THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Simultaneous Identification of Multiple Binding Sites in Proteins: A Statistical Mechanics Approach

Published as part of The Journal of Physical Chemistry virtual special issue "Dave Thirumalai Festschrift".

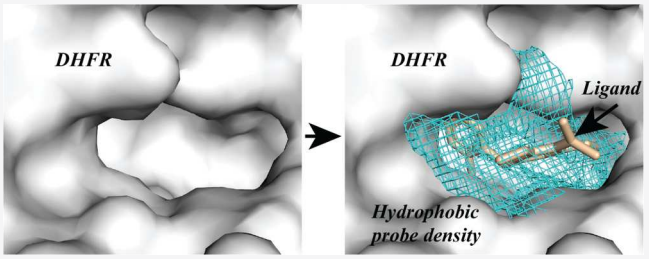
Patrice Koehl,* Marc Delarue,* and Henri Orland*

Cite This: *J. Phys. Chem. B* 2021, 125, 5052–5067

Read Online

ACCESS | Metrics & More | Article Recommendations

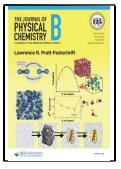
ABSTRACT: We present an extension of the Poisson−Boltzmann model in which the solute of interest is immersed in an assembly of self-orienting Langevin water dipoles, anions, cations, and hydrophobic molecules, all of variable densities. Interactions between charges are controlled by electrostatics, while hydrophobic interactions are modeled with a Yukawa potential. We impose steric constraints by assuming that the system is represented on a cubic lattice. We also assume incompressibility; i.e., all sites of the lattice are occupied. This model, which we refer to as the Hydrophobic Dipolar Poisson−Boltzmann Langevin (HDPBL) model, leads to a system of two equations whose solutions give the water dipole, salt, and hydrophobic molecule densities, all of them in the presence of the others in a self-consistent way. We use those to study the organization of the ions, cosolvent, and solvent molecules around proteins. In particular, peaks of densities are expected to reveal, simultaneously, the presence of compatible binding sites of different kinds on a protein. We have tested and validated the ability of HDPBL to detect pockets in proteins that bind to hydrophobic ligands, polar ligands, and charged small probes as well as to characterize the binding sites of lipids for membrane proteins.



1. INTRODUCTION

Proteins are unique among biomolecules in that their function is modulated by a wide variety of small molecules that use different types of interactions in their binding sites. These binding sites can bind natural substrates, as observed for example in enzyme active sites, as well as in allosteric regulatory sites. They are also the target of artificial or nonnatural ligands, developed by the pharmaceutical industry as drugs that can inhibit or activate protein function. The identification and characterization of those binding sites are therefore essential steps for understanding and controlling protein function by structure-based drug design, and as such have been central themes in structural biology. Many experimental methods have been developed for this purpose, based on NMR,[1] X-ray crystallography,[2] and now even cryo-EM.[3] These methods mostly rely on the concept of fragment-based drug discovery (FBDD),[4] in which, first, small chemical fragments are identified as possibly weak binders to the biological target and then grown or combined to produce a lead with a higher affinity.[5] Those techniques, however, are time-consuming and often expensive. This has led to parallel developments in computational structural biology with the same goals of identifying druggable binding sites in proteins (for reviews, see refs 6−8). The corresponding methods are usually fast and easy to implement. Their success, however, is often mitigated by the fact that they rely on drastic simplifications. Geometric methods assume a static structure of the protein; they identify pockets and clefts within the protein and do not take into account the geometry and/or stereochemistry of the putative ligand.[6,9] Energy-based methods combine geometry to position a putative ligand near the protein of interest with energy calculations to estimate the likeliness of their interactions. Q-SiteFinder[9] for example coats the protein surface with a layer of methyl probes and calculate van der Waals interaction energies between the protein and those probes. Probes with favorable interaction energies are retained and clusters of these probes are deemed to define putative binding sites. Other methods with more elaborate energy functions have been developed, such as GRID,[10] MCSS,[11] and FTMAP.[12,13] Those methods rely on sophisticated sampling procedures (usually based on Fourier analysis), and they also include solvent in the computation of the energy. They all follow a concept similar to FBDD as they consider small probes that can then be combined.[14] These

Received: March 25, 2021
Revised: April 27, 2021
Published: May 11, 2021



https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

techniques, however, often yield a large number of false-positive energy minima. A successful computational technique for identifying drug-binding sites requires adequate and efficient sampling of the conformation of the ligand, of the protein, and of the environment within the putative binding site, including the presence of ordered water molecules and salt. Molecular dynamics (MD) simulations provide a rigorous framework to implement such sampling.[15−18] Of particular interest, the incorporation of cosolvents to mimic the ligand in such simulations improves the sampling and the identification of binding sites.[19−24] Molecular dynamics, however, is time-consuming as proper sampling requires long simulation times. In this paper, we propose a complementary technique based on statistical mechanics for identifying simultaneously one or several binding sites of different kinds in a given protein.

In an MD simulation interactions between atoms are usually described by semiempirical "force fields", with different levels of approximations (for recent reviews, see refs 25−28). Applications of these force fields imply that the positions of all atoms be known. While this seems to be a simple requirement, it is unfortunately difficult to meet when modeling solvated large biomolecular systems. This is mostly due to the inherent difficulties in accounting for the mobile solvent molecules and ions that surrounds the solute and of the size of the system that increases significantly when one includes solvent and ions in the simulation. To circumvent these difficulties, there have been continuous efforts to develop simplified models that remain physically accurate. Many of these models consider the solvent implicitly, reducing the solute−solvent interactions to their mean field characteristics. In such models, the solvent is treated as a dielectric continuum; they are referred to as continuum dielectric models. The Poisson−Boltzmann (PB) theory is one such model, unquestionably the most popular, which provides a framework for calculating the electrostatics solvation free energy of a solute in such a dielectric continuum. The corresponding PB equation (PBE) is not however the remedy to all problems associated with characterizing the electrostatics interactions for a biomolecule: it remains a mean field approximation, with known limitations. Many improvements have been proposed to the PB equation. The novel technique proposed in this paper is one such extension. We first review the current theoretical developments around the PB model and their relationships with our new model.

PBE is only a mean-field approximation to the multibody problem of solvent−solute electrostatics interactions. It is based on several approximations that have proved to be limitations in some cases. For example, PBE does not include effects due to ion size or ion−ion correlations in its treatment (for review, see Grochowski and Trylska[29]). Solutions have been proposed to account for at least ion size using either a single size[30] or two different sizes,[31] yielding a size-modified Poisson−Boltzmann equation (SMPB), which is used to study ion channels[32] and electrolyte solutions.[33] Classical PBE handles solvent implicitly, with a fixed dielectric constant inside the biomolecule (usually set to 2−4) that abruptly jumps to 80 at the interface between the biomolecule and the solvent. This approximation leads to too much importance to the definition of this interface, usually set to the molecular surface or the vdW surface of the solute.[34] Nicholls and colleagues[35] proposed to solve this problem by using a Gaussian representation of the atoms of the solutes. Their solution, however, does not circumvent an even more important limitation of the PB equation: because of polarization effects in the vicinity of charges, a representation of the solvent as a homogeneous dielectric medium is bound to be erroneous close to the interface. We have developed an extension to the PB equation in which the solvent is described as an assembly of interacting dipoles on a lattice gas to account for the nonuniform dielectric property of the solvent.[36−40] We referred to the corresponding equation set as the dipolar Poisson−Boltzmann-Langevin (DPBL) equations. Here we describe an extension to DPBL, which we refer to as HDPBL, in which we allow for the addition of an hydrophobic cosolvent in the environment of the solute of interest. These cosolvent molecules interact with each other and with hydrophobic "charges" set on hydrophobic groups inside the solute (usually CH2 and CH3 groups). Those interactions are represented with a Yukawa potential. The electrostatic field and an additional attractive field associated with the Yukawa potential are found to be solution of a system of two PB-like partial differential equations (PDE). Those fields are used to compute the water, ions, and cosolvent densities that are then used to map the characteristics of pockets in the neighborhood of the protein of interest.

The HDPBL model proposed in this paper is an alternate approach to the grid-based drug mapping procedures such as GRID and FTMap, as well as to the cosolvent-enhanced MD simulations. Its novelty and possible advantages compared to these methods are based on

(i) *Improved Sampling*. The conformational space and degrees of freedom of solvent, ions, and cosolvent molecules within and around the target biomolecule are sampled efficiently and their densities are computed self-consistently.

(ii) *Multiprobe Exploration*. HDPBL enables incorporation of multiple probe types in the analyses. The output of HDPBL are the densities of dipoles representing a polar solvent, anions, cations, and hydrophobic cosolvent molecules as a function of position, thereby allowing for the identification of polar, positively charged, negatively charged, and hydrophobic binding sites simultaneously.

(iii) *Efficiency*. MD simulations solve Newton's equations at the atomic level numerically to construct a picture of the dynamics of the biomolecule of interest. Those simulations require significant computing time to provide adequate sampling. In contrast, HDPBL relies on solving once a system of coupled second order PDEs. We will show that this can be performed efficiently using standard techniques for solving elliptic PDEs.

This paper is organized as follows. The next section provides a complete overview of the HDPBL model. The following section describes our implementation of a solver of HDPBL in a program we refer to as AquaVit. AquaVit is a continuation of AquaSol;[40] it is heavily based on the package MG developed by Michael Holst.[41] The next section provides examples of the usefulness of HDPBL for detecting and characterizing binding sites in proteins. We conclude the paper with a discussion on the limitations of HDPBL, of possible improvements that would circumvent those limitations, as well as on possible developments around HDPBL.

2. THE HDPBL MODEL

2.1. The System: Solute, Salt, Solvent, and Cosolvent.

We consider a fixed solute molecule (protein) in a solution containing water, a z:z salt (e.g., NaCl for which the valence z is 1), and a cosolvent, i.e. neutral hydrophobic molecules. Those hydrophobic molecules will serve as probes to assess the

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

hydrophobic environment within and around the solute of interest. Methane, xenon, krypton, or any other small hydrophobic molecule may be considered.

We assume that the system comprises N_w water molecules, N_s salt molecules (that is N_s cations of charge $+ze_c$ and N_s anions of charge $-ze_c$ where e_c is the elementary electronic charge) and N_h inert hydrophobic molecules. Those molecules are only present in the region outside of the solute, which is characterized by its molecular surface. The water molecules are modeled as dipoles with fixed dipole moment p_0. The charged ions and the water dipoles interact through the Coulomb interaction, and all particles are subject to steric repulsion. To model this steric repulsion, we assume that the system is represented on a cubic lattice gas of lattice spacing a and that at each site of the lattice there is either no particle, or one ion or one water molecule or one hydrophobic molecule. Each mobile particle is thus modeled as a hard sphere with radius $a/2$. We denote respectively by $n_+(\mathbf{r})$, $n_-(\mathbf{r})$, $n_w(\mathbf{r})$, and $n_h(\mathbf{r})$, the occupation numbers of site $\mathbf{r}$ by a cation, an anion, a water or a hydrophobic molecule, where each of these numbers are 1 or 0 depending whether the corresponding particle is present or absent. Note that $\mathbf{r}$ represents one of the cuboids of the cubic lattice. the steric constraint imposes that there is at most one particle at any site $\mathbf{r}$

$$n_+(\mathbf{r}) + n_-(\mathbf{r}) + n_w(\mathbf{r}) + n_h(\mathbf{r}) = 0 \text{ or } 1 \tag{1}$$

If the system is incompressible, the above sum is strictly equal to 1, meaning that there is exactly one particle at each site. The relationships between the occupation numbers of all species and their number of molecules are given by the additional constraints

$$\begin{aligned}
\sum_{\mathbf{r}} n_+(\mathbf{r}) &= N_s \\
\sum_{\mathbf{r}} n_-(\mathbf{r}) &= N_s \\
\sum_{\mathbf{r}} n_w(\mathbf{r}) &= N_w \\
\sum_{\mathbf{r}} n_h(\mathbf{r}) &= N_h
\end{aligned} \tag{2}$$

The corresponding particle densities are given by

$$\begin{aligned}
\rho_{\pm}(\mathbf{r}) &= \frac{n_{\pm}(\mathbf{r})}{a^3} \\
\rho_w(\mathbf{r}) &= \frac{n_w(\mathbf{r})}{a^3} \\
\rho_h(\mathbf{r}) &= \frac{n_h(\mathbf{r})}{a^3}
\end{aligned} \tag{3}$$

We emphasize that all the occupation numbers of mobile particles are zero inside the solute, since we assume that mobile particles cannot penetrate the solute. Let c_s, c_w, and c_h be the bulk concentrations of salt, water, and hydrophobic probes, respectively. We introduce the volume fraction Φ for each type of particle

$$\begin{aligned}
\Phi_s &= 2c_s a^3 \\
\Phi_w &= c_w a^3 \\
\Phi_h &= c_h a^3
\end{aligned} \tag{4}$$

In addition, we consider a volume fraction for vacancies:

$$\Phi_v = 1 - (\Phi_s + \Phi_w + \Phi_h) \tag{5}$$

Φ_v should be understood as follows. If the system is incompressible, $\Phi_v = 0$ and the concentrations of salt, water, and hydrophobic probes are necessarily dependent. Otherwise, Φ_v is positive, and vacancies are possible in the environment of the solute. In this case, c_s, c_w, and c_h are independent, although they still need to satisfy $\Phi_s + \Phi_w + \Phi_h = 2c_s a^3 + c_w a^3 + c_h a^3 \leq 1$.

We denote by $v(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r}$ the Coulomb potential, where ε_0 is the dielectric permittivity of the vacuum and $r = |\mathbf{r}|$.

The hydrophobic interactions between the hydrophobic molecules are captured by an attractive Yukawa potential

$$w(\mathbf{r}) = -\frac{w_0}{4\pi}\frac{e^{-\kappa r}}{r} \tag{6}$$

where $\kappa = 1/l_h$ defines the range of the hydrophobic interaction, and $w_0 > 0$ defines its strength. The negative sign in eq 6 denotes the attractive nature of the interaction.

2.2. An Effective Free Energy for the System. The canonical partition function of the system on the lattice can be written as

$$\begin{aligned}
Z_c(N_s, N_w, N_h) = &\sum_{\{n_{\pm}(\mathbf{r})=0,1\}} \sum_{\{n_w(\mathbf{r})=0,1\}} \sum_{\{n_h(\mathbf{r})=0,1\}} \delta \\
&\left(\sum_{\mathbf{r}} n_+(\mathbf{r}) - N_s\right)\delta\left(\sum_{\mathbf{r}} n_-(\mathbf{r}) - N_s\right) \\
&\times \delta\left(\sum_{\mathbf{r}} n_w(\mathbf{r}) - N_w\right)\delta\left(\sum_{\mathbf{r}} n_h(\mathbf{r}) - N_h\right) \\
&\times \int \prod_{\mathbf{r}} \mathrm{d}\mathbf{p}_w(\mathbf{r})\, \delta(\mathbf{p}_w{}^2(\mathbf{r}) - p_0{}^2) \\
&\exp\left(-\frac{\beta}{2}\sum_{\mathbf{r},\mathbf{r}'} \rho_c(\mathbf{r}) v(\mathbf{r}-\mathbf{r}') \rho_c(\mathbf{r}')\right. \\
&\left. - \frac{\beta}{2}\sum_{\mathbf{r},\mathbf{r}'} \rho_H(\mathbf{r}) w(\mathbf{r}-\mathbf{r}') \rho_H(\mathbf{r}')\right)
\end{aligned} \tag{7}$$

where $\beta = 1/k_B T$ with T being the temperature and k_B the Boltzmann constant. The notation $\sum_{\{n_{\pm}(\mathbf{r})=0,1\}}$ denotes a sum over all possible combinations of values of the occupation numbers (0 or 1) at all lattice sites. The total charge density ρ_c in (7) is the sum of the charge densities of the ions, of the point-like water dipoles, and of the charges of the solute

$$\rho_c(\mathbf{r}) = ze_c(\rho_+(\mathbf{r}) - \rho_-(\mathbf{r})) - \mathbf{p}_w(\mathbf{r})\cdot\nabla\rho_w(\mathbf{r}) + \rho_f(\mathbf{r}) \tag{8}$$

where $\rho_f(\mathbf{r})$ is the charge density of the fixed charges of the solute at point $\mathbf{r}$ and $\mathbf{p}_w(\mathbf{r})$ is the dipole moment of the water molecule at the same point. These dipole moments have a fixed magnitude p_0 but can take all possible orientations. This is accounted for by the integral over all dipole orientations in (7).

Finally, $\rho_H(\mathbf{r})$ is the sum of the density $\rho_h(\mathbf{r})$ of mobile hydrophobic particles and of the density $\rho_p(\mathbf{r})$ of the hydrophobic sites of the fixed solute

$$\rho_H(\mathbf{r}) = \rho_h(\mathbf{r}) + \rho_p(\mathbf{r}) \tag{9}$$

Going to the grand canonical ensemble and introducing the chemical potentials μ and fugacities λ of the various species

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

$$\lambda_s = e^{\beta\mu_s}$$

$$\lambda_w = e^{\beta\mu_w}$$

$$\lambda_h = e^{\beta\mu_h}$$

the grand partition function can be written as

$$\Xi = \sum_{N_s=0}^{+\infty}\sum_{N_w=0}^{+\infty}\sum_{N_h=0}^{+\infty} \frac{e^{2\beta\mu_s N_s+\beta\mu_w N_w+\beta\mu_h N_h}}{(N_s!)^2 N_w! N_h!} Z_c(N_s, N_w, N_h) \tag{10}$$

where the canonical partition function Z_c is given in (7).

Following the formalism introduced in[36−39] we perform two Hubbard−Stratonovich transforms within eq 10 and integrate over the dipole moments. After a few standard manipulations, the partition function can be written in an exact manner as an integral over two fields, $\varphi(\mathbf{r})$ and $\psi(\mathbf{r})$, corresponding to the electrostatic and the Yukawa interactions, respectively:

$$\begin{aligned}\Xi = &\int \mathcal{D}\varphi(\mathbf{r})\mathcal{D}\psi(\mathbf{r})\\ &e^{-(\beta/2\int d\mathbf{r}d\mathbf{r}'\phi(\mathbf{r})v^{-1}(\mathbf{r}-\mathbf{r}')\phi(\mathbf{r}')+\beta/2\int d\mathbf{r}d\mathbf{r}'\psi(\mathbf{r})w^{-1}(\mathbf{r}-\mathbf{r}')\psi(\mathbf{r}'))}\\ &\times \exp\left(-i\beta\int d\mathbf{r}\,\varphi(\mathbf{r})\rho_f(\mathbf{r}) - \beta\int d\mathbf{r}\,\psi(\mathbf{r})\rho_p(\mathbf{r})\right)\\ &\times \prod_{\mathbf{r}}\left(\lambda_v + 2\lambda_s\cosh(i\beta z e_c\varphi(\mathbf{r})) + \lambda_w\frac{\sinh(ip_0\beta|\nabla\varphi(\mathbf{r})|)}{ip_0\beta|\nabla\varphi(\mathbf{r})|}\right.\\ &\left. + \lambda_h e^{-\beta\psi(\mathbf{r})}\right)^{\gamma(\mathbf{r})}\end{aligned} \tag{11}$$

where we have introduced a pseudofugacity for vacancies λ_v, such that $\lambda_v = 0$ if the system is incompressible and $\lambda_v = 1$ otherwise, and $\gamma(\mathbf{r})$ is the indicator function for the points available to the mobile particles: namely, $\gamma(\mathbf{r}) = 1$ outside the solute, and $\gamma(\mathbf{r}) = 0$ inside.

The operator $w^{-1}(\mathbf{r})$ is the inverse of the Yukawa interaction (6), and is given by

$$w^{-1}(\mathbf{r}-\mathbf{r}') = -\frac{1}{w_0}(-\nabla^2 + \kappa^2)\delta(\mathbf{r}-\mathbf{r}')$$

To simplify the notations and the equations, we take the continuous limit $a \to 0$ in the lattice product above, and we treat the fields $\varphi(\mathbf{r})$ and $\psi(\mathbf{r})$ as defined in the whole space. The sums are replaced by integrals according to

$$\sum_{\mathbf{r}} = \frac{1}{a^3}\int d\mathbf{r} \tag{12}$$

where the integral on the right side is over the whole 3D space, and we obtain

$$\begin{aligned}\Xi = &\int \mathcal{D}\varphi(\mathbf{r})\mathcal{D}\psi(\mathbf{r})\\ &e^{\beta\varepsilon_0/2\int d\mathbf{r}(\nabla\varphi(\mathbf{r}))^2 - \beta/2w_0\int d\mathbf{r}((\nabla\psi(\mathbf{r}))^2+\kappa^2\psi^2(\mathbf{r}))}\\ &\times \exp\left(-i\beta\int d\mathbf{r}\,\varphi(\mathbf{r})\rho_f(\mathbf{r}) - \beta\int d\mathbf{r}\,\psi(\mathbf{r})\rho_p(\mathbf{r})\right)\\ &\times \exp\left(\frac{1}{a^3}\int d\mathbf{r}\,\gamma(\mathbf{r})\ln\left(\lambda_v + 2\lambda_s\cosh(i\beta z e_c\varphi(\mathbf{r}))\right.\right.\\ &\left.\left. + \lambda_w\frac{\sinh(ip_0\beta|\nabla\varphi(\mathbf{r})|)}{ip_0\beta|\nabla\varphi(\mathbf{r})|} + \lambda_h e^{-\beta\psi(\mathbf{r})}\right)\right)\end{aligned} \tag{13}$$

The functional integral in eq 13 is evaluated by the Saddle-Point Approximation. This method, which is also called Mean-Field Theory, consists of minimizing the exponent in the above equation with respect to the two fields φ and ψ. The field φ is pure imaginary at the saddle-point, while ψ is real, and the exponent above can be written as an effective free energy as

$$\begin{aligned}\mathcal{F} = &-\frac{\varepsilon_0}{2}\int d\mathbf{r}\,(\nabla\varphi(\mathbf{r}))^2 + \frac{1}{2w_0}\int d\mathbf{r}\,((\nabla\psi(\mathbf{r}))^2\\ &+ \kappa^2\psi^2(\mathbf{r})) + \int d\mathbf{r}\,\varphi(\mathbf{r})\rho_f(\mathbf{r}) + \int d\mathbf{r}\,\psi(\mathbf{r})\rho_p(\mathbf{r})\\ &- \frac{k_B T}{a^3}\int d\mathbf{r}\,\gamma(\mathbf{r})\ln\left(\lambda_v + 2\lambda_s\cosh(\beta z e_c\varphi(\mathbf{r}))\right.\\ &\left. + \lambda_w\frac{\sinh(p_0\beta|\nabla\varphi(\mathbf{r})|)}{p_0\beta|\nabla\varphi(\mathbf{r})|} + \lambda_h e^{-\beta\psi(\mathbf{r})}\right)\end{aligned} \tag{14}$$

2.3. Optimizing the Free Energy. The mean field equations are the Euler−Lagrange equations obtained by minimizing 14 with respect to φ and ψ

$$\begin{aligned}-\varepsilon_0\nabla^2\varphi(\mathbf{r}) &= \rho_f(\mathbf{r}) - \frac{2\lambda_s}{a^3}z e_c\gamma(\mathbf{r})\frac{\sinh(\beta z e_c\varphi(\mathbf{r}))}{\mathcal{D}(\mathbf{r})}\\ &+ \frac{p_0}{a^3}\lambda_w\gamma(\mathbf{r})\nabla\cdot\left(\frac{\nabla\varphi(\mathbf{r})}{|\nabla\varphi(\mathbf{r})|}\frac{g(p_0\beta|\nabla\varphi(\mathbf{r})|)}{\mathcal{D}(\mathbf{r})}\right)\\ \frac{1}{w_0}(-\nabla^2+\kappa^2)\psi(\mathbf{r}) &= -\rho_p(\mathbf{r}) - \frac{\lambda_h}{a^3}\gamma(\mathbf{r})\frac{e^{-\beta\psi(\mathbf{r})}}{\mathcal{D}(\mathbf{r})}\\ \mathcal{D}(\mathbf{r}) &= \lambda_v + 2\lambda_s\cosh(\beta z e_c\varphi(\mathbf{r})) + \lambda_w\frac{\sinh(p_0\beta|\nabla\varphi(\mathbf{r})|)}{p_0\beta|\nabla\varphi(\mathbf{r})|}\\ &+ \lambda_h e^{-\beta\psi(\mathbf{r})}\end{aligned} \tag{15}$$

where

$$g(x) = \frac{\cosh x}{x} - \frac{\sinh x}{x^2}$$

Note that $\varphi(\mathbf{r}) \to 0$ and $\psi(\mathbf{r}) \to \psi_0$ as $\mathbf{r} \to +\infty$, i.e. in the bulk, far from the solute. All coefficients in those equations are computed either from physical constants, or from input information describing the system, with the exception of the fugacities and ψ_0, which we derive now.

The fugacities are determined by the equations

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

$$-\lambda_s \frac{\partial(\beta F)}{\partial \lambda_s} = N_s$$
$$-\lambda_w \frac{\partial(\beta F)}{\partial \lambda_w} = N_w$$
$$-\lambda_h \frac{\partial(\beta F)}{\partial \lambda_h} = N_h \tag{16}$$

These equations translate into

$$\int d\mathbf{r}\, \frac{2\lambda_s}{\mathcal{D}(\mathbf{r})} \cosh(\beta z e_c \varphi(\mathbf{r})) = a^3 N_s$$
$$\int d\mathbf{r}\, \frac{\lambda_w}{\mathcal{D}(\mathbf{r})} \frac{\sinh(p_0 \beta |\nabla \varphi(\mathbf{r})|)}{p_0 \beta |\nabla \varphi(\mathbf{r})|} = a^3 N_w$$
$$\int d\mathbf{r}\, \frac{\lambda_h}{\mathcal{D}(\mathbf{r})} e^{-\beta \psi(\mathbf{r})} = a^3 N_h \tag{17}$$

Assuming that the volume of the solution is large compared to that of the solute,

$$\frac{\lambda_s}{\lambda_v + 2\lambda_s + \lambda_w + \lambda_h e^{-\beta\psi_0}} = a^3 c_s = \frac{\Phi_s}{2}$$
$$\frac{\lambda_w}{\lambda_v + 2\lambda_s + \lambda_w + \lambda_h e^{-\beta\psi_0}} = a^3 c_w = \Phi_w$$
$$\frac{\lambda_h e^{-\beta\psi_0}}{\lambda_v + 2\lambda_s + \lambda_w + \lambda_h e^{-\beta\psi_0}} = a^3 c_H = \Phi_h \tag{18}$$

where Φ_s, Φ_w, and Φ_h are the volume fractions defined in eq 4. Recall that $\Phi_v = 1.0 - \Phi_s - \Phi_w - \Phi_h$ is the molar fraction of vacancies. We consider two cases:

(i) *Compressible System.* In a compressible system, we allow for vacancies and $\Phi_v \neq 0$ and $\lambda_v = 1$. Solving the system in eq 18, we get:

$$2\lambda_s = \frac{\Phi_s}{\Phi_v}$$
$$\lambda_w = \frac{\Phi_w}{\Phi_v}$$
$$\lambda_h e^{-\beta\psi_0} = \frac{\Phi_h}{\Phi_v} \tag{19}$$

(ii) *Incompressible System.* In the case of an incompressible system, there are no vacancies, and $\Phi_s+\Phi_w+\Phi_h = 1$. The fugacities are not independent, and it is possible to chose for example $\lambda_w = 1$. The fugacities are then given by

$$\lambda_w = 1$$
$$2\lambda_s = \frac{\Phi_s}{\Phi_w}$$
$$\lambda_h e^{-\beta\psi_0} = \frac{\Phi_h}{\Phi_w}$$
$$\Phi_w = 1 - \Phi_s - \Phi_h \tag{20}$$

The bulk value ψ_0 is given by

$$\psi_0 = -\frac{w_0}{\kappa^2}\frac{\Phi_h}{a^3} \tag{21}$$

Using the values derived above for the fugacities, in all cases, the meanfield equations given in eq 15 can be rewritten as

$$-\varepsilon_0 \nabla^2 \varphi(\mathbf{r}) = \rho_f(\mathbf{r}) - \frac{\Phi_s}{a^3} z e_c \gamma(\mathbf{r}) \frac{\sinh(\beta z e_c \varphi(\mathbf{r}))}{\mathcal{D}_a(\mathbf{r})} + \frac{p_0}{a^3}\Phi_w \gamma(\mathbf{r}) \nabla\cdot\left(\frac{\nabla\varphi(\mathbf{r})}{|\nabla\varphi(\mathbf{r})|}\frac{g(p_0\beta|\nabla\varphi(\mathbf{r})|)}{\mathcal{D}_a(\mathbf{r})}\right) \tag{22a}$$

$$\frac{1}{w_0}(-\nabla^2 + \kappa^2)\psi(\mathbf{r}) = -\rho_p(\mathbf{r}) - \frac{\Phi_h}{a^3}\gamma(\mathbf{r})\frac{e^{-\beta(\psi(\mathbf{r})-\psi_0)}}{\mathcal{D}_a(\mathbf{r})} \tag{22b}$$

where

$$\mathcal{D}_a(\mathbf{r}) = \Phi_v + \Phi_s \cosh(\beta z e_c \varphi(\mathbf{r})) + \Phi_w \frac{\sinh(p_0\beta|\nabla\varphi(\mathbf{r})|)}{p_0\beta|\nabla\varphi(\mathbf{r})|} + \Psi_H e^{-\beta\psi(\mathbf{r})} \tag{23}$$

Note that $\mathcal{D}_a(\mathbf{r}) \to 1$ when $\mathbf{r} \to +\infty$.

2.4. The HDPBL System of Equations. The meanfield equations eqs 22a and 22b given above fully describe the system under study. Equation 22a is a dipolar Poisson−Boltzmann Langevin (DPBL) equation,[36,38,40] while eq 22b is a Poisson−Boltzmann-like equation that relates to the hydrophobic interactions involving the hydrophobic probes in the solvent and the hydrophobic charges on the solute. As a consequence, we refer to this system of equations as the Hydrophobic Dipolar Poisson−Boltzmann Langevin equations, or HDPBL equations in short. In the following, we provide modified equations involving dimensionless potentials that are more amenable to a numerical solution, and we derive all constants necessary for those equations.

2.4.1. Revisiting the DPBL-like Equation 22a. The electrostatic potential $\varphi(\mathbf{r})$ and the field $\psi(\mathbf{r})$ are expressed in volts and Joules in the SI unit system, respectively. It is common to consider instead the dimensionless potentials $u(\mathbf{r})$ and $v(\mathbf{r})$ defined as

$$u(\mathbf{r}) = \frac{e_c\varphi(\mathbf{r})}{k_B T} = \beta e_c \varphi(\mathbf{r}) \tag{24a}$$

$$v(\mathbf{r}) = \frac{\psi(\mathbf{r}) - \psi_0}{k_B T} = \beta(\psi(\mathbf{r}) - \psi_0) \tag{24b}$$

Equation 22a can then be rewritten as

$$-\varepsilon_0 \nabla^2 u(\mathbf{r}) = \beta e_c^{\,2} \rho_{fd}(\mathbf{r}) - \frac{\Phi_s}{a^3}\beta z e_c^{\,2}\gamma(\mathbf{r})\frac{\sinh(z u(\mathbf{r}))}{\mathcal{D}_1(\mathbf{r})} + \frac{\beta e_c^{\,2} p_e}{a^3}\Phi_w\gamma(\mathbf{r})\nabla\cdot\left(\frac{\nabla u(\mathbf{r})}{|\nabla u(\mathbf{r})|}\frac{g(p_e|\nabla u(\mathbf{r})|)}{\mathcal{D}_1(\mathbf{r})}\right) \tag{25}$$

where we have defined $p_e = p_0/e_c$, and

$$\mathcal{D}_1(\mathbf{r}) = \Phi_v + \Phi_s \cosh(z u(\mathbf{r})) + \Phi_w \frac{\sinh(p_e|\nabla u(\mathbf{r})|)}{p_e|\nabla u(\mathbf{r})|} + \Phi_h e^{-v(\mathbf{r})} \tag{26}$$

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

In eq 25, ρ_{fd} is the density of fixed charges expressed as fraction of electrons, hence the ${e_c}^2$ as a coefficient, where the first e_c comes from the change to a dimensionless potential, and the second e_c is factored from the solute charges.

Let us introduce the function $F(x) = \frac{g(x)}{x}$. Note that

$$F(x) = \frac{\cosh(x)}{x^2} - \frac{\sinh(x)}{x^3} = \frac{\sinh(x)}{x^2}\mathcal{L}(x) \tag{27}$$

where $\mathcal{L}(x) = \frac{1}{\tanh(x)} - \frac{1}{x}$ is the Langevin function and $F(x) \to {}^1/_3$ when $x \to 0$.

After a few standard manipulations, taking into account that $\nabla\gamma(\mathbf{r}) = \mathbf{0}$, eq 25 can be rewritten as

$$-\nabla\cdot\left(\left(1 + \gamma(\mathbf{r})C_w\Phi_w\frac{F(p_e|\nabla u(\mathbf{r})|)}{\mathcal{D}_1(\mathbf{r})}\right)\nabla u(\mathbf{r})\right) + \gamma(\mathbf{r})C_s\Phi_s\frac{\sinh(\mathrm{zu}(\mathbf{r}))}{\mathcal{D}_1(\mathbf{r})} = C_f\rho_{fd}(\mathbf{r}) \tag{28}$$

where we have introduced the three constants

$$C_w = \frac{4\pi l_B {p_e}^2}{a^3} \tag{29a}$$

$$C_s = \frac{z4\pi l_B}{a^3} \tag{29b}$$

$$C_f = 4\pi l_B \tag{29c}$$

where l_B is the Bjerrum length in vacuum, namely

$$l_B = \frac{\beta {e_c}^2}{4\pi\varepsilon_0} \tag{30}$$

As written, eq 28 is a PB-like second order differential equation, with a relative permittivity $\varepsilon(\mathbf{r}, u, v)$ defined as

$$\varepsilon(\mathbf{r}, u, v) = 1 + \gamma(\mathbf{r})C_w\Phi_w\frac{F(p_e|\nabla u(\mathbf{r})|)}{\mathcal{D}_1(\mathbf{r})} \tag{31}$$

$\varepsilon(\mathbf{r}, u, v)$ is 1 inside the solute, and depends on both the position $\mathbf{r}$ and on the values of the potentials $u(\mathbf{r})$ and $v(\mathbf{r})$ at that position, for $\mathbf{r}$ outside the molecule. Equation 31 gives a self-consistent representation of the dielectric permittivity of the system.

2.4.2. Revisiting the PB-like Equation 22b. Using the dimensionless potentials $u(\mathbf{r})$ and $v(\mathbf{r})$ defined in eq 24, eq 22b becomes

$$\frac{1}{w_0}(-\nabla^2 + \kappa^2)(v(\mathbf{r}) + v_0) = -\beta\rho_p(\mathbf{r}) - \frac{\beta\Phi_h}{a^3}\gamma(\mathbf{r})\frac{\mathrm{e}^{-v(\mathbf{r})}}{\mathcal{D}_1(\mathbf{r})} \tag{32}$$

where

$$v_0 = \beta\psi_0 = -\frac{\beta w_0}{\kappa^2}\frac{\Phi_h}{a^3} \tag{33}$$

Note that $\beta\, w_0$ is a length, which we write as l_Y, and

$$v_0 = \frac{l_Y}{\kappa^2}\frac{\Phi_h}{a^3} \tag{34}$$

We rewrite eq 32 as

$$-\nabla^2 v(\mathbf{r}) + \kappa^2(v(\mathbf{r}) + v_0) + \gamma(\mathbf{r})\frac{l_Y}{a^3}\Phi_h\frac{\mathrm{e}^{-v(\mathbf{r})}}{\mathcal{D}_1(\mathbf{r})} = -l_Y\rho_p(\mathbf{r}) \tag{35}$$

2.5. The Densities or Water Dipoles, Ions, and Hydrophobic Probes. Once the dimensionless fields $u^{MF}(\mathbf{r})$ and $v^{MF}(\mathbf{r})$ have been derived as solutions of the HDPBL system of equations, the densities of the various molecules are given by

(i) Anions and cations:

$$\rho_{\pm}(\mathbf{r}) = \frac{1}{a^3}\frac{\Phi_s e^{\mp z u^{MF}(\mathbf{r})}}{2\mathcal{D}_1^{MF}(\mathbf{r})} \tag{36}$$

(ii) Salt:

$$\rho_s(\mathbf{r}) = \frac{1}{a^3}\frac{\Phi_s\cosh(z u^{MF}(\mathbf{r}))}{\mathcal{D}_1^{MF}(\mathbf{r})} \tag{37}$$

(iii) Water dipoles:

$$\rho_w(\mathbf{r}) = \frac{1}{a^3}\frac{\Phi_w}{\mathcal{D}_1^{MF}(\mathbf{r})}\frac{\sinh(p_e|\nabla u^{MF}(\mathbf{r})|)}{p_e|\nabla u(\mathbf{r})|} \tag{38}$$

(iv) Hydrophobic Probes:

$$\rho_h(\mathbf{r}) = \frac{1}{a^3}\frac{\Phi_h}{\mathcal{D}_1^{MF}(\mathbf{r})}\mathrm{e}^{-v(\mathbf{r})} \tag{39}$$

where $\mathcal{D}_1^{\mathrm{MF}}(\mathbf{r}) = \Phi_v + \Phi_s\cosh(\mathrm{zu}^{\mathrm{MF}}(\mathbf{r})) + \Phi_w\frac{\sinh(p_e|\nabla u^{\mathrm{MF}}(\mathbf{r})|)}{p_e|\nabla u^{\mathrm{MF}}(\mathbf{r})|}\cdot + \Phi_h\mathrm{e}^{-v^{\mathrm{MF}}(\mathbf{r})}$

Equation 38 defines the local densities of water around the solute. In parallel, we can also compute the polarization density

$$\mathbf{P}(\mathbf{r}) = \frac{p_0}{a^3}\Phi_w\frac{g(p_0\beta|\mathbf{E}^{MF}(\mathbf{r}|)}{\mathcal{D}_1^{MF}(\mathbf{r})}\hat{\mathbf{E}}^{MF}(\mathbf{r}) \tag{40}$$

where $\hat{\mathbf{E}}$ is the unit vector of the electric field $\mathbf{E}$. Using the expression for the water density, we have

$$\mathbf{P}(\mathbf{r}) = \rho_w(\mathbf{r})\mathcal{L}(p_0\beta|\mathbf{E}^{MF}(\mathbf{r})|)\hat{\mathbf{E}}^{MF}(\mathbf{r}) \tag{41}$$

where $\mathcal{L}$ is the usual Langevin function (see above). For small electric field $\mathbf{E}$, eq 41 becomes the standard linear relation

$$\mathbf{P}(\mathbf{r}) = \frac{{p_0}^2\beta c_w}{3}\mathbf{E}^{MF}(\mathbf{r}) = \alpha\mathbf{E}^{MF}(\mathbf{r}) \tag{42}$$

Note that if we use the expression for the polarization density $\mathbf{P}(\mathbf{r})$, we can rewrite the DPBL-like eq 28 as

$$\nabla\cdot(\varepsilon_0\mathbf{E}(\mathbf{r}) + \mathbf{P}(\mathbf{r})) = \rho_f(\mathbf{r}) + \rho_{ions}(\mathbf{r}) \tag{43}$$

3. NUMERICAL SOLUTIONS TO THE HDPBL SYSTEM OF EQUATIONS

Let us first recall the system of equations HDPBL:

$$-\nabla\cdot(\varepsilon(\mathbf{r}, u, v)\nabla u(\mathbf{r})) + \gamma(\mathbf{r})C_s\Phi_s\frac{\sinh(\mathrm{zu}(\mathbf{r}))}{\mathcal{D}_1(\mathbf{r}, u, v)} = C_f\rho_{fd}(\mathbf{r}) \tag{44a}$$

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

$$-\nabla^2 v(\mathbf{r}) + \kappa^2(v(\mathbf{r}) + v_0) + \gamma(\mathbf{r})\frac{l_Y}{a^3}\Phi_h\frac{e^{-v(\mathbf{r})}}{\mathcal{D}_1(\mathbf{r}, u, v)} = -l_Y\rho_p(\mathbf{r}) \tag{44b}$$

where $\mathcal{D}_1(\mathbf{r}, u, v)$ and $\varepsilon(\mathbf{r},u,v)$ are functions of the position, $\mathbf{r}$, and of the fields u and v:

$$\varepsilon(\mathbf{r}, u, v) = 1 + \gamma(\mathbf{r})C_w\Phi_w\frac{F(p_e|\nabla u(\mathbf{r})|)}{\mathcal{D}_1(\mathbf{r}, u, v)}$$

$$\mathcal{D}_1(\mathbf{r}, u, v) = \Phi_v + \Phi_s\cosh(zu(\mathbf{r})) + \Phi_w\frac{\sinh(p_e|\nabla u(\mathbf{r})|)}{p_e|\nabla u(\mathbf{r})|} + \Phi_h e^{-v(\mathbf{r})} \tag{45}$$

The two equations in this system are dependent as they both involve the two fields u and v. They are PB-like, but they cannot be solved directly using a PB solver. Equation 44 for example is a second order elliptic nonlinear PDE, like PB; however, its response coefficients (the coefficients in the divergence term) are not constant, as they are nonlinear functions of the two fields u and v. For eq 44, it is the Helmholtz term that is a nonlinear function of the two fields u and v. Instead of considering a new specific solver for each of those equations, we propose to use a standard inexact Newton method developed for the PB equation by Hold and Saied[42] within a self-consistent algorithm for solving the HDPBL system, as described in Algorithm 1.

Algorithm 1 *AquaVit: Self Consistent Newton method for solving the DPBL equation*

Initialize $u_0(\mathbf{r}) = 0$ and $v_0(\mathbf{r}) = 0$, $\forall\mathbf{r}$

for $n = 0, \ldots$ until convergence **do**

(1) Solve equation 44a for $u(\mathbf{r})$ with $v(\mathbf{r})$ fixed

(1a) Initialize a field $\psi(\mathbf{r}) = 0, \forall\mathbf{r}$;
Define $F(\mathbf{r}, \psi) = -\nabla\cdot(\varepsilon(\mathbf{r}, \psi, v_n)\nabla\psi(\mathbf{r})) + \gamma(\mathbf{r})C_s\Phi_s\frac{\sinh(z\psi(\mathbf{r}))}{\mathcal{D}_1(\mathbf{r},\psi,v_n)} - C_f\rho_{fd}(\mathbf{r})$
for $m = 0, \ldots$ until convergence **do**
(1b) Set $\varepsilon_m(\mathbf{r}) = \varepsilon(\mathbf{r}, \psi_m, v_n)$
(1c) Set $\mathcal{D}_m(\mathbf{r}) = \mathcal{D}_1(\mathbf{r}, \psi_m, v_n)$
Define $H_m(\mathbf{r}, \psi) = \gamma(\mathbf{r})C_s\Phi_s\frac{\sinh(z\psi(\mathbf{r}))}{\mathcal{D}_m(\mathbf{r})}$
(1d) Solve the PB-like PDE:
$\nabla\cdot(\varepsilon_n(\mathbf{r})\nabla\psi(\mathbf{r})) + H_n(\mathbf{r}, \psi(\mathbf{r})) = C_f\rho_{fd}(\mathbf{r})$
for ψ_{sol}, using the inexact Newton method of Holst and Saied[42]
(1e) Update ψ:
$\psi_{m+1}(\mathbf{r}) = \lambda\psi_{sol}(\mathbf{r}) + (1-\lambda)\psi_m(\mathbf{r}), \quad \forall\mathbf{r}$
(1f) Check for convergence: if $\frac{\sum_{\mathbf{r}}\|\mathbf{F}(\mathbf{r}, \psi_{m+1}\|}{\sum_{\mathbf{r}}\|\mathbf{F}(\mathbf{r}, \psi_0)\|} < TOL$, stop
end for
(1g) Update $u_{n+1}(\mathbf{r}) = \psi_{m+1}(\mathbf{r}), \quad \forall\mathbf{r}$

(2) Solve equation 44b for $v(\mathbf{r})$ with $u(\mathbf{r})$ fixed

(2a) Initialize a field $\phi(\mathbf{r}) = 0, \forall\mathbf{r}$;
(2b) Define $G(\mathbf{r}, \phi) = \kappa^2(\phi(\mathbf{r}) + v_0) + \gamma(\mathbf{r})\frac{l_Y}{a^3}\Phi_h\frac{e^{-\phi(\mathbf{r})}}{\mathcal{D}_1(\mathbf{r}, u_{n+1}, \phi)}$
(2c) Solve the PB-like PDE:
$-\nabla^2 v(\mathbf{r}) + G(\mathbf{r}, \phi(\mathbf{r})) = -l_Y\rho_p(\mathbf{r})$
for ϕ_{sol}, using the inexact Newton method of Holst and Saied[42]
(2d) Update $v_{n+1}(\mathbf{r}) = \lambda\phi_{sol}(\mathbf{r}) + (1-\lambda)v_n(\mathbf{r}), \quad \forall\mathbf{r}$
(3) Check for convergence: if $\sum_{\mathbf{r}}|v_{n+1}(\mathbf{r}) - v_n(\mathbf{r})| < TOL$, stop
end for

This algorithm alternatively solves for u (part 1), given v, and then solves for v (part 2), given u, until convergence, i.e., until those fields do not change anymore. To solve for u, step 1b sets the diffusion coefficient ε_n independent of the fields u and v. Similarly, step 1c defines a Helmholtz-like term H_m whose value at position $\mathbf{r}$ only depends on the value of the potential at that position. The PDE in step 1d is then a PB equation that can be solved directly with a PB algorithm without modification. The updates in steps 1e and 2d follow a typical trick for self-consistent methods that removes oscillations in the convergence behavior. Note that here is no need to solve the PDE in step 1d exactly. As its solution $\psi(\mathbf{r})$ is used as an correction for the solution of the DPBL equation (step 1e), it is appropriate to use an approximation: the number of total iterations may then be higher, but this is compensated for by the fact that the amount of work per iteration is smaller.

The algorithm described above was implemented in a software package, AquaVit. AquaVit is written in Fortran and is designed specifically to solve the HDPBL system of equations. AquaVit is mostly inspired from AquaSol,[40] and it uses many routines from the Fortran package MG developed by Michael Holst.[41] More information on the implementation is available in the paper describing Aquasol in length.[40]

4. METHODS

4.1. System Setup. The coordinates of the atoms of the solute(s) as well as their partial charges are read from a single file under the PQR format used by APBS. For large biomolecules, PQR files can be readily generated from the correspondent PDB[43] files using the service PDB 2PQR.[44] For all examples described below, we used the PARSE parameter data set[45] to assign charges. The PQR file may contain several molecules. The PQR file was then modified to add hydrophobic charges to selected subsets of atoms. While those are not charges per se, they enable interactions between the hydrophobic probes in the solvent and the solute. Using the nomenclature of CHARMM, all atoms identified as CT (aliphatic carbon), CH1E (extended carbon, with one hydrogen), CH2E (extended carbon, with two hydrogens), CH3E (extended carbon, with three hydrogens), CR1E (ring carbons, such as those found in the side chains of Phe, Tyr, and Trp residues), S (sulfur), and SH1E (extended atom S with one hydrogen) were assigned a nonzero hydrophobic charge of +2, while all other atoms have no nonpolar charges. Note that these choices are considered as the default and can easily be changed.

AquaVit starts by building a regular mesh around the solutes. The mesh is positioned such that its center matches with the center of the solute. The user provides the number of points and the mesh spacing in each directions. AquaVit checks that there is at least a distance of $2l_B^w$, (l_B^w being the Bjerrum length in water at 300 K, i.e., approximately 7 Å) from any point on the surface of the solute to the closest edge of the mesh; if this condition is not met, the mesh size is adjusted accordingly.

The interface between the interior and exterior of the solutes is modeled based on their molecular surface. The molecular surface is the lower envelope obtained by rolling a water probe of radius R_{probe} on the vdW surface of the molecule. A full description of how this surface is computed can be found in ref 40.

Classical treatment of electrostatics assigns a point charge to each atom, usually located at the center of the sphere representing this atom. The mesh considered in AquaVit is Cartesian; as such, the centers of the atoms of the solute(s) will most likely not coincide with its vertices. We therefore need to project the atomic charges on the vertices of the mesh; we have used trilinear interpolation.

4.2. Parameterizing the HDPBL System. The HDPBL system of equations include 10 parameters: the number of vertices in the mesh in each dimension, the lattice size a, the temperature T, the concentrations of water c_w, salt, c_s, and hydrophobic probes, c_h, the valence z of the anions and cations

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

from the salt, the strength of the water dipole, p_0, and the parameters of the Yukawa potential, l_Y (=β w_0) and l_h (= $1/\kappa$).

The mesh was set with 193 vertices in each direction, x, y, and z. Those vertices are equally spaced, and the distance between two vertices is computed automatically based on the size of the solute and the fact that the borders of the mesh are set to be at least $2l_B^w$ away from the solute.

Assuming incompressibility, in the presence of pure water, we expect $\Phi_w = 1$, i.e., that $c_w a^3 = 1$, where c_w is the concentration of bulk water, namely 55 M. This leads to a = 3.11. In all the simulations described in the Results and Discussion, we have considered monovalent (i.e., z = 1) salt at 0.2 M and hydrophobic probes at 1 M. This amount of hydrophobic probes is equivalent to the amount used in ligand mapping molecular dynamics programs such as SILCS[46] and SWISH.[47] Again, assuming incompressibility, the concentration of water is then fixed as we have the relation (see above):

$$2c_s a^3 + c_w a^3 + c_h a^3 = 1$$

Using the prescribed concentrations of salt and hydrophobic probes, and the lattice size a = 3.11, this gives us an apparent concentration of water c_w = 53.6 M.

The parameter l_h defines the range of the Yukawa potential. We have set it equal to the size of the lattice: i.e., l_h = 3.1 Å. l_Y is a characteristic length for the Yukawa potential that directly relates to its strength. We have set it to l_Y = 4 Å.

The temperature is set to 300 K.

The experimental dipole moment of water is 1.85D in the gas phase. We have observed previously that with this value for p_0, we could not obtain the correct dielectric permittivity of water in the DPBL model.[38] As HDPBL is based on DPBL, we follow the recommendation of increasing p_0, which we set at 2.8 D.

In Table 1, we give the values of the three constants C_w, C_s, and C_f and p_e for a typical run of AquaVit with T = 300 K and the values of the different parameters given above.

Table 1. Typical Values for the Constants in eq 28

name	expression	value[a]	unit
l_B	$\frac{\beta e_c^2}{4\pi\varepsilon_0}$	556.99995	Å
C_w	$\frac{4\pi l_B p_e^2}{a^3}$	78.51664	dimensionless
C_s	$\frac{z4\pi l_B}{a^3}$	231.83464	Å^{-2}
C_f	$4\pi l_B$	6999.46779	Å
p_e	$\frac{p_0}{e_c}$	0.58195	Å

[a] T = 300K, a = 3.11 Å (size of the lattice), p_D = 2.8 D (dipole moment of water), and z = 1 (valence of salt ions).

4.3. Output Format. The output of AquaVit are the maps corresponding to the densities of the different species in the solvent, namely the water dipoles, the anions, the cations, and the hydrophobic probes. While those maps are initially expressed as relative densities with respect to the bulk concentrations of the respective species, we have re-expressed those maps as Z-maps, using

$$Z(\mathbf{r}) = \frac{\rho(\mathbf{r}) - \mu_\rho}{\sigma_\rho}$$

where μ_ρ and σ_ρ are the mean and standard deviation of the densities computed over all positions $\mathbf{r}$ outside of the solute, respectively.

The choice of Z-maps instead of raw maps is motivated by the popular use of such maps when representing electron density maps in X-ray crystallography. We do note however that our maps are different. First, we consider four types of coexisting maps, based on concentrations of hydrophobic probes, anions, cations, and water, respectively. The bulk values for those species, which usually correspond to the mean values μ_ρ defined above, differ significantly: from 0.2 M for anions and cations to 55 M for water. The fluctuations around those mean values, namely the σ_ρ, do differ also significantly. In addition, as we assume incompressibility, the concentrations and therefore the Z-values for the different species are not independent. As such, meaningful values for those maps, i.e., values that differ significantly from bulk, may be found at different σ levels. We have not been able to provide a general framework for defining those levels, which are then defined through trial and error. This will be discussed in the next section.

5. RESULTS AND DISCUSSION

HDPBL differs from a standard PB model, as we have included hydrophobic probes around the solute of interest. Our primary goal is to use those probes to detect hydrophobic pockets in this solute. We have therefore tested the ability of HDPBL to detect pockets that bind to hydrophobic ligands as well as to characterize the environment of membrane proteins. HDPBL also provides the densities of anions, cations, and water dipole around the solute. We have included a test on detecting binding sites characterized with electrostatics interactions to illustrate their usefulness. Finally, we will describe limitations of HDBPL when it comes to detecting cryptic binding sites.

5.1. Detecting Hydrophobic Pockets in Proteins. The set of proteins used for validation includes four proteins from different families and functions: a dihydrofolate reductase (DHFR) from *Staphylococcus aureus* (PDB code 2W9H[48]), a maize lipid-transfer protein (PDB code 1FK2[49]), a human retinol binding protein 1 (PDB code 5HBS[50]), and an insect Takeout 1 protein (PDB code 3E8W[51]) . The PDB structures correspond to those proteins bound to a hydrophobic ligand. All ligands, ions, and crystallographic water molecules, however, were removed prior to running AquaVit. In Figures 1 and 2, we show the resulting hydrophobic probe occupancy maps superimposed on the PDB structures for all four proteins, with and without the ligand.

DHFR is the enzyme responsible for the NADPH-dependent reduction of dihydrofolate to tetrahydrofolate, an essential cofactor in the synthesis of purines, methionine, and other key metabolites. Because of its importance in a wide range of cellular functions, DHFR has been the subject of much research targeting the enzyme for anticancer, antibacterial, and antimicrobial agents. Clinically used compounds targeting DHFR include methotrexate for the treatment of cancer and trimethoprim (TMP) for the treatment of bacterial infections. The active site of DHFR is comprised of a large hydrophobic pocket which serves as the folate-binding site. This pocket was successfully detected by AquaVit, as illustrated in Figure 1a. This high density region of hydrophobic probes overlaps well with the TMP ligand that was cocrystallized with DHFR and bound within the hydrophobic active site[48] (see Figure 1b).

Lipid binding proteins (LBP) facilitate the transfer of lipids between membranes. We consider a nonspecific LBP from

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

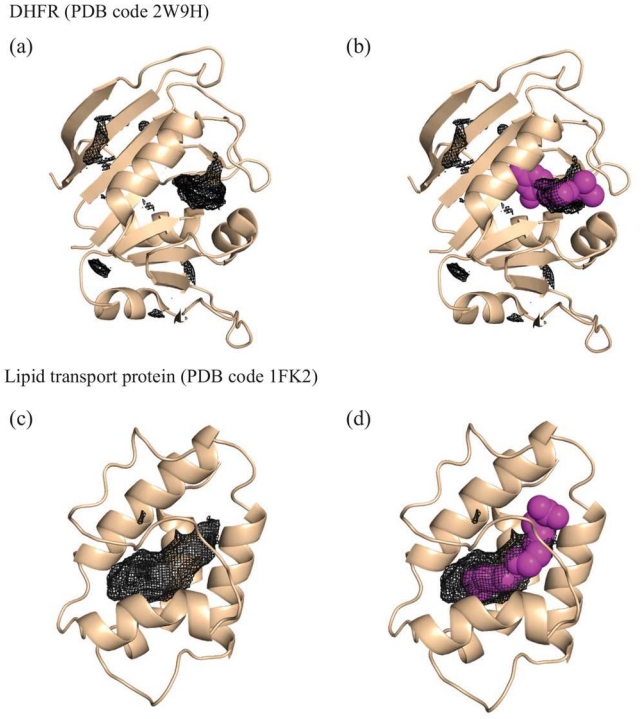


Figure 1. Hydrophobic occupancy maps, black mesh, superimposed onto the PDB structures for DHFR (PDB code 2W9H) and a lipid binding protein (LBP, PDB code 1FK2). The maps are derived from the densities of hydrophobic probes computed by AquaVit, and represented at +20 σ. Those maps are derived from the apo structure of the protein, i.e. in the absence of all crystallographic ligands and water molecules. In panels a and c, we show the hydrophobic maps for DHFR and LBP superposed to the apo PDB structure, while in panels b and d, we visualize the hydrophobic ligand (trimethoprim for DHFR and myristic acid for LBD) in magenta. Note that all images in the figure and in the subsequent figures were generated using Pymol.[52]

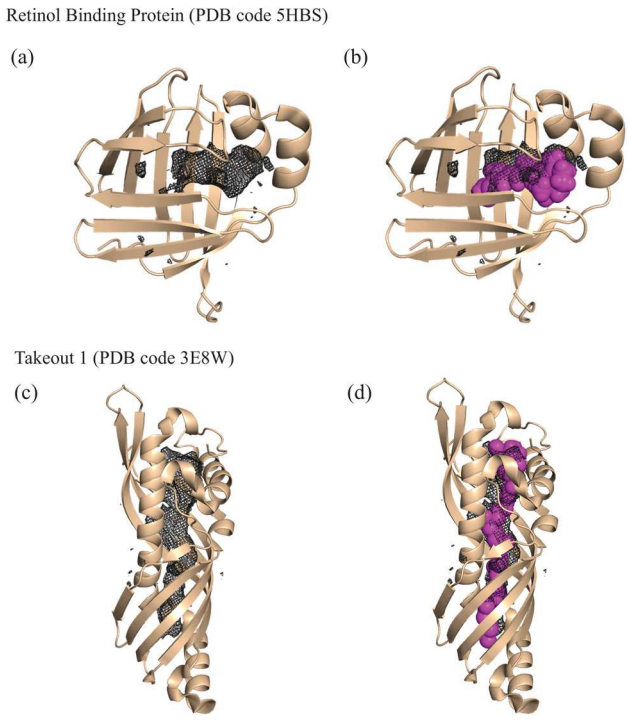


Figure 2. Hydrophobic occupancy maps, black mesh, superimposed onto the PDB structures for a retinol binding protein (RBP, PDB code 5HBS) and a TakeOut 1 protein (To1, PDB code 3E8W). The maps are derived from the densities of hydrophobic probes computed by AquaVit, and represented at +20 σ. Those maps are derived from the apo structure of the protein, i.e. in the absence of all crystallographic ligands and water molecules. In panels a and c, we show the hydrophobic maps for RBP and To1 superposed to the apo PDB structure, while in panels b and d, we visualize the hydrophobic ligand (trans retinol for RBP and 8-ubiquinone for To1) in magenta.

maize, whose hydrophobic cavity can accommodate various lipids from C10 to C18. AquaVit was successful in identifying this cavity, as illustrated in Figure 1c. Interestingly, as we superimpose the ligand found in the PDB structure, myristic acid, onto the hydrophobic occupancy map, we find that the high density region of hydrophobic probes identified by AquaVit extends beyond this ligand (see Figure 1d). Myristic acid is C14, while the pocket found by AquaVit indicates that the hydrophobic cavity of the maize LBP can accommodate larger ligands.[49]

The cellular retinol-binding protein 1(CRBP1) is another example of a protein with a large hydrophobic active site. CRBP1 is important in regulating the uptake, storage and metabolism of retinoids (vitamin A and its derivatives). Its vitamin A binding site is lined with hydrophobic residues that are well conserved among retinol binding proteins. Those residues provide the nonpolar interactions that stabilize the retinoid ligand. This pocket was successfully detected by AquaVit, as illustrated in Figure 2a. The high density region of hydrophobic probes overlaps well with the all trans retinol ligand that was cocrystallized with CRBP1,[50] as observed in Figure 2b.

Takeout (To) proteins are found exclusively in insects, in which they have been proposed to play important roles in their physiology and behavior.[53] Of particular interest to us, they have been suggested to bind to hydrophobic ligands. We considered the To 1 protein from *Epiphyas postvittana*, a light brown apple moth, whose structure was solved by crystallography at 1.3 Å[51] in the presence of ubiquinone-8. The crystal structure revealed a 45 Å long hydrophobic internal tunnel that extends to the full length of the protein. This pocket was successfully detected by AquaVit, as illustrated in Figure 2c. Note that the same pocket was originally characterized based on geometry only.[53] AquaVit provides the additional information that this pocket is amenable to interaction with an hydrophobic ligand. This is confirmed as the high density region of hydrophobic probes overlaps well with the hydrophobic ubiquinone ligand; see Figure 2d.

5.1.1. Comparison with FTMAP. FTMAP[12,13] is another computational mapping technique that identifies binding hot spots in proteins that bears some similarity with the HDPBL model. It performs a global search of the entire protein surface for regions that can potentially bind a number of small organic probe molecules (currently FTMAP considers 16 different small probes, ethanol, 2-propanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and *N*,*N*-dimethylformamide). The search takes into account global translational and rotational degrees of freedom for the probe (i.e., internal degrees of freedom are ignored). Each position is scored using an energy function based on vdW interactions, a cavity term to account for the hydrophobic environment, a statistical potential that indirectly includes solvent effects, and electrostatics computed using the Poisson−Boltzmann formalism.[12] Favorable positions are then clustered, and overlapping

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

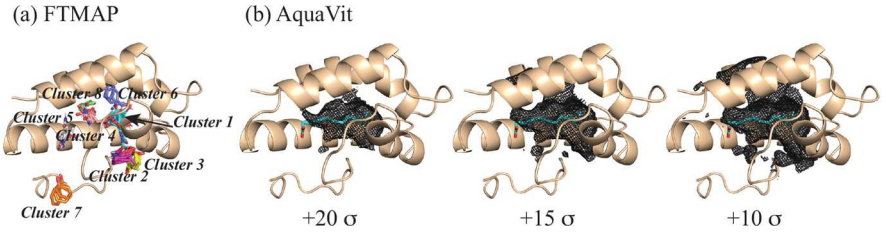


Figure 3. Predicting the ligand binding side of a lipid binding protein (LBP, PDB code1FK2). (a) Results of the prediction generated by FTMAP.[12] The centers of the eight top probe clusters are superimposed onto the structure of LBP. The probes are color-coded to distinguish between the different clusters. (b) Hydrophobic occupancy maps computed with AquaVit, black mesh, superimposed onto the PDB structure for LBP. The maps are represented at +20 σ, +15 σ, and +10 σ, from left to right. Myristic acid (the ligand found in 1FK2) is shown as sticks with carbon atoms colored yellow.

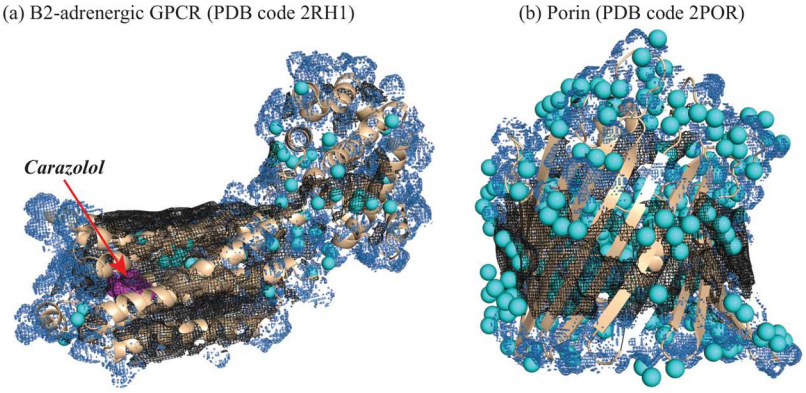


Figure 4. Hydrophobic occupancy maps, black mesh, and water dipole occupancy maps, blue mesh superimposed onto the PDB structures for (a) the β_2 adrenergic GPCR (PDB code 2RH1) and (b) a porin (PDB code 2POR). The crystallographic waters are shown as cyan spheres. The maps are derived from the densities of hydrophobic probes and water dipoles computed by AquaVit, respectively, and represented at +10 σ for the hydrophobic probes and at +0.3 σ for the water dipoles. Those maps are derived from the apo structure of the proteins, i.e., in the absence of all crystallographic ligands and water molecules.

clusters of different probes are then defined as putative binding sites. We compared the predictions of FTMAP with those of AquaVit on the maize lipid-transfer protein considered above, with PDB code 1FK2. We used the FTMAP server at https://ftmap.bu.edu/, with default values for all its parameters. Results are presented in Figure 3.

FTMAP predicted eight clusters of hydrophobic probes on the surface and on the cavity of 1FK2 (Figure 3a). Clusters 1, 4, 6, 8, and to some extent cluster 5 overlap with the position of the ligand, myristic acid, found in 1FK2. However, clusters 2, 3, and 7 are found in the neighborhood of the relatively unstructured C-terminal region of 1FK2; those clusters can be considered as being false positive. In contrast, AquaVit identifies one major hydrophobic region within 1FK2 in which the highest concentrations of hydrophobic probes (identified at the highest σ cutoff, + 20) are found in the close vicinity of the ligand myristic acid. At cutoffs lower than those shown in Figure 3b), the hydrophobic pocket keeps increasing in size and auxiliary pockets appear (results not shown). At this stage, the right cutoff for visualizing those highest densities is found by trial and error. We are working on designing a more automatic method for identifying meaningful cutoffs.

5.2. Environments of Membrane Proteins. The subsection above illustrates that AquaVit is able to locate hydrophobic pockets in proteins. However, the solution of the HDPBL system of equations is more comprehensive and also provides information on dipolar density in the presence of a given salt concentration. Here we assess its ability to characterize the polar and nonpolar environments of membrane proteins. We consider two types of such protein, a member of the G protein coupled receptor (GPCR) family, with a transmembrane domain that consists of a helical bundle (PDB code 2RH1[54]), and a porin that consist of a β-pleated sheet (PDB code 2POR[55]). All ligands, ions, and crystallographic water molecules were removed prior to running AquaVit on those structures. In Figure 4, we show the resulting hydrophobic probe occupancy maps and water dipole occupancy maps, superimposed on the PDB structures, including the crystallographic water molecules.

β_2-Adrenergic receptors (β_2AR) are members of the GPCR family that reside predominantly in smooth muscle. Their antagonists are used in particular in the treatment of asthma.[56] To study the structure of this membrane protein, Cherezov et al. designed studied a chimera consisting of β_2AR and the T4 lysozyme (T4L).[54] The crystal structure of this chimera reveals a fold for β_2AR composed of a transmembrane domain with a 7 helix bundle, and a standard all-helix fold for the T4L. Interactions between β_2AR and T4L are minimal. AquaVit provides a consistent image of the environment of this chimera, with a mostly hydrophobic environment for the transmembrane domain, and a mostly hydrophilic environment for T4L (see Figure 4a). Interestingly, the water dipole occupancy map superimposes well with the crystallographic water molecules detected in the structure.

Porins are integral membrane proteins that are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. The crystal structure of the porin from *Rhodobacter capsulatus* reveals a 16-stranded β-barrel, with all β-strands antiparallel and connected to their neighbors.[55] AquaVit reveals

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

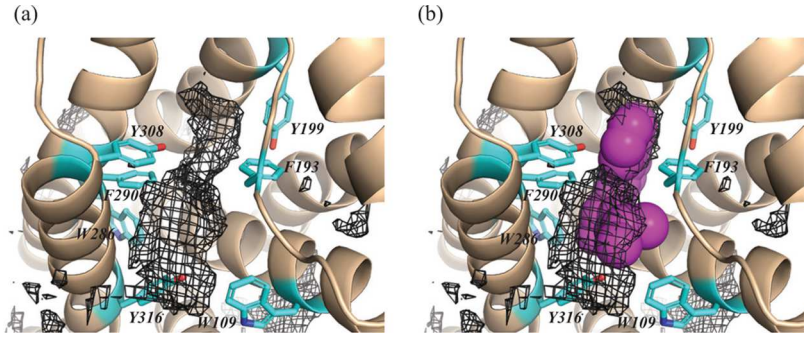


Figure 5. Hydrophobic occupancy map, black mesh, superimposed onto the PDB structure of β_2AR (PDB code 2RH1). The map is derived from the densities of hydrophobic probes computed by AquaVit, and represented at +20 σ. In panel a, we show the hydrophobic map superposed to the apo PDB structure, while in panel b, we visualize the hydrophobic ligand (carazozol) in magenta. The main residues of β_2AR contributing to carazolol binding are shown in stick mode with carbons colored in cyan and are labeled.

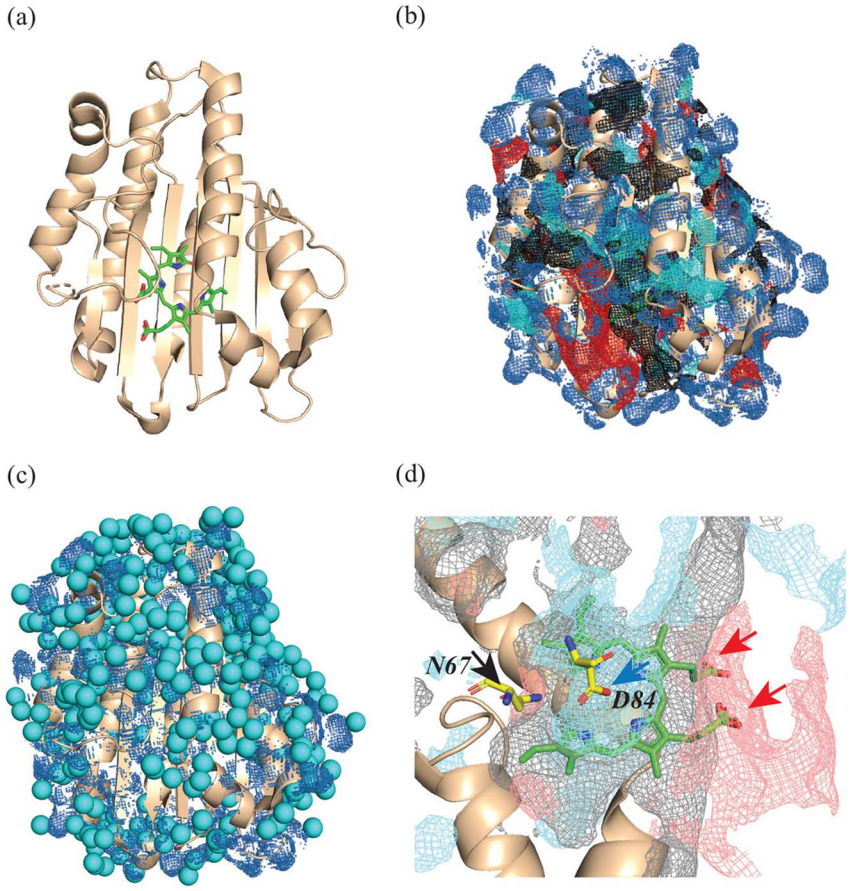


Figure 6. (a) Overall structure of PcbA, a ferredoxin-dependent bilin reductase (cartoon mode) in the presence of its substrate, biliverdin (BV): PDB code 2X9O.[59] (b) Superimposition of the occupancy maps of water (+0.3 σ, blue), hydrophobic probes (+10 σ, gray), anion (+20 σ, red), and cation (+20 σ, cyan) on the structure of PcbA, in the absence of BV. (c) PcbA structure, the water occupancy map, and the crystallographic water molecules (shown as blue spheres). (d) Superposition of the occupancy maps of hydrophobic probes (gray), anion (red), and cation (cyan) with the structure of BV (sticks). We also show the position of residues Asp84 and Asn67 that are known to play an important role in placing the ligand in its binding pocket.[59] Note the presence of anions density overlapping with the carboxyl groups of BV (red arrows) and forming a pocket in front of the N of the terminal group of Asn67. Note also the presence of cations around the carboxyl group of Asp84 (blue arrow).

a hydrophobic environment on the outside of the β-pleated sheet, with the loops between the strands in a more hydrophilic environment (Figure 4b). This is consistent with the positions of the crystallographic water molecules that map with the water dipole density map.

5.2.1. Ligand Binding in β_2AR. The crystal structure of β_2AR was solved in the presence of a partial inverse agonist, carazolol, which binds in an hydrophobic pocket of the transmembrane region of the protein.[54] In Figure 4, we showed a density of hydrophobic probes in the surrounding of the transmembrane region of β_2AR, when the density map is represented at +10 σ. When the density map for those hydrophobic probes is displayed at a higher cutoff (+20 σ), we observed one main

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

region of high density within the transmembrane region (Figure 5a) that corresponds to the position of carazozol (Figure 5b).

5.3. Multiprobe Analysis of a Complex Active Site. All examples presented above focused on the identification of hydrophobic pockets and characterization of the hydrophobic environment of proteins, based on the analyses of the hydrophobic occupancy maps generated by AquaVit. However, much akin to MixMD[57] and mLMMD,[58] AquaVit can be seen as a multiprobe analysis of the environment of a protein. It generates the densities of water dipole, anion, cation, and hydrophobic molecules surrounding the protein that serve as probes to identify and characterize pockets in the protein. We use the ferredoxin oxidoreductase system to illustrate this functionality of AquaVit.

Ferredoxin-dependent bilin reductases (FBDR) are enzymes that are involved in the reduction of biliverdin (BV) to form phycobilins used for light-perception or light harvesting in plants and cyanobacteria.[60] Several members of this family have been identified in multiple species.[61] Among those, PebA reduces BV at its C15−C16 double bond to produce 15−16 dihydrobiliverdin (DHBV). The structure of PebA from the cyanobacterium *Synechococcus* sp. WH8020 with its substrate BV was determined at 1.55 Å (PDB code 2X9O[59]). This structure consists of a seven-stranded antiparallel β-sheet surrounded by six α helices (Figure 6a). We used AquaVit to solve the HDPBL system of equations that characterize the environment of PeBA. The calculation was performed on the protein structure alone, i.e., in the absence of the BV ligand and crystallographic water molecules. Figure 6b illustrates the densities of water dipoles, anions, cations, and hydrophobic probes around the structure of PebA in the form of occupancy maps. The whole protein is surrounded with water dipoles. This is in agreement with the fact that many crystallographic water molecules have been identified. Those molecules superimpose well with the water dipole occupancy map Figure 6c. Of significant interest are the anion, cation, and hydrophobic probe occupancy maps in the active site of PeBA; those maps are illustrated in Figure 6d. The superposition of the ligand structure and of the two residues (Asp84 and Asn67) that define the central polar centering "pin" of the binding pocket[59] on those maps shows that the anion densities (in red) map well with the two carboxyl groups on BV (red arrows). We also observe an anion pocket in front of the nitrogen of the terminal group of Asn67 (black arrow). Note that the densities are expected to match with the chemical structure of the ligand, and be complementary to the chemical properties of the solute. The hydrophobic probe density map superimposes well with the hydrophobic parts of BV. Finally, we observe a cation occupancy in the region surrounding the four nitrogen of the rings of BV, as well as a pocket in from of the carboxyl moiety of Asp84 (blue arrow). These observations highlight the advantage of AquaVit to account for the multiple species that form the environment of a protein.

5.4. Computing Time. In Figure 7 we report the computing times for solving the HDPBL system of equations using AquaVit for three systems, DHFR (PDB code 2W9H), GPCR (PDB code 2RH1), and PEBA (PDB code 2X9O) under the standard conditions defined in the Methods. In addition to the inputs specific to the system under consideration (structure of the solute, concentrations of water dipoles, salt, and hydrophobic probes, dipole moment of the water dipole, and parameters of the Yukawa potential for hydrophobic interactions), AquaVit relies on the parameters of Algorithm 1 to solve the HDPBL system. These parameters include on the size of the Cartesian

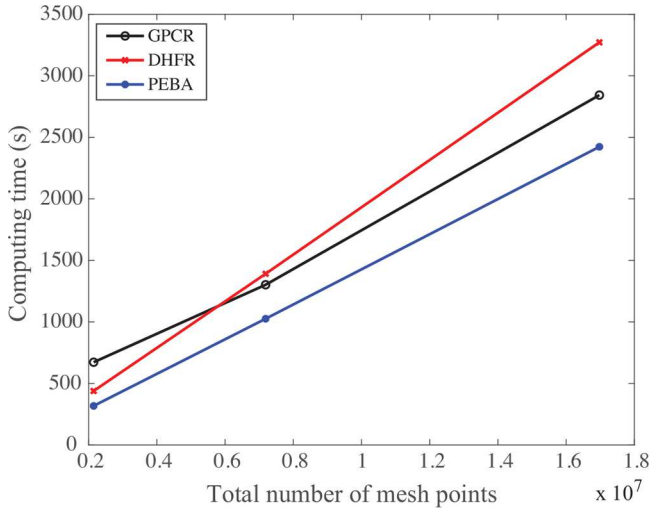


Figure 7. Computing time required to solve the system of equations HDPBL using AquaVit with a residual error lower than TOL = 10^{-4} (see text for details) is plotted versus the total number of points in the mesh for the three proteins DHFR (PDB code 2W9H), GPCR (PDB code 2RH1), and PEBA (PDB code 2X9O). All computations were performed on an Intel Core i7 processor with 8 cores running at 4.00 GHz, and 64GB of memory, although AquaVit was compiled without parallelization options.

grid used to solve the PDEs, the tolerance TOL that serves as stopping criteria when solving the system self-consistently, and the parameter λ used to update the fields u and v (see Algorithm 1 presented in section 3). In all calculations presented above, TOL is set to 10^{-4} and λ is set to 0.7 (this number is somewhat arbitrary and could in fact be optimized for each system considered). We have also used consistently a grid of size 193^3, but in Figure 7, we report the computing time of AquaVit as a function of this grid size. We observe a near linear dependence of the computing times of AquaVit with respect to the total number of points in the grid, as expected.[40] These computing time do not differ significantly between the three proteins, despite the fact that they are very different in size (157 residues for 2W9H, 237 residues for 2X9O, and 465 residues for 2RH1). The average computing time for a grid of 193^3 points is 1240 s, i.e., approximately 21 min. We note that the computing times reported are both CPU and clock time; i.e., AquaVit ran on a single core and does not benefit from parallelization. We acknowledge that there is much room for improvement in the implementation of Algorithm 1 in AquaVit.

5.5. Limitations of the HDBPL Model: Absence of Flexibility. We have shown above that the HDPBL model enables the discovery and characterization of binding sites on proteins. In addition, HDPBL is fast (see Figure 7) or at least competitive with respect to computing time compared to the ligand mapping molecular dynamics simulations. It has one major limitation, however, as it considers the structure of the protein to be static. While it was not an issue in the examples described in the previous sections, we illustrate here a case in which dynamics matter.

Polo-like kinases (PLKs) are a family of serine/threonine kinases related to the polo gene product of *Drosophila melanogaster*. Most PLKs have multiple functions which map with their organization in domains. Their C-terminal regions, for example, contain the polo box domain (PBD), which helps in their subcellular localization by binding to serine- or threonine-

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

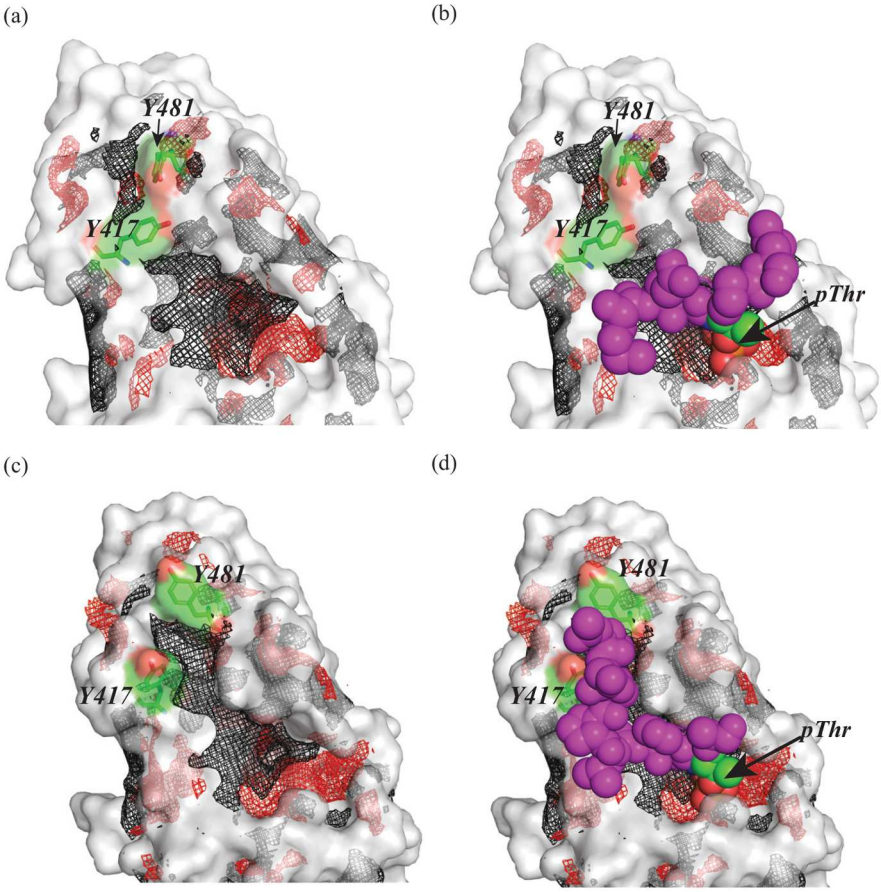


Figure 8. (a) Superimposition of the occupancy maps of hydrophobic probes (+20 σ, black) and anions (+20 σ, red) on the structure of the polo binding domain (PBD) of human polo like kinase 1 (PLK1) (PDB code 1Q4K). The occupancy maps are computed using AquaVit, on the PBD structure in the absence of the ligand and crystallographic water molecules. (b) Same as in part a, but with the ligand (i.e., the phosphopeptide MetGln *Ser*(pThr)ProLeu) shown in magenta, with the phosphorylated Thr colored according to atom type, with green for carbon, yellow for phosphate, and red for oxygen. Note that the peptide overlaps well with one of the hydrophobic pockets identified by AquaVit, with the phosphorylated group on the Thr fitting inside a pocket identified from the anion density. (c) Superimposition of the occupancy maps of hydrophobic probes (+20 σ, black) and anions (+20 σ, red) on the structure of the polo binding domain (PBD) of human polo like kinase 1 (PLK1) (PDB code 3P37). The occupancy maps are computed using AquaVit, on the PBD structure in the absence of the ligand and crystallographic water molecules. Note the longer hydrophobic pocket compared to 1Q4K, as a result of the change of conformation of Tyr417 and Tyr481. (d) Same as in part a, but with the ligand (the phosphopeptide PheAspProProLeuHisSerp(pThr)Ala) shown in magenta. The ligand overlaps well with the long hydrophobic pocket,with the phosphorylated group on the Thr fitting inside a pocket identified from the anion density.

phosphorylated sequences on target proteins. Cheng et al. have determined the structure of the PBD domain of the human PLK1 in the presence of a phosphopeptide with sequence MetGln *Ser*(pThr)ProLeu, where pThr indicates that the threonin is phosphorylated (PDB code 1Q4K[62]). The phosphopeptide was found to bind on the surface of the protein, in an hydrophobic pocket. We ran AquaVit on a single copy of the PBD domain in the absence of the peptide ligand and of all crystallographic water molecules. We can identify the ligand binding hydrophobic pocket of PBD from the resulting hydrophobic probe density map, as illustrated in parts a and b of Figure 8. Interestingly, the phosphate group of pThr is found to fit within a pocket that is identified by AquaVit as a region with high concentration of anions; see Figure 8b.

The binding of MetGln *Ser*(pThr)ProLeu onto PBD was found to induce very little conformational changes.[62] Two subsequent independent studies,[63,64] however, identified a cryptic hydrophobic binding site that is close to the original phosphate binding site identified by Cheng et al.[62] The opening of this cryptic binding site is the result of a change in the conformation of the side chains of Tyr417 and Tyr481. This opening enables PBD to bind longer phosphophopeptides. Sledz et al.[64] for example presented the structure of the complex of PBD with the phosphopeptide PheAspProProLeuHisSer-(pThr)Ala (PDB code 3P37). When we ran AquaVit on a single copy of the PBD domain in the absence of the peptide ligand and of all crystallographic waters in its configuration from the PDB code 3P37, we were able to identify the longer ligand binding hydrophobic pocket of PBD which includes the cryptic pocket. The long phosphopeptide PheAspProProLeuHisSer-(pThr)Ala fits well in the hydrophobic density, again with the phosphate group of pThr fitting inside a deep pocket identified as a region with high concentration of anions by AquaVit. This is illustrated in parts c and d of Figure 8. AquaVit, however, was not able to detect the longer hydrophobic pocket from the structure in PDB code 1Q4K, as residues Tyr417 and Tyr481 were prohibiting access to the cryptic pocket.

6. CONCLUSION

We have developed the HDBPL model and implemented it in the program AquaVit as a tool for identifying and characterizing

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

binding sites in proteins. A protein of interest is immersed in a lattice gas containing water dipoles, anions, cations, and hydrophobic molecules. The charged molecules interact between themselves and with the solute charges through electrostatic interactions, while the hydrophobic molecules (including the nonpolar groups on the solute) interact based on a Yukawa potential. We impose steric constraints on the lattice, as well as incompressibility; i.e., all sites of the lattice are occupied. The system (protein of interest and its environment) is then characterized with an effective free energy that depends on two fields $\phi(\mathbf{r})$ and $\psi(\mathbf{r})$ corresponding to the electrostatics and hydrophobic interactions, respectively. The Euler−Lagrange equations obtained by minimizing the free energy with respect to those two fields form a system of two Poisson−Boltzmann like PDEs, HDPBL, which we solve using a self-consistent approach, implemented in the program AquaVit. The outputs of HDPBL are the densities of the different species, and peaks of densities are expected to reveal the presence of compatible binding sites. We have tested and validated the ability of HDPBL to detect pockets that bind to hydrophobic ligands (the DHFR, a lipid binding protein, HIV protease, and a retinol binding protein), polar ligands (biliverdin), as well as to characterize the environment of membrane proteins such as GPCR and a porin.

The HDPBL equations form a system of two coupled second order elliptic nonlinear PDEs. While those equations are akin to the PB equation, they cannot be solved directly with a PB solver, mostly because the two equations are strongly dependent. We have proposed, however, an algorithm that makes use of a standard PB solver by solving those equations self-consistently. The same approach was used previously to solve the DBPL equation,[40] as well as the YULP system of equations.[39] This algorithm is relatively simple to implement and can be adapted to any PB solver, including those based on finite elements methods, which we did not consider here. We are currently developing a more versatile solver based on those methods.

In the current implementation of Aquavit, the lattice is populated with point-like electric charges, dipoles and hydrophobic molecules. One could also include finite-size dipoles made of an electric and an hydrophobic charge, as well as an electric dipole attached to a hydrophobic charge. In this case, these entities will react to an hydrophobic field $\nabla\Psi(\mathbf{r})$, using the same formalism. This could be well adapted to study the solvation of proteins in the presence of large and more complex cosolvents such as acetonitrile or DMSO.

AquaVit, our program for solving the HDPBL system of equations, is fast, and as such it compares favorably with the ligand mapping molecular dynamics simulations that have been designed with the same goal of detecting and characterizing binding sites in proteins. The latter, however, have the significant advantage that they account for the dynamics of the solute. As such, they have been shown to detect cryptic bind sites in proteins,[20,23,24,58] namely sites that are not accessible unless a structural change occur in the protein. In its current formulation, the HDPBL system of equations assumes that the protein is static; such conformational changes are then inaccessible. As an extension to HDBPL, one could model the dynamics of the solute protein with its low frequency normal modes, using for example a simple elastic model to compute those modes.[65−67] Ultimately, we want to develop AquaVit as a tool for structure-based as well as dynamics-based drug design.

AUTHOR INFORMATION

Corresponding Authors

Patrice Koehl − *Department of Computer Science and Genome Center, University of California, Davis, California 95616, United States;* orcid.org/0000-0002-0908-068X; Email: koehl@cs.ucdavis.edu

Marc Delarue − *Architecture et Dynamique des Macromolécules Biologiques, Département de Biologie Structurale et Chimie, UMR 3528 du CNRS, Institut Pasteur, 75015 Paris, France*; Email: delarue@pasteur.fr

Henri Orland − *Institut de Physique Théorique, Université Paris-Saclay, CEA, 91191 Gif/Yvette Cedex, France;* orcid.org/0000-0002-6983-2951; Email: henri.orland@ipht.fr

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c02658

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work discussed here originated from a visit by P.K. at the Institut de Physique Théorique, CEA Saclay, France, during the fall of 2019. He thanks them for their hospitality and financial support. P.K. acknowledges support from the University of California Multicampus Research Programs and Initiatives (Grant No. M21PR3267).

REFERENCES

(1) Shuker, S.; Hajduk, P.; Meadows, R.; Fesik, S. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.

(2) Mattos, C.; Ringe, D. Locating and characterizing binding sites on proteins. *Nat. Biotechnol.* **1996**, *14*, 595−599.

(3) Saur, M.; Hartshorn, M.; Dong, J.; Reeks, J.; Bunkoczi, G.; Jhoti, H.; Williams, P. Fragment-based drug discovery using cryo-EM. *Drug Discovery Today* **2020**, *25*, 485−490.

(4) Murray, C.; Verdonk, M. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 741−753.

(5) Murray, C.; Rees, D. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187−192.

(6) Henrich, S.; Salo-Ahen, O.; Huang, B.; Rippmann, F.; Cruciani, G.; Wade, R. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2009**, *23*, 209−219.

(7) Konc, J.; Janežič, D. Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.* **2014**, *25*, 34−39.

(8) Zhao, J.; Cao, Y.; Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417−426.

(9) Laurie, A.; Jackson, R. Q-SiteFinder: an energy-based method for theprediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908−1916.

(10) Goodford, P. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(11) Miranker, A.; Karplus, M. Functionality maps of binding-sites - a multiple copy simultaneous search method. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.

(12) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25*, 621−627.

(13) Kozakov, D.; Grove, L.; Hall, D.; Bohnuud, T.; Mottarella, S.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **2015**, *10*, 733−755.
(14) Jacquemard, C.; Kellenberger, E. A bright future for fragment-based drug delivery: what does it hold? *Expert Opin. Drug Discovery* **2019**, *14*, 413−416.
(15) Ivetac, A.; McCammon, J. A molecular dynamics ensemble-based approach for the mapping of druggable binding sites. *Methods Mol. Biol.* **2012**, *819*, 3−12.
(16) Feng, T.; Barakat, K. Molecular dynamics simulation and prediction of druggable binding sites. *Methods Mol. Biol.* **2018**, *1762*, 87−103.
(17) Śledź, P.; Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93−102.
(18) Bissaro, M.; Sturlese, M.; Moro, S. The rise of molecular simulations in fragment-based drug design (FBDD): an overview. *Drug Discovery Today* **2020**, *25*, 1693−1701.
(19) Basse, N.; Kaar, J.; Settanni, G.; Joerger, A.; Rutherford, T.; Fersht, A. Toward the rational design of p53-stabilizing drugs: Probing the surface of the oncogenic Y220C mutant. *Chem. Biol.* **2010**, *17*, 46−56.
(20) Tan, Y.; Sledz, P.; Lang, S.; Stubbs, C.; Spring, D.; Abell, C.; Best, R. Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of Polo-Like kinase 1. *Angew. Chem., Int. Ed.* **2012**, *51*, 10078−10081.
(21) Tan, Y. S.; Spring, D. R.; Abell, C.; Verma, C. The use of chlorobenzene as a probe molecule in molecular dynamics simulations. *J. Chem. Inf. Model.* **2014**, *54*, 1821−1827.
(22) Kalenkiewicz, A.; Grant, B.; Yang, C.-Y. Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. *Biology* **2015**, *4*, 344−366.
(23) Kimura, S.; Hu, H.; Ruvinsky, A.; Sherman, W.; Favia, A. Deciphering cryptic binding sites on proteins by mixed-solvent molecular dynamics. *J. Chem. Inf. Model.* **2017**, *57*, 1388−1401.
(24) Schmidt, D.; Boehm, M.; McClendon, C.; Torella, R.; Gohlke, H. Cosolvent-enhanced sampling and unbiased identification of cryptic pockets suitable for structure-based drug design. *J. Chem. Theory Comput.* **2019**, *15*, 3331−3343.
(25) Nerenberg, P.; Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 129−138.
(26) Huang, J.; MacKerell, A. D Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2018**, *48*, 40−48.
(27) Inakollu, V. S. S.; Geerke, D. P; Rowley, C. N; Yu, H. Polarisable force fields: what do they add in biomolecular simulations? *Curr. Opin. Struct. Biol.* **2020**, *61*, 182−190.
(28) van der Spoel, D. Systematic design of biomolecular force fields. *Curr. Opin. Struct. Biol.* **2021**, *67*, 18−24.
(29) Grochowski, P.; Trylska, J. Continuum molecular electrostatics, salt effects, and counterion binding - A review of the Poisson-Boltzmann theory and its modifications. *Biopolymers* **2008**, *89*, 93−113.
(30) Borukhov, I.; Andelman, D.; Orland, H. Steric effects in electrolytes: A modified Poisson-Boltzmann equation. *Phys. Rev. Lett.* **1997**, *79*, 435−438.
(31) Chu, V.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. Evaluation of ion binding to DNA duplexes using a size-modified Poisson-Boltzmann theory. *Biophys. J.* **2007**, *93*, 3202−3209.
(32) Xie, D.; Audi, S.; Dash, R. A size modified Poisson-Boltzmann ion channel model in a solvent of multiple ionic species: application to voltage-dependent anion channel. *J. Comput. Chem.* **2020**, *41*, 218−230.
(33) Stein, C.; Herbert, J.; Head-Gordon, M. The Poisson-Boltzmann model for implicit solvation of electrolyte solutions: Quantum chemical implementation and assessment via Sechenov coefficients. *J. Chem. Phys.* **2019**, *151*, 224111.
(34) Pang, X.; Zhou, H.-X. Poisson-Boltzmann calculations: van der Waals or molecular surface? *Commun. Comput. Phys.* **2013**, *13*, 1−12.
(35) Grant, J.; Pickup, B.; Nicholls, A. A smooth permittivity function for Poisson-Boltzmann solvation methods. *J. Comput. Chem.* **2001**, *22*, 608−640.
(36) Azuara, C.; Lindahl, E.; Koehl, P.; Orland, H.; Delarue, M. Incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. *Nucleic Acids Res.* **2006**, *34*, W38−W42.
(37) Abrashkin, A.; Andelman, D.; Orland, H. Dipolar Poisson-Boltzmann equation: ions and dipoles close to charge interfaces. *Phys. Rev. Lett.* **2007**, *99*, 77801.
(38) Azuara, C.; Orland, H.; Bon, M.; Koehl, P.; Delarue, M. Incorporating dipolar solvents with variable density in Poisson-Boltzmann electrostatics. *Biophys. J.* **2008**, *95*, 5587−5605.
(39) Koehl, P.; Orland, H.; Delarue, M. Beyond Poisson-Boltzmann: modeling biomolecule-water and water-water interactions. *Phys. Rev. Lett.* **2009**, *102*, 087801.
(40) Koehl, P.; Delarue, M. Aquasol: an efficient solver for the dipolar Poisson-Boltzmann-Langevin equation. *J. Chem. Phys.* **2010**, *132*, 064101.
(41) Holst, M. Multilevel Methods for the Poisson-Boltzmann Equation. Ph.D. thesis; University of Illinois: Urbana−Champaign, IL, 1993.
(42) Holst, M.; Saied, F. Numerical solution of the nonlinear Poisson-Boltzmann equation: developing more robust and efficient methods. *J. Comput. Chem.* **1995**, *16*, 337−364.
(43) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acids. Res.* **2000**, *28*, 235−242.
(44) Dolinsky, T.; Nielsen, J.; McCammon, J. M.; Baker, N. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.
(45) Sitkoff, D.; Sharp, K.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.
(46) Guvench, O.; MacKerell, A., Jr. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435.
(47) Oleinikovas, V.; Saladino, G.; Cossins, B.; Gervasio, F. Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc.* **2016**, *138*, 14257−14263.
(48) Heaslet, H.; Harris, M.; Fahnoe, K.; Sarver, R.; Putz, H.; Chang, J.; Subramanyam, C.; Barreiro, G.; Miller, J. Structural comparison of chromosomal and exogenous dihydrofolate reductase from Staphylococcus aureus in complex with the potent inhibitor trimethoprim. *Proteins: Struct., Funct., Genet.* **2009**, *76*, 706−717.
(49) Han, G.; Lee, J.; Song, H.; Chang, C.; Min, K.; Moon, J.; Shin, D.; Kopka, M.; Sawaya, M.; Yuan, H.; et al. Structural basis of non-specific lipid binding in maize lipid transfer protein complexes revealed by high-resolution X-ray crystallography. *J. Mol. Biol.* **2001**, *308*, 263−278.
(50) Silvaroli, J.; Arne, J.; Chelstowska, S.; Kiser, P.; Banerjee, S.; Golczak, M. Ligand binding induces conformational changes in human cellular retinol-binding protein 1 (CRBP1) revealed by atomic resolution crystal structures. *J. Biol. Chem.* **2016**, *291*, 8528−8540.
(51) Hamiaux, C.; Stanley, D.; Greenwood, D.; Baker, E.; Newcomb, R. D. Crystal structure of Epiphyas postvittana takeout 1 with bound ubiquinone supports a role as ligand carriers for takeout proteins in insects. *J. Biol. Chem.* **2009**, *284*, 3496−3503.
(52) *The PyMOL Molecular Graphics System*, Ver. 2.4, Schrödinger, LLC: New York, 2020.
(53) Saurabh, S.; Vanaphan, N.; Wen, W.; Dauwalder, B. High functional conservation of takeout family members in a courtship model system. *PLoS One* **2018**, *13*, e0204615.
(54) Cherezov, V.; Rosenbaum, D.; Hanson, M.; Rasmussen, S.; Thian, F.; Kobilka, T.; Choi, H.; Kuhn, P.; Weis, W.; Kobilka, B.; et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258−1265.

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067

(55) Weiss, M.; Schulz, G. Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* **1992**, *227*, 493−509.
(56) Bai, T. Beta 2 adrenergic receptors in asthma: a current perspective. *Lung* **1992**, *170*, 125−141.
(57) Graham, S.; Leja, N.; Carlson, H. MixMD Probeview: robust binding site prediction from cosolvent simulations. *J. Chem. Inf. Model.* **2018**, *58*, 1426−1433.
(58) Tan, Y.; Verma, C. Straightforward incorporation of multiple ligand types into molecular dynamics simulations for efficient binding site detection and characterization. *J. Chem. Theory Comput.* **2020**, *16*, 6633−6644.
(59) Busch, A.; Reijerse, E.; Lubitz, W.; Frankenberg-Dinkel, N.; Hofmann, E. Structural and mechanistic insight into the ferredoxin-mediated two-electron reduction of bilins. *Biochem. J.* **2011**, *439*, 257−264.
(60) Dammeyer, T.; Frankenberg-Dinkel, N. Function and distribution of bilin biosynthesis enzymes in photosynthetic organisms. *Photochem. Photobiol. Sci.* **2008**, *7*, 1121−1130.
(61) Frankenberg, N.; Mukougawa, K.; Kohchi, T.; Lagarias, J. Functional genomics analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxyhenic photosynthetic organisms. *Plant Cell* **2001**, *13*, 965−978.
(62) Cheng, K.-Y.; Lowe, E.; Sinclair, J.; Nigg, E.; Johnson, L. The crystal structure of the human polo-like kinase-1 polo box domain and its phospho-peptide complex. *EMBO J.* **2003**, *22*, 5757−5768.
(63) Liu, F.; Park, J.; Qian, W.-J.; Lim, D.; Graeber, M.; Berg, T.; Yaffe, M.; Lee, K.; Burke, T., Jr. Serendipitous alkylation of a Plk1 ligand uncovers a new binding channel. *Nat. Chem. Biol.* **2011**, *7*, 595−601.
(64) Śledź, P.; Stubbs, C.; Lang, S.; Yang, Y.-Q.; McKenzie, G.; Venkitaraman, A.; Hyvönen, M.; Abell, C. From crystal packing to molecular recognition: prediction and discovery of a binding site on the surface of polo-like kinase 1. *Angew. Chem., Int. Ed.* **2011**, *50*, 4003−4006.
(65) Tirion, M. Large amplitude elastic motions in proteins from a single parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(66) Sanejouand, Y. Elastic network models: theoretical and empirical foundations. *Methods Mol. Biol.* **2013**, *924*, 601−616.
(67) Sinitskiy, A.; Voth, G. Coarse-graining of proteins based on elastic network models. *Chem. Phys.* **2013**, *422*, 165−174.

https://doi.org/10.1021/acs.jpcb.1c02658
J. Phys. Chem. B 2021, 125, 5052−5067