Abstract

An important characteristic that determines the behavior of a solute in water is whether it is hydrophobic or hydrophilic. The traditional classification is based on chemical experience and heuristics. However, this does not reveal how the local environment modulates this important property. We present a new local fingerprint for hydrophobicity and hydrophilicity inspired by a term of the entropy multiparticle correlation expansion of the system. This fingerprint is an inexpensive, quantitative and physically meaningful way of studying hydrophilicity and hydrophobicity that only requires as input the water-solute radial distribution functions. As such, it can adapt to the changes in the local environment. We apply our fingerprint to systems such as octanol, benzene and the 20 proteinogenic amino acids. Our measure of hydrophilicity is coherent with chemical experience and, moreover, it also shows how the character of an atom can change as its environment is changed.

1 Introduction

Like dissolves like is one of the earliest chemical rules a scientist learns in relation to solvation. It implies that solutes that are chemically similar to water have a favorable interaction with water and are hydrophilic. On the other hand, solutes that are not like water will tend to repel water and be hydrophobic. Typically one assigns to each atom its own hydrophobicity or hydrophilicity based on chemical experience and heuristics. Despite the importance of these intuitive classifications, none of them is quantitative, nor takes into account thermodynamics or solvent structure. Processes like protein folding, the assembly of molecules, or crystallization depend crucially on their interaction with water. Thus it would be of great help to have a measure of the hydrophobicity and hydrophilicity of the atoms in a solute molecule and understand how these parameters change as the environment changes.

In the context of protein science, many hydrophobicty scales for aminoacids have been proposed based on empirical or computational...
data without any definitive consensus. Scales have been proposed that focus on the local hydrophobicity of selected heavy atoms. Some of them are based on local compressibility or density fluctuations of the hydration layers of proteins and surfaces.

Here we propose to use a related concept, namely the local structure environment of the different atoms. A quantity that ascribes local variation in density is the radial distribution function (RDF). Thus we suggest here a fingerprint that is a function of the RDF between solute atoms and water oxygens. This fingerprint has been inspired by our previous work on using approximated expressions for the entropy which has proven to be useful in distinguishing between solid-like and liquid-like environments and in nucleation processes. In a complex liquid like water angular correlation, ignored here, plays a major role as amply discussed in the literature. The fact however that some relation exists between the translational part of the entropy and our fingerprint can be considered an additional bonus. Strictly speaking our fingerprint measures the deviation of the local water structure from that of a perfect gas.

With respect to other entropy based hydrophobicity measures, our fingerprint has the advantage of being easy to compute and to be specified to each individual atom in the context of its molecular environment. One can thus assess the hydrophobicity of each individual atom and the modifications that result from changes in the environment. We apply this fingerprint to water and methane as representatives of optimal hydrophilicity and hydrophobicity, and to more complex systems such as octanol, benzene and the 20 proteinogenic amino acids.

2 Fingerprint for hydrophobicity and hydrophilicity

As discussed in the introduction, theory provides an expansion of the entropy of a liquid as a sum of many-body correlation functions. Inspired by this theoretical framework we propose the following term of the expansion as a local fingerprint for hydrophobicity and hydrophilicity of atom $i$.

$$S_i^s = -2\pi \rho_{w,loc} \int_0^{\infty} \left\{ g_{iw}(r) \ln[g_{iw}(r)] - g_{iw}(r) + 1 \right\} r^2 dr$$

where $\rho_{w,loc}$ is the local number density of water, and $g_{iw}(r)$ is the radial distribution function of the atom of the solute, $i$, with water oxygen atoms. The reader should bear in mind that this is not an expression for the excess entropy of the system but rather one of its contributions. Calculating the entropy requires including higher order terms and angular correlations as explained in the literature at a much higher computational cost. This defeats our purpose of having an inexpensive semiquantitative fingerprint useful also in enhanced sampling simulations.

Equation can be reexpressed in term of what is called the Bregmann divergence which is a measure of the distance between two positive definite functions. In the case of Equation it measures the divergence between $g_{iw}(r)$ and the perfect gas RDF, $g(r) = 1$.

Since the volume occupied by solutes can be comparable to that of a sphere of radius $r_{max}$, water has a big excluded volume. Therefore at $r_{max}$ the density is lower than at the bulk, and the RDF does not converge to 1. For this reason, the total density is replaced by the average density within a sphere of radius $r_{max}$ centered on the atom of interest. This ensures that the RDFs are all equivalently normalized regardless of the excluded volume of the solute.

Our hydrophobic fingerprint bridges the complex entropy theories and the simpler density based theories. The RDF is the basic ingredient of the our fingerprint. The simpler density based theories, use the solute density fluctuations and solute compressibility which are directly connected to the RDF. On the other hand, the more complex entropy calculation methods, require the calculation of angular contributions to the entropy and the translational terms cannot be radially averaged.

The fingerprint value in the simple case of a spherical cavity of radius $R$ embedded in an ideal solvent is instructive in this regard. In
this particular case the $g_{iw}(r)$ in equation $[1]$ is:

$$g_{iw}(r) = \begin{cases} 
0 & \text{if } r \leq R \\
1 & \text{if } r > R 
\end{cases} \quad (2)$$

If we introduce this step function in equation $[1]$, the following formula for the fingerprint of a cavity of volume $V = \frac{4}{3}\pi R^3$ is obtained:

$$S^i_{cav} = -\frac{\rho k_B V}{2} \quad (3)$$

This expression is the leading term of the solvation entropy in the information theory model of hydrophobic interactions, if one assumes that the solvent behaves ideally.

### 3 Computational Methods

All the systems used in this work were solutions of a single solute molecule with 1000 SPC/E water molecules at water density $0.997 \text{ g cm}^{-3}$. The solutes studied were: an SPC/E water molecule, methane, n-octanol, benzene and the 20 proteinogenic amino acids. The amino acids were simulated in their standard physiological protonation state and with N-methylated and C-acetylated termini. The OPLS force field was used for methane and octanol. AMBER03 was used for the amino acids and benzene. The partial charges of benzene were calculated at the B3LYP/cc-PVTZ level using the ESP method and the polarizable continuum model was used to mimic the aqueous environment. Water molecules were kept rigid using the SETTLE algorithm. For the rest of the solutes the bonds involving hydrogen were constrained with the P-LINKS algorithm. Lennard-Jones cross-term parameters were assigned using $\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{1/2}$ and $\sigma_{ij} = \frac{\sigma_{ii}\sigma_{jj}}{2}$, except in the case of AMBER03 where $\sigma_{ij} = (\sigma_{ii} + \sigma_{jj})^{1/2}$ was used.

All molecular dynamics (MD) simulations were run with GROMACSv5.1.1 in the NVT ensemble using the stochastic velocity rescaling thermostat at 298 K and a relaxation time $\tau=0.1\text{ ps}$. The equations of motion were integrated using the leapfrog algorithm with a 2 fs time step for a total time of 5 ns. Periodic boundary conditions were used and long-range electrostatic interactions were calculated with the PME method. Short range van der Waals interactions were truncated at 10 Å.

The calculations of the fingerprint were done using a development version of PLUMED 2. The RDF is calculated using a kernel density estimation of the radial distribution function. Which for a gaussian kernel is:

$$g_{iw}(r) = \frac{1}{4\pi\rho_{w,\text{loc}}} \sum_{j \in w} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(r-r_j)^2}{2\sigma^2}} \quad (4)$$

where $r_j$ is the distance between the fingerprinted atom, $i$, and the $j$th water molecule where $j$ runs over the set of water molecules. $\sigma$ is the gaussian kernel bandwidth. Kernel density estimation ensures that $g_{iw}(r)$ is continuous and differentiable with respect to atomic positions for future use as a collective variable in enhanced sampling simulations. In addition, this decreases the noise when the statistics is poor. Nevertheless, a conventional RDF would give identical results. The value of $\sigma$ was $0.05 \text{ Å}$ producing RDFs that are smooth but yet preserve all the relevant features. The fingerprint was integrated using the trapezoid rule. The upper integration limit was chosen to be $r_{\text{max}} = 10 \text{ Å}$. Equation $[1]$ corresponds to the single configuration $g_{iw}(r)$ which is averaged over the full length of the simulation before its use in Equation $[1]$ unless otherwise stated.

### 4 Results and Discussion

#### 4.1 Simple Solutes

Water and methane are paradigmatic cases of hydrophilic and hydrophobic solutes. Thus, their fingerprint values provide reference values. Water has an $S_s$ of $-1.56 \pm 0.01$ and methane of $-2.73 \pm 0.02$. Figure $[1]$ clarifies the physics behind these numbers. The top graph shows the radial distribution functions of the solutes and the bottom graph the integrand $I(r)$ of the fingerprint:

$$I(r) = -2\rho_w \{ g_{iw}(r) \ln [g_{iw}(r)] - g_{iw}(r) + 1 \} r^2 \quad (5)$$
The figure shows how $S_s$ varies with the radial structure of the solvent around the solute. In essence, $S_s$ becomes more negative the larger the deviation of the RDF from one. The more the solvent is structured around the solute, the smaller $S_s$. Because of the $r^2$ factor, the structuring at larger distances is especially effective in decreasing $S_s$. At short distances, for $r$ less than a distance $r_c$ of the order of the molecular radius, $g_{w}(r) \approx 0$ and this small $r$ region gives a contribution proportional to $r_c^3$. This contribution to $S_s$ corresponds to the cavity formation entropy. Methane has a lower $S_s$ than water for two reasons. First, it generates a larger cavity. Second, although its first hydration shell peak is less structured, it is wider, it is located at distances larger than the first hydration shell of water, and contains 4 times more water molecules.

We shall use the $S_s$ values for water and methane as representative of extreme hydrophilicity and hydrophobicity. It is therefore convenient to rescale the values of $S_s$ introducing an adimensional index $h$ that is $+1$ for water and $-1$ for methane. Thus, in this scale the sign of $h$ determines whether the atom is hydrophobic or hydrophilic.

![Figure 2: Octanol, water and methane molecules with their heavy atoms colored according to the $h$ index. The scale ranges from hydrophobic (blue), to intermediate (white) and to hydrophilic (orange).](image)

Figure 1: Top: RDFs of C-W for the aqueous methane simulation (blue) and of W-W of a pure water simulation (red). Bottom: for the same pairs, the integrand, $I(r)$, of the fingerprint is plotted.
solvation shells of neighboring CH\textsubscript{2} groups in the aliphatic chain overlap. This shifts the RDF first solvation peak to higher distances thus increasing \( h \). Since CH\textsubscript{3} has only one neighbor, this effect is less pronounced. Figure S1 of the Supporting Information (SI) illustrates this by analyzing the RDFs of primary, secondary, tertiary carbon atoms and methane. Figure S2 of the SI includes the numeric values of the pair entropies of the atoms in octanol.

4.2 Amino acids

The fingerprint for the heavy atoms of the 20 proteinogenic amino acids were computed, offering the possibility of testing our fingerprint on a wide range of chemical groups. This is a first step for future use in the study of hydrophobic and hydrophilic interactions in proteins. Figure 3 shows the different amino acid molecules with the heavy atoms in the side chains colored according to their \( h \) value. The backbone atoms are shown only for glycine but a similar picture is obtained for the other amino acids. As in the case of octanol, \( h \) assigns a hydrophobic value to aliphatic carbons and an hydrophilic value to polar N and O atoms of hydrophilic residues. All the heavy atoms of the backbone have \( h \) values adequate to the hydrophilicity or hydrophobicity that chemical intuition suggests. A list of \( h \) values can be found in Figure S3 of the SI.

While most of the \( h \) values reflect the expected behavior, some apparently surprising values can be seen. For instance the aromatic C are placed in the middle of the \( h \) scale and thus they are classified as neither properly hydrophilic or hydrophobic. In reality this result is in line with the known solvation behavior of benzene which is much more soluble than its aliphatic counterpart cyclohexane. The reasons for this effect have been discussed in the literature.\textsuperscript{3,11} As seen from the point of view of our fingerprint, this results from the fact that the other atoms in the ring exclude some of the solvation water leading to a reduction in the RDF peak height. In order to confirm that this behavior is not an artifact of our force field, we have calculated \( S\textsubscript{s} \) using the benzene RDF kindly provided to us by Choudary et al obtained using ab initio MD.\textsuperscript{42} The ab initio value, \( S\textsubscript{s} = -1.9 \, k_B \) is very close to that of the AMBER force field. Here we did not scale the \( S\textsubscript{s} \) values since we do not have the ab initio reference point for methane.

Another \( h \) value that deserves some discussion is that of the sulfur atoms with a \( h \sim 0 \). This can be linked to the fact that the electronegativity of sulfur is intermediate between carbon and oxygen, and to the ability of sulfur to accept weak H bonds.\textsuperscript{43,44}

Since we relate \( h \) to the water solvation structure and the water structure around each atom can fluctuate as a function of time, we also looked at the temporal variation of this index. In order to reduce the noise in the calculation of \( h \), we averaged the RDF over a 50 ps moving window. This time was chosen in order to obtain smooth RDFs. The data thus obtained were put in a histogram in which we considered separately aliphatic C, aromatic C, S, and O and N of the side chains. The histograms are shown in Figure 4.

In the histogram, the hydrophobic aliphatic C are clearly separated from the hydrophilic O and N of the side chains, proving the usefulness of the fingerprint. As discussed previously, the distribution of the aromatic C and S are centered around \( h \sim 0 \). The distribution of the hydrophilic O and N of the side chains (shown in red in Figure 4) presents two peaks and a shoulder. The peak at \( h \sim 0.8 \) corresponds to all the hydrophilic O and N of charged amino acids with the exception of arginine, while the other peak at \( h \sim 1.4 \) corresponds to hydrophilic O and N of neutral amino acids and arginine. Charged residues have a lower \( h \) than neutral ones because they induce more structure in water. Arginine is an exception to this rule due to its higher charge delocalization and therefore leads to a less well-defined solvation structure. The shoulder at \( 0 < h < 0.6 \) in the histogram of O and N of the side chains corresponds to glutamate since carboxylate oxygens have a very negative effective charge with respect to the rest of hydrophilic atoms. The histogram of hydrophobic aliphatic C has two peaks. The peak at \( h \sim -0.4 \) corresponds to CH\textsubscript{3} carbon atoms,
Figure 3: Structures of the proteinogenic amino acids with their heavy atoms colored according to the $h$ index. The scale ranges from hydrophobic (blue), to intermediate (white) and to hydrophilic (orange). Unlabeled atoms are carbon and hydrogen atoms are omitted. Since all backbone atoms have similar $h$ index, only sidechain atoms are considered. Backbone atoms are visible for glycine (gray box). The boxes organize the amino acids by families: hydrophilic (red), glycine (gray), sulfur-containing (green), aromatic (black) and hydrophobic (blue).
Figure 4: Probability densities of the hydrophobicity fingerprint, $h$, of different groups of atoms in their respective simulations. The lines are the distributions of $S_s$ for: C atoms of hydrophobic amino acids (blue), N and O atoms of hydrophilic amino acids (red), aromatic C atoms (purple) and S atoms (green).

while the broad peak at $h \sim -1.75$ corresponds to CH$_2$ and CH carbon atoms. This behavior has been discussed earlier in Section 4.1 for octanol. The histogram for aromatic C shows three peaks. The two peaks around $h \sim 0.25$ correspond to the more solvent exposed aromatic C while the remaining peak centered at $h \sim -1.25$ corresponds to C closer to the C$\beta$.

5 Conclusions

We have developed a fingerprint for hydrophobicity and hydrophilicity. The fingerprint is inspired by the two body solute water contributions to the entropy which is a function of the RDF. Therefore whether an atom is hydrophobic or hydrophilic is a consequence by the structure of water around it. This feature allows to understand how the character of a solute is modulated by its environment. We have also introduced an index of hydrophilicity $h$ that uses methane and water as representatives of hydrophobic and hydrophilic behavior. In the future we plan to use this fingerprint as a collective variable in enhanced sampling simulations.

We expect that the fingerprint could provide insight into phenomena where hydrophobicity plays an important role such as protein folding.

Acknowledgement The authors thank Ashu Choudhary and Amalendu Chandra for kindly providing us with their ab initio benzene RDF. The authors also thank Enrique Sánchez Marcos for the helpful discussion. S.P.C. also acknowledge the Spanish Ministry of Education, Culture and Sports for the Ph.D. grant (FPU14/02100). P.M.P and M.P acknowledge supported from the NCCR MARVEL, funded by the Swiss National Science Foundation. The computational time for this work was provided by the Swiss National Supercomputing Center (CSCS) under Project ID p503.

Supporting Information Available

The following files are available free of charge.

Supporting Information (PDF): Additional information including extra RDFs, fingerprint integrands and numeric values of the pair en-
tropies not displayed in the text.

References

(1) Simm, S.; Einloft, J.; Mirus, O.; Schleiff, E. Biological research 2016, 49, 31.

(2) Rossky, P. J. Faraday Discussions 2010, 146, 13–18.

(3) Xi, E.; Venkateshwaran, V.; Li, L.; Rego, N.; Patel, A. J.; Garde, S. Proceedings of the National Academy of Sciences 2017, 114, 20170092.

(4) Piaggi, P. M.; Parrinello, M. Journal of Chemical Physics 2017, 147.

(5) Piaggi, P. M.; Valsson, O.; Parrinello, M. Physical Review Letters 2017, 119, 015701.

(6) Lazaridis, T.; Paulaitis, M. E. The Journal of Physical Chemistry 1992, 96, 3847–3855.

(7) Lazaridis, T.; Paulaitis, M. E. The Journal of Physical Chemistry 1994, 98, 635–642.

(8) Lazaridis, T. The Journal of Physical Chemistry B 2000, 104, 4964–4979.

(9) Kinoshita, M.; Matubayasi, N.; Harano, Y.; Nakahara, M. The Journal of chemical physics 2006, 124, 24512.

(10) Liu, M.; Besford, Q. A.; Mulvaney, T.; Gray-Weale, A. The Journal of chemical physics 2015, 142, 114117.

(11) Lynden-Bell, R. M.; Rasaiah, J. C. The Journal of chemical physics 1997, 107, 1981–1991.

(12) Bergman, D. L.; Lyubartsev, A. P.; Laaksonen, A. Physical Review E 1999, 60, 4482.

(13) Lazaridis, T. The Journal of Physical Chemistry B 1998, 102, 3531–3541.

(14) Nguyen, C. N.; Kurtzman Young, T.; Gilson, M. K. Journal of Chemical Physics 2012, 137, 973–980.

(15) Li, Z.; Lazaridis, T. Computational Drug Discovery and Design; Springer, 2012; pp 393–404.

(16) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Journal of the American Chemical Society 2008, 130, 2817–2831.

(17) Huggins, D. J.; Payne, M. C. Journal of Physical Chemistry B 2013, 117, 8232–8244.

(18) Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. Journal of Chemical Theory and Computation 2014, 10, 2769–2780.

(19) Green, H. S. The molecular theory of fluids; North-Holland Publishing Company Amsterdam, 1952.

(20) Nettleton, R. E.; Green, M. S. The Journal of Chemical Physics 1958, 29, 1365–1370.

(21) Baranyai, A.; Evans, D. J. Physical Review A 1989, 40, 3817.

(22) Piaggi, P. M.; Parrinello, M. 2018, 1–7.

(23) Garde, S.; Hummer, G.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. Physical review letters 1996, 77, 4966.

(24) Hummer, G.; Garde, S.; Garcia, A. E.; Pohorille, A.; Pratt, L. R. Proceedings of the National Academy of Sciences 1996, 93, 8951–8955.

(25) Chandler, D. Nature 2005, 437, 640–647.

(26) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. Journal of Physical Chemistry 1987, 91, 6269–6271.

(27) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Journal of Physical Chemistry B 2001, 105, 6474–6487.
(28) Duan, Y.; Wu, C.; Chowdhury, S. S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *Journal of computational chemistry* **2003**, *24*, 1999–2012.

(29) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comp. Chem.* **1990**, *11*, 431–439.

(30) Tomasi, J.; Tomasi, J.; Mennucci, B.; Mennucci, B.; Cammi, R.; Cammi, R. *Chemical Reviews* **2005**, *105*, 2999–3094.

(31) Miyamoto, S.; Kollman, P. A. *Journal of computational chemistry* **1992**, *13*, 952–962.

(32) Hess, B. *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122.

(33) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. *SoftwareX* **2015**, *1-2*, 19–25.

(34) Bussi, G.; Donadio, D.; Parrinello, M. *Journal of Chemical Physics* **2007**, *126*.

(35) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(36) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

(37) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Computer Physics Communications* **2009**, *180*, 1961–1972.

(38) McAuliffe, C. *The Journal of Physical Chemistry* **1966**, *70*, 1267–1275.

(39) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *Journal of solution chemistry* **1981**, *10*, 563–595.

(40) Ben-Amotz, D. *Annual Review of Physical Chemistry* **2016**, *67*, 617–638.

(41) Harris, R. C.; Pettitt, B. M. *Journal of Physics: Condensed Matter* **2016**, *28*, 083003.

(42) Choudhary, A.; Chandra, A. *Physical Chemistry Chemical Physics* **2016**, *18*, 6132–6145.

(43) Biswal, H. S.; Shirhatti, P. R.; Wategaonkar, S. *J. Phys. Chem. A* **2010**, *114*, 6944–6955.

(44) Rao Mundlapati, V.; Ghosh, S.; Bhattacherjee, A.; Tiwari, P.; Biswal, H. S. *Journal of Physical Chemistry Letters* **2015**, *6*, 1385–1389.
