HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2016 May 23.

Published in final edited form as:

J Chem Theory Comput. 2015 November 10; 11(11): 5090–5102. doi:10.1021/acs.jctc.5b00439.

Estimation of Solvation Entropy and Enthalpy via Analysis of Water Oxygen–Hydrogen Correlations

Camilo Velez-Vega[†,*], **Daniel J. J. McKay**[†,*], **Tom Kurtzman**[‡,§,*], **Vibhas Aravamuthan**[†], **Robert A. Pearlstein**[†], and **José S. Duca**[†]

[†]Computer-Aided Drug Discovery, Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 100 Technology Square, Cambridge, Massachusetts 02139, United States

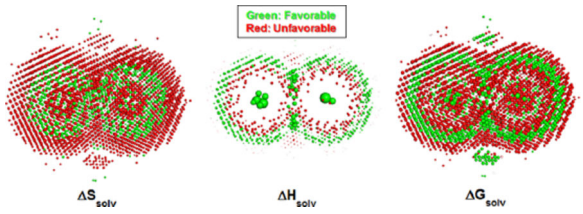
[‡]Department of Chemistry, Lehman College, The City University of New York, 250 Bedford Park Boulevard West, Bronx, New York 10468, United States

[§]Ph.D. Program in Chemistry, The Graduate Center of the City University of New York, New York, New York 10016, United States

Abstract

A statistical-mechanical framework for estimation of solvation entropies and enthalpies is proposed, which is based on the analysis of water as a mixture of correlated water oxygens and water hydrogens. Entropic contributions of increasing order are cast in terms of a Mutual Information Expansion that is evaluated to pairwise interactions. In turn, the enthalpy is computed directly from a distance-based hydrogen bonding energy algorithm. The resulting expressions are employed for grid-based analyses of Molecular Dynamics simulations. In this first assessment of the methodology, we obtained global estimates of the excess entropy and enthalpy of water that are in good agreement with experiment and examined the method's ability to enable detailed elucidation of solvation thermodynamic structures, which can provide valuable knowledge toward molecular design.

Graphical abstract



[*]**Corresponding Authors**. ; Email: camilo.velez-vega@novartis.com (C.V.-V.). ; Email: daniel.mckay@novartis.com (D.J.J.M.). ; Email: thomas.kurtzman@lehman.cuny.edu (T.K.)

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00439.

Appendix A: Normalization of the solvation entropy differences relative to bulk solvent, Appendix B: Solvation entropy: Water as a polyatomic fluid, Appendix C: Estimation of the error in the entropy calculations upon disregarding contributions of voxels around the center voxel (PDF)

The authors declare no competing financial interest.

INTRODUCTION

In silico approaches aimed at the analysis of structural and thermodynamic properties of water have in recent years shown promise as tools to aid ligand design. The palpable interest in these methods is naturally linked to the key role that water plays in most biophysical processes, which can in principle be elucidated through information extracted from solvent structural dynamics. Indeed, we have previously developed and reported[1] a novel approach yielding detailed time-averaged solute and solvent structures that have been used to reveal relevant insights pertaining to biomolecular mechanisms and, more importantly, have facilitated rational ligand design. Yet, further development of this approach, aimed at bridging the gap between structure-free energy relationships within the context of solvation-based analyses, is a necessary advancement toward more efficient drug discovery.

Several methods have been proposed for estimation of thermodynamic quantities from solvation dynamics, which are either rooted in Inhomogeneous Solvation Theory (IST)[2–8] (GIST,[9,10] STOW,[11] WaterMap,[12,13] and others[14,15]) or alternate statistical mechanical fundamentals (3D-RISM,[16] GCT,[17] SZMAP,[18] and WATsite[19]). Although these approaches show great potential for expediting ligand design, their systematic use has been largely hampered by the limited accuracy of the resulting solvation thermodynamic quantities. The latter may be in part due to the apparent drawbacks of some of these methodologies, which include the following:

1. The use of fixed-solute simulations to study solvation dynamics,[2–9,12,17] which may lead to nonrepresentative solvation structures for flexible systems, and can mask inconsistencies in the system/simulation setup.
2. The study of high-occupancy hydration sites only,[12,14,19,20] disregarding low-occupancy solvation and its corresponding solvation thermodynamics; the latter contribution can have critical impact on ligand design and optimization.
3. The demanding computational effort required to properly evaluate second order solvation entropy terms (e.g., for implementations derived from IST[2–9,11]), difficulty that escalates upon estimation of third and higher order correlations.

This work is part of an ongoing effort aimed at further developing solvation dynamics-based methodologies toward increasingly dependable and efficient tools for drug design.

We present here a new theoretical platform for calculating water entropy, in which increasing order contributions are expressed in terms of a Mutual Information Expansion[21] (MIE). This treatment allows for the straightforward discretization of space, leading to an efficient, grid-based, all-density occupancy algorithm for analysis of solvent structures extracted from flexible solute simulations. The proposed entropy formulation is centered on the breakdown of water into a mixture of correlated oxygens and hydrogens, a perspective initially prompted by the mobile nature of water's constituent atoms. This approach entails handling only relative atomic positions, thus offering a more tractable way to compute solvation entropies relative to the practice (as in IST implementations) of estimating distinct terms for the translational and orientational entropy of water molecules.

Water enthalpy values are also computed within this grid-based analysis protocol. In brief, nonbonded interactions between waters are approximated by their corresponding hydrogen bonding energies. Free energies are then calculated as the sum of entropic cost and the enthalpic contributions.

Overall, the methodology introduced provides both a global estimate and a detailed structural description of the solvation entropy, enthalpy, and free energy of a system.

As a key validation step, the proposed thermodynamics expressions were used to compute global estimates of the excess enthalpy and entropy of pure water under physiological conditions. The ability of our method to portray local solvation structures is then illustrated via analysis of pure water systems whose perturbed solvent fields (generated upon fixing the position of two water molecules) are either interacting or noninteracting.

METHODS

Theoretical Formulations

Solvation Entropy – Integral Expressions—The total entropy S of a solvated system can be expressed in terms of the joint probability density (ρ) of the degrees of freedom describing the solute (***A***) and the solvent (***B***)[22]

$$S = -k_{\mathrm{B}} \int \int \rho(\boldsymbol{A},\boldsymbol{B})^{*}\ln(\rho(\boldsymbol{A},\boldsymbol{B}))d\boldsymbol{A}d\boldsymbol{B}+k_{\mathrm{B}}\ \ln(C) \quad (1)$$

where k_{B} is the Boltzmann constant, and the last term is a constant that eliminates the units of the probability density inside the logarithm and cancels out exactly upon computing differences in entropy.

By virtue of the conditional probability law, ρ(***A,B***) = ρ(***A***)ρ(***B***|***A***) = ρ(***B***)ρ(***A***|***B***), eq 1 can be denoted as

$$S=-k_{\mathrm{B}} \int \rho(\boldsymbol{A})^{*}\ln(\rho(\boldsymbol{A}))d\boldsymbol{A}-k_{\mathrm{B}} \int \rho(\boldsymbol{A})d\boldsymbol{A}\int \rho(\boldsymbol{B}|\boldsymbol{A})\ ^{*}\ln(\rho(\boldsymbol{B}|\boldsymbol{A}))d\boldsymbol{B}+k_{\mathrm{B}}\ \ln(C) \quad (2)$$

The first term corresponds to the entropy of the solute degrees of freedom, whereas the second term has been elegantly interpreted by Zhou and Gilson[22] as the entropy of the solvent degrees of freedom (***B***) averaged over the equilibrium distribution of the solute degrees of freedom (***A***). Thus, eq 2 can be further cast as the sum of separate contributions to the entropy arising from motions of the solute molecules (S_{solute}) and from motions of the solvent (S_{solv}) due to the presence of solute:

$$S=S_{\mathrm{solute}}+S_{\mathrm{solv}} \quad (3)$$

The first and second terms in eq 2 are entirely included in S_{solute} and S_{solv}, respectively. This work does not consider the estimation of S_{solute} in eq 3. For that purpose, the reader is referred to previous research[21,23–26] relevant to the estimation of configurational entropies.

We are primarily concerned with computing solvation entropy differences (ΔS_{solv}) relative to the bulk solvent at conditions relevant to biological systems. To this end, we propose ΔS_{solv} expressions that are cast in terms of normalized probability densities (relative to bulk), denoted by $\bar{\rho}$ (as described in Appendix A, Supporting Information). Note again that calculation of solvation entropy differences ensures cancellation of the constant contribution to S_{solv} arising from the last term in eq 2.

Below, we derive expressions for ΔS_{solv} in which water is considered as a mixture of N oxygens (x) and $2N$ correlated hydrogens (y), whose reference state corresponds to a uniform distribution of oxygens and hydrogens. For this case

$$\begin{aligned}\Delta S_{\mathrm{solv}} = & -k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{r}_{x1},\boldsymbol{p}_{x1},\boldsymbol{r}_{y1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{p}_{xN},\boldsymbol{r}_{y2N},\boldsymbol{p}_{y2N}) \\ & * \int\cdots\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{p}_{x1},\boldsymbol{r}_{y1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{p}_{xN},\boldsymbol{r}_{y2N},\boldsymbol{p}_{y2N}) \\ & {}^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{p}_{x1},\boldsymbol{r}_{y1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{p}_{xN},\boldsymbol{r}_{y2N},\boldsymbol{p}_{y2N}))d\boldsymbol{r}_{x1}d\boldsymbol{p}_{x1}d\boldsymbol{r}_{y1}d\boldsymbol{p}_{y1} \\ & \ldots d\boldsymbol{r}_{xN}d\boldsymbol{p}_{xN}d\boldsymbol{r}_{y2N}d\boldsymbol{p}_{y2N}\end{aligned} \tag{4}$$

where $\boldsymbol{r}$ and $\boldsymbol{p}$, respectively, correspond to the Cartesian coordinates and momenta of all oxygens (x) and hydrogens (y) in the system. The normalized joint probability density $\bar{\rho}$ is a function of all degrees of freedom of the solvent atoms, whereas ρ_{bulk} can be computed as the product of the marginal bulk probability density of all such quantities.

When the dynamics of a system is modeled via a Hamiltonian in Cartesian coordinates (as is usually the case for MD simulations), the momentum and spatial contributions are uncorrelated; i.e., the kinetic and potential energy terms depend only on the momenta and the spatial coordinates, respectively. Therefore, eq 4 can be separated into

$$\begin{aligned}\Delta S_{\mathrm{solv}} = & -k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N}) \int\cdots\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N}) \\ & {}^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1},\ldots,\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N}))d\boldsymbol{r}_{x1}d\boldsymbol{r}_{y1}\ldots d\boldsymbol{r}_{xN}d\boldsymbol{r}_{y2N} \\ & -k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{p}_{x1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{p}_{xN},\boldsymbol{p}_{y2N}) \int\cdots\int \overline{\rho}(\boldsymbol{p}_{x1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{p}_{xN},\boldsymbol{p}_{y2N}) \\ & {}^{*}\ln(\overline{\rho}(\boldsymbol{p}_{x1},\boldsymbol{p}_{y1},\ldots,\boldsymbol{p}_{xN},\boldsymbol{p}_{y2N}))d\boldsymbol{p}_{x1}d\boldsymbol{p}_{y1}\ldots d\boldsymbol{p}_{xN}d\boldsymbol{p}_{y2N}\end{aligned} \tag{5}$$

Furthermore, given that the momentum term depends only on temperature and the weighted atomic mass $\overline{m}=\left(\prod_{i=1}^{S} m_i\right)^{1/s}$ computed from all s atoms in the system (which is invariant to changes in the number of waters),[27] the last term in eq 5 is negligible for thermostated simulations. Moreover, this term cancels out when taking differences relative to the reference bulk state. Thus, with $\boldsymbol{r}_{\boldsymbol{x}} = \{\boldsymbol{r}_{x1},\ldots,\boldsymbol{r}_{xN}\}$ and $\boldsymbol{r}_{\boldsymbol{y}} = \{\boldsymbol{r}_{y1},\ldots,\boldsymbol{r}_{y2N}\}$, we obtain

$$\Delta S_{\mathrm{solv}}(\boldsymbol{r}_{\boldsymbol{x}},\boldsymbol{r}_{\boldsymbol{y}}) = -k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{r}_{\boldsymbol{x}},\boldsymbol{r}_{\boldsymbol{y}}) \int\int \overline{\rho}(\boldsymbol{r}_{\boldsymbol{x}},\boldsymbol{r}_{\boldsymbol{y}})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{\boldsymbol{x}},\boldsymbol{r}_{\boldsymbol{y}}))d\boldsymbol{r}_{\boldsymbol{x}}d\boldsymbol{r}_{\boldsymbol{y}} \tag{6}$$

In this work, we estimate the solvation entropy defined by eq 6 in terms of increasingly higher order terms of the Mutual Information Expansion for this quantity, as adopted from the formulation proposed by Matsuda[28] and Killian et al.[21]

$$S^{(D)}(\alpha_1,\ldots,\alpha_m)=\sum_{i=1}^{m}S_1(\alpha_i)-\sum_{C_2^m}I_2(\alpha_i,\alpha_j)+\sum_{C_3^m}I_3(\alpha_i,\alpha_j,\alpha_k)-\sum_{C_4^m}I_4(\alpha_i,\alpha_j,\alpha_k,\alpha_l)+\ldots \quad (7)$$

where D is the order of the truncation in the entropy expansion, I is the mutual information (MI), variables α are the degrees of freedom defining the probability density functions (e.g., positions, angles, torsions), and C_d^m are all distinct order $d \leq D$ combinations of m degrees of freedom.

Expansion of eq 6 based on eq 7 leads to

$$\Delta S_{\mathrm{solv}} \equiv S^{(D)}(\boldsymbol{r}_{xi},\boldsymbol{r}_{yi})=\sum_{i=1}^{N}S_1(\boldsymbol{r}_{xi})+\sum_{i=1}^{2N}S_1(\boldsymbol{r}_{yi})-\left[\sum_{C_2^{2N}}I_2(\boldsymbol{r}_{xi},\boldsymbol{r}_{yj})+\frac{1}{2}\sum_{C_2^{N}}I_2(\boldsymbol{r}_{xi},\boldsymbol{r}_{xj})+\frac{1}{2}\sum_{C_2^{2N}}I_2(\boldsymbol{r}_{yi},\boldsymbol{r}_{yj})\right]+O(3)+\ldots \quad (8)$$

The first and second summations of eq 8 correspond to the marginal (S_1) entropies, which collectively represent an upper bound for the entropy.[21] The 1/2 factor preceding the second and third pairwise MI (I_2) summations prevents double counting of correlations of the same species.

Eq 8 can be directly expanded in terms involving joint probability distributions:

$$
\begin{aligned}
\Delta S_{\text{solv}} \equiv S^{(D)}(\boldsymbol{r}_{xi},\boldsymbol{r}_{yi}) = &-k_{\text{B}}\rho_{\text{bulk}}(\boldsymbol{r}_x)\Big[\int \overline{\rho}(\boldsymbol{r}_{x1})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x1}))d\boldsymbol{r}_{x1}+ \\
&\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{xN}))d\boldsymbol{r}_{xN}\Big]-k_{\text{B}}\rho_{\text{bulk}}(\boldsymbol{r}_y)\Big[\int \overline{\rho}(\boldsymbol{r}_{y1}) \\
&{}^{*}\ln(\overline{\rho}(\boldsymbol{r}_{y1}))d\boldsymbol{r}_{y1}+\ldots+\int \overline{\rho}(\boldsymbol{r}_{y2N})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{y2N}))d\boldsymbol{r}_{y2N}\Big] \\
&-k_{\text{B}}\rho_{\text{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)\times\Bigg[\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{y1})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{y1}+\ldots+ \\
&\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y2N})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y2N})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{y2N})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{y2N}+ \\
&\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y1})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{y1})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{y1}+\ldots+ \\
&\ldots+\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{y2N})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{y2N}\Bigg]-\frac{1}{2}k_{\text{B}}\rho_{\text{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_x) \\
&\times\Bigg[\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{x2})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{x2})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{x2})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{x2}+\ldots+\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{xN}) \\
&{}^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{xN})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{xN})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{xN}+\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{x1}) \\
&{}^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{x1})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{x1})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{x1}+\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{xN-1}) \\
&{}^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{xN-1})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{xN-1})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{xN-1}\Bigg]-\frac{1}{2}k_{\text{B}}\rho_{\text{bulk}}(\boldsymbol{r}_y,\boldsymbol{r}_y) \\
&\times\Bigg[\int \overline{\rho}(\boldsymbol{r}_{y1},\boldsymbol{r}_{y2})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{y1},\boldsymbol{r}_{y2})}{\overline{\rho}(\boldsymbol{r}_{y1})\overline{\rho}(\boldsymbol{r}_{y2})}\right)d\boldsymbol{r}_{y1}d\boldsymbol{r}_{y2}+\ldots+ \\
&\int \overline{\rho}(\boldsymbol{r}_{y1},\boldsymbol{r}_{y2N})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{y1},\boldsymbol{r}_{y2N})}{\overline{\rho}(\boldsymbol{r}_{y1})\overline{\rho}(\boldsymbol{r}_{y2N})}\right)d\boldsymbol{r}_{y1}d\boldsymbol{r}_{y2N}+ \\
&\ldots+\int \overline{\rho}(\boldsymbol{r}_{y2N},\boldsymbol{r}_{y1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{y2N},\boldsymbol{r}_{y1})}{\overline{\rho}(\boldsymbol{r}_{y2N})\overline{\rho}(\boldsymbol{r}_{y1})}\right)d\boldsymbol{r}_{y2N}d\boldsymbol{r}_{y1} \\
&+\ldots+\int \overline{\rho}(\boldsymbol{r}_{y2N},\boldsymbol{r}_{y2N-1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{y2N},\boldsymbol{r}_{y2N-1})}{\overline{\rho}(\boldsymbol{r}_{y2N})\overline{\rho}(\boldsymbol{r}_{y2N-1})}\right)d\boldsymbol{r}_{y2N}d\boldsymbol{r}_{y2N-1}\Bigg] \\
&+O(3)+\ldots
\end{aligned}
\tag{9}
$$

Eq 9 is a general expression for calculation of ΔS_{solv} referenced to water as a mixture of two monatomic fluids. For completeness, Appendix B, Supporting Information offers a MIE-based formulation akin to the one presented in eq 9, resulting from considering water as a polyatomic molecule. The latter may be of interest to researchers studying implementations derived from or related to IST. Moreover, it is anticipated that further exploration of the derivation shown in Appendix B, Supporting Information will lead to entropy formulations closely reminiscent of those proposed by IST.

We now formulate discrete expressions that allow estimation of ΔS_{solv} from eq 9, via analysis of MD simulations.

Solvation Entropy − Discrete Expressions—Capitalizing on the indistinguishable nature of the oxygen and hydrogen atoms comprising the solvent, it is recognized that the marginal contribution to the solvation entropy, computed as the sum of the integrals over phase space of all individual particles (e.g., *N* particles for oxygen; see term 1 in eq 9), can be computed via the Riemann sum of the solvation entropy contributions of all individual volume elements (e.g., voxels) used to discretize the simulation space. For instance, the term corresponding to the oxygen marginal translational solvation entropy in eq 9 can be denoted as

$$\begin{aligned} -k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{r}_x)[\int \overline{\rho}(\boldsymbol{r}_{x1})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x1}))d\boldsymbol{r}_{x1}+\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{xN}))d\boldsymbol{r}_{xN}] \\ =-k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x)\sum_{k=1}^{K}\int_{V_k}\overline{\rho}(\boldsymbol{r}_{x,k})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x,k}))d\boldsymbol{r}_{x,k} \end{aligned} \tag{10}$$

where *k* is the voxel index, *K* is the total number of voxels, $\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_{x,k})$ is the bulk density of oxygen in Å^{-3}, *V* is the voxel volume (in this work, it is equal for all voxels) in Å^3, and $\overline{\rho}(\boldsymbol{r}_{x,k})$ is the normalized probability density of finding an oxygen atom (*x*) at position ***r*** within voxel *k*.

For small voxel sizes such as the one employed in this work (i.e., [0.5 Å]^3), it is reasonable to assume that the variation in the probability density within each of these subregions is small.[9] As a result, the integral over each voxel can be represented by its average value, and the right-hand side of eq 10 is then expressed as

$$-k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x)\sum_{k=1}^{K}\int_{V_k}\overline{\rho}(\boldsymbol{r}_{x,k})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x,k}))d\boldsymbol{r}_{x,k}\ \approx -k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x)V\sum_{k=1}^{K}\overline{\rho}(\boldsymbol{r}_{x,k})^{*}\ln(\overline{\rho}(\boldsymbol{r}_{x,k})) \tag{11}$$

The pairwise correlation integrals can be discretized in an analogous fashion; e.g., the first pairwise MI term in eq 9 would be expressed as

$$\begin{aligned} k_{\mathrm{B}}\rho_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)\Bigg[&\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y1})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{y1})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{y1}+ \\ &\ldots+\int \overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y2N})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x1},\boldsymbol{r}_{y2N})}{\overline{\rho}(\boldsymbol{r}_{x1})\overline{\rho}(\boldsymbol{r}_{y2N})}\right)d\boldsymbol{r}_{x1}d\boldsymbol{r}_{y2N}+ \\ &\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y1})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y1})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{y1})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{y1}+ \\ &\ldots+\int \overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{xN},\boldsymbol{r}_{y2N})}{\overline{\rho}(\boldsymbol{r}_{xN})\overline{\rho}(\boldsymbol{r}_{y2N})}\right)d\boldsymbol{r}_{xN}d\boldsymbol{r}_{y2N}\Bigg] \\ \approx\ & k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)V^{2}\sum_{k=1}^{K}\left[\sum_{n\neq k}^{K}\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})}{\overline{\rho}(\boldsymbol{r}_{x,k})\overline{\rho}(\boldsymbol{r}_{y,n})}\right)\right] \end{aligned} \tag{12}$$

where *n* is the index for all voxels around the central voxel *k*.

Incorporating all discretized terms into eq 9, we obtain

$$
\begin{aligned}
\Delta S_{\mathrm{solv}} \approx & -k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x)V\sum_k \overline{\rho}(\boldsymbol{r}_{x,k})\ln(\overline{\rho}(\boldsymbol{r}_{x,k})) \\
& - k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_y)V\sum_k \overline{\rho}(\boldsymbol{r}_{y,k})\ln(\overline{\rho}(\boldsymbol{r}_{y,k})) \\
& - k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)V^2\sum_k \left[\sum_{n\neq k}^{K}\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})}{\overline{\rho}(\boldsymbol{r}_{x,k})\overline{\rho}(\boldsymbol{r}_{y,n})}\right)\right] \\
& - \frac{1}{2}k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_x)V^2\sum_k \left[\sum_{n\neq k}^{K}\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{x,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{x,n})}{\overline{\rho}(\boldsymbol{r}_{x,k})\overline{\rho}(\boldsymbol{r}_{x,n})}\right)\right] \\
& - \frac{1}{2}k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_y,\boldsymbol{r}_y)V^2\sum_k \left[\sum_{n\neq k}^{K}\overline{\rho}(\boldsymbol{r}_{y,k},\boldsymbol{r}_{y,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{y,k},\boldsymbol{r}_{y,n})}{\overline{\rho}(\boldsymbol{r}_{y,k})\overline{\rho}(\boldsymbol{r}_{y,n})}\right)\right] \\
& + O(3) + \ldots
\end{aligned}
\tag{13}
$$

Finally, casting eq 13 in terms of quantities computed in simulation leads to

$$
\begin{aligned}
\Delta S_{\mathrm{solv}} \approx & -k_{\mathrm{B}}\rho^{ox}_{\mathrm{bulk}}V\sum_k \left(\frac{N^{ox}_k}{\rho^{ox}_{\mathrm{bulk}}VN_f}\right)\ln\left(\frac{N^{ox}_k}{\rho^{ox}_{\mathrm{bulk}}VN_f}\right) \\
& - k_{\mathrm{B}}\rho^{hy}_{\mathrm{bulk}}V\sum_k \left(\frac{N^{hy}_k}{\rho^{hy}_{\mathrm{bulk}}VN_f}\right)\ln\left(\frac{N^{hy}_k}{\rho^{hy}_{\mathrm{bulk}}VN_f}\right) \\
& - k_{\mathrm{B}}\rho^{ox}_{\mathrm{bulk}}\rho^{hy}_{\mathrm{bulk}}V^2\sum_k\sum_{n\neq k}\left[\frac{N^{hy}_{n,k_{ox}}}{\rho^{ox}_{\mathrm{bulk}}\rho^{hy}_{\mathrm{bulk}}V^2N_f}{}^{*}\ln\left(\frac{N^{hy}_{n,k_{ox}}}{\rho^{ox}_{\mathrm{bulk}}\rho^{hy}_{\mathrm{bulk}}V^2N_f}*\frac{\rho^{hy}_{\mathrm{bulk}}VN_f}{N^{hy}_n}\right.\right. \\
& \left.*\frac{\rho^{ox}_{\mathrm{bulk}}VN_f}{N^{ox}_k} - \frac{1}{2}k_{\mathrm{B}}(\rho^{ox}_{\mathrm{bulk}}V)^2\sum_k\sum_{n\neq k}\left[\frac{N^{ox}_{n,k_{ox}}}{(\rho^{ox}_{\mathrm{bulk}}V)^2N_f}\ln\left(\frac{N^{ox}_{n,k_{ox}}}{(\rho^{ox}_{\mathrm{bulk}}V)^2N_f}\right.\right. \\
& \left.*\frac{\rho^{ox}_{\mathrm{bulk}}VN_f}{N^{ox}_n}*\frac{\rho^{ox}_{\mathrm{bulk}}VN_f}{N^{ox}_k}\right) - \frac{1}{2}k_{\mathrm{B}}(\rho^{hy}_{\mathrm{bulk}}V)^2\sum_k\sum_{n\neq k}\left[\frac{N^{hy}_{n,k_{hy}}}{(\rho^{hy}_{\mathrm{bulk}}V)^2N_f}\right. \\
& \left.{}^{*}\ln\left(\frac{N^{hy}_{n,k_{hy}}}{(\rho^{hy}_{\mathrm{bulk}}V)^2N_f}*\frac{\rho^{hy}_{\mathrm{bulk}}VN_f}{N^{hy}_n}*\frac{\rho^{hy}_{\mathrm{bulk}}VN_f}{N^{hy}_k}\right)\right] + O(3) + \ldots
\end{aligned}
\tag{14}
$$

where N_f is the total number of frames in the simulation, $N_k^{ox(hy)}$ is the number of frames with an O or H in voxel *k*, and $N^{ox(hy)}_{n,k_{ox(hy)}}$ is the number of frames for which voxel *n* is occupied by an O or H given that an O or H is found in voxel *k*.

Note that second-plus order contributions depend on the joint probability between particles (i.e., the probability of finding these particles at specific positions at the same time/frame); therefore, the per-voxel evaluation of $N^{ox(hy)}_{n,k_{ox(hy)}}$ in eq 14 requires a frame-by-frame analysis. It is also important to keep in mind that the space should be divided into voxels whose volume is sufficiently small to harbor only one particle per simulation frame (i.e., side ≤0.55 Å, based on a 0.96 Å covalent OH bond). This ensures that the mutual information is zero within each voxel, given that there are no intravoxel correlations.

Solvation Enthalpy—The total enthalpy H of a solvated system can be expressed in terms of the pairwise contributions arising from solute–solute, water–water, and solute–water nonbonded interactions:

$$H = H_{\mathrm{sol-sol}} + H_{\mathrm{wat-wat}} + H_{\mathrm{sol-wat}} \quad (15)$$

Defining the enthalpy difference with respect to a reference state comprised of pure (bulk) water and each of the solutes isolated in vacuum leads to

$$\Delta H = (H^{\mathrm{solvated}}_{\mathrm{sol-sol}} - H^{\mathrm{vacuum}}_{\mathrm{sol-sol}}) + (H^{\mathrm{solvated}}_{\mathrm{wat-wat}} - H^{\mathrm{bulk}}_{\mathrm{wat-wat}}) + H_{\mathrm{sol-wat}} \quad (16)$$

In this work, we focus on the estimation of water–water enthalpy differences (i.e., second parentheses in eq 16, henceforth called ΔH_{solv}). Yet, solute–solute and solute–water enthalpies can in principle be estimated in a way analogous to that presented here for the water–water contributions.

Specifically, given that the dominant contribution to ΔH_{solv} comes from short-range nonbonded interactions between oxygen and hydrogen atoms, we can approximate this quantity as the hydrogen bonding (henceforth, H-bonding) energy of the system. For this purpose, we capitalize on the grid-based nature of the methodology introduced to compute ΔH_{solv} as the deviation from bulk water, of the sum of the per-voxel ensemble-averaged H-bonding energy. We compute this energy for every frame in which a water oxygen is in voxel k (k_{ox}) and an H-bonding partner is within a spherical ring volume ($V^{k_{ox}}_{H\mathrm{bound}}$) of 1.4–2.4 Å around the exact position of that water oxygen. Restated in mathematical terms

$$\Delta H_{\mathrm{solv}} \approx \sum_{k} \left\langle \sum_{hy \in V^{k_{ox}}_{H\mathrm{bond}}} f(r) \right\rangle_{\mathrm{frames}} - H^{\mathrm{bulk}}_{\mathrm{solv}} \quad (17)$$

$H^{\mathrm{bulk}}_{\mathrm{solv}}$ is the reference value, corresponding to the total enthalpy of a bulk water simulation at the same temperature and density, and can be computed by setting $\Delta H_{solv} = 0$ in eq 17. $f(r)$ is an energy function that is dependent upon the distance r between each donor–acceptor pair, and whose parameters were obtained by fitting to water–water interaction energies computed from quantum mechanical (QM) calculations at varying water–water distances.

In brief, a two-water-molecule potential curve was computed in gas phase via the counter-poise corrected MP2/6–311+ +G(d,p) formalism. We restrained the distance between one of the water's O atom and an H atom on the other water, allowing the waters to adopt the lowest energy conformation at each of the fixed distances studied. Nine pairwise potentials were obtained and collapsed into a single potential curve that reached 0 kcal/mol at a separation of 13.95 Å. This curve was then fit to a standard Morse potential

$$f(r) = \left[D_e \left(1 - e^{(-a[r - r_e])}\right)^2 \right] - C_w \quad (18)$$

where C_w = 5.7642, D_e = 5.3441, a = 1.4952, and r_e = 1.9070. Note that r_e corresponds to the average O–H H-bonding distance for bulk water, which changes depending on the water

model. A very good fit to the QM calculations was achieved (R^2 = 0.9995) for the distance range of interest.

This simple potential exploits the cancellation of errors due to the overestimation of interactions at short-range and the neglect of interactions longer than the H-bond cutoff, to provide a fast, yet accurate, estimate of the solvation enthalpy (see Results).

Considerations toward Practical Implementation of the Proposed Expressions

Solvation Entropy - Approximations—The evaluation of all contributions to the solvation entropy is currently a very challenging computational task. This complexity calls for the use of practical implementations that can still maintain accuracy. Accordingly, we identify approximate solutions by neglecting the third-and-higher order terms in eq 8. The adequacy of this approximation is supported by related work,[6] which within the context of entropy calculations concludes that third-and-higher order terms appear to account only for ~10% of the thermodynamic properties of water.

Furthermore, calculation of the discretized pairwise MI contributions (see eq 12) can readily become intractable with the current computational capabilities. Therefore, the summation on the right-hand side of eq 12 (as well as those corresponding to the remaining I_2 terms) is limited here to the vicinity of each voxel

$$\begin{aligned} k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)V^2\sum_{k=1}^{K}\left[\sum_{n\neq k}^{K}\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})}{\overline{\rho}(\boldsymbol{r}_{x,k})\overline{\rho}(\boldsymbol{r}_{y,n})}\right)\right] \\ \approx -k_{\mathrm{B}}\rho^{0}_{\mathrm{bulk}}(\boldsymbol{r}_x,\boldsymbol{r}_y)V^2\sum_{k=1}^{K}\left[\sum_{n\in W}\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})^{*}\ln\left(\frac{\overline{\rho}(\boldsymbol{r}_{x,k},\boldsymbol{r}_{y,n})}{\overline{\rho}(\boldsymbol{r}_{x,k})\overline{\rho}(\boldsymbol{r}_{y,n})}\right)\right] \end{aligned} \tag{19}$$

where W is a voxel subspace of the entire space comprised by K voxels. The optimal value of W is that which provides a satisfactory approximation to the exact per-voxel entropy at a reasonable computational cost. This is contingent upon the rate at which the correlations decay as W is extended, which can differ for each entropy contribution.

Accordingly, eq 14 is reduced to

$$\begin{aligned}\Delta S_{\text{solv}} \approx & -k_{\text{B}}\rho_{\text{bulk}}^{ox}V\sum_{k}\left(\frac{N_k^{ox}}{\rho_{\text{bulk}}^{ox}VN_f}\right)\ln\left(\frac{N_k^{ox}}{\rho_{\text{bulk}}^{ox}VN_f}\right) \\ & -k_{\text{B}}\rho_{\text{bulk}}^{hy}V\sum_{k}\left(\frac{N_k^{hy}}{\rho_{\text{bulk}}^{hy}VN_f}\right)\ln\left(\frac{N_k^{hy}}{\rho_{\text{bulk}}^{hy}VN_f}\right) \\ & -k_{\text{B}}\rho_{\text{bulk}}^{ox}\rho_{\text{bulk}}^{hy}V^2\sum_{k}\sum_{n\in W}\left[\frac{N_{n,k_{ox}}^{hy}}{\rho_{\text{bulk}}^{ox}\rho_{\text{bulk}}^{hy}V^2N_f}\ln\left(\frac{N_{n,k_{ox}}^{hy}}{\rho_{\text{bulk}}^{ox}\rho_{\text{bulk}}^{hy}V^2N_f}*\frac{\rho_{\text{bulk}}^{hy}VN_f}{N_n^{hy}}\right.\right. \\ & \left.*\frac{\rho_{\text{bulk}}^{ox}VN_f}{N_k^{ox}}\right)-\frac{1}{2}k_{\text{B}}(\rho_{\text{bulk}}^{ox}V)^2\sum_{k}\sum_{n\in W}\left[\frac{N_{n,k_{ox}}^{ox}}{(\rho_{\text{bulk}}^{ox}V)^2N_f}\right. \\ & \left.\left.*\ln\left(\frac{N_{n,k_{ox}}^{ox}}{(\rho_{\text{bulk}}^{ox}V)^2N_f}*\frac{\rho_{\text{bulk}}^{ox}VN_f}{N_n^{ox}}*\frac{\rho_{\text{bulk}}^{ox}VN_f}{N_k^{ox}}\right)\right] \\ & -\frac{1}{2}k_{\text{B}}(\rho_{\text{bulk}}^{hy}V)^2\sum_{k}\sum_{n\in W}\left[\frac{N_{n,k_{hy}}^{hy}}{(\rho_{\text{bulk}}^{hy}V)^2N_f}\ln\left(\frac{N_{n,k_{hy}}^{hy}}{(\rho_{\text{bulk}}^{hy}V)^2N_f}*\frac{\rho_{\text{bulk}}^{hy}VN_f}{N_n^{hy}}\right.\right. \\ & \left.\left.*\frac{\rho_{\text{bulk}}^{hy}VN_f}{N_k^{hy}}\right)\right]\end{aligned} \tag{20}$$

Pairwise MI calculations are limited here to a subspace W comprised of n = 15,624 voxels, namely, those that are fully or partially encompassed by a sphere of 6 Å radius around center voxel k (i.e., r_{dist} = 6 Å). The adequacy of this cutoff (after which pairwise correlations are assumed to be negligible) is supported by the marked dampening observed for radial distribution functions of pure water (herein RDFs), gOO, gOH, and gHH, at such distance from the center particle. Figure 1 shows plots of these RDFs, resolved experimentally[29] at 298 K; the corresponding RDFs from simulations at 300 K, carried out using the same water model employed in this work, have been reported by Huggins[30] in Figures 3–5 of the cited publication. Ensuing studies involving solutes might require adjustment of this value.

Reduction in the Uncertainty of Calculations—The noise introduced as a result of deficient sampling leads to fluctuations/deviations from the actual thermodynamic values for bulk and nonbulk solvation. The concomitant uncertainty can both hamper the identification of hot spots and swamp the overall values obtained upon summation over the millions of small voxels comprising the entire grid for a solvated protein system. The latter effect is related to the well-known challenge of accurately computing small differences between large numbers (e.g., when carrying out free energy calculations of biomolecular systems) and encourages the use of approaches to increase the signal-to-noise ratio of our calculations.

We have developed a “filtering” scheme in which we isolate the signal from higher-than-bulk and lower-than-bulk solvation by zeroing out bulk-like entropy and enthalpy contributions. Two alternatives have been evaluated toward this end. In the first, we

1. Carry out a reference unrestrained bulk water simulation.
2. Calculate per-voxel solvation entropies and enthalpies across the whole grid, via eqs 20 and 17, respectively.

3. Fit a Gaussian distribution to the solvation entropy data and another one to the solvation enthalpy data.

4. For each case, identify a minimum cutoff of the standard deviation (expressed as the normalized quantity over the absolute standard deviation; i.e., $\overline{\sigma}_{\mathrm{cut}}^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}}=(\sigma_{\mathrm{cut}}^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}})/(\sigma^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}})$ and $\overline{\sigma}_{\mathrm{cut}}^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}}=(\sigma_{\mathrm{cut}}^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}})/(\sigma^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}})$ for which the summation of the values that lie outside of such cutoff is effectively zero ($<10^{-2}$ kcal/mol).

5. Perform a simulation of interest under the same conditions as the reference bulk water run.

6. Calculate the per-voxel solvation entropies and enthalpies of this new simulation, via eqs 20 and 17.

7. Use the bulk-water cutoffs obtained on step 4 to filter the solvation entropy and enthalpy data computed on step 6, by neglecting the contribution of those voxels whose ΔS_{solv} and ΔH_{solv} fall within the range delimited by $\overline{\sigma}_{\mathrm{cut}}^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}}$ and $\overline{\sigma}_{\mathrm{cut}}^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}}$, respectively.

As a second more efficient alternative, we extract $\overline{\sigma}_{\mathrm{cut}}^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}}$ and $\overline{\sigma}_{\mathrm{cut}}^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}}$ directly from the simulation of interest and use these cutoffs to zero out the bulk-like solvation of that same simulation. For this scheme, the box size should be sufficiently large so that the majority of the waters have bulk-like thermodynamic properties (ensuring that the dominant peak for each term, to which a Gaussian distribution is fit, corresponds to bulk). Indeed, the latter requirement is naturally fulfilled when considering a volume of voxel subspace ($\mathcal{W}$) such that second order entropic correlations are captured with satisfactory accuracy.

Both filtering schemes were compared via preliminary analyses of enthalpies and first and second order entropies obtained from 100 ns MD simulations (results not shown). The latter led to excellent agreement between both alternatives, readily suggesting the use of the second one as the standard choice for all ensuing calculations. The previous analysis also led to the selection of $\overline{\sigma}_{\mathrm{cut}}^{\Delta H_{\mathrm{solv}}^{\mathrm{bulk}}}=6$ and $\overline{\sigma}_{\mathrm{cut}}^{\Delta S_{\mathrm{solv}}^{\mathrm{bulk}}}=5$ as suitable values for the present study.

It is important to mention that filtering should be carried out separately for each entropy term (e.g., first order O entropy, second order O–H entropy, and so forth), prior to adding these quantities to obtain a value of the total entropy. This is because straight summation of the “unfiltered” terms, each of which contains a considerable amount of bulk-like noise, will lead to loss of the nonbulk-like signal contributed by each individual term, resulting in an inaccurate value of total entropy.

Altogether, the practice described above filters out the contribution from voxels where the signal is indistinguishable from the noise (i.e., singles-out those voxels that are statistically significant relative to bulk solvent), allowing not only for easier visualization of the effectively relevant solvent regions within the grid but also improving the accuracy of the global ΔS_{solv} and ΔH_{solv} estimates (i.e., those obtained by summation of all pervoxel

quantities). The effect of filtering thermodynamics distributions is discussed in the Results section.

Computational Details—Systems of 10,012 TIP3P water molecules with 0, 1, or 2 water oxygens restrained by means of a stiff force constant k = 100 kcal/mol/Å^2 (for all cases) were arranged into a cubic periodic box, minimized, and equilibrated at 300 K for 500 ps. These were used as starting points for production MD simulations at 300 K within the NPT ensemble, which were run for either 50 or 100 ns. All MD simulations were performed using the Amber12[31,32] software package, with a frame rate of 0.1 ps and a time step of 2 fs. Constant temperature and pressure were, respectively, maintained by means of the Berendsen thermostat and the Berendsen barostat with isotropic position scaling. The time constant for heat bath coupling and the pressure relaxation time were both set to 2 ps.

The resulting MD trajectories were used to compute solvation entropies and enthalpies via eqs 20 and 17, respectively. For this purpose, we developed algorithms for efficient extraction of the required probability distributions, which we integrated into an in-house software package called WATMD. This package is customarily employed in our lab to characterize solute–solvent dynamics and has proven useful for ligand design and elucidation of mechanistic insights of biophysical processes. Details on WATMD can be found in ref 1.

Our thermodynamic calculations were carried out on MD trajectories representing a single image of the periodic solvent box. Hence, those voxels located near the edges of the box exhibit occupancies (both marginal and joint) that are distorted by the absence of contiguous images. Accordingly, to ensure proper calculation of the quantities of interest, we focus our analysis on a subvolume of the solvent box, corresponding in this work to those voxels at least 14 Å away from any box edge; this amounts to 2,570 waters. Complementary analyses (data not shown) confirmed that the filtering procedure leads to estimates that are highly robust to changes in the size of this subvolume, granted that the cutoff is larger than 10 Å from any box edge. MATLAB[33] was used to isolate and filter the subvolume, yielding the results reported in the following section.

RESULTS

The current work focuses on the estimation of ΔS_{solv} from water-only simulations.

To assess the accuracy of the approach presented above, we initially calculate overall values of the excess enthalpy and entropy of pure water and obtain an estimate of the solvation free energy of a water molecule in bulk. This is achieved by means of sets of simulations with one stiffly restrained water oxygen or without restraints.

Subsequently, we study two systems (each with two restrained water oxygens) whose solvation fields are either correlated or uncorrelated, to highlight the wealth of information contained in their detailed solvation structures, which for the general case can be critical to improve the efficacy of drug design.

Excess Enthalpy and Entropy of Pure Water

An excess property represents the difference between the value of a thermodynamic property M of a real solution at a specified T, P, and composition and that of an ideal solution under analogous conditions:

$$M^E = M - M^{\text{ideal}} \quad (21)$$

The excess entropy, S^E, is equivalent to

$$S^E = S - S^{\text{ideal}} \approx S^{(2)} \quad (22)$$

We approximate S^E as the second-order truncation of the solvation entropy (i.e., $D = 2$ in eq 8).

Theoretical estimates of S^E for pure water have previously been obtained by integration of the RDFs.[6,30,34] In these analyses, the change in entropy relative to the ideal molecular fluid is computed based on the order established upon fixing the position of a single water molecule of the system. Accordingly, in this work we compute the per-water S^E as the filtered value (see Methods) of the solvation entropy extracted from pure water simulations in which a single water oxygen has been restrained via a stiff harmonic potential. Specifically, a 100 ns run was used to obtain the statistics up to 1 million frames (0.1 ps frame rate), and 20 50 ns simulations were used to gather data up to 10 million frames. The use of multiple shorter (in this case 50 ns) simulations, each of which is assigned a different seed for the initial velocities at the production stage, can lead to better sampling[1] and improved computational efficiency. In the present case, the value of S^E at 1 million frames extracted from the 100 ns simulation coincides with that obtained from two 50 ns simulations, suggesting that sampling is comparable in both instances (such that the latter strategy primarily aids computational efficiency).

Our calculations are referenced to a uniform distribution of uncorrelated oxygens and hydrogens, which we believe provides a more comprehensive standpoint relative to other methods, given that it inherently accounts for intramolecular entropy (see Discussion). Yet, direct comparison between our estimate of S^E and available experimental values[35,36] requires the exclusion of the entropic cost of forming all water molecules, which corresponds to the difference between our higher entropy reference state and an ideal fluid of uniformly distributed water molecules. We account for this difference by neglecting the following: 1) the first order H contribution of those voxels that are one or two voxels away from the voxel containing the restrained particle and 2) the O–H and H–H MI terms of those voxels that are at most two or three voxels away from the central voxel, respectively. Each voxel is either included or excluded in the calculation, leading to a cubic exclusion volume around the central voxel. This procedure effectively excludes the translational entropic cost of forming the two O–H bonds and the orientational entropic cost arising from the H–O–H angle; i.e., the first peak of the O–H and the H–H RDFs (see Figure 1). Note that the number of voxels to be disregarded depends on the voxel size employed (in this case, $[0.5\ \text{Å}]^3$) and involves an added approximation to our calculation. Appendix C, Supporting Information

presents an analysis from which we obtained a rough estimate of the error in S^E resulting from ignoring the aforementioned entropic contributions of neighboring voxels; we estimate the error to be around 0.3 kcal/mol.

In addition, comparison with the experimental S^E requires that the entropic cost attributable to fixing the reference particle is also omitted from the calculation. For a stiff restraint, the latter corresponds predominantly to the marginal entropy of a very small number of voxels that are visited by the restrained atom.

Figure 2 displays convergence plots of the marginal entropies and mutual information terms, as well as the sum of all such quantities to obtain an estimate of the excess entropy (TS^E). The result at 10 million frames, −3.76 kcal/mol, is in good agreement with the corresponding experimental value of −4.22 kcal/mol (i.e., −14.05 cal/mol/K) at 300 K.[35] This outcome supports the validity of the proposed formulations for estimation of ΔS_{solv}. Note also that all of the pairwise MI terms are non-negative after sufficient sampling, as required by a corollary of Mutual Information Theory. We highlight that our calculations correspond to the solvation entropy of the entire system in response to fixing a water oxygen, which acts as a solute immersed in pure water. Thus, the uncertainty in this quantity upon solvating other solutes is anticipated to be similar to the one reported above.

The entropic contributions to S^E (i.e., S_{ox}, S_{hy}, $I_{(ox,hy)}$, $I_{(ox,ox)}$, and $I_{(hy,hy)}$) converge at a varying pace. Figure 2 shows that by 2 million frames the faster converging first order entropies and the H–H MI term have stabilized within 90% of the value at 10 million frames. In contrast, the O–H and O–O MI terms have reached 75% and 81% of the corresponding values at 10 million frames. Potential ways to improve convergence are discussed later in this work. Yet, we highlight here the benefit of the filtering approach relative to an "unfiltered" case in which S^E is simply calculated as the difference with respect to a bulk water simulation, *i.e.* $S^E \approx \Delta S_{\mathrm{solv}} - \Delta S_{\mathrm{solv}}^{\mathrm{bulk}}$. Figure 3 shows that this choice leads to a value of TS^E that by 2 million frames is already 3.7 kcal/mol lower than the experimental value, and whose O–O MI term is still negative. As expected, the effect of filtering is less dramatic for faster-converging quantities, such as marginal entropies.

The excess enthalpy, H^E, is in turn equal to

$$H^E = H - H^{\mathrm{ideal}} \quad (23)$$

The enthalpy of the real solution (H) is equivalent to H^E since all energetic interactions are zero in an ideal solution. Thus, for pure water under physiological conditions, H^E should correspond to the value extracted from bulk water simulations at 300 K. We initially computed this quantity via brute force (i.e., no filtering) estimation of $H_{\mathrm{solv}}^{\mathrm{bulk}}$ via eq 17. We then normalized the result by the amount of waters in the subvolume analyzed to enable comparison with the experimental per-water value reported by Wagner and Pruβ.[35] Figure 4 shows fast and accurate convergence of our H^E estimate to the target value of −10.51 kcal/mol;[35] the equilibrium value from simulation is −10.54 kcal/mol.

Alternatively, a more stringent calculation of H^E can be performed to test the ability of our approach to filter out bulk-like enthalpy contributions. To accomplish this, we analyzed the trajectories with one fixed water oxygen that were used for calculation of S^E. The filtered enthalpy in this case should still be equal to H^E, since bulk-like enthalpy (and thus zero contribution to H^E) is expected for all but the translationally fixed water. We note that fixing a water oxygen is equivalent to aligning all frames with respect to that particle, a practice that clearly does not alter the per-frame enthalpy of the system (i.e., the restraining work is solely entropic in nature). Figure 5 illustrates the convergence of H^E for this scenario, which by 10 million frames leads to $H^E = -10.78$ kcal/mol, in good agreement with experiment. Moreover, the error in our calculations maps directly to the noise contributed by bulk-like voxels (see inset of Figure 5), which is zero by 10 million frames. Note that Figure 5 shows much slower convergence than Figure 4 because the former portrays the time average of a single molecule whereas the latter represents a time-dependent ensemble average over all waters in the system (the error contributed by each water cancels out upon averaging). Overall, the sizably longer convergence time observed for this case relative to our first estimate (see Figure 4) is a good indicator of the cost of converging global thermodynamic quantities to high accuracy via the current histogramming approach.

The excess free energy (G^E) computed from our filtered one-oxygen restraint simulations is $G^E = H^E - TS^E = -10.78 + 3.76 = -7.02$ kcal/mol at 10 million frames, whereas the corresponding experimental value is −6.33 kcal/mol.[35]

Thermodynamic Properties of Two-Oxygen Restrained Water Systems

We evaluated the ability of the proposed methodology to enable solvation thermodynamics interpretations at the local level (e.g., within a binding site), while still providing consistent estimates of those quantities for the total system. To this end, two pure water systems were studied, each with two stiffly restrained water oxygens ($k = 100$ kcal/mol/Å^2), which still allows for free rotation of these molecules. The restrained particles are either ~10.8 Å (case 1) or ~4.3 Å (case 2) apart from each other. For the first case, the constrained particles are sufficiently far that their respective solvation fields are effectively independent, whereas for the second case both solvation fields would be highly correlated. Each system was run for 100 ns (1 million frames), an extent that is considered suitable for portraying relevant features of the solvation structures at the local level. The resulting trajectories were postprocessed to obtain the corresponding thermodynamic values.

Concerning case 1, the fixed particles should produce decorrelated nonbulk solvation fields; therefore, the global thermodynamic properties must correspond to two times the excess quantities computed for the same number of frames. Indeed, for 1 million frames we found that $T\Delta S_{solv} = 2.07 * S^E = -5.32$ kcal/mol, $\Delta H_{solv} = 2.02 * H^E = -26.18$ kcal/mol, and $\Delta G_{solv} = -20.86$ kcal/mol (note that the excess quantities at 1 million frames, used here for comparison, differ from the converged values; see Figures 2 and 5). The fact that both fields are decoupled is verified by visualization of the nonbulk pervoxel entropy and enthalpy of the system. Figure 6 also shows that the enthalpic contributions (ΔH_{solv}), which are favorable relative to bulk, stem primordially from the direct interaction between those voxels encompassing the restrained oxygens and the first solvation shell (i.e., large translucent

green spheres). Conversely, the entropy is distributed radially away from each of the fixed particles.

Visualization of the different entropic contributions offers relevant details on how the solvation structures are shaped by interatomic correlations and facilitates mapping of these structures to experimental RDFs for pure water under the same conditions (see Figure 1). Figure 7 illustrates the solvation entropy field produced by one of the restrained water oxygens, dissected into marginal (Figures 7A,B), pairwise MI (Figures 7C–E) and total (Figure 7F) contributions. The higher-than-bulk occupancy structure for nonbulk entropy voxels is also included in Figure 7G, as a reference for the entropy structures. Figure 7A (S_{ox}) displays an isotropic ring of unfavorable (i.e., red color) solvation entropy associated with the first peak of the O–O RDF. In turn, S_{hy} (Figure 7B) is characterized by an inner region of favorable solvation conforming to the area between the first minimum and the third barrier of the O–H RDF (structure that is consistent with the absence of higher-than-bulk occupancy in this region; see Figure 7G), and an outermost shell comprised of mainly unfavorable entropy corresponding to the third barrier of the O–H RDF. The $I_{(ox,hy)}$ term (Figure 7C) is comprised of two rings, the innermost resulting from favorable lower-than-bulk correlations (i.e., $g < 1$ in the O–H RDF; see Figure 1), and the outermost due to entropically unfavorable H-bonding between surrounding Hs/Os and the O/Hs of the restrained water (the innermost dark blue spheres in Figure 7G represent hydrogen occupancy of the restrained water). Similarly, the first shell of the $I_{(ox,ox)}$ term (Figure 7D) accounts for favorable correlations in the lower-than-bulk occupancy region before the first peak of the O–O RDF, whereas the second shell further increases (i.e., adds to S_{ox}) the entropic cost at the first peak of the O–O RDF. Lastly, the $I_{(hy,hy)}$ structure (Figure 7E) is characterized by an inner region of marked entropic cost and an outer shell, both representing unfavorable intermolecular correlations between the hydrogens of the translationally restrained water and those of surrounding water molecules (2nd peak of H–H RDF). Figure 7F depicts the overall structure resulting from the per-voxel sum of all entropic contributions. An inner region of unfavorable entropy is apparent, which is enclosed by two lower-magnitude shells of increasing radius, contributing favorable and unfavorable entropy, respectively.

The second case, in which two water oxygens are restrained at ~4.3 Å from each other, allowed us to study how global estimates and local structures change when both solvation fields interact. We remark that certain specifics of these results may vary with more thorough sampling; yet, our analysis at 1 million frames is considered suitable for describing overarching differences between the two cases inspected. We find that the overall entropy and enthalpy of the system are lower than the previous case. This denotes an enthalpy gain that is offset by a stronger correlation between the fields: $T\Delta S_{solv} = -6.03$ kcal/mol and $\Delta H_{solv} = -27.67$ kcal/mol. Hence, some compensation is observed, leading to $\Delta G_{solv} = -21.64$ kcal/mol, which is only $\Delta\Delta G_{solv} = -0.78$ kcal/mol more favorable than the fully decorrelated system (see case 1, where $\Delta G_{solv} = -20.86$ kcal/mol). We now examine the per-voxel structures of the total entropy (Figure 8A), enthalpy (Figure 8B), and free energy (Figure 8C). These are referenced to the higher-than-bulk occupancy of those voxels with non-negligible free energy (Figure 8D).

Figure 8A illustrates two symmetric colliding entropy fields with a vertical symmetry interface characterized by voxels with overall unfavorable entropy of ~1.37 kcal/mol. Conversely, Figure 8B shows that this interface affords an important enthalpy gain, estimated at around −3.42 kcal/mol. This is consistent with the high oxygen occupancy in that region (see Figure 8D), resulting from persistent H-bonding between the hydrogens of the restrained waters and the oxygens of neighboring molecules. The net result is a favorable free energy interface (see Figure 8C).

A second interesting feature arises along the plane bisecting the solvation fields; namely, the existence of two peripheral concentric rings with net favorable free energy (two outermost rings in Figure 9C), the smaller one resulting from favorable enthalpy (outermost ring in Figure 9B) and unfavorable entropy (second largest ring in Figure 9A), and the larger one solely from favorable entropy (outermost ring in Figure 9A). These can also be discerned in Figures 8A–C as clusters of spheres directly above and below the vertical interface discussed in the previous paragraph. These rings represent the propagation of nonbulk correlations/interactions into the second solvation shell, effects that are nearly absent when the fields are decorrelated (e.g., case 1). Note that Figures 8 and 9 display the logarithm of each quantity; thus, it can be inferred from these figures that the two second-shell rings provide a markedly lower contribution to the overall free energy, relative to the enthalpy of the fixed particles or the enthalpy of the direct interface between the fields. Nevertheless, in the general case, second and higher-shell effects that disseminate to large portions of space may have a sizable impact on ligand design and should therefore be carefully scrutinized.

As a final observation, the phenomenon of entropy-enthalpy compensation is suggested in Figure 9 by the palpable superposition of shells of favorable enthalpy and unfavorable entropy, which offset each other with the concomitant effect on the free energy.

DISCUSSION

We have introduced a methodology for calculating solvation entropies, enthalpies, and free energies, which is rooted in the study of water oxygen and hydrogen correlations/interactions. Our results suggest that the discretized expressions derived from our postulates are suitable for estimation of these thermodynamic quantities, both globally and around local regions of interest. Furthermore, for the general case, the viewpoint introduced offers particular advantages over related grid-based or site-based techniques. These include the following:

1. A more tractable algorithm for calculation of second and higher-order solvation entropies, relative to those based on the analysis of molecular positions and orientations.[2–8,16] In particular, potential inaccuracies resulting from the summation of entropies that are computed via two distinct distance metrics (e.g., positional and orientational) are avoided. Moreover, we anticipate that convergence of ΔS_{solv} for large systems will be faster when circumventing the highly complex process of computing accurate probability densities from relative orientations between waters. In general, the derived thermodynamics expressions can also be employed to

postprocess water oxygen–hydrogen density data obtained from other techniques (e.g., 3D-RISM[16]).

2. The entropy expressions derived in this work are not founded on (but are naturally related to) IST; more explicitly, our derivation is not conceptually built around the use of radial distribution functions. Thus, future attempts to extend our theory to the estimation of solute–solvent free energies (ongoing work) would likely be independent of the fixed-solute requirement inherent in IST. In general, allowing solute motions should lead to a more representative solute–solvent structural ensemble relative to the use of fixed solute. In addition, flexible-solute simulations can expedite recognition of incorrect solvent structures (arising, for instance, from inappropriate charge assignments) given that the resulting ensemble-averaged structures should overlay satisfactorily onto the corresponding high-resolution crystals, when available. More laborious fixed-solute schemes in which various representative states are sampled in parallel might be an appealing alternative to flexible-solute analyses. Nevertheless, these involve costly preliminary steps including initial sampling runs, clustering toward identification of a representative structure for each state, and proper weighting of the probability density of each state. It may be argued that a single flexible-solute simulation affords deficient sampling of solvent states relative to multiple fixed-solute analyses of diverse conformations. To overcome this potential limitation, our standard protocol includes running multiple parallel simulations with perturbed initial velocities, as implemented in this work and elsewhere.[1]

3. The ability to capture relative differences in intramolecular entropy resulting from molecular vibrations, which are directly linked to the first peak of the O–H and H–H RDFs. This is only relevant when using flexible-water models (which should lead to more detailed results due to their ability to capture intramolecular motions), given that intramolecular contributions cancel out when using rigid-water models to compute $\Delta\Delta S_{solv}$. Note that the use of our approach to compute $\Delta\Delta G_{solv}$ from flexible-water simulations would require both retaining the unfavorable entropic cost of covalently forming all water molecules and including all favorable covalent bonding energies (formation of a water molecule is an exothermic process), which are not encompassed by eq 17.

The value of the excess entropy computed for pure water, resulting from the sum of two marginal and three pairwise terms, exhibits good agreement with experiment. Our results support the notion that third-plus order terms do not contribute significantly to the overall entropy estimate; yet, plausible scenarios for which higher order terms are relevant should not be discounted, and follow-up efforts to characterize these circumstances may be advantageous. Note that the lower entropic cost observed in our calculations is likely due to the fact that for voxel sizes of 0.5 Å the actual probability density of each voxel is still smoothed out. A simple analysis of the experimental RDFs of pure water, in which the one-dimensional data is binned into distances of varying magnitude, suggests that highly uniform probability densities will be achieved only for voxel sizes that are ≤0.1 Å. The use of such a high-resolution grid is, however, not feasible in light of our current computational capabilities.

Concerning the excess enthalpy, it is encouraging to find that a simplified H-bonding distance algorithm leads to values that are in good agreement with experiment. Interestingly, other approaches encompassing the full spectrum of interaction energies[9,30] can potentially lead to higher uncertainty, due to their direct dependence on force field energies and/or due to noise present in the brute-force sum (i.e., nonfiltered) of energetic contributions. Forthcoming studies will be carried out to test the robustness of our approach and to suggest more efficient protocols for the more general case of solute–solvent analysis.

The methodology introduced supports detailed (i.e., per-voxel) visualization of all-occupancy solvation thermodynamics structures, capability that is also accessible via rigid-solute approaches such as GIST.[9,10] Yet, structures obtained by means of oxygen–hydrogen decomposition not only facilitate identification of the directionality of H-bonding networks within and around a protein[1] but can also promote better understanding of interatomic correlations by direct inspection of solvation structures derived from individual contributions.

The present study suggests that neglect of per-voxel bulk-like solvation entropies and enthalpies can significantly improve the convergence and accuracy of the overall estimates. In addition, filtering has favorable implications for visualization toward design, given that only a limited number of thermodynamically relevant voxels needs to be considered. Conversely, we observed that uncertainty can emerge from this procedure if the normalized standard deviation cutoffs are smaller than 4 (in which case some unwanted bulk-like noise will be included in the calculation) or larger than 6 (in which case some nonbulk-like solvation will be excluded from the calculation). Alternate, potentially superior algorithms for effective reduction of the noise in our calculations may be explored in the future. More so, we consider that other solvation analysis techniques may also benefit from this, or similar, practices.

Comparison between our estimate of S^E and the corresponding experimental value entails the exclusion of the entropic cost arising from the presence of intramolecular interactions. In this first implementation of the proposed theory, we carry this out by ignoring the contribution of two or three-voxel shells around the center voxel, depending on the case. The error inherent in this approximation appears to be small, based on the good agreement with experiment and on the magnitude of the uncertainty estimates presented in Appendix C, Supporting Information. Nevertheless, it is possible that this practice leads to appreciable inaccuracies upon using this approach to validate experimental solvation thermodynamic measurements of more complex systems. Therefore, other alternatives that can effectively minimize or even eliminate this uncertainty will be explored in ensuing publications. Possible avenues toward improving accuracy include reducing the voxel size (this requires increasing the efficiency of our calculations), keeping track of the covalently bonded atoms on a per-frame basis in order to more precisely disregard the corresponding contributions, and directly subtracting the intramolecular entropy of the homogeneous fluid from any simulation of interest. The latter option would altogether obviate the need to employ exclusion volumes. Overall, it is important to remark that it is not necessary to omit intramolecular interactions when computing relative entropy or, more generally, free energy differences between two systems, which is the dominant practice in *in silico* drug discovery.

Convergence of thermodynamic quantities proved to be a challenging task. The computational cost of reasonable convergence of the pairwise MI terms (see Figure 2) and thorough noise reduction (see Figure 5) is on the order of millions of frames. This bound is not expected to be much higher when studying more complex cases, because bulk-like voxels are potentially more difficult to equilibrate (i.e., the limiting case) than regions experiencing stronger correlations, granted the absence of long-scale conformational changes. For the systems studied (comprising 10,012 water molecules, which is enough to solvate a medium-sized globular protein) we find that an ~2 million-frame ensemble is necessary to reach satisfactory convergence (e.g., both excess entropy and enthalpy are within 1 kcal/mol of experiment). This process requires approximately 2 days upon intensive use of our existing capabilities (parallel MD simulations on 20 NVIDIA GTX 780 GPUs, followed by fully parallel postprocessing runs on ~100 Intel Xeon E5–2660 @2.20 GHz processors). Ongoing studies in our lab are aimed at identifying the minimum simulation frame rate and run-time needed for optimal parallelization of our calculations. As a straightforward alternative, efficiency could in principle be boosted by increasing voxel size and/or reducing the value of r_{dist}. Regarding the voxel size, a value of 0.5 Å is currently considered appropriate given that it guarantees the presence of only one atom within a single voxel at each frame (therefore ensuring the absence of mutual information inside the voxel itself), while still leading to reasonable postprocessing times and memory requirements. In turn, decreasing r_{dist} might be applicable only on a case-by-case basis; to this end, supporting studies on systems of interest for drug discovery should be performed. More broadly, enhanced data processing schemes enabling improved convergence will also be tested in future work.

FINAL REMARKS

Several advancements have been made in recent years concerning the efficacy of well-established free energy methods,[37] enabling their systematic use as virtual screening tools.[38] Nevertheless, most of these techniques have limited use in rational design because they are restricted to estimation of the system's total free energy, which provides only indirect insights concerning the effects of ligand modifications.

A conceivably more attractive avenue toward efficient *in silico* identification of potent molecules is the exploration of methodologies that afford thermodynamics-based phenomenological understanding of solute and solvent interactions at local regions of interest (e.g., within a binding pocket and between the pocket and ligand substructures). Indeed, alternative methods are available, which appear to be useful for ligand design[10,39–41] despite their focus on analysis of solvation thermodynamics.[9–12,14–17] However, several challenges (discussed in the Introduction section) need to be addressed prior to extending such approaches to account for all contributions to the free energy, particularly concerning improvements in accuracy/efficiency without the use of empirical parametrization (which can hamper the general applicability of these techniques). This work represents a step forward in this direction.

The methodology presented, which was evaluated herein for pure water, can provide both local and global estimates of the solvent free energy using all-occupancy statistics extracted

from flexible-solute simulations (these will be considered in future work). Solvent free energy estimates are derived using an innovative perspective of solvating water wherein interatomic interactions/correlations between oxygens and hydrogens are evaluated, which facilitates calculations relative to analyses built upon translational and orientational correlations between water molecules.[2–10] The excess enthalpy and entropy of pure water is properly reproduced via postprocessing of MD runs with thousands of molecules; this is a fundamental test that lays a solid foundation for the efficacious study of solvated systems relevant to drug discovery. Variations in the local and global enthalpy, entropy, and free energy between water systems with correlated versus decorrelated inhomogeneous solvation structures were also examined. The ability of our approach to isolate the contributions of individual solvating water molecules and networks thereof provides, in the general case, the basis for detailed structural and mechanistic studies essential for understanding flexible protein–ligand binding.

The expressions proposed herein will be adopted as a basis for iterative design practices, in which local solvation thermodynamic quantities are used to guide changes in the solute/ligand (e.g., aimed at displacing free energetically unfavorable or avoiding favorable solvation), whose effect is assessed by means of iterative simulations/analysis, design (based on solute–solvent reorganization), synthesis, and experimental testing. This process, when carried out until a satisfactory solute–solvent structure has been attained, can noticeably improve the chance of identifying promising candidates.

Looking forward, additional work is necessary to further assess the accuracy of the proposed methodology in its current state. This involves exploring improved filtering protocols as well as practices that would allow us to relax some of the assumptions/approximations that are presently in place. In addition, the computational cost required for convergence of thermodynamic quantities via the current histogramming protocol may, to some extent, hinder the systematic use of our methodology for virtual screening. Therefore, research in progress includes exploration of alternate statistical analysis techniques, such as the *k*-Nearest-Neighbors algorithm which has been previously evaluated within the framework of IST-derived approaches.[9,42]

Prospective research also includes the estimation of solvation enthalpies and entropies of small solutes, host–guest complexes, and protein–ligand systems, with the primary goal of partitioning the relative contributions of solvation free energies into the association and dissociation barriers (i.e., to k_{on} and k_{off}). Related efforts have been previously carried out by our group,[13,40,43] suggesting that water–water interactions play a dominant role (as a “free energy reservoir”) during the binding process. More general theoretical inquiries, such as the degree of coupling between solute and solvent entropies and entropy-enthalpy compensation, can also be explored by complementing the proposed methodology with techniques such as alchemical or PMF free energy methods or algorithms for estimation of configurational entropies and energies.

Overall, we believe that this type of structure-free energy analyses can enhance the role of computational chemistry in the pharmaceutical industry, with the potential to more

consistently deliver drug-like ligand designs than those efforts supported by more traditional molecular modeling techniques.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Prof. Mike Gilson, Dr. Viktor Hornak, and Dr. Mitsunori Kato for fruitful discussions leading to the work presented in this document. Tom Kurtzman's contribution was funded, in part, by NIH grant GM095417.

REFERENCES

1. Velez-Vega C, McKay DJJ, Aravamuthan V, Pearlstein R, Duca JS. Time-Averaged Distributions of Solute and Solvent Motions: Exploring Proton Wires of GFP and PfM2DH. J. Chem. Inf. Model. 2014; 54(12):3344–3361. [PubMed: 25405925]
2. Wallace DC. On the Role of Density Fluctuations in the Entropy of a Fluid. J. Chem. Phys. 1987; 87(4):2282–2284.
3. Ashbaugh HS, Paulaitis ME. Entropy of Hydrophobic Hydration: Extension to Hydrophobic Chains. J. Phys. Chem. 1996; 100(5):1900–1913.
4. Lazaridis T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. J. Phys. Chem. B. 1998; 102(18):3531–3541.
5. Morita T, Hiroike K. A New Approach to the Theory of Classical Fluids. III General Treatment of Classical Systems. Prog. Theor. Phys. 1961; 25(4):537–578.
6. Lazaridis T, Karplus M. Orientational Correlations and Entropy in Liquid Water. J. Chem. Phys. 1996; 105(10):4294–4316.
7. Baranyai A, Evans DJ. Direct Entropy Calculation from Computer Simulation of Liquids. Phys. Rev. A: At., Mol., Opt. Phys. 1989; 40(7):3817–3822.
8. Huggins DJ. Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations. Biophys. J. 2015; 108(4):928–936. [PubMed: 25692597]
9. Nguyen CN, Young TK, Gilson MK. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor cucurbit[7]uril. J. Chem. Phys. 2012; 137(4): 044101. [PubMed: 22852591]
10. Nguyen CN, Cruz A, Gilson MK, Kurtzman T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. J. Chem. Theory Comput. 2014; 10(7):2769–2780. [PubMed: 25018673]
11. Li Z, Lazaridis T. Computing the Thermodynamic Contributions of Interfacial Water. Methods Mol. Biol. 2012; 819:393–404. [PubMed: 22183549]
12. Young T, Abel R, Kim B, Berne BJ, Friesner RA. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein–ligand Binding. Proc. Natl. Acad. Sci. U. S. A. 2007; 104(3): 808–813. [PubMed: 17204562]
13. Pearlstein RA, Hu Q-Y, Zhou J, Yowe D, Levell J, Dale B, Kaushik VK, Daniels D, Hanrahan S, Sherman W, Abel R. New Hypotheses about the Structure-Function of Proprotein Convertase Subtilisin/kexin Type 9: Analysis of the Epidermal Growth Factor-like Repeat A Docking Site Using WaterMap. Proteins: Struct., Funct., Genet. 2010; 78(12):2571–2586. [PubMed: 20589640]
14. Huggins DJ. Application of Inhomogeneous Fluid Solvation Theory to Model the Distribution and Thermodynamics of Water Molecules around Biomolecules. Phys. Chem. Chem. Phys. 2012; 14(43):15106–15117. [PubMed: 23037989]
15. Haider K, Huggins DJ. Combining Solvent Thermodynamic Profiles with Functionality Maps of the Hsp90 Binding Site to Predict the Displacement of Water Molecules. J. Chem. Inf. Model. 2013; 53(10):2571–2586. [PubMed: 24070451]

16. Sindhikara DJ, Hirata F. Analysis of Biomolecular Solvation Sites by 3D-RISM Theory. J. Phys. Chem. B. 2013; 117(22):6718–6723. [PubMed: 23675899]
17. Gerogiokas G, Calabro G, Henchman RH, Southey MWY, Law RJ, Michel J. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. J. Chem. Theory Comput. 2014; 10(1):35–48. [PubMed: 26579889]
18. Grant JA, Pickup BT, Nicholls A. A Smooth Permittivity Function for Poisson–Boltzmann Solvation Methods. J. Comput. Chem. 2001; 22(6):608–640.
19. Hu B, Lill MA. WATsite: Hydration Site Prediction Program with PyMOL Interface. J. Comput. Chem. 2014; 35(16):1255–1260. [PubMed: 24752524]
20. Abel R, Young T, Farid R, Berne BJ, Friesner RA. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. J. Am. Chem. Soc. 2008; 130(9):2817–2831. [PubMed: 18266362]
21. Killian BJ, Kravitz JY, Gilson MK. Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation. J. Chem. Phys. 2007; 127(2):024107. [PubMed: 17640119]
22. Zhou H-X, Gilson MK. Theory of Free Energy and Entropy in Noncovalent Binding. Chem. Rev. 2009; 109(9):4092–4107. [PubMed: 19588959]
23. Head MS, Given JA, Gilson MK. Mining Minima": Direct Computation of Conformational Free Energy. J. Phys. Chem. A. 1997; 101(8):1609–1618.
24. Hnizdo V, Darian E, Fedorowicz A, Demchuk E, Li S, Singh H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. J. Comput. Chem. 2007; 28(3):655–668. [PubMed: 17195154]
25. King BM, Silver NW, Tidor B. Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. J. Phys. Chem. B. 2012; 116(9):2891–2904. [PubMed: 22229789]
26. Karplus M, Kushick JN. Method for Estimating the Configurational Entropy of Macromolecules. Macromolecules. 1981; 14(2):325–332.
27. Hnizdo V, Gilson MK. Thermodynamic and Differential Entropy under a Change of Variables. Entropy. 2010; 12(3):578–590. [PubMed: 24436633]
28. Matsuda H. Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration. Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top. 2000; 62(3):3096–3102.
29. Soper AK. The Radial Distribution Functions of Water and Ice from 220 to 673 K and at Pressures up to 400 MPa. Chem. Phys. 2000; 258(2–3):121–137.
30. Huggins DJ. Correlations in Liquid Water for the TIP3P-Ewald, TIP4P-2005, TIP5P-Ewald, and SWM4-NDP Models. J. Chem. Phys. 2012; 136(6):064518. [PubMed: 22360206]
31. Case D, Darden T, Cheatham T, Simmerling C, Wang J, Duke R, Luo R, Walker R, Zhang W, Merz K, Roberts B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong K, Paesani F, Vanicek J, Wolf R, Liu J, Wu X, Brozell S, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh M-J, Cui G, Roe D, Mathews D, Seetin M, Salomon-Ferrer R, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, Kollman P. Amber. 12
32. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 2013; 9(9):3878–3888. [PubMed: 26592383]
33. MATLAB and Statistics Toolbox Release 2007b. Natick, Massachusetts, United States: The MathWorks, Inc.; 2007.
34. Rosenfeld Y. Relation between the Transport Coefficients and the Internal Entropy of Simple Systems. Phys. Rev. A: At., Mol., Opt. Phys. 1977; 15(6):2545–2549.
35. Wagner W, Pruß A. The IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use. J. Phys. Chem. Ref. Data. 2002; 31(2):387–535.
36. Wertz DH. Relationship between the Gas-Phase Entropies of Molecules and Their Entropies of Solvation in Water and 1-Octanol. J. Am. Chem. Soc. 1980; 102(16):5316–5322.
37. Hansen N, van Gunsteren WF. Practical Aspects of Free-Energy Calculations: A Review. J. Chem. Theory Comput. 2014; 10(7):2632–2647. [PubMed: 26586503]

38. Christ CD, Fox T. Accuracy Assessment and Automation of Free Energy Calculations for Drug Design. J. Chem. Inf. Model. 2014; 54(1):108–120. [PubMed: 24256082]

39. Weldon DJ, Shah F, Chittiboyina AG, Sheri A, Chada RR, Gut J, Rosenthal PJ, Shivakumar D, Sherman W, Desai P, Jung J-C, Avery MA. Synthesis, Biological Evaluation, Hydration Site Thermodynamics, and Chemical Reactivity Analysis of α-Keto Substituted Peptidomimetics for the Inhibition of Plasmodium Falciparum. Bioorg. Med. Chem. Lett. 2014; 24(5):1274–1279. [PubMed: 24507921]

40. Pearlstein RA, Sherman W, Abel R. Contributions of Water Transfer Energy to Protein-Ligand Association and Dissociation Barriers: Watermap Analysis of a Series of p38α MAP Kinase Inhibitors. Proteins: Struct., Funct., Genet. 2013; 81(9):1509–1526. [PubMed: 23468227]

41. Mondal J, Friesner RA, Berne BJ. Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. J. Chem. Theory Comput. 2014; 10(12):5696–5705. [PubMed: 25516727]

42. Huggins DJ. Estimating Translational and Orientational Entropies Using the K-Nearest Neighbors Algorithm. J. Chem. Theory Comput. 2014; 10(9):3617–3625. [PubMed: 26588506]

43. Tran Q-T, Williams S, Farid R, Erdemli G, Pearlstein R. The Translocation Kinetics of Antibiotics through Porin OmpC: Insights from Structure-Based Solvation Mapping Using WaterMap. Proteins: Struct., Funct., Genet. 2013; 81(2):291–299. [PubMed: 23011778]

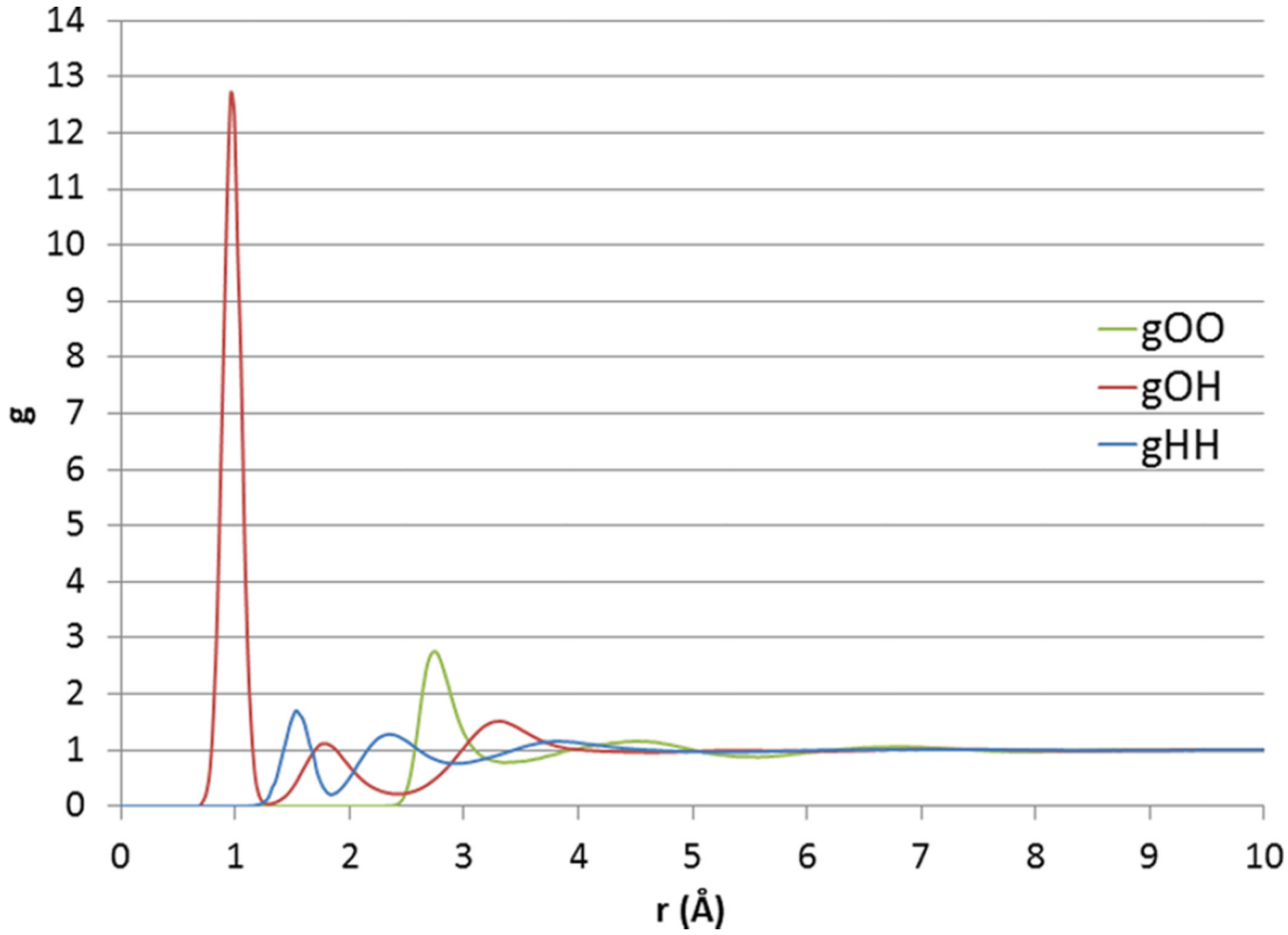


Figure 1.
Radial distribution functions of pure water at 298 K, as resolved by Soper.[29]

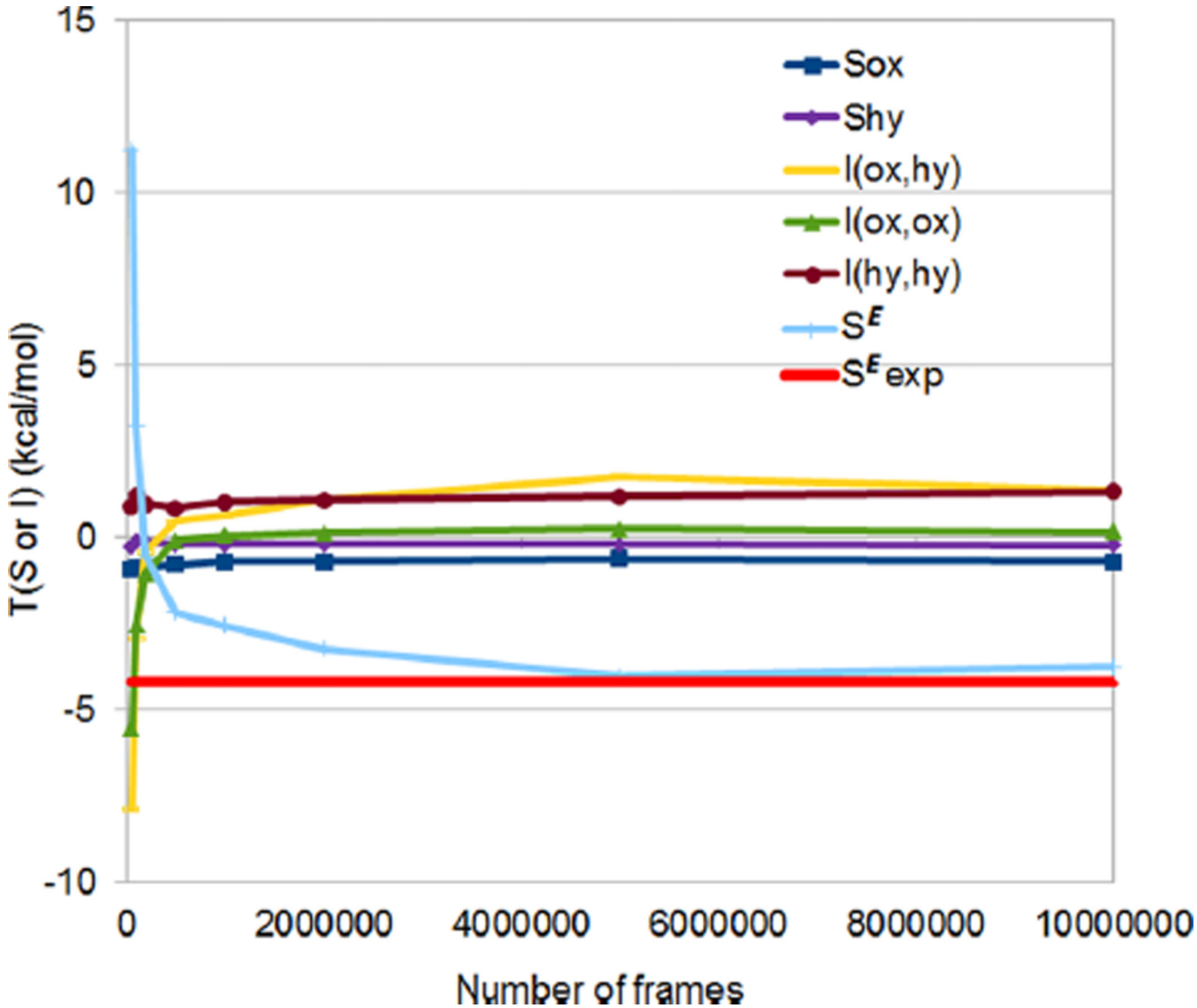


Figure 2.

Convergence of the filtered marginal (S_{ox} and S_{hy}) and pairwise ($I_{(ox,hy)}$, $I_{(ox,ox)}$, and $I_{(hy,hy)}$) contributions to S^E and the total value of $S^E = S_{ox} + S_{hy} - I_{(ox,hy)} - I_{(ox,ox)} - I_{(hy,hy)}$ as a function of number of frames. S_{ox}, S_{hy}, $I_{(ox,hy)}$, $I_{(ox,ox)}$, and $I_{(hy,hy)}$ correspond, respectively, to the 1st, 2nd, 3rd, 4th, and 5th terms in eq 20. The experimental value of S^E is also included for reference.

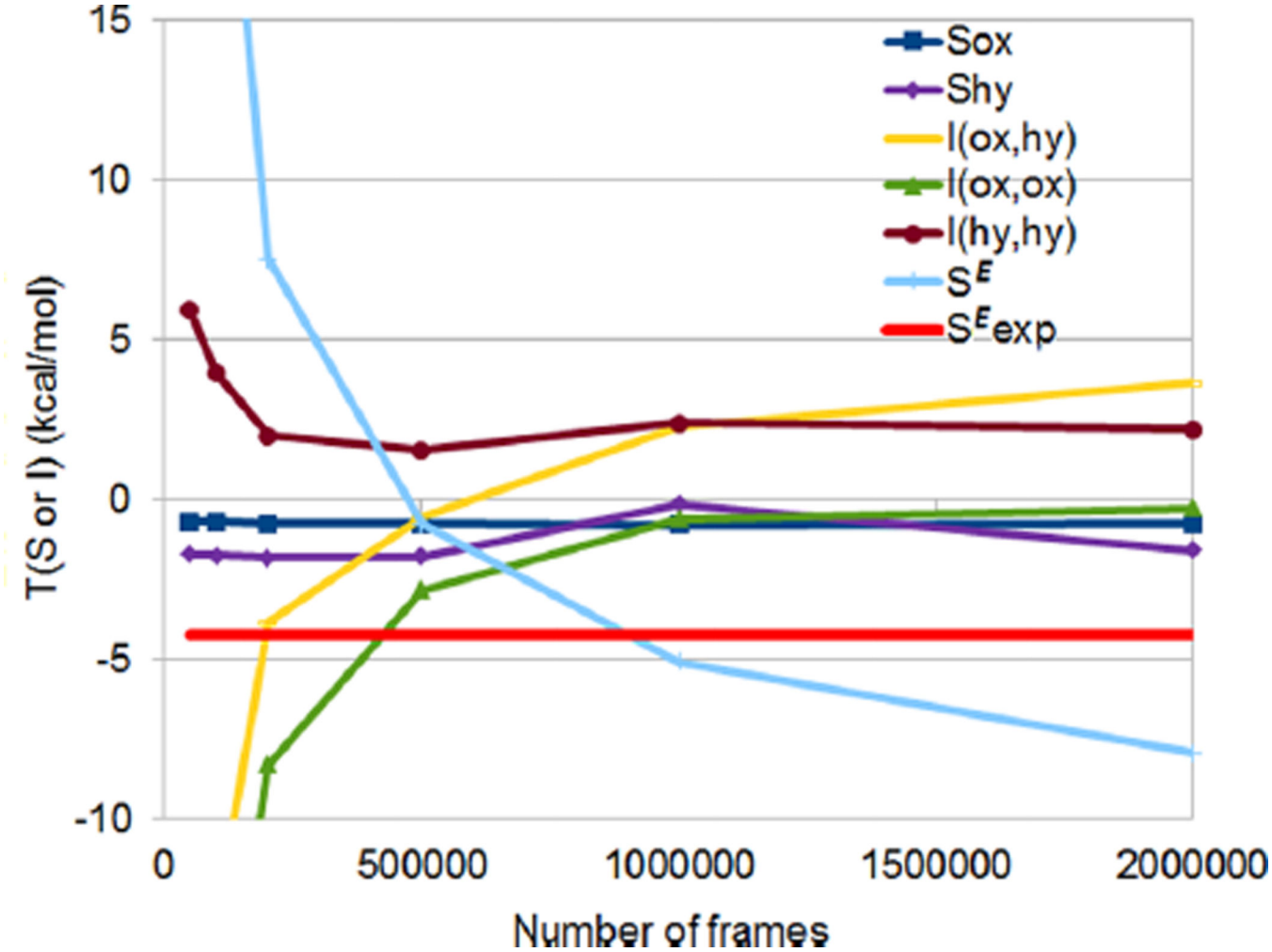


Figure 3.

Convergence of the unfiltered marginal and pairwise contributions to S^E and total value of S^E, as a function of number of frames. The scale of the *y*-axis has been matched to Figure 2 to facilitate comparison.

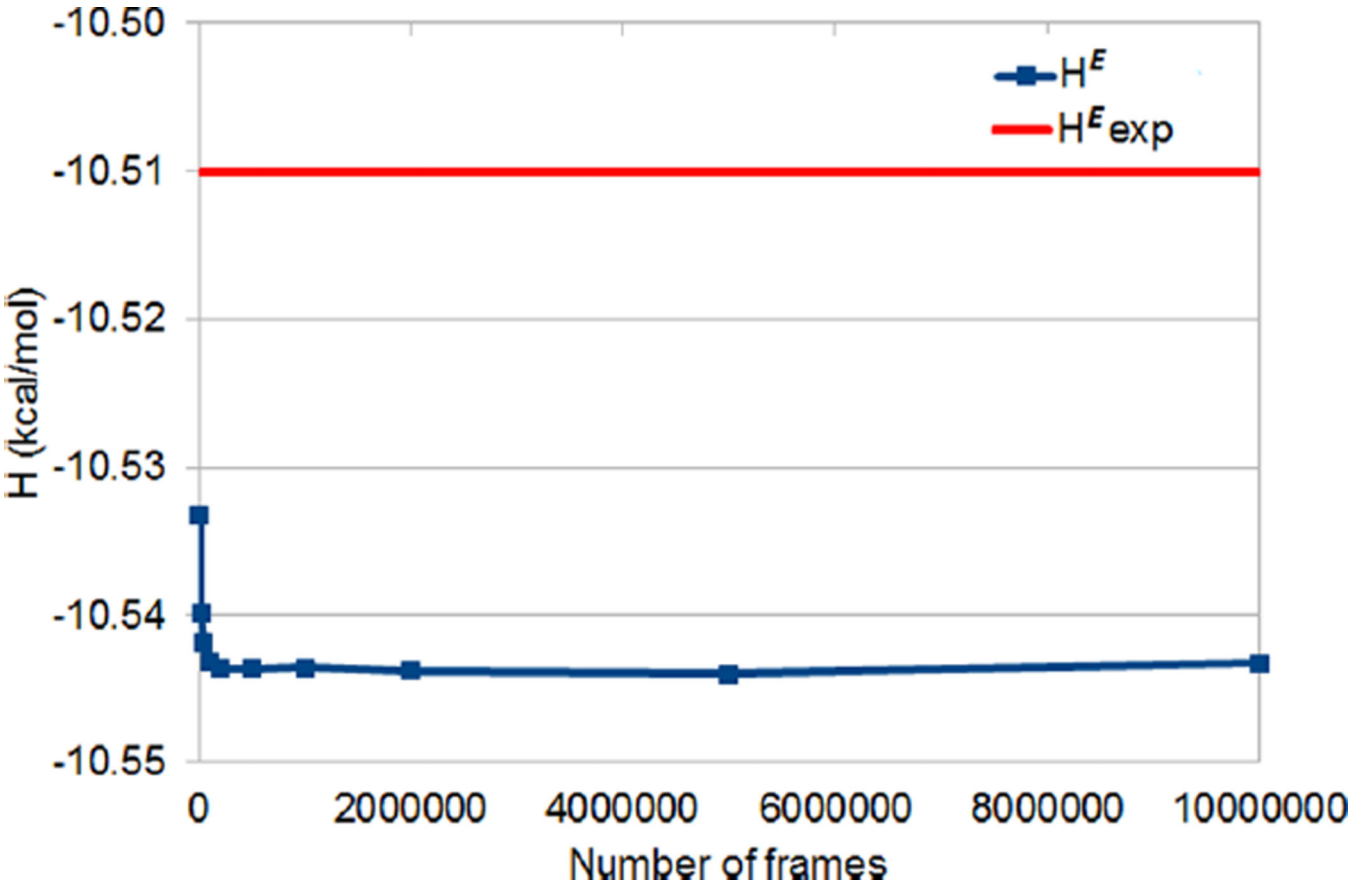


Figure 4.

Convergence of the per-water H^E as a function of number of frames for bulk (unrestrained) water. The experimental value of H^E is also included for reference.

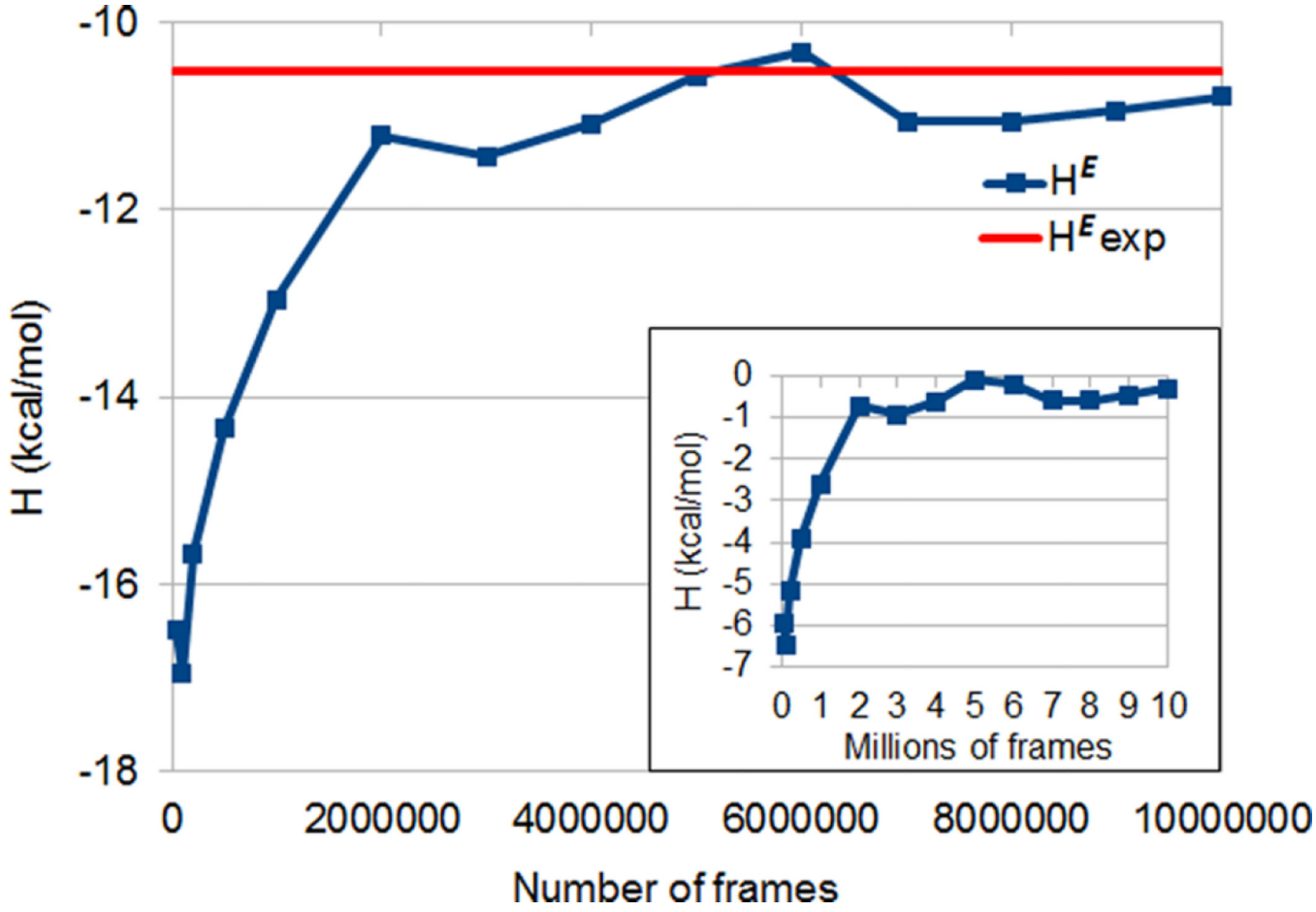


Figure 5.

Convergence of H^E as a function of number of frames for simulations with one water oxygen stiffly restrained. To produce the inset plot, we subtracted the enthalpy of the restrained water (which for this case was found to be concentrated in four voxels) from H^E at each data point; in this way, we isolated the convergence of the noise arising from bulk-like enthalpy voxels.

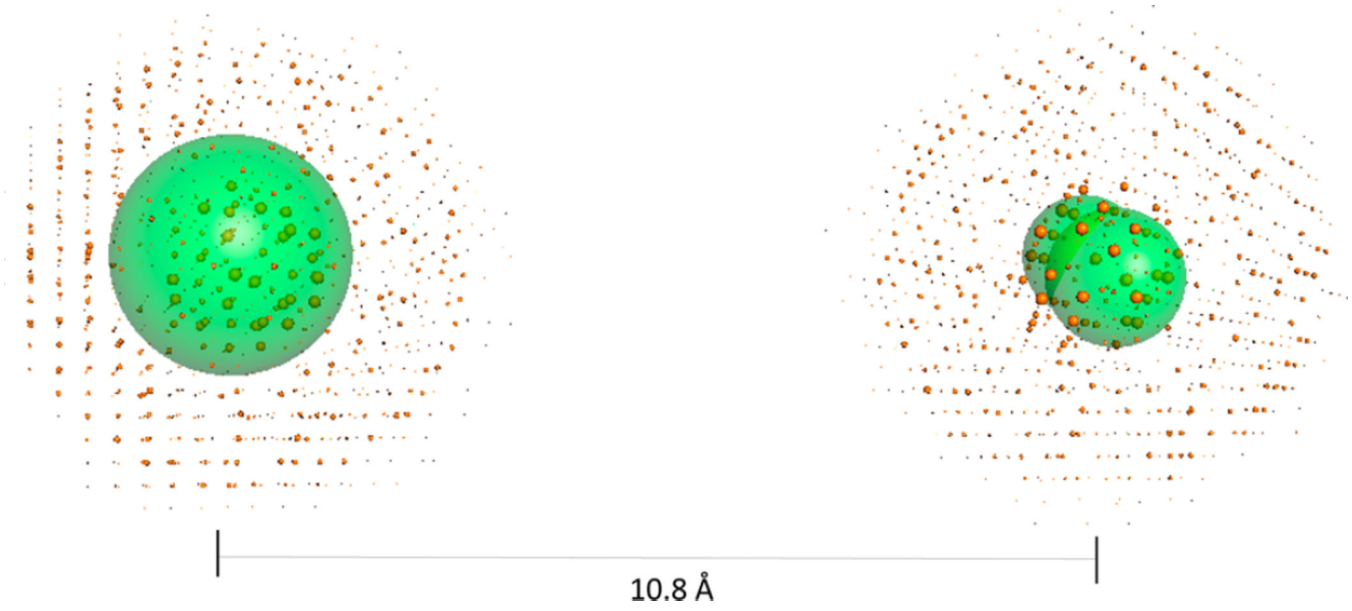


Figure 6.

Per-voxel nonbulk entropy (solid orange) and enthalpy (translucent green) for a water simulation with two restrained water oxygens that are ~10.8 Å apart from each other. Each voxel is represented by a sphere whose size correlates with the magnitude of the thermodynamic quantity estimated for that volume element. For better visualization, both enthalpies and entropies have been independently rescaled based on the range of their magnitudes. We note that the volume of the spheres representing the enthalpy of the voxels harboring the restrained oxygens is larger for the left field because the tethered particle remains in a single voxel, whereas the one on the right-hand side visits more than one voxel.

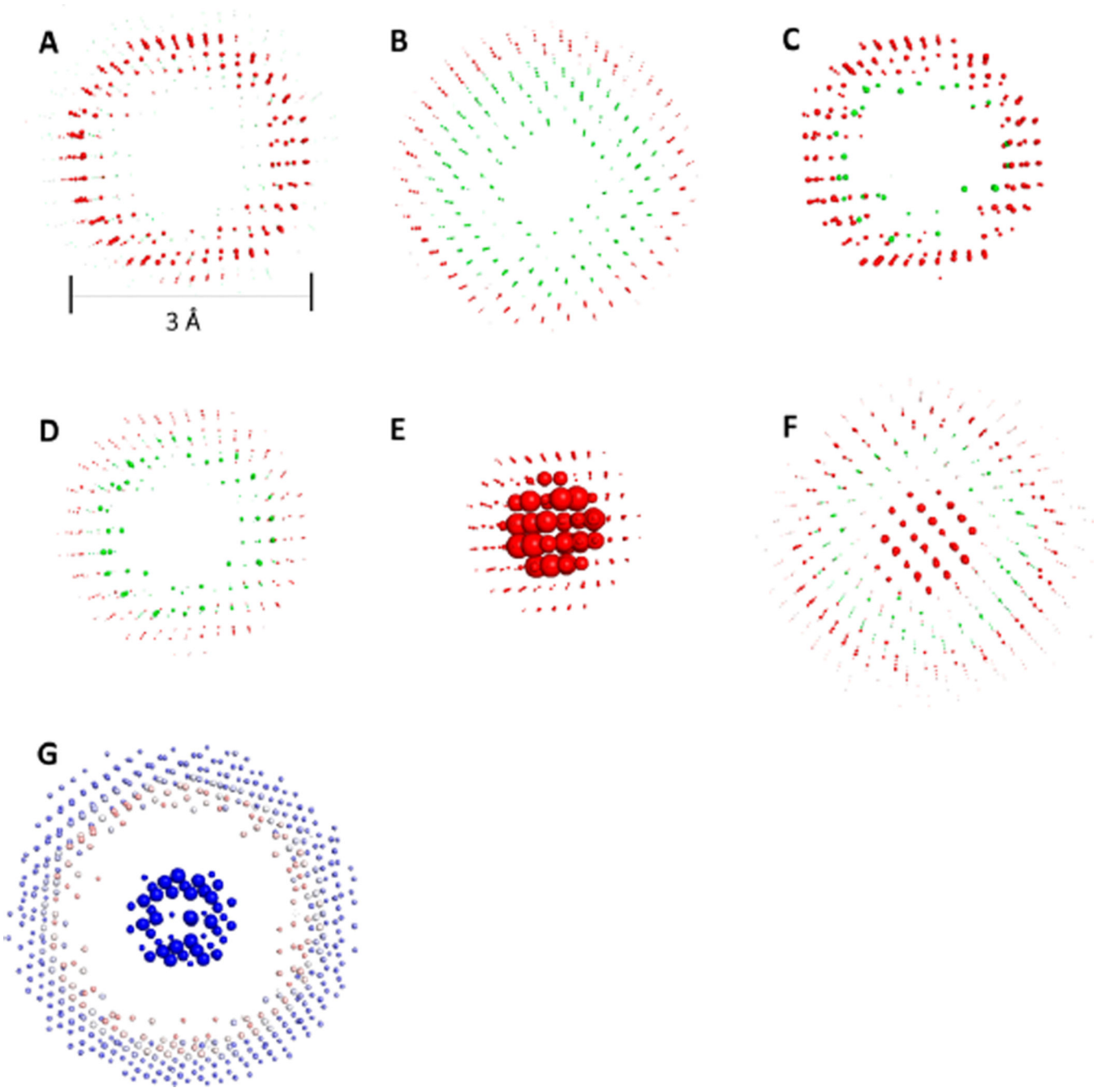


Figure 7.

Per-voxel nonbulk entropy structures produced upon fixing a water oxygen within a bulk water system. (A) S_{ox}, (B) S_{hy}, (C) $I_{(ox,hy)}$, (D) $I_{(ox,ox)}$, (E) $I_{(hy,hy)}$, and (F) S_{tot}. For reference purposes, the higher-than-bulk occupancy of those voxels with nonbulk entropy is also included in (G). To facilitate visualization, the plane of view has been clipped in all cases. In (A)–(F), the green and red colors correspond to favorable and unfavorable entropy relative to bulk solvent, respectively. The coloring of (G) correlates with the "charge" of that voxel: In brief, spheres are colored in red-white-blue spectrum, where intense red

corresponds to oxygen-only occupancy, deep blue to hydrogen-only occupancy, and white a 2:1 hydrogen:oxygen (i.e., water-like) occupancy. Sphere size conventions are as defined in Figure 6. Additional information concerning visualization of such structures can be found in ref 1. A reference length-scale has been provided in (A).

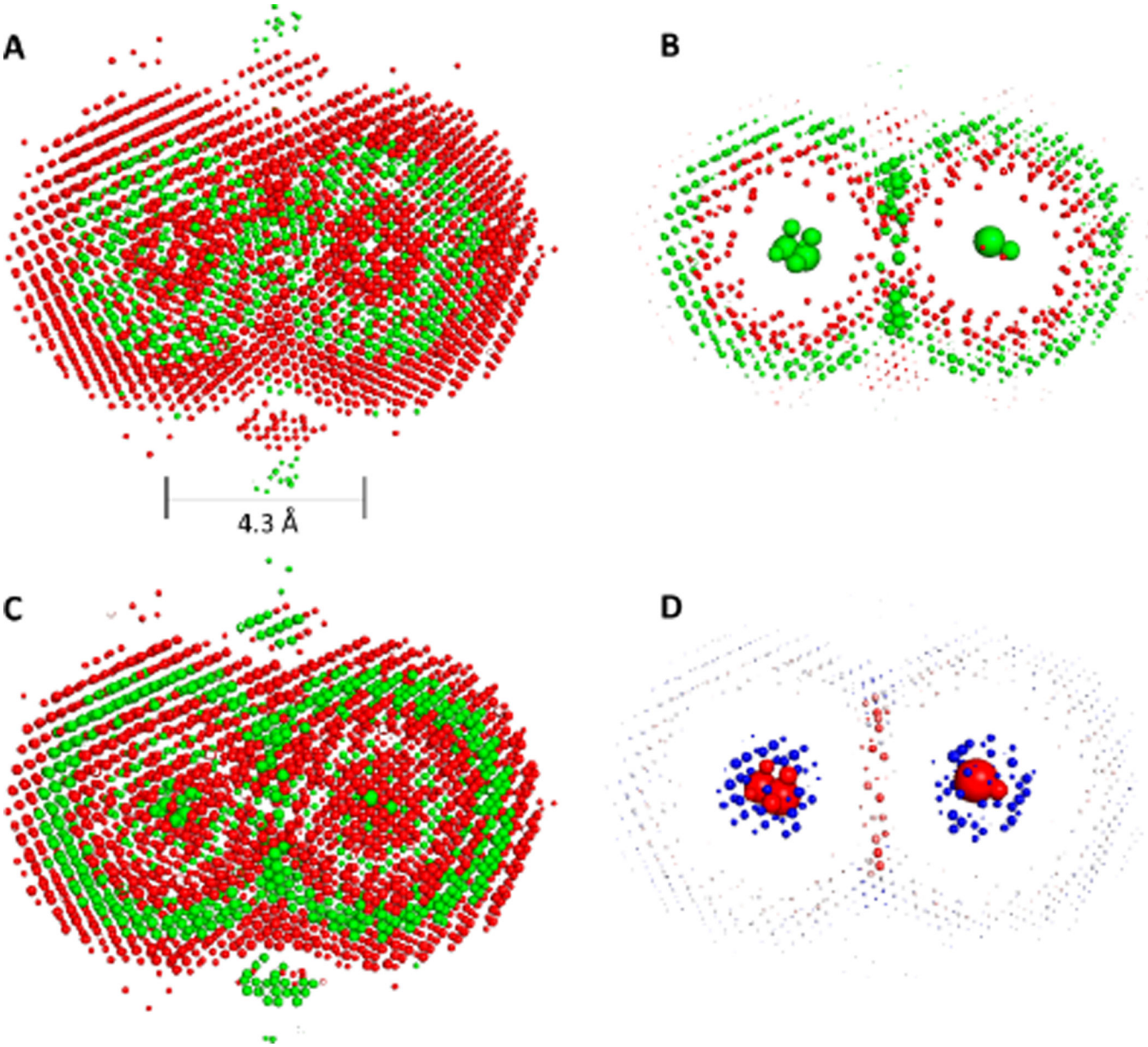


Figure 8.

Per-voxel nonbulk structure of (A) ΔS_{solv}, (B) ΔH_{solv}, (C) ΔG_{solv}, and (D) higher-than-bulk occupancy of those voxels with nonzero ΔG_{solv}, for a water system with two water oxygens ~4.3 Å apart. For each structure, all quantities were normalized with respect to the minimum (making the magnitude of all values greater than or equal to 1). The natural logarithm of the resulting quantities was then taken. Each field was subsequently rescaled and clipped to obtain the visually tractable fields shown. Sphere color and size conventions are the same as in Figures 6 and 7.

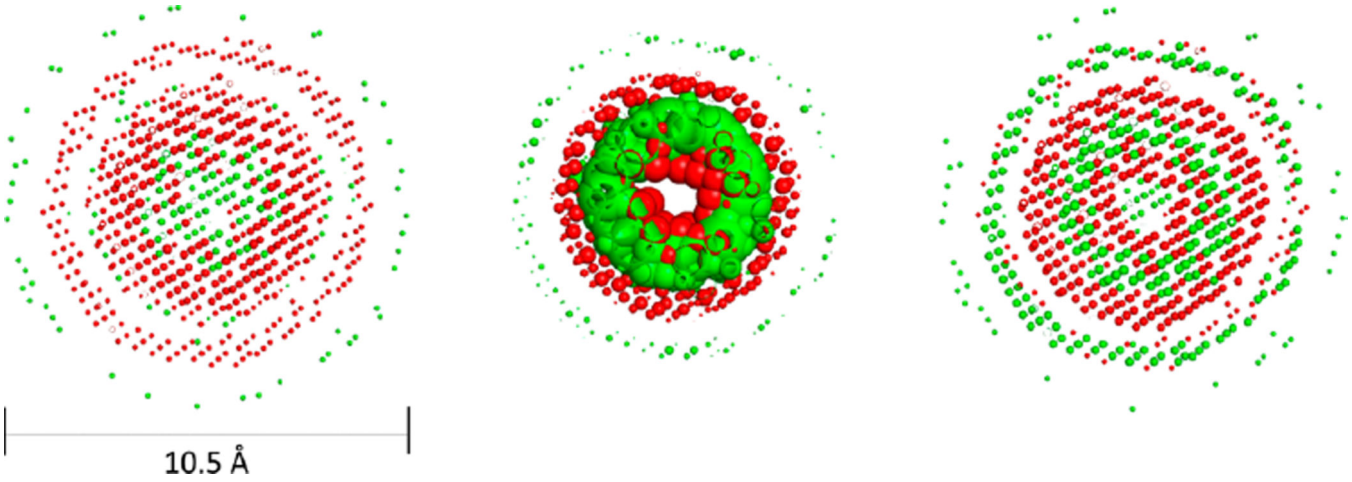


Figure 9.

Per-voxel nonbulk structure of (A) ΔS_{solv}, (B) ΔH_{solv}, and (C) ΔG_{solv}. These views correspond to a 90-degree rotation of Figures 8A–C, along the *x*-axis. The enthalpy has been scaled by a smaller factor relative to Figure 8B, to allow visualization of the outermost ring. In accordance with this rescaling, the enthalpically favorable inner ring in (B) corresponds to the enthalpically favorable outer rings in Figure 8B. Sphere color and size is the same as in Figures 6 and 7. A reference length-scale has been provided in (A).