On the estimation of the molecular inaccessible volume and the molecular accessible surface of a ligand in protein ligand systems

Konstantinos Konstantinidis[1], Ioannis Karakasiliotis[1]. Kostas Anagnostopoulos[2] ,

*Georgios C.Boulougouris[3]**

[1] Laboratory of Biology, Department of Medicine, Democritus University of Thrace, Alexandroupolis, Greece

[2] Laboratory of Biochemistry, Department of Medicine, Democritus University of Thrace, Alexandroupolis, Greece

[3] Laboratory of Computational Physical Chemistry, Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Greece

KEYWORDS: Drug Design, Modeling , Protein Ligand Binding, Accessible surface, Molecular Simulation, Free energy.

** corresponding author: gbouloug@mbg.duth.gr*

Abstract

The increasing availability of computational resources over the last years has significantly facilitated computational studies which model protein ligand interactions in the atomistic level. In many of those *in silico* studies, the process of trying to "fit" a ligand in a protein cavity is an important first step, especially in the expanding field of computer aided drug design (CADD). In this work, a novel approach is proposed based on the accurate computation of a protein's inaccessible volume and the corresponding surface area as regards to a ligand, , where the ligand can be placed so that it "touches" the protein without any overlaps. The proposed approach can be thought as an extension of the widely used concept of the Solvent-Accessible Surface Area (SASA), evaluating the surface generated by the ligand while being rolled over all the atoms of the protein without penetrating them. Identification of the inaccessible volume of each candidate protein-ligand pair is also provided in the context of this study, along with the boundary surface where the ligand can be placed so as to be in "contact" with the protein, which is expected to significantly enhance the ability to investigate specific protein drug interactions. Several trials have been conducted upon implementation of the proposed algorithm using the analytical method of Dodd and Theodorou leading to accurate volume and surface area measurements of an arbitrary set of fused spheres in test systems of various scales.

Introduction

Contributing to the cell's structure, metabolism, cycle, communication or response to stimuli, molecular interactions lie in the core of all fundamental biological processes. The scientific community has put a great effort in investigating such interactions, especially among molecules like proteins, also known as protein-protein interactions (PPIs), or between small molecules (ligands) and proteins, highlighting computational methods apart from experimental techniques (e.g. X-ray crystallography, NMR). Docking computation is considered a significant approach for the study of protein-protein or protein-ligand interactions, guided by several theories behind binding phenomena, such as the "lock-key" model[1], the "induced-fit" theory[2], the "conformational selection" mechanism[3] and similar established approaches. Development of structure-based virtual screening and construction of novel therapeutic agents via computer-aided drug design (CADD) have all been achieved with molecular docking software applications[4]. The algorithms implemented in molecular simulations – docking software are intended to predict the structure via conformation ensemble (sampling algorithms). They can also predict the binding affinity of the tested biomolecules during the interactions by scoring functions (scoring algorithms) under certain docking methodologies as shown in **Table 1**. These algorithms rely on a variety of theoretical, chemical and geometrical approaches to visualize molecular structures and processes. Interactions are handled based on the properties of the amino acid residues found on the surfaces of the molecules. Examination of amino acid charge, polarity, shape, potential for intercalation with other molecules, high evolutionary conservation of surface amino acids and the estimated energy of molecular interactions, constitute primary elements for the functional interpretation and calculation of molecular surfaces. Molecular surfaces may have a dual use; heir graphical representations can provide a prediction of the possible function and interactions which may take place by

visualizing the shape, electron distribution or evolutionary conservation of molecular surface sequences. Moreover, quantification of surfaces is mainly used as a descriptor in an attempt to quantify the binding Gibbs free energy.

Table 1 Molecular docking software classified by implemented algorithm nature and docking methodology

Sampling algorithms categories	**Software - Programs**
Matching algorithms	DOCK[5], FLOG[6], LibDock[7], SANDOCK[8]
Incremental construction methods	DOCK 4.0[9], FlexX[10], Hammerhead[11], SLIDE[12], eHiTS[13]
Monte Carlo	Early versions of AutoDock[14], ICM[15], QXP[16]
Genetic algorithms	AutoDock[17], GOLD[18], DIVALI[19], DARWIN[20]
Scoring functions categories	**Software - Programs**
Force-field-based	DOCK[5,21,22], GOLD[18], AutoDock[17]
Empirical	LUDI[23], PLP[24,25,26], ChemScore[27]
Knowledge-based	PMF[28], DrugScore[29], SMoG[30], Bleep[31]
Consensus scoring	CScore[32] (combination of DOCK, ChemScore, PMF, GOLD, and FlexX)
Docking methodologies	**Software - Programs**
Rigid ligand - Rigid receptor	DOCK[5,21,22], FLOG[6], FTDOCK[33]
Flexible ligand - Rigid receptor	AutoDock[17], FlexX[10]
Flexible ligand - Flexible receptor	Glide[34], IFREDA[35], QXP[16]

Calculating accessible molecular surfaces is of high significance to docking methodologies. The concept of the accessible surface area was firstly described by Lee & Richards[36] in 1971 as

Solvent-Accessible Surface Area (SASA). SASA traces the geometrical locus derived from the centre of a hypothetical probe-sphere rolling on the van der Waals surface of the molecule without penetrating its atoms. It is also equivalent to the van der Waals surface, with the difference that the atomic radii r_i have been substituted with the total of r_i+r_p (r_p equal to the atomic radius of the hypothetical probe-sphere, typically 1.4Å). Various approaches have been developed for calculating accessible surface areas, with the "rolling ball" algorithm by Shrake-Rupley[37] being one of the earliest and most popular methods among others. Additional improvements to these methods delimited the solvent-excluded surface (SES) or widely known molecular/Connolly surface[38,39,40,41], which consists of two segments. The first is the contact surface, (part of the van der Waals surface of the atoms), tangent to the hypothetical "rolled-over" probe sphere. Thee second is the reentrant surface, which comprises the inward-facing surface of the probe sphere when it is simultaneously tangent to two or more atoms. Analytical calculation of Connolly surfaces is founded upon numerical algorithms retrieving data from the atomic coordinates, van der Waals radii and the probe radius, thus generating finite sets of points constructing a network of convex, saddle-shaped and concave faces defined in terms of vertices, circular arcs, spheres and tori so as to compute the solvent-excluded surface.

Theoretical Basis

Extending the typical approaches for calculating accessible surface area and volume confined to the use of probe-spheres, this paper proposes a novel approach based on the analytical calculation of the accessible volume-area to a hard-sphere poly-atomic molecule. The core idea has been developed originally as part of the staged particle deletion (SPD) method[42,43], in an effort to accurately estimate the free energy of cavity formation and its contribution to the chemical po-

tential of small molecules in molecular simulations. According to the SPD method, an intermediate state is introduced from the initial N-molecule system, which consists of N-1 molecules and an inserted hard-sphere molecule of given complexity. Generally, the accessible volume of a molecule is equal to the difference of the total volume of the system minus the excluded volume. When inserting a mono atomic molecule that is modeled as a single hard-sphere into a system composed of atoms (represented also as hard spheres), the excluded volume limns the geometric locus of points where a hypothetical insertion hard-sphere center would cause overlap with any of the existing hard spheres in the system. More specifically, this geometric locus is a set of fused spheres, whose centers coincide with these of the spheres of the atoms in the system but with radii augmented by the radius of the inserted hard-sphere. Provided that a single sphere is inserted and the system consists of molecules made of atoms modeled as spheres, the accessible volume calculation can formally be mapped to evaluation of the volume of fused spheres, even when periodic boundary conditions are implemented. This approach works irrespective of the presence of inter-molecular connectivity, whereas the computational task is expected to depend mainly on the number of atoms in the system and the actual size of the spheres. Furthermore, the estimation of the accessible volume can become computationally quite demanding as the size of the system increases and the actual accessible volume starts to diminish , including the case of inserting a mono atomic molecule. On the other hand, what may not be so straight forward is the ability to estimate the inaccessible and in extension the accessible volume after inserton of an arbitrary polyatomic molecule in a similar manner[43]. Upon rationalizing the process, a possible solution is to consider the interaction of each sphere in the system with the inserted polyatomic molecule under fixed internal degrees of freedom. It turns out[43] that the volume of the loci of points where a trial insertion of the chain molecule will result in overlap can also be estimated as

the volume of a set of fused spheres. This is a problem that can be handled efficiently by the method of Dodd and Theodorou[44]. In order to express the problem of inserting a chain hard molecule as a problem of fused spheres the following steps can be followed:

- For each atom in the system enumerate a number of fused spheres equal to the number of atoms in the inserted molecule.
- Given the conformation of the inserted molecule with fixed internal degrees of freedom, an auxiliary sphere is inserted in the system for each atom in the inserted molecule.
- For each system atom, the center of the first auxiliary sphere coincides with the center of the respective atom
- The rest of the auxiliary spheres are displaced one by one, applying the negative or antiparallel corresponding bond vector on the inserted molecule.

As a result, the set of auxiliary spheres for each inserted molecule and every atom in the system constitutes a parallel translated "mirror" image of the inserted molecule. As in SASA, the radius of each sphere is the original radius augmented by the radius of the corresponding atom in the inserted molecule. Therefore, the problem of estimating the molecular inaccessible volume and accessible surface area has now been expressed as a problem of evaluating the volume and surface of a set of fused spheres. On the negative side, the number of fused spheres that one has to consider is now equal to the number of atoms in the system times the number of atoms of the inserted molecule.

In **Figure 1**, a graphical representation of the basic concept is depicted, referring to a system of two water molecules, with the former acting as the protein molecule of the system and the latter as the inserted ligand molecule. The method is founded on creation of multiple images of the inserted atoms by maintaining the internal degrees of freedom and relative orientation. The algo-

rithm generates 9 (9=3 protein atoms x 3 ligand atoms) auxiliary spheres (six of them depicted in pink plus the three that are placed at the same position as the molecule in the system). The gray 3D surface created by the 9 auxiliary spheres delineates the geometrical locus where the center of the first atom, as ordered in the inserted molecule (here the oxygen atom as red colored sphere), can be placed so as for the two molecules of the system to be in "touch". The annotated distance of 1.45Å is equal to the sum of the atomic radii multiplied by the algorithm's scaling factor f_R (here adjusted at 0.5). This scaling factor is used to describe the excluded volume interactions of the closest atoms between the inserted and native molecule of the system (the oxygen and hydrogen atoms in this example), possessing atomic radii of 1.7Å and 1.2Å respectively. Additionally, placing the center of the oxygen of the inserted molecule at different points on the generated gray 3D surface surrounding the auxiliary atoms, brings the hypothetical ligand and protein molecules in contact with out overlap. Notably, the connectivity between the inserted atoms does not add significant complication at this computational stage. This allows the insertion of two or more molecules simultaneously, as long as the relative position between atoms is maintained during the geometrical calculation and the relative inter-molecular degrees of freedom are sampled in an outer loop. Furthermore regarding SASA, the proposed method is expected to be used in ensembles, where the system configurations are created based on desirable statistical ensembles. Similarly, the internal degrees of freedom of the ligand could be sampled by simulating the inserted molecule at the ideal gas conditions. The geometrical calculation would then be performed over a double nested loop over the configurations of the ligand and the system ensembles.

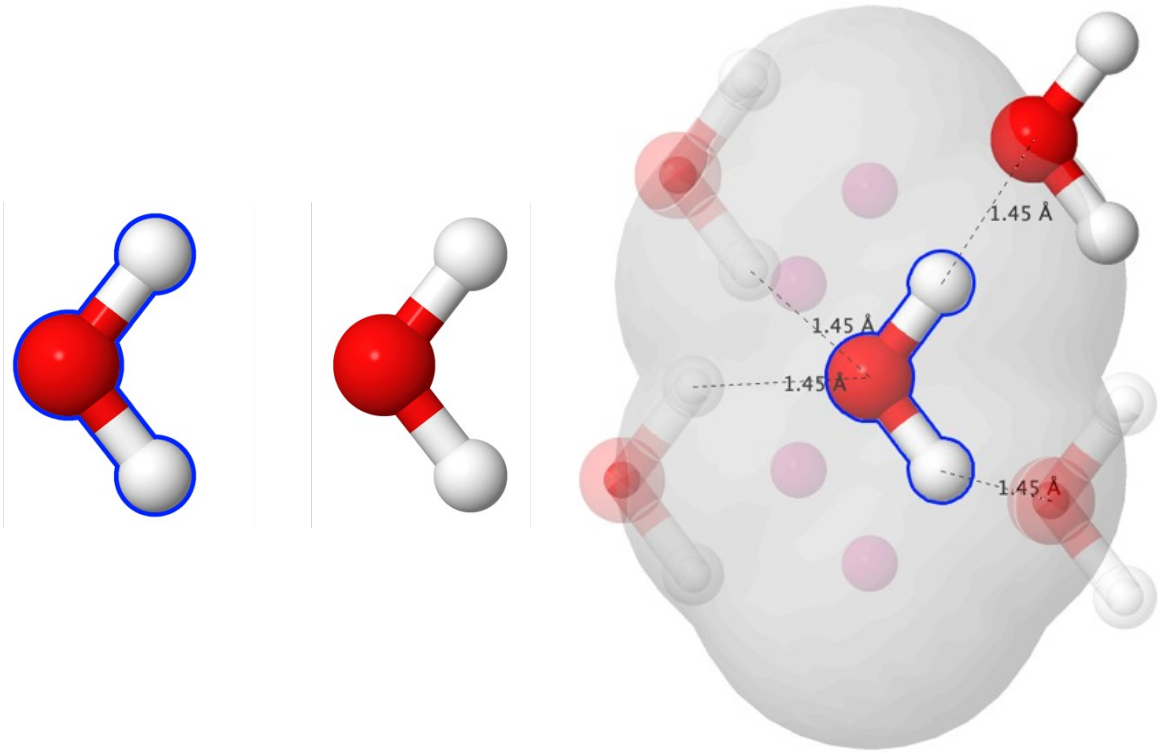


Figure 1. Graphical representation of the molecular accessible surface volume methodology in Jmol[45]. On the left side, a showcase of the test system consisting of two water molecules, the former representing a molecule in the system (blue outline) and the latter displaying the inserted molecule. On the right side, an illustration of the excluded volume around the water molecule of the test system (blue outline), where the generated gray surface points coincide with the center of the oxygen atoms of the 3 semi-transparent and 1 opaque inserted surrounding water molecules, without penetrating the generated gray 3D surface, at a fixed 1.45Å distance between the closest atoms of the inserted water molecules and the center water molecule of the system.

In this work, sampling of the relative protein-ligand orientation is a crucial step, independent of the way that the rest of degrees of freedom are sampled. Among the various methods expressing relative molecular orientation, quaternion usage was preferred in order to generate random

molecular orientations. The generation of random molecular orientations has been based on the Marsaglia G. method[46], implemented as follows:

- Firstly, two numbers x_1 and y_1 are selected from a random uniform distribution between (-1, 1), until $s_1 = x_1^2 + y_1^2 < 1$ is satisfied.
- Similarly, two more numbers x_2 and y_2 are selected respectively from a random uniform distribution between (-1, 1), until $s_2 = x_2^2 + y_2^2 < 1$ is satisfied.
- The generated values of s_1 and s_2 are used for the production of a random unit quaternion q as $q = [\, y_2\sqrt{\frac{(1-s_1)}{s_2}},\ x_1,\ y_1,\ x_2\sqrt{\frac{(1-s_1)}{s_2}}\,]$

By generating sets of unitary quaternions and applying the corresponding rotations to the inserted ligand molecules, it is possible to create a set of protein-ligand relative orientations.

However, from a technical perspective , the greatest challenge and major concern in developing a computational tool capable of estimating the molecular inaccessible volume and molecular accessible area in protein-ligand systems, has been the memory usage due to the large size of the resulting system. In order to be able to use Dodd and Theodorou's analytical approach[44] as a black-box library, the distribution of the computational load using Message Passage Interface (MPI)[47] over a number of processors was compulsory. This was so as to ascertain the efficient handling of the memory load. As a result, the user can perform analytical calculations of the molecular inaccessible volume and molecular accessible area in realistic protein-ligand systems usign reasonable computational resources.

Results and Discussion

In order to assess the proposed method, several tests were performed on different systems of varying size. Ligand and protein molecules constituting the main focus systems were mainly downloaded from the Protein Data Bank PDB[48] in .pdb format except simpler molecules, which were retrieved from Github[49] in .xyz format. PDB files with bound molecules underwent conversion so as to separate ligand and protein components into different .xyz files for more efficient and convenient file manipulation. All molecules and their generated volume-area surfaces were visualized by Jmol[45]. Initial tests of the algorithm were performed between simple molecules like water[50], methane[50] and ethane[50], following progression to more complex molecular systems downloaded from PDB and analyzed. More specifically, the 1zp8[51], 2bpw[52] and 4wtg[53] PDB files were selected as representatives of small-, medium- and large-size scale molecular systems respectively. 1zp8 and 2bpw PDB entries refer to HIV-1 protease-inhibitor complexes. former 1zp8 demonstrates an effective replacement of a peptide group in HIV-1 protease inhibitors with 1,2,3-triazole[51]. 2bpw demonstrates the ability to replace a putative inhibitor bound to the HIV-1 protease in single crystals[52.] The third PDB entry (4wtg) includes a modified version of the Hepatitis C virus (HCV) RNA-dependent RNA polymerase (RdRp) in complex with the clinically active metabolite formed by sofosbuvir, Mn^{2+} and a primer-template RNA[53].

Given a PDB file that contains both ligand and protein, estimations were performed on the inaccessible and accessible volume and surface area of the ligand against each desired protein receptor at different relative orientations by separating the protein and ligand input molecules from PDB[54]. Having assigned an atomic radius to each type of atom of both the ligand and protein molecules, several calculations were carried out by scaling all distances with a common factor. This was done in order to investigate the effect of uniformly scaling the contact distance between atoms. In our attempt to validate the algorithm, we used the default values of van der

Waals radii as starting values for each atom in Jmol. Depending on the practical application, the potential user of our computational tool may choose to alter the assignment of each atomic radius, taking into consideration the difference between ions and uncharged atoms for instance. Nevertheless in this study, given that the main concern is to provide validation of our approach, the simplest reproducible cases were selected while the option of changing the values of atomic radii was deferred for future versions purposes. Due to this reason, no further processes were performed on protein molecules extracted from the downloaded PDB files, like restoring missing atoms or imposing the proper protonated state under a given pH.

Despite the development of this computational tool taking advantage of the analytical calculation of Dodd and Theodorou[44] to a large extent, the proposed calculations of molecular accessible surface and molecular inaccessible volume can also be carried out by making use of any other computational tool capable of calculating SASA. To achieve this, one has to generate the set of auxiliary spheres in the same manner as described in the previous paragraph (**Figure 1**) and then, perform the calculations with the tool of choice. In the context of this research, visual representation was accomplished by using Jmol and its ability to draw 3D isosurfaces. It should be noted that visual rendering of the aforementioned isosurfaces is an arduous computational task, with memory requirements increasing significantly as the size of the molecular systems grows. Nevertheless, most of the available visualization tools output significantly less accurate results when compared to the analytical estimation of Dodd and Theodorou[44]. However, due to graphical representation necessities in many studies, the best strategy is to combine both approaches. Subsequently throughout this work, we report our estimations using the analytical method of Dodd and Theodorou[44,] whereas Jmol is used for visualization purposes. Finally, in order to provide better insight into molecular accessible surfaces, ligand placement on a point of

the protein surface is also presented, highlighting the contact between the test molecules given the selected relative ligand-protein orientation.

In **Figure 2** , the estimations of the molecular accessible surface area (**Figure 2a**) and molecular inaccessible volume (**Figure 2b**) are presented for various ligand-protein orientations of the 1zp8 test system (HIV-1 protease with its AB-2 inhibitor[51]). Relative orientations were randomly produced via quaternion formulation of Marsaglia on the ligand-protein pair found in the 1zp8 PDB file downloaded from the Protein Data Bank. The estimated values of molecular accessible surface area (**Figure 2a**) and molecular inaccessible volume (**Figure 2b**) are plotted versus the quaternion distance. The baseline against which the quaternion distance was calculated is the ligand-protein orientation of the original input 1zp8 PDB file configuration. Since plotting against quaternion distance constitutes projection onto one-dimensional space, the reader should bear in mind that only distance relevant to the original orientation retains the properties of distance, meaning that any of the expressed orientations depicted as points in close proximity on **Figure 2,** may actually be far apart. Nonetheless, the above representation style was selected since the deviation of 50 sampled orientations relative to the original one found in the 1zp8 PDB file is better illustrated .

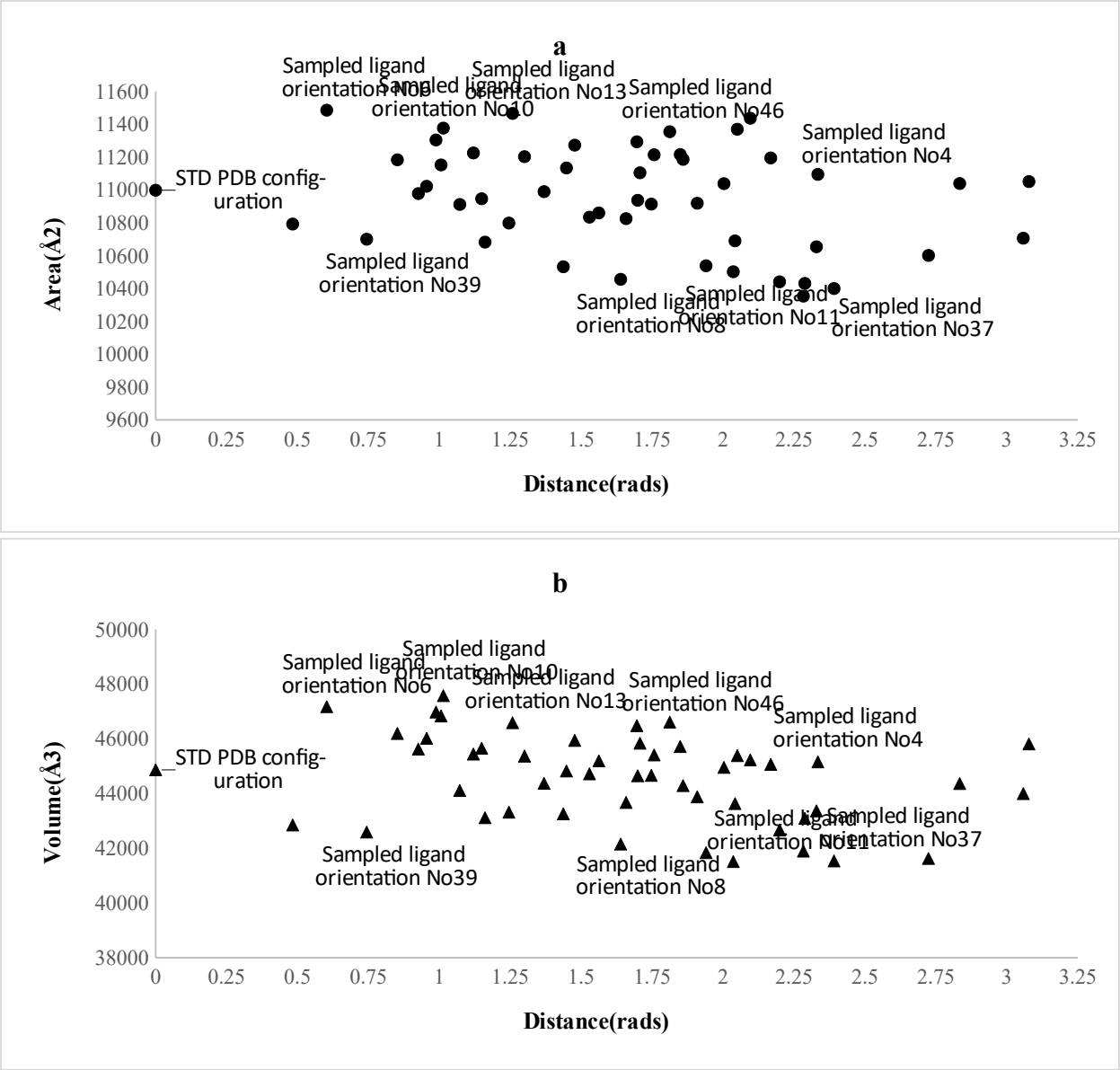
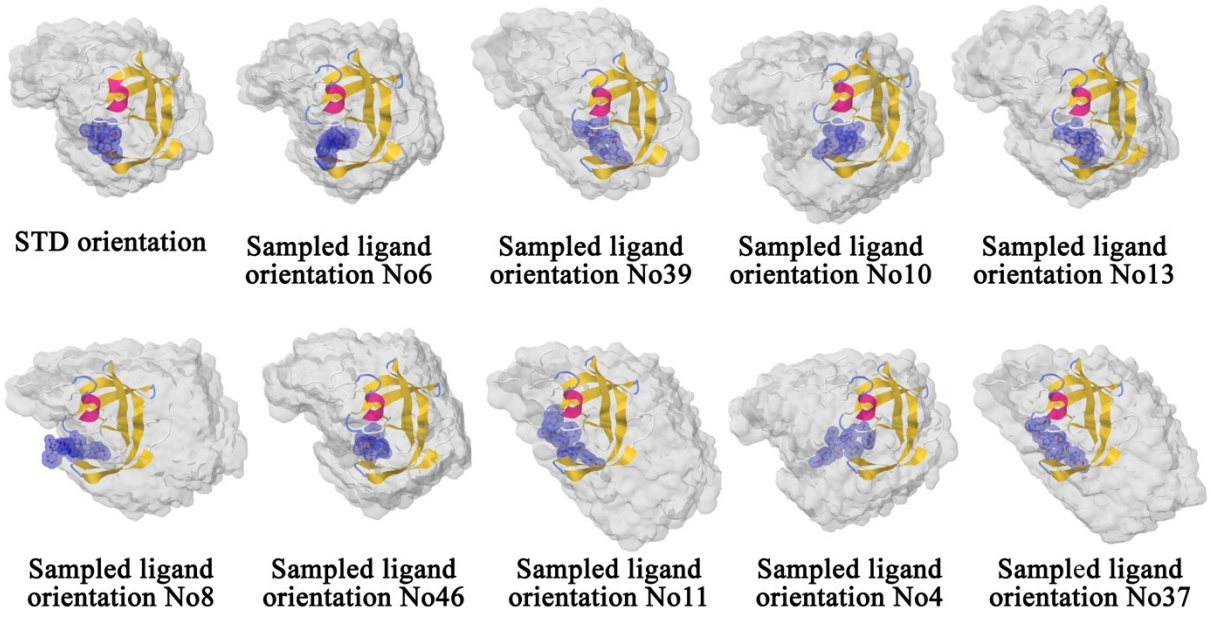


Figure 2. Analytical calculation of ligand-protein molecular accessible surface area (a) and molecular inaccessible volume (b) of the HIV-1 protease and AB-2 inhibitor complex retrieved from the 1zp8 PDB entry, at different orientations relative to the ligand-protein orientation of the standard PDB file configuration (marked as STD PDB configuration).In both charts, the original input molecular configuration is show at x=0, followed by 50 random ligand-protein orientations sampled by Marsaglia's method[46]. The estimations are plotted as a function of the quaternion distance (in radians) between each orientation and the relative ligand-protein orientation of

the original PDB molecular configuration. The labeled data points indicate the range of calculations, while their annotated numbers correspond to the order in which the random orientations were sampled and aim to help the reader compare the plotted information against the corresponding modelled structures shown in **Figure 3**.

In **Figure 2**, the values of the accessible volume and surface area range from considerably high to low levels, depending on the various ligand configurations compared to the original one. In **Figure 3a**, a sample of 3D representations of the protein-ligand molecular accessible surface are shown , mainly for configurations retrieved from the minima or maxima of **Figure 2a** and **2b** using Jmol. According to the displayed molecular states, binding of the inhibitor to the protein can be achieved with significant changes in the relative orientation. Notably, several of the sampled ligand orientations could potentially bind in the opposite direction, reverse to the ligand configuration of the original 1zp8 PDB file (**Figure 3b**). We should point out that calculations in **Figure 2** pertain solely to excluded volume interactions. Therefore such observations may serve exclusively for initial screening. Moreover, in the calculations of **Figure 2**, we do not distinguish between placing the ligand into pocket cavities or onto the outer surface of the protein. Nevertheless, once the total molecular accessible surface is evaluated, it is also possible to partition the area based on concavity, charge, polarity, or hydrophobicity of the protein contact atom utilizing the tools which have been developed for SASA and are available in visualization software like Jmol. It should be noted that in the molecular accessible surface, each point corresponds to a specific atom-atom interaction between the ligand and protein molecules, with more than 3 body contacts, mapped onto lines and points which form from the intersections of spheres at the surface.

a


STD orientation
Sampled ligand orientation No6
Sampled ligand orientation No39
Sampled ligand orientation No10
Sampled ligand orientation No13
Sampled ligand orientation No8
Sampled ligand orientation No46
Sampled ligand orientation No11
Sampled ligand orientation No4
Sampled ligand orientation No37

b

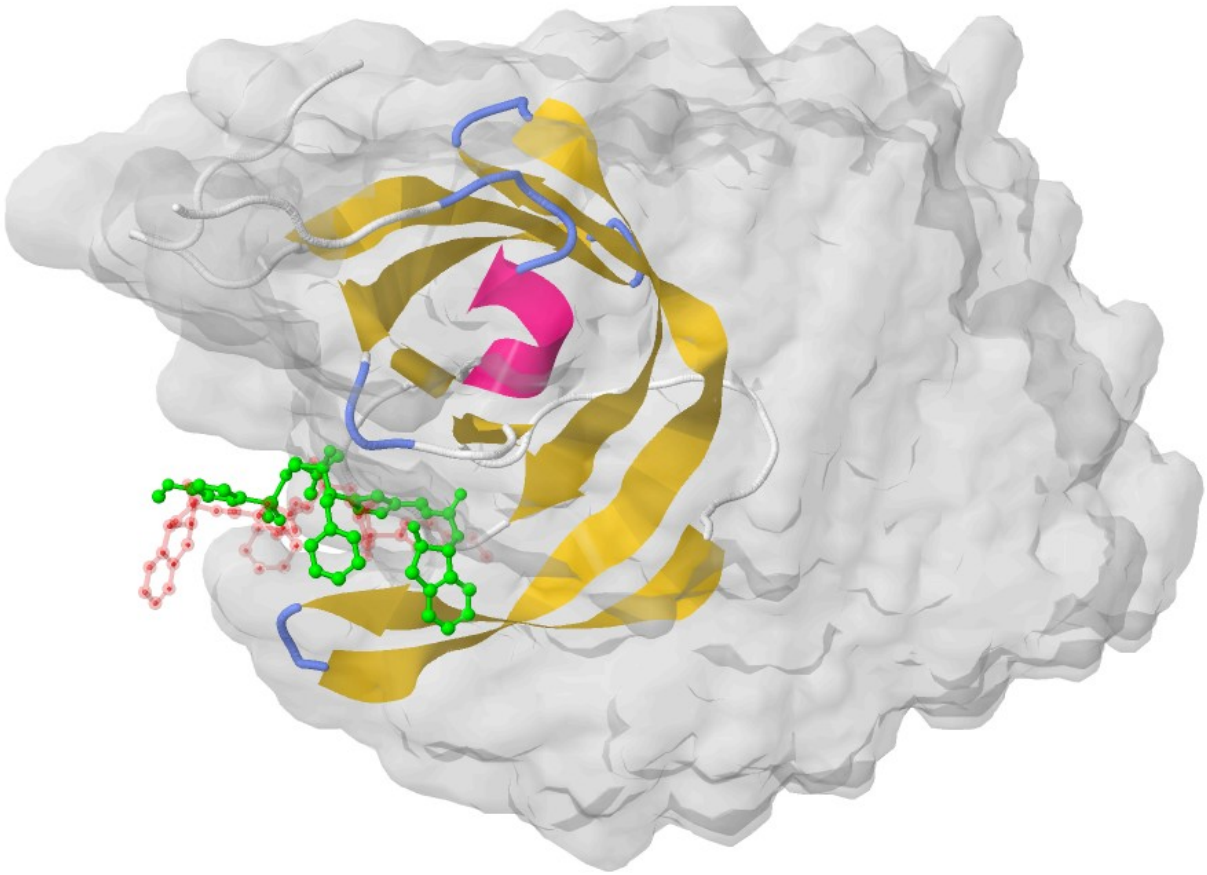


Figure 3. (a) 3D illustrations using Jmol of the 1zp8 PDB protein-ligand molecular accessible surfaces extracted from a selected set of sampled relative orientations presented in **Figure 2** (labeled by their sample number). Ligand and protein molecules are presented at the corresponding relative orientation by placing the sampled ligand configuration onto the molecular accessible surface close to the original biding cavity. Unlike SASA, our molecular accessible surface is a function of both the actual ligand and the ligand-protein relative orientation. **(b)** A more detailed view of the sampled ligand configuration No 8 versus the ligand configuration of the original 1zp8 PDB file. Interestingly in this sampled orientation, the No8 ligand can bind in the opposite direction, reverse to the ligand configuration of the original 1zp8 PDB file which is depicted transparently in red, while the sampled ligand configuration No 8 is colored bright green.

In an effort to verify and validate the accuracy of the proposed approach, the estimate of the molecular accessible surface area and an arithmetic finite difference estimate of the inaccessible volume are shown in **Figure 4**. The numerical derivative has been estimated by performing volume calculations over slight increments of the radius parameter δR incorporated in the algorithm. Confirming the consistency between our estimations of molecular accessible surface and inaccessible molecular volume, the analytical calculation of the molecular accessible surface can be estimated using finite differences provided that the alteration in the radius parameter is neither too small nor too big as it is the case with most numerical estimations based on finite differences.

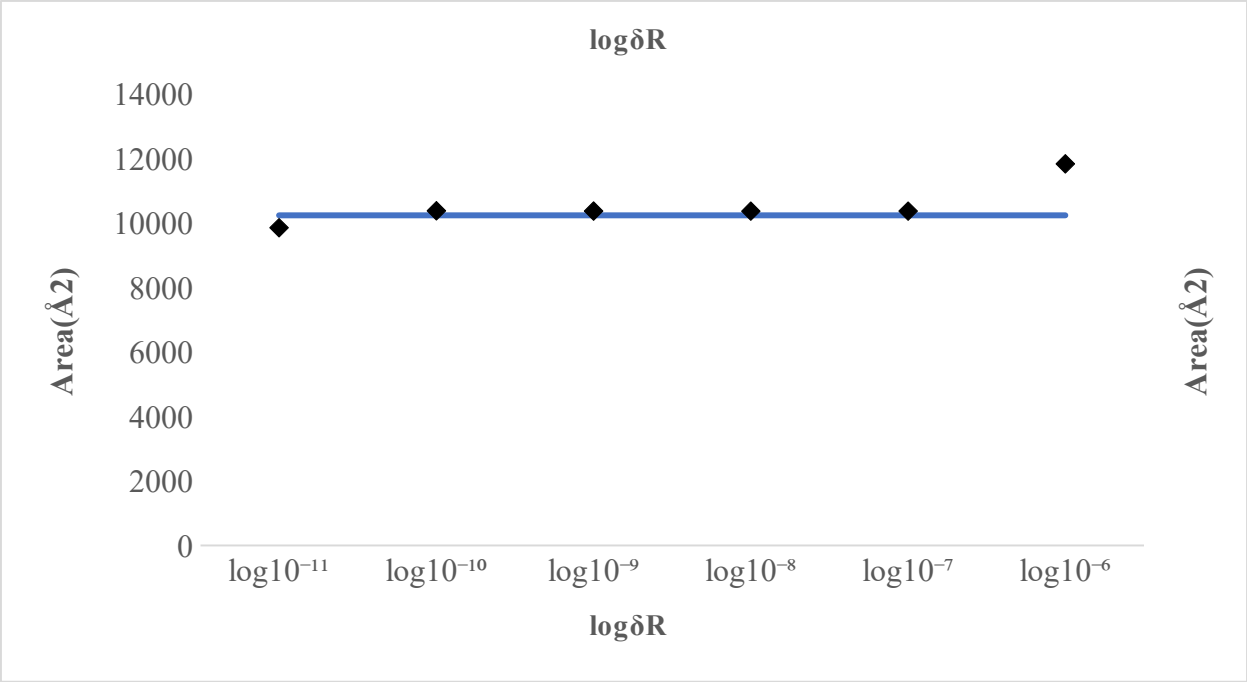


Figure 4. Comparison of the implemented estimation of the molecular accessible area (blue continuous line) against its derivative equivalent based on numerical differentiation of the inaccessible molecular volume (black diamonds). The numerical estimate is presented as a function of the gradually increasing radius parameter δR used for the calculation of forward finite difference (in logarithmic scale). The estimations have been performed upon the original input configuration of the protein-ligand complex inside the 1zp8 PDB file.

Having established the consistency between molecular inaccessible volume and accessible molecular surface, in **Figure 5** we demonstrate the validity of molecular inaccessible volume calculation by comparing the proposed analytical calculation with the estimation based on random "Widom"-like test insertions[55] under the original input relative orientation regarding a simple test system, where methane and caffeine act as ligand and protein molecules, respectively. To perform the stochastic estimation, we initially enclosed the molecule of caffeine in a box and then, measured the ratio of attempts which failed to place the methane molecule without overlap in the box, given the original relative orientation. An estimate of the inaccessible volume was

produced after multiplying the volume of the box by the ensemble average of the ratio of failed "test" insertions. In **Figure 5,** the stochastic estimation is reported as a function of the number of insertions, alongside the analytical volume estimation at the same original relative ligand-protein orientation, where the results from the stochastic method coincide with our analytical calculation output.

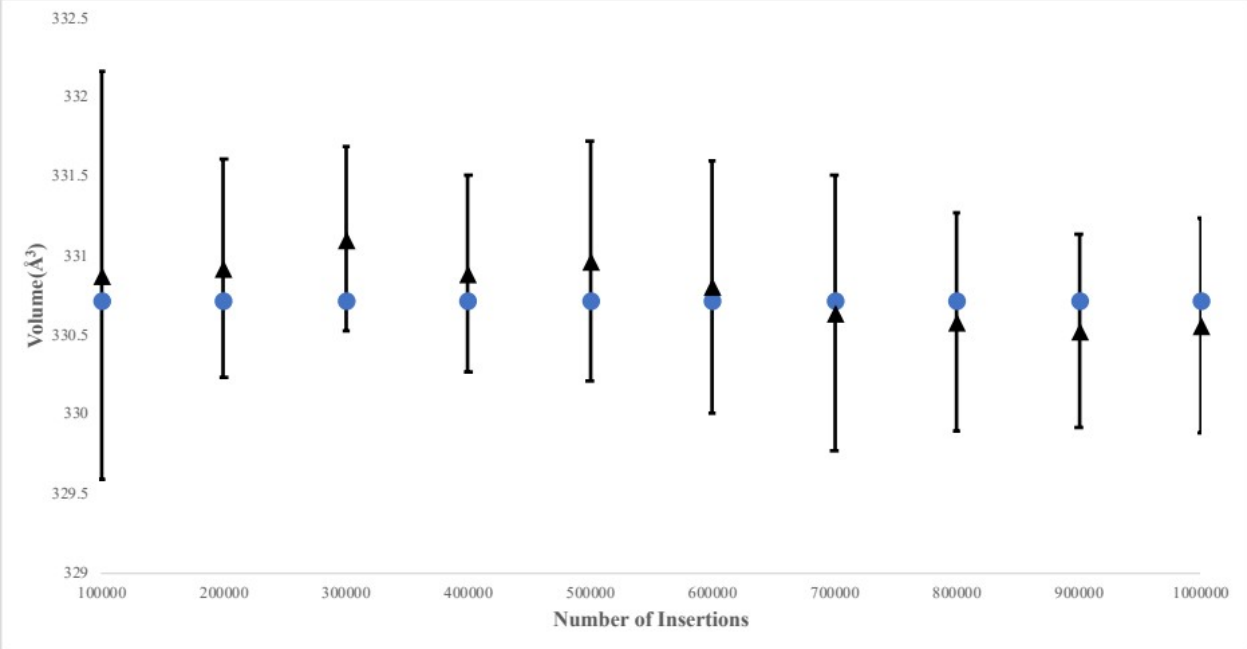


Figure 5 Comparison of analytical inaccessible volume calculation (blue circular points) versus the stochastic evaluation based on test "Widom"-like insertions in a simple molecular system, consisting of methane and caffeine as ligand and protein molecules respectively (black triangular points). The stochastic estimation results were acquired after 5 repetitive runs on the aforementioned test system at the original relative orientation with different seed numbers for each given number of insertions. All calculations coincided with the analytical estimation of inaccessible volume, within the 95% confidence interval (depicted as error bars in the above graph).

In this work, accurate evaluation of the molecular inaccessible volume and molecular accessible surface is rendered possible, given the values of inter-atomic distances where two atoms are

expected to exclude each other. As it has already been shown in the development of particle deletion[43,42,56,57,58] and staged insertion methods[59], it is also possible to use estimations of the accessible volume. These estimations are based on hard core interactions as part of the evaluation of the chemical potential in the case of molecules interacting via "soft" potential., In this case, the free energy difference related to the transformation of hard cavities formed through the hard core interactions is added to the final "soft" molecule. The overall calculation becomes independent of the actual choice by considering the range of hard core interactions smaller or larger than the minimum distance of two atoms expected to contact each other under the given conditions. If the free energy difference is estimated via the staged insertion method, hard core interactions should be smaller than the minimum distance. If the particle deletion method is used, hard core interactions should be larger than the minimum distance. In any case, the range of hard core interactions is expected to be smaller than the distance at the first pick of inter-atomic radius of gyration. Consequently, the proposed algorithm for the estimation of molecular inaccessible volume of ligands is expected to have practical uses in the estimation of protein-ligand binding affinities via staged insertion or particle deletion methods. It is therefore interesting to assess the effect that the range of hard core interactions may have on the proposed estimation.

In **Figure 6,** we observe the effect of scaling all inter-atomic contact distances by a common factor f_R covering the range which is expected to be used by the method as part of a staged insertion and the particle deletion scheme, regarding 3 molecular test systems of different sizes based on the 1zp8, 2bpw, 4wtg files downloaded from PDB. More specifically, the smallest 1zp8 system consists of 812 atoms in total, 765 of which form the HIV-1 protease while the remaining 47 atoms form its ligand inhibitor AB-2[51]. The mid-sized scale 2bpw system contains 1559 atoms, 1514 constituting the HIV-1 protease and 45 its potent ligand inhibitor[52]. Lastly, the

largest 4wtg system consists of 4357 atoms, 4327 of which belong to the modified version of HCV RdRp and the remaining 30 atoms are found within its ligand, the clinically active metabolite formed by sofosbuvir, Mn^{2+} and a primer-template RNA[53]. Examination of the accessible surfaces dependency on the algorithm's parameter f_R promotes an interesting perspective. There is a certain range where increasing the scaling factor f_R leads to reduction of the accessible area, strongly indicating presence of concave parts on the protein surface which shrink as the radius expands . However, one may conceive of an approach that uses such observations to identify the presence of cavities but to our knowledge, there is no such method. This is probably due to the usual alternative methods being quite sufficient in identifying cavities or due to the fact that similar calculations would require significant accuracy in the estimation of accessible surfaces. This would not be a practical choice since most of the available methods are of a stochastic nature. On the other hand, implementing the analytical calculation of Dodd and Theodorou[44] leads to accurate estimations which can be used to estimate partial differences from finite differences. Finally, for users that would like to use our approach in combination with existing (or newer methods) for partitioning the surface area based on concavity, we should note that the correlation between accessible area of a concave cavity formed out of spheres can be affected by the actual definition of the criteria used to separate concave from convex regions.

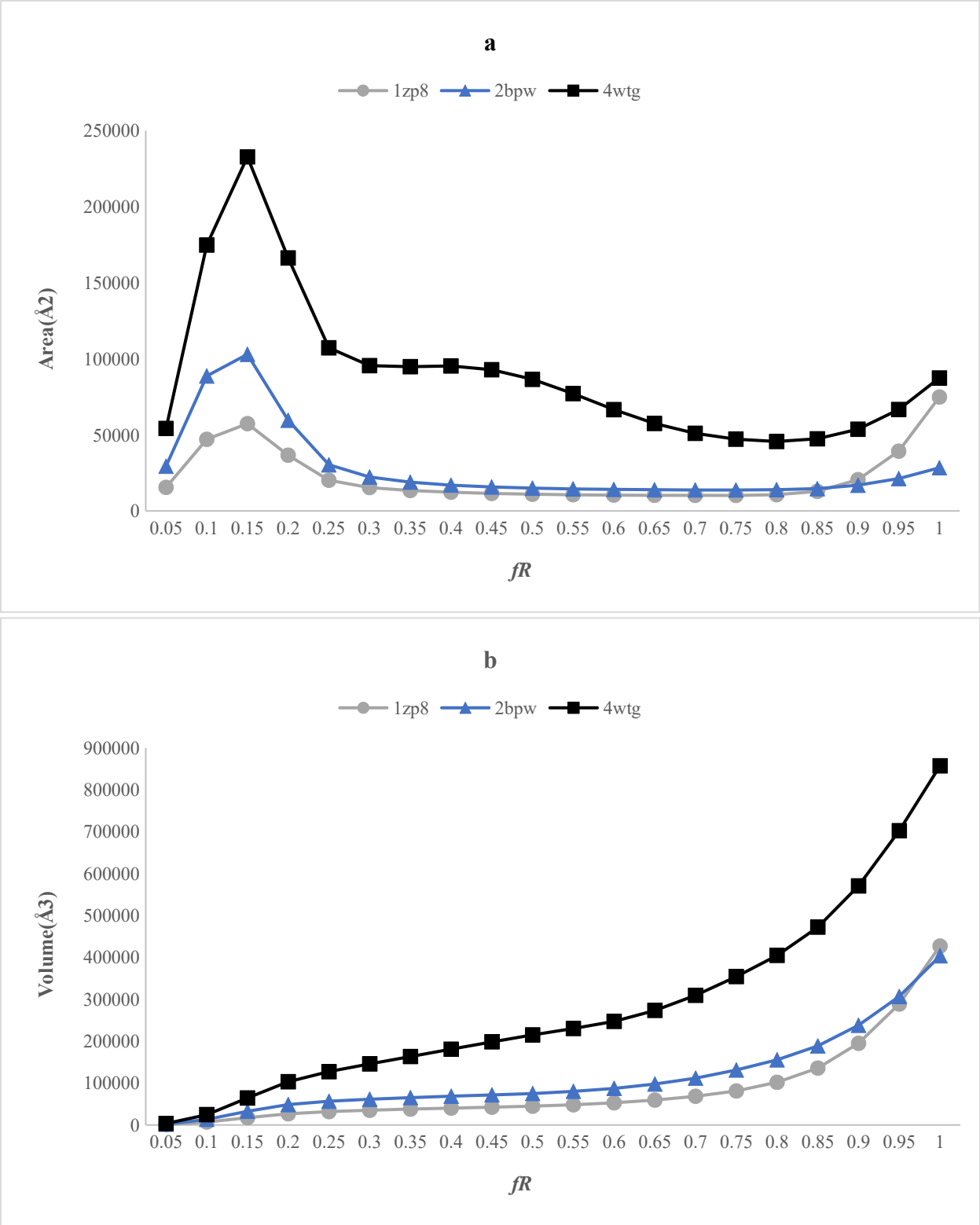


Figure 6. Estimations of the accessible surface area **(a)** and inaccessible volume **(b)**, both expressed as functions of the algorithm's parameter scaling factor f_R. The radii of the auxiliary

spheres which determine the range of hard core inter-atomic interactions have been estimated by scaling by a common factor, the sum of the van der Waals radii for each atom pair that is used in the formation of the auxiliary sphere. Tests were performed upon 3 molecular systems of varying size (1zp8, 2bpw, 4wtg).

Finally, as mentioned previously, a considerable amount of effort has been put in decomposing analytical calculations into independent sub-calculations which can be performed in parallel, since dealing with all of the auxiliary spheres using a single processor may not be feasible for most of the protein-ligand complexes of interest. Aiming to distribute the memory load into multiple processors even at the expense of performing more arithmetic calculations, in **Figure 7** we present the algorithm's execution time as a function of the number of processors used in our parallel decomposition (**Figure 7a**) as well as a function of f_R alterations utilizing all processors of our computational nodes through MPI[47] (**Figure 7b**). The system examined in **Figure 7a** consists of the protein-ligand complex retrieved from the 1zp8 PDB entry where all radii have been scaled at half of their van der Waals value by setting the f_R parameter to 0.5, while in **Figure 7b,** the largest in size 4wtg system (4357 atoms) is tested with increasing f_R values exploiting plenty of computational resources (38 processors). As the size of the molecular system increases relevant to the available resources per processor, the necessity and parallel efficiency of actual calculations may differ, but our approach is expected to be applicable provided that sufficient computational resources are allocated. Subsequently, parallelization through MPI makes our approach suitable both for super-computers and homemade clusters alike.

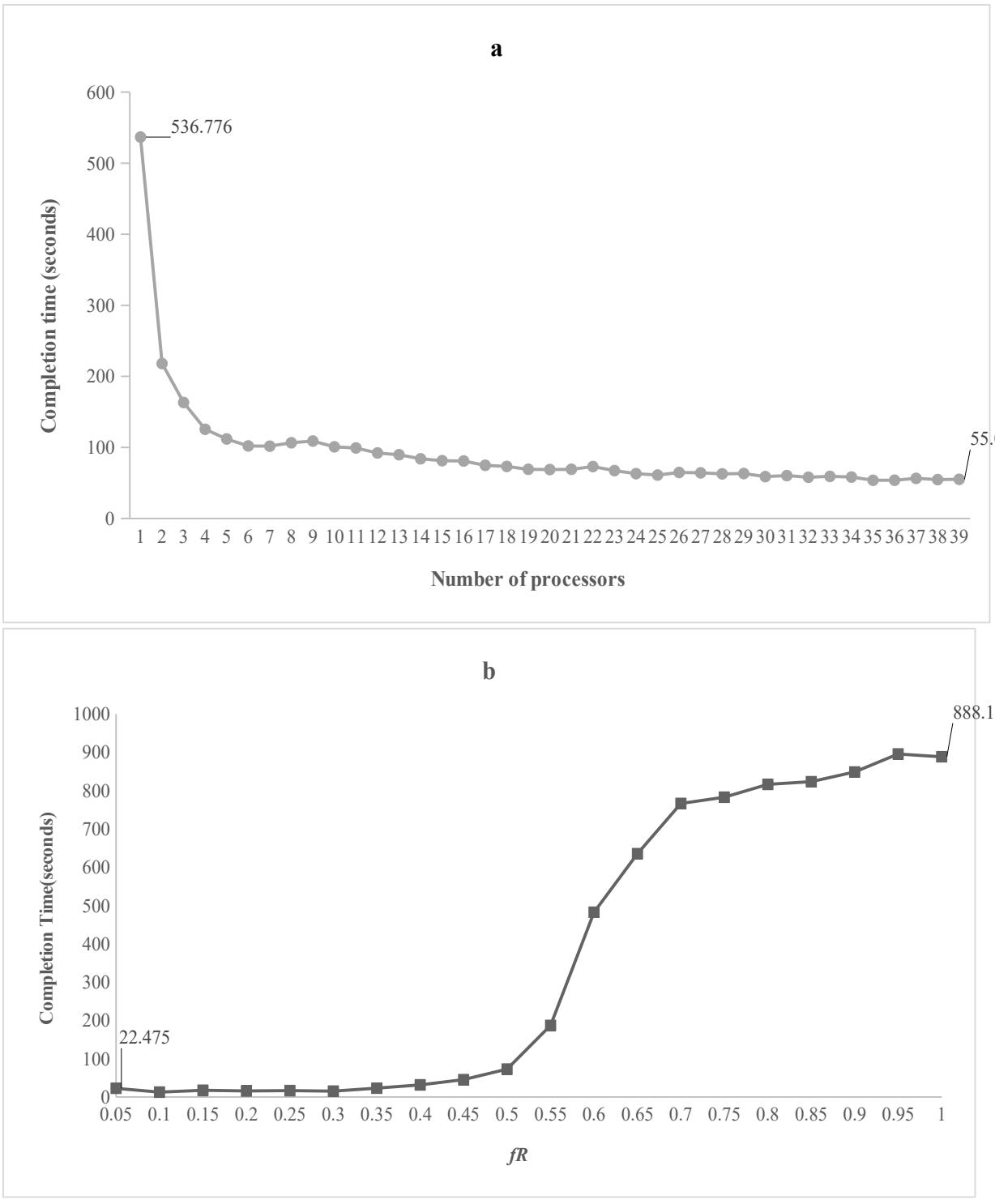


Figure 7. Inaccessible volume and accessible surface area calculations completion time as a function of the number of processors tested on a small size-scale molecular system (1zp8) (**a**) as

well as a function of *fR* alterations upon the largest in size 4wtg molecular system utilizing all 38 processors of our computational nodes through MPI (**b**).

Conclusion

In this work, the estimation of the molecular inaccessible volume and accessible surface area are proposed as a generalization of SASA. We implemented the proposed approach for estimation of protein-ligand inaccessible volume and accessible surface area upon a set of molecular systems of various sizes. We demonstrated how it is possible to estimate the proposed molecular volumes and surfaces using any available tool that can be used for the estimation of SASA by adding the proper set of auxiliary spheres as in the included example of Jmol. Furthermore, by utilizing the power of analytical calculation of the volume of fused spheres by Dodd and Theodorou plus distributing the computational load via MPI, it is possible to make very accurate estimations in a variety of protein-ligand systems. The validity of our approach was assessed firstly by estimating the inaccessible volume via stochastic Widom-like test insertion method and secondly, by comparing the molecular accessible surface with a numerical finite difference calculation. Finally, by drawing the connection between the proposed molecular inaccessible volume and free energy difference estimations via the staged insertion and deletion schemes, the molecular inaccessible volume ought to be used in the future for estimation of protein-ligand binding affinities. Alternatively, it is expected to constitute an additional visualization tool, providing more specificity to the examination of protein-ligand interactions.

Data and Software Availability

We provide a FORTRAN based program that is able to perform the inaccessible volume and accessible surface calculations reported in this work by performing calls to a static library that deploys Dodd and Theodorou estimation of the volume of fused spheres, kindly provided to us by Professor Theodorou. The program requires a minimal input of two .xyz files, the first one constituting the protein molecule and the second one possessing the ligand coordinates at the desired relative orientation. Additionally, example files are provided as supplementary material which can be used for verification purposes. For more details, please check Supporting Information below.

ACKNOWLEDGMENTS

The support in form of computational time granted from the Greek Research & Technology Network (GRNET) at the National HPC facility - ARIS - under projects "ToraDrug" and "ProLiq" (project IDs 001040 and 002035), are kindly acknowledged. The accessibility to the MAPS platform via Scienomics, Groups of Scientific Excellence initiative is also acknowledged. Finally, Professor Theodorou is acknowledged for providing us with an implementation of the Dodd and Theodorou method.

References:

(1) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte Dtsch. Chem. Ges.* **1894**, *27* (3), 2985–2993. https://doi.org/10.1002/cber.18940270364.

(2) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44* (2), 98–104.

(3) Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R. Folding and Binding Cascades: Dynamic Landscapes and Population Shifts. *Protein Sci. Publ. Protein Soc.* **2000**, *9* (1), 10–19.

(4) Lionta, E.; Spyrou, G.; Vassilatis, D.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1923–1938. https://doi.org/10.2174/1568026614666140929124445.

(5) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288. https://doi.org/10.1016/0022-2836(82)90153-X.

(6) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: A System to Select 'Quasi-Flexible' Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8* (2), 153–174. https://doi.org/10.1007/BF00119865.

(7) Diller, D. J.; Merz, K. M. High Throughput Docking for Library Design and Library Prioritization. *Proteins* **2001**, *43* (2), 113–124. https://doi.org/10.1002/1097-0134(20010501)43:2<113::aid-prot1023>3.0.co;2-t.

(8) Burkhard, P.; Taylor, P.; Walkinshaw, M. D. An Example of a Protein Ligand Found by Database Mining: Description of the Docking Method and Its Verification by a 2.3 A X-Ray Structure of a Thrombin-Ligand Complex. *J. Mol. Biol.* **1998**, *277* (2), 449–466. https://doi.org/10.1006/jmbi.1997.1608.

(9) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Aided Mol. Des.* **2001**, *15* (5), 411–428. https://doi.org/10.1023/a:1011115820450.

(10) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489. https://doi.org/10.1006/jmbi.1996.0477.

(11) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: Fast, Fully Automated Docking of Flexible Ligands to Protein Binding Sites. *Chem. Biol.* **1996**, *3* (6), 449–462. https://doi.org/10.1016/s1074-5521(96)90093-9.

(12) Schnecke, V.; Kuhn, L. A. Virtual Screening with Solvation and Ligand-Induced Complementarity. *Perspect. Drug Discov. Des.* **2000**, *20* (1), 171–190. https://doi.org/10.1023/A:1008737207775.

(13) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. EHiTS: An Innovative Approach to the Docking and Scoring Function Problems. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 421–435. https://doi.org/10.2174/138920306778559412.

(14) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* **1990**, *8* (3), 195–202. https://doi.org/10.1002/prot.340080302.

(15) ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation - Abagyan - 1994 - Journal of

Computational Chemistry - Wiley Online Library
https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540150503 (accessed Mar 23, 2021).

(16) McMartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms for Structure-Based Drug Design. *J. Comput. Aided Mol. Des.* **1997**, *11* (4), 333–344. https://doi.org/10.1023/a:1007907728892.

(17) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662. https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

(18) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.

(19) Clark, K. P.; Ajay. Flexible Ligand Docking without Parameter Adjustment across Four Ligand–Receptor Complexes. *J. Comput. Chem.* **1995**, *16* (10), 1210–1226. https://doi.org/10.1002/jcc.540161004.

(20) Taylor, J. S.; Burnett, R. M. DARWIN: A Program for Docking Flexible Molecules. *Proteins* **2000**, *41* (2), 173–191.

(21) Leach, A. R.; Kuntz, I. D. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **1992**, *13* (6), 730–748. https://doi.org/10.1002/jcc.540130608.

(22) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-Based Discovery of Inhibitors of Thymidylate Synthase. *Science* **1993**, *259* (5100), 1445–1450. https://doi.org/10.1126/science.8451640.

(23) Böhm, H. J. LUDI: Rule-Based Automatic Design of New Substituents for Enzyme Inhibitor Leads. *J. Comput. Aided Mol. Des.* **1992**, *6* (6), 593–606. https://doi.org/10.1007/BF00126217.

(24) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2* (5), 317–324. https://doi.org/10.1016/1074-5521(95)90050-0.

(25) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. *J. Comput. Aided Mol. Des.* **2000**, *14* (8), 731–751. https://doi.org/10.1023/a:1008158231558.

(26) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De Novo Design of Enzyme Inhibitors by Monte Carlo Ligand Generation. *J. Med. Chem.* **1995**, *38* (3), 466–472. https://doi.org/10.1021/jm00003a010.

(27) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided Mol. Des.* **1997**, *11* (5), 425–445. https://doi.org/10.1023/a:1007996124545.

(28) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791–804. https://doi.org/10.1021/jm980536j.

(29) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–356. https://doi.org/10.1006/jmbi.1999.3371.

(30) DeWitte, R.; Shakhnovich, E. SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J Am Chem Soc* **1996**, *118* (47), 11733–11744.

(31) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP—Potential of Mean Force Describing Protein–Ligand Interactions: I. Generating Potential. *J. Comput. Chem.* **1999**, *20* (11), 1165–1176. https://doi.org/10.1002/(SICI)1096-987X(199908)20:11<1165::AID-JCC7>3.0.CO;2-A.

(32) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graph. Model.* **2002**, *20* (4), 281–295. https://doi.org/10.1016/s1093-3263(01)00125-5.

(33) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. Modelling Protein Docking Using Shape Complementarity, Electrostatics and Biochemical Information. *J. Mol. Biol.* **1997**, *272* (1), 106–120. https://doi.org/10.1006/jmbi.1997.1203.

(34) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(35) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337* (1), 209–225. https://doi.org/10.1016/j.jmb.2004.01.003.

(36) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379–400. https://doi.org/10.1016/0022-2836(71)90324-x.

(37) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79* (2), 351–371. https://doi.org/10.1016/0022-2836(73)90011-9.

(38) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16* (5), 548–558. https://doi.org/10.1107/S0021889883010985.

(39) Richards, F. M. Areas, Volumes, Packing and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176. https://doi.org/10.1146/annurev.bb.06.060177.001055.

(40) Richmond, T. J. Solvent Accessible Surface Area and Excluded Volume in Proteins. Analytical Equations for Overlapping Spheres and Implications for the Hydrophobic Effect. *J. Mol. Biol.* **1984**, *178* (1), 63–89. https://doi.org/10.1016/0022-2836(84)90231-6.

(41) Connolly, M. L. The Molecular Surface Package. *J. Mol. Graph.* **1993**, *11* (2), 139–141. https://doi.org/10.1016/0263-7855(93)87010-3.

(42) BOULOUGOURIS, G. C.; ECONOMOU, I. G.; THEODOROU, D. N. On the Calculation of the Chemical Potential Using the Particle Deletion Scheme. *Mol. Phys.* **1999**, *96* (6), 905–913. https://doi.org/10.1080/00268979909483030.

(43) Boulougouris, G.; Economou, I.; Theodorou, D. Calculation of the Chemical Potential of Chain Molecules Using the Staged Particle Deletion Scheme. *J. Chem. Phys.* **2001**, *115*, 8231–8237. https://doi.org/10.1063/1.1405849.

(44) Dodd, L. R.; Theodorou, D. N. Analytical Treatment of the Volume and Surface Area of Molecules Formed by an Arbitrary Collection of Unequal Spheres Intersected by Planes. *Mol. Phys.* **1991**, *72* (6), 1313–1345. https://doi.org/10.1080/00268979100100941.

(45) Jmol: an open-source Java viewer for chemical structures in 3D http://jmol.sourceforge.net/ (accessed Mar 24, 2021).

(46) Marsaglia, G. Choosing a Point from the Surface of a Sphere. **1972**. https://doi.org/10.1214/AOMS/1177692644.

(47) Clarke, L.; Glendinning, I.; Hempel, R. The MPI Message Passing Interface Standard. In *Programming Environments for Massively Parallel Distributed Systems*; Decker, K. M., Rehmann, R. M., Eds.; Monte Verità; Birkhäuser: Basel, 1994; pp 213–218. https://doi.org/10.1007/978-3-0348-8534-8_21.

(48) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **2003**, *10* (12), 980–980. https://doi.org/10.1038/nsb1203-980.

(49) GitHub: Where the world builds software https://github.com/ (accessed Mar 24, 2021).

(50) nutjunkie/IQmol https://github.com/nutjunkie/IQmol (accessed Mar 24, 2021).

(51) Brik, A.; Alexandratos, J.; Lin, Y.-C.; Elder, J. H.; Olson, A. J.; Wlodawer, A.; Goodsell, D. S.; Wong, C.-H. 1,2,3-Triazole as a Peptide Surrogate in the Rapid Synthesis of HIV-1 Protease Inhibitors. *Chembiochem Eur. J. Chem. Biol.* **2005**, *6* (7), 1167–1169. https://doi.org/10.1002/cbic.200500101.

(52) Munshi, S.; Chen, Z.; Li, Y.; Olsen, D. B.; Fraley, M. E.; Hungate, R. W.; Kuo, L. C. Rapid X-Ray Diffraction Analysis of HIV-1 Protease-Inhibitor Complexes: Inhibitor Exchange in Single Crystals of the Bound Enzyme. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (Pt 5), 1053–1060. https://doi.org/10.1107/s0907444998003588.

(53) Appleby, T. C.; Perry, J. K.; Murakami, E.; Barauskas, O.; Feng, J.; Cho, A.; Fox, D.; Wetmore, D. R.; McGrath, M. E.; Ray, A. S.; Sofia, M. J.; Swaminathan, S.; Edwards, T. E. Structural Basis for RNA Replication by the Hepatitis C Virus Polymerase. *Science* **2015**, *347* (6223), 771–775. https://doi.org/10.1126/science.1259210.

(54) Bank, R. P. D. RCSB PDB: Homepage https://www.rcsb.org/ (accessed Mar 24, 2021).

(55) Widom, B. Some Topics in the Theory of Fluids. *J. Chem. Phys.* **1963**, *39* (11), 2808–2812. https://doi.org/10.1063/1.1734110.

(56) Boulougouris, G. C. On the Estimation of the Free Energy, from a Single Equilibrium Statistical Ensemble, via Particle Reinsertion. *J. Phys. Chem. B* **2012**, *116* (3), 997–1006. https://doi.org/10.1021/jp2036185.

(57) Boulougouris, G. C. Multidimensional Direct Free Energy Perturbation. *J. Chem. Phys.* **2013**, *138* (11), 114111. https://doi.org/10.1063/1.4795319.

(58) Gc, B. Free Energy Calculations, Enhanced by a Gaussian Ansatz, for the “Chemical Work” Distribution. *J. Comput. Chem.* **2014**, *35* (13), 1024–1035. https://doi.org/10.1002/jcc.23590.

(59) Kofke, D. A.; Cummings, P. T. Precision and Accuracy of Staged Free-Energy Perturbation Methods for Computing the Chemical Potential by Molecular Simulation. *Fluid Phase Equilibria* **1998**, *150–151*, 41–49. https://doi.org/10.1016/S0378-3812(98)00274-X.