Effect of the Solute Cavity on the Solvation Energy and its Derivatives within the Framework of the Gaussian Charge Scheme

Miquel Garcia-Ratés | Frank Neese

Abstract
The treatment of the solvation charges using Gaussian functions in the polarizable continuum model results in a smooth potential energy surface. These charges are placed on top of the surface of the solute cavity. In this article, we study the effect of the solute cavity (van der Waals-type or solvent-excluded surface-type) using the Gaussian charge scheme within the framework of the conductor-like polarizable continuum model on (a) the accuracy and computational cost of the self-consistent field (SCF) energy and its gradient and on (b) the calculation of free energies of solvation. For that purpose, we have considered a large set of systems ranging from few atoms to more than 200 atoms in different solvents. Our results at the DFT level using the B3LYP functional and the def2-TZVP basis set show that the choice of the solute cavity does neither affect the accuracy nor the cost of calculations for small systems (< 100 atoms). For larger systems, the use of a vdw-type cavity is recommended, as it prevents small oscillations in the gradient (present when using a SES-type cavity), which affect the convergence of the SCF energy gradient. Regarding the free energies of solvation, we consider a solvent-dependent probe sphere to construct the solvent-accessible surface area required to calculate the nonelectrostatic contribution to the free energy of solvation. For this part, our results for a large set of organic molecules in different solvents agree with available experimental data with an accuracy lower than 1 kcal/mol for both polar and nonpolar solvents.

KEYWORDS
C-PCM, density functional theory, implicit solvation

1 | INTRODUCTION
The use of implicit solvation models is nowadays one of the most adopted strategies to include the effect of the solvent into quantum-mechanical (QM) calculations. In this approach, the interaction between the solute and the solvent (continuum) is represented by a charge density defined on the surface of the cavity where the solute is accommodated. The most adopted implicit solvation model is the polarizable continuum model (PCM), first formulated by Tomasi et al. and refined afterward. A particular case of the PCM is the conductor-like polarizable continuum model (C-PCM), where the solvation charge density is calculated assuming a scaled conductor boundary condition. This approach was first proposed by Klamt within the framework of the conductor-like screening model (COSMO). In this approach, like in most of the PCM variants, the polarization charge density is defined in terms of discrete point charges spread...
over the surface of the solute cavity. There exist two main types of solute cavities used to treat the contribution of solvation to the electrostatics of a system: (a) the vdW-type cavity, (b) the solvent-excluded surface (SES)-type cavity. The first kind of cavity is the one contained inside the external surface of the vdW spheres centered on each atom (see Figure 1b). The latter, on the other hand, is obtained by following the inward part of a solvent sphere rolling over the solute (see Figure 1c), which makes its outer surface smoother than the vdW surface. The SES-type cavity is, in most of the cases, generated through the GEt Pol (GEnerating POLyhedra) algorithm. Other strategies to generate the SES are based on an implicit function derived from equivalence statements and Voronoi-type diagrams. The use of solvation point charges leads, however, to discontinuities in the Coulomb potential that affect the self-consistent field (SCF) energy and its derivatives with respect to nuclear displacements. This is firstly due to the fact that the interaction between surface elements depends on $r_{ij}^{-1}$, $r_{ij}$ being the distance between charges $i$ and $j$. Then, when $r_{ij}$ tends to zero (solvation charges overlap), this directly affects the SCF energy, as it is computed from the solvation charges. A very efficient way to overcome this problem is to represent the effect of the solvent in our system by a charge density made of spherical Gaussians centered at the point charge coordinates. This strategy was first proposed by York and Karplus and further refined by Scalmani and Frisch as well as Lange and Herbert. In all cases, together with the Gaussian treatment of the solvation charges, a switching function is used to attenuate the contribution of the charges to the solvation properties. This last point is crucial in order to avoid discontinuities in the solvation potential due to the charges that may appear or disappear when optimizing the geometry of a molecule. This whole strategy, that from now we call the Gaussian charge scheme (GCS), results in a rigorously smooth potential energy surface possessing continuous derivatives.

The use of Gaussian charges together with the use of a switching function permits to model the contribution of solvation to the electrostatic energy of the system. However, a rigorous treatment of solvation should also include nonelectrostatic contributions to the energy. There exist different ways to consider the nonelectrostatic solvation energy, $\Delta G_{\text{sol}}$. A simple way to do so is to assume this contribution to linearly depend on the solvent-accessible surface area (SASA) and on the contribution of each atom to the SASA. The SASA is obtained by following the center of a solvent probe sphere rolling over the vdW surface (see Figure 1a). A well-known continuum solvation model that involves an accurate description of the Solvation Model based on Density (SMD) is modeled as a sum of contributions that are proportional to the SASA of the atoms forming the solute via a set of geometry-dependent proportionality constants called atomic surface tensions. In this model, the SASA is generated by rolling a solvent sphere with a fixed radius $R_S$ of 0.4 Å over the vdW surface of the solute. In principle, the fact of not considering $R_S$ to be solvent-dependent is somehow hidden in the careful parametrization of the atomic surface tensions. Regarding this last point, there is a lack of studies in the literature that consider a different value for $R_S$ depending on the solvent of interest. Although the main goal of our study is to report the effect of the cavity used for electrostatics on the properties of a system using the GCS, we also want to prove that the use of a solvent-dependent $R_S$ to construct the SAS leads to the accurate prediction of solvation energies. The present work is divided into two parts. In the first part, we study the effect of the solute cavity used to compute the solvation contribution to electrostatics within the C-PCM using a GCS for the polarization charges. For that purpose, the GCS has been implemented into the ORCA quantum chemistry suite. We focus on the results for the first derivatives of the SCF energy for small and big systems (up to 300 atoms). In order to verify the correctness of our implementation we compare our results using the GCS with those for systems where the point charge scheme results in singularities in the SCF energy and its derivatives. In the second part of this study, we have parametrized the nonelectrostatic contribution to the free energy of solvation in order to predict experimental data for this last quantity for organic solutes in different solvents. Our aim in this last part is to show that a strategy that considers a solvent-dependent $R_S$ for nonelectrostatic solvation effects together with a Gaussian treatment of the solvation charges leads to accurate results for solvation data.

This article is organized in the following way. In the theory section, we present the basics of the C-PCM and the GCS, we discuss the different types of solute cavities and we introduce the equations to compute the free energy of solvation. The computational details of our calculations, together with the procedure used to parametrize the free energy of solvation are provided afterward. In the results section, we show the outcome regarding the efficiency and computational cost associated to the GCS, the performance of this scheme for large systems, and the results for the parametrization of the solvation free energy. Finally, the last section is devoted to the main conclusions that can be drawn from this study.

**FIGURE 1** The different type of cavity surfaces for a given solute (a), van der Waals surface (b), solvent-excluded surface (c), and solvent-accessible surface (d). The solvent molecule is represented as a dashed sphere labeled as "S"
2 | THEORY

2.1 | C-PCM equations

For a solvated system within the PCM, the Hamiltonian of the solute in vacuum, $H_0$, is perturbed by a solvation term that depends on the $N_q$ charges, $q_i$, spread out over the surface of the solute cavity,

$$H = H_0 + \sum_{i=1}^{N_q} \frac{q_i}{|r_i - r|}$$  \hspace{1cm} (1)

with $r_i$ the position of the $i$th point charge.

Within the C-PCM, the charge vector $q$ is calculated by solving the following system of equations:

$$Aq = -f(x)V$$  \hspace{1cm} (2)

where $V$ stands for the electrostatic potential at the position of the charges, $A$ is the LHS matrix, and $f(x)$ is a scaling function that depends on the dielectric permittivity, $\epsilon$, of the solvent

$$f(x) = \frac{\epsilon - 1}{\epsilon + x}$$  \hspace{1cm} (3)

The parameter $x$ is equal to 0 in the case of the C-PCM. The elements of the matrix $A$ have a different functional form depending on how the charges on the surface of the cavity are treated, as we discuss later on.

The SCF energy for the solvated system reads as

$$E = \sum_{\mu\nu\sigma} P_{\mu\nu\sigma} h_{\mu\nu\sigma} + \frac{1}{2} \sum_{\mu\nu\sigma\tau\eta\zeta} (P_{\mu\nu\sigma} P_{\tau\eta\zeta} - c_{\sigma} \delta_{\nu\tau} P_{\mu\eta\zeta}) \times (\mu|\nu|\sigma) + V_{\text{NN}}$$

$$+ \frac{1}{2} q^T V X + E_{\text{XC}}[\rho_x(r)]$$  \hspace{1cm} (4)

where $P_{\mu\nu\sigma} = \sum_{\eta\zeta} c_{\eta\zeta} c_{\sigma}$ is the solute density matrix for spin $\alpha$, $\eta = \alpha, \beta$, with $c_{\eta\zeta}$ being the molecular orbital (MO) coefficients. We adopt the Mulliken notation for the two-electron integrals over the atom-centered basis functions $|\mu\rangle$. The term $V_{\text{NN}}$ accounts for the nuclear repulsion energy while the fourth term describes the interaction between the solute and the solvent. Finally, $E_{\text{XC}}[\rho_x(r)]$ is the exchange-correlation energy with $c_{\eta}$ equal to the fraction of Hartree–Fock exchange and $\rho_x(r) = \sum_{\mu\eta} P_{\mu\eta}(r)\nu(r)$ being the electron density for spin $\alpha$ at position $r$.

The first derivative of $E$ with respect to an infinitesimal nuclear displacement $X$, is equal to

$$E_X = \frac{\partial E}{\partial X} = \sum_{\mu\nu\sigma} P_{\mu\nu\sigma} h_{\mu\nu\sigma} X + \frac{1}{2} \sum_{\mu\nu\sigma\tau\eta\zeta} (P_{\mu\nu\sigma} P_{\tau\eta\zeta} - c_{\sigma} \delta_{\nu\tau} P_{\mu\eta\zeta}) \times (\mu|\nu|\sigma) X$$

$$- \sum_{\mu\nu\sigma} S_{\mu\nu\sigma} h_{\mu\nu\sigma} + V_{\text{NN}} + q^T V X + \frac{1}{2f(x)} q^T A_X^T A_X q + E_{\text{XC}}[\rho_x(r)]$$  \hspace{1cm} (5)

The matrices $S$ and $W$ are the overlap and energy-weighted density matrices, respectively. If the superscript $X$ is enclosed in parenthesis, then the derivative does not involve the calculation of $d\mathbf{P}/dX$. Notice that, by definition, $W = P F P^T$ with $F$ being the Fock matrix. When we adopt the C-PCM approach, a solvation contribution is also added to the Fock matrix elements over basis functions, $F_{\mu\nu\sigma}$,

$$F_{\mu\nu\sigma} = h_{\mu\nu\sigma} + \sum_{\xi} P_{\sigma\xi\eta} \left( \phi_{\xi}(x|\mu|\nu) - c_{\sigma} \delta_{\nu\tau} \phi_{\xi}(x|\mu|\tau) \right) + V_{\text{XC}}^{\text{NC}} + q^T V_{\text{NC}}$$  \hspace{1cm} (6)

where $V_{\text{XC}}^{\text{NC}} = \int \phi_{\nu}(x) f_{\text{NC}}(\epsilon_{\nu}(x)) \phi_{\nu}(x) dr$ is the exchange–correlation potential and $V_{\text{NC}}$ is the potential due to shells $\chi_{\nu}$ and $\chi_{\nu}'$ at the position of the charges.

2.2 | Discontinuities in the SCF energy

The left-hand side (LHS) of Equation (2) represents the electrostatic potential caused by the distribution of solvation charges at their position on the surface of the solute cavity. That means that the non-diagonal elements of the LHS matrix, let us say $A_{ij}$, depend on the inverse of $r_{ij}$, the distance between charges $i$ and $j$. If we adopt the point charge scheme, $A_{ij} = 1/r_{ij}$. This means that $\lim_{r_{ij} \to 0} A_{ij} = \infty$, that is $A_{ij}$ diverges when charges $i$ and $j$ overlap resulting in a discontinuous potential energy surface. Of course, the larger the molecule, the larger the number of charges and then, the larger the probability of having point charges almost overlapping. This problem gets even worse when performing geometry optimizations, as the gradient of the SCF energy depends on the derivatives of $A$ (see Equation (5)) that diverge faster ($A^3$ depends on $r_{ij}^{-5}$). Then, even if we have a small system, when optimizing it, it is very likely that some of the charges get very close to each other causing numerical instabilities in the SCF energy that will lead to erroneous optimized geometries. If one calculates vibrational frequencies, this systematic error is accumulated again, due to two main reasons. The first one is associated to the fact that we are using an optimized geometry that can be biased. Second, the Hessian, if calculated analytically, does not only depend on $r_{ij}^{-4}$ and $r_{ij}^{-2}$ but also on $r_{ij}^{-5}$. If we compute it numerically through finite differences, we need to generate the cavity as many times as we compute the gradient.

Thus, there is a significant probability to end up with point charges overlapping for many of the Hessian elements. A rudimentary way to solve this problem is to discard the point charges that are closer to each other than a certain cutoff. This basically removes all the charges overlapping for many of the Hessian elements. However, by doing so we are eliminating an important part of the physics of the problem. An efficient way to remove the singularities in the Coulomb potential caused by the solvation charges, already discussed earlier, is based on replacing the point charges by a distribution of spherical Gaussian functions.

However, although using Gaussian charges produces finite solvation contributions to the SCF energy, it does not completely remove the discontinuity in the derivatives of the solvation potential with respect to nuclear displacements. This issue arises from those charges changing discontinuously in the intersection between cavity spheres from one molecular configuration to the other during geometry
2.3 Gaussian Charge Scheme

Within the GCS, the effect of the solvent is represented by a charge density made of spherical Gaussian functions centered at positions \( r_i \) on the surface of the solute cavity\([15]\)

\[
g_i(r) = q_i \left( \frac{2\zeta}{\pi} \right) \exp \left( -\frac{\zeta^2}{2} |r - r_i|^2 \right)
\]

Here, \( q_i \) is the magnitude of the charge located at \( r_i \) and \( \zeta \) is the width of the Gaussian. If we consider that the surface of each of the spheres that compose the solute cavity is discretized following a Lebedev quadrature scheme, then each charge \( i \) with width \( \zeta_i \) is represented by a position \( r_i \), its Lebedev weight \( w_i \) and the radius \( R_i \) of the sphere on which the charge is placed. In this case, \( \zeta_i \) is computed as

\[
\zeta_i = \frac{\zeta}{R_i \sqrt{w_i}}
\]

where the exponent \( \zeta \) is a dimensionless parameter adjusted to obtain the exact Born solvation energy for a conductor and a uniform surface charge distribution. This parameter depends on the Lebedev grid chosen\([15]\).

The elements of the LHS matrix \( A \) in Equation (2) have the following functional form under the GCS.

\[
A_{ij} = \frac{\zeta_j}{\sqrt{2} F_i} \quad (9)
\]

\[
A_{ij} = \frac{\text{erf}(\zeta_i r_i)}{r_i} \quad (10)
\]

with \( \zeta_i = \zeta_i / \left( \zeta_i^2 + \zeta_j^2 \right)^{1/2} \). Regarding the diagonal matrix elements, \( A_{ii} \), they are calculated from the so-called switching function \( F_i \). This function weights the contribution of each charge to the solvation properties and can be calculated as a product of elementary switching functions \( f(r, R_J) \)

\[
F_i = \prod_{J=1}^{N_{sph}} f(r_i, R_J)
\]

where \( N_{sph} \) is the number of spheres that define the solute cavity and \( R_J \) is the position vector of sphere \( J \). We adopt the improved switching-Gaussian (ISWIG) function proposed in Reference [17].

\[
f(r_i, R_J) = 1 - \frac{1}{2} \left\{ \text{erf}(\zeta_i (r_i - r_{i0})) + \text{erf}(\zeta_i (r_i + r_{i0})) \right\}
\]

with \( R_J \) being the radius of sphere \( J \) and \( r_{i0} \) the distance between charge \( i \) and the \( J \)th sphere center.

It is especially interesting to analyze the functional form of the non-diagonal elements of the LHS \( A \) matrix in Equation (10). Differently than for the point charge scheme, where \( A_{ii} = R_i^{-1} \), we have this quantity weighted by \( \text{erf}(\zeta_i r_i) \). Then,

\[
\lim_{r_i \to 0} A_{ij} = \frac{2\zeta_i}{\sqrt{\pi}}
\]

That is, \( A_{ij} \) does not diverge when two charges overlap, as it is the case for the point charge scheme. This fact prevents discontinuities in the SCF energy. Indeed, with the choice of \( A_{ii}, A_{ij} \) and \( F_i \) we have a strictly continuous potential energy surface.

The contribution to the surface area of the solute cavity by the charge \( i \), \( a_i \), is calculated as

\[
a_i = w_i F_i R_i^2
\]

Then, the total surface area of the chosen cavity is just

\[
S_A = \sum_{i=1}^{N_sph} a_i
\]

The total volume of the cavity, \( V_T \), is

\[
V_T = \frac{1}{3} \sum_{i=1}^{N_{sph}} a_i (r_i - r_{i0}) \frac{1}{R_i}
\]

In order to perform geometry optimizations within the C-PCM using the GCS we need the derivatives of \( A_{ij} \) and \( A_{ii} \) with respect to nuclear displacements. This involves calculating the gradient of \( F_i \). If we consider the nuclear coordinates of sphere \( A, R_A = (X_A, Y_A, Z_A) \), it can be demonstrated that \( \frac{\partial F_i}{\partial R_A} = F_i \sum_{j=1}^{N_{sph}} \frac{\partial a_j}{\partial R_A} \). Then,

\[
\frac{\partial A_{ij}}{\partial R_A} = \sqrt{\frac{2}{\pi}} \frac{\zeta_j}{F_i} \sum_{j=1}^{N_{sph}} \frac{1}{R_{ij}} \left\{ (R_j + r_{i0}) \times \exp \left[ -\zeta_j^2 (R_j + r_{i0})^2 \right] + (R_j - r_{i0}) \times \exp \left[ -\zeta_j^2 (R_j - r_{i0})^2 \right] \right\} \frac{\partial \zeta_j}{\partial R_A}
\]

\[
+ \frac{2 \zeta_j^2}{\pi} \sum_{j=1}^{N_{sph}} \frac{1}{r_{ij}} \exp \left[ -\zeta_j^2 (r_{ij})^2 \right] \left( \frac{\partial R_j}{\partial R_A} - \frac{\partial r_{ij}}{\partial R_A} \right)
\]
On the other hand,

\[
\frac{\partial A_{ij}}{\partial R_{ij}} \frac{\partial}{\partial R_{ij}} \left[ \frac{\text{erf}(\frac{\zeta R_{ij}}{\tau})}{\tau} \right] = \frac{2}{\sqrt{\pi}} \exp\left(-\frac{\zeta^2 r_{ij}^2}{\tau^2}\right) \frac{\partial^2 \zeta}{\partial R_{ij}^2} + \left[ \frac{2}{\sqrt{\pi} \tau r_{ij}} \exp\left(-\frac{\zeta^2 r_{ij}^2}{\tau^2}\right) - \frac{\text{erf}(\frac{\zeta R_{ij}}{\tau})}{\tau^2} \right] (r_i - r_j) \cdot \left( \frac{\partial r_i}{\partial R_{ij}} - \frac{\partial r_j}{\partial R_{ij}} \right)
\]

The terms that depend on \(\frac{\partial \zeta}{\partial R_{ij}}\) and \(\frac{\partial^2 \zeta}{\partial R_{ij}^2}\) are equal to 0 for a vdW-type cavity because they are proportional to the derivatives of the radius of the cavity spheres with respect to nuclear displacements. In the case of the SES, the equations used in the present study to compute these terms are provided in Data S1. The quantities \(\frac{\partial^2 \zeta}{\partial R_{ij}^2}\) in Equations (17) and (18) are calculated slightly different depending on the solute cavity chosen, as we discuss in the following subsection.

### 2.4 Type of cavities used for the solvation contribution to electrostatics

We consider two types of solute cavities to be used together with the GCS to compute the solvation contribution to electrostatics, (a) a vdW-type cavity and (b) a SES-type cavity.

#### 2.4.1 VdW-type cavity

This type of cavity is generated by scaling the vdW radius of each atom \(R_{\text{vdW}j}\) by a factor \(f\), that is, \(R_i = f R_{\text{vdW}j}\) and taking the external surface of these spheres. We use Bondi’s vdW radii for those elements where this quantity is available,\(^{24}\) except for hydrogen where we adopt the radii proposed by Mantina et al.\(^{26}\) This is the case of the following elements: Be, B, Al, Ca, Ge, Rb, Sr, Sb, Cs, Ba, Bi, Po, At, Rn, Fr, and Ra. For those elements that are covered neither by Bondi nor by Reference \(^{26}\), we consider a value of 2 Å. Although previous studies have shown that the use of a solvent-dependent \(f\) improves the electrostatic interaction between solute and solvent,\(^{27–30}\) in our approach, \(f = 1.2\), as done in other studies.\(^{17}\) The Gaussian charges are finally spread over the scaled vdW surface using a Lebedev grid as pointed out in the previous subsection.

#### 2.4.2 SES-type cavity

The second type of solute cavity is the one that has the SES as its external surface. In this case, we use the GEPOL algorithm to generate it but not to discretize its surface. The GEPOL algorithm iteratively considers all possible pairs of existing spheres and checks if there is space in between them that should be inaccessible to the solvent. If it is the case, a new sphere (N) is added in between the pair of spheres. There exist three different situations depending on the distance between the pair of spheres, their radii and the radius of the solvent probe sphere. These different cases, A, B, and C, are shown in Figure 2. The equations used to compute the position vector of the new spheres, \(r_{N}\), are provided in Data S1. Regarding the radii of the spheres for the SES-type cavity, we adopt the same strategy that for the vdW-type cavity, which is scaling the vdW radius of each atom by a scaling factor \(f\), with \(f\) being equal to 1.2. The radii of the new GEPOL spheres (those that are not associated to atoms) are calculated from the radii of the pair of spheres considered to create it, as well as from the distance between them and from the radius of the solvent molecule \(R_s\). In order to determine \(R_s\) for each solvent, we start from the Stokes-Einstein formula,\(^{31}\) that gives the radius, \(R_{SE}\), of a sphere that diffuses at the same rate as the solvent under consideration,

\[
R_{SE} = \frac{k_B T}{6 \pi \eta D}
\]

where \(D\) is the self-diffusion coefficient of the solvent, \(\eta\) its viscosity, \(k_B\) the Boltzmann’s constant, and \(T\) is the temperature. In the case of water at ambient conditions \((T = 298.15 \text{ K})\), \(D = 2.26 \times 10^{-9} \text{ m}^2/\text{s}\), \(\eta = 8.9 \text{ cP}\), which results in \(R_{SE}^w = 1.09 \text{ Å}\). This value is slightly lower than the experimentally determined radius for a water molecule, which is \(R_s^w = 1.38 \text{ Å}\). In the present study, we adopt the later value of \(R_s\) for water. If we consider a different solvent than water, we calculate the radius of the solvent molecules with the following equation,

\[
R_s^s = \left( \frac{R_s^w}{R_{SE}^w} \right) R_{SE}^s
\]

where \(R_s^w / R_{SE}^w \approx 1.27\). This prefactor slightly corrects for the fact that \(R_{SE}\) underestimates the value of \(R_s\). The values of \(R_s\) for the solvents

![FIGURE 2](Image 309x101 to 549x317)
considered in this study are shown in Table 1. Although an estimation of $R_s$ based on $R_{SE}$ can seem a crude approximation, this strategy permits us to take into account the size of the solvent molecule, something that is ignored in most of the implicit solvation models.

Once the GEPOL spheres are generated (position and radius) we spread the Gaussian charges over the SES using a Lebedev grid. A quantity that requires special attention are the derivatives of the charge positions, $r_i$, with respect to nuclear displacements. Let us consider, for instance, the first derivative of $r_i$ with respect to the infinitesimal displacement of atom $A$ along the $x$ direction, that is $\partial r_i/\partial x_A$. If the charge $i$ belongs to atom $A$, then $\partial r_i/\partial x_A = (1, 0, 0)$. If it is not the case, and $i$ belongs to a GEPOL sphere $N$ (not associated to any atom) then we should check if the atom $A$ is one of the two spheres that was used to create $N$. If it is the case, due to the fact that $r_i$ depends on the coordinates of atom $A$, $R_A$, then $\partial r_i/\partial x_A \neq 0$. The equations used to compute the gradient for cases A, B, and C in Figure 2 are provided in Data S1.

### 2.5 Free energy of solvation

The solvation free energy, $\Delta G_{solv}$, is the free energy of transfer of a solute from the gas phase to the condensed phase. If we denote by $R_s$ and $R_i$ the equilibrium coordinates of a solute $M$ in the gas phase and in solution, respectively, the fixed-concentration free energy of solvation of $M$ can be written as,[32]

$$\Delta G_{solv} = E_{solv}(R_s, R_i) + \Delta G_c(R_i) + G_{solv,lib} + G_{solv,rv} \tag{21}$$

The first term, $E_{solv}(R_s, R_i)$, accounts for the difference between the liquid-phase expectation value of the gas-phase Hamiltonian, $E(R_i)$, and the gas-phase potential energy surface $E(R_s)$

$$E_{solv}(R_s, R_i) = E(R_i) - E(R_s) \tag{22}$$

The second contribution to Equation (21), $\Delta G_c(R_i)$, represents the changes in the free energy due to the coupling between the solute and the solvent. This term can be approximated by

$$\Delta G_c(R_i) = \Delta G_{el} + \Delta G_{pol} \tag{23}$$

with $\Delta G_{el}$ and $\Delta G_{pol}$ being the electronic-polarization and non-electrostatic contributions to $\Delta G_{solv}$, respectively. Regarding $\Delta G_{pol}$, it can be split into two terms: (a) a cavitation energy, $\Delta G_{cav}$, defined as the reversible work to create a cavity inside the solvent to accommodate the solute, and (b) a term, $\Delta G_{disp}$, that accounts for the changes in the dispersion energy when solvating the solute.

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{disp} \tag{24}$$

With respect to the term $G_{solv,lib} - \Delta G_{disp}$ in Equation (21), it accounts for the difference between the liberaional energy of the system in solution, $G_{solv,lib}$, and the translational free energy of the solute in the gas phase, $G_{solv,lib}$. If we choose standard states in which the concentration of the solute in the gas-phase is the same as that in solution, $G_{solv,lib} = G_{solv,lib}$, and then $G_{solv,lib} - \Delta G_{disp} = 0$. Finally, the term $G_{solv,rv}$ is defined as the difference between the rotational-vibrational free energy of the solvated system, $G_{solv,rv}$, and that for the system in vacuum, $G_{solv,rv}$.

$$G_{solv,rv} = G_{solv,rv} - G_{solv,rv} \tag{26}$$

In this case, it is also customary to consider $G_{solv,rv} = 0$. This approximation is based on the fact that the contribution of $G_{solv,rv}$ to $\Delta G_{solv}$ is in general small with respect to the other terms in Equation (21) and expected to be smaller than the mean intrinsic error of the continuum solvation model.

If we then consider Equations (25) and (26) to be equal to zero, and combine Equations (22), (23), and (24) with Equation (21), then $\Delta G_{solv}$ reads as,

$$\Delta G_{solv} = E(R_s) - E(R_i) + \Delta G_{el} + \Delta G_{cav} + \Delta G_{disp} \tag{27}$$

The term $\Delta G_{el}$ is equal to the fourth term in Equation (4), and is calculated from the charges spread over the surface of the vdW-type cavity or the SES. Regarding the last two terms in Equation (27), they are calculated using the SASA (see Figure 1d). The radius of the solvent is calculated using the approach described in the previous subsection. For the cavitation energy $\Delta G_{cav}$, we employ the equation proposed by Pierotti,[33] based on scaled particle theory for a hard-sphere solute in a hard-sphere fluid of solvent molecules. In this case, $\Delta G_{cav}$ is expanded in powers of the ratio $R$ between the radius of the solute sphere $R_i$ and that of the solvent spheres $R_s$.

$$\Delta G_{cav} = \Delta G_{cav} = -\ln(1 - y) + \left( \frac{3y}{1 - y} \right) R_s^3 + \left( \frac{3y}{1 - y} \right) \left[ \frac{y}{1 - y} \right]^2 R_s^3 + \frac{y}{p_s} \frac{R_s^3}{R} \tag{28}$$

with $y = \delta / \rho_s R_s^2 / 6$, being $\delta$ the density of the solvent, and $R = R_i / R_s$. The last term in Equation (28) depends on the thermodynamic conditions, that is, temperature and pressure. Equation (28) was extended by Claverie et al. for a cavity of complex shape made of several

### Table 1

| Molecule     | $D$ ($10^{-9}$ m$^2$/s) | $\eta$ (cP) | $R_{SE}$ (Å) | $R_s$ (Å) |
|--------------|--------------------------|-------------|--------------|-----------|
| Benzene      | 2.207                    | 6.030       | 1.641        | 2.086     |
| Chloroform   | 2.320                    | 5.360       | 1.756        | 2.232     |
| Cyclohexane  | 1.470                    | 8.945       | 1.661        | 2.111     |
| Octanol      | 2.207                    | 6.030       | 1.641        | 2.086     |
| Toluene      | 2.274                    | 5.600       | 1.715        | 2.177     |
| Water        | 2.260                    | 8.901       | 1.086        | 1.380     |

*Note: The value for the self-diffusion coefficient, $D$, the viscosity, $\eta$, and the Stokes-Einstein radius, $R_{SE}$, are also provided.*
In this case, the contribution to the total $\Delta G_{\text{cav}}$ by sphere $I$ is weighted by a prefactor that depends on the exposed SASA$_I$ of that sphere.

$$\Delta G_{\text{cav}} = \sum_{I=1}^{N_{sph}} \sigma_I \text{SASA}_I = \sum_{I=1}^{N_{sph}} \sigma_I \text{SASA}_I$$  \hspace{1cm} (29)

Here, the summation over the total number of spheres can be replaced by a summation over the total number of surface elements ($N_d$) in which the SASA is divided into. This number is, in general, larger than the number of charges placed on top of either the vdW and the SES surfaces. Finally, $I$ stands for the index of the sphere to which the surface element $i$ belongs to.

Regarding the last term in Equation (27), $\Delta G_{\text{disp}}$, we assume it to depend linearly on the contribution of each atom to the SASA$^{[18,19]}$

$$\Delta G_{\text{disp}} = \sum_{I=1}^{N_{at}} \sigma_I \text{SASA}_I = \sum_{I=1}^{N_{at}} \sigma_I \text{SASA}_I$$  \hspace{1cm} (30)

Here, the factor $\sigma_I$ is the atomic surface tension of the sphere $I$. The equations for the gradient of $\Delta G_{\text{cav}}$ and $\Delta G_{\text{disp}}$ are provided in Data S1.

### 2.6 Computational details

All the calculations in this study have been performed using a development version of the ORCA 4.2 electronic structure package at the DFT level of theory. The GCS has been implemented in ORCA, as described in the theory section. We denote by GVDW and GSES the cases where we use the scaled vdW-type cavity and the SES-type cavity for the contribution of solvation to the electrostatics of the system, respectively. When the nonelectrostatic contributions to the free energy of solvation are included in the SCF energy and its derivatives, we label the solvation approach as GVDW$_{\text{nel}}$ and GSES$_{\text{nel}}$. In all calculations using the GCS, we place 110 charges per sphere. Regarding the SASA, which is used in the calculation of $\Delta G_{\text{cav}}$, and $\Delta G_{\text{disp}}$, we consider 434 Lebedev points, due to the fact that the SASA is larger than the vdW surface or the SES. We have also performed a set of calculations using point charges in order to prove the validity of the GCS to treat the contribution to the electrostatics. In this case, we use a SES-type cavity generated through the GEPOL algorithm$^{[11-13]}$ with 60 triangles per sphere, discarding surface points that are closer than 0.1 Å from each other. The other GEPOL settings are NDIV = 5 and RMIN = 0.5 Å. We refer to this last method as PSES. The different approaches used in this study, with the type of charges and cavities employed are shown in Table 2.

In all calculations, we have employed the B3LYP functional$^{[25-27]}$ with the def2-TZVP basis set$^{[28]}$ using the resolution of the identity and the chain of spheres approximation (RIJCOSX)$^{[29]}$ With regard to the DFT integration grid, we consider the Grid7 settings that correspond to $N_{\text{ang}}$ = 770 and $\xi$ = 5.67, being $N_{\text{ang}}$ the number of points of the largest angular grid, and $\xi$ the number of radial shells for a given atom. In order to speed-up the calculation of molecular integrals, ORCA uses the Libint library, version 2.0.2-stable.$^{[40]}$ The energy convergence tolerance for the SCF cycle is equal to $10^{-8}$ $E_h$ (VeryTightSCF). Geometry optimizations were done using the VeryTightOpt settings, which involve an energy tolerance equal to $2 \times 10^{-7}$ $E_h$, $3 \times 10^{-5}$ a.u. for the maximum component of the gradient vector, $8 \times 10^{-6}$ a.u. for the root mean square (RMS) gradient, $2 \times 10^{-4}$ Bohr for the maximum allowed displacement, and $10^{-4}$ a.u. for the maximum RMS displacement. The atom-pairwise dispersion correction with Becke-Johnson damping has been used in DFT calculations.$^{[41-45]}$

To parametrize $\Delta G_{\text{cav}}$, we use part of the SMD training set,$^{[21]}$ that involves experimental free energies of solvation, $\Delta G_{\text{exp}}$, for 318 solutes in 91 different solvents. The total number of considered solutes per solvent in our study is listed in Table 3, while the information on the type of solutes considered is provided in the Table S1. The training set involves organic solutes that just contain H, C, N, and O atoms. In order to derive the values for the atomic surface tensions in Equation (30), we proceed in the following way:

1. We perform a geometry optimization of the system in the gas-phase. We get a final energy $E_0$ (it is equal to the second term in Equation (27)).
2. We perform a geometry optimization of the same system in solution. In this case, we include the contribution of $\Delta G_{\text{cav}}$ to the SCF energy (Equations (28) and (29)) and its gradient but the effect of $\Delta G_{\text{disp}}$ is not included neither in the SCF nor its gradient. We get a final energy $E_0^\text{cav-disp}$. 
3. At the end of Step 2 we calculate the contribution of each atom to the SASA. Then we have a vector SASA$_I$ where $I$ runs over the different atoms in our system.
4. We compute $\Delta G_{\text{cav}} = \Delta G_{\text{cav}}^\text{Solv} - (E_0^\text{cav-disp} - E_0)$.
5. We fit $\Delta G_{\text{disp}}$ vs SASA$_I$ in order to get the different atomic surface tensions $\sigma_I$ via multivariate regression (Equation (30)).
6. Once we get the different $\sigma_I$, we perform Step 2 but now including the effect of $\Delta G_{\text{disp}}$ in the SCF energy and its gradient and calculate $\Delta G_{\text{disp}}$ as shown in Equation (27). If it is needed, we refine slightly $\sigma_I$.

### Table 2 Different schemes adopted to include solvation effects in our calculations

| Method   | Solvation charges | Type of surface (electrostatics) | Type of surface (nonelectrostatics) |
|----------|------------------|----------------------------------|-------------------------------------|
| PSES     | Punctual         | SES                              | -                                   |
| GSES     | Gaussian         | SES                              | -                                   |
| GVDW     | Gaussian         | vdW                              | -                                   |
| GSES$_{\text{nel}}$ | Gaussian         | SES                              | SAS                                 |
| GVDW$_{\text{nel}}$ | Gaussian         | vdW                              | SAS                                 |
In addition to the systems listed in Table S1, we have tested our approach for \( \Delta G_{\text{solv}} \) for a set of cations and anions in water (see Table S2). We have also carried geometry optimizations of two large systems, the oxygen evolving complex of photosystem II, and vancomycin in protein and water, respectively. The corresponding computational details are provided in the results section.

3 | RESULTS AND DISCUSSION

3.1 | Gaussian Charge Scheme

3.1.1 | Dependence on the number of charges

In order to determine the influence of the number of Gaussian charges per sphere (\( N_{\text{leb}} \)) on the calculated SCF energy, we have optimized the set of molecules described in Table S1 in water for the GVDW and GSES models with six different values for \( N_{\text{leb}} \): 50, 110, 194, 302, 434, 770. We have also considered two more molecules: (a) penicillin (42 atoms), and (b) sildenafil (63 atoms). This results in a total of 139 solutes. For each molecule, the initial structure to optimize in solution corresponds to the optimized geometry in vacuum. In Figure 3 we show the distribution of Gaussian charges for acetone for different values of \( N_{\text{leb}} \) using the GVDW and the GSES models. We define MAE\(_{770}\) as

\[
\text{MAE}_{770} = \frac{1}{N_k} \sum_{i=1}^{N_k} |E_i^k - E_i^{770}|
\]  \hspace{1cm} (31)

where \( k \) runs over \( N_{\text{leb}} \), \( E_i^k \) is the SCF energy of molecule \( i \) for the subset \( k \), and \( N_k \) is equal to the number of molecules for that subset, which is independent of \( k \). By definition MAE\(_{6,770}\) = 0, as \( k = 6 \) corresponds to \( N_{\text{leb}} = 770 \). The values of MAE\(_{770}\) for the GVDW and GSES models are shown in Figure 4. As expected, MAE\(_{770}\) converges to 0 when we increase \( k \). It is interesting to notice the small deviation between calculations using \( N_{\text{leb}} = 50 \) with respect to \( N_{\text{leb}} = 770 \) for both types of solute cavities, being MAE\(_{770}\) equal to 0.104 kcal/mol and 0.109 kcal/mol for the GVDW and the GSES models, respectively. For both models MAE\(_{770}\) < 0.05 kcal/mol, for \( k > 1 \), which is a very small error (for the GSES model MAE\(_{770}\) ≈ 0.052 kcal/mol). Then, based on that, we should choose \( N_{\text{leb}} \) to be larger or equal to 110.

However, to choose \( N_{\text{leb}} \) we have to look not only at the accuracy of the results but also at the computational cost of the calculations. In this respect, we have computed the three following quantities for each \( N_{\text{leb}} \): (a) the total time spent in the SCF iterations, \( t_{\text{SCF}} \), (b) the total time spent in the calculation of the gradient, \( t_{\text{grad}} \), and (c) the average time to compute the gradient per optimization cycle, \( t_{\text{cycle}} \). In Figure 5, we show the ratio between \( t_{\text{SCF}} \), \( t_{\text{grad}} \), and \( t_{\text{cycle}} \) for \( N_{\text{leb}} = 770 \) for both the GVDW and GSES models. All three quantities increase with \( N_{\text{leb}} \). If we focus on \( t/t_{770} \) for the gradient, we observe a linear increase with respect to \( N_{\text{leb}} \). In this case, \( t_{\text{grad}}/t_{\text{grad},770} \) is equal to 0.718 and 0.705 for the GVDW and the GSES models, respectively. When \( N_{\text{leb}} \) increases from 110 to 194, there is an increase of 5% of \( t_{\text{grad}} \) and \( t_{\text{cycle}} \) with respect to the case of \( N_{\text{leb}} = 770 \). For the SCF the same increase is of 2 and 4% for the GVDW and the GSES model. If we increase \( N_{\text{leb}} \) a bit more, from 110 to 302, then the

| Solvent  | Benzene | Chloroform | Cyclohexane | Octanol | Toluene | Water  |
|---------|---------|------------|-------------|---------|---------|--------|
| N\(_{\text{solutes}}\) | 51      | 59         | 54          | 96      | 35      | 137    |
| \( \varepsilon \) | 2.28    | 4.90       | 2.02        | 10.30   | 2.40    | 80.40  |

FIGURE 3 | Distribution of Gaussian charges for the GVDW model (a, b, c) and the GSES model (d, e, f). Left figures (a, d) correspond to \( N_{\text{leb}} = 110 \), middle figures (b, e) correspond to \( N_{\text{leb}} = 302 \), and right figures (c, f) correspond to \( N_{\text{leb}} = 770 \) [Color figure can be viewed at wileyonlinelibrary.com]
The efficiency of the GVDW and the GSES models has been studied. For this purpose, we compare both schemes to each other and to the PSES model. The PSES model uses 60 charges per sphere, while for the GVDW and the GSES models 110 charges are placed on the surface of each sphere. This makes $N_{q}^{GVDW}, N_{q}^{GSES} > N_{q}^{PSES}$, with $N_{q}$ being the total number of charges (after discarding charges inside other spheres), introducing a small bias in the comparison. However, the next level of tessellation for the PSES model would involve $60 \times 4$ point charges per sphere, which is quite large according to other studies that employ this same methodology. The ratio between $t_{SCF}$, $t_{grad}$, and $t_{cycle}$ for one of the methods with respect to the same quantity for the other is provided in Table 4 for the same systems studied in the previous subsection in water. If we focus first on $t_{cycle}$, we observe that this quantity is slightly larger for the GVDW and the GSES models with respect to the PSES model (1.035 and 1.049, respectively). This very small difference is due to the reason explained above, that is, $N_{q}$ is larger when the GCS is employed and then there are more operations to compute per each gradient cycle. Due to the fact that $N_{q}^{GSES} > N_{q}^{GVDW}$, $t_{cycle}$ is slightly larger for the GSES model as compared to the GVDW model. If we focus on $t_{SCF}$ and $t_{grad}$ these quantities are slightly smaller when adopting the GVDW and GSES models with respect to the PSES model. In this case, although the cost per SCF iteration and gradient cycle should increase when adopting the GCS scheme, the total amount of cycles is slightly smaller in total. This is also the case when we compare the GVDW and the GSES models, being $t_{SCF}$ and $t_{grad}$ slightly smaller for the first model. From the results shown in Table 4 we do not observe a significant difference in terms of computational cost between the three schemes. This is especially interesting, as it shows that adopting the GCS does not make calculations more expensive as compared to the point charge scheme.

For completeness, we have also computed the ratio between $t_{cycle}$ and $t_{SCF}$ (the average time spent in the SCF iterations before each optimization cycle) for the GVDW and GSES with respect to the vacuum case. The corresponding results, presented in Table 5, show that the GCS only adds an overhead of 7–8% relative to the gas-phase calculations. Regarding $t_{cycle}$, we have studied how this quantity increases with the size of the system. In Figure 6, we plot $t_{cycle}$ as a function of the product $N_{\text{atomic}} \times N_{q}$ for the GVDW and the GSES. In both cases, $t_{cycle} \propto (N_{\text{atomic}} \times N_{q})^{\beta}$, being $\beta = 1.414$ and 1.387 for the GVDW and the GSES, respectively. Consequently, an increase in system size, does not lead to a severe increase in computational cost for the calculation of the gradient with respect to the case of the same system in vacuum.
3.1.3 Performance

In the present subsection, we study the effect of the GCS on the convergence of the SCF energy and its gradient. For that purpose, we focus on the oxygen evolving complex (OEC) of photosystem II (PSII). The cluster model considered has 238 atoms (see Figure 7) and is taken from Reference [46] (optimized $S_2$ state model A). It contains an inorganic Mn$_4$O$_5$Ca core and its protein environment. For more details on the structure see References 46 and 47. The OEC of PSII cluster model has been optimized at the DFT level using the BP86 functional with the Zeroth-order regular approximation (ZORA) \cite{48-50} and ZORA-contracted basis sets \cite{51} (TZVP for Mn, Ca, O, and N atoms and the SARC-SVP for C and H atoms). The atom-pairwise dispersion correction with Becke-Johnson damping is used. \cite{41-43} The system is implicitly solvated ($\varepsilon = 8.0$). For the SCF convergence we choose the tight criterium, while for the gradient optimization we adopt the opt criterium, according to the nomenclature used in ORCA. The resolution of the identity (RI) is used to accelerate the SCF convergence.

The results for $E - E_{\text{conv}}$ for the GVDW, GSES, and PSES models are shown in Figure 8. If we focus first on the results for the GVDW, we observe a very smooth convergence of the energy, without oscillations. This situation is slightly different when using a SES-type cavity, where $E - E_{\text{conv}}$ has some jumps, much smaller for the case of the GSES model than for the PSES model. If we focus on the PSES model, we observe a sudden change in the energy of around 0.125 a.u. (~78 kcal/mol) after 25 optimization cycles, together with several small oscillations in the range of 0.010 and 0.015 a.u. (~6 and 10 kcal/mol). The numerical instabilities associated to the PSES are also displayed in the maximum component of the gradient vector, $\text{Grad}_{\text{MAX}}$, which is plotted in Figure 9 as a function of the optimization cycle.

| TABLE 5 | Ratio between the times spent in computing different properties for the GVDW, GSES with respect to the vacuum case |
|-----------------|-----------------|-----------------|
| Property       | $t_{\text{GVDW}} / t_{\text{vac}}$ | $t_{\text{GSES}} / t_{\text{vac}}$ |
| SCF (cycle)    | 1.076           | 1.078           |
| Gradient (cycle) | 1.068           | 1.081           |

FIGURE 6 Time spent to compute the gradient per optimization cycle, $t_{\text{grad}}$, as a function of $N_{\text{atoms}} \times N_q$. The fits $t_{\text{grad}} = \delta \times (N_{\text{atoms}} \times N_q)\beta$ are also shown for clarity (black line, $\beta = 1.414$, corresponds to the GVDW while the red line, $\beta = 1.387$, to the GSES) [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 7 The 238-atom OEC of PSII cluster model (optimized $S_2$ state model A) used as initial structure for the geometry optimizations using the GVDW, the GSES, and the PSES models. H white, C gray, N blue, O red, Mn purple, Ca yellow [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 8 SCF energy as a function of the optimization cycle for the GVDW, GSES, and PSES models for the OEC of PS-II. The inset shows the same data increasing the range of the y-axis [Color figure can be viewed at wileyonlinelibrary.com]
models. In this case, $\text{Grad}_{\text{MAX}}^{\text{GSES}}$ shows changes of two orders of magnitude. Because of that, some atom coordinates may be substantially affected making the energy to converge later on to a local minimum. This situation is worse in calculations where the initial structure is far away from its final counterpart (data not shown for simplicity) or for solvents where the geometry of the solute is prone to change substantially during its optimization. For these cases, the SCF energy and the gradient present several oscillations of large amplitude for the PSES. This is something that occurs more frequently for large systems but is also observed for small systems.

Regarding, the GSES model, $\text{Grad}_{\text{MAX}}^{\text{GSES}} \approx 3 \times \text{Grad}_{\text{MAX}}^{\text{GVDW}}$. In our implementation of the GSES, all contributions to the gradient are calculated, with no approximations for any of the terms in Equations (17) and (18). Therefore, the results shown in Figures 8 and 9 indicate that when performing a geometry optimization of a big system, the use of the GSES is not as adequate as using the GVDW model. To prove that we have also optimized the structure of Vancomycin (176 atoms) in water. Calculations have been carried out at the DFT level using the B3LYP functional with the def2-SVP basis set, the RI-J approximation and exact analytical integration for the exchange term (RIJDX), and the atom-pairwise dispersion correction with Becke-Johnson damping. The DFT grid is chosen to be Grid7 and verytight criteria are chosen for the SCF energy and the gradient. The number of contracted basis functions is 1,797. The structure of Vancomycin, optimized in vacuum, is shown in Figure 10. This structure is used as a starting point for the calculations using the GVDW and the GSES models.

The convergence of the SCF energy with respect to the geometry optimization cycle is shown in Figure 11. Due to the fact that the calculation using GSES converges much slower than that using the GVDW model, for the GSES model we plot $E - E'$ being $E = (E_{50} - E_{50})/51$, with $E_{50}$ and $E_{100}$ the SCF energies at the 50th and 100th optimization cycle, respectively. Again, as observed for the OEC of PSII, the SCF energy converges relatively smooth for the GVDW model (convergence is achieved after 284 cycles). On the other hand, when the GSES model is adopted, convergence is not reached before 100 optimization cycles with verytight criteria for the SCF and the gradient. If we analyze the values of $\text{Grad}_{\text{MAX}}$ for vancomycin shown in Figure 12, we observe that while for the GVDW model, $\text{Grad}_{\text{MAX}}$ oscillates initially to decrease two orders of magnitude later on, when the GSES model is adopted, $\text{Grad}_{\text{MAX}}$ oscillates indefinitely with the same amplitude. This fact shows that, for this system, there is always an atom for which the module of the gradient vector is larger than the convergence criterium. Then, for the GSES model, if we want the geometry optimization to converge, we
have to decrease the convergence threshold for the maximum component of the gradient vector (also check the criterium for the RMS gradient).

From the results exposed in Figures 8, 9, 11 and 12, we conclude that (a) the GCS should be adopted to prevent numerical instabilities in the SCF and gradient, and (b) for large systems, the use of the GVDW model is recommended instead of the GSES model. Regarding this last point, there are some exceptions. On the one hand, small systems, and on the other hand, systems which contain few explicit solvation layers. For this last set of systems, the GSES model prevents the implicit solvent to fill spaces between solvent molecules and between the solute and the explicit solvent molecules and, therefore, its use is recommended.

3.2 Free energy of solvation

Following the steps described in the computational details section, we have parametrized the dispersion contribution to $\Delta G_{solv}$ for organic molecules containing H, C, N, and O atoms. In order to make the analysis of the results easier, we divide our results into two subsets: (a) solvents with low $\epsilon$ ($\epsilon < 10$) and (b) solvents with mid and large values of $\epsilon$ ($\epsilon > 10$). For completeness, the performance of our approach to compute $\Delta G_{solv}$ has also been tested for a set of cations and anions in water.

3.2.1 Solvents with low $\epsilon$

This subset considers four nonpolar solvents: benzene, chloroform, cyclohexane, and toluene. The results for the fittings of the atomic surface tensions using the GVDW_nel and the GSES_nel models are shown in Table 6. In all cases the R Square coefficients for the multivariate regressions are higher than 0.95, being in most of the cases equal or larger than 0.99. There is almost no difference between the results for the GDW_nel and the GSES_nel models. Indeed, if we take as a reference the value for the GSES_nel model, then the average relative error between both sets of data is equal to 1.7%. In Figure 13, we show the correlation plot between $\Delta G_{solv}^{GVDW}$ using the atomic surface tensions in Table 6 and $\Delta G_{solv}^{GSES}$. The level of agreement between the experimental and the calculated free energies of solvation is very good, being the mean absolute error (MAE) lower than 1.0 kcal/mol (see Table 7). In general, MAE$^{GES}$ < MAE$^{vdW}$ (except for toluene where the number of solutes is slightly smaller than for the other solvents); however, the difference is too small to extract any trend. It is interesting to see the improvement of the calculated energies when adopting the GDVW_nel and GSES_nel with respect to the case of the GVDW and GSES models (see data in parenthesis in Table 7). For this last two models, the MAE is around three times larger than when considering nonelectrostatic contributions to the free energy of solvation showing that $\Delta G_{rel}^{rel}$ represents an important percentage of $\Delta G_{solv}$ for nonpolar solvents.

Although the agreement between calculated and experimental $\Delta G_{solv}$ is good, this quantity should not be the only guideline for the parametrization of a solvation model. Indeed, large polarization effects can occur within a continuum model but be counterbalanced by the term $\Delta G_{rel}$ in Equation (27), giving a good prediction of $\Delta G_{solv}$. Such a large electrostatic contribution to $\Delta G_{solv}$ is due in general, to the choice of a too small cavity for $\Delta G_{rel}$ in Equation (27). A variable in our model that controls the polarization of the solute is the parameter $f$ that scales the vdW radii of the solute atoms. For simplicity, we have considered a solvent-independent value for $f$ equal to 1.2. Previous studies by Luque, Orozco et al. showed, however, that a choice of $f$ in the range 1.20–1.25 for solvents with low $\epsilon$ results in an overestimation of the free energy of polarization and the solvent induced dipole.\[28–30\] For instance, values of $f = 1.6$ and 1.8 are suggested for neutral solutes in carbon tetrachloride ($\epsilon = 2.24$), and chloroform ($\epsilon = 4.90$), respectively.\[28,29\] In order to evaluate our assumption of a solvent-independent $f$, we compare the value of the solute dipole moment in solution, $\mu_s$, to its counterpart in vacuum, $\mu_{vac}$. We restrict the analysis to the GVDW_nel model, as the obtained results are almost the same than those for the GSES_nel model. The corresponding plots $\mu_s$ versus $\mu_{vac}$ are shown in Figure 14, together with the linear fits $\mu_s = a + b \times \mu_{vac}$. For solutes in cyclohexane, we observe an increase of around 16% in the dipole moment. Similar results are observed for benzene and toluene (increase of 19%). These results agree with the increase observed in Reference [29]. A larger enhancement is observed, however, for chloroform where $b = 1.281$ (increase of 28%). This last value does not agree with the value provided in Reference [29] for a series of solutes in chloroform (increase of 12% in $\mu$). Indeed, if we focus on the particular case of ethanol and methanol solvated in chloroform, our results reveal a value for $\Delta \mu = \mu_s - \mu_{vac}$ of 0.37 D and 0.32 D, respectively. These results differ from the QM/MM data provided in Reference [29], with $\Delta \mu$ equal to 0.12 D.
both for ethanol and methanol. If we increase $f$ from 1.2 to 1.8, the calculated $\Delta \mu$ with our approach equal 0.16 D and 0.13 D for ethanol and methanol, respectively. These results confirm that for solutes in chloroform, an increase in $f$ is needed. A similar refinement, could be done for benzene, chloroform, and cyclohexane although it lies outside the scope of this article.

### 3.2.2 | Solvents with $\varepsilon > 10$

For this subset, we consider two solvents: octanol and water. In the case of octanol, the value of $\varepsilon$ is still low ($\varepsilon = 10.30$). Indeed, octanol (C$_8$H$_{18}$O) can be considered as nonpolar due to the fact that the polar nature of the molecule only comes from the O─H bond (being the rest of bonds nonpolar). Then, although $\varepsilon$ is slightly larger than for the case of pure nonpolar solvents, there should not be significant effects from the nonconsideration of a first solvation explicit shell on the computed $\Delta G_{\text{solv}}$. The situation for water is very different. Water is a highly polar molecule with a large value of $\varepsilon$, being able to form hydrogen bonds with other molecules, forming quite rigid solvation shells around solutes and consequently screening out the interaction between them. Then, the derivation of a set of $\sigma_i$ for water that are independent of the protic or aprotic character of the solute is not so clear. The multivariate fittings of $\sigma_i$ for octanol and water for the GVDW$_{\text{nel}}$ and the GSES$_{\text{nel}}$ models are shown in Table 8, while the correlation plots $\Delta G_{\text{calc}}^{\text{solv}}$ vs $\Delta G_{\text{exp}}^{\text{solv}}$ are provided in Figure 15.

The agreement between calculated and experimental solvation energies for octanol is slightly worse than that for the case of nonpolar solvents, as seen in Figure 15a. Nevertheless, most of the points lie in the 1 kcal/mol error region, being the MAE equal to 1.09 kcal/mol and 1.10 kcal/mol for the GVDW$_{\text{nel}}$ and the GSES$_{\text{nel}}$ models, respectively (see Table 9). Again, the effect of the type of solute cavity is negligible on $\Delta G_{\text{solv}}$. The electrostatic contribution to $\Delta G_{\text{solv}}$ plays a larger role for octanol than for nonpolar solvents. This can be seen from the value of the MAE for the GVDW and GSES models, which equals 1.58 kcal/mol and 1.54 kcal/mol, respectively, that is, 1.5 times
larger than the same quantity for the GVDW_nel and the GSES_nel (while it was three times larger for nonpolar solvents).

For water, the number of calculated points that deviate more than 1 kcal/mol from their experimental counterpart is larger than for octanol (see Figure 15b), being the MAE equal to 1.27 kcal/mol for both type of solute cavities (see Table 9). This value is not so far from the MAE for the GVDW and GSES models (1.68 and 1.64 kcal/mol), showing that for water, the nonelectrostatic contributions play a smaller effect than for solvents with low $\varepsilon$.

Indeed, if we plot $\mu_S$ versus $\mu_{vac}$ for water using the GVDW model (see Figure 16), we observe an increase of around 36% in $\mu$. This enhancement agrees with the value of 35% reported in Reference [30]. If we assume $\Delta G_{\text{sol}}$ to be linearly proportional to the SASA, as it is suggested from experimental transfer data for hydrocarbons into water,[52] then we obtain a MAE equal to 1.67 kcal/mol. This confirms that the strategy described in Equations (29) and (30) is more suitable than simply taking $\Delta G_{\text{sol}}$ to depend on the total SASA. However, as mentioned before, part of the MAE shown in Table 9 for water may come from the fact that our approach considers the same values for $\sigma_H$, $\sigma_C$, $\sigma_N$, and $\sigma_O$, whatever is the content of the solute. In order to prove that, we split out our training set for water into four different subsets: solutes that just contain (a) H and C atoms, (b) H, C, and N atoms (c) H, C, and O atoms, and (d) H, C, N, and O atoms. The number of solutes per each subset is shown in Table 10. More information about the solutes used per subset is provided in Table S1. After having defined the subsets, we fit the $\sigma_i$ for each of them and compare $\Delta G_{\text{calc}}^{\text{sol}}$ with $\Delta G_{\text{exp}}^{\text{sol}}$. The corresponding results for the $\sigma_i$ are shown in Table 11, the correlation plots for the solvation energy are provided in Figure 17, and finally the MAE for each subsets and solute cavity are shown in Table 12. The improvement in the results is rather important, as seen from Figure 17 and Table 12, with all subsets having now a MAE lower than 1 kcal/mol, regardless of the solute cavity used. In order to understand the reasons behind the improvement, we analyze the values of $\sigma_i$ per each subset (Table 11). First, we observe that $\sigma_N$ for solutes that contain N atoms is twice the same quantity for solutes that do not contain N atoms. Indeed, molecules with N atoms have N–H bonds,

**FIGURE 14** Correlation plot between the dipole moment in solution, $\mu_S$, and the same quantity in vacuum, $\mu_{vac}$, for solvents with low $\varepsilon$ for the GVDW_nel model. The slope $b$ of the linear fit ($\mu_S = a + b \times \mu_{vac}$) and the corresponding correlation coefficient $r$ are shown for completeness.

**TABLE 8** Atomic surface tension $\sigma_{vdW}$ and $\sigma_{SES}$ for H, C, N and O for solvents with $\varepsilon > 10$

| Solvent  | $\sigma_{vdW,H}^{\text{vdW}}, \sigma_{H}^{\text{vdW}} \times 10^5$ (a.u.) | $\sigma_{vdW,C}^{\text{vdW}}, \sigma_{C}^{\text{vdW}} \times 10^5$ (a.u.) | $\sigma_{vdW,N}^{\text{vdW}}, \sigma_{N}^{\text{vdW}} \times 10^5$ (a.u.) | $\sigma_{vdW,O}^{\text{vdW}}, \sigma_{O}^{\text{vdW}} \times 10^5$ (a.u.) |
|---------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Octanol | $-0.349/-0.353$                                 | $-0.466/-0.487$                                 | $1.092/1.089$                                    | $0.286/0.240$                                    |
| Water   | $-1.174/-1.184$                                 | $-1.222/-1.234$                                 | $-2.788/-2.780$                                 | $-1.272/-1.329$                                 |
which form hydrogen bonds easily with water. That should be something to account for when computing $\Delta G_{solv}$. That is, considering a different value for $\sigma_H$ depending on the fact that it is bonded to N. This point is important, especially considering that hydrogen atoms are in general the atoms that are present in a larger amount in organic molecules, and then whose area contribute the most to the SASA. The values of $\sigma_C$ are also slightly different depending on whether N is present in the solute. If it is not the case, then we observe a similar value for $\sigma_C$. Indeed, for the case of solutes containing just H and C atoms, which do not form strong hydrogen bonds with water, the values of $\sigma_H$ and $\sigma_C$ are very similar to those observed in Table 6 for nonpolar solvents. The fact that for the last subset in Table 10 (solutes with H, C, N, and O atoms) we just have 12 solutes prevents us to really infer the effect of having a N atom on $\sigma_C$. The same can be argued for $\sigma_O$ and $\sigma_C$. Of course, the reported values for $\sigma_I$ would slightly vary if other atom types were considered in the fitting. However, as pointed out before, the aim of this manuscript regarding the free energy of solvation is not to derive here a universal implicit solvation model but (a) to show that an approach that considers a solvent-dependent $R_S$ to construct the SASA can lead to accurate results for $\Delta G_{solv}$ and (b) to stress the fact that the atomic surface tensions $\sigma_I$ should depend on the atomic content of the solute for solvent with a large value of $\varepsilon$. With the results shown above, we believe that both points are sufficiently addressed.

### 3.2.3 | Ions in water

Solvation energies of ions in water are larger than those for their neutral counterparts. In general, implicit solvation models perform worse...
for ions than for neutrals.\textsuperscript{[21,27]} One of the main sources of error can be found in the use of the same cavity to treat electrostatic solvation effects for both uncharged and charged solutes. Ab initio computational studies aiming at the calculation of free energies of hydration of ions suggest that the use of a scaling factor $f$ equal to 1.20–1.25 for the vdW radii results in an underestimation of the electrostatic part of $\Delta G_{\text{solv}}$.\textsuperscript{[27]} A small reduction of $f$ to 1.15 improves the agreement between the calculated and the experimental $\Delta G_{\text{solv}}$, substantially, suggesting that both for anions and cations a shrinkage in the solute cavity occurs in solution. In order to prove this statement, we have calculated $\Delta G_{\text{solv}}$ for a set of cations and anions in water using the GVDW\textsubscript{nel} and GSES\textsubscript{nel} models with three different values for $f$ (1.10, 1.15, and 1.20) using the values for the atomic surface tensions provided in Table 8. The corresponding results are shown in Table 13. The results for cations show a significant improvement for both the MAE and the percentage error (100 × $\Delta G_{\text{calc,solv}} - \Delta G_{\text{exp,solv}}$ / $\Delta G_{\text{exp,solv}}$ ) when decreasing $f$, in agreement with Reference [27]. The use of $f = 1.15$ results in a MAE of 4.41 and 4.74 kcal/mol for the GVDW\textsubscript{nel} and the GSES\textsubscript{nel} models, respectively. These values are similar to those reported for the SMD model for cations.\textsuperscript{[21]} A further reduction of $f = 1.10$ gives very accurate results for $\Delta G_{\text{solv}}$ with a MAE lower than 3 kcal/mol and a percentage error lower than 5%. Although decreasing $f$ also improves the agreement between calculated and experimental data for the case of anions, the magnitude of the MAE and the percent error is larger than for cations. For $f = 1.10$, both the MAE and the percent error for anions using the GVDW\textsubscript{nel} and the GSES\textsubscript{nel} models are around 2.5 times larger than the same quantities for cations. A further decrease of $f$ ($f = 1.00$) would improve the results for anions but it is not recommended as it can lead to unrealistic estimates of the electrostatic part of $\Delta G_{\text{solv}}$.\textsuperscript{[27]} Refining the atomic surface tensions in Equation (30) for anions should not change the accuracy substantially as electrostatic effects dominate over non-electrostatic effects in ions. A better approach would be to refine slightly the vdW radii for the solute atoms used to construct the vdW-type or the SES-type cavity for the treatment of solvation electrostatic effects. However, such a change would involve a careful

### Table 11: Atomic surface tension $\sigma_{\text{vdW}}$ and $\sigma_{\text{SES}}$ for solutes with different atomic content in water

| Content | $\sigma_{\text{vdW}}$ (H) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{ SES}}$ (H) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{vdW}}$ (C) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{ SES}}$ (C) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{vdW}}$ (N) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{ SES}}$ (N) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{vdW}}$ (O) × 10\textsuperscript{5} (a.u.) | $\sigma_{\text{ SES}}$ (O) × 10\textsuperscript{5} (a.u.) |
|---------|---------------------------------|---------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| HC      | $-0.831/-0.829$                | $-1.463/-1.485$                | -              | -              | -              | -              | -              | -              |
| HCN     | $-1.936/-1.945$                | $-0.970/-0.992$                | $-0.346/-0.365$ | -              | -              | -              | -              | -              |
| HCO     | $-1.021/-1.044$                | $-1.345/-1.385$                | -              | -              | -              | -              | $-1.796/-1.781$ | -              |
| HCNO    | $-1.838/-1.865$                | $-2.102/-2.055$                | $-3.870/-4.000$ | 0.675/0.618    | 0.675/0.618    | 0.675/0.618    | 0.675/0.618    | 0.675/0.618    |

### Table 12: Mean absolute error (MAE) of the calculated solvation energies $\Delta G_{\text{calc,solv}}$ with respect to their experimental counterpart $\Delta G_{\text{exp,solv}}$ for solutes with different atomic content in water for the GVDW\textsubscript{nel} and GSES\textsubscript{nel} models

| Content | MAE\textsubscript{GVDW\textsubscript{nel}} (kcal/mol) | MAE\textsubscript{GSES\textsubscript{nel}} (kcal/mol) |
|---------|-----------------------------------------------|----------------|
| HC      | 0.74                                          | 0.75            |
| HCN     | 0.87                                          | 0.86            |
| HCO     | 0.99                                          | 0.97            |
| HCNO    | 0.46                                          | 0.42            |
parametrization of both the GVDW_nel and the GSES_nel models that will not be addressed in the present study.

4 | CONCLUSIONS

In this work, we have studied the effect of the solute cavity (vdW-type and SES-type cavities) on the properties of solvated systems using the GCS within the framework of the C-PCM. First, we have analyzed the influence of the total number of charges spread over the surface of the solute cavity on the accuracy of the final SCF energy and the cost of geometry optimizations for a large set of solutes in water. We observe that for a number of charges per sphere that forms the solute cavity (Nq) equal or larger than 110, the deviation between the SCF energy with respect to the case of Nq = 770 (high density of charges) is lower than 0.05 kcal/mol, irrespective of the type of solute cavity employed. With respect to the computational cost of the calculations, there is a reduction of 20 and 30% in the calculation of the SCF cycles and the gradient, respectively, when Nq ≤ 194 as compared to the case Nq = 770 when using either the vdw-type or the SES-type solute cavities. For the aforementioned reasons, we recommend Nq to be chosen equal to 110 or 194, especially for big systems. The cost of the calculations using the GCS versus the same quantity for the point charge scheme has also been studied. Starting with the case of small systems (Natoms ~ 100), we observe no change in the time spent in computing the SCF iterations and the gradient. The same occurs when we compare the results using the vdw-type and the SES-type cavities within the GCS between each other. Regarding the geometry optimizations of big systems (Natoms > 100), the SCF energy of the oxygen evolving complex of photosystem II (238 atoms) and vancomycin (176 atoms) converges smoothly for the case of the vdw-type solute cavity using the GCS. When the SES-type cavity is adopted within the GCS, the SCF energy converges much slower. In this case, the maximum component of the gradient vector oscillates almost indefinitely with a small amplitude. Therefore, when SES-type cavities are used, it is recommended to soften slightly the thresholds for the gradient in order to reach convergence. In any case, the use of the GCS prevents sudden jumps in the SCF energy observed when adopting the point charge scheme, and which are on the order of 100 kcal/mol.

Finally, we have parametrized the nonelectrostatic part of the free energy of solvation, ΔG_solv, for a large set of organic molecules in different polar and nonpolar solvents within the GCS. The solvent-accessible surface, needed to compute ΔG_solv, is generated by considering the radius of the solvent probe sphere to be dependent on thermophysical data for the solvent under interest. We have assumed a solvation-dispersive term in the ΔG_solv proportional to the contribution of each atom to the SASA via a set of atomic surface tensions, σ. The agreement between calculated and experimental ΔG_solv for the nonpolar solvents is very good, with a deviation lower than 1 kcal/mol between both sets of data, regardless of the type of solute cavity used. For polar solvents like water and octanol, the situation is slightly different with a deviation between simulation and experiment larger than 1 kcal/mol (1.10 and 1.27 for octanol and water, respectively). Again, there is no influence of the type of solute cavity on the calculated energies. For water, we demonstrate that in order to obtain accurate enough solvation energies one should consider a different value for σ_H depending on the immediate environment of the H atom and which type of bond it is forming. This conclusion is extendable to other type of atoms. In order to predict accurate ΔG_solv for the case of ions, one needs to reduce the scaling factor for the vdw radii used to construct the solute cavity for electrostatics from 1.20 to 1.15–1.10. Our results show that the GCS can be used together with different functional forms for the nonelectrostatic contribution to ΔG_solv to predict accurate solvation data.

ACKNOWLEDGMENTS

FN and MG gratefully acknowledge financial support of this work by the Max Planck Society, the DFG SPP 1601. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – EXC 2033 – Projektnummer 390677874 (funded by the Deutsche Forschungsgemeinschaft [DFG, German Research Foundation] under Germany’s Excellence Strategy – EXC 2033 – Projektnummer 390677874).

ORCID

Miquel Garcia-Ratès
https://orcid.org/0000-0002-6315-0845

REFERENCES AND NOTES

1. B. Mennucci, R. Cammi, In Continuum Solvation Models In Chemical Physics: From Theory To Applications, John Wiley & Sons, Chichester, U.K. 2007, p. 1.
2. S. Miertus, E. Scrocco, J. Tomasi, Chem. Phys. 1981, 55, 117.
3. S. Miertus, J. Tomasi, Chem. Phys. 1982, 65, 239.
4. R. Cammi, J. Tomasi, J. Comput. Chem. 1995, 16, 1449.
5. E. Cancè, B. Mennucci, J. Tomasi, J. Chem. Phys. 1997, 107, 3032.
6. B. Mennucci, E. Cancè, J. Tomasi, J. Phys. Chem. B 1997, 101, 10506.
7. E. Cancè, B. Mennucci, J. Math. Chem. 1998, 23, 309.
[30] R. S. Rowland, R. Taylor, J. Phys. Chem. A 2005, 109, 3927.
[31] D. V. Pantazis, W. Ames, C. J. Marenich, D. G. Truhlar, F. Neese, D. A. Pantazis, Chem. Sci. 2015, 6, 1678.
[32] A. Klamt, G. Schüürmann, V. Barone, M. Cossi, Angew. Chem. Int. Ed. 2009, 48, 761.
[33] V. Barone, S. Grimme, J. Comput. Chem. 2001, 22, 1180.
[34] A. J. Short, M. Karplus, J. Chem. Phys. 2003, 119, 3827.
[35] A. J. Short, M. Karplus, J. Chem. Phys. 2003, 119, 3827.
[36] F. Neese, D. A. Pantazis, J. Comput. Chem. 2009, 30, 922.
[37] How to cite this article: Garcia-Ratés M, Neese F. Effect of the Solute Cavity on the Solvation Energy and its Derivatives within the Framework of the Gaussian Charge Scheme. J Comput Chem. 2020;41:922–939. https://doi.org/10.1002/jcc.26139

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.