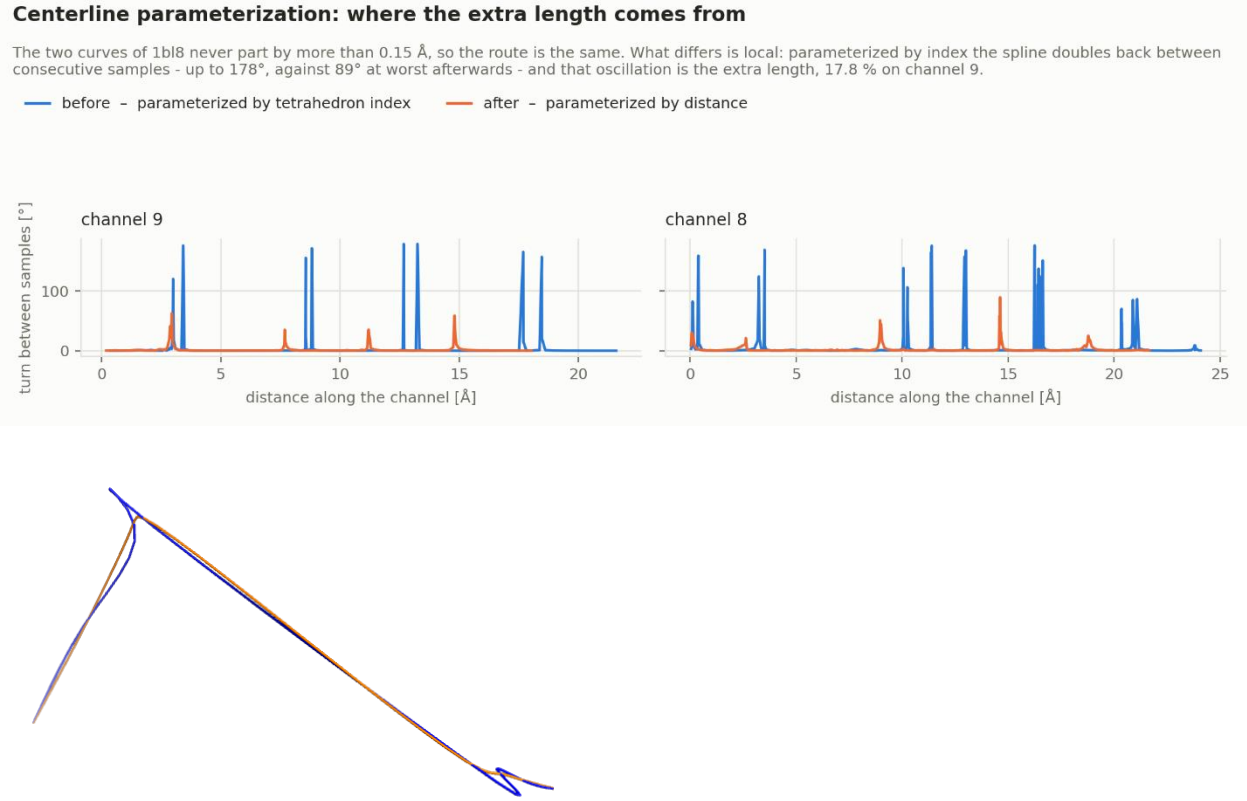
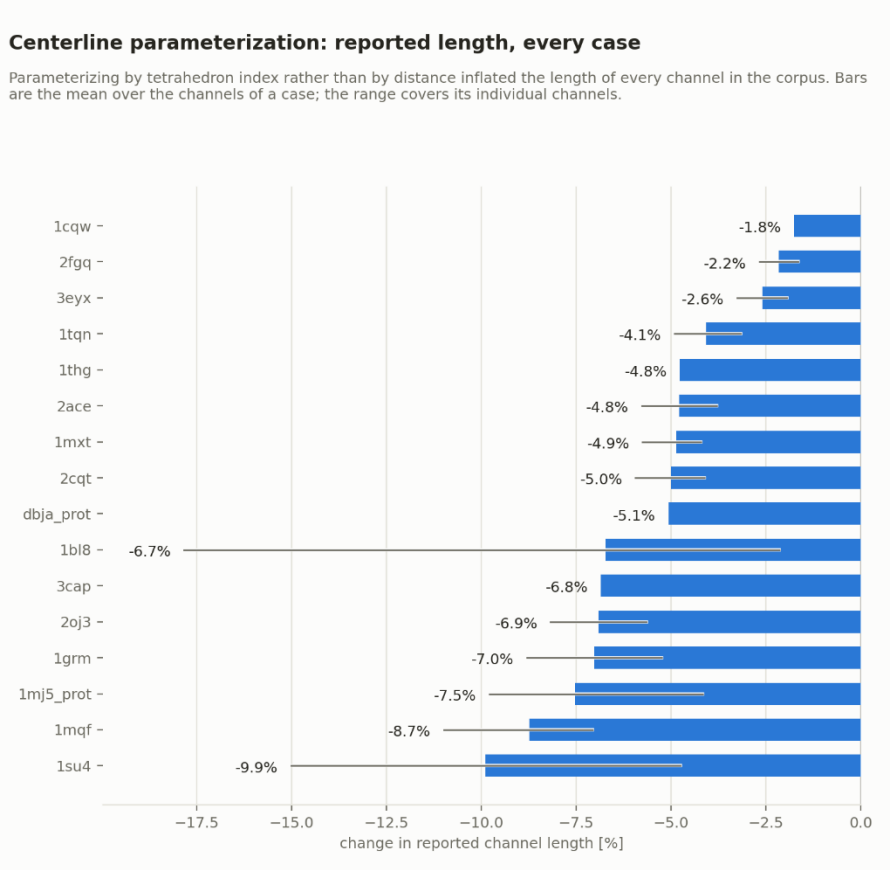


Figure 4. Illustration of the actual centerline with the knots and deviations, before is in blue, after in orange.

Across the corpus: every channel in every case got shorter, by 1.8 % (1cqw) to 9.9 % (1su4) on average per case, up to 17.8 % on a single channel. No case changes its channel count, and the widest bottleneck moves by at most 0.06 Å (1mqf 1.93 -> 1.99), almost always upward - a correctly spaced curve samples the clearance where it is actually widest.



4. Recalibrated defaults

inner_radius 0.9 -> 1.2 Å (f8539989): 1.2 Å is the smallest probe that gives similar channels with and without explicit hydrogens, so the default no longer sits inside the range the code itself warns about. As the decomposition in section 0 shows, this has large effect on the reported channel count

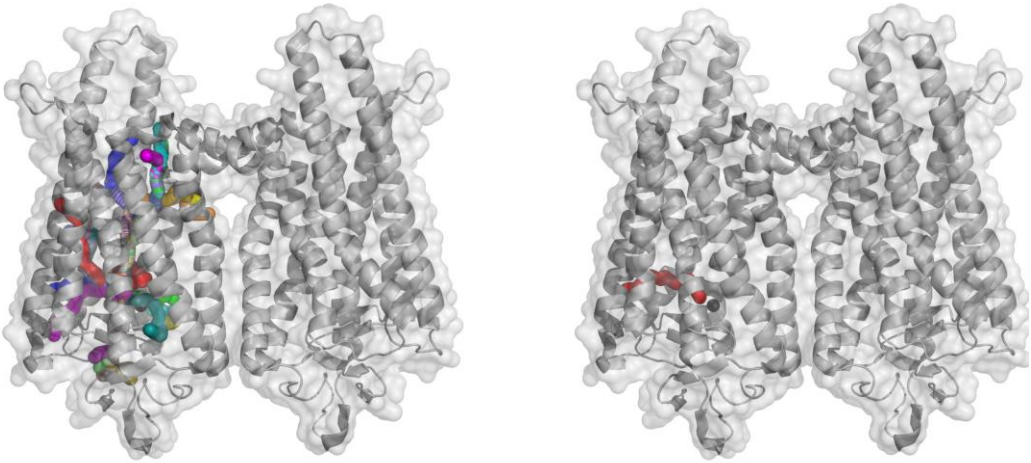


Figure 6. 3cap (rhodopsin), left with its own inner_radius=0.9 (27 channels), right at 1.2 Å (1). Twenty-six of the 27 routes exist only for a probe smaller than water, through gaps that a structure without explicit hydrogens leaves open. Nothing about deduplication is visible here - this is the probe alone.

surf_radius 3 -> 15 (5b93bcd6), which is what recovers the porin lumen and the 2bg9 vestibule. Because of the surface peel, the results for all globular proteins with smaller enterances are same, tested with radius from 3-20

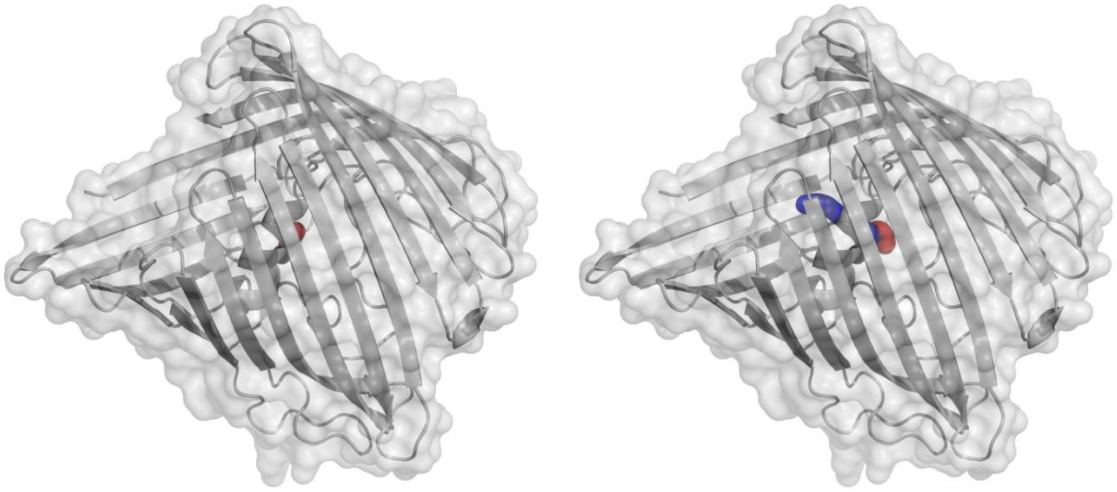


Figure 7. 2fgq (porin), left main with its defaults (1 channel, bottleneck 0.83 Å), right this branch (2 channels, 2.20 Å). With surf_radius=3 the barrel mouth is already treated as outside, and what is reported is a narrow side route
rather than the pore.

5. Protonation-aware input check

The hydrogen check asked only whether any hydrogen was present. A structure carrying its polar hydrogens and nothing else therefore passed as protonated and could be probed below water size - exactly the input an MD frame stripped of non-polar hydrogens gives you. The test moved into _reportAtomsInputComposition, which already reports what the input is made of, and now scores each kind of chain against the hydrogens per heavy atom it should carry: near 1.0 for protein, near 0.55 for nucleic acid (measured on 1JJ2_prot: protein 0.987, RNA 0.505 - a nucleotide is far richer in heavy atoms that carry no hydrogen). A mixed structure counts as protonated only when every kind of chain in it is, and the message names whichever is short.

Nucleic acid is also named in the composition report rather than counted among the components to warn about, so a protein-nucleic acid complex no longer draws advice to select the protein alone. 2oj3 (774 protein + 1552 DNA atoms).