*Bioinformatics*, 35(23), 2019, 4986–4993
doi: 10.1093/bioinformatics/btz386
Advance Access Publication Date: 11 May 2019
Original Paper

OXFORD

Structural bioinformatics

# CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels

**Ondrej Vavra[1,2,‡], Jiri Filipovic[3,‡], Jan Plhak[3], David Bednar[1,2], Sergio M. Marques[1,2], Jan Brezovsky[1,†], Jan Stourac[1,2], Ludek Matyska[3] and Jiri Damborsky[1,2,*]**

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, Brno 625 00, Czech Republic, [2]International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic and [3]Institute of Computer Science, Masaryk University, Brno 602 00, Czech Republic

*To whom correspondence should be addressed.
[†]Present 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
[‡]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Yann Ponty
Received on December 31, 2018; revised on April 11, 2019; editorial decision on April 30, 2019; accepted on May 5, 2019

## Abstract

**Motivation:** Protein tunnels and channels are key transport pathways that allow ligands to pass between proteins' external and internal environments. These functionally important structural features warrant detailed attention. It is difficult to study the ligand binding and unbinding processes experimentally, while molecular dynamics simulations can be time-consuming and computationally demanding.
**Results:** CaverDock is a new software tool for analysing the ligand passage through the biomolecules. The method uses the optimized docking algorithm of AutoDock Vina for ligand placement docking and implements a parallel heuristic algorithm to search the space of possible trajectories. The duration of the simulations takes from minutes to a few hours. Here we describe the implementation of the method and demonstrate CaverDock's usability by: (i) comparison of the results with other available tools, (ii) determination of the robustness with large ensembles of ligands and (iii) the analysis and comparison of the ligand trajectories in engineered tunnels. Thorough testing confirms that CaverDock is applicable for the fast analysis of ligand binding and unbinding in fundamental enzymology and protein engineering.
**Availability and implementation:** User guide and binaries for Ubuntu are freely available for non-commercial use at https://loschmidt.chemi.muni.cz/caverdock/. The web implementation is available at https://loschmidt.chemi.muni.cz/caverweb/. The source code is available upon request.
**Contact:** jiri@chemi.muni.cz
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Proteins are macromolecules that have myriads of functions in cells and uses in the chemical, biotechnological and pharmaceutical industries (Clouthier and Pelletier, 2012; Koeller and Wong, 2001; Soetaert and Vandamme, 2006). The majority of enzymes have their active site buried inside their core, connected with the external environment by access tunnels. Protein *tunnels* are characterized by a single opening. They enable the transport of substrates, products, solvent, and ions in and out of the active site (Brezovsky *et al.*, 2013; Gora *et al.*, 2013). Tunnels are essential for the natural function of enzymes, affecting their substrate specificity, stability and activity (Gora *et al.*, 2013). The shape and physicochemical properties of the tunnels may also protect proteins' hydrophobic core by restricting the access of solvent molecules and inhibitors. Protein *channels* are characterized by two openings. They are often connecting different cellular environments and play an essential role in the transport of various ligands, solvent molecules and ions. The rational modification of protein tunnels and channels is an important paradigm in protein engineering (Damborsky and Brezovsky, 2009). Tunnel- or channel-lining residues directly interact with the passing ligands and therefore represent hot spots for the optimization of various enzymatic properties (Bendl *et al.*, 2016).

The process of ligand transport cannot be studied easily by experimental techniques at the molecular level. Characterization of the transport processes is usually carried out indirectly by evaluation of enzymatic activity by steady state or transient kinetics (Biedermannova *et al.*, 2012; Hu *et al.*, 2016). Experimental methods offering a direct molecular description of the access pathways, like crystallography under xenon pressure (Milani *et al.*, 2005; de Sanctis *et al.*, 2004; Tilton *et al.*, 1984) or time-resolved protein crystallography (Schmidt *et al.*, 2005; Schotte *et al.*, 2003), are still very demanding and can only be applied to a narrow spectrum of proteins. Therefore, computational approaches provide an important insight into the molecular transport. Many of these methods involve perturbed molecular dynamics (MD) simulations (Arroyo-Mañez *et al.*, 2011) and other enhanced sampling methods (Rydzewski and Nowak, 2017). Methods such as Protein Energy Landscape Exploration (Borrelli *et al.*, 2005), Binding Free Energy Landscape (Bai *et al.*, 2013) or IterTunnel (Kingsley and Lill, 2014) were developed to simplify the setup and assessment of MD-based simulations. Nevertheless, MD-based methods are still difficult to use for interactive analyses, comparative studies or virtual screening campaigns due to the long simulation times and high numbers of repetitions required.

To analyse the ligand unbinding more rapidly, without the need for such computationally demanding MD methods, two alternative tools have been previously developed. SLITHER (Lee *et al.*, 2009) uses an iterative docking scheme to generate protein–ligand complexes and calculates corresponding binding free energies. This tool focuses on the study of ligands passing through channels inside a protein. The computational core of this method is molecular docking using AUTODOCK or MEDock (Chang *et al.*, 2005; Morris *et al.*, 2009). MoMA-LigPath (Devaurs *et al.*, 2013) uses a steric representation of molecules and a robotic Manhattan-like RRT algorithm (Cortes *et al.*, 2007) to explore the conformational space, but does not evaluate the free energy of the system. Therefore, an external method must be applied to quantify energy changes that occur during protein–ligand interactions along the tunnel. Here we present a novel method for simulating ligand binding and unbinding, implemented in the software tool CaverDock. The software is based on the step-wise movement of the ligand along the pre-calculated tunnel. CaverDock uses the docking algorithm of AutoDock Vina (Trott and Olson, 2010) enriched by the restraints, which serve to: (i) hold a selected atom of a ligand at a specific disc located along the tunnel or channel, i.e. position restraint; and (ii) dock the ligand in the upper-bound vicinity of a previous ligand conformation in order to maintain continuous ligand movement along the tunnel, i.e. pattern restraint.

## 2 Materials and Methods

### 2.1 CaverDock

In this section, we introduce the basic principles of CaverDock computation. A more thorough methodology is described in detail in Supplementary Material S1. The complete mathematical and algorithmic description of the method, which is beyond the scope of this study, is provided in Filipovic *et al.* (2019). The method is based on the step-wise movement of the ligand along the tunnel. The tunnel geometry, approximated by a sequence of spheres, is used as an input. This sequence of spheres can be obtained from tools providing the PDB file of the tunnel represented by spheres, such as CAVER 3.02 (Chovancova *et al.*, 2012), for whose output file format CaverDock was optimized. The sequence of spheres is then discretized into a sequence of discs (cross-section slices of a maximal thickness set by the user).

First, the selected ligand's atom is positioned at the disc by a position restraint. Second, CaverDock minimizes the ligand conformation and evaluates its binding free energy by using the scoring function from AutoDock Vina (Trott and Olson, 2010). Third, the ligand trajectory is produced by aggregating the docked poses of the ligand at each consecutive disc. Such a trajectory samples the tunnel thoroughly, but the movement of the ligand may be non-continuous. This non-continuous (lower-bound) trajectory is used to estimate the lower-bound (lowest) energy profile of the ligand's transport through the tunnel. The actual energy may be higher since the non-continuous movement can avoid small bottlenecks by rapid changes in the orientation or the conformation of the ligand.

Finally, the pattern constraint is used to compute the continuous (upper-bound) trajectory. In each step, the ligand is docked in the vicinity of its previous position allowing only small changes in the ligand conformation. The number of possible continuous trajectories grows exponentially with the number of discs, because each transition to a new disc may lead to changes in the ligand's position, orientation and conformation. Therefore, a heuristic method is employed to search for a continuous trajectory. When the binding free energy of a given docked conformation is significantly higher than the binding free energy of the conformation obtained from the lower-bound trajectory, backtracking is turned on. The ligand conformation is changed (e.g. to a conformation explored when lower-bound trajectory was computed) and the ligand is moved successively backward to previous discs. The backtracking ends when the forward and backward trajectories converge, or it is stopped if the starting disc is reached. As there is no guarantee that the resulting continuous trajectory is optimal, we call it the upper-bound trajectory as the actual energy may be lower than the computed energy.

The practical differences between lower-bound and upper-bound trajectories are the following: the lower-bound trajectory is able to completely sample the ligand trajectory. The information from the lower bound is sufficient for comparison purposes but its main limitation is that it can miss small bottlenecks by rapid changes of the ligand orientation. However, the sudden changes in orientation could potentially mimic the natural flexibility of the protein and lower the

unnatural energy barriers caused by the receptor rigidity during binding or unbinding. On the other hand, the upper-bound trajectory is completely smooth. However, it can create unrealistic conformations in very tight parts of the tunnel, which are signified by sudden sharp peaks in the binding energy profile. The energy profile from the lower-bound calculation shows the best-case scenario of the binding energy along the tunnel, while the upper-bound can report exaggerated energies because the respective trajectory may not be optimal.

### 2.2 Input preparation

With one exception (for Dataset III, Section 2.3.3), all the simulations described in this manuscript were performed using the following settings: the PDB files of the proteins were obtained from the RCSB Protein Data Bank (Berman *et al.*, 2000). MOL2 files of the ligands were either downloaded from the ZINC database (Irwin and Shoichet, 2005) or built in Avogadro (Hanwell *et al.*, 2012) and minimized with the MMFF94 force field (Halgren, 1996). The receptor and ligand PDBQT files were prepared using scripts from MGLtools (Morris *et al.*, 2009) with the default parameters. The tunnel calculation was performed by CAVER 3.02 (Chovancova *et al.*, 2012), with the size of the probe set to 0.7 Å, and the other parameters at the default values. The tunnels were discretized with 0.3 Å steps and extended by 2 Å in the direction of the vector calculated from the last two spheres in the original tunnel. The script for the tunnel extension is provided in the CaverDock package. The configuration files and calculation of the grid box containing the whole tunnel geometry were prepared by the provided preparation script. The dragged atom, i.e. the atom attracted to the middle of the disc at the beginning of each calculation step, was chosen using the default auto-selection (the closest atom to the centroid of the molecule). All the simulations were performed in the default (unbinding) direction. To simulate the binding process, the user has to invert the discretized tunnel file, e.g. using the bash command *tac*. Side-chain flexibility of selected residues can be prepared using MGLtools. Detailed information about the CaverDock setup is provided in the manual available at https://loschmidt.chemi.muni.cz/caverdock/.

### 2.3 Testing datasets

All datasets (Supplementary Material S20) together with the figures of the ligand's geometries (Supplementary Material S19) are provided as the Supplementary Material.

#### 2.3.1 Dataset I: benchmarking

CaverDock was compared with the two existing tools for prediction of the ligand passage SLITHER (Lee *et al.*, 2009) and MoMA-LigPath (Devaurs *et al.*, 2013). These tools were compared using 10 cases. The dataset consists of six example cases presented at the websites of SLITHER and MoMA-LigPath, complemented with other systems found in the literature (Cui *et al.*, 2015; Koudelakova *et al.*, 2011; Peräkylä, 2009; Wang *et al.*, 2005) (Supplementary Material S2). SLITHER and MoMA-LigPath are available as web servers. SLITHER calculations were conducted with the AUTODOCK algorithm and the default rigid receptor. MoMA-LigPath was used with the default settings. The side chains are treated as flexible by default only in the case of MoMA-LigPath. CaverDock was used with a rigid receptor and the calculations were set up in the same manner as described in the Section 2.2.

#### 2.3.2 Dataset II: geometry of tunnels

This dataset was used to test the ability of CaverDock to model ligand trajectories through tunnels with a broad range of geometries. The data for proteins and their corresponding tunnels were collected from the literature (Chovancova *et al.*, 2012; Koudelakova *et al.*, 2013). Information about the proteins' native substrates was obtained from the UniProt (The UniProt Consortium, 2017) and BRENDA databases (Schomburg *et al.*, 2004). The complete dataset consists of 26 proteins with 113 identified tunnels and 33 natural substrates, creating altogether 136 cases (Supplementary Material S3).

#### 2.3.3 Dataset III: geometry of substrates

The correspondence between the binding energies from CaverDock and experimentally measured kinetics data were validated using this dataset. The haloalkane dehalogenase LinB (PDB ID: 1K63) and the set of 25 halogenated substrates with experimentally determined $K_M$ values (Kmunícek *et al.*, 2005) were used (Supplementary Material S4). To ensure the complete unbinding of each ligand, we used specific settings to calculate the tunnels in CAVER 3.02, with the shell radius and shell depth set to 20 and 4 Å, respectively. We selected the tunnels corresponding to the p1 and p2 tunnels of the LinB dehalogenase and extended them by 20 Å.

#### 2.3.4 Dataset IV: tunnel engineering

This dataset was assembled to test the ability of CaverDock to describe the differences in enzymes with rationally engineered access tunnels (Supplementary Material S5). We analysed the wild-type dehalogenase LinBWT (PDB ID: 1K63), the variant LinB32 (PDB ID: 4WDQ) with closed main p1 tunnel (LinB-closed$^W$), and the variant LinB86 (PDB ID: 5LKA) with newly open p3 tunnel (LinB-open$^W$). The goal was to compare the energy profiles from CaverDock (i) with the tunnels detected in the crystal structures and (ii) the frequency of product (2-bromoethan-1-ol) release through p1 and p3 tunnels obtained in previously published MD simulations (Brezovsky *et al.*, 2016).

## 3 Results

### 3.1 Illustration of CaverDock output

CaverDock generates an output in the form of two PDBQT files. One file provides a smoothed upper-bound trajectory while the other represents the lower-bound trajectory of the ligand. Information about the binding energies and tunnel radii is listed in the REMARK lines of the respective ligand trajectories, and can be extracted and plotted using the scripts provided with the package. The visualization of the results obtained using the CaverDock is presented in Figure 1.

### 3.2 Comparison of CaverDock with state-of-the-art methods SLITHER and MoMA-LigPath

We studied the robustness of SLITHER, MoMA-LigPath and CaverDock using the Dataset I (Table 1). SLITHER was able to predict the unbinding trajectory for half of the tested systems. Its main limitation is that the ligands are moved in a direction parallel to the *y*-axis only, making the analysis of curved and narrow tunnels difficult. Moreover, the ligand trajectories calculated by SLITHER are discontinuous with significant gaps between the predicted ligand positions. MoMA-LigPath was more successful, providing a continuous trajectory for 6 of the 10 test cases, but the tool does not provide any energy information. CaverDock was the most robust and provided results for all 10 test cases.

A critical comparison of the features of individual tools revealed that the main advantage of CaverDock over SLITHER is its ability to calculate the ligand transport in any direction with a simple setup

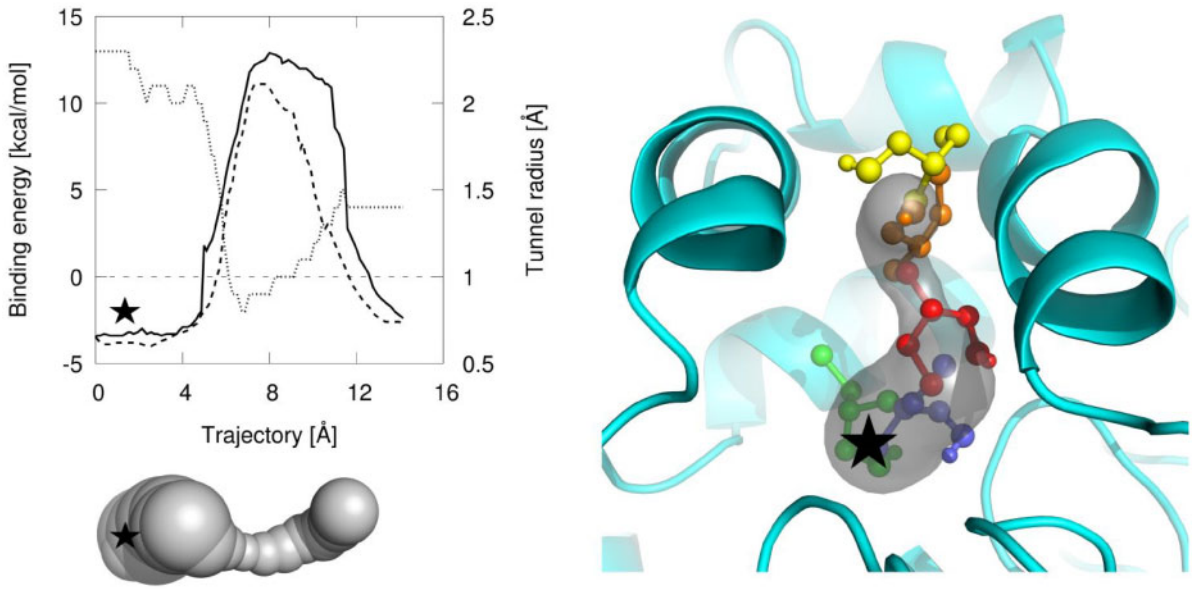


**Fig. 1.** Illustration of the results obtained using the CaverDock. Top left: Examples of the energy profiles for the haloalkane dehalogenase LinB (PDB ID: 1K63) and 2,3-dichloropropan-1-ol. The binding energy (left vertical axis) of a smoothed continuous upper-bound trajectory, the lower-bound trajectory and tunnel radius (right vertical axis) are indicated by the full line, dashed line and dotted line, respectively. The direction of the trajectory in the plot is from the active site (marked by the star symbol) to the surface of the protein. Bottom left: The three-dimensional surface of the corresponding tunnel calculated by CAVER 3.02. Right: Visualization of a part of a CaverDock trajectory. The protein is displayed as the cyan cartoon with the tunnel shown as the grey transparent surface. Selected snapshots of the ligand are shown in ball-and-stick representation: 1 (green), 10 (blue), 45 (red), 60 (orange) and 85 (yellow). The snapshot 45 (red) corresponds to the binding energy maximum of the energy profile

**Table 1.** Comparative study of CaverDock, SLITHER and MoMA-LigPath

| PDB ID | Protein | Ligand | Ligand passage | | |
|---|---|---|---|---|---|
| | | | CaverDock | SLITHER | MOMA-LigPath |
| 1BN7 | Haloalkane dehalogenase | 1-Chlorobutane | Yes | Yes | Yes |
| 1MAH | Acetylcholinesterase | Acetylcholine | Yes | Yes | Yes |
| 2A65 | Leucine transporter | Leucine | Yes | Yes | Yes |
| 1PV7 | Lactose permease | Lactose | Yes | Yes | Yes |
| 1SUK | Glucose transporter | α-D-Glucopyranose | Yes | Yes | No |
| 1TCC | Lipase B | 4-Methyloctanoic acid | Yes | No | Yes |
| 1ZNJ | Insulin hexamer | Phenol | Yes | No | Yes |
| 1RC2 | Aquaporin Z | Glycerol | Yes | No | No |
| 1IE9 | Vitamin D receptor | 1, 25-Dihydroxyvitamin D3 | Yes | No | No |
| 3LC4 | Cytochrome P450 2E1 | Arachidonic acid | Yes | No | No |

*Note*: Yes and No describes the result of the qualitative test whether the tool was able to predict a ligand's trajectory.

(Supplementary Material S6). The resolution of CaverDock trajectories is much higher since the ligand has to move through each disc of the discretized tunnel or channel so there are no large gaps in the trajectory. On the other hand, CaverDock currently cannot analyse multiple protein conformations simultaneously as it is possible with the relaxed receptor mode of SLITHER. The main differences between CaverDock and MoMA-LigPath are that CaverDock is able to simulate also the binding trajectory of a ligand and gives the information about the binding energy along the pathway. The advantage of MoMA-LigPath is that it can treat the flexibility of many side chains simultaneously, while implementation of the side-chain flexibility is still rather limited in CaverDock and SLITHER. Finally, CaverDock is the only software which is provided as a web application as well as a standalone tool, making it suitable for extensive virtual screening campaigns. The features and the setup options of the tested tools are summarized in Supplementary Material S6.

### 3.3 Impact of tunnel geometries on CaverDock calculations

Dataset II was constructed to test the predictive power of CaverDock with proteins possessing various geometries of tunnels with their native substrates. CaverDock was tested on 26 proteins with 33 substrates (some proteins had more than one native substrate) and 113 tunnels, 136 calculations altogether. Out of 136 CaverDock runs, 81 finished with lower-bound and upper-bound trajectories, 44 finished only with lower-bound and in 11 cases the ligands were not able to pass through the tunnels (Supplementary Material S7).

Although a smoothed (upper-bound) trajectory was not calculated for almost half of the cases, this does not mean that CaverDock could not properly simulate the ligand unbinding. The ligand unbinding process was still sufficiently sampled along the whole tunnel in the lower-bound trajectory, although the transition

from one conformation to the next was not smooth. This assertion is corroborated by the manifestation of increases in energy caused by the tunnel bottlenecks in the lower-bound energy profiles alone. Therefore, providing data for the lower-bound trajectory alone is a valid result. Further analysis of the 11 cases in which the ligands could not pass through the tunnels in CaverDock simulations revealed that the failure was due to the tunnels being too narrow for the ligands. In all except one case, these tunnels were graded by CAVER 3.02 as being 'lower throughput', meaning they are apparently less important than others for transport and thus unlikely to be functionally relevant for transport of the respective ligands. Plots of the energy profiles can be found in Supplementary Material S8.

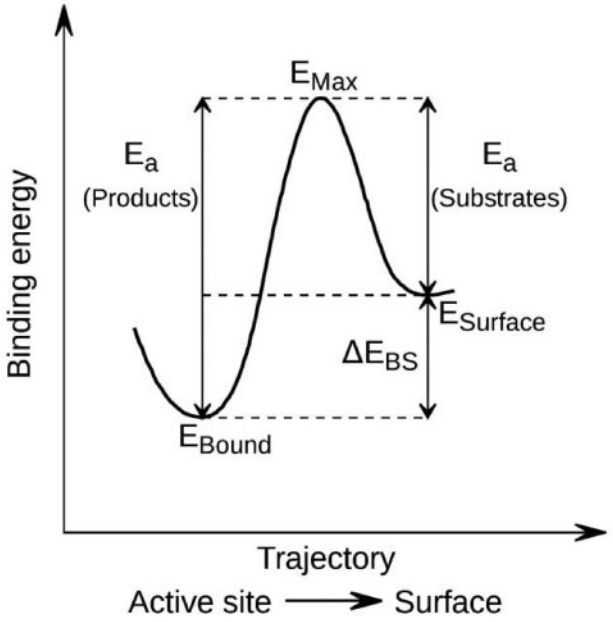


**Fig. 2.** Schematic energy profile with marked energy values. $E_{Bound}$, the binding energy of a ligand located inside the active site; $E_{Max}$, the highest binding energy in the trajectory; $E_{Surface}$, the binding energy of the ligand located at the protein surface; $E_a$, the activation energy of association for the products ($E_{Max} - E_{Bound}$) and for the reactants ($E_{Max} - E_{Surface}$), corresponds to the kinetics of a ligand passing through the tunnel; $\Delta E_{BS}$, difference of the binding energies in the bound state and at the surface corresponds to the enthalpy of binding, and is related to the equilibrium constant

### 3.4 Validation of CaverDock calculations against experimental data

CaverDock was used to analyse the p1 and p2 tunnels of the haloalkane dehalogenase LinB with the set of halogenated substrates from Dataset III, for which the values of Michaelis constants have been determined experimentally in our laboratory (Kmunícek *et al.*, 2005). The impact of the ligand and tunnel geometry on the energy profiles was studied. Selected energy values (Fig. 2) were extracted from the energy profiles: (i) the energy minimum close to the start of the trajectory corresponding to the ligand bound into the active site ($E_{Bound}$), (ii) the maximum energy from the profile ($E_{Max}$) and (iii) the last minimum related with the surface-bound ligand ($E_{Surface}$). For the analysis of substrates, the main focus was devoted to the evaluation of the height of the activation energy of association ($E_a$) calculated for the ligands going through the tunnel into the active site. Therefore, the activation energy of association was calculated as $E_a = E_{Max} - E_{Surface}$. $E_a$ can be related to the binding kinetics by the Arrhenius law ($k = Ae^{-\frac{E_a}{RT}}$), and thus $E_a$ is expected to vary linearly with the logarithm of the association rate, $k_{on}$. The energy difference between the active site and the surface-bound energy ($\Delta E_{BS}$) was calculated as the difference between the corresponding minima. $\Delta E_{BS}$ quantifies the enthalpy of binding, which is, according to the van't Hoff equation, negatively correlated with the logarithm of the equilibrium constant. Even though CaverDock provided the smoothed trajectories for all the test cases, we analysed the binding energies from the lower-bound trajectories, which provide more reasonable profiles.

Comparison of $E_a$ values for the two tunnels (Fig. 3) indicates that the energy barriers for the ligand passage through the p2 tunnel

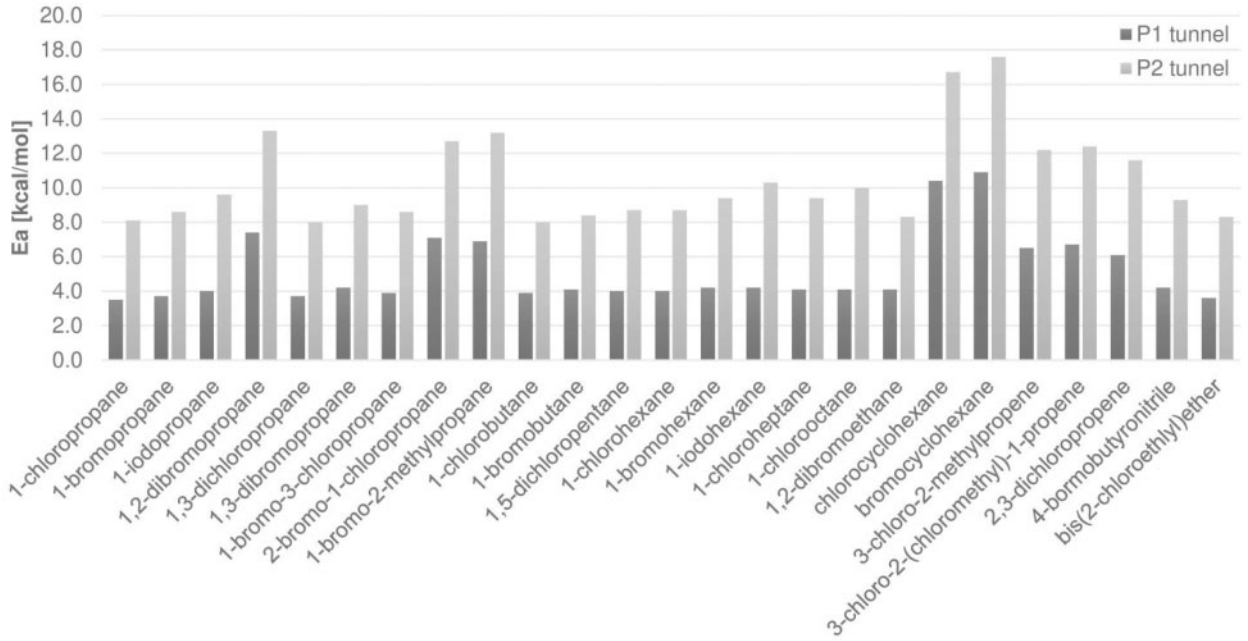


**Fig. 3.** Comparison of the activation energy of association $E_a$ for the p1 and p2 tunnels of the haloalkane dehalogenase LinB with the set of 25 substrates

are typically two times higher than the corresponding barriers for the p1 tunnel (Supplementary Material S9 and S10). This is likely due to the fact that p2 is narrower, longer and more curved than p1 (Supplementary Material S11). These results suggest the preference of the substrates for binding into the buried active site of that protein using the p1 tunnel, which is the main tunnel observed in the crystal structures. Without exception, the energy minima of the bound states were lower than the surface-bound minima, showing the tendency of all substrates to bind into the active site cavity rather than any other part of the tunnel or surface.

We calculated the correlations between the analysed energy values and the experimentally measured Michaelis constants ($K_M$) and catalytic constants ($k_{cat}$). With crude approximation, $K_M$ values should be related to the binding affinity, and it may be expected to correlate with $\Delta E_{BS}$. The interpretation of $K_M$ in various systems is complex, as it is composed of multiple steps in the enzymatic reaction, and not only composed of the binding process. Pearson's correlation coefficient of $\log(K_M)$ with $\Delta E_{BS}$ showed the values of 0.6 and 0.7 for the p1 and p2 tunnel, respectively. This statistically significant correlation shows that CaverDock can describe the binding trajectory and find a proper binding mode. The level of the observed correlation is in agreement with the nature of Michaelis constant for the haloalkane dehalogenase LinB, which is defined by a combination of the substrate binding and the rate of the follow-up $S_N2$ reaction step resulting in the covalently bound intermediate (Prokop *et al.*, 2003).

Regarding $k_{cat}$, this kinetic parameter is limited by the slowest step in the catalytic cycle. In the haloalkane dehalogenases, this cycle is rather complex, and the rate-limiting step can easily vary from substrate to a substrate (Prokop *et al.*, 2003). Therefore, $k_{cat}$ is expected to correlate with the $E_a$ barriers *only* in the systems where the binding of a substrate or unbinding of a product is the rate-limiting step of the catalysis. Pearson's correlation coefficient of $k_{cat}$ and $E_a$ is −0.2 for both p1 and p2 tunnels. These statistically insignificant correlations are in agreement with the transient kinetic analysis of the haloalkane dehalogenase LinB (Prokop *et al.*, 2003), demonstrating that the substrate binding is not the rate-limiting step in the catalytic cycle. The linear regressions of these correlations are provided in Supplementary Material S12.

### 3.5 Analysis of proteins with computationally designed access tunnels

We used CaverDock to analyse Dataset IV, the unbinding of the 2-bromoethan-1-ol product from three different LinB variants. The lower-bound energy profiles for the p1 and the p3 tunnels in LinB wild-type (LinBWT) and two variants carrying tunnel mutations LinB32 (LinB-closed$^W$) and LinB86 (LinB-open$^W$) are shown in Figure 4. The results from CaverDock calculation correspond well with the properties of the tunnels found in the crystal structures, supporting the blockage of the main p1 tunnel by the bulky Trp residue, intentionally introduced to the LinB32 and LinB86 variants (Brezovsky *et al.*, 2016). The narrowing of the p1 tunnel by this engineering step resulted in an increased energy barrier for the transport of the 2-bromoethan-1-ol from the active site to protein surface (Fig. 4A). The follow-up step of the project was opening *de novo* p3 tunnel in the protein LinB86. Calculation of the energy barriers for the release of 2-bromoethan-1-ol by this route clearly illustrates removal of the first barrier and significant lowering of the second barrier of the energetic profile (Fig. 4B), which is again in perfect agreement with crystallographic data. The calculated energy barriers are matching the diameters of p1 and p3 tunnels calculated in each

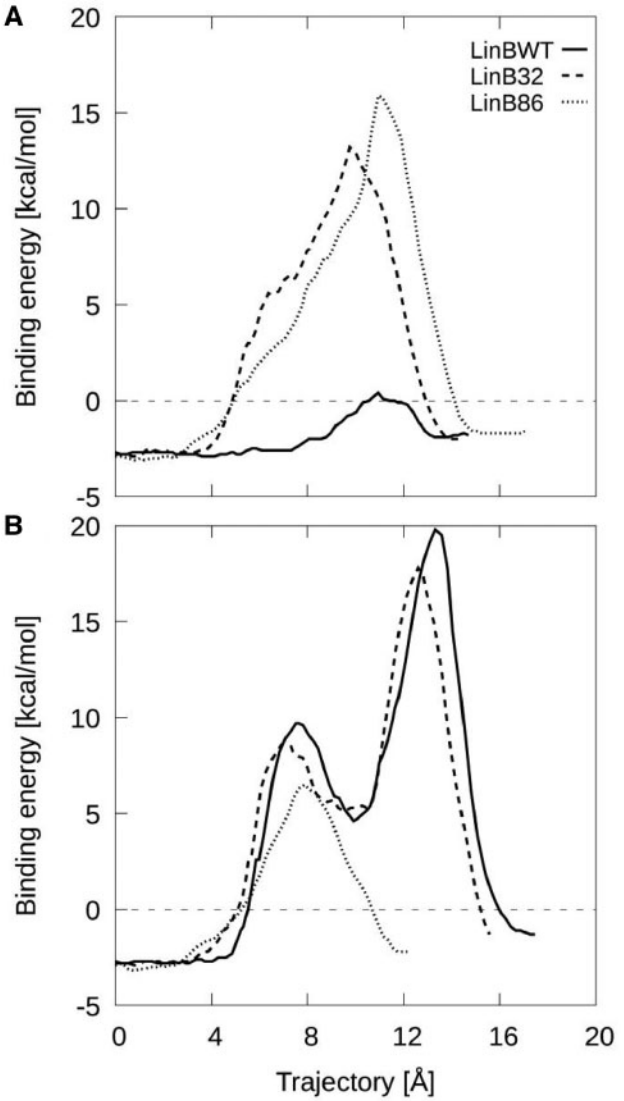


**Fig. 4.** Analysis of 2-bromoethan-1-ol unbinding through p1 (**A**) and p3 (**B**) tunnels in the LinB variants with rationally engineered tunnels (Brezovsky *et al.*, 2016). The p1 tunnel is blocked by a bulky Trp residue in LinB32 and LinB86, resulting in an increase in the energy barrier. The p3 tunnel was opened in LinB86 by three point mutations, resulting in removal of the first barrier and lowering the second barrier

of the experimental structures of LinBWT, LinB32 and LinB86 (Supplementary Material S13), suggesting that CaverDock calculations can reproduce tunnel engineering exercises and has a great potential for computational protein design targeting protein tunnels and channels.

## 4 Discussion

Tunnels and channels facilitate the transport of ligands through diverse proteins, so understanding the processes underlying the ligand transport is a cornerstone of biochemistry, structural biology and medicinal chemistry. The characteristics of these transport pathways are difficult to study using the currently available experimental techniques, which are not trivial to set up and are very time-consuming (Mittermaier and Meneses, 2013; Schotte *et al.*, 2003). Moreover, it can be difficult to study them with the currently available

computational tools, as they typically involve the use of MD simulations (Barducci *et al.*, 2010; Grubmüller *et al.*, 1996; Lüdemann *et al.*, 2000), which require substantial knowledge of the methods and too extensive computational resources for screening large number of ligands.

These limitations led us to develop CaverDock, a fast computational tool based on molecular docking for simulating ligand transition through protein tunnels and channels. CaverDock can be used to infer whether the studied ligand will likely pass through a particular protein tunnel. It can evaluate the passage of different ligands, or semi-quantitatively compare the difficulty of ligands passing through several different tunnels. The method is very easy to setup with the calculation times typically in the order of minutes. This makes it suitable for virtual screening purposes or for the enrichment of widely used virtual screening results by molecular docking (Daniel *et al.*, 2015). In comparison with MD simulations, CaverDock does not require extensive knowledge of the studied system. CaverDock is able to sample the binding energy throughout the whole protein tunnel and identify unfavourable binding interactions, which can then be optimized by site-directed mutagenesis (Kaushik *et al.*, 2018; Liskova *et al.*, 2017). Such places would be missed by traditional docking techniques. The easy setup and execution of the calculations may easily provide trajectories of the a ligand passage through a protein of interest, which can be used as educational materials in the biochemistry courses, assisting teachers with visualization of the process of ligand binding or unbinding. Finally, the advanced settings in CaverDock also enable constrained and pattern docking calculations.

The comparison presented here showed that, when aiming at exploring the properties of the ligand transport through molecular tunnels, CaverDock displayed better performance than the other tested tools SLITHER (Lee *et al.*, 2009) and MoMA-LigPath (Devaurs *et al.*, 2013). We thoroughly tested CaverDock using 69 ligands, 130 tunnel geometries and 40 protein structures. In most cases, the ligands successfully passed through the tunnels. In some cases, the steric hindrances prevented the calculation of a smoothed (upper-bound) continuous ligand trajectory. However, in these cases CaverDock was still able to calculate the non-continuous lower-bound trajectory. Further analysis of the lower-bound energy profiles showed that they reflect the increases in the energy associated with the ligands' passage through a more restricted and narrow spaces in a tunnel. Thus, the lower-bound trajectory alone is suitable for sampling all the binding energies through a tunnel. CaverDock's ability to calculate smoothed upper-bound trajectories could potentially be improved by choosing a dragged atom close to the edge of the ligand rather than the default atom closest to its centroid (especially for large ligands with high degrees of freedom). Another problem that may occur is that the energy at the end of the simulation (at the tunnel mouth) sometimes did not converge to zero. This implies that the ligand did not reach the fully unbound state. To ensure further unbinding of the ligand, the tunnel geometry obtained from CAVER may be prolonged or recalculated with different settings.

The most important limitation of the first version of CaverDock is that it cannot robustly address conformational dynamics of the protein structure. We have analysed the current implementation of flexibility of the sidechains (Trott and Olson, 2010) in CaverDock (Supplementary Material S14). The application of flexibility brought an overall lowering of the energy profile but at the same time it produced unlikely high-energy conformations of the protein structure in some instances (Supplementary Material S15, Supplementary Material S16). Moreover, the introduction of multiple side-chain flexibility significantly increased the calculation time. For now, we advise users to use the current implementation of sidechain flexibility cautiously and take practical measures such as minimizing the number of flexible sidechains and checking the generated protein conformations for steric clashes. We also looked at the importance of backbone dynamics (Supplementary Material S17). Using the snapshots from previously published accelerated molecular dynamics simulations, we have observed expected changes in the CaverDock energy profiles calculated with the structures of proteins possessing different conformations (Supplementary Material S18). These structures represent highly valuable benchmark for the rigorous treatment of protein flexibility, which is currently under development. The most important part in the development will be to balance the trade-off between the systematic description of protein conformations and the speed of CaverDock calculations.

## Acknowledgements

The authors thank Antonin Pavelka (Masaryk University) for stimulating discussions during the initial phase of the project.

## Funding

For financial support, the authors thank the Czech Ministry of Education [LM2015047, LM2015055, LM2015042, LM2015085, CZ.02.1.01/0.0/0.0/16_026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868, CZ.02.1.01/0.0/0.0/16_013/0001761] and European Commission [720776, 722610 and 814418]. Computational resources were provided by the CESNET and CERIT Scientific Cloud (LM2015042 and LM2015085). O.V. is the recipient of a Ph.D. Talent award provided by Brno City Municipality.

*Conflict of Interest*: none declared.

## References

Arroyo-Mañez,P. *et al.* (2011) Protein dynamics and ligand migration interplay as studied by computer simulation. *Biochim. Biophys. Acta*, **1814**, 1054–1064.

Bai,F. *et al.* (2013) Free energy landscape for the binding process of Huperzine A to acetylcholinesterase. *Proc. Natl. Acad. Sci. USA*, **110**, 4273–4278.

Barducci,A. *et al.* (2010) Linking well-tempered metadynamics simulations with experiments. *Biophys. J.*, **98**, L44–6.

Bendl,J. *et al.* (2016) HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.*, **44**, W479–W487.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Biedermannova,L. *et al.* (2012) A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB. *J. Biol. Chem.*, **287**, 29062–29074.

Borrelli,K.W. *et al.* (2005) PELE: protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.*, **1**, 1304–1311.

Brezovsky,J. *et al.* (2013) Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol. Adv.*, **31**, 38–49.

Brezovsky,J. *et al.* (2016) Engineering a de novo transport tunnel. *ACS Catal.*, **6**, 7597–7610.

Chang,D.T.-H. *et al.* (2005) MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.*, **33**, W233–W238.

Chovancova,E. *et al.* (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8**, e1002708.

Clouthier,C.M. and Pelletier,J.N. (2012) Expanding the organic toolbox: a guide to integrating biocatalysis in synthesis. *Chem. Soc. Rev.*, **41**, 1585.

Cortes,J. *et al.* (2007) Molecular disassembly with Rrt-like algorithms. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation, Rome, Italy*. IEEE, pp. 3301–3306.

Cui,Y.-L. *et al*. (2015) Molecular basis of the recognition of arachidonic acid by cytochrome P450 2E1 along major access tunnel. *Biopolymers*, **103**, 53–66.

Damborsky,J. and Brezovsky,J. (2009) Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol*., **13**, 26–34.

Daniel,L. *et al*. (2015) Mechanism-based discovery of novel substrates of haloalkane dehalogenases using *in silico* screening. *J. Chem. Inf. Model*., **55**, 54–62.

Devaurs,D. *et al*. (2013) MoMA-LigPath: a web server to simulate protein-ligand unbinding. *Nucleic Acids Res*., **41**, W297–W302.

Filipovic,J. *et al*. (2019) CaverDock: a novel method for the fast analysis of ligand transport. *IEEE/ACM Trans. Comput. Biol. Bioinform*., 1–1. doi: 10.1109/TCBB.2019.2907492.

Gora,A. *et al*. (2013) Gates of enzymes. *Chem. Rev*., **113**, 5871–5923.

Grubmüller,H. *et al*. (1996) Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science*, **271**, 997–999.

Halgren,T.A. (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem*., **17**, 490–519.

Hanwell,M.D. *et al*. (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform*., **4**, 17.

Hu,H. *et al*. (2016) Transient kinetics define a complete kinetic model for protein arginine methyltransferase 1. *J. Biol. Chem*., **291**, 26722–26738.

Irwin, J.J. and Shoichet, B.K. (2005) ZINC: a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*., **45**, 177–182.

Kaushik,S. *et al*. (2018) Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *Febs J*., **285**, 1456–1476.

Kingsley,L.J. and Lill,M.A. (2014) Including ligand-induced protein flexibility into protein tunnel prediction. *J. Comput. Chem*., **35**, 1748–1756.

Kmunícek,J. *et al*. (2005) Quantitative analysis of substrate specificity of haloalkane dehalogenase LinB from *Sphingomonas paucimobilis* UT26. *Biochemistry*, **44**, 3390–3401.

Koeller,K.M. and Wong,C.-H. (2001) Enzymes for chemical synthesis. *Nature*, **409**, 232–240.

Koudelakova,T. *et al*. (2011) Substrate specificity of haloalkane dehalogenases. *Biochem. J*., **435**, 345–354.

Koudelakova,T. *et al*. (2013) Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. *Angew. Chem. Int. Ed*., **52**, 1959–1963.

Lee,P.-H. *et al*. (2009) SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters. *Nucleic Acids Res*., **37**, W559–W564.

Liskova,V. *et al*. (2017) Different structural origins of the enantioselectivity of haloalkane dehalogenases toward linear $\beta$-haloalkanes: open-solvated versus occluded-desolvated active sites. *Angew. Chem. Int. Ed*., **56**, 4719–4723.

Lüdemann,S.K. *et al*. (2000) How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol*., **303**, 797–811.

Milani,M. *et al*. (2005) Structural bases for heme binding and diatomic ligand recognition in truncated hemoglobins. *J. Inorg. Biochem*., **99**, 97–109.

Mittermaier,A. and Meneses,E. (2013) Analyzing protein–ligand interactions by dynamic NMR spectroscopy. In: John,W. (ed.) *Methods in Molecular Biology*, Springer, Clifton, NJ, pp. 243–266.

Morris,G.M. *et al*. (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem*., **30**, 2785–2791.

Peräkylä,M. (2009) Ligand unbinding pathways from the vitamin D receptor studied by molecular dynamics simulations. *Eur. Biophys. J*., **38**, 185–198.

Prokop,Z. *et al*. (2003) Catalytic mechanism of the maloalkane dehalogenase LinB from *Sphingomonas paucimobilis* UT26. *J. Biol. Chem*., **278**, 45094–45100.

Rydzewski,J. and Nowak,W. (2017) Ligand diffusion in proteins via enhanced sampling in molecular dynamics. *Phys. Life Rev*., **22–23**, 58–74.

de Sanctis,D. *et al*. (2004) Crystal structure of cytoglobin: the fourth globin type discovered in man displays heme hexa-coordination. *J. Mol. Biol*., **336**, 917–927.

Schmidt,M. *et al*. (2005) Ligand migration pathway and protein dynamics in myoglobin: a time-resolved crystallographic study on L29W MbCO. *Proc. Natl. Acad. Sci. USA*, **102**, 11704–11709.

Schomburg,I. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*., **32**, 431D–4433.

Schotte,F. *et al*. (2003) Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science*, **300**, 1944–1947.

Soetaert,W. and Vandamme,E. (2006) The impact of industrial biotechnology. *Biotechnol. J*., **1**, 756–769.

Tilton,R.F. *et al*. (1984) Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 A. *Biochemistry*, **23**, 2849–2857.

Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem*., **31**, 455–461.

The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, **45**, D158–D169.

Wang,Y. *et al*. (2005) What makes an aquaporin a glycerol channel? A comparative study of AqpZ and GlpF. *Structure*, **13**, 1107–1118.