Supplementary Material

CaverDock: A Molecular Docking-Based Tool to Analyse Ligand Transport through Proteins

Vavra, O.[#1,2], Filipovic, J. [#3], Plhak, J.[3], Bednar, D.[1,2], Marques, S.M.[1,2], Brezovsky, J.[1,4], Stourac, J.[1,2], Matyska, L.[3], Damborsky, J.[1,2]*

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic;
[2]International Centre for Clinical Research, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno, Czech Republic;
[3]Institute of Computer Science, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic;
[4]Current address: Laboratory of Biomolecular Interactions and Transport, Department of Gene Expression, Institute of Molecular Biology and Biotechnology Faculty of Biology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan, Poland; International Institute of Molecular and Cell Biology in Warsaw, Ks Trojdena 4, 02-109, Warsaw, Poland

* To whom correspondence should be addressed. [#] These authors contributed equally to this work

List of supplementary data

SI-1: Description of CaverDock calculation.

SI-2: Dataset I – Benchmarking.

SI-3: Dataset II – Geometry of tunnels.

SI-4: Dataset III – Geometry of substrates.

SI-5: Dataset IV - Tunnel engineering.

SI-6: Comparison of settings and functionalities in CaverDock, SLITHER, and MoMA-LigPath.

SI-7: The results obtained for the Dataset II - Geometry of tunnels.

SI-8: Energy profiles from the Dataset II - Geometry of tunnels.

SI-9: Lower-bound energies of the ligands passing through the p1 tunnel of LinB.

SI-10: Lower-bound energies of the ligands passing through the p2 tunnel of LinB.

SI-11: Lower-bound energy profiles of the halogenated substrates passing through p1 and p2 tunnels of LinB.

SI-12: Linear regression plots of the lower-bound energies and the experimental data.

SI-13: Parameters of the tunnels found in rationally engineered LinB variants.

SI-14: Analysis of side-chain flexibility.

SI-15: Lower-bound energy profiles from the iterative flexibility.

SI-16: Rosetta scores of the ligand-free protein structures from the iterative flexibility simulations.

SI-17: Analysis of backbone dynamics.

SI-18: CaverDock tunnel radii and energy profiles based on snapshots from accelerated molecular dynamics.

SI-19: Molecular structures of all used ligands.

SI-20: Compiled datasets used for validation of CaverDock (ZIP).

SI-1: Description of CaverDock calculation.

CaverDock computation consists of three essential steps: (i) tunnel discretization, (ii) definition of restraints, and (iii) computation of lower-bound and upper-bound trajectories.

I. Tunnel discretization

The input tunnel is represented by a set of spheres $T = \{S_i\}_{i=1}^{k}$, obtained from CAVER 3.02 (Chovancova et al., 2012) or a similar tool. To be further processed by CaverDock, the set of spheres is discretized into a set of discs. A disc is defined as $\theta = (A, u, r)$, where $A \in R^3$ is a centre, $u \in R^3$ is a normal and $r > 0$ is a radius. Let $\Theta = \{\theta\}_{i=0}^{n}$ be a sequence of discs cutting tunnel T.

During docking with position restraint, the selected atom of a ligand is placed to consecutive discs. We need to sample the tunnel finely and also generate a continuous trajectory. Therefore, we need to upper-bound the distance between the discs, so the disc cannot force ligand to change its position too much. Let δ be the highest distance of the discs allowed and θ_i, $\theta_{i+1} \in \Theta$. Then we require:

$$\begin{aligned} \forall x \in \theta_i \Rightarrow \exists y \in \theta_{i+1} \Rightarrow \|x - y\| \leq \delta \\ \forall y \in \theta_{i+1} \Rightarrow \exists x \in \theta_i \Rightarrow \|x - y\| \leq \delta \end{aligned} \quad (1)$$

Moreover, we require the cuts not to intersect each other in more than a single point:

$$\theta_i, \theta_j \in \Theta \Rightarrow |\theta_i \cap \theta_j| \leq 1 \quad (2)$$

II. Definition of restraints

We used the docking implemented in AutoDock Vina and extended it with restraints. During the docking, the position of a protein-ligand complex is minimized. Let the binding energy of the protein-ligand complex computed with AutoDock Vina be E_{vina}. The restraints are implemented as the new energy $E_{position}$ and $E_{pattern}$. The original AutoDock Vina is modified in the way that a sum $E_{vina} + E_{position} + E_{pattern}$ is minimized, so the ligand finds a position, where binding-free energy is minimized in a space restricted by restraints. However, the output energy presented by CaverDock is only E_{vina}, so the restraints energy keeps the ligand within the defined position, but does not influence the numerical value of the energy output. The ligand conformation λ is defined by the Cartesian position of its atoms: $\lambda = \{a_i\}_{i=1}^{m}$. The position restraint snaps atom $a_c \in \lambda$ to any position at disc θ:

$$a_c \in \theta \quad (3)$$

To keep the ligand at the disc, the restraint energy $E_{position}$ is computed as follows:

$$E_{position} = p_{max} - p_{max} e^{-\frac{|a_c - t|^2}{0.5}} \quad (4)$$

where p_{max} is the maximum penalization energy and for $t \in \theta$ holds $\forall u \in \theta$, u ≠ t, $|a_c - u| > |a_c - t|$ (i. e. t is the point in θ, which is the nearest to a_c). The bell-shaped function is used to compute $E_{position}$ to avoid a too strong penalization of small distances between a_c and θ, which keeps the good numerical

stability of the AutoDock Vina's optimization method. The pattern restraint keeps the ligand λ in the vicinity of the pattern position $\lambda_{pattern}$:

$$\forall j \in [1,m] : \left[a_j - b_j\right] < \delta \tag{5}$$

where $a_j \in \lambda$, $b_j \in \lambda_{pattern}$.

The energy of the pattern restraint is computed as:

$$E_{pattern} = c \cdot \sum_{a \in \lambda, b \in \lambda_{pattern}} max(0, |a - b| - \delta) \tag{6}$$

where c is a constant determining the strength of the pattern (it has been empirically set to 40). Apparently, Eq. 6 is not differentiable in the area where $|a - b| = \delta$. We define the derivative at these points to be 0. This simple definition of pattern energy keeps the computation of the pattern restraint computationally efficient and the gradient-based optimization in Vina possible. During the minimization of $E_{vina} + E_{position} + E_{pattern}$, some restraint can be broken when the binding-free energy E_{vina} is too high. The restraint is considered broken when its energy is higher than 5 kcal/mol and 8 kcal/mol for the position and pattern restraint, respectively. CaverDock does not use the results in such a case.

III. Trajectory search

The *lower-bound trajectory* is determined by docking a ligand at each disc $\theta \in \Theta$, using only the position restraint. Computation of the *upper-bound trajectory* is more difficult. It is obtained by iterative docking onto consequent discs with restricted changes in the position of all atoms by a pattern restraint. However, such a trajectory may be sub-optimal. The ligand movement prefers the transition where the ligand follows the strongest energy gradient locally between the following steps. It may be more preferable for the ligand to make a transition to some different conformation, which may allow it to pass the energy barrier with lower energy. Therefore, CaverDock needs to search for multiple variants of ligand trajectories. We implemented a simple heuristic. The ligand is moved in one direction in the tunnel, e.g. from the binding site to the protein surface, producing a trajectory $\lambda_1, \lambda_2, \ldots$ When the binding free energy at some disc θ_i is significantly higher than the binding free energy of some known conformation λ_{low} at the same disc (λ_{low} may be obtained during lower-bound trajectory computation), we set $\lambda_i = \lambda_{low}$, and search the trajectory moving the ligand backwards to previous disks $\theta_{i-1}, \theta_{i-2}, \ldots$ The backtracking ends when the forward and backward trajectories converge or when the beginning of the tunnel is reached. Note that the resulting trajectory still follows one direction only. When the backtracking is used the trajectory is reversed and integrated into a forward trajectory.

SI-2: Dataset I – Benchmarking.

PDB	Enzyme	Ligand
1BN7	Haloalkane dehalogenase	1-Chlorobutane
1MAH	Acetylcholinesterase	Acetylcholine
2A65	Leucine transporter	Leucine
1PV7	Lactose permease	Lactose
1SUK	Glucose transporter	α-D-Glucopyranose
1TCC	Lipase B	4-Methyloctanoic acid
1ZNJ	Insulin hexamer	Phenol
1RC2	Aquaporin Z	Glycerol
1IE9	Vitamin D receptor	1,25-Dihydroxyvitamin D3
3LC4	Cytochrome P450 2E1	Arachidonic acid

SI-3: Dataset II – Geometry of tunnels.

PDB ID	Enzyme	Substrate number	Substrate	Number of tunnels
1AKD	Cytochrome P450CAM	1	(*R*)-camphor	3
1MQF	Compound I from *Proteus mirabilis* catalase	1	Formic acid	3
		2	Hydrogen peroxide	3
1MXT	Cholesterol oxidase	1	Cholesterol	3
1THG	Lipase	1	1,2-dihexadecanoyl-3-(9Z-octadecenoyl)-sn-glycerol	3
2ACE	Acetylcholinesterase	1	Acetylcholine	3
2BG9	Nicotinic acetylcholine acceptor	1	Acetylcholine	1
2BJV	Choline oxidase	1	Choline	3
1A9X	Carbamoyl phosphate synthetase	1	Adenosine triphosphate	3
		2	Bicarbonate	3
		3	L-glutamine	3
1B37	Polyamine oxidase	1	Spermidine	11
1BRT	Bromoperoxidase	1	Acetate	3
1EA5	Acetylcholinesterase	1	Acetylcholine	2
1HNJ	Beta-ketoacyl-ACP synthase III	1	Acetyl-CoA	3
1I88	Chalconesynthase	1	Coumaroyl-CoA	5
		2	Malonyl-CoA	5
1M1N	Nitrogenase MoFe protein	1	Adenosine triphosphate	3
1PBE	P-hydroxybenzoate hydroxylase	1	4-hydroxybenzoate	5
1V4A	Aryl esterase	1	Phenylacetate	3
1YGE	Lipoxygenase	1	Linoleate	8
1YRC	Cytochrome P450CAM	1	(*R*)-camphor	3
2AYL	Prostaglandin H2 synthase 1	1	Arachidonic acid	9
2C3N	Glutathione-S-transferase	1	Glutathione	3
2IID	L-amino acid oxidase	1	L-phenylalanine	12
2INC	Toluene/o-xylene monooxygenase hydroxylase	1	Toluene	4
2Q9O	Laccase	1	2,6-dimethoxyphenol	3
		2	2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid)	3
		3	Guaiacol	3

		4	Syringaldazine	3
25QC	Squalene-Hopene cyclase	1	Squalene	7
3C6X	Hydroxynitrile lyase	1	Acetone cyanohydrin	1
3TTV	Catalase	1	Hydrogen peroxide	6

SI-4: Dataset III – Geometry of substrates.

Substrate group	Substrate code	Substrate
Propanes	m003	1-chloropropane
	m017	1-bromopropane
	m028	1-iodopropane
	m072	1,2-dibromopropane
	m038	1,3-dichloropropane
	m048	1,3-dibromopropane
	m052	1-bromo-3-chloropropane
	m076	2-bromo-1-chloropropane
	m084	1-bromo-2-methylpropane
Butanes, pentanes, hexanes, heptanes, octanes	m004	1-chlorobutane
	m018	1-bromobutane
	m040	1,5-dichloropentane
	m006	1-chlorohexane
	m020	1-bromohexane
	m031	1-iodohexane
	m007	1-chloroheptane
	m008	1-chlorooctane
Miscellaneous alkanes and alkenes	m047	1,2-dibromoethane
	m115	chlorocyclohexane
	m117	bromocyclohexane
	m209	3-chloro-2-methylpropene
	m222	3-chloro-2-(chloromethyl)-1-propene
	m225	2,3-dichloropropene
	m141	4-bromobutyronitrile
	m111	bis(2-chloroethyl)ether

SI-5: Dataset IV – Tunnel engineering.

LinB variant	PDB ID	Tunnel	Ligand
LinBWT	1K63	P1	2-bromoethanol
		P3	2-bromoethanol
LinB32	4WDQ	P1	2-bromoethanol
		P3	2-bromoethanol
LinB86	5LKA	P1	2-bromoethanol
		P3	2-bromoethanol

SI-6: Comparison of settings and functionalities in CaverDock, SLITHER, and MoMA-LigPath.

Tool	CaverDock	SLITHER	MoMA-LigPath
Settings	Vina docking parameters	Slither settings, docking algorithm, docking settings	Standard settings, advanced RRT algorithm settings
Additional options	Flexible receptor	Flexible receptor, relaxed receptor	Block side-chain flexibility
Input preparation	Receptor and ligand, tunnel geometry, tunnel discretization, grid box, Vina configuration file	Receptor, ligand, tunnel aligned with y-axis, grid box surrounding the tunnel	Ligand docked in the active site
Input files	Protein PDBQT, ligand PDBQT, discretized tunnel	Protein PDB or PDBQ, ligand PDB or PDBQ	Complex PDB
Runtime	Minutes	Minutes	Minutes
Output	Ligand trajectory in single PDBQT file, energy information in REMARK	Ligand trajectory in a single PDB file, energy information in REMARK	Ligand trajectory in multiple PDB files, no energy information

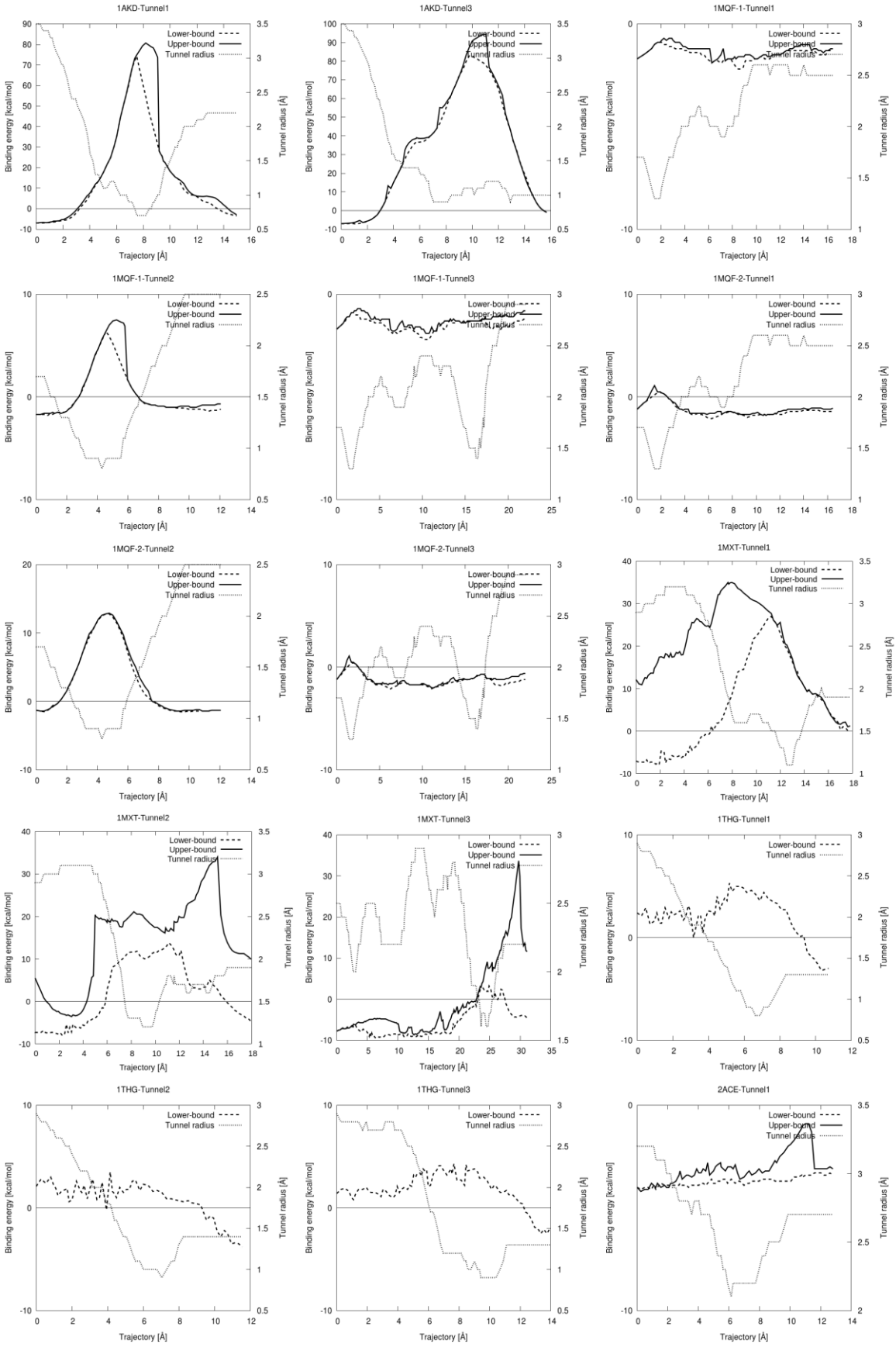
SI-7: The results obtained for the Dataset II – Geometry of tunnels.

PDB ID	Substrate number	Number of tunnels	Ligand passage		
			Upper and lower-bound[a]	Lower-bound only[b]	Did not pass[c]
1AKD	1	3	2	0	1
1MQF	1	3	3	0	0
	2	3	3	0	0
1MXT	1	3	3	0	0
1THG	1	3	0	3	0
2ACE	1	3	2	1	0
2BG9	1	1	0	1	0
2BJV	1	3	3	0	0
1A9X	1	3	1	2	0
	2	3	3	0	0
	3	3	3	0	0
1B37	1	11	8	3	0
1BRT	1	3	3	0	0
1EA5	1	2	2	0	0
1HNJ	1	3	0	1	2
1I88	1	5	1	3	1
	2	5	0	5	0
1M1N	1	3	1	0	2
1PBE	1	5	5	0	0
1V4A	1	3	3	0	0
1YGE	1	8	1	7	0
1YRC	1	3	2	0	1
2AYL	1	9	1	8	0
2C3N	1	3	2	1	0
2IID	1	12	12	0	0
2INC	1	4	4	0	0
2Q9O	1	3	3	0	0
	2	3	0	0	3
	3	3	3	0	0
	4	3	0	2	1
25QC	1	7	0	7	0
3C6X	1	1	1	0	0
3TTV	1	6	6	0	0

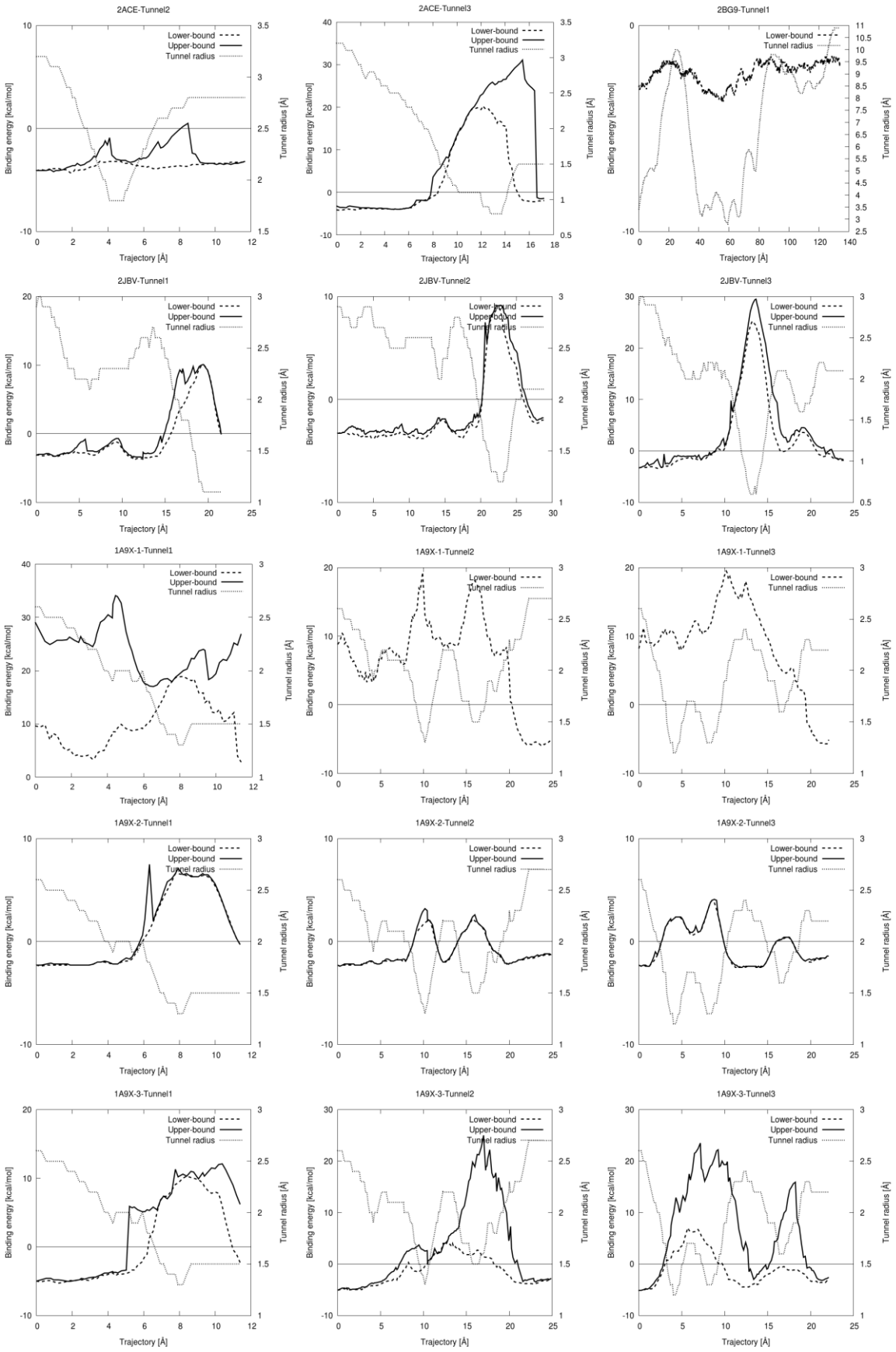
[a] Ligand passed without any hindrance

[b] Ligand passed with difficulties, the upper-bound trajectory was not calculated

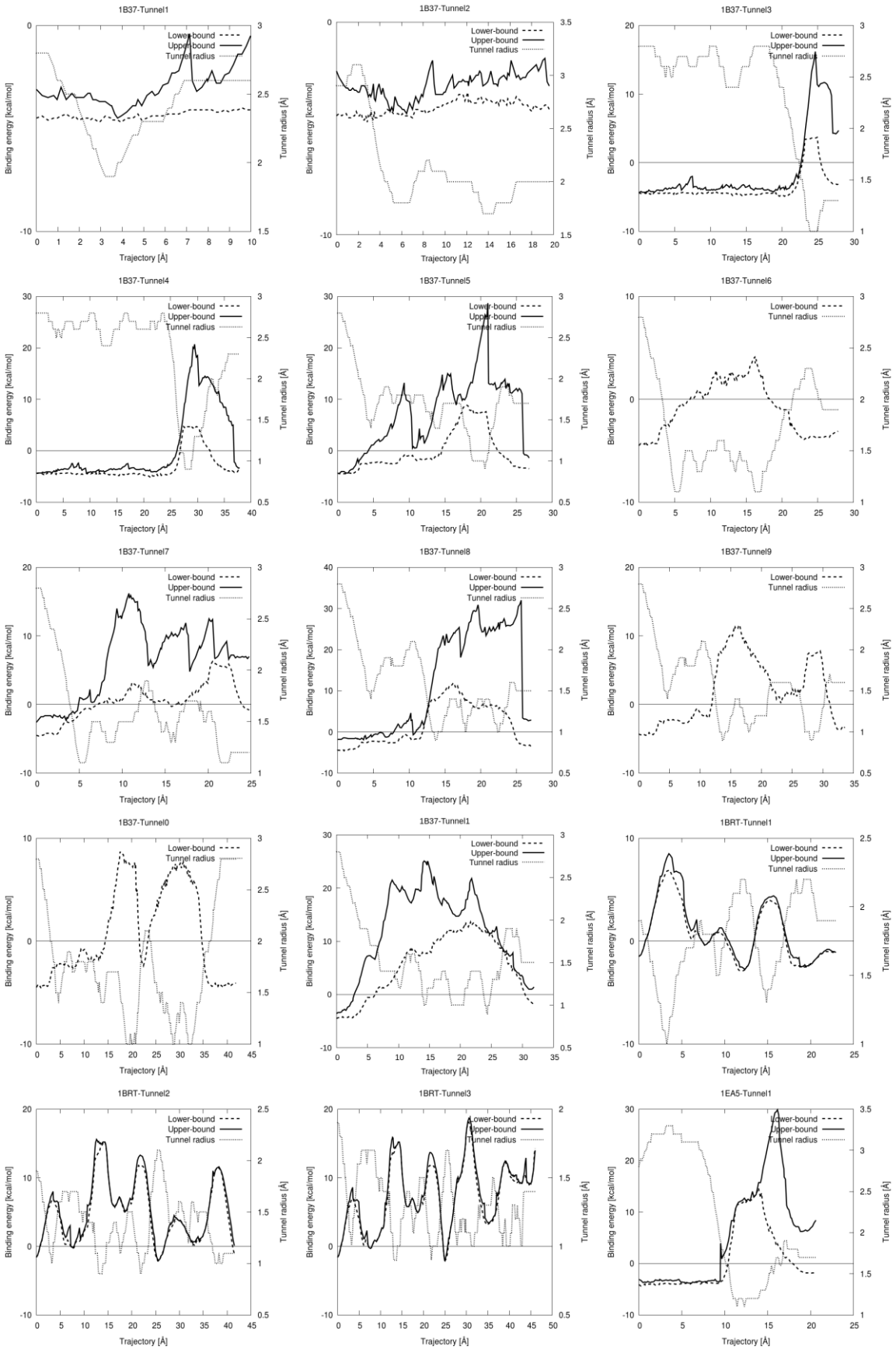
[c] Ligand was too large to pass through



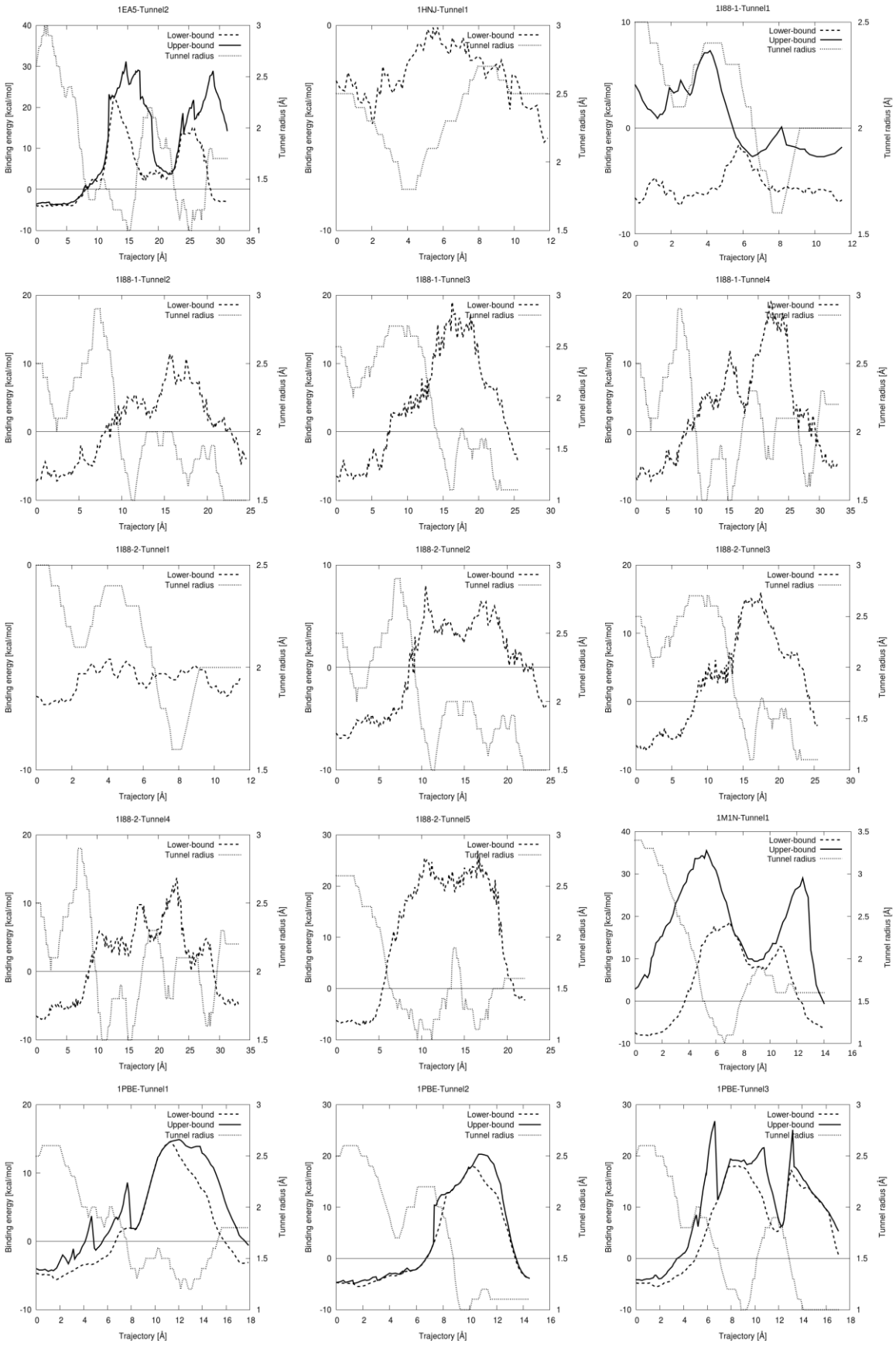
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 1).



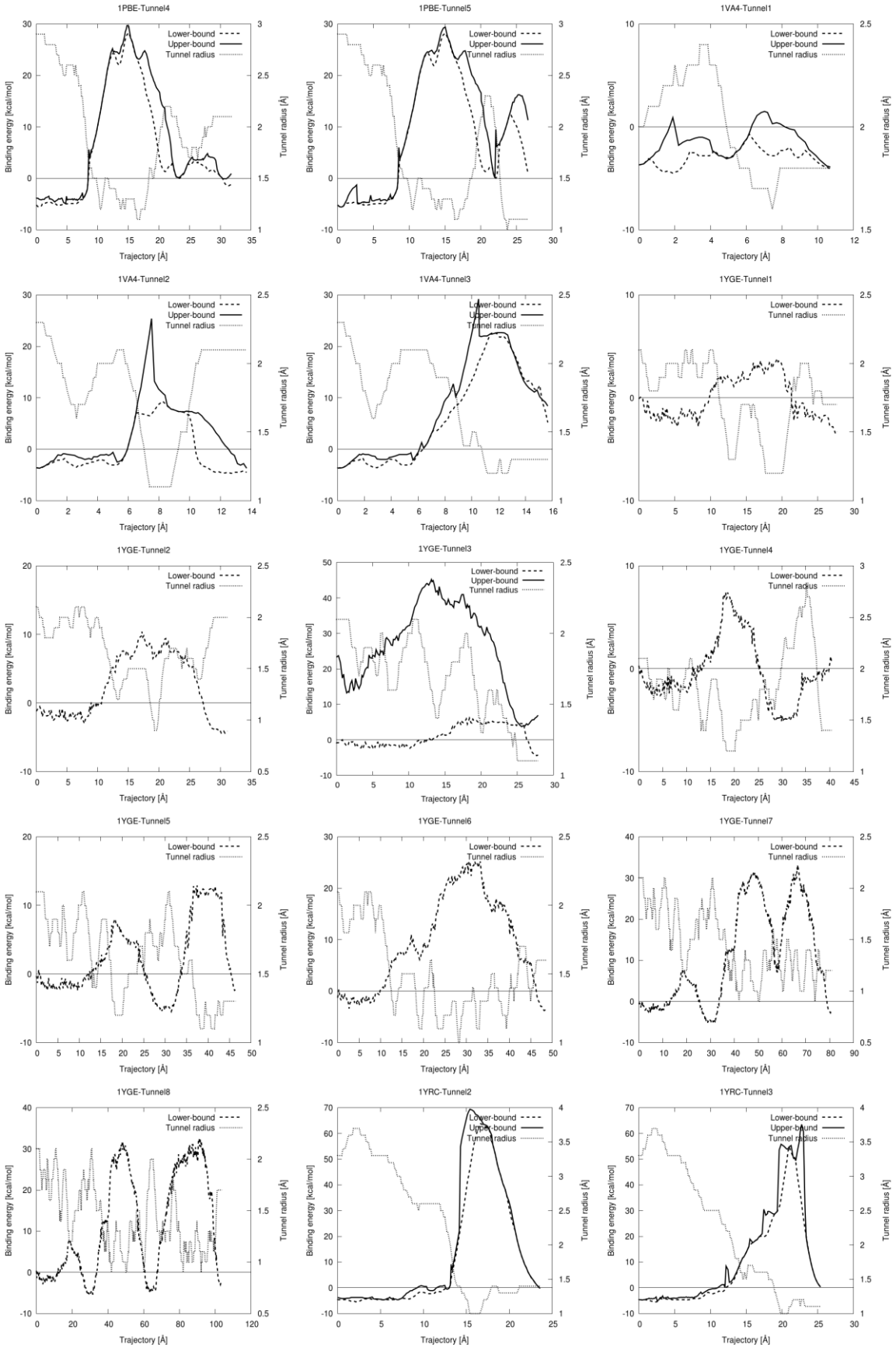
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 2).



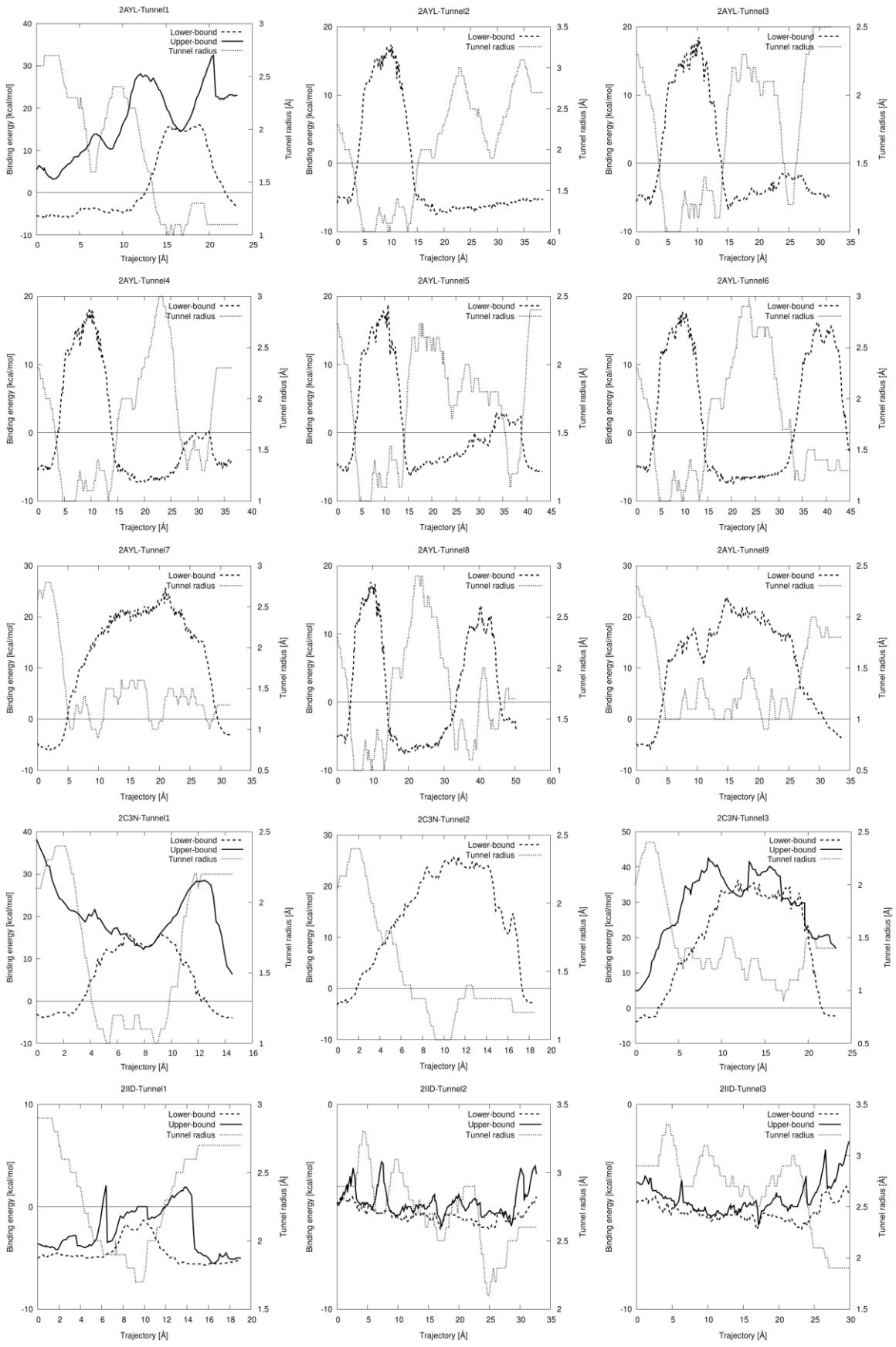
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 3).



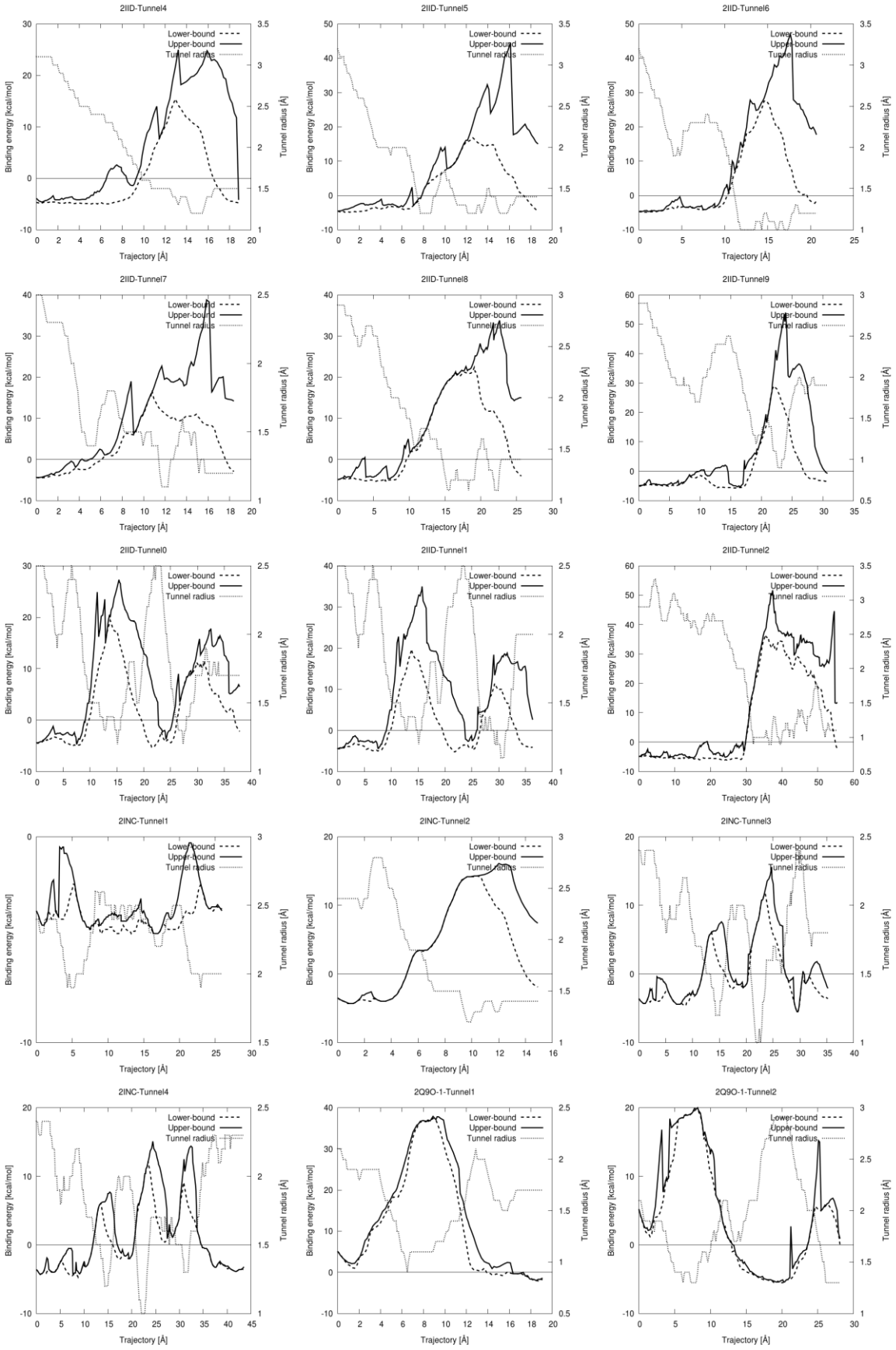
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 4).



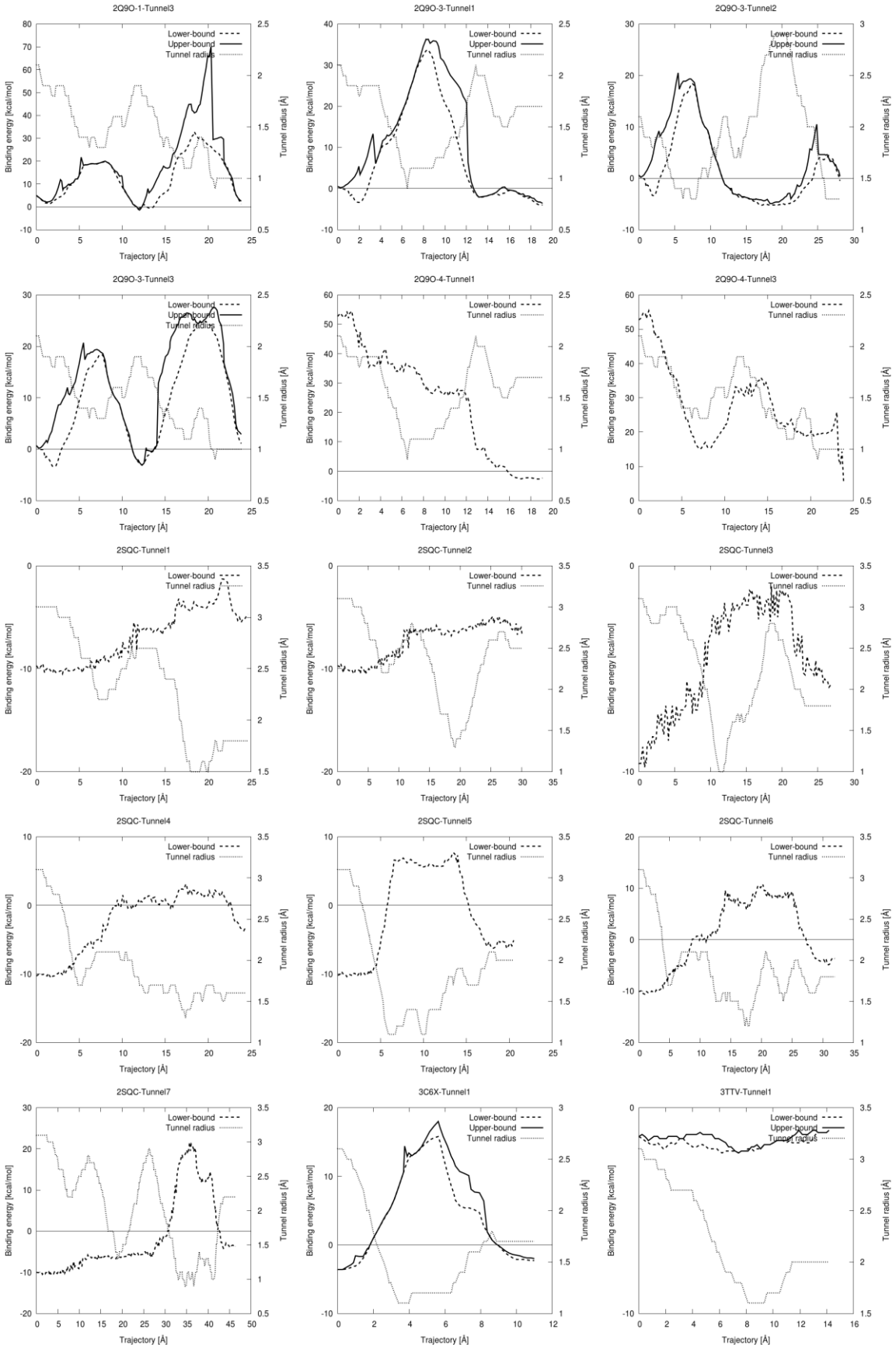
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 5).



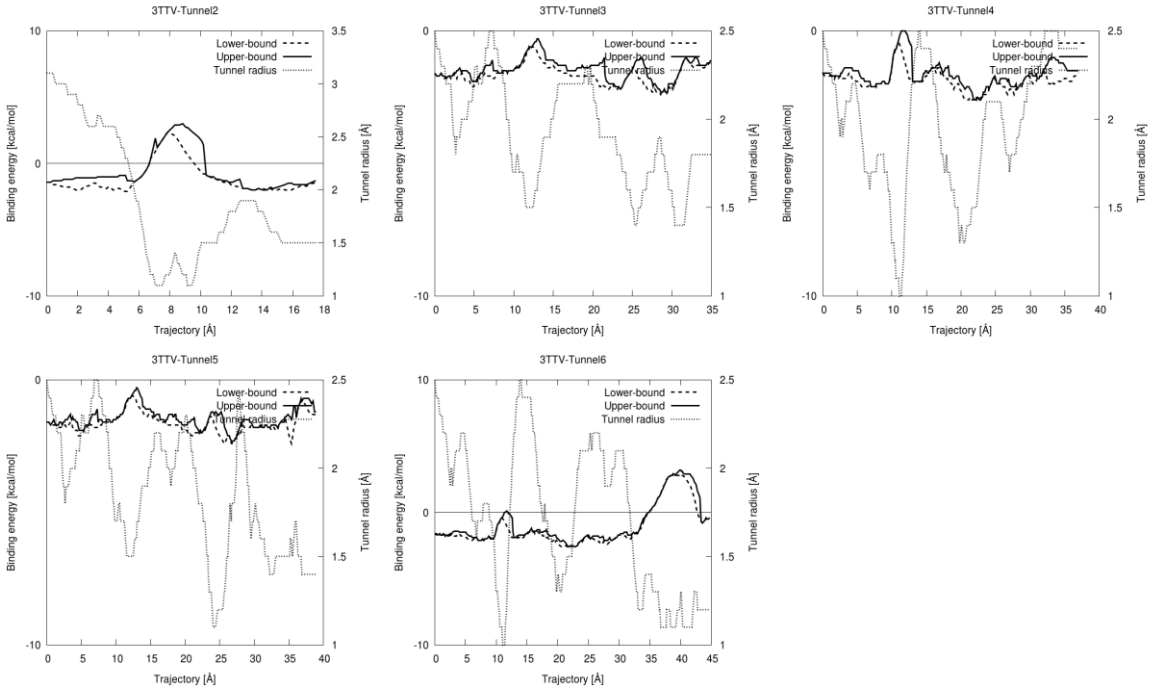
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 6).



SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 7).



SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 8).



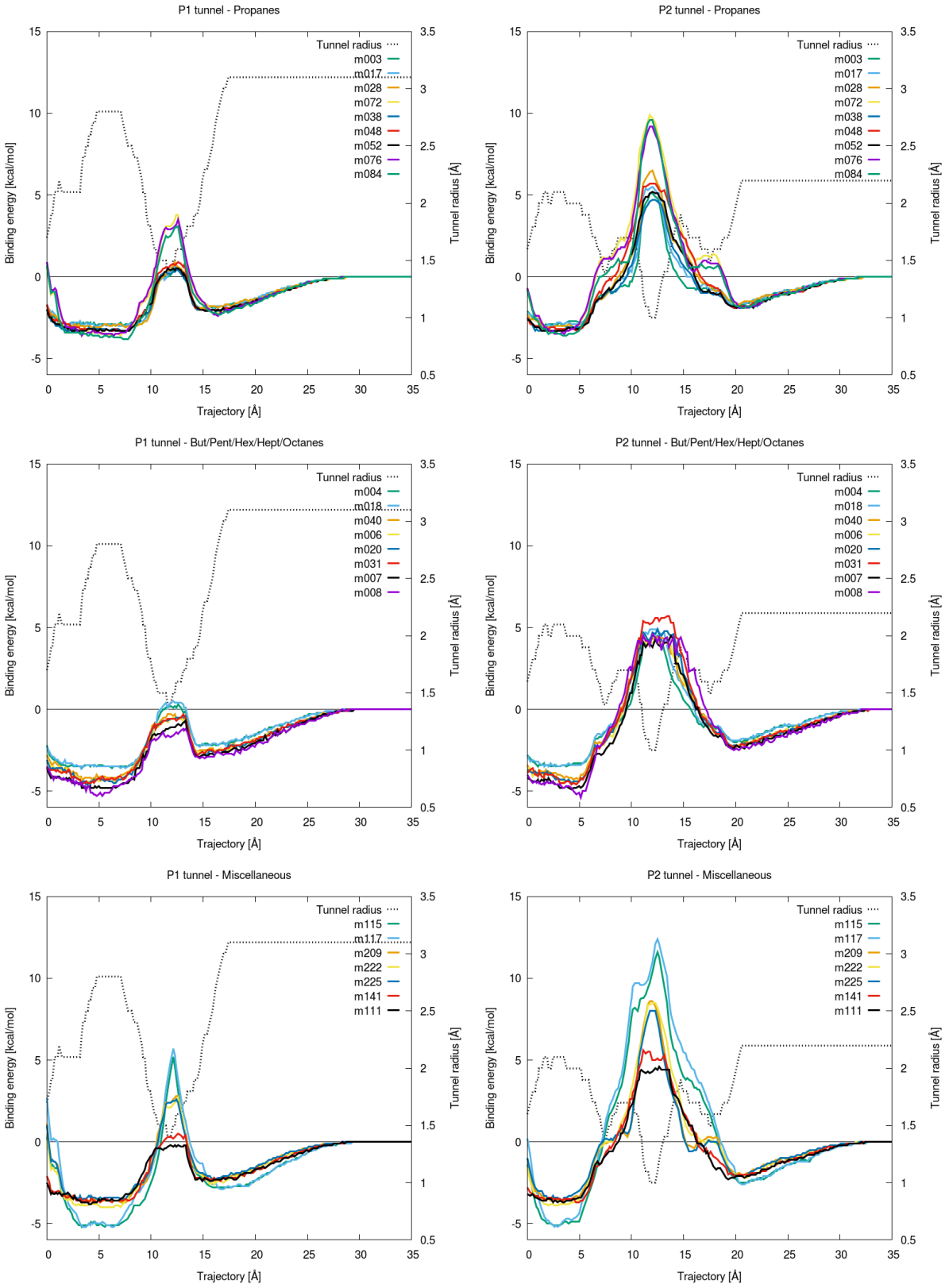
SI-8: Energy profiles from the Dataset II – Geometry of tunnels (part 9).

SI-9: Lower-bound energies of the ligands passing through the p1 tunnel of LinB (PDB ID 1K63).

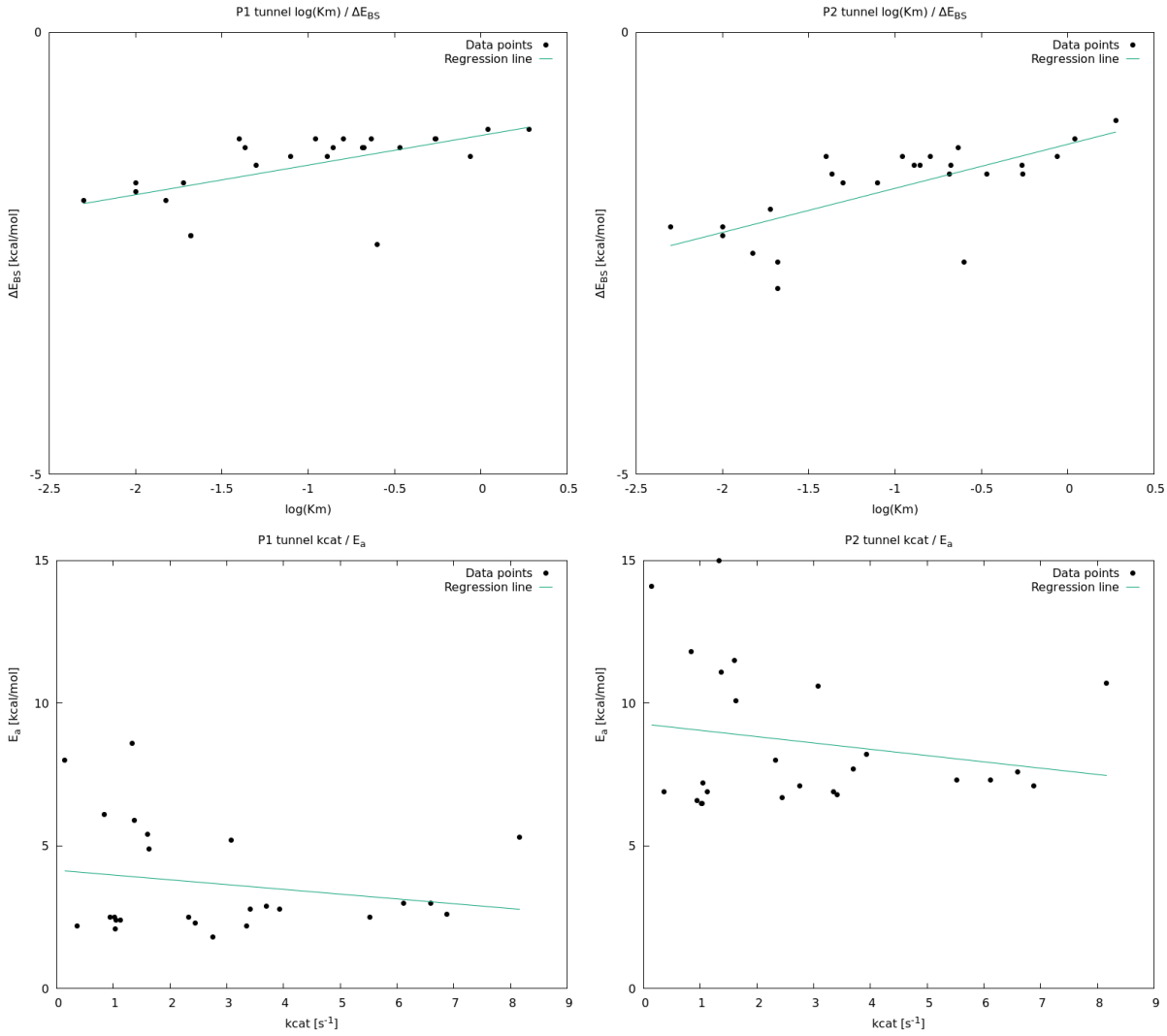
Substrate group	Substrate code	Substrate	K_M [mM]	log(K_M)	k_{cat} [s^{-1}]	E_{Bound}	E_{Max}	$E_{Surface}$	E_a [kcal/mol]	ΔE_{BS} [kcal/mol]
Propanes	m003	1-chloropropane	1.100	0.041	1.124	-3.0	0.5	-1.9	2.4	-1.1
	m017	1-bromopropane	0.231	-0.636	5.519	-3.1	0.6	-1.9	2.5	-1.2
	m028	1-iodopropane	0.110	-0.959	3.935	-3.1	0.9	-1.9	2.8	-1.2
	m072	1,2-dibromopropane	0.140	-0.854	0.843	-3.6	3.8	-2.3	6.1	-1.3
	m038	1,3-dichloropropane	0.160	-0.796	0.950	-3.3	0.4	-2.1	2.5	-1.2
	m048	1,3-dibromopropane	0.040	-1.398	6.600	-3.3	0.9	-2.1	3.0	-1.2
	m052	1-bromo-3-chloropropane	0.210	-0.678	6.886	-3.4	0.5	-2.1	2.6	-1.3
	m076	2-bromo-1-chloropropane	0.551	-0.259	1.374	-3.6	3.5	-2.4	5.9	-1.2
	m084	1-bromo-2-methylpropane	0.050	-1.301	1.599	-3.8	3.1	-2.3	5.4	-1.5
Butanes, pentanes, hexanes, heptanes, octanes	m004	1-chlorobutane	0.129	-0.889	1.020	-3.6	0.3	-2.2	2.5	-1.4
	m018	1-bromobutane	0.043	-1.367	3.414	-3.6	0.5	-2.3	2.8	-1.3
	m040	1,5-dichloropentane	0.019	-1.721	2.450	-4.3	-0.3	-2.6	2.3	-1.7
	m006	1-chlorohexane	0.005	-2.301	1.033	-4.5	-0.5	-2.6	2.1	-1.9
	m020	1-bromohexane	0.010	-2.000	1.050	-4.6	-0.4	-2.8	2.4	-1.8
	m031	1-iodohexane	0.010	-2.000	2.327	-4.5	-0.3	-2.8	2.5	-1.7
	m007	1-chloroheptane	0.015	-1.824	3.345	-4.8	-0.7	-2.9	2.2	-1.9
	m008	1-chlorooctane	0.021	-1.678	2.750	-5.3	-1.2	-3.0	1.8	-2.3
Miscellaneous alkanes and alkenes	m047	1,2-dibromoethane	1.900	0.279	6.120	-2.9	1.2	-1.8	3.0	-1.1
	m115	chlorocyclohexane	0.252	-0.599	0.144	-5.2	5.2	-2.8	8.0	-2.4
	m117	bromocyclohexane	0.021	-1.678	1.335	-5.2	5.7	-2.9	8.6	-2.3
	m209	3-chloro-2-methylpropene	0.340	-0.469	3.080	-3.7	2.8	-2.4	5.2	-1.3
	m222	3-chloro-2-(chloromethyl)-1-propene	0.079	-1.102	8.165	-4.0	2.7	-2.6	5.3	-1.4
	m225	2,3-dichloropropene	0.542	-0.266	1.632	-3.5	2.6	-2.3	4.9	-1.2
	m141	4-bromobutyronitrile	0.207	-0.684	3.701	-3.7	0.5	-2.4	2.9	-1.3
	m111	bis(2-chloroethyl)ether	0.870	-0.060	0.363	-3.8	-0.2	-2.4	2.2	-1.4

SI-10: Lower-bound energies of the ligands passing through p2 tunnel of LinB (PDB ID 1K63).

Substrate group	Substrate code	Substrate	K_M [mM]	log(K_M)	k_{cat} [s^{-1}]	E_{Bound}	E_{Max}	$E_{Surface}$	E_a [kcal/mol]	ΔE_{BS} [kcal/mol]
Propanes	m003	1-chloropropane	1.100	0.041	1.124	-3.0	5.1	-1.8	6.9	-1.2
	m017	1-bromopropane	0.231	-0.636	5.519	-3.1	5.5	-1.8	7.3	-1.3
	m028	1-iodopropane	0.110	-0.959	3.935	-3.1	6.5	-1.7	8.2	-1.4
	m072	1,2-dibromopropane	0.140	-0.854	0.843	-3.4	9.9	-1.9	11.8	-1.5
	m038	1,3-dichloropropane	0.160	-0.796	0.950	-3.3	4.7	-1.9	6.6	-1.4
	m048	1,3-dibromopropane	0.040	-1.398	6.600	-3.3	5.7	-1.9	7.6	-1.4
	m052	1-bromo-3-chloropropane	0.210	-0.678	6.886	-3.4	5.2	-1.9	7.1	-1.5
	m076	2-bromo-1-chloropropane	0.551	-0.259	1.374	-3.5	9.2	-1.9	11.1	-1.6
	m084	1-bromo-2-methylpropane	0.050	-1.301	1.599	-3.6	9.6	-1.9	11.5	-1.7
Butanes, pentanes, hexanes, heptanes, octanes	m004	1-chlorobutane	0.129	-0.889	1.020	-3.5	4.5	-2.0	6.5	-1.5
	m018	1-bromobutane	0.043	-1.367	3.414	-3.5	4.9	-1.9	6.8	-1.6
	m040	1,5-dichloropentane	0.019	-1.721	2.450	-4.3	4.4	-2.3	6.7	-2.0
	m006	1-chlorohexane	0.005	-2.301	1.033	-4.5	4.2	-2.3	6.5	-2.2
	m020	1-bromohexane	0.010	-2.000	1.050	-4.5	4.9	-2.3	7.2	-2.2
	m031	1-iodohexane	0.010	-2.000	2.327	-4.6	5.7	-2.3	8.0	-2.3
	m007	1-chloroheptane	0.015	-1.824	3.345	-4.9	4.5	-2.4	6.9	-2.5
	m008	1-chlorooctane	0.021	-1.678	2.750	-5.4	4.6	-2.5	7.1	-2.9
Miscellaneous alkanes and alkenes	m047	1,2-dibromoethane	1.900	0.279	6.120	-2.7	5.6	-1.7	7.3	-1.0
	m115	chlorocyclohexane	0.252	-0.599	0.144	-5.1	11.6	-2.5	14.1	-2.6
	m117	bromocyclohexane	0.021	-1.678	1.335	-5.2	12.4	-2.6	15.0	-2.6
	m209	3-chloro-2-methylpropene	0.340	-0.469	3.080	-3.6	8.6	-2.0	10.6	-1.6
	m222	3-chloro-2-(chloromethyl)-1-propene	0.079	-1.102	8.165	-3.9	8.5	-2.2	10.7	-1.7
	m225	2,3-dichloropropene	0.542	-0.266	1.632	-3.6	8.0	-2.1	10.1	-1.5
	m141	4-bromobutyronitrile	0.207	-0.684	3.701	-3.7	5.6	-2.1	7.7	-1.6
	m111	bis(2-chloroethyl)ether	0.870	-0.060	0.363	-3.7	4.6	-2.3	6.9	-1.4



SI-11: Lower-bound energy profiles of the halogenated substrates passing through p1(left) and p2 (right) tunnels of LinB.



SI-12: Linear regression plots of the lower-bound energies and the experimental data.

SI-13: Parameters of the tunnels found in rationally engineered LinB variants.

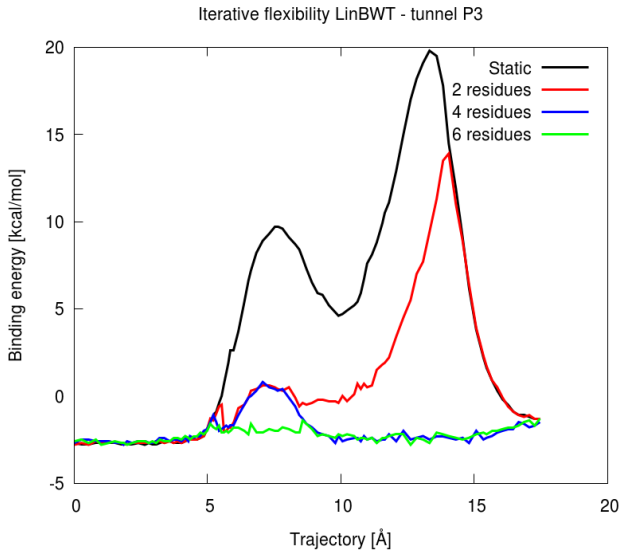
LinB variant	PDB ID	Tunnel	Average Bottleneck Radius	Average Length	Average Curvature	Priority
LinBWT	1K63	p1	1.44	15.02	1.21	0.68
		p3	0.71	17.05	1.35	0.26
LinB32[a]	4WDQ	p1	0.79	13.61	1.25	0.46
		p3	0.61	14.47	1.29	0.27
LinB86[b]	5LKA	p1	0.73	16.38	1.31	0.32
		p3	1.04	11.7	1.19	0.54

[a] variant denoted as LinB-closed[W]; [b] variant denoted as LinB-open[W]

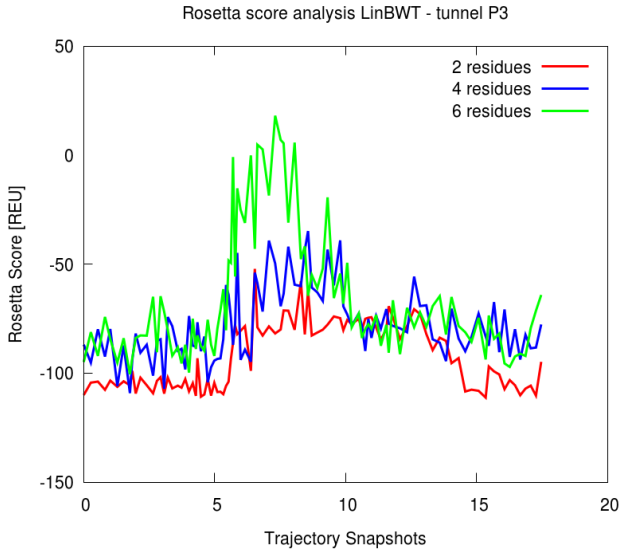
SI-14: Analysis of side-chain flexibility.

Analysis of the Dataset IV revealed that the results of MD simulations (Brezovsky *et al.*, 2016) provided a different picture than those obtained by the analysis of static structures. The simulations showed higher flexibility of the residues lining the p1 tunnel in LinB86 with three inserted mutations. As a result, only 10% of the unbinding trajectories preferred the p3 tunnel more than the p1 tunnel, supporting the importance of protein dynamics for the product release. Therefore, we tested the current implementation of flexible sidechains based on AutoDock Vina with the p3 tunnel in LinBWT. For this, we introduced flexibility to CaverDock calculations in three iterations by adding two flexible bottleneck residues in each iteration: Ile211+Ile213, Trp140+Arg155 and Phe143+Phe151. The flexibility helped to open parts of the tunnel with high energy barriers. In the last iteration, all obvious barriers were removed and the ligand was able to exit the protein with no spatial or energetic hindrance (SI-15).

We noticed that the flexible sidechains moved away from the ligand to let it escape, but these new sidechain conformations were partially clashing with the rest of the protein. Therefore, the local decrease in the energy barrier was offset by an increase in the whole protein's energy. To assess the likelihoods of the resulting structures, the protein conformations without the ligand from each trajectory snapshot were analysed by the Rosetta scoring function (Rohl *et al.*, 2004) (SI-16). The score calculated for the crystal structure was -126.43 Rosetta Energy Units (REU), while average scores for variants with two, four, and six flexible residues were -92.43, -78.04, and -67.17 REU, respectively. During the simulations with introduced sidechain flexibility, the protein opens for the ligand transport at the cost of introducing clashes. Particularly, the run with six flexible residues resulted in unrealistic protein conformation. The implementation of the flexibility is still under development and the users are advised to use rigid simulations or include flexibility to just a few (1-4) residues. Improved treatment of protein flexibility will be addressed in the next version of CaverDock.



SI-15: Lower-bound energy profiles from the iterative flexibility. Each profile represents a different number of flexible sidechains.



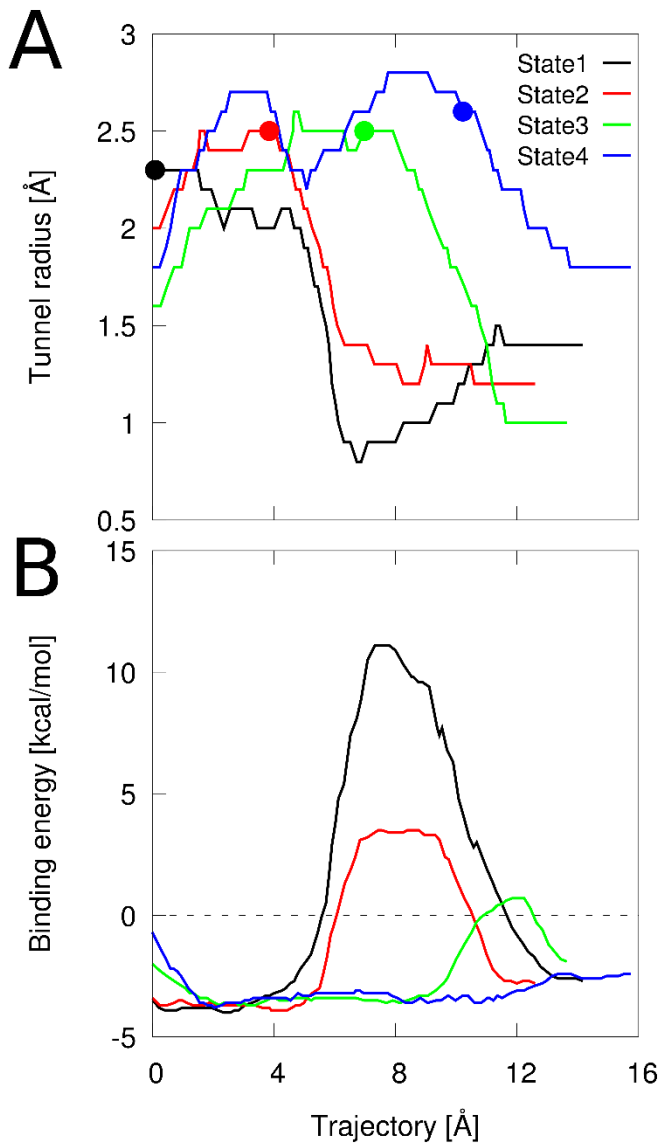
SI-16: Rosetta scores of the ligand-free protein structure for all snapshots from the iterative flexibility simulations.

SI-17: Analysis of backbone dynamics

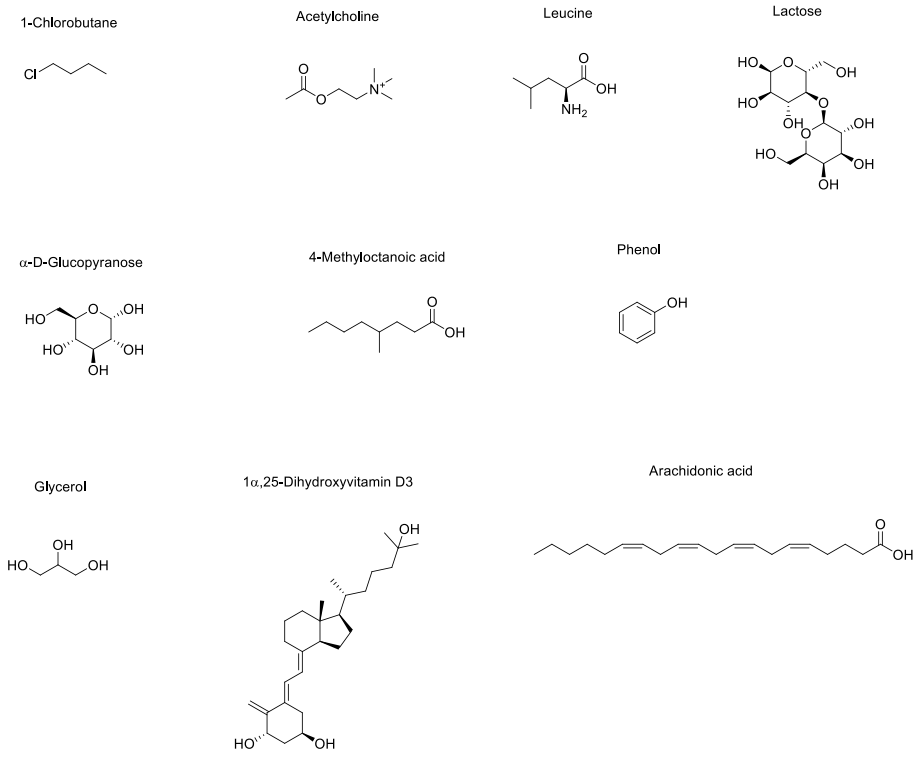
We analysed the effects of the presence of a ligand in a tunnel on the protein's structure during unbinding. Four snapshots from an accelerated molecular dynamics simulation with the catalytic product 2,3-dichloropropan-1-ol (DCP) at different positions in the p1 tunnel of haloalkane dehalogenase mutant DhaA31 (PDB ID: 3RK4) were used for CaverDock calculation. The accelerated molecular dynamics is an enhanced sampling method and was used to promote the release of DCP in reasonable simulation time (Marques *et al.*, 2017). The release of the product DCP is known to be a rate-limiting step in the enzymatic reaction of DhaA31 enzyme. The original 200 ns of accelerated molecular dynamics simulation was carried out with the PMEMD.CUDA (Salomon-Ferrer *et al.*, 2013; Le Grand *et al.*, 2013) module of AMBER 12 (Case *et al.*, 2012).

Our results show that the tunnel is flexible and deforms to accommodate the ligand as it moves through the protein by turning aside residues that contribute to the bottleneck regions in the tunnel (SI-18). We clearly see that the dynamics of the protein is important during the ligand passage. Unfortunately, the current version of CaverDock can only analyse protein-ligand interactions; it cannot infer any information about the energy of the whole system. For example, in the last state of the release process, there is no barrier and the ligand can escape without any steric hindrance since that protein conformation displays the tunnel in a fully open state.

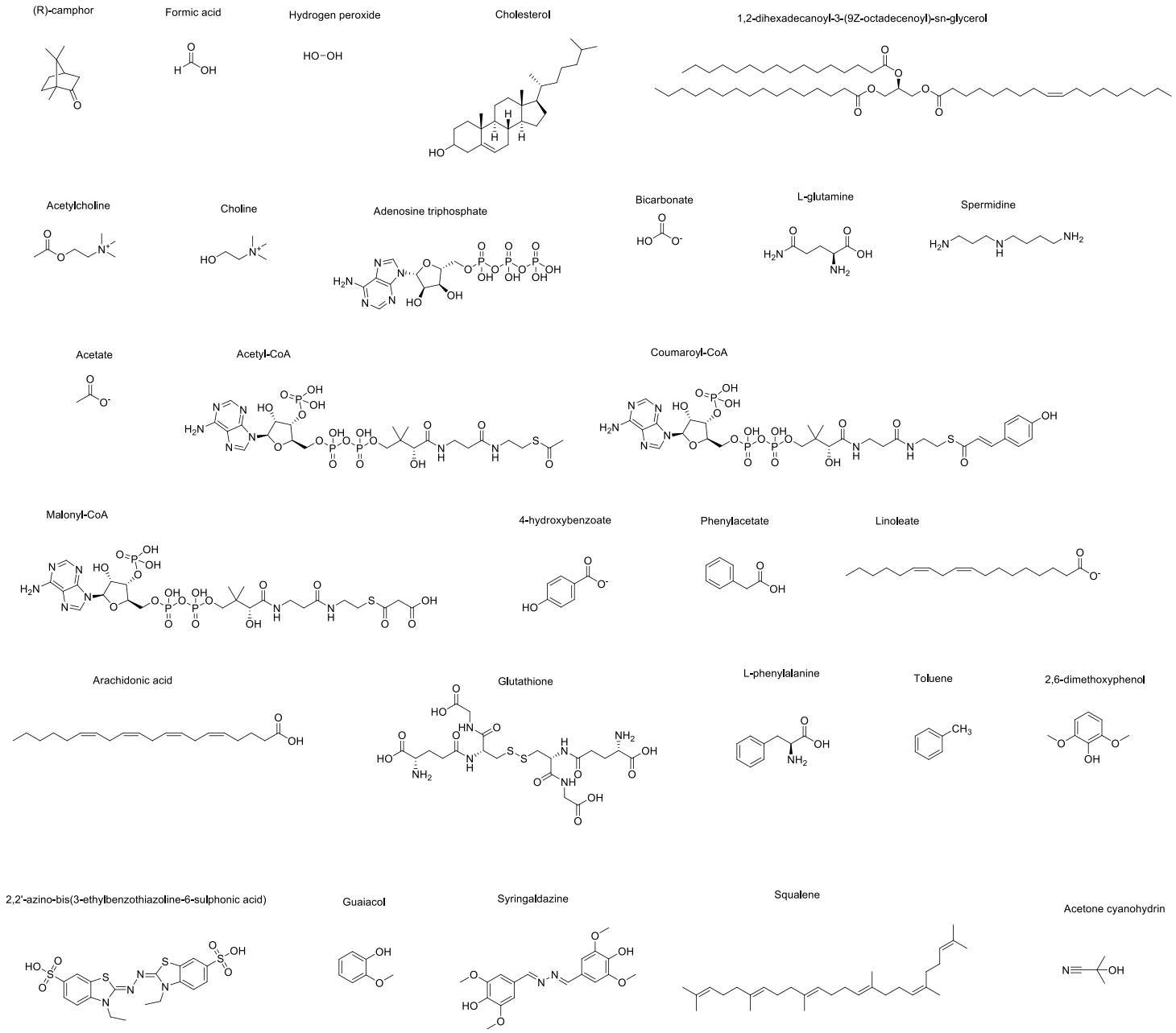
We tried to estimate how much the protein's energy has risen to accommodate this conformational state. We analysed the quality of the structures with the Rosetta scoring function (Rohl *et al.*, 2004). The score calculated for the crystal structure was -495.19 and 910.62, 1134.66, 1168.49, 1005.73 for the State1, State2, State3, State4 snapshots respectively. It is clear that the protein undergoes large conformation changes depending on the position of the ligand. In comparison with the flexibility of side chains, the changes in the backbone lead to the structures with even higher energy although these snapshots replicate the natural behaviour of the protein. This information is valuable for further development of CaverDock, which will require finding a reasonable way to introduce flexibility and establish a reasonable ratio between the accuracy and the speed of the calculation.



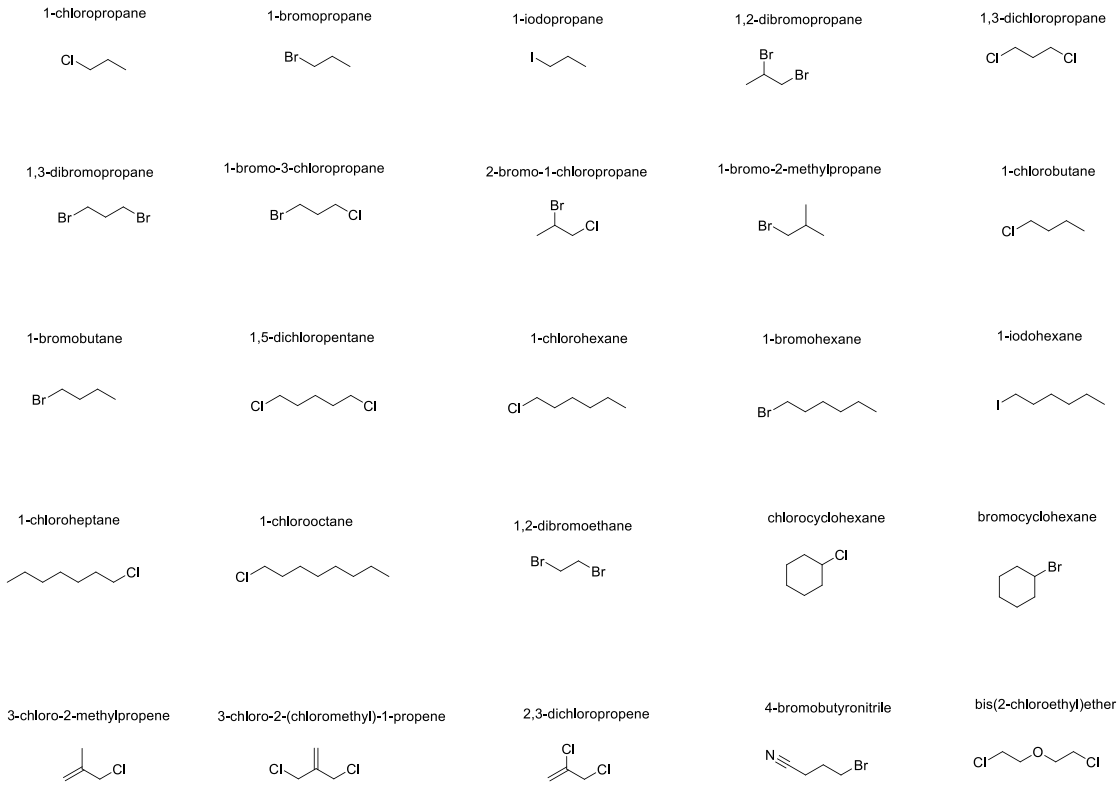
SI-18: CaverDock tunnel radii (A) and corresponding energy profiles (B) based on snapshots from accelerated molecular dynamics simulations. The simulate system is haloalkane dehalogenase DhaA31 interacting with the product 2,3-dichloropropan-1-ol. Each state refers to a position of the product during the accelerated molecular dynamics simulation (marked by a dot in the respective plot of the tunnel radii). This analysis illustrates that the tunnel gradually adapts to the passing ligand.



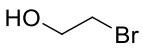
SI-19a: Molecular structures of the ligands used in Dataset I – Benchmarking.



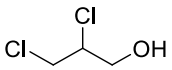
SI-19b: Molecular structures of the ligands used in Dataset II – Geometry of tunnels.



SI-19c: Molecular structures of the ligands used in Dataset III – Geometry of substrates.



SI-19d: Molecular structure of the 2-bromoethanol used in Dataset IV – Tunnel engineering.



SI-19e: Molecular structure of the 2,3-dichloropropan-1-ol used in the analysis of backbone dynamics.

SI-20: Compiled datasets used for validation of CaverDock. All data sets used for the validation of CaverDock are provided as compressed ZIP files.

References

Brezovsky,J. *et al.* (2016) Engineering a de Novo Transport Tunnel. *ACS Catal.*, **6**, 7597–7610.

Case,D. *et al.* (2012) AMBER 12, University of California, San Francisco. **79**, 926–935.

Le Grand,S. *et al.* (2013) SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.*, **184**, 374–380.

Marques,S.M. *et al.* (2017) Catalytic Cycle of Haloalkane Dehalogenases Toward Unnatural Substrates Explored by Computational Modeling. *J. Chem. Inf. Model.*, **57**, 1970–1989.

Rohl,C.A. *et al.* (2004) Protein Structure Prediction Using Rosetta. In *Methods in Enzymology*, pp. 66–93.

Salomon-Ferrer,R. *et al.* (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.