JCTC Journal of Chemical Theory and Computation

Biomolecular Systems

Protein docking using a single representation for protein surface, electrostatics and local dynamics

Lucas S.P. Rudden, and Matteo T. Degiacomi

J. Chem. Theory Comput., **Just Accepted Manuscript** • DOI: 10.1021/acs.jctc.9b00474 • Publication Date (Web): 07 Aug 2019

Downloaded from pubs.acs.org on August 19, 2019

is published by the American Chemical Society. 1155 Sixteenth Street N.W.,
Washington, DC 20036

Protein docking using a single representation for protein surface, electrostatics and local dynamics

Lucas S. P. Rudden [1], Matteo T. Degiacomi [1]*

[1] Department of Chemistry, Durham University, South Road, DH1 3LE, UK

* matteo.t.degiacomi@durham.ac.uk

Abstract

Predicting the assembly of multiple proteins into specific complexes is critical to the understanding of their biological function in an organism, and thus the design of drugs to address their malfunction. Proteins are flexible molecules, and this inherently poses a problem to any protein docking computational method, where even a simple rearrangement of the side chain and backbone atoms at the interface of binding partners complicates the successful determination of the correct docked pose. Herein, we present a means of representing protein surface, electrostatics and local dynamics within a single volumetric descriptor. We show that our representations can be physically related to the surface accessible solvent area and mass of the protein. We then demonstrate that the application of this representation into a protein-protein docking scenario bypasses the need to compensate for, and predict, specific side chain packing at the interface of binding partners. This representation is leveraged in our *de novo* protein docking software, JabberDock, which we show can accurately and robustly predict difficult target complexes with an average success rate of >54%, which is comparable to or greater than currently available methods.

Keywords: protein-protein interaction; protein docking; molecular dynamics; JabberDock; particle swarm optimisation; CAPRI

Introduction

Most functions in an organism are governed by interactions of proteins with specific substrates. To achieve their task, proteins often form homo- and heteromultimeric complexes. A plethora of genetic diseases are connected to mutations altering protein structures, and consequently their capacity to interact with their binding partners[1]. Consequently, a great body of research and development focuses on methods for the elucidation of protein structures. In this context, computational techniques devised to predict the complex formed when two proteins bind can be of great help. *In silico* techniques can often be significantly cheaper and quicker than experimental methods, and their predictive capability can be harnessed to guide subsequent targeted experiments.

Protein-protein docking is a highly complex optimisation problem, requiring the generation of a considerable number of candidate arrangements. To accurately discriminate between correct and incorrect docked poses, an accurate scoring function is essential. Typical scoring functions used in this context involve a set of non-trivial physical or empirical terms, combined with custom weightings. An ideal scoring function should feature minimum mathematical uncertainty while accounting for protein structure and dynamics. In order to navigate the landscape of possible conformations in search of the specific arrangements that minimise this scoring function, a highly efficient exploration method is also required. The two are intimately linked, with the scoring function guiding the behaviour of the navigator.

The simplest and most widespread approaches for protein-protein docking involve a global, systematic, rigid-body docking search. Typically, a large number of solutions are generated from a pair of static molecular structures[2–4], and a scoring function is then used to identify the most favourable arrangement. Treating proteins as fully rigid objects, while simultaneously using a scoring function that accounts for the specific position of each atom, leads to a modelling process that is excessively sensitive to the specific packing of atoms at the interface. In order to cater to protein dynamics, alterations to how models are built or how the scoring function assesses a protein arrangement are required. Possible strategies include: rigid-body docking of ensembles of structures[5–7], additional refinement stages that take place after rigid-body docking[8,9], scoring functions that feature soft potentials to allow minor molecular overlaps[10,11], pseudo coarse-grained protein representations[12], docking subunits connected by potentials[13], matching protein surfaces represented as a collection of patches[14], using normal modes to account for flexible conformational switches[15] and relaxing the interface of docking poses using techniques such as molecular dynamics (MD), Monte Carlo (MC) or simulated annealing[10,12,16]. Some methods, such as HADDOCK[17] and IMP[18], feature scoring functions that utilise a combination of terms that describe physical interactions and penalise models that do not recapitulate available experimental data. Overall, two approaches are used to describe amino acid side chains at the interface. They are either represented explicitly, thus requiring the docking method to determine their correct packing, or their presence is described by means of pseudo-coarse-grained representations. The first method requires highly sophisticated optimisation procedures that may still yield suboptimal arrangements while the second usually ignores the uncertainty in the position of the side chains that comes naturally with any time dynamics.

In this work, we describe a new protein volumetric representation, named Spatial and Temporal Influence Density (STID) maps, capable of simultaneously describing protein shape, electrostatics, and local dynamics (see Figure 1(c)). We demonstrate that complementarity of these isosurfaces can help create suitable solutions in a protein-protein docking scenario. While surface complementarity techniques have been used for many years[19], the representations from STID maps are superior to any surface method used to date. STID maps inherently consider any side chain flexibility by using the general motion of atomic point charges in time as a building block for the model. We demonstrate that the key consequence of this is the retention of accuracy between identified docking models regardless of their difficulty.

Our STID map-based scoring method is embodied in JabberDock, a protein *de novo* docking software. JabberDock explores the surface complementarity space of two binding partners by means of a Particle Swarm Optimisation (PSO) algorithm supplied by the POW^{er} optimisation environment[20].

POW^{er} features a modified version of the PSO algorithm, explicitly adapted to prevent premature convergence and maximise the diversity of solutions.

Hereafter, we first outline the theory behind our protein representation method and then present a set of benchmarks aimed at testing the accuracy of our protein-protein docking method in line with the CAPRI blind docking competition guidelines[21]. Results demonstrate that JabberDock can return models matching the quality of top *de novo* docking algorithms currently available.

Theory

The following theory is built upon the principle that a volumetric map can consider both the structural and dynamical properties of a protein. It is constructed from the inherent motion of charged atoms *via* the time-averaged dipoles forming within a localised space. The complete methodology, represented in Figure 1, is summarized as follows:

1. The PDB file of a protein structure is input along with the desired atomistic force field.

2. The protein is immersed in a water box with Na^+ or Cl^- acting as counter-ions and automatically subjected to energy minimisation, followed by an MD equilibration and production protocol.

3. A dipole map is derived based on the produced MD trajectory.

4. The dipole map is converted into a 3-dimensional grid of points, each containing a pseudo-atom with a characteristic van der Waals radius.

5. Each pseudo-atom is used to define a local Gaussian distribution, with a standard deviation determined by the van der Waals radius of the pseudo-atom.

6. A volumetric map is produced by summing, on each grid point, the value of local and neighbouring Gaussians. The resulting map is finally normalised.

From Protein Structure to Dipole Map

The overall dynamics of a protein emerge from a combination of slow large-scale motions and fast local rearrangements. We start by sampling the fast motions of the side chains by means of a short MD simulation. Consistent with experimental NMR evidence[22], we have found that 500 ps is enough for this purpose (see supplementary data and Figures S1-2). We align the resulting protein trajectory according to the centre of mass of the molecule, and arrange it within a stationary, cubic grid, wherein each voxel is 1 Å across in x, y and z (a parameter determined quantitatively in a benchmark shown in Figures S3-5). We use this information to calculate local dipoles on each grid point. To this end, we are expanding upon the theory laid out by Kirkwood[23], Fröhlich[24] and Neumann *et al.*[25], which describe the fundamental theory of dielectrics.

Following the Onsager theory of dielectric polarisation, we represent each voxel, v, as a spherical solute with volume, V_v, with an internal permittivity, ε_v, embedded within a uniform dielectric continuum with permittivity ε_Ex. The charge distribution inside the voxel is that of several point charges, with a dipole associated with the centre. Point charges within the neighbouring voxels on each Cartesian edge and corner (*i.e.*, a total of 26 neighbours, a quantity determined in a benchmark shown in Figures S3-5) are also associated with the central voxel. A sliding window is applied spatially such that a point charge at a time, t, will contribute to 27 different voxels in total. Given the fluctuations of the dipole moment of the solute, $\mathbf{M}_\mathrm{v}$, observed over the simulation in a voxel, it is possible to calculate ε_v.

The Fröhlich-Kirkwood model states that ε_v is a function of the probability distribution of the total dipole's second moment, with $\mathbf{M}_\mathrm{v}$ given by:

$$\mathbf{M}_\mathrm{v} = \sum_{i=0}^{N} q_{i,\mathrm{v}} \mathbf{r}_{i,\mathrm{v}} \ , \tag{1}$$

where $q_{i,\mathrm{v}}$ is the charge of atom i at distance $\mathbf{r}_{i,\mathrm{v}}$ from the geometrical centre of voxel v. N is the number of atoms that contribute to a voxel's dipole moment. The charges are obtained from the force field used for the simulation. For a solute with a net charge, which most voxels have, $\mathbf{M}_\mathrm{v}$ is dependent on the origin, thus the grid is fixed in time and space. Therefore, we can produce a dipole map delivering a representation of local vectorial electrostatic characteristics of a region of space occupied by a molecule (see Figure 1(b)).

From Dipole Map to Spatial and Temporal Influence Density Map

We can now leverage on the obtained dipole map to derive volumetric information on a molecule, *i.e.*, a quantity that is easier to visualise and to use in a protein-protein docking context. To this end, it is necessary to convert its dipolar vectorial representation into a scalar quantity. Under the Fröhlich-Kirkwood model, we can relate the fluctuations of each dipole moment, $\mathbf{M}_\mathrm{v}$, to the voxel's dielectric, ε_v:

$$\frac{\langle {\mathbf{M}_\mathrm{v}}^2 \rangle - \langle \mathbf{M}_\mathrm{v} \rangle^2}{3\epsilon_0 V_\mathrm{v} k_\mathrm{B} T_\mathrm{v}} = \frac{(2\epsilon_\mathrm{Ex} + 1)(\epsilon_\mathrm{v} - 1)}{2\epsilon_\mathrm{Ex} + \epsilon_\mathrm{v}} \, . \tag{2}$$

Where T_v is the temperature in a voxel, which we approximate as the temperature of the system. Solving for ε_v:

$$\epsilon_{\mathrm{v}} = \frac{1 + \dfrac{\langle \mathbf{M}_{\mathrm{v}}{}^{2}\rangle - \langle \mathbf{M}_{\mathrm{v}}\rangle^{2}}{3\epsilon_0 V_{\mathrm{v}} k_B T_{\mathrm{v}}} \dfrac{2\epsilon_{\mathrm{Ex}}}{(2\epsilon_{\mathrm{Ex}}+1)}}{1 - \dfrac{\langle \mathbf{M}_{\mathrm{v}}{}^{2}\rangle - \langle \mathbf{M}_{\mathrm{v}}\rangle^{2}}{3\epsilon_0 V_{\mathrm{v}} k_B T_{\mathrm{v}}} \dfrac{1}{(2\epsilon_{\mathrm{Ex}}+1)}}. \qquad (3)$$

Each ε_{v} value derived from the dipole map now encodes information on the local dynamics and atomic charges. Our next step is to convert the resulting dielectric map into a quantity that relates to a pseudo-electron density. To do so, we place a pseudo-atom at the centre of each voxel and calculate its polarizability, α_{v}, using the Clausius-Mossotti equation:

$$\alpha_{\mathrm{v}} = \frac{3\epsilon_0}{N_{\mathrm{v}}}\left(\frac{\epsilon_{\mathrm{v}} - 1}{\epsilon_{\mathrm{v}} + 2}\right), \qquad (4)$$

where N_{v} is the number density inside the voxel. Since N_{v} is derived from the number of pseudo-atoms inside a voxel; *i.e.*, one, we can simply set it as the inverse of V_{v}. α_{v} can then be related to a van der Waals radius R_{vdW} by a scaling relationship identified by Fedorov *et al.*[26], based on the quantum Drude oscillator model:

$$R_{\mathrm{vdW}} = 2.54\alpha_{\mathrm{v}}{}^{(1/7)}, \qquad (5)$$

where the constant 2.54 is a universal scaling factor between electron density and atomic volume at R_{vdW} in atomic units. While Fedorov *et al.* note that a full derivation of this constant is still incomplete, they demonstrate that the relationship in Equation 5 gives theoretical quantities closer to experimental data than previous models based on classical hard-sphere representations.

The electrostatic and dynamic information encapsulated in the local R_{vdW} values is now suitable to be transformed into a quantity encoding a pseudo-electron density. To this end, we assume that each pseudo-atom radius is equal to the full width at half maximum of a decaying function, here defined as a 3-dimensional Gaussian, with the maximum at the voxel's centre. This allows each pseudo-electron density to ‘leak’ into neighbouring voxels, which is reasonable given that a central voxel’s behaviour is characterised by the atoms in its neighbourhood.

Contributions from any Gaussians with a non-zero value present within a voxel are then summed, and the resulting map is finally normalised (an isosurface example is shown in Figure 1(c)), yielding a molecular representation we call Spatial and Temporal Influence Density (STID) maps. Because of this methodology, regions inside the protein’s conformational space that are visited often or are highly charged will have greater associated STID values. This property makes STID maps a useful representation in a protein-protein docking scenario, with the electrostatics arising from rapid side chain motions, often ignored with other docking software, now accounted for.

Results

STID Map Cut-off Values and Solvent Accessible Surface Area of the Protein are Related

The global average STID, $D_{\mathrm{avg.}}$, provides us with a direct comparison between the different structural and dynamic characteristics of a protein. The presence of both rigid and highly-charged regions within a protein contribute greater STID values to their respective voxels than a flexible or apolar residue. This is shown in Figure 2(a), where only the core regions of the protein are observed at greater isosurface cut-offs, but the more flexible regions can be seen at lower isovalues. Furthermore, we found that, while having a greater relative quantity of charged or polar residues did indeed increase the $D_{\mathrm{avg.}}$, time-averaged dynamics and the structure had a considerably larger impact on the maps’ topography.

We sought to determine whether a link exists between the characteristic $D_{\mathrm{avg.}}$ of each protein and any of their physical quantities that are easily measurable. For this test, we select 118 proteins with various size, shape and secondary structure, and observed that the ratio between SASA and molecular weight,

S_{m}, is anticorrelated with D_{avg} (see Figure 2(b)). The relationship could be fitted via a linear least square fit (Pearson correlation coefficient equal to -0.80):

$$D_{\mathrm{avg.}} = 0.34 S_{\mathrm{m}} + 0.59\,, \qquad (6)$$

Thus, each STID map is associated with a characteristic cut-off value, determined by the SASA and molecular weight of the protein. Important topographical features of the STID maps are entirely independent of the protein: core secondary structure features are always visible in and around an isovalue of 0.8, and highly charged atoms become isolated from the body of the protein at more stringent cut-offs beyond 0.9. This direct link between the structural characteristics of a protein and its associated volumetric isosurface's shape makes STID maps an appropriate way of representing how a protein will be perceived by its immediate surroundings, such as a binding partner.

STID Maps are Effective to Score Protein-Protein Interactions

STID maps encapsulate information on the local dynamics of the atomic charges in a protein. This feature is particularly attractive in a protein-protein docking context, as it circumvents the need of determining the specific atomic position of each side chain at the interface between two binding partners. We, therefore, used this representation within a docking protocol, where the scoring function is determined by the surface complementarity of ligand and receptor STID map isosurfaces (see Methods). Benefitting from the fact that the structural characteristics reported by each STID map isovalue are protein-independent, we determined that an isovalue of 0.43 is the most appropriate to report on all electrostatic and dynamic features of any protein within our surface complementarity scheme (see benchmark in Figures S6-7).

We implemented this calibrated STID map-based scoring into our *de novo* protein docking algorithm: JabberDock. The program utilised the PSO algorithm within the POW$^{\mathrm{er}}$ environment to explore the energy landscape associated with the arrangement of two binding partners, in search of the arrangement maximising our complementarity score. We assessed the performance of JabberDock against all the 230 test cases featured in the most recent iteration of the standard protein-protein interaction benchmark[27] (6 cases excluded for the presence of non-standard amino acids). According to the RMSD between unbound and known bound state, 151 of these cases are classified as rigid-body (easy), 45 as medium, and 34 as difficult (see Methods). To gather further information on the relationship between docking quality and the conformational change proteins undergo upon binding, we selected a diverse subset of 32 cases (20 easy, 7 medium and 5 difficult) that were also treated as bound cases. In these cases, the subunits used to predict the assemblies were proteins extracted from the known complex. We classified the quality of all our modelling runs according to the three CAPRI categories - acceptable, intermediate and high (see definition in Methods). Hereon, we qualify a test case as a success if at least one model in the top 10 ranked solutions is at least of acceptable quality.

Against the 32 bound cases, JabberDock was successful in 85.0% of the easy, 71.4% of the medium, and 20.0% of the difficult cases. Challenged with the full unbound benchmark set, JabberDock yielded successful predictions for 56.3% of the easy, 60.0% of the medium, and 54.9% of the difficult cases. While no high-quality predictions were found for any of the test cases, intermediate quality results were found for 29.2% of the easy, 22.2% of the medium, and 25.8% of the difficult cases. Overall, these results indicate that JabberDock performance is mostly unaffected by the case difficulty (full details are provided in Table S1). These results compare favourably against four of the most commonly used protein-protein docking algorithms: SwarmDock[15], pyDock[28], ZDOCK[2] and HADDOCK[17]. As reported by Vreven *et al.* while setting the benchmark set used in this work[27], their acceptable success rate for rigid-body cases ranges between 31% and 50%, whereas for the medium and difficult cases substantially lower success rates (between 4% and 22%) are observed. Regarding intermediate success rates, 13 to 18% success rates are reported. It is only when considering the percentage of high-quality models, where success rates <6% are reported, that JabberDock is outperformed.

The reason behind JabberDock's consistent performance throughout cases of different difficulty lies in its ability to correctly identify interfacial amino acids. Indeed, while the RMSD of models *versus* the known complex is lower for cases with more flexible subunits, the average ratio of correct contact residues (f_{nat}) remains nearly unaltered (Figures 3(b) and S8). The relationship between the number of candidate models selected from JabberDock's ranked solutions and the resulting success rate features an initial steep gradient (Figure 3(c)). This indicates that the ranking of JabberDock's first successful model is most likely to be high. Still, by increasing the number of candidate models to 100, results with significantly smaller RMSD and higher f_{nat} can be found (Figure S8). Thus, while most successful models usually rank high according to our scoring function, better models may well be available when we consider a larger pool of solutions. For instance, in 98.6% of easy cases, our full datasets of 300 solutions always contain at least one acceptable pose (Figure 3(c)).

By aligning each monomeric subunit to their counterpart in the complex and assessing the score achieved by such a pose, we observed that four of the easy and one of the difficult cases yielded scores higher than anything found by JabberDock (see Table S1). One example is the xyloglucan-specific endo-beta-1,4-glucanase (PDB: 3VLB), where the two binding partners are highly interlocked. In this case, unsuccess was not caused by an unsuitable scoring function, but by an underperforming optimiser, which was unable to navigate into the complex binding site. Many of the successful unbound cases feature interlocked arrangements. In such cases, if the optimiser can identify the narrow set of roto-translation allowing the binding partners to interlock, the resulting model will have a high score. A successful example is that of the *β*-Lactamase TEM1 (PDB: 1BTL) – Ribonuclease A (PDB: 9RSA) complex, involving a significantly large and complex contact region, whereby almost the entire circumference of *β*-Lactamase's STID map is buried (see Figure 3(a)).

Unsuccessful cases, such as the Profilin – *β*-actin complex (PDB: 2BTF), most often feature a flat binding site. In these cases, the surface complementarity score alone struggles to discriminate between binding and non-binding regions due to a lack of characteristic surface features, and thus successful models do not rank high. Addressing these cases requires capturing additional properties of protein-protein interactions. To this end, we explored the possibility of reranking JabberDock models accounting to the vectoral alignment of neighbouring dipoles at the interface *via* the dipole maps used to build the STID maps (see Supplementary Data). Preliminary results indicate that such a post-processing reranking, while not increasing the overall success rate, significantly improved the quality of poses for 12% of the dataset.

The most flexible model for which we had a successful prediction was the histone chaperone CIA/ASF1-double bromodomain complex (PDB: 3AAD), with an RMSD between known unbound and bound state of 4.37 Å. The 46 kDa complex formed by thioredoxin reductase (thioredoxin) and the NADP+ analogue AADP+ (PDB: 1F6M) was more flexible, exhibiting domain movements associated to an RMSD of 4.9 Å. As no top 10 model produced by JabberDock had an RMSD lower than 10 Å from the known bound state, this case was declared unsuccessful. However, a top 10 model featured a f_{nat} of 0.419 (rank 7, see Table S1), indicating that the binding site was partially identified. Thus, while in terms of RMSD several cases were unsuccessful, JabberDock could still identify their binding site (see Figure 3b).

Following from this, an interesting unbound case is that of the UBA domain from Cbl-b ubiquitin ligase (PDB: 2OOA), a small 11 kDa protein. Although it is crystallised as a homodimer, it is its monomer that participates in the formation of a heteromultimer (PDB: 2OOB) with ubiquitin. Simulating a 2OOA monomer and using its associated STID map within JabberDock to predict the 2OOB complex yielded no successful results. On the other hand, generating a STID map using a monomer extracted from the simulation of its dimer gave intermediate quality results. This indicates that the dynamics of a Cbl-b ubiquitin ligase as part of a homodimer or a heteromultimer were similar, and this similarity could be harnessed to improve the predictive power of our surface complementarity scoring. This approach was also tested with the significantly larger Integrin I domain of complement receptor 3 complex (PDB: 4M76), but in this scenario, no good pose was found. Thus, protein docking involving small proteins, and possibly very flexible ones, may benefit from information about their bound dynamics extracted from other known complexes these proteins are part of.

Discussion and Conclusion

We have presented STID maps, a strategy to represent how a molecule is perceived by its immediate surroundings. Our physical formalism encompasses the localised electrostatic nature of the space occupied by a molecule, and the dynamics of the protein itself, into a series of local dipole vectors, which is ultimately cast into a volumetric representation.

We have demonstrated that the average STID quantity of each protein is linearly anticorrelated with the ratio between protein SASA and molecular weight, S_{m}, and that typical structural elements are always discernible at the same isovalue, independently from the protein under study (Figure 2(b)). This means that proteins with similar mass and aspect, but different secondary structure will have a different $D_{\mathrm{avg.}}$ value. This is because different secondary structure elements contribute to the STID voxel system in different ways, determined by their characteristic structure and dynamics. For instance, a greater number of unstructured coils will produce a greater number of occupied voxels as the protein explores a relatively greater region of the available space, but these will have smaller associated non-zero STID values, decreasing D_{avg}. In previous electron density modelling and 3D reconstruction software yielding volumetric representations, choices of isovalue cut-off to display isosurfaces have been arbitrary[29]. Their choice is often chosen based on what is deemed by the authors to be most appropriate for the work, with no clear link made between a defining characteristic of a protein and the isosurface shown. In contrast, this work has shown that our STID map-based representations can be directly related to a physical, and easily measurable quantity.

We have then shown that STID maps representations are suitable for the definition of an accurate protein-protein docking scoring function. To this end, we have performed a comprehensive set of benchmarks to determine the optimal value of each parameter required for the construction and usage of STID maps for protein docking. The results show that JabberDock can provide predicted complexes on par with a competitive range of blind protein-protein docking software and is highly robust across a range of difficult cases – an achievement not observed in other docking algorithms.

The strength of JabberDock to yield comparable results across the dataset indicates that the ability of the STID maps to encapsulate high-frequency atomic motions accommodates for different amounts of flexibility in interacting proteins. In case of complexes characterized by flat and relatively featureless binding sites, the surface complementarity function is likely to fail in highlighting a single most suitable docking position. On the other hand, when an interface is exceedingly complex, small perturbations about the docked pose are likely to lead to clashes, hindering the optimiser in its exploration of this region of the energy landscape. These represent JabberDock's boundary conditions. The successful (and most typical) docking cases feature topographical complexities that enable both the scoring function and the optimiser to work harmoniously and effectively. This is the significant middle ground where the coupling of POW^{er} and STID maps provide excellent results, as indicated by the prediction of accurate protein complexes for most of the benchmark. These observations are expected to hold for any complex not requiring refolding or domain-level movements at the interface between binding partners.

JabberDock utilises two individual PDB files with a well-parameterised force field to generate STID maps. These are used to guide the docking process, at the end of which all candidate models, typically several thousand, are clustered and returned to the user (in this work, 300 solutions are returned). The models themselves are built using the last snapshot from the pool of conformations explored by the binding partners in their respective MD simulations. Combinations of other conformations within the monomers' simulations (see Figure S9), and dimeric arrangements within the full collection of candidate assemblies may be closer to that found in the crystallised bound state. In future versions of JabberDock, we will explore the possibility of leveraging on this additional source of structural information to provide the user with more accurate models.

We have observed that using a protein's STID map reporting on its dynamics when bound to an alternative complex can improve JabberDock's performance. While our benchmark shows that docking proteins represented by the STID map of their monomeric state yield a good number of successes, more accurate predictions may be obtained by harnessing the dynamics of the bound complex. Other areas of future investigation will include the adoption of different functions as a model for the pseudo-electron density (e.g., a Lorentzian), using JabberDock for rescoring models predicted by other protein docking methods, a reranking process building upon our preliminary results on the usage of a dipole complementarity score, the use of a different atomistic forcefield (including a polarizable one) to explore the impact on the STID maps, and an additional post-processing step based on MD or MC techniques to refine the best docking poses. In this context, we also foresee that the use of optimisation algorithms requiring no weighting could be beneficial[30]. Overall, enhancements in the scoring function and solutions reranking will help improve the performance of JabberDock against cases with low interface complexity, whilst refining the optimisation engine will reinforce its performance against cases with highly complex ones.

Supporting Information

- Analysis of side chain motion convergence during MD simulations; determination of voxel size parameters for STID maps; benchmark on the best choice of isovalue for the STID map and cut-off for the distance between two surfaces during docking; mathematical details for the dipole alignment re-ranking method that can be employed; information on computational resources used and associated execution times; analysis of the quality of the best docked pose within the top 100 results for each protein; analysis of RMSD variation observed during the MD simulations; comparison of docking score vs RMSD and f_{nat} for three example protein complex cases; analysis of relationship between mass of protein complex and time to complete the POW$^{\mathrm{er}}$ optimisation search; Figures S1-10 (file: Supp_Info.pdf).

- Table of results for all cases in the protein docking benchmark. It contains, for each of the easy, medium and difficult cases (and a subset of bound cases therein): the highest ranking acceptable and intermediate results (shown in Figure 3); the lowest RMSD in the top 10 poses, with corresponding rank, f_{nat} and interfacial RMSD; the largest f_{nat} in the top 10 poses, with corresponding rank, RMSD and interfacial RMSD (file: Table_S1.xlsx).

This information is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

We wish to thank Basile Curchod, Chloe Fuller, Valentina Erastova, Ehmke Pohl and Venkata Krishnan Ramaswamy for useful discussions and for critically reviewing this manuscript. We also thank Dr. Basile Curchod for having generously shared his computational resources, and Durham University for computer time on its HPC facility, Hamilton. The work was supported by the Engineering and Physical Sciences Research Council (grant EP/P016499/1).

Methods

Molecular Dynamics

All simulations are run on the Gromacs[31] molecular dynamics engine, with Amber14sb force field[32]. Systems are prepared by immersing the protein of interest in a TIP3P water box, neutralised with Na^+ or $Cl^–$ counterions. The system is then energy minimised using a steepest descent algorithm, with a tolerance threshold set to 200 kJ $mol^{–1}$ $nm^{–1}$. The initial step size is set to 1 pm, the maximum number of allowed steps to 5×10^6. The cut-offs for both Coulombic and van der Waals interactions are set to 1.2 nm.

The protein is then equilibrated for 500 ps within a canonical ensemble, with T set to 310.15 K with 2 fs step size, and the constraint algorithm LINCS applied to the bonds[33]. A particle mesh Ewald summation is used to treat long-range interactions, and a velocity-rescale temperature coupling method applied separately to protein and non-protein atoms, the coupling constant is set to 0.1 ps. Velocities are randomly assigned from a Boltzmann distribution of velocities at T.

Finally, production occurs over a 500 ps timescale, for reasons shown in Figure S1, in an isothermal-isobaric ensemble. T is set as above; the pressure is set to 1 bar. Berendsen temperature and pressure coupling methods are used, again keeping the protein and non-protein groups separate. The temperature coupling is as above, with the pressure coupling constant set to 10 ps. The compressibility for both is set to $4.5 \times 10^{–5}$ $bar^{–1}$. Atomic coordinates are saved every 5 ps.

Particle Swarm Optimisation

An initial starting point with the two input monomers' centres of mass centred at the origin is used prior to generating any models. JabberDock uses a seven-dimensional space for implementation comfort when roto-translating the STID maps. Three dimensions define ligand translation in the Cartesian space, three dimensions define an axis of rotation for this ligand, and one dimension defines a rotation angle around this axis. Translation values are limited by the size of the receptor, the axis of rotation is normalised (and thus has values ranging from -1 to 1), and the rotation angle in radians ranges between 0 and 2π.

In order to navigate the potential energy surface (PES) associated with the scoring function (see next section) and produce an ensemble of possible docked poses, JabberDock leverages a distributed heuristic global optimization algorithm featured in the POW^{er} optimisation environment – particle swarm optimisation "kick and reseed" (PSO-KaR). [20] PSO-KaR was used to explore the PES over 300 iterations using 80 randomly initialised agents ("particles"). According to the "kick and reseed" procedure, particles converging to a local minimum (i.e. with a velocity decaying to less than 4% of the search space dimension in each direction) were randomly restarted, and a repulsion potential placed at their convergence location. The whole optimization process was repeated three times, with the memory of previous repulsion potentials retained from one repetition to the next. In sum, this docking procedure requires the evaluation of 72000 docking poses. To obtain a diverse ensemble of solutions, 300 poses were finally selected as representatives from the pool of poses having a positive score using a *K*-means clustering algorithm on the 7-dimensional coordinates associated with each model.

JabberDock's Scoring Function

JabberDock uses a surface complementarity assessment that takes advantage of the STID maps to generate the PES explored by the particle swarm optimisation algorithm implemented in the POW^{er} optimisation engine.

Following a roto-translation of a model requested by the optimiser, a quick test is first performed to identify poses featuring no contact, or unphysical atomic overlaps between the ligand and the receptor. Suitable poses, featuring a negative Lennard-Jones potential between the alpha carbon atoms of receptor and ligand, are scored according to their surface complementary *S*. The shape of the isosurfaces analysed by JabberDock are determined by an isovalue cut-off, an appropriate value of 0.43 was chosen based on the benchmark discussed in Figures S6-7. The score between the STID of the receptor and that of the ligand is given by:

$$S = \frac{\{S_{\mathrm{AB}}\} + \{S_{\mathrm{BA}}\}}{2}\ , \qquad (7)$$

where the curly brackets indicate that we used the median of the score for protein A into B and *vice versa*, where the scores are given by:

$$S_{\mathrm{AB}} = \left(\mathbf{n}_{\mathrm{A}} \cdot \mathbf{n}_{\mathrm{B}}'\right)\exp\left(-w\left|\mathbf{x}_{\mathrm{A}} - \mathbf{x}_{\mathrm{A}}'\right|^2\right)\frac{v_{\mathrm{A}}}{V_{\mathrm{A}}}\ , \qquad (8)$$

where $\mathbf{n}_{\mathrm{A}}$ is the normal from a region of interest on A's surface, $\mathbf{n}_{\mathrm{B}}$' the anti-normal from the closest point on B to that point on A. w (0.5 Å^{-2}) is an arbitrary weighting found by Lawrence & Coleman[34], $|\mathbf{x}_{\mathrm{A}} - \mathbf{x}_{\mathrm{A}}'|$ is the physical distance between the two points. v_{A} is the total number of successful contact points on A in contact with B inside some arbitrary distance cut-off, an optimal cut-off of 1.6 Å was chosen based on the work discussed in Figures S3-5. V_{A} is the total number of points describing the surface of A. These last two terms are used to avoid minor contact points providing good scores. The larger *S*, the better the fit, thus the optimiser is set up to maximise the score. Only positive scores are accepted by POW$^{\mathrm{er}}$. Figure S10 provides examples of how these scores match with a corresponding RMSD and f_{nat} for three complexes.

STID benchmark

118 non-redundant proteins (maximum 30% homology) were extracted from the PDB-REDO databank[34]. All structures were soluble proteins featuring solely standard amino acids, none required the application of a biomatrix, and all were composed of more than 30 amino acids.

The SASA of each structure in the benchmark set was calculated using the Shrake-Rupley algorithm[35], with the solvent probe radius set to 1.4 Å to represent that of water. For each protein, we report the average SASA over 500 ps production cycle (one structure every 5 ps, excluding the first 50 ps). Molecular weights were calculated accounting for all atoms present in the atomic structures.

Case Difficulty Classification

Protein-protein docking cases are classified under three levels of difficulty which is associated with their flexibility, and RMSD difference between the Cα atoms at the interface after superposing the bound and unbound interfaces. Cases can be classified as either rigid-body (or easy), medium difficulty or difficult. Easy cases are those with minimal difference between the unbound crystallised structures and the bound: usually < 1 Å difference. In medium cases, the RMSD difference is between 1 Å and ~ 2.5 Å. Finally, difficult cases can be anything greater than 2.5 Å. Thus, the difficult cases are accordingly significantly more difficult than the other two, particularly given that the requirements for an acceptable success are close to the upper boundaries that define the difficult cases.

Assessment of Models Accuracy

We use three metrics to determine the quality of a model: the ratio of correct contact residues (a valid contact defined as an atom within 5 Å of the binding partner) to the number of residues in the predicted complex, f_{nat}, the RMSD between the alpha carbons of the known crystal pose and the predicted pose, and the RMSD of the two poses between the alpha carbons at the interface (defined as within 10 Å of the binding partner). CAPRI guidelines specify four levels of possible success criteria: (1) incorrect, where RMSD > 10.0 Å and interfacial RMSD > 4.0 Å OR $f_{\mathrm{nat}} < 0$; (2) acceptable quality, where RMSD ≤ 10.0 Å or interfacial RMSD ≤ 4.0 Å and $0.1 \leq f_{\mathrm{nat}} < 0.3$ OR $f_{\mathrm{nat}} \geq 0.3$ and RMSD > 5.0 Å and interfacial RMSD > 2.0 Å; (3) intermediate quality, where RMSD ≤ 5.0 Å or interfacial RMSD ≤ 2 Å and $0.3 \leq f_{\mathrm{nat}} < 0.5$ OR $f_{\mathrm{nat}} \geq 0.5$ and RMSD > 1.0 Å and interfacial RMSD > 1.0 Å; (4) high quality, where RMSD ≤ 1.0 Å and interfacial RMSD ≤ 1.0 Å and $f_{\mathrm{nat}} \geq 0.5$. The protocol for applying this list of inequalities follows the order provided, beginning with defining the incorrect predictions. In the text, we qualify the result of a test as of high, intermediate or acceptable quality if at least one in the top 10 ranked models matches the criteria above.

Software implementation

Software to generate STID maps is developed in Python, using numpy, scipy and cython packages. An automated bash script that prepares all the necessary Gromacs files and runs them is used to generate the trajectories. JabberDock is implemented as a POW$^{\mathrm{er}}$ Python module. All software is freely available at github.com/degiacom/JabberDock.

References

(1) Winograd-Katz, S. E.; Fässler, R.; Geiger, B.; Legate, K. R. The Integrin Adhesome: From Genes and Proteins to Human Disease. *Nat. Rev. Mol. Cell Biol.* **2014**, *15* (4), 273–288.

(2) Chen, R.; Li, L.; Weng, Z. ZDOCK: An Initial-Stage Protein-Docking Algorithm. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 80–87.

(3) Tovchigrechko, A.; Vakser, I. A. GRAMM-X Public Web Server for Protein-Protein Docking. *Nucleic Acids Res.* **2006**, *34* (WEB. SERV. ISS.), W310–W314.

(4) Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S. PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (2), 392–406.

(5) Smith, G. R.; Sternberg, M. J. E.; Bates, P. A. The Relationship between the Flexibility of Proteins and Their Conformational States on Forming Protein–Protein Complexes with an Application to Protein–Protein Docking. *J. Mol. Biol.* **2005**, *347* (5), 1077–1101.

(6) Grünberg, R.; Leckner, J.; Nilges, M. Complementarity of Structure Ensembles in Protein-Protein Binding. *Structure* **2004**, *12* (12), 2125–2136.

(7) Król, M.; Chaleil, R. A. G.; Tournier, A. L.; Bates, P. A. Implicit Flexibility in Protein Docking: Cross-Docking and Local Refinement. *Proteins: Structure, Function, and Bioinformatics* **2007**, *69* (4), 750–757.

(8) Jackson, R. M.; Gabb, H. A.; Sternberg, M. J. E. Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem. *J. Mol. Biol.* **1998**, *276* (1), 265–285.

(9) Król, M.; Tournier, A. L.; Bates, P. A. Flexible Relaxation of Rigid-Body Docking Solutions. *Proteins Struct. Funct. Bioinforma.* **2007**, *68* (1), 159–169.

(10) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and

Side-Chain Conformations. *J. Mol. Biol.* **2003**, *331* (1), 281–299.

(11) Zacharias, M. Protein-Protein Docking with a Reduced Protein Model Accounting for Side-Chain Flexibility. *Protein Sci.* **2003**, *12* (6), 1271–1282.

(12) Zacharias, M. ATTRACT: Protein-Protein Docking in CAPRI Using a Reduced Protein Model. *Proteins Struct. Funct. Bioinforma.* **2005**, *60* (2), 252–256.

(13) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Geometry-Based Flexible and Symmetric Protein Docking. *Proteins Struct. Funct. Bioinforma.* **2005**, *60* (2), 224–231.

(14) Lesk, V. I.; Sternberg, M. J. E. 3D-Garden: A System for Modelling Protein-Protein Complexes Based on Conformational Refinement of Ensembles Generated with the Marching Cubes Algorithm. *Bioinformatics* **2008**, *24* (9), 1137–1144.

(15) Moal, I. H.; Bates, P. A. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int. J. Mol. Sci.* **2010**, *11* (10), 3623–3648.

(16) Fernández-Recio, J.; Totrov, M.; Abagyan, R. ICM-DISCO Docking by Global Energy Optimization with Fully Flexible Side-Chains. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 113–117.

(17) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737.

(18) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biol.* **2012**, *10* (1), e1001244.

(19) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102.

(20) Degiacomi, M. T.; Dal Peraro, M. Macromolecular Symmetric Assembly Prediction Using

Swarm Intelligence Dynamic Modeling. *Structure* **2013**, *21* (7), 1097–1106.

(21) Lensink, M. F.; Méndez, R.; Wodak, S. J. Docking and Scoring Protein Complexes: CAPRI 3rd Edition. *Proteins Struct. Funct. Bioinforma.* **2007**, *69* (4), 704–718.

(22) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature*. December 13, **2007**, *450* (7172), 964–972.

(23) Kirkwood, J. G. The Dielectric Polarization of Polar Liquids. *J. Chem. Phys.* **1939**, *7* (10), 911–919.

(24) Fröhlich, H. *Theory of Dielectrics: Dielectrics Constant and Dielectric Loss.*; Cole, R. H., Ed.; Clarendon Press: Oxford, 1958; Vol. 80.

(25) Neumann, M.; Steinhauser, O.; Pawley, G. S. Consistent Calculation of the Static and Frequency-Dependent Dielectric Constant in Computer Simulations. *Mol. Phys.* **1984**, *52* (1), 97–113.

(26) Fedorov, D. V; Sadhukhan, M.; Stöhr, M.; Tkatchenko, A. Quantum-Mechanical Relation between Atomic Dipole Polarizability and the van Der Waals Radius. *Phys. Rev. Lett.* **2018**, *121* (18)..

(27) Vreven, T.; Moal, I. H.; Vangone, A.; Pierce, B. G.; Kastritis, P. L.; Torchala, M.; Chaleil, R.; Jiménez-García, B.; Bates, P. A.; Fernandez-Recio, J.; et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **2015**, *427* (19), 3031–3041..

(28) Cheng, T. M.-K.; Blundell, T. L.; Fernandez-Recio, J. PyDock: Electrostatics and Desolvation for Effective Scoring of Rigid-Body Protein-Protein Docking. *Proteins Struct. Funct. Bioinforma.* **2007**, *68* (2), 503–515. https://doi.org/10.1002/prot.21419.

(29) Wriggers, W. Conventions and Workflows for Using Situs. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2012**, *68* (4), 344–351..

(30) Tamò, G.; Maesani, A.; Träger, S.; Degiacomi, M. T.; Floreano, D.; Peraro, M. D. Disentangling Constraints Using Viability Evolution Principles in Integrative Modeling of Macromolecular Assemblies. *Sci. Rep.* **2017**, *7* (1), 235..

(31) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91* (1–3), 43–56..

(32) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.

(33) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(34) Lawrence, M. C.; Colman, P. M. Shape Complementarity at Protein/Protein Interfaces. *J. Mol. Biol.* **1993**, *234* (4), 946–950.

(35) Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. The PDB-REDO Server for Macromolecular Structure Model Optimization. *IUCrJ* **2014**, *1* (4), 213–220.

(36) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79* (2), 351–371.

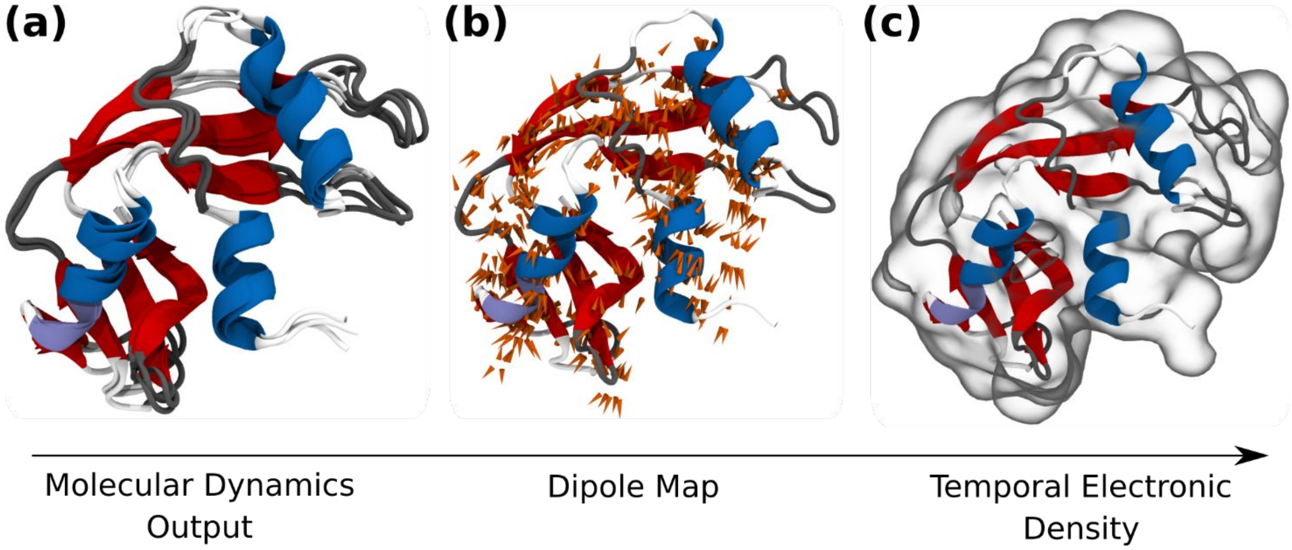


Figure 1: The pipeline for the generation of the STID maps. **(a)** Superimposition of multiple structures from the molecular dynamics simulation of Ribonuclease A (PDB: 9RSA), coloured by secondary structure (alpha helices as blue, 3_{10} helices as light purple, beta sheets as red, unstructured coils as white and turns as grey). **(b)** Superimposed dipole map generated from the simulation. For clarity, only dipoles greater than 0.8 D are shown **(c)** The final STID map, derived from the dipole map.

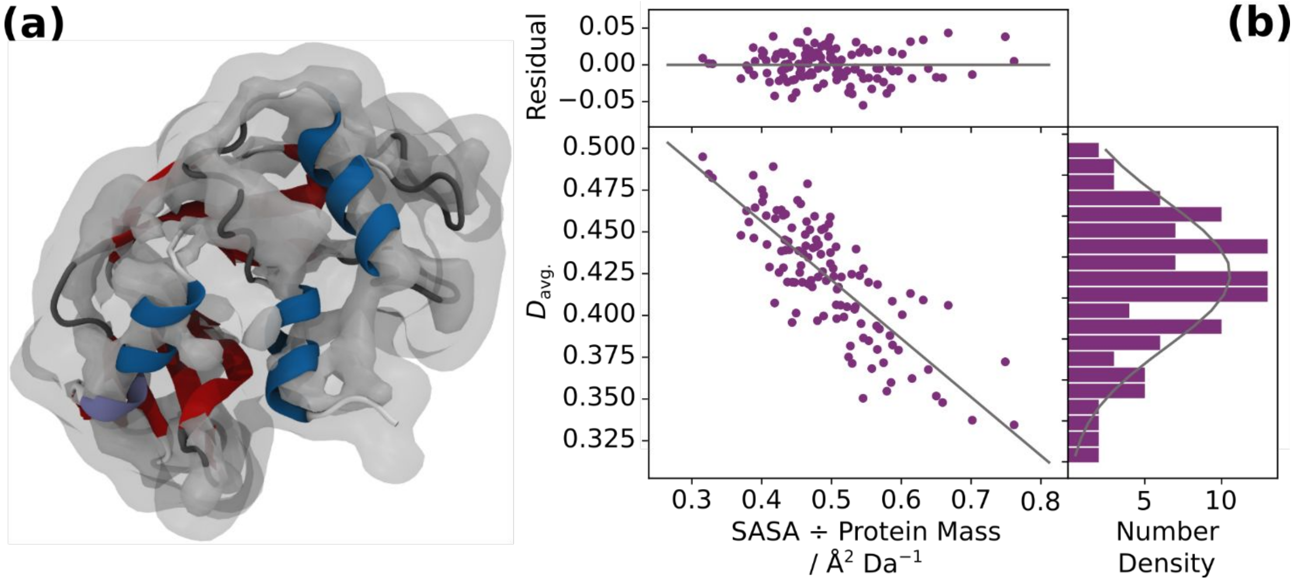


Figure 2: (a) Ribonuclease A (PDB: 9RSA) embedded in its associated STID map. Two isosurface selections are shown. The transparent isosurface, at an isovalue of 0.43, shows how the local side chains contribute to the isosurface's topography. The opaque one, at 0.8, illustrates primarily the core secondary structure features and charged residues. **(b)** Bottom left: Variation of average non-zero STID value *versus* the protein's SASA divided by its molecular weight (shown in palatinate). The fitted grey line was found via a linear least square fit, with a Pearson correlation coefficient of -0.80. Top: Residual between the points and the fitted line. Bottom right: Representation of the points as the STID average against number density in palatinate, with a non-linear least square fitted Gaussian shown in grey.

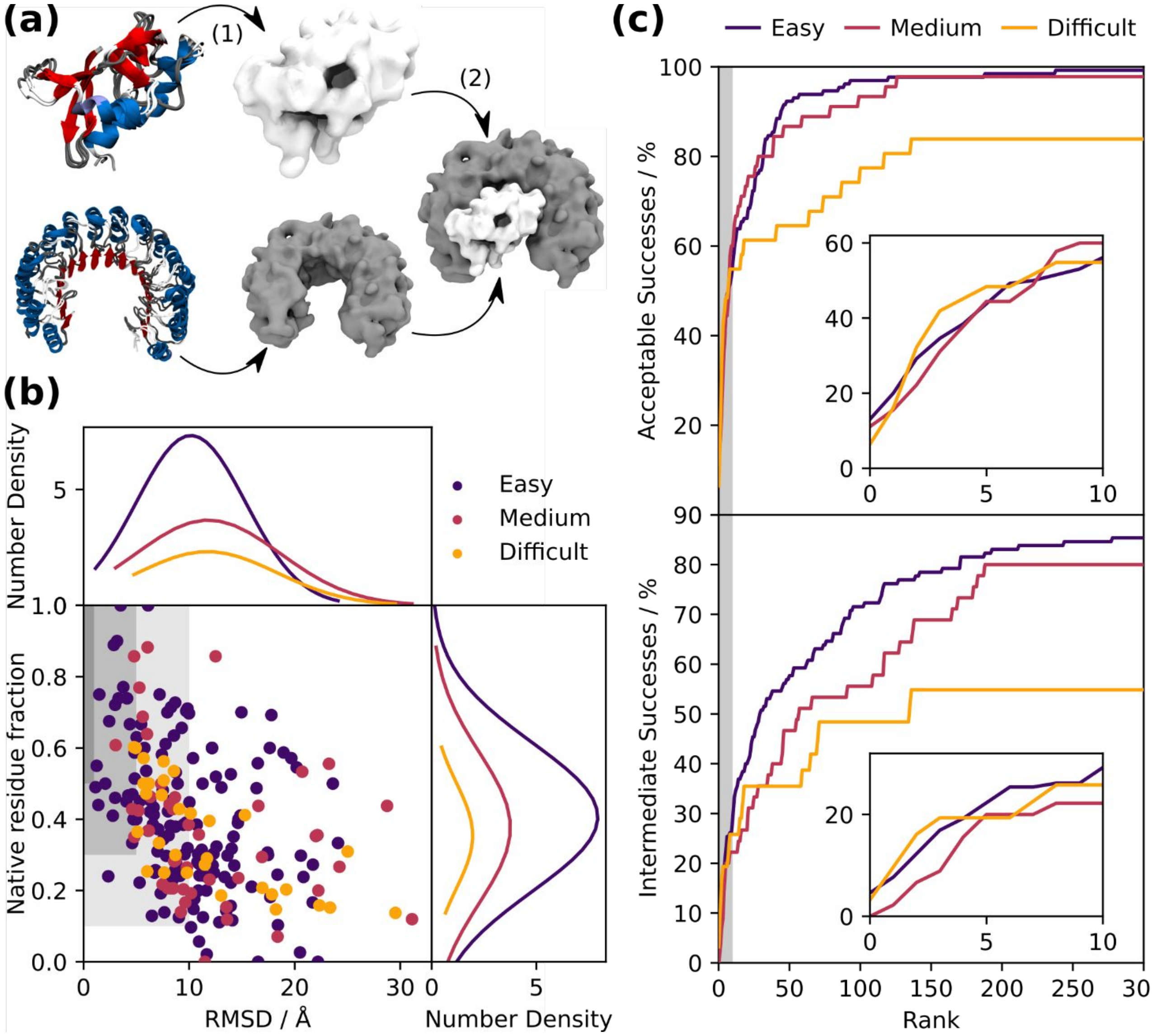
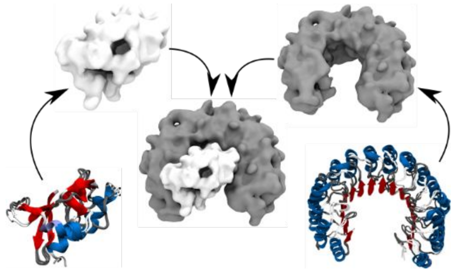
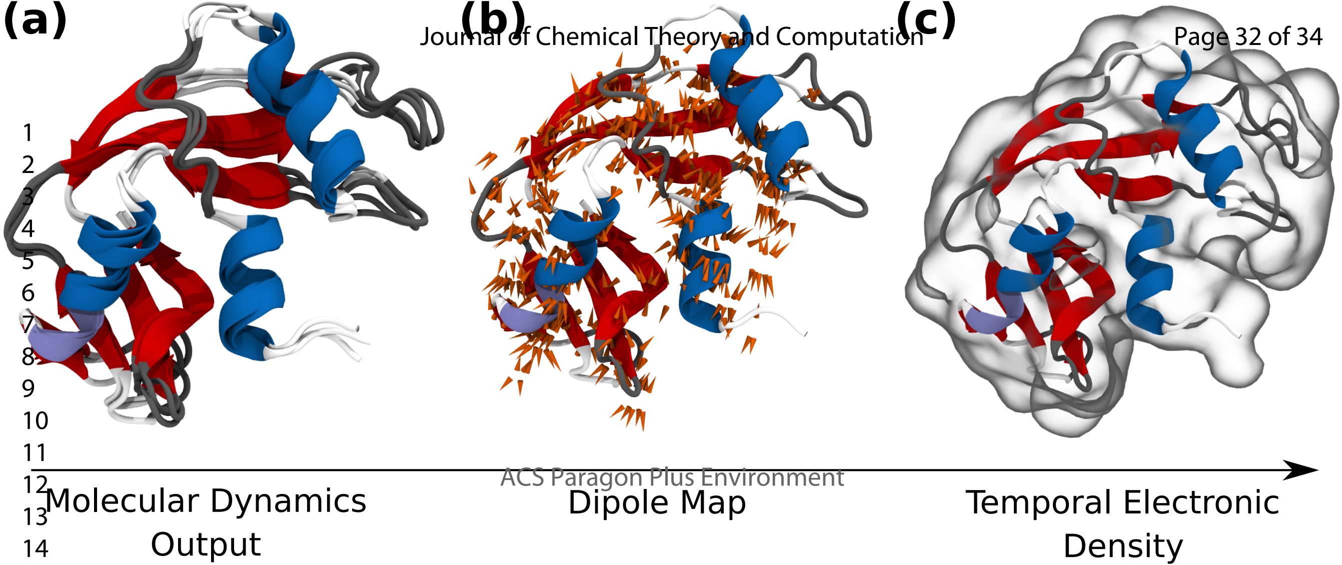


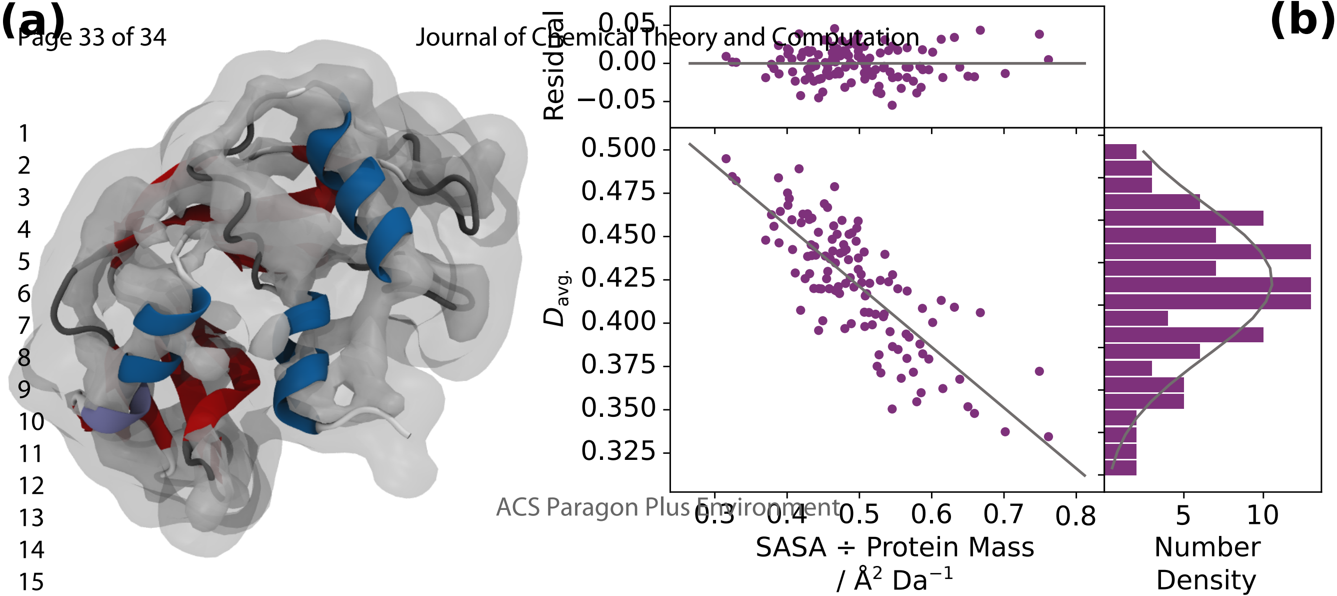
Figure 3: (a) (1) The STID map of two binding partners is calculated using their respective MD simulations. (2) The STID map representation of both binding partners is leveraged by JabberDock, our *de novo* protein docking algorithm, to accurately predict the complex. The image shows the intermediate quality model of Ribonuclease A complexed with its inhibitor (PDB: 1DFJ). **(b)** Quality of best models within the top 10 results for every docking case. For each case, the lowest alpha carbon RMSD between prediction and crystallised complex is presented, against their associated native residue fraction (f_{nat}). Point colours indicate the case difficulty, while the dark to light shaded regions represents the criteria for high, intermediate and acceptable quality results, respectively. Thus, a point landing in one of these regions indicates that the corresponding success was found within the top 10 ranked JabberDock solutions. The top and right adjoining subplots show, respectively, the distribution of RMSDs and f_{nat} across the models. **(c)** Percentage of test cases yielding an acceptable

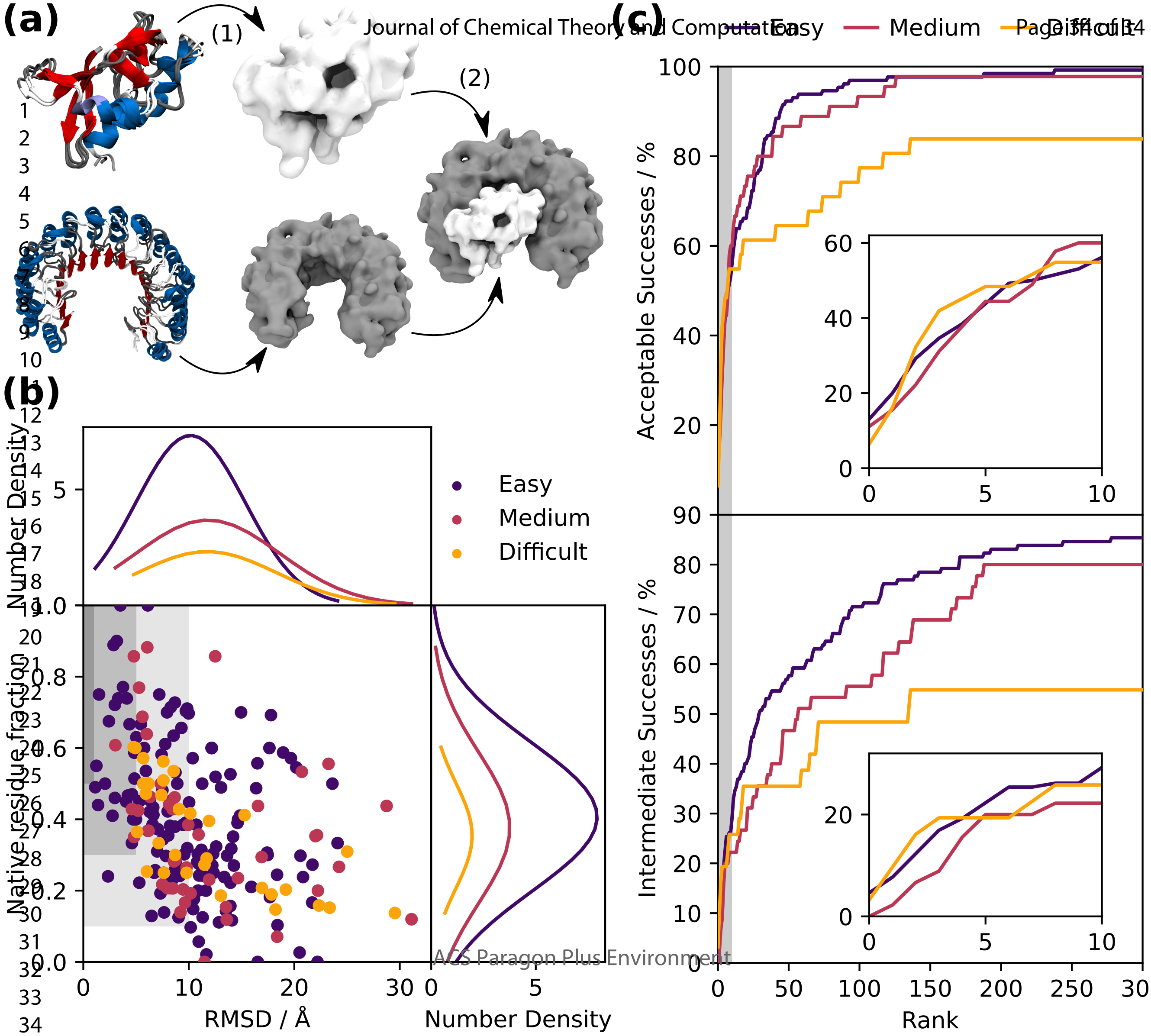
(top) and intermediate (bottom) success, as a function of the number of ranked structures considered as candidate models. Data is reported independently, in different colours, per case difficulty. The region corresponding to the top 10 models is shaded and magnified in the insets. In this region, JabberDock's success rate is consistent *versus* easy, medium and difficult docking cases. In the larger pool of 300 models, an acceptable solution is always found for the easy cases.

TOC




(a)
(b)
(c)
Molecular Dynamics Output
Dipole Map
Temporal Electronic Density


(a)
(b)
Residual
0.05
0.00
−0.05
$D_{avg.}$
0.500
0.475
0.450
0.425
0.400
0.375
0.350
0.325
0.3
0.4
0.5
0.6
0.7
0.8
SASA ÷ Protein Mass
/ Å^2 Da^{-1}
5
10
Number
Density


(a)
(1)
(2)
(b)
Number Density
5
Easy
Medium
Difficult
Native residue fraction
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
RMSD / Å
0
5
Number Density
(c)
Easy
Medium
Difficult
Acceptable Successes / %
100
80
60
40
20
60
40
20
0
0
5
10
90
Intermediate Successes / %
80
70
60
50
40
30
20
10
0
20
0
0
5
10
0
50
100
150
200
250
300
Rank