# Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics and composition

Benjamin Bouvier, Raik Grünberg, Michael Nilges, Frédéric Cazals

# Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics and composition

Benjamin Bouvier *Raik Grünberg †Michael Nilges ‡Frederic Cazals §

January 5, 2009

**Abstract**

The accurate description and analysis of protein-protein interfaces remains a challenging task. Traditional definitions, based on atomic contacts or changes in solvent accessibility, tend to over- or underpredict the interface itself and cannot discriminate active from less relevant parts. We here extend a fast, parameter-free and purely geometric definition of protein interfaces and introduce the shelling order of Voronoi facets as a novel measure for an atom's depth inside the interface. Our analysis of 54 protein-protein complexes reveals a strong correlation between Voronoi Shelling Order (VSO) and water dynamics. High Voronoi Shelling Order coincides with residues that were found shielded from bulk water fluctuations in a recent molecular dynamics study. Yet, VSO predicts such "dry" residues without consideration of forcefields or dynamics at dramatically reduced cost. More central interface positions are often also increasingly enriched for hydrophobic residues. Yet, this hydrophobic centering is not universal and does not mirror the far stronger geometric bias of water fluxes. The seemingly complex water dynamics at protein interfaces appears thus largely controlled by geometry. Sequence analysis supports the functional relevance of both dry residues and residues with high VSO, both of which tend to be more conserved. Upon closer inspection, the spatial distribution of conservation argues against the arbitrary dissection into core or rim and thus refines previous results. Voronoi Shelling Order reveals clear geometric patterns in protein interface composition, function and dynamics and facilitates the comparative analysis of protein-protein interactions.

**Keywords:** Protein-protein complex, interface activity, hotspots, conservation, Voronoi models.

# 1 INTRODUCTION

Specific recognition between proteins plays a crucial role in almost all cellular processes and most proteins are embedded in highly connected and dynamic networks of interaction partners [1]. Despite much progress in docking methods [2], identifying the exact interface between two proteins remains difficult. On the one hand, exact predictions are hindered by the complex and dynamic nature of proteins [3, 4]; on the other hand, current methods for delineating and describing even a known interface may be inaccurate or ill-chosen.

Given the structure of a complex, a protein-protein interface is traditionally defined by the 'geometric footprint', which refers to all atoms within a certain distance of the interaction partner. Somewhat more precise definitions rely on the loss of solvent accessibility (SA) upon binding [5]. Yet, it has been shown experimentally that as much as half of this footprint can seemingly be irrelevant to binding [6]. As contributions to specificity and affinity appear very unevenly distributed, substantial effort has been spent on the identification of areas or residue patches that are more actively involved in molecular recognition

*Institut de Biologie et de Chimie des Protéines, CNRS/Lyon Univ., France

†EMBL-CRG Systems Biology Unit, Barcelona, Spain. Note: R. Grünberg and B. Bouvier equally contributed to this work.

‡Unit de Bioinformatique Structurale, Institut Pasteur Paris, France

§INRIA Sophia-Antipolis, Algorithms-Biology-Structure, France. Corresponding author: Frederic.Cazals@sophia.inria.fr

[7, 8, 9, 10, 11]. Single residue mutation studies thus pointed to an usually rather small [12] number of 'hotspot' [13] residues with a dominating influence on binding free energy. Beyond this focus on isolated residues, recent studies have revealed strong non-additive, collective effects [14] indicating a modular organization of interfaces into interaction clusters [15].

The functional relevance of an interface residue or patch is also expected to leave traces in the evolutionary record. Guharoy and Chakrabarti have shown that an interface core tends, on average, to be more conserved than the rim [16]. Nevertheless, the difference in average conservation between rim and core observed by Guharoy and Chakrabarti is rather subtle and becomes significant only when studying large sets of complexes. Unlike catalytic sites, which are highly unlikely to transform in a series of discrete steps without complete loss of activity [17], the assembly of proteins involves a continuous scale of binding modes, from transient to stable, leaving more freedom for evolution to proceed in incremental steps [18, 19, 20]. Picking up such a subtle evolutionary signal is difficult. Furthermore, the coarse partitioning of interfaces into either rim or core is likely overlooking finer trends. In a recent molecular dynamics study, Mihalek and coworkers [21] tried to relate evolutionary conservation with a more biophysically meaningful property and demonstrated that conserved residues tend to be excluded from direct solvent exchange. They thus suggested to define interface activity based on the simulation of water dynamics.

The dynamics of the surrounding water is unarguably a decisive factor in protein protein binding. Binding free energies are dominated by entropic terms which arise from changes in the dynamics of the solvent (hydrophobic effect) and the protein [4, 22], both of which are intimately coupled [23, 24]. The removal of water from partially solvated backbone hydrogen atoms has been suggested as a driving force of binding [25, 26]. Along the same lines, hotspots tend to be isolated from bulk solvent [27], and interaction modules are lined by interfacial water [15]. Considering the importance of the issue we still know very little about the interplay of solvation, water dynamics, and interface structure. How does water dynamics relate to the geometry of the interface? And which proportion of it can be directly inferred from a structural picture?

We here address this issue by overcoming a central limitation of previous works in the field: the lack of a precise and rigorous measure for protein-protein interface geometries with which to correlate other observables, whether experimental or computed. Building upon the framework of Voronoi models [28], we divide protein interfaces into concentric shells. This yields a fast, robust, accurate and parameter-free measure for the depth of any atom or residue inside an interface, which we call its Voronoi Shelling Order (VSO). We demonstrate the advantages of this measure by comparing VSO with water dynamics, evolutionary conservation, and residue type on a set of 54 known complexes.

Despite the minimalist character of our model, we show that residues remote from the rim very closely match those identified as "dry" in the complex simulations of Mihalek et al. Residues with high VSO also tend to exhibit higher conservation. However, the detailed rim to core distribution of conserved residues reveals deviations from this overall trend, which were not captured by previous, more arbitrary, rim-core partitions.

Apart from providing a novel and cost-effective approach to the prediction of interfacial water dynamics, our model thus sheds new light on the relationship of structure, activity and solvation. Various physico-chemical, functional and dynamic properties point to a general structuring of protein interfaces from center to rim. The Voronoi Shelling Order measure allows us to examine this architecture with more accuracy compared to methods that solely partition the interface into either core or rim.

# 2 THEORY

## 2.1 Voronoi description of protein-protein interfaces

Since the early work of Gerstein and Richards [29, 30], Voronoi models and related constructions such as the Delaunay triangulation or the $\alpha$-complex have been prominent in modeling proteins and their interactions. Example applications of these geometric complexes have been reviewed in [31] and include the calculation of packing properties of atoms [30], the definition of molecular surfaces [32], the enumeration of atom contacts for statistical potentials [33], the investigation of pockets within macro-molecules [34, 35], but also the definition of the Voronoi interface of a protein complex – the selected Voronoi facets separating the two partners. Although all these applications elaborate upon Voronoi diagrams, the actual models differ in, at least, two major areas. The first one is the type of Voronoi diagram used, be it an affine diagram, i.e. a power diagram [36], or a curved diagram [37]. Moreover, different solutions have been proposed for the treatment of unbound or ill-defined Voronoi cells. The problem may be tackled by explicit solvation [38] or by more elaborate mechanisms. Ban et al. [39] defined the interface from the $\alpha$-complex associated to the Voronoi diagram, based upon a retraction process mixed with topological persistence. But the interface atoms selected this way are not qualified in terms of solvent accessibility, and structural water molecules are not handled. Our current study is based on a more recent interface model [28] that resolves these limitations. Our model accommodates structural water, it allows the processing of large Voronoi facets based on the orthogonality properties of spheres encoded in the Voronoi diagram, but also affords a fine description of the geometry/topology as well as the biochemistry of the interface.

For the sake of completeness, we now briefly recall this interface model. Assume we wish to model the interface between two proteins $A$ and $B$. The $AB$ interface consists of the Delaunay edges found in the 0-complex – the $\alpha$-complex for $\alpha = 0$, and whose endpoints are such that one atom lies in protein $A$ and the other in protein $B$. (An edge belongs to the 0-complex if the two Voronoi regions are neighbors, and the balls clipped to their respective regions have a non empty intersection.) Because of the duality between the Delaunay and Voronoi representations, the interface can also be described using the Voronoi facets dual to the aforementioned edges. The interface model can be extended to accommodate interface water molecules $W$, defined as sharing at least one edge with each partner in the 0-complex. This allows for the definition of the following interfaces: $AB$ between the protein partners; $AW$ (resp. $BW$) between partner $A$ (resp. $B$) and interface water; $AW - BW$ as the union of the interfaces $AW$ and $BW$; $ABW$ as the union of the interfaces $AB$ and $AW - BW$. Like methods based on the loss of solvent accessibility, our model correctly identifies any atom losing solvent accessibility as an interface atom. Unlike these methods however, it also detects interface atoms that do not lose solvent accessibility – essentially buried backbone atoms – which represent a non-negligible 13% of the interface [28].

## 2.2 Shelling the $ABW$ interface

We attribute a Voronoi Shelling Order (VSO) to each facet of the $ABW$ interface. VSO represents the number of 'jumps' between adjacent facets that needs to be performed, from the currently considered location, to reach the rim of the interface (Figures 1a and 2a). The Voronoi interface is thus partitioned into concentric shells of increasing shelling order.

The calculation of VSO values for all interface facets requires two passes. During the first pass, boundary Voronoi facets located at the rim of the interface are enumerated and given a VSO of one. Voronoi facets are bounded by Voronoi edges, each of which is incident to exactly three Voronoi facets in the Voronoi diagram; however, some of these facets may not belong to the interface (their dual Delaunay edges are not in the 0-complex). This allows us to detect rim Voronoi facets as the ones featuring at least one Voronoi edge that is incident to one interface Voronoi facet only. The second pass explores the interface breadth-first starting from the previously identified rim facets. Given an interface Delaunay edge (of shelling order $n$), the algorithm checks all incident Delaunay triangles, as each such triangle contributes zero, one or two additional interface edges. If these have not already been shelled, they are given a VSO of $n + 1$.

The outcome of this process is the association of an integer VSO value to each Delaunay edge (or equivalently, Voronoi facet) of the $ABW$ interface. However, our ultimate goal is to quantify the depth of any given atom inside the interface. This is done by tagging the atom with the minimum value among the shelling orders of the Delaunay edges to which the atom contributes (Figures 1b and 2b). The maximum or average values have also been considered, but their variation throughout the interface were found to closely mimic that of the minimum. Finally, the shelling order of a residue is defined as the average VSO value over its constituent atoms contributing to the Voronoi interface. Figure 2 provides an example for a protein-protein interface described in terms of Voronoi shells.

# 3 RESULTS

## 3.1 Voronoi shelling order, water dynamics, conservation and polarity

Voronoi Shelling Order (VSO), despite being a purely geometric measure, appears to correlate with more complex biophysical and functional traits. In Figure 3 we have color-coded three different residue properties on the homodimer interface of complex 2DOR – the shelling of the same interface is shown in Figure 2. On this example, high shelling order seems to coincide with high conservation. Indeed, the three "core" patches of the interface seem to bear the highest selective pressure. Residues with high shelling order also tend to be excluded from exchange with bulk water. In their recent simulation study [21], Mihalek et al. concluded that residues that are shielded from mobile water molecules are more conserved and thus related to the active part of the interface. We here adopt the same classification of residues into dry (shielded) or wet and, in the example of 2DOR, dry residues cluster towards patches with high VSO. Presumably, this water dynamics should be strongly affected or even controlled by the pattern of residue polarity. However, in case of 2DOR, some of the high VSO positions are indeed held by unpolar residues but the correlation is far from perfect.

We now extend our analysis to the full set of 54 homo- and heterodimer complexes initially studied by Mihalek et al. [21] and start out by quantifying how well Voronoi Shelling Order (VSO) is able to predict the rate at which residues in protein-protein interfaces exchange surrounding water molecules. We also examine the correlation between VSO and conservation, to gather information on the spatial distribution of conserved residues at the interface. We then compare these figures to the previously established correlation between conservation and dryness and, finally, explore the structuring of residue polarity across the interface.

In most of these cases, we have to correlate a continuous measure (VSO or conservation) with a binary classification of residues (dry or unpolar). Such connections are typically assessed with ROC plots (Receiver Operating Characteristic) [40] which evaluate the sensitivity and specificity of a prediction over a range of threshold values. The Area Under the ROC Curve (AUC) corresponds to the probability that the continuous measure will correctly rank a randomly chosen positive residue higher than a randomly chosen negative one [41]. It thus quantifies the predictive power or correlation of the score (VSO, conservation) with respect to the binary classification (dry, unpolar). An area of 1.0 corresponds to a perfect prediction, which in the case of Voronoi shelling order predicting dryness would mean that the $n$ dry residues in the interface perfectly match the $n$ residues with highest shelling order. By contrast, a ROC area of 0.5 corresponds to the performance of a pure random score. A p-value quantifies the statistical significance of each AUC, which is influenced by the total as well as the number of dry residues. See also section 5.4 for details.

We generate four ROC plots for each complex, describing the performance of Voronoi shelling order as predictor of dryness, of conservation as predictor of dryness, of conservation as predictor of shelling order and of shelling order as predictor of residue polarity, respectively. Our results are compiled in tables 1 and 2 for heterodimers and homodimers, respectively, and summarized in Figure 4. Evidently, Voronoi shelling order is a very good predictor of dryness and always performs better than a purely random classifier. Moreover, the relation between VSO and water dynamics is of high statistic significance, not only for the overall set of heterodimers ($P=6*10^{-74}$) and homodimers ($P=2*10^{-265}$), but even for each of the 18 homodimer and each of the 36 heterodimer complexes considered alone (see the two last rows of tables 1 and 2). Note, higher significance of the overall signal among homodimers is merely owed to the larger number of these complexes. Our analysis does not, in fact, reveal systematic differences between hetero- and homodimer interfaces, as becomes apparent from the very similar average ROC areas.

For comparison, we also reproduce and quantify the previously established relation between conservation and "dryness" of a residue [21]. As expected, the overall signal is significant for both heterodimers ($P=6*10^{-14}$) and homodimers ($P=6*10^{-43}$). Nevertheless, there are also many individual complexes for which conservation fails to predict water-shielding significantly better than a random classifier. The null hypothesis is rejected *only* 8 out of 18 heterodimers, and 25 out of 36 homodimers.

Tables 1 and 2 also quantify the ability of sequence conservation to predict VSO which would indicate

higher evolutionary constraints on central interface positions. We define the $n_{core}$ residues with highest VSO as 'core' and the remainder as 'rim' and test the ability of conservation to discriminate between the two. For a fair comparison, we adjust $n_{core}$ for each complex so as to exactly match the number of residues classified as dry. We thus tie ourselves to a threshold chosen by Mihalek et al. [21] rather than optimizing our own. Moreover, unlike Mihalek et al., we do not exclude catalytic residues from the analysis. Such as dry residues, also residues with high VSO tend to be more conserved than the ones found closer to the rim of an interface. Average ROC areas for the prediction of water shielding and of Voronoi shelling order from conservation are comparable and the overall trend is statistically significant (P=$2 * 10^{-09}$ and P=$4 * 10^{-20}$ for VSO prediction in heterodimers and homodimers, respectively). The AUC reported is statistically significant 8 out of 18 heterodimers and 14 out of 36 homodimers. For homodimers, the relatively poor prediction of high Voronoi Shelling Order from conservation may indicate a somewhat more direct connection of sequence conservation to water shielding but may as well be a consequence of the arbitrary threshold or different treatment of catalytic residues.

Unpolar and aromatic amino acids are less likely interacting with water and are expected to cluster towards the center of protein interfaces. Indeed, high VSO is often predictive of unpolar or aromatic residues. As shown in tables 1 and 2, this is the case for 11 out of 18 heterodimers and for 27 out of 36 homodimers. Overall, this trend is statistically significant (P=$1 * 10^{-21}$ and P=$2 * 10^{-63}$ for VSO predictions in heterodimers and homodimers, respectively) but it, evidently, does not hold for every single interface. The correlation is therefore again considerably weaker than the connection from Voronoi Shelling Order to water shielding.

Sequence conservation thus supports the functional relevance of both water shielding and Voronoi Shelling Order but cannot outline core or dry residues in all individual interfaces. By contrast, we observe a very strong connection between the structure-based Voronoi Shelling Order and the simulation-derived water shielding of a residue. This trend cannot be explained by a simple clustering of hydrophobic residues and has implications for the dynamics of interfacial water, as discussed further below.

## 3.2 Spatial distribution of conserved residues

Our ROC curve analysis of conservation reduces VSO to a binary classifier (akin to core or rim) and tests the hypothesis that core residues should coincide with the most conserved and, reciprocally, rim residues with the least conserved part of the interface. It thus provides insight into the location of extreme conservation values and confirms previous findings [16]. However, beyond simplified classifications into core and rim, the VSO measure also allows for a finer analysis which we expect to better capture the spatial distribution of conservation. Figures 5 and 6 show the distribution of conservation scores across Voronoi shells. For comparison, both conservation and VSO were normalized to their respective maximum. The relationship between residue conservation on the one hand and depth within the interface on the other, is evidently not a simple one. The original values (crosses) highlight the scattering of conservation across shells: highly conserved residues are found even at the very rim. We therefore filter the signal through a running average over a window comprising 1/4 of all interface residues (black line). The curves remain very similar for window sizes between 1/8 and 1/2 of the interface (data not shown). This running average indeed reveals correlations between increases in shelling order and residue conservation. In line with the ROC analysis above, the correlation holds for many but not all complexes. Nevertheless, apart from the few obvious exceptions, closer inspection also reveals some interesting systematic deviations: (i) Conservation density often reaches its maximum before the innermost shell – the interface center thus appears under less constraint than a surrounding outer core; (ii) contrary to the overall trend, a secondary peak of conservation is sometimes apparent at the very edge of the interface.

The in-depth examination of average conservation thus confirms the general trend of higher conservation towards core shells but also hints at a more complex fine structure. It demonstrates the added value of a continuous rim-to-core measure over an arbitrary binary classification.

### 3.3 Case-studies: best and worst case scenarios for shelling order

To identify in more detail the incentives and shortcomings of using shelling order for the description of interfaces and as a predictor of water dynamics, we focus on three extreme cases of application, which are presented in Figure 7.

*The ideal case.* The interface of the homodimer complex 1E2D (left) features a compact and planar core composed of a single patch of atoms with high shelling orders (large panel), which the MD simulations of Mihalek and coworkers also identify as dry (lower left-hand panel). Such compact interfaces with disk-like topologies and no holes represent best case scenarios for the predictive power of our model. Central and dry residues are also more conserved. However, in contrast to shelling order, the conservation score delimitates a patch which extends far beyond the dry residues, resulting in a good sensitivity but a poor selectivity.

*Stacks of water molecules.* The interface of the homodimer 1L5W is quite extensive and highly non planar, consisting of two 'prongs' separated by a cleft. Two high-VSO patches are found on either of the prongs. The *ABW* interface is discontinuous in the region of the cleft, due to the presence of more than one layer of solvent molecules sandwiched between the partners (Figure 8); this resets the shelling order to low values in that area. On the other hand, MD simulations find a much smaller patch of dry residues that extends inside the cleft, which means that some of the aforementioned solvent molecules are in fact structural in nature, and do not move during the simulation. A remarkable example of this occurs for tryptophane 203 (located inside the cleft, not visible on figure), which is classified as dry by Mihalek and coworkers but is surrounded by numerous water molecules on Figure 8. Here we are confronted with the main advantage of MD simulations over our model: they are able to discriminate structural water on the basis of residence times, whereas our static model relies on the fact that buried interfacial water does not usually form multiple layers. Nevertheless, it is clear from the high correlations in Tables 1 and 2 that situations featuring water molecules structured along more than one layer rarely occur; we discuss this issue further in section 4.

*Discontinuities of the interface.* Figure 7 shows a graphical representation of shelling, conservation and dryness for complex 1A59. 1A59 has an intricate topology, consisting of two monomers of predominantly globular nature linked by long 'tails' wrapped around the partner. Dry residues appear both on the globular part and on the first segment of the tail (Figure 7). Voronoi shelling order very accurately predicts the latter patch of dry residues, but over-predicts the entire tail as being dry or active, too. More interestingly, it also misses the lower part of the dry patch on the globular side of the protein. A careful inspection of the interface reveals two holes in the AB interface which reset the Voronoi Shelling Order there, preventing it from peaking in this region (Figure 9). The fact that such holes are visible in the AB interface hints at a sizable packing issue: minute defects do not usually result in such discontinuities of the AB interface [28]. Indeed, the gaps between the atoms of the two monomers span the range 5.2-6.2 Å and 5.9-6.3 Å, respectively, and could accommodate a water molecule each. (Hole 1: residues 209 to 213 (chain A) and 583 to 587 (chain B); hole 2: residues 206 to 210 (chain A) and 586 to 590 (chain B).) Since the crystal structure does not contain structural water, we cannot ascertain whether this is the case and our fast solvation procedure merely proved unable to fill the holes – even though it did successfully place isolated water molecules in three other locations.

# 4 DISCUSSION AND CONCLUSION

## 4.1 A quantitative interface definition

Among the various definitions of what exactly constitutes a protein-protein interface, the planar facets obtained from a Voronoi tessellation [42, 39] arguably come closest to the literal meaning of the term 'interface'. Indeed, such facets stem from pairs of directly interacting atoms, and provide a simpler definition of the interaction area than that required by analytical interface models [43]. The Voronoi model shows excellent correlation with classically defined curvature and solvent accessible area [28]. In comparison, the widely used geometric footprint (based on residue contacts) yields an ambiguous interaction layer which is biased towards large residues and subject to an arbitrary distance cut-off, as was further discussed in [3]. On the other hand, interface descriptions based on changes in solvent accessibility still tend to overlook certain atoms that are, in fact, direct neighbors [28]. The Voronoi definition of interfaces thus strikes a balance between a slight underprediction by solvent accessibility measures and a massive overprediction inherent to geometric footprinting. See also [31] for an in-depth review on the use of Voronoi diagrams in protein structure and interface analysis.

Here, we go beyond the binary classification of whether or not a given atom is part of the interface and furthermore quantify how many facets separate it from the edge of the interface. The idea is related to the concept of residue or atom depth [44, 45] which shows some correlation with thermodynamic properties [44] and residue conservation [46] in globular proteins. Previous studies have defined atomic depth as the simple Euclidean distance to the closest solvent molecule. By contrast, Voronoi shelling order partitions the interface into concentric shells, accounting for both the geometry and topology of the interface and appears closer to physical reality. Other studies have dissected protein interfaces into 'inner' and 'outer' or 'core' and 'rim' residues (for example, [47, 48, 49, 16]). Although a number of general trends emerge, conclusions from these works are hindered by distinct definitions of the interface combined with different classifications for core and rim. Voronoi shelling order provides a quantitative, parameter-free and unambiguous alternative to the ad-hoc classifications previously employed.

## 4.2 Shelling order and water dynamics

Voronoi Shelling Order can be interpreted as the number of atomic shells a water molecule must pass on the shortest path to a given position (facet) in the interface. This description is particularly valuable for highly curved interfaces (1A59, 1L5W...) which the Euclidean distance cannot correctly measure. We have here revealed a clear correlation between Voronoi shelling order and the 'dryness' of a residue, that is, its shielding from itinerant bulk solvent molecules. While one would expect some ties between the two measures, the extent of the agreement over a representative set of complexes is intriguing. After all, dryness had been derived from exhaustive molecular dynamics simulations which considered hundreds of additional parameters and details that are not considered by our model. On the contrary, Voronoi Shelling Order is a purely geometric property, calculated from a static set of atomic positions without any further parameter. In particular, we do *not* consider: electrostatic charges and polarity, hydrogen bonds, or any kind of fluctuations – all of which are expected to influence water dynamics. This suggests that the seemingly complex dynamic exchange of bulk solvent with interfacial water primarily depends on a simple path length and could tentatively be approximated by an analytical model of diffusion. While the geometry of diffusion fronts has been under study for two decades [50], recent developments in the realm of percolation in general and the achievements of the Fields medalist W. Werner in particular could serve as starting point for such a minimal model. (See e.g. `http://www.icm2006.org/dailynews/fields_werner_info_en.pdf`.)

Until such a more quantitative model is available, VSO can only yield qualitative predictions of water dynamics. As we show in supplemental Figures 11, 12 and 13, interface atoms with VSO higher than 4 are very likely to be dry. However, while Mihalek et al. provide an appealingly simple wet/dry binary tag for each interface residue, their definition introduces a threshold in water residence time and requires averaging over atoms of residues which often span several shells. Obtaining per-atom water residence

times to match our per-atom burial depth measure would probably provide a much finer picture of the correlation between both series. The comparatively weaker relation between VSO and hydrophobicity adds further weight to path length and interface geometry as main determinants of water dynamics.

### 4.3 Spatial distribution of conservation at the interface

The evolutionary conservation signal is of particular interest, in that it is not restricted to the interface residues of a protein-protein complex but is available for all the amino acids of the partners. However, because the conservation of a residue can have multiple causes, the evolutionary record cannot typically be used to predict which residues on the isolated partners will form an interface in the complex – hence the necessity to complement it with some other measure (like geometric footprint or change in solvent accessibility).

The quantification of evolutionary signals itself is far from trivial. Pfam sequence alignments are considered high quality but are not guaranteed to be homogeneously distributed between protein families, hereby introducing bias. Sequence families are often defined very broadly and may contain many members with similar structure but different interaction partners and surfaces. This may lead to a sizeable background of conservation signals that are not actually relevant for the particular interface under study. Moreover, for lack of relevant information in the databases, some protein stretches cannot be aligned at all: such sequences had to be pruned out of our analysis of conservation. Finally, several methods can be employed to quantify conservation. We use an entropy-based measure that has been shown to outperform other conservation scores [51]; though the choice of methods seems to have only limited impact on the specific issue of correlation with dryness [21], it should be kept in mind that alternative approaches could possibly yield diverging results.

Sequence conservation can, nevertheless, provide independent testimony of an area's importance. Mihalek and coworkers have shown that the evolutionary signal tends to peak at dry residues, confirming the notion of water shielding as an indicator of binding activity. Since our VSO measure is an excellent indicator of dryness, the correlation observed by these authors naturally translates into a tendency for the conserved residues to occur towards the center(s) of the interface, in agreement with the findings of Guharoy and Chakrabarti[16]. In contrast to the study of Guharoy et al., we are moreover able to quantify this trend also for individual complexes and find that, at least in half of the cases, it remains significant even on this single complex level.

However, the correlation between VSO and conservation, while clearly relevant (quantitatively similar to that between conservation and dryness), is much weaker than that between VSO and dryness. This hints at a complex spatial distribution for conserved residues, as shown in figures 5 and 6 and questions the simple core - rim partitioning of previous studies.

### 4.4 Methodological improvements

As previously discussed, discrepancies between dryness and shelling order arise for cases where structural (slow moving) water molecules form more than one layer inside a cavity. This is due to the fact that in our current model, interfacial water molecules must make simultaneous contact with both protein partners; any additional layer of water molecules not fulfilling this criterion will be considered as bulk and lead to the splitting of the $ABW$ interface. However, 'trapped' water molecules are known to stabilize turns and bends through hydrogen bonding with main-chain atoms in otherwise unstructured regions [52], and cannot be ignored. Their behavior is so different from that of bulk water that it is debatable whether they should be considered as delimiters for the interface, even when stacked in more than one layer – dryness results from MD simulations tend to show that they shouldn't.

A straightforward approach to alleviate discrepancies between dryness and shelling order in these difficult cases may be to optimize the threshold separating 'dry' from 'wet', instead of using Mihalek's choice [21]. Our model could also be extended so as to declare as interface water all solvent molecules $W_i$ found on a path $AW_1 \dots W_k B$ joining both partners. Using $k = 2$ or $k = 3$ could allow to infer similar properties for water molecules organized in layers, as in complex 1L5W. Nevertheless, the current

interface model, despite using $k = 1$, demonstrates that it is legitimate to infer dryness/activity from a purely geometric perspective. This effectively replaces a costly MD simulation by a very fast computation on a structure taken directly from the PDB.

Another worthwhile methodological improvement would address rare cases where discontinuities in the interface appear due to packing or solvation defects. An example thereof is the previously discussed 1A59 interface (Figure 9). Regardless of the quality of the structure or the equilibration procedure, such cases could be accommodated by using a water probe radius larger than 1.4 Å, or by devising an adaptive scheme for the value of $\alpha$ ($\alpha > 0$) employed to construct the $\alpha$-complex.

## 4.5 Conclusion

In this paper, we present a novel method to explore protein-protein interfaces. The interface is defined using the Voronoi diagram of interacting atom pairs; unlike geometric footprinting methods, atoms involved in the interface are identified with little to no over-prediction and without resorting to a distance threshold. We have shelled this Voronoi interface from rim to core, thus associating an interface depth to each atom. This Voronoi Shelling Order (VSO) shows a strong and universal correlation with the protection of residues from itinerant water fluxes, as computed by Mihalek and coworkers [21] which, in turn, can be considered a measure of residue activity. The calculation of shelling orders, however, is about five orders of magnitude faster than a typical MD simulation. Moreover, the rather accurate prediction from a simplistic and purely geometric model hints at the possibility of approximating the complex dynamics of interfacial water by analytic diffusion models. Of particular interest would be the development of a quantitative model to estimate the parameters of these dynamics as a function of the shelling order. A majority of complexes also feature a rim-to-center accumulation of hydrophobic residues but water fluxes seem, nevertheless, to be primarily shaped by the geometry of an interface rather than its composition. Comparison with evolutionary signals confirms the functional relevance of 'dry' residues and, likewise, reveals a general increase of conservation towards inner interface shells. Systematic deviations from this trend may inform about distinct binding mechanisms, catalytic activities but also modeling errors.

Much experimental and theoretical effort has been – and still is – invested into the decomposition of interfaces, for instance, by evolutionary conservation or binding affinity. By contrast, the geometrical descriptors that these measures are correlated with have hardly evolved. With a descriptor such as Voronoi Shelling Order, we can now quantify remarkable geometric patterns in the composition, function and dynamics of protein interfaces. The new measure will, hopefully, facilitate and stimulate the further study of protein interface architectures.

# 5 METHODS

## 5.1 Complex preparation

The coordinates for the homo- and heterodimer complexes listed in Tables 1 and 2 originate from the PDB database. Crystallographic water molecules were removed in order to exclude bias from different structure qualities. Missing atoms, including polar hydrogens, were added and briefly minimized. The structure was surrounded by a 9 Å layer of water molecules from an equilibrated TIP3P box. The water was briefly minimized by 3 rounds of conjugate-gradient optimization of 40 steps each with, initially (round 1), frozen and later (rounds 2 and 3) harmonically restrained protein coordinates. Keeping this restraint, the water was then further relaxed by 100 2-fs steps of molecular dynamics at 100 K, followed by 40 steps conjugate gradient minimization. Optimizations and simulations were performed using the CHARMM19 force field [53] and an electrostatic cutoff of 12 Å with force shifting [54] inside the X-PLOR package. This structure preparation protocol is automated by the pdb2xplor.py program which is part of the open source Biskit package [55]. The final structure was stripped of its hydrogen atoms and used as input for the Voronoi interface calculations (see below).

To test the legitimacy of this economical solvation procedure, a more costly, state-of-the-art approach was employed on complex 1M0S. Section A.1 of Supplemental Material describes this procedure and compares the Voronoi interfaces obtained using the two equilibration protocols. The very similar results, both in terms of interface topology and the identification of interfacial water, justify the economical solvation method and indicate the robustness of our model against minor changes both in protein conformation and hydration patterns.

## 5.2 Calculation of shelling orders

The program Intervor, responsible for the actual computation and shelling of the Voronoi interface, is based on the CGAL computational geometry library [56]; an online version of Intervor is available [57]. On an Intel Pentium IV 3 GHz CPU, an Intervor run for a typical complex takes less than 5 seconds. We also provide a wrapper (Biskit.Intervor) for integrating the stand-alone program in Biskit workflows. Residue shelling orders were calculated by averaging over a residue's interface atoms.

## 5.3 Dryness, conservation and polarity

Dryness results were those discussed in [21] and were kindly provided to us by O. Lichtarge and coworkers.

Multiple sequence alignments were obtained from the Pfam database [58] of HMMER profiles [59] using the HMMER software version 2.3.1. Protein family profiles matching a given sequence were identified with hmmpfam using a conservative E-value and bit score cutoff of 1e-8 and 60, respectively. The sequence was then aligned to the matching profile with the hmmalign program. Following [51], the conservation of each alignment position was quantified by the Kullback-Leibler divergence (relative entropy) between the HMM emission probabilities $p$ and the background distribution of amino acids in SwissProt $q$:

$$s = \sum_{i=1}^{20} p_i \log \frac{p_i}{q_i}.$$

The complete procedure is automated in the Hmmer.py module of Biskit. Before further analysis, residues outside the interface (average $VSO = 0$) or lacking conservation scores were removed and conservation scores were independently normalized to the maximum of each monomer face.

The following amino acid residues (3-letter code) were considered unpolar or aromatic: Ala, Gly, Ile, Leu, Met, Phe, Trp, Tyr, Val.

## 5.4 ROC curves

A Receiver Operating Characteristics (ROC) curve [40] evaluates the ability of a continuous score to pick a the true positive items out of a set of positives and negatives. It is obtained by plotting sensitivity

versus specificity for all possible values of a threshold. Sensitivity and specificity are defined as

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

and

$$Specificity = 1 - False\ Alert\ Rate = \frac{True\ Negative}{True\ Negative + False\ Positive}.$$

A perfect predictor maximizes sensitivity and specificity for at least one threshold value, for which its ROC curve passes through the point (1,1). Therefore, the closer the ROC plot comes to the upper right corner, the higher the overall accuracy of the classifier [60]. This can be quantified by measuring the area under the ROC curve, which ranges from 0 to 1.

ROC curve and ROC area calculations were performed with the Biskit.ROCalyzer module. The statistical significance of each ROC area was determined from a Mann-Withney U test, as described in [61]. The stats.stats module of SciPy [62] provides an implementation of this test. Its application to ROC curves was implemented as part of the ROCalyzer module in the Biskit.Statistics package. The inverse ChiSquare method (also known as Fisher method) [63, 64] was used to calculate the significance of the overall trend from the individual P-values.

### 5.5 Miscellaneous

The Biskit python package [55] was also used for various other scripting tasks and the collation of results. All parts of Biskit are open source and available at http://biskit.sf.net. Pymol [65], Ipe [66] and CGAL-Ipelets [67] were employed for the rendering of figures.

**Acknowledgments.** *We would like to express our gratitude to Olivier Lichtarge and Tuan Anh Tran for providing us with their detailed dryness results. The automatic generation of conservation profiles was implemented by Johan Leckner. B. Bouvier acknowledges funding from the INRIA cooperative project ReflexP. R. Grünberg is supported by the Human Frontiers Science Program.*

## References

[1] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery., Nature 440 (7084) (2006) 631–636.
URL `http://dx.doi.org/10.1038/nature04532`

[2] J. J. Gray, High-resolution protein-protein docking., Curr Opin Struct Biol 16 (2) (2006) 183–193.
URL `http://dx.doi.org/10.1016/j.sbi.2006.03.003`

[3] R. Grünberg, J. Leckner, M. Nilges, Complementarity of structure ensembles in protein-protein binding., Structure 12 (12) (2004) 2125–2136.
URL `http://dx.doi.org/10.1016/j.str.2004.09.014`

[4] R. Grünberg, M. Nilges, J. Leckner, Flexibility and conformational entropy in protein-protein binding., Structure 14 (4) (2006) 683–693.
URL `http://dx.doi.org/10.1016/j.str.2006.01.014`

[5] C. Chotia, J. Janin, Principles of protein-protein recognition, Nature 256 (1975) 705–708.

[6] J.-L. K. Kouadio, J. R. Horn, G. Pal, A. A. Kossiakoff, Shotgun alanine scanning shows that growth hormone can bind productively to its receptor through a drastically minimized interface., J Biol Chem 280 (27) (2005) 25524–25532.
URL `http://dx.doi.org/10.1074/jbc.M502167200`

[7] S. Jones, J. Thornton, Analysis of protein-protein interaction sites using surface patches, J. Mol. Biol. 272.

[8] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites, Journal of Molecular Biology 285 (1999) 2177–2198.
URL http://www.sciencedirect.com/science/article/B6WK7-45R8%84M-RY/2/6f4a9866e2495a34273695de046893dc

[9] S. A. Teichmann, Principles of protein-protein interactions, Bioinformatics 18 (suppl. 2) (2002) S249–.
URL http://bioinformatics.oxfordjournals.org/cgi/content/ab%stract/18/suppl_2/S249

[10] B. Ma, T. Elkayam, H. Wolfson, R. Nussinov, Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces, Proceedings of the National Academy of Sciences 100 (10) (2003) 5772–5777.
URL http://www.pnas.org/cgi/content/abstract/100/10/5772

[11] Y. Ofran, B. Rost, Protein-protein interaction hotspots carved into sequences, PLoS Computational Biology 3 (2007) e119.
URL http://dx.doi.org/10.1371%2Fjournal.pcbi.0030119

[12] I. S. Moreira, P. A. Fernandes, M. J. Ramos, Hot spots – a review of the protein-protein interface determinant amino-acid residues, Proteins: Structure, Function, and Bioinformatics 68 (4).
URL http://dx.doi.org/10.1002/prot.21396

[13] T. Clackson, J. Wells, A hot spot of binding energy in a hormone-receptor interface, Science 267 (5196) (1995) 383–386.
URL http://www.sciencemag.org/cgi/content/abstract/267/5196%/383

[14] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, G. Schreiber, The modular architecture of protein-protein binding interfaces, PNAS 102 (1) (2005) 57–62.
URL http://www.pnas.org/cgi/content/abstract/102/1/57

[15] D. Reichmann, O. Rahat, M. Cohen, H. Neuvirth, G. Schreiber, The molecular architecture of protein-protein binding sites, Current Opinion in Structural Biology 17 (2007) 67–76.
URL http://www.sciencedirect.com/science/article/B6VS6-4MVD%VF3-3/2/60d484544e9900423f219f319b0936b5

[16] M. Guharoy, P. Chakrabarti, Conservation and relative importance of residues across protein-protein interfaces., Proc Natl Acad Sci U S A 102 (43) (2005) 15447–15452.
URL http://dx.doi.org/10.1073/pnas.0505425102

[17] B. E. Shakhnovich, N. V. Dokholyan, C. DeLisi, E. I. Shakhnovich, Functional fingerprints of folds: Evidence for correlated structure-function evolution, Journal of Molecular Biology 326 (1) (2003) 1–9.
URL http://www.sciencedirect.com/science/article/B6WK7-47RC%32Y-3/2/780a30e8b2d84bdefba0ea80b0a6b6b2

[18] A. Valencia, Automatic annotation of protein function, Current Opinion in Structural Biology 15 (3) (2005) 267–274.
URL http://www.sciencedirect.com/science/article/B6VS6-4G7X%9HS-3/2/90d887674957c53860854c3254a94748

[19] P. Aloy, H. Ceulemans, A. Stark, R. B. Russell, The relationship between sequence and interaction divergence in proteins, Journal of Molecular Biology 332 (5) (2003) 989–998.
URL http://www.sciencedirect.com/science/article/B6WK7-49HM%CGT-4/2/6d13750d3d40cc10b07ef096ef54961b

[20] R. P. Bahadur, P. Chakrabarti, F. Rodier, J. Janin, A dissection of specific and non-specific protein-protein interfaces, Journal of Molecular Biology 336 (4) (2004) 943–955.
URL http://www.sciencedirect.com/science/article/B6WK7-4BRT%KR9-C/2/ce716dee1132d2605f8e1c96d1154253

[21] I. Mihalek, I. Res, O. Lichtarge, On itinerant water molecules and detectability of protein-protein interfaces through comparative analysis of homologues, Journal of Molecular Biology 369 (2) (2007) 584–595.

[22] K. K. Frederick, M. S. Marlow, K. G. Valentine, A. J. Wand, Conformational entropy in molecular recognition by proteins., Nature 448 (7151) (2007) 325–329.
URL http://dx.doi.org/10.1038/nature05959

[23] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, F. G. Parak, Slaving: solvent fluctuations dominate protein dynamics and functions., Proc Natl Acad Sci U S A 99 (25) (2002) 16047–16051.
URL http://dx.doi.org/10.1073/pnas.212637899

[24] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, R. D. Young, Bulk-solvent and hydration-shell fluctuations, similar to alpha- and beta-fluctuations in glasses, control protein motions and functions., Proc Natl Acad Sci U S A 101 (40) (2004) 14408–14413.
URL http://dx.doi.org/10.1073/pnas.0405573101

[25] A. Fernandez, R. S. Berry, Extent of Hydrogen-Bond Protection in Folded Proteins: A Constraint on Packing Architectures, Biophys. J. 83 (5) (2002) 2475–2481.
URL http://www.biophysj.org/cgi/content/abstract/83/5/2475

[26] A. Fernandez, H. A. Scheraga, Insufficiently dehydrated hydrogen bonds as determinants of protein interactions, Proceedings of the National Academy of Sciences 100 (1) (2003) 113–118.
URL http://www.pnas.org/cgi/content/abstract/100/1/113

[27] A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces, Journal of Molecular Biology 280 (1998) 1–9.
URL http://www.sciencedirect.com/science/article/B6WK7-45S4%9GB-9C/2/b3d9c6f299c1eec3933d2774dffaf67d

[28] F. Cazals, F. Proust, R. P. Bahadur, J. Janin, Revisiting the Voronoi description of protein-protein interfaces, Protein Sci 15 (9) (2006) 2082–2092.
URL http://www.proteinscience.org/cgi/content/abstract/15/9%/2082

[29] F. M. Richards, Areas, volumes, packing and protein structure, Ann. Rev. Biophys. Bioeng. 6 (1977) 151–176.

[30] M. Gerstein, F. Richards, Protein geometry: volumes, areas, and distances, The international tables for crystallography (Vol F, Chap. 22) F (Chapter 22.1.1) (2001) 531–539.

[31] A. Poupon, Voronoi and voronoi-related tessellations in studies of protein structure and interaction., Curr Opin Struct Biol 14 (2) (2004) 233–241.
URL http://dx.doi.org/10.1016/j.sbi.2004.03.010

[32] H. Edelsbrunner, M. Facello, J. Liang, On the definition and the construction of pockets in macromolecules, Discrete Appl. Math. 88 (1998) 83–102.

[33] R. Singh, A. Tropsha, I. Vaisman, Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues, J. Comput. Biol. 3 (1996) 213–222.

[34] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, S. Subramaniam, Analytical shape computing of macromolecules ii: identification and computation of inaccessible cavities inside proteins, Proteins 33 (1998) 18–29.
URL http://cast.engr.uic.edu/cast

[35] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, S. Subramaniam, Analytical shape computing of macromolecules i: molecular area and volume through alpha shape, Proteins 33 (1998) 1–17.
URL http://cast.engr.uic.edu/cast

[36] F. Aurenhammer, Power diagrams: properties, algorithms and applications, SIAM J. Comput. 16 (1987) 78–96.

[37] H.-M. Will, Fast and efficient computation of additively weighted voronoi cells for applications in molecular biology, in: SWAT, 1998, pp. 310–32.

[38] J. Bernauer, J. Aze, J. Janin, A. Poupon, A new protein-protein docking scoring function based on interface residue properties., Bioinformatics (2007) 652–658.
URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi%?cmd=prlinks\&dbfrom=pubmed\&retmode=ref\&id=17237048

[39] Y.-E. A. Ban, , H. Edelsbrunner, J. Rudolph, Interface surfaces for protein-protein complexes, in: RECOMB, San Diego, 2004, pp. 205–212.

[40] D. M. Green, J. M. Swets, Signal detection theory and psychophysics, John Wiley and Sons Inc., New York, 1966.

[41] S. Mason, N. Graham, Areas beneath the relative operating characteristics (roc), and levels (rol) curves: statistical significance and interpretation, Quarterly Journal of the Royal Meteorological Society 128 (2002) 2145–2166.

[42] A. Varshney, F. P. Brooks, D. C. Richardson, W. V. Wright, D. Manocha, Defining, computing, and visualizing molecular interfaces, in: IEEE Visualization, Atlanta, USA, 1995, pp. 36–43.

[43] R. R. Gabdoulline, R. C. Wade, D. Walther, Molsurfer: A macromolecular interface navigator., Nucleic Acids Res 31 (13) (2003) 3349–3351.

[44] S. Chakravarty, R. Varadarajan, Residue depth: a novel parameter for the analysis of protein structure and stability., Structure 7 (7) (1999) 723–732.

[45] A. Pintar, O. Carugo, S. Pongor, Atom depth in protein structure and function., Trends Biochem Sci 28 (11) (2003) 593–597.

[46] A. Pintar, O. Carugo, S. Pongor, Atom depth as a descriptor of the protein interior., Biophys J 84 (4) (2003) 2553–2561.

[47] P. Chakrabarti, J. Janin, Dissecting protein-protein recognition sites, Proteins 47.

[48] I. M. A. Nooren, J. M. Thornton, Structural characterisation and functional significance of transient protein-protein interactions., J Mol Biol 325 (5) (2003) 991–1018.

[49] A. Bordner, R. Abagyan, Statistical analysis and prediction of protein-protein interfaces, Proteins 60 (3) (2005) 353–66.

[50] M. Rosso, J. F. Gouyet, B. Sapoval, Gradient percolation in three dimensions and relation to diffusion fronts, Phys. Rev. Lett. 57 (25) (1986) 3195–3198.

[51] K. Wang, R. Samudrala, Incorporating background frequency improves entropy-based residue conservation measures., BMC Bioinformatics 7 (2006) 385.
URL http://dx.doi.org/10.1186/1471-2105-7-385

[52] J. G. S. Sheldon Park, Statistical and molecular dynamics studies of buried waters in globular proteins, Proteins: Structure, Function, and Bioinformatics 60 (2005) 450–463.
URL http://dx.doi.org/10.1002/prot.20511

[53] B. Brooks, R. Bruccoleri, Olafson B.D., D. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization and dynamics calculations., J Comp Chem 4 (1983) 187–217.

[54] P. Steinbach, R. Loncharich, B. Brooks, The effects of environment and hydration on protein dynamics: A simulation study of myoglobin., Chem Phys 158 (1991) 383–94.

[55] R. Grünberg, M. Nilges, J. Leckner, Biskit–A software platform for structural bioinformatics, Bioinformatics 23 (6) (2007) 769–770.
URL http://bioinformatics.oxfordjournals.org/cgi/content/ab%stract/23/6/769

[56] CGAL, Computational Geometry Algorithms Library, http://www.cgal.org.

[57] http://cgal.inria.fr/Intervor.

[58] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucl. Acids Res. 34 (suppl. 1) (2006) D247–251.
URL http://nar.oxfordjournals.org/cgi/content/abstract/34/s%uppl_1/D247

[59] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998, Ch. The theory behind profile HMMs.

[60] M. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine [published erratum appears in Clin Chem 1993 Aug;39(8):1589], Clin Chem 39 (4) (1993) 561–577.
URL http://www.clinchem.org/cgi/content/abstract/39/4/561

[61] S. Mason, N. E. Graham, Areas beneath the relative operating characteristics (roc), and levels (rol) curves: statistical significance and interpretation, Quarterly Journal of the Royal Meteorological Society 128 (2002) 2145–2166.

[62] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python (2001–).
URL http://www.scipy.org/

[63] R. A. Fisher, Combining independent tests of significance., American Statistician 2 (5) (1948) 30.

[64] W. W. Piegorsch, A. J. Bailer, Analyzing Environmental Data, Wiley, 2005.

[65] W. DeLano, The Pymol molecular graphics system, http://www.pymol.org (2002).

[66] O. Cheong, The Ipe extensible drawing editor, http://tclab.kaist.ac.kr/ipe/ (1993-2007).

[67] http://cgal-ipelets.gforge.inria.fr/.

[68] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, Gromacs: fast, flexible, and free., J Comput Chem 26 (16) (2005) 1701–1718.
URL http://dx.doi.org/10.1002/jcc.20291

[69] W. Damm, A. Frontera, J. Tirado-Rives, W. L. Jorgensen, Opls all-atom force field for carbohydrates, Journal of Computational Chemistry 18 (1997) 1955–1970.
URL http://dx.doi.org/10.1002/(SICI)1096-987X(199712)18:16<%1955::AID-JCC1>3.0.CO;2-L

# 6 FIGURES LEGENDS

**Legend of Fig. 1.** (a) Shelling of the Voronoi interface of a dimer complex, seen from the top. Solid dots represent protein atoms' centers, hollow squares water atoms' centers; for clarity, all atomic radii have been taken equal and the corresponding spheres omitted. The Voronoi facets composing the protein-protein interface are colored according to their shelling order: one (light gray, at the rim), two (middle gray), three (dark gray). (b) Two-dimensional illustration of the Voronoi interface shelling of a dimer complex. Red and blue circles represent the atoms of each partner, the green circle a water molecule. Interface Delaunay edges, which connect atoms on different partners, are shown as dashed black ($AB$ interface) or green ($AW-BW$ interface) lines; the Voronoi facets are shown as solid lines. Black numerals denote the shelling order of each Delaunay edge/Voronoi facet, from which the atomic shelling orders (red, blue and green numerals) can be derived (refer to text for details). On this simple illustration, the high curvature of the $AW-BW$ interface due to the water molecule accounts for the high Voronoi shelling order of the blue atoms.
**Legend of Fig. 2.** (a) Voronoi interface of the 2DOR homodimer complex, superimposed on the solvent accessible surface representation of one of the monomers (gray); for clarity, the second monomer is not shown. The Voronoi shelling order varies from 1 (blue) to 6 (red). (b) Solvent accessible surface of one monomer of the 2DOR complex, showing the Voronoi shelling order of interface atoms (color-coded as in panel b).
**Legend of Fig. 3.** Properties of the 2DOR homodimer interface. Conservation, exposure to bulk water, and residue polarity are color-coded onto the solvent accessible surface of one monomer. The surface not involved in the interface is colored grey and, for clarity, the second monomer is not shown.
**Legend of Fig. 4.** Performance of Voronoi shelling order (circles, solid line) and conservation (squares, dashed line) as predictors of dryness, for all studied heterodimer (left panel) and homodimer (right panel) complexes. Scores are measured as the area underneath the corresponding ROC curve; complexes are sorted by decreasing Voronoi shelling order score. Values lower than 0.5 (hatched area) denote a performance that is no better than that of a purely random classifier.
**Legend of Fig. 5.** Spatial distribution of conservation across heterodimer interfaces. The normalized conservation score for each interface residue is plotted against its normalized Voronoi shelling order. VSO ranges from 0 (interface edge) to 1 (interface center); Conservation ranges from 0 (lowest conservation) to 1 (highest conservation). x: all data points; –: running average with a window covering 1/4 of interface residues. The gray area outlines the expected variation of the running average when the same conservation values are randomly distributed along the VSO axis ($\pm\ \sigma$ from 500-fold shuffling of conservation versus VSO values).
**Legend of Fig. 6.** Spatial distribution of conservation across homodimer interfaces. See figure 5 and text for a detailed description.
**Legend of Fig. 7.** Projection of Voronoi shelling order (large panels), dryness (lower left-hand panel) and conservation (lower right-hand panel) on the molecular surface of homocomplexes 1E2D (left), 1L5W (center) and 1A59 (right); one of the monomers was removed for clarity. Cold (resp. hot) colors represent low (resp. high) values; gray areas denote residues for which conservation information was unavailable.
**Legend of Fig. 8.** View of the cleft region of the 1L5W interface, showing the two protein partners as solid and mesh surfaces, respectively. Colors code for Voronoi shelling order, which is low inside the cleft due to the presence of numerous water molecules which fragment the interface.
**Legend of Fig. 9.** Boundary of the $AB$ interface of complex 1A59 (red line), interfacial water (gray spheres), and $AW-BW$ interface (grey and green Voronoi polygons). The holes pointed out by arrows prevent the Voronoi shelling order from peaking in the middle of the interface patch –compare to the bottom left panel of complex 1A59 on Fig. 7.

# 7 TABLES

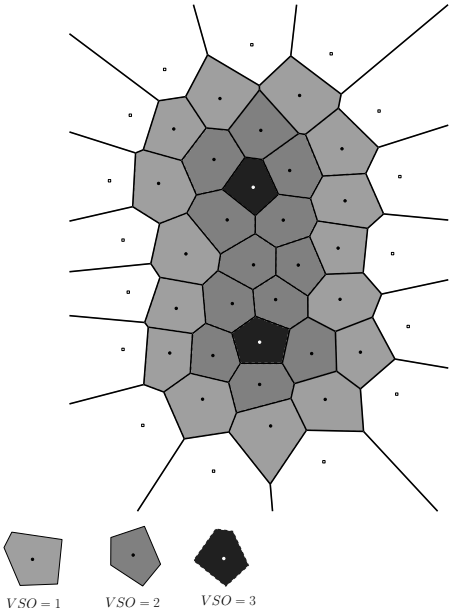
| PDB Id. | VSO→dryness | | conserv.→dryness | | conserv.→VSO | | VSO→unpolar | |
|---|---|---|---|---|---|---|---|---|
| | AUC | P-value | AUC | P-value | AUC | P-value | AUC | P-value |
| 1HE1 | 0.92 | 4e-10 | 0.78 | 1e-04 | 0.52 | 4e-01 | 0.54 | 1e-01 |
| 1CXZ | 0.89 | 5e-06 | 0.74 | 2e-03 | 0.69 | 7e-02 | 0.60 | 1e-01 |
| 1CEE | 0.89 | 5e-06 | 0.62 | 8e-02 | 0.61 | 2e-01 | 0.74 | 2e-05 |
| 1C1Y | 0.86 | 3e-04 | 0.67 | 7e-02 | 0.55 | 4e-01 | 0.55 | 4e-01 |
| 1RRP | 0.84 | 2e-08 | 0.72 | 3e-04 | 0.71 | 3e-03 | 0.65 | 4e-04 |
| 1FIN | 0.84 | 1e-09 | 0.60 | 5e-02 | 0.68 | 2e-02 | 0.71 | 5e-05 |
| 1E96 | 0.84 | 9e-04 | 0.48 | 9e-01 | 0.65 | 1e-01 | 0.54 | 4e-01 |
| 1ZBD | 0.83 | 2e-06 | 0.59 | 1e-01 | 0.69 | 3e-02 | 0.75 | 4e-05 |
| 1FOE | 0.83 | 1e-07 | 0.69 | 2e-03 | 0.77 | 2e-03 | 0.66 | 2e-03 |
| 1A0O | 0.82 | 4e-03 | 0.73 | 4e-02 | 0.62 | 2e-01 | 0.67 | 2e-02 |
| 2TRC | 0.82 | 6e-10 | 0.42 | 6e-01 | 0.61 | 8e-02 | 0.67 | 8e-05 |
| 1GOT | 0.82 | 2e-06 | 0.63 | 4e-02 | 0.73 | 7e-03 | 0.68 | 2e-03 |
| 1WQ1 | 0.81 | 2e-09 | 0.69 | 9e-05 | 0.58 | 2e-01 | 0.62 | 2e-02 |
| 1IBR | 0.80 | 7e-09 | 0.51 | 5e-01 | 0.36 | 5e-01 | 0.66 | 1e-03 |
| 1A2K | 0.76 | 4e-03 | 0.65 | 6e-02 | 0.78 | 7e-03 | 0.64 | 7e-03 |
| 1LFD | 0.75 | 4e-03 | 0.76 | 7e-03 | 0.65 | 2e-01 | 0.55 | 2e-01 |
| 1AGR | 0.69 | 3e-03 | 0.60 | 8e-02 | 0.75 | 9e-03 | 0.60 | 6e-02 |
| 1YCS | 0.66 | 4e-02 | 0.66 | 6e-02 | 0.79 | 1e-02 | 0.54 | 3e-01 |
| Reject $H_0$ | | 18/18 | | 8/18 | | 8/18 | | 11/18 |
| Global | 0.81 | 6e-74 | 0.64 | 3e-14 | 0.65 | 2e-09 | 0.63 | 1e-21 |

Table 1: Heterodimers. Performance for the prediction of dryness from Voronoi Shelling Order (VSO→dryness); of dryness from conservation (conserv.→dryness); of VSO from conservation (conserv.→VSO); and of unpolar+aromatic residues from VSO (VSO→unpolar) for each of the heterodimer complexes. The predictive power in each direction is quantified in terms of Area Under the ROC curve (AUC) between 0 and 1, and the associated P-value (see text for details). The last two rows respectively feature (i) the number of predictive cases with respect to a random classifier (null hypothesis rejected at a threshold of $P = 0.05$), and (ii) Averages AUC and and combined P-values.

| PDB Id. | VSO→dryness | | conserv.→dryness | | conserv.→VSO | | VSO→unpolar | |
|---|---|---|---|---|---|---|---|---|
| | AUC | P-value | AUC | P-value | AUC | P-value | AUC | P-value |
| 2BIF | 0.95 | 4e-11 | 0.59 | 1e-01 | 0.52 | 4e-01 | 0.73 | 2e-04 |
| 1E5Q | 0.95 | 2e-07 | 0.65 | 4e-02 | 0.81 | 1e-03 | 0.71 | 6e-04 |
| 1E2D | 0.95 | 8e-08 | 0.87 | 1e-05 | 0.88 | 5e-04 | 0.74 | 5e-04 |
| 1H7T | 0.95 | 9e-09 | 0.62 | 8e-02 | 0.67 | 5e-02 | 0.69 | 1e-03 |
| 1TB5 | 0.93 | 9e-06 | 0.64 | 1e-01 | 0.52 | 5e-01 | 0.77 | 1e-04 |
| 2DOR | 0.92 | 3e-15 | 0.69 | 3e-04 | 0.63 | 6e-02 | 0.64 | 2e-03 |
| 1QIN | 0.92 | <1e-16 | 0.64 | 7e-03 | 0.64 | 2e-02 | 0.70 | 1e-07 |
| 1E98 | 0.92 | 8e-07 | 0.90 | 5e-06 | 0.95 | 3e-04 | 0.83 | 2e-05 |
| 1J79 | 0.90 | 7e-06 | 0.41 | 6e-01 | 0.42 | 7e-01 | 0.61 | 4e-02 |
| 1NYW | 0.90 | 8e-10 | 0.41 | 6e-01 | 0.54 | 3e-01 | 0.54 | 3e-01 |
| 1BTO | 0.88 | 1e-09 | 0.77 | 2e-05 | 0.62 | 8e-02 | 0.66 | 5e-04 |
| 1Y6R | 0.88 | 4e-09 | 0.67 | 6e-03 | 0.53 | 4e-01 | 0.57 | 8e-02 |
| 1KER | 0.87 | 2e-08 | 0.64 | 2e-02 | 0.58 | 2e-01 | 0.61 | 6e-02 |
| 1EK4 | 0.87 | <1e-16 | 0.65 | 4e-04 | 0.71 | 3e-03 | 0.49 | 9e-01 |
| 1LBX | 0.87 | 6e-06 | 0.71 | 1e-02 | 0.61 | 1e-01 | 0.66 | 1e-03 |
| 1L9W | 0.86 | 7e-06 | 0.79 | 3e-04 | 0.77 | 1e-02 | 0.72 | 1e-03 |
| 1AI2 | 0.86 | <1e-16 | 0.68 | 5e-06 | 0.45 | 7e-01 | 0.62 | 4e-04 |
| 1W1U | 0.85 | <1e-16 | 0.57 | 4e-02 | 0.47 | 8e-01 | 0.63 | 3e-04 |
| 1DQX | 0.83 | 1e-09 | 0.60 | 4e-02 | 0.41 | 6e-01 | 0.61 | 2e-02 |
| 1E7Y | 0.82 | 2e-10 | 0.74 | 2e-06 | 0.44 | 7e-01 | 0.64 | 1e-03 |
| 1HKV | 0.82 | 2e-15 | 0.59 | 1e-02 | 0.54 | 2e-01 | 0.52 | 3e-01 |
| 1M0S | 0.82 | 3e-06 | 0.57 | 1e-01 | 0.84 | 3e-04 | 0.72 | 5e-05 |
| 1KC3 | 0.82 | 8e-06 | 0.85 | 2e-06 | 0.82 | 3e-04 | 0.49 | 9e-01 |
| 1M4N | 0.81 | 9e-09 | 0.67 | 1e-03 | 0.64 | 2e-02 | 0.67 | 2e-05 |
| 1A59 | 0.81 | 1e-14 | 0.65 | 1e-04 | 0.69 | 8e-04 | 0.68 | 5e-07 |
| 1DQR | 0.81 | <1e-16 | 0.59 | 3e-03 | 0.58 | 4e-02 | 0.64 | 1e-07 |
| 1AN9 | 0.80 | 1e-06 | 0.61 | 4e-02 | 0.56 | 3e-01 | 0.67 | 2e-03 |
| 1M7P | 0.79 | 3e-06 | 0.51 | 4e-01 | 0.58 | 2e-01 | 0.51 | 4e-01 |
| 1TC2 | 0.79 | 9e-07 | 0.49 | 9e-01 | 0.67 | 7e-02 | 0.64 | 3e-03 |
| 1AD3 | 0.78 | 3e-14 | 0.47 | 7e-01 | 0.66 | 4e-03 | 0.68 | 1e-07 |
| 1ALN | 0.77 | 1e-07 | 0.64 | 5e-03 | 0.54 | 3e-01 | 0.52 | 3e-01 |
| 1H16 | 0.77 | 8e-07 | 0.44 | 6e-01 | 0.48 | 9e-01 | 0.65 | 8e-04 |
| 1M9N | 0.76 | 1e-14 | 0.59 | 6e-03 | 0.70 | 2e-05 | 0.62 | 2e-05 |
| 1L5W | 0.74 | 7e-05 | 0.68 | 4e-03 | 0.75 | 3e-04 | 0.60 | 5e-03 |
| 1CG0 | 0.72 | 7e-08 | 0.62 | 2e-03 | 0.55 | 2e-01 | 0.60 | 4e-03 |
| 1LXY | 0.71 | 1e-03 | 0.60 | 7e-02 | 0.61 | 1e-01 | 0.61 | 5e-02 |
| Reject $H_0$ | | 36/36 | | 25/36 | | 14/36 | | 27/36 |
| Global | 0.84 | 2e-265 | 0.63 | 2e-43 | 0.62 | 4e-20 | 0.64 | 2e-63 |

Table 2: Homodimers. Performance for the prediction of dryness from Voronoi Shelling Order (VSO→dryness); of dryness from conservation (conserv.→dryness); of VSO from conservation (conserv.→VSO); and of unpolar+aromatic residues from VSO (VSO→unpolar) for each of the homodimer complexes. See also description of table 1.

# 8 FIGURES



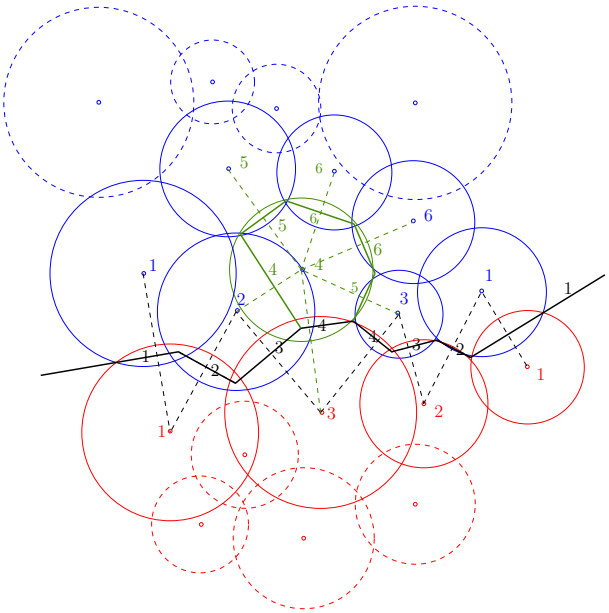



Figure 1: (a) and(b)

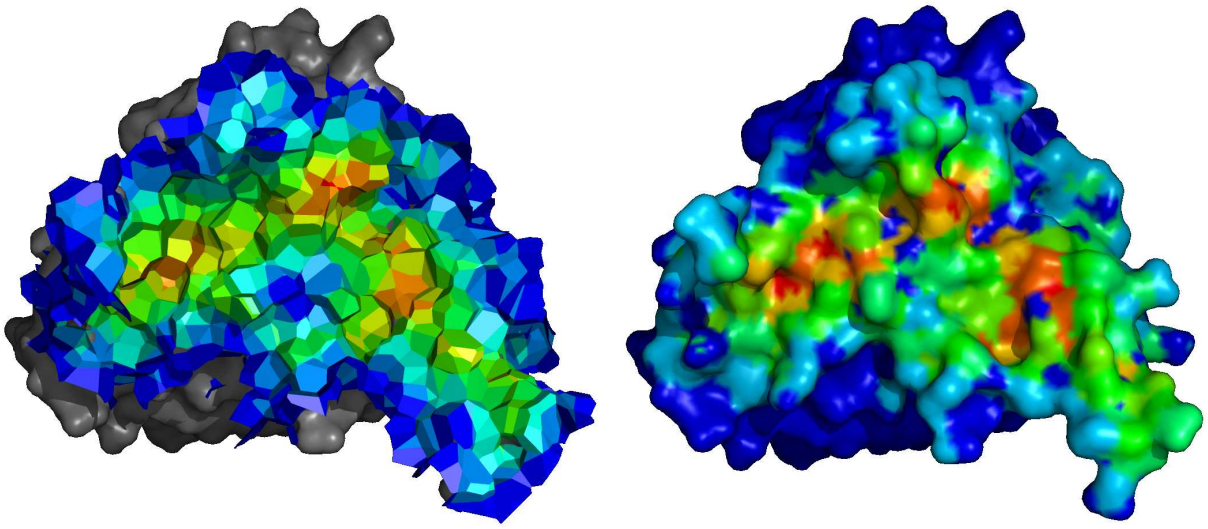


Figure 2: (a) and(b)

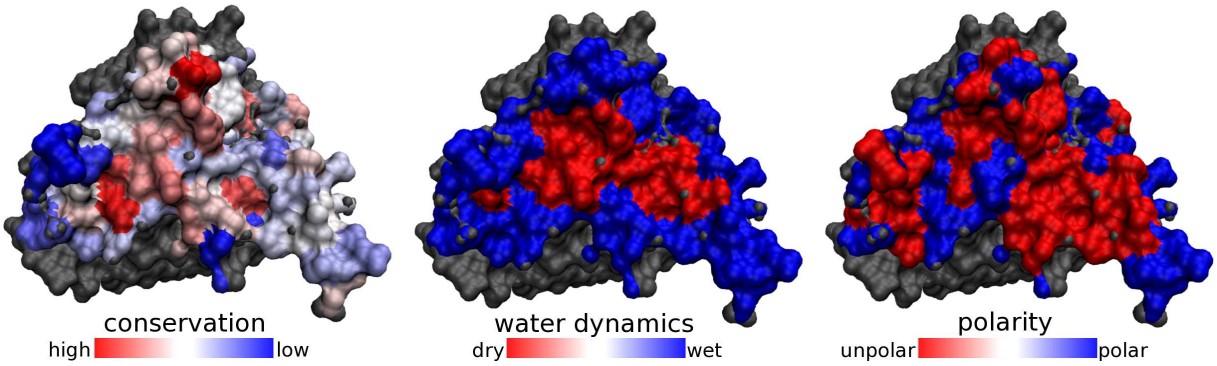



Figure 3:

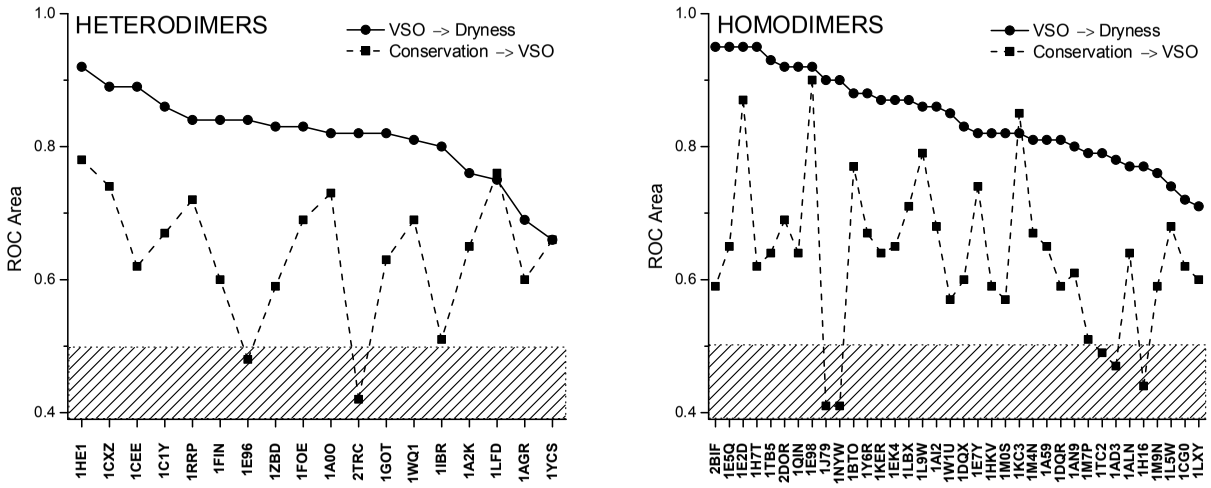

HETERODIMERS
VSO -> Dryness
Conservation -> VSO
ROC Area
1.0
0.8
0.6
0.4
1HE1
1CXZ
1CEE
1C1Y
1RRP
1FIN
1E96
1ZBD
1FOE
1A0O
2TRC
1GOT
1WQ1
1IBR
1A2K
1LFD
1AGR
1YCS
HOMODIMERS
VSO -> Dryness
Conservation -> VSO
ROC Area
1.0
0.8
0.6
0.4
2BIF
1E5Q
1E2D
1H7T
1TB5
2DOR
1QIN
1E98
1J79
1NYW
1BTO
1Y6R
1KER
1EK4
1LBX
1L9W
1AI2
1W1U
1DQX
1E7Y
1HKV
1M0S
1KC3
1M4N
1A9J
1DQR
1AM9
1M7P
1TC2
1AO3
1ALN
1H16
1M9N
1L5W
1CG0
1LXY

Figure 4:

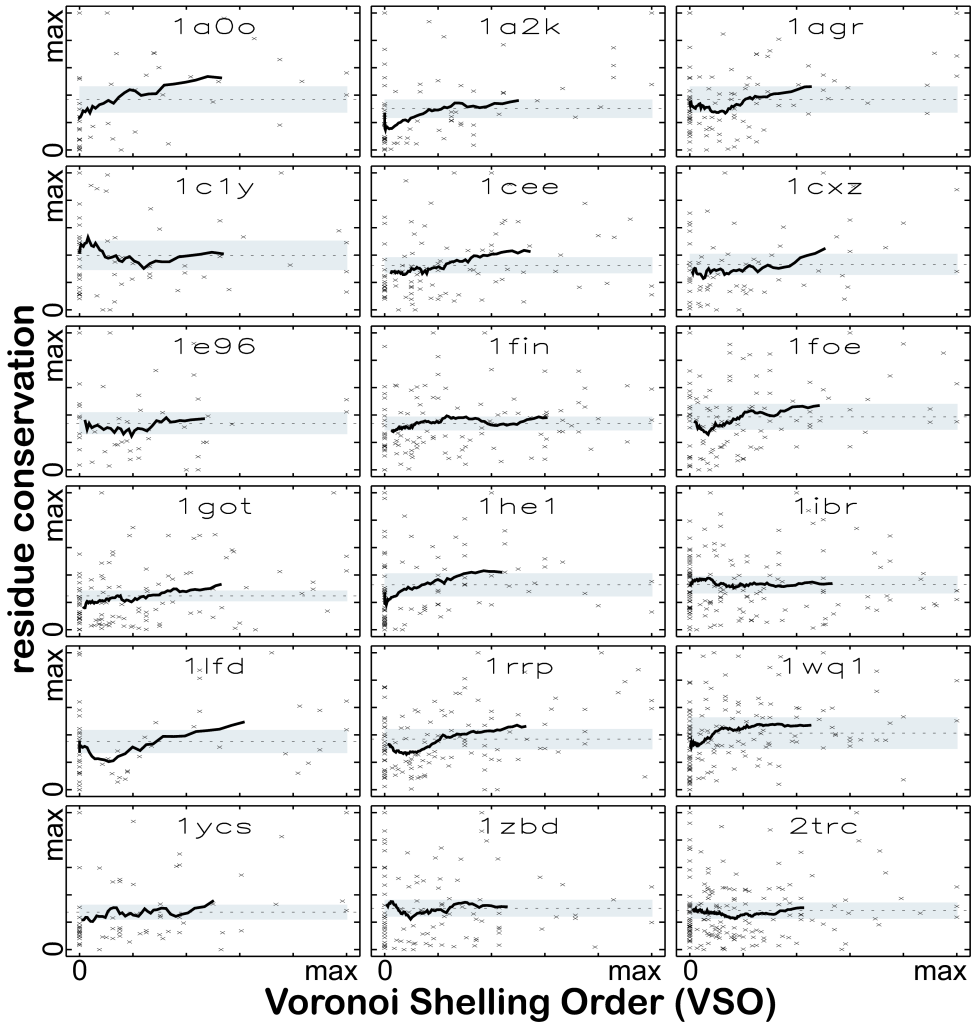
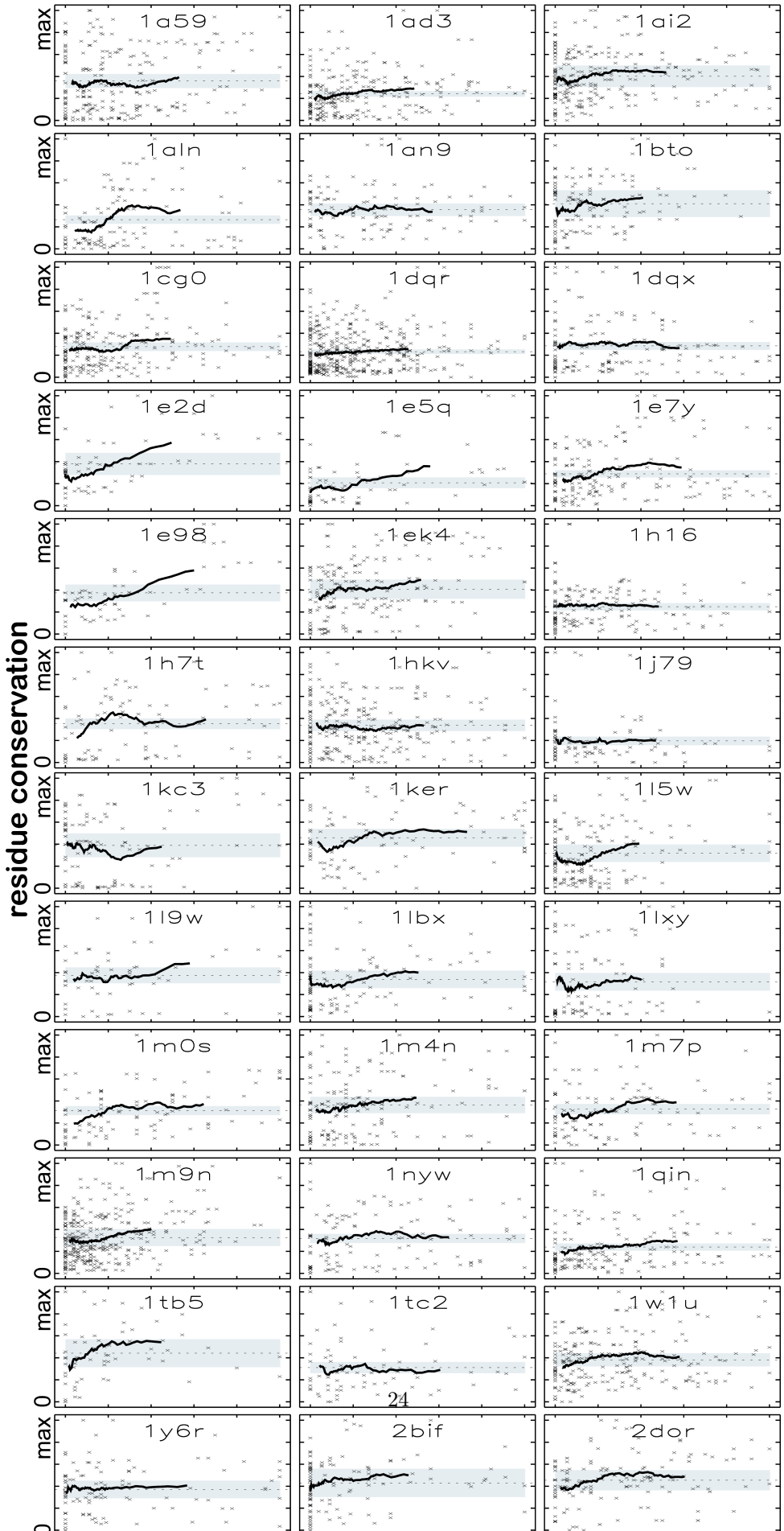

1a0o
1a2k
1agr
1c1y
1cee
1cxz
1e96
1fin
1foe
1got
1he1
1ibr
1lfd
1rrp
1wq1
1ycs
1zbd
2trc
max
0
residue conservation
Voronoi Shelling Order (VSO)

Figure 5:


residue conservation
1a59
1ad3
1ai2
1aln
1an9
1bto
1cg0
1dqr
1dqx
1e2d
1e5q
1e7y
1e98
1ek4
1h16
1h7t
1hkv
1j79
1kc3
1ker
1l5w
1l9w
1lbx
1lxy
1m0s
1m4n
1m7p
1m9n
1nyw
1qin
1tb5
1tc2
1w1u
1y6r
2bif
2dor
max
0

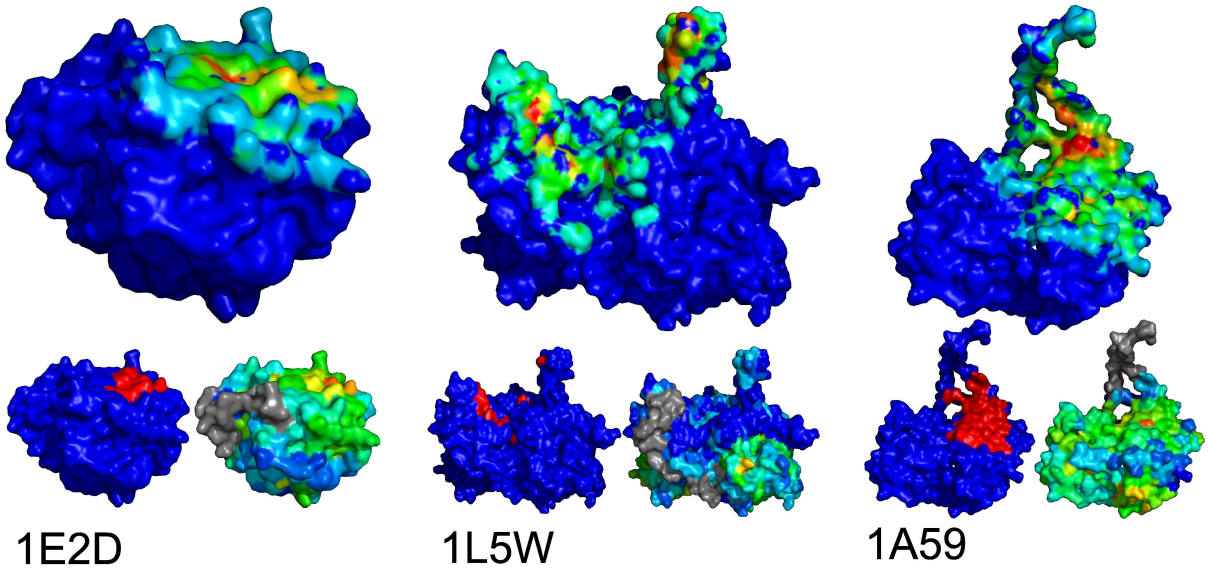



Figure 7:

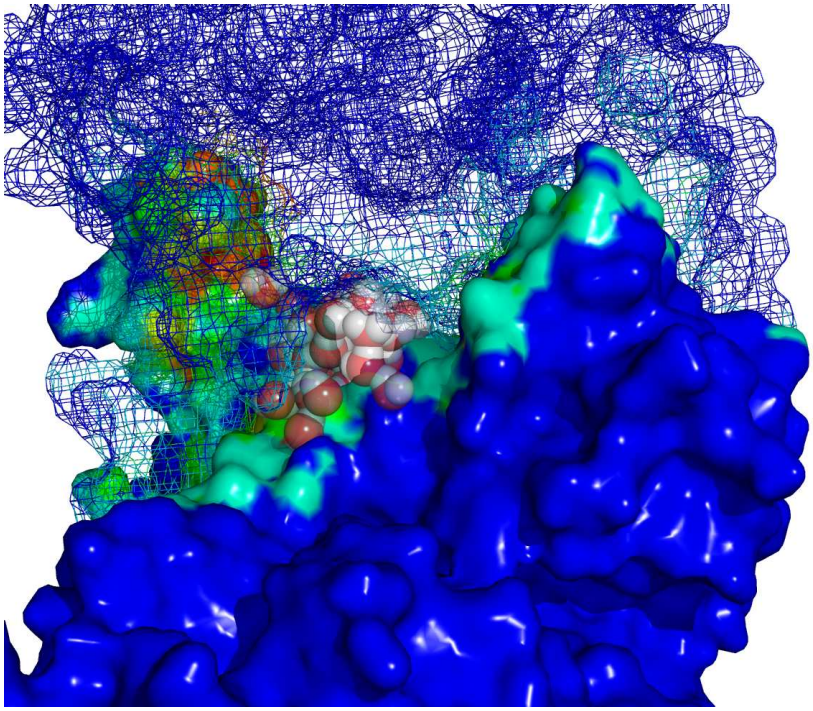


Figure 8:

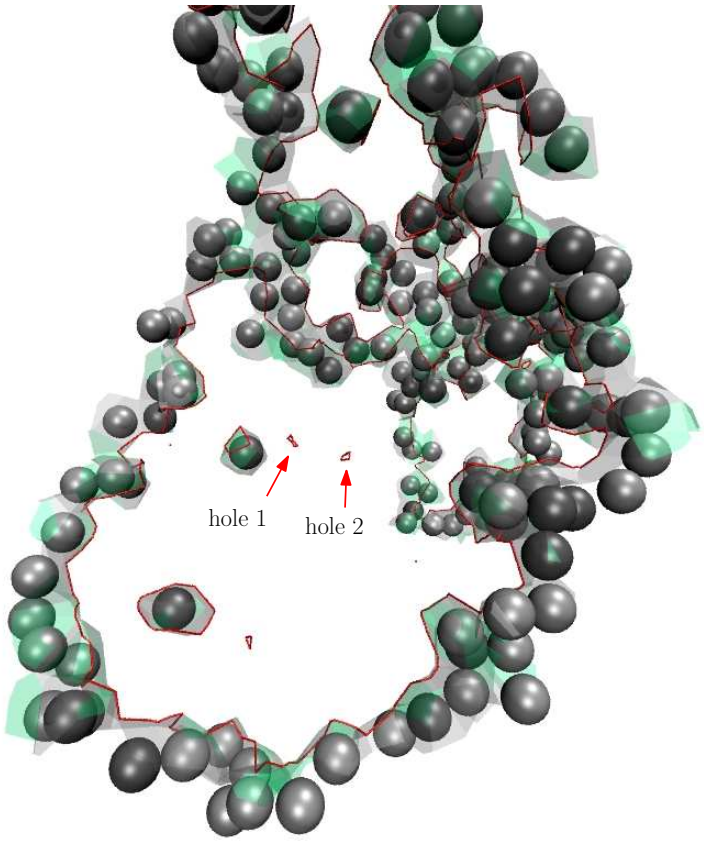



Figure 9:

# A Supplemental Material

## A.1 Validation of the sample preparation procedure

The procedure employed for the rehydration and equilibration of each of the complexes (Section 5) has deliberately been kept short, and can be run in minutes on a desktop computer. In this paragraph, we ascertain whether the placement and equilibration of the water molecules added using this fast protocol are of sufficient quality for the current application. Of particular interest are the interfacial water molecules. When in simultaneous contact with both protein partners, they form the $AW - BW$ interface (Figure 1b and 9); but several layers of water inside a larger pocket will create holes in the interface, possibly splitting it into several connected components. The implications for shelling orders are crucial: in the first case, the water molecules will not affect the SO, while in the second scenario a boundary is created and the SO consequently reset to 1.

The complex 1M0S, which features a large pocket filled with crystal water molecules, was used for the test. A rigorous equilibration procedure, retaining the crystal water molecules and involving a 5 ns molecular dynamics simulation with state-of-the-art algorithms and parameters (Section A.2), provided us with a reference structure. Both this structure and the one from the fast procedure were used as input to Intervor. Figure 10 shows the tessellation of the AB interface and the interfacial water molecules for both cases. Due to minor conformational transitions that have occurred during the 5 ns MD simulation, the two interfaces are not superposable. However, they retain the same shape and number of connected components. In both cases, the central cavity is filled with interfacial water that participates to the ABW interface. Both interfaces feature boundaries of comparable lengths and topologies.

This difficult test case provides justification for our sample preparation methodology. It also represents a tribute to the robustness of our model, which delivers stable results upon variation of the solvation of the complex within a reasonable range.

## A.2 Details of the thorough sample preparation procedure

After an initial re-optimization of the crystal structure (retaining crystal water), the complex was placed inside a triclinic box, solvated with SPC water molecules from an equilibrated box and neutralized by 8 $Na^+$ ions. The solvent molecules were then relaxed around the fixed solute by a steepest-descent optimization followed by 100 ps of molecular dynamics (MD) simulation with position restraints on the solute. The entire system was then simulated for 5 ns without restraints, with a 300 K Maxwellian distribution of initial velocities. MD simulations employed the particle-mesh Ewald treatment of long-range electrostatics and periodic boundary conditions, as well as couplings to heat (300 K, 1 ps) and pressure (1 bar, 1 ps) baths; they were performed with GROMACS 3.3.2 [68] using the OPLS all-atom force field [69]. The final equilibrated box had dimensions 76x92x69 Å and comprised 13460 water molecules. Convergence of the protein structure was reached after 2 ns of simulation, at a mean RMSD of 1.90 Å from the crystal structure.

## A.3 Figure legends

**Legend of Fig. 10.** The $AB$ interface (colored Voronoi facets) and the interfacial water molecules $W$ (grey spheres) for two distinct rehydration and equilibration procedures – a fast (a) and a more exhaustive one (b); see text for details. Boundaries of the $AB$ and $AW - BW$ interfaces are shown as red and green sticks, respectively.

**Legend of Fig. 11.** Distribution of Voronoi facets with respect to Voronoi Shelling Order (heterodimers). Fractions are given on an absolute scale from 0 to 100%. The histogram of actual Voronoi facets is shown in black – most facets belong to the outer shells with VSO=1 or 2. The red line quantifies the fraction of dry facets within each shell – which increases towards inner shells. Since Mihalek et al. did not determine dryness values for each atom, facets were classified as dry if they belong to a residue that has no contact to itinerant water molecules. The fraction of dry facets thus represents a lower bound and does not consider the many facets that, even though belonging to a "wet" residue, have no direct water contact either. Note also that innermost shells comprise only very few facets and their classification becomes yet more abitrary. Keeping in mind these caveats, interface regions that are more than four Voronoi shells away from the rim ($VSO > 4$) can generally be assumed to have little contact with bulk water.

**Legend of Fig. 12 and 13.** Distribution of Voronoi facets with respect to Voronoi Shelling Order (homodimers). See description of Fig. 11.

## A.4 Figures

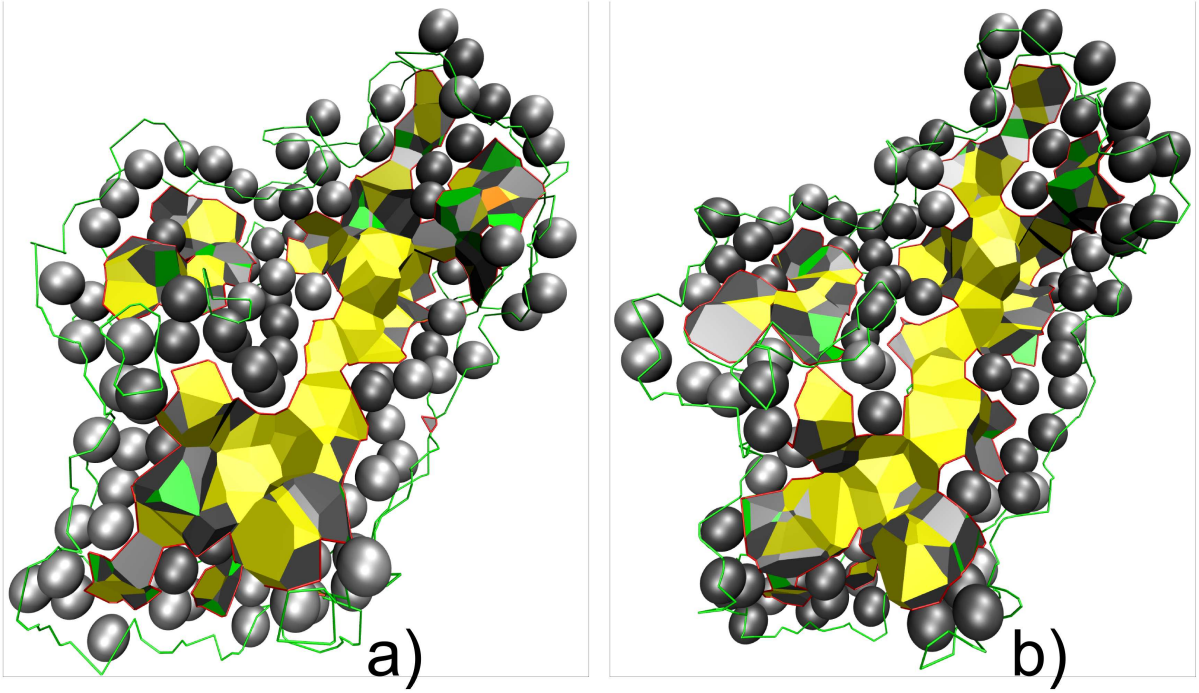



Figure 10:

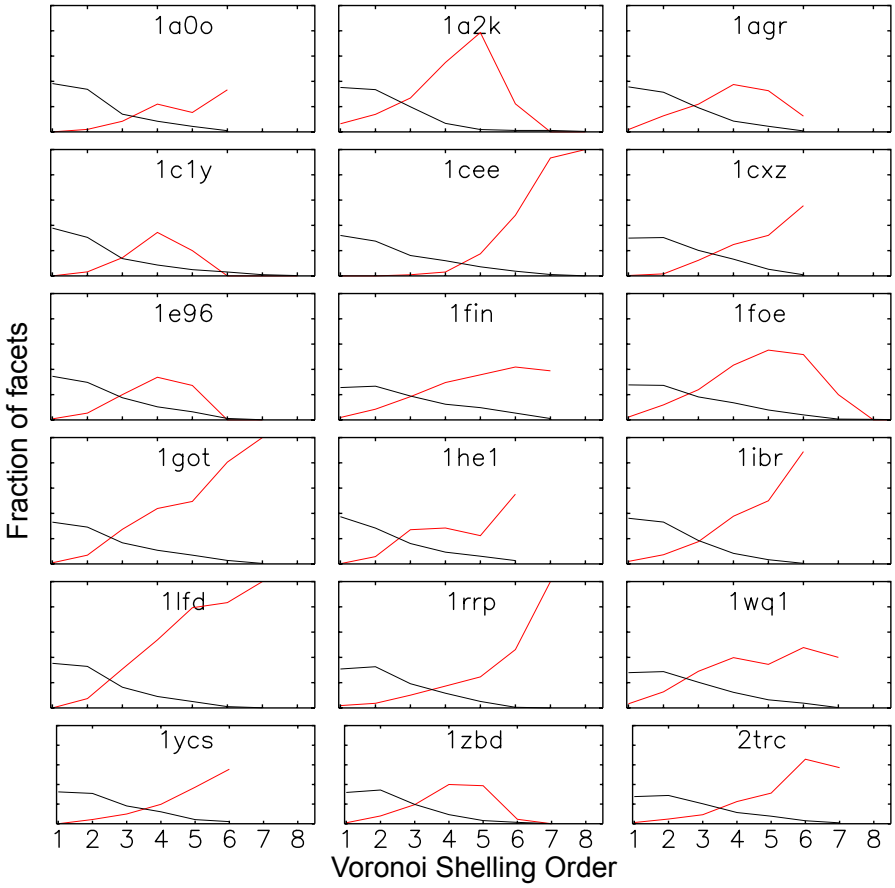

1a0o
1a2k
1agr
1c1y
1cee
1cxz
1e96
1fin
1foe
1got
1he1
1ibr
1lfd
1rrp
1wq1
1ycs
1zbd
2trc
Fraction of facets
1 2 3 4 5 6 7 8
Voronoi Shelling Order



Figure 11:

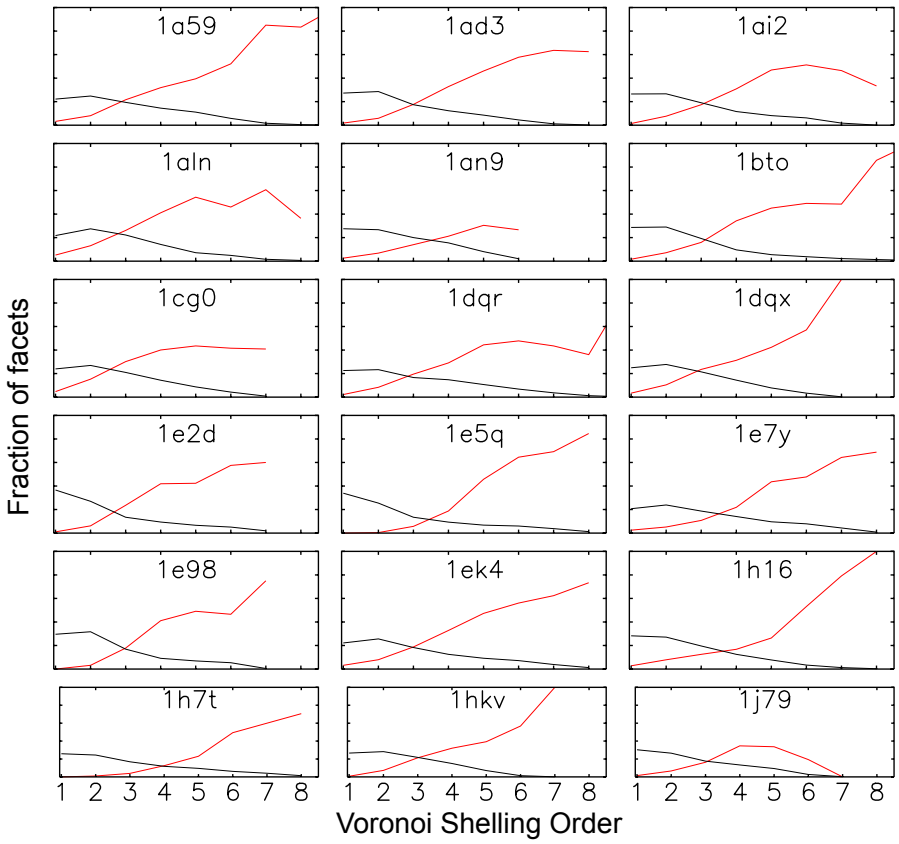

1a59
1ad3
1ai2
1aln
1an9
1bto
1cg0
1dqr
1dqx
1e2d
1e5q
1e7y
1e98
1ek4
1h16
1h7t
1hkv
1j79
Fraction of facets
1 2 3 4 5 6 7 8
Voronoi Shelling Order



Figure 12:

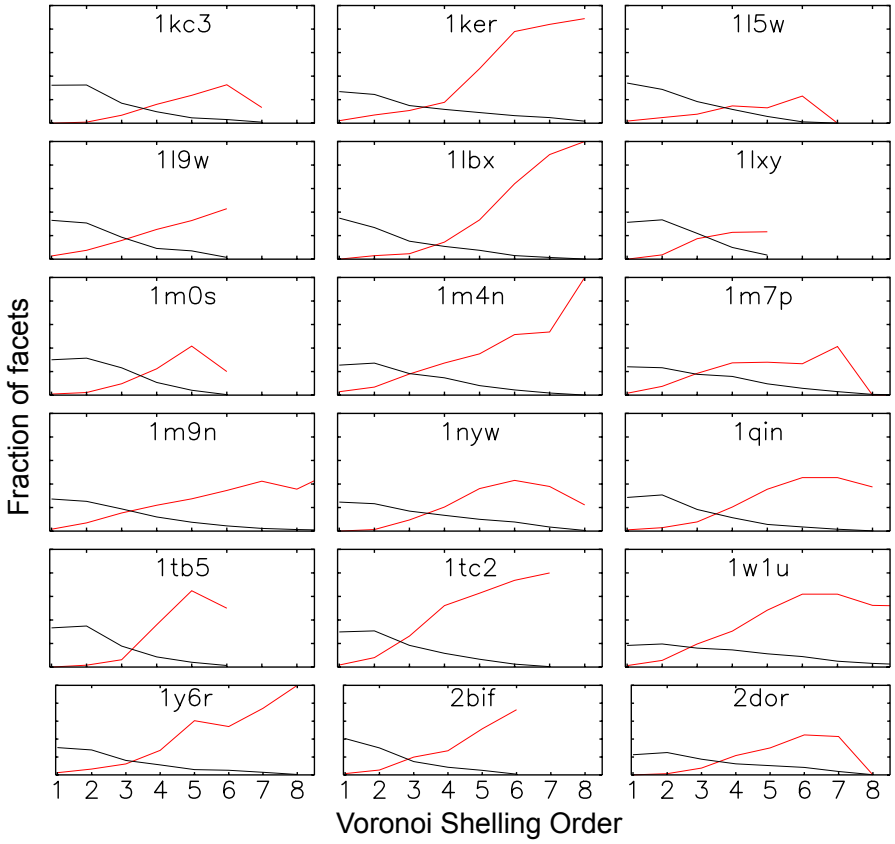

1kc3
1ker
1l5w
1l9w
1lbx
1lxy
1m0s
1m4n
1m7p
1m9n
1nyw
1qin
1tb5
1tc2
1w1u
1y6r
2bif
2dor
Fraction of facets
1 2 3 4 5 6 7 8
Voronoi Shelling Order



Figure 13: