Origin of parameter degeneracy and molecular shape relationships in geometric-flow calculations of solvation free energies

Michael D. Daily,[1] Jaehun Chun,[2] Alejandro Heredia-Langner,[3] Guowei Wei,[4] and Nathan A. Baker[5]

[1]*Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA*
[2]*Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA*
[3]*National Security Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA*
[4]*Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, USA*
[5]*Computational and Statistical Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA*

(Received 9 August 2013; accepted 31 October 2013; published online 27 November 2013)

Implicit solvent models are important tools for calculating solvation free energies for chemical and biophysical studies since they require fewer computational resources but can achieve accuracy comparable to that of explicit-solvent models. In past papers, geometric flow-based solvation models have been established for solvation analysis of small and large compounds. In the present work, the use of realistic experiment-based parameter choices for the geometric flow models is studied. We find that the experimental parameters of solvent internal pressure $p = 172$ MPa and surface tension $\gamma = 72$ mN/m produce solvation free energies within 1 *RT* of the global minimum root-mean-squared deviation from experimental data over the expanded set. Our results demonstrate that experimental values can be used for geometric flow solvent model parameters, thus eliminating the need for additional parameterization. We also examine the correlations between optimal values of p and γ which are strongly anti-correlated. Geometric analysis of the small molecule test set shows that these results are inter-connected with an approximately linear relationship between area and volume in the range of molecular sizes spanned by the data set. In spite of this considerable degeneracy between the surface tension and pressure terms in the model, both terms are important for the broader applicability of the model. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4832900]

I. INTRODUCTION

Implicit solvent models have received much attention in the past two decades due to their low computational cost and relatively high accuracy. Such models consist of a "nonpolar" free energy functional that accounts for cavity creation and dispersive interactions and a polar free energy functional that accounts for the difference in charging free energies of the solute between vacuum and solvent. While both of these terms depend on the solute-solvent boundary position and the resulting position-dependent dielectric, the polar and nonpolar functionals are often optimized independently. For example, different arbitrary choices of the boundary (e.g., the van der Waals surface[1] or molecular surface[2]) may be used for calculating solvent-accessible surface area (SASA) and the position-dependent local dielectric coefficient, respectively. To address this problem, some groups have recently developed methods that couple formalisms for the two functionals so that a single, optimal solvent-solute boundary can be estimated. For example, Dzubiella *et al.* proposed minimization of the solvation free energy with respect to a solvent volume exclusion function[3,4] and Bates *et al.* introduced surface definitions via surface free energy minimization.[5] Recently, we have developed an approach to describe the solute-solvent interface using a potential-driven geometric flow model.[6,7] The key parameters in the geometric flow approach, such as solute and solvent dielectric constants ε, solvent internal pressure p, and surface tension γ can be systematically optimized for any training set of small molecules.[8] However, such parameters would ideally be based on experimental measurements to provide more physical relevance and to remove unnecessary free parameters from the model to improve robustness and generalizability. Optimal choices for parameters such as solvent pressure and surface tension have been shown to vary significantly over a range of possible parameters with strong anti-correlation between these two quantities.[8]

Although a wide range of values are used in practice, a reasonable value of the solute dielectric constant ε_m has been estimated at 1.8 based on the high-frequency contributions to molecular polarizability.[9] As explained by Marcus,[10] the solvent internal pressure p is a nebulous concept that represents incremental isothermal stretching of local interactions, but without breaking the solvent intermolecular attractive forces. In particular, the internal pressure is defined as $p = (\frac{\partial U}{\partial V})_T$, the rate of change of the internal energy U with respect to the volume V at given temperature T. Experimental measurements estimate the solvent internal pressure at 172 MPa (0.0248 kcal mol^{-1} Å^{-3}) which is about 2000 times higher than atmospheric pressure.[10] Surface tension γ is characterized by the energy required to change the area of an interface and is often associated with the energetics of hydrogen bonding structure of water molecules near the solute/solvent

interface. The experimental measurement for the water-air interface at 25 °C is 72 mN/m (0.103 kcal mol^{-1} Å^{-2}).[11] Because of water's high number and strength of hydrogen bonds, its surface tension is larger relative to its internal pressure than for organic liquids.[10]

While models considering only the area of the cavity have been traditionally popular,[12,13] the recent inclusion of solute volume and nonpolar dispersive interactions in implicit solvation models is an important improvement.[3,4,6,14,15] Considerations from scaled particle theory[16,17] and multiple recent studies concerning the solute size-dependence of the hydrophobic effect[18–20] motivated this innovation which improves predictions of solvation forces and energies.[3,4,15,21,22] In the present work, we combine these improvements with the use of realistic experiment-based parameters in the context of the geometric flow solvation model. We find that the experimental parameters of solvent internal pressure $p = 172$ MPa and surface tension $\gamma = 72$ mN/m produce solvation free energies within 1 *RT* of the global minimum root-mean-squared deviation from experimental data over a set of 58 molecules. Our results demonstrate that experimental values can be used for geometric flow solvent model parameters, thus eliminating the need for additional parameterization. It is worth noting that we focus on the applicability of the experimental-based solvation parameters with relevant physics in the geometric flow solvation model; the optimization of the force-field charge and radius parameters is not pursued.

II. METHODS

A. The geometric flow solvation model

The geometric flow based solvation model is briefly summarized below; more details are provided in previous publications.[6–8,23,24] The total free energy functional for the solute-solvent system (G) can be written as the sum, $G[\chi, \phi] = G_{\rm p}[\chi, \phi] + G_{\rm np}[\chi, \phi]$, of the polar free energy functional ($G_{\rm p}$) and a nonpolar free energy functional ($G_{\rm np}$). In the absence of mobile ions, the polar free energy functional is described by

$$G_{\rm p}[\chi, \phi] = \int_\Omega \left(\chi \left(\varrho_f \phi - \frac{1}{2} \varepsilon_m \|\nabla \phi\|^2 \right) + (1-\chi) \left(-\frac{1}{2} \varepsilon_s \|\nabla \phi\|^2 \right) \right) d\mathbf{x}, \quad (1)$$

where Ω is the problem domain, χ is a solvent accessibility indicator or characteristic function varying smoothly from 1 at the solute van der Waals surface to 0 in the bulk solvent. More specifically, χ defines a smooth interface between van der Waals and solvent accessible surfaces in a thermodynamically self-consistent manner, coupled with the local charges and electrostatic potential. The distribution function ϱ_f denotes the fixed charge distribution of the fixed solute molecule, the scalars ε_m and ε_s are the dielectric constants of the solute and solvent, respectively, and ϕ is the electrostatic potential. The nonpolar free energy functional is described by

$$G_{\rm np}[\chi, \phi] = \gamma A + pV + \rho_0 \int_\Omega (1-\chi) U_{\rm vdW}^{\rm att} d\mathbf{x}, \quad (2)$$

where γ is the solvent surface tension, A is the surface area of the solute, p is the solvent pressure, V is the volume of the solute, and ρ_0 is the solvent bulk density. The function $U_{\rm vdW}^{\rm att}$ is the attractive potential of the van der Waals dispersion interaction between the solute and the solvent, which can be represented by a summation of the attractive interaction potential (using a Weeks-Chandler-Anderson decomposition[25]) for each atom. The area and volume can be calculated directly from the characteristic function χ via

$$A = \int_\Omega \|\nabla \chi\| d\mathbf{x}, \quad (3)$$

$$V = \int_\Omega \chi d\mathbf{x}. \quad (4)$$

The polar and the nonpolar free energy functionals are coupled via the characteristic function χ. Therefore, extremizing the total free energy G with respect to ϕ and χ leads to two coupled partial differential equation. The first equation is a generalized Poisson equation which governs ϕ,

$$-\nabla \cdot (\epsilon(\chi) \nabla \phi) = \chi \varrho_f, \quad (5)$$

where the dielectric function $\varepsilon(\chi)$ is defined as

$$\epsilon(\chi) = \varepsilon_m \chi + \varepsilon_s (1-\chi) \quad (6)$$

such that it achieves the solute dielectric constant value ε_m in the solute interior and the solvent dielectric constant value ε_s in the exterior. The second equation resulting from variation of G is the generalized geometric flow equation which governs χ,

$$-\nabla \cdot \left(\gamma \frac{\nabla \chi}{\|\nabla \chi\|} \right) + w(\phi) = 0, \quad (7)$$

where w is a driving potential for the flow

$$w(\phi) = p - \rho_0 U_{\rm vdW}^{\rm att} + \varrho_f \phi - \frac{1}{2} (\varepsilon_m - \varepsilon_s) \|\nabla \phi\|^2. \quad (8)$$

Solving Eqs. (5) and (7) together provide a self-consistent definition of both the electrostatic potential ϕ and the solvent density, defined via the solvent accessibility indicator function as $1 - \chi$. The solvation energy can be determined from these functions,

$$\Delta G_{\rm solv}[\chi, \phi] = G[\chi, \phi] - \int_\Omega \varrho_f \psi d\mathbf{x}, \quad (9)$$

where ψ is the electrostatic potential in the presence of a medium with the same dielectric constant as that of the solute.

B. Numerical methods

A grid-based optimization was carried out in (p, γ) space with p ranging from 0.001 to 0.055 kcal mol^{-1} Å^{-3} and γ ranging from 0.055 to 0.165 kcal mol^{-1} Å^{-2}, and a spacing of 0.005 along both of these axes. ε_m was held constant at 1.8 per the work of Leontyev and Stuchebrukhov,[9] ε_s (solvent dielectric) at 80, solvent density at 0.0334 Å^{-3}, and the minimum molecule-box edge distance at 3.8 Å. The equations were solved using the second-order central finite difference scheme discussed in Chen *et al.*,[6] and the solver grid spacing was set at 0.25 Å. The experimental solvent internal pressure at 172 MPa[10] converts to 0.0248 kcal mol^{-1} Å^{-3}, and

the experimental surface tension of 72 mN/m[11] converts to 0.103 kcal mol^{-1} Å^{-2}. Linear regression fits of $\gamma_{\min}(p)$ vs. p were performed with scipy (www.scipy.org).

C. Small molecule test sets

We investigate three different sets of molecules to provide a diverse range of molecular sizes and chemical properties. First, we re-examine the SAMPL0 set compiled originally by Nicholls *et al.*[26] that was analyzed by Thomas *et al.*[8] Second, the linear, branched, and cyclic alkane set of Levy and Gallicchio[27] provides a basic set of nonpolar molecules with a range of geometries. Third, the SAMPL2 set[28] provides twice as many molecules as SAMPL0, with a broader range of experimental solvation energies, from −25 to +5 kcal/mol, for more robust testing of the methods. Prominent types of molecules in this set include uracils, parabens, and carboxylic acids.

For alkane and SAMPL0 sets, the charges, van der Waals radii, and well depth parameters were taken directly from the OPLS-AA (Optimized Potentials for Liquid Simulations - All Atom) force field.[29] For SAMPL2, these parameters were taken directly from Klimovich and Mobley,[28] who used Generalized Amber Forcefield (GAFF)[30,31] van der Waals parameters and computed charges using AM1-BCC.[32,33] We used the same approach to generate GAFF parameters for the SAMPL0 set molecules in Antechamber.[30]

Although the ZAP-9 forcefield performed well in our previous analysis of SAMPL0,[8] the OPLS-AA force field was chosen because it employs van der Waals interactions between solvent and solute, and thus experimental p and γ are likely to produce reasonable solvation free energies given the importance of the van der Waals term in the nonpolar free energy functional (Eq. (2)).

III. RESULTS AND DISCUSSION

A. Overall performance of the geometric flow method for a range of pressures and surface tensions

To test the hypothesis that the geometric flow solvation model can predict reasonable solvation free energies with experimental or near-experimental parameters, we analyzed the root-mean-squared error (RMSE) for small molecule solvation energy in the space of p and γ parameter values for different sets of molecules. Figure 1 shows that there is a linear "valley" region in (p, γ) space along which the RMSE varies by less than RT by comparison to the minimum value for the entire surface. This linear valley covers a wide range of pressures, $(0.001 < p < 0.055)$ kcal mol^{-1} Å^{-3}, but a narrow range of surface tensions ±0.01 kcal mol^{-1} Å^{-2} for γ at any given p.

For a more quantitative analysis, we calculated linear fits for $\gamma_{\min}(p)$, the value of γ which minimizes the RMSE at a given p, for each set of molecules. Table I shows that the resulting slopes are −0.74 to −0.78 Å, while the intercepts range from 0.11 to 0.12 kcal mol^{-1} Å^{-2}. The Pearson correlation coefficient was $R^2 > 0.98$ for all three sets and the pooled set. As described below, we also analyzed linear correlations for random subsets of the pooled set. Figure 2 shows that these parameter estimates are robust in cross-validation tests. Specifically, we examined fitting parameter distributions among a large number (10 000) of random subsamples of n molecules from the pooled set of $N = 58$ compounds, at varying levels of n. For $n = N - 5 = 53$, among the 10 000 random samples, the intercept estimate varies by less than 0.001 from the pooled-set value of 0.119 and the slope varies by less than 0.04 from the pooled-set value of −0.78. Pearson R^2 values for the $\gamma_{\min}(p)$ vs. p fits vary from 0.985 to 0.991 among the 10 000 subsamples of size $n = 53$. Even for $n = N/2 = 29$, the estimated intercept and slope vary by less than 0.005 and less than 0.05, respectively, relative to the full set (Figure S1 of the supplementary material[34]). Furthermore, the R^2 for $\gamma_{\min}(p)$ vs. p is 0.94 or higher for each molecule in the pooled set, with the intercept ranging from 0.09 to 0.14 (see Table S1 of the supplementary material[34]). These high R^2 values indicate that the negative linear correlation between p and γ is an inherent property of small molecule solvation in water and not merely an average phenomenon.

Table I also presents the errors $\mathrm{RMSE}_{\mathrm{exp}}$ for solvation free energy using experimental values for the pressure, $p = 0.025$ kcal mol^{-1} Å^{-3}, and surface tensio,n $\gamma = 0.103$ kcal mol^{-1} Å^{-2}, parameters.[10] For the pooled set, $\mathrm{RMSE}_{\mathrm{exp}} = 3.72RT$ which is very close to the global minimum error, $\mathrm{RMSE}_{\min} = 3.21RT$ at $(p, \gamma) = (0.045, 0.085)$. The difference between RMSE values is denoted by $\Delta\mathrm{RMSE} = \mathrm{RMSE}_{\mathrm{exp}} - \mathrm{RMSE}_{\min}$; the largest ΔRMSE is 1.15 RT for the SAMPL0 set. In addition, Figure S2 of the supplementary material[34] shows that the small ΔRMSE is robust in cross-validation tests. In 10 000 size $n = 53$ samplings, ΔRMSE ranges from 0.35 to 0.75 for the majority of subsets and is less than $1RT$ for all sets. Even when the sample size is decreased from $n = 53$ to $n = 38$, ΔRMSE never exceeds $1.5RT$, and n has to be reduced to 18 $(0.31N)$ to find any subset for which ΔRMSE exceeds the approximate thermal noise level of $2RT$.

Furthermore, our results are consistent with several previous investigations which estimated optimal water internal pressures of 0.03–0.09 kcal mol^{-1} Å^{-3} for fitting implicit solvent models with a pressure-volume energy term to molecular dynamics (MD) simulation predictions of solvation forces on proteins.[15,22,35] These observations, and our optimal surface tension and pressure estimates, are well explained by an internal pressure of $p = 0.025$ kcal mol^{-1} Å^{-3} but poorly explained by the use of atmospheric pressure.[10]

In addition to statistical validation, we can demonstrate that these results are robust to the choice of force field used to model the small molecules. For the SAMPL0 set, we generate GAFF[30,31] parameters and compare the solvation free energy predictions from these parameters to those obtained from OPLS parameters. The resulting RMSE vs. (p, γ) landscape is shown in Figure S3 of the supplementary material,[34] and Table I shows that, for the SAMPL0 set, the slope and intercept of $\gamma_{\min}(p)$ vs. p are very similar regardless of whether OPLS or GAFF charges and radii are used. In addition, $\mathrm{RMSE}_{\mathrm{exp}}$ and ΔRMSE are lower for SAMPL0-GAFF than SAMPL0-OPLS, which suggests that the auto-generated GAFF parameters are actually superior to OPLS parameters when used in the geometric flow model. These results show

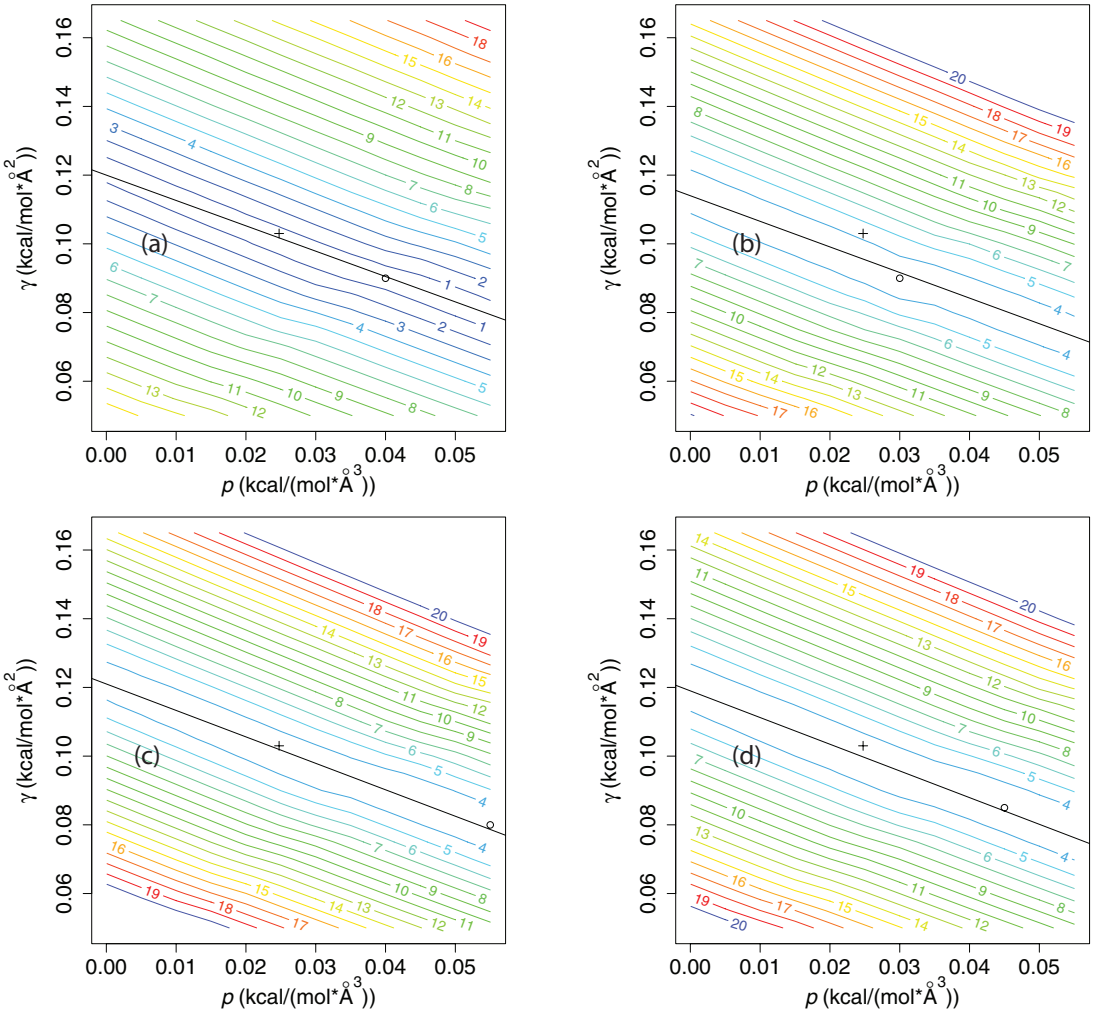

FIG. 1. Root-mean-squared error (RMSE) in solvation free energy for different sets of molecules as a function of solvent internal pressure p and surface tension γ for a solute dielectric constant $\varepsilon_m = 1.8$ using the OPLS-AA force field.[29] (a) Linear, branched, and cyclic alkane set of Levy and Gallicchio;[27] (b) SAMPL0 set;[26] (c) SAMPL2 set;[28] (d) pooled set. The RMSE is normalized by $RT = 0.592$ kcal mol^{-1} at 298 K, shown as contours. The linear regression fit of $\gamma_{\min}(p)$ vs. p is indicated in black, where $\gamma_{\min}(p)$ is the choice of γ at any given p which minimizes the RMSE (values provided in Table I). The experimental values for the pressure[10] $p = 0.0248$ kcal mol^{-1} Å^{-3} and surface tension[11] $\gamma = 0.103$ kcal mol^{-1} Å^{-2} are indicated with a cross on each plot and the minimum (γ, p) values are indicated with a circle.

TABLE I. Solvent pressure and surface tension relationships. The intercept and slope were determined from a linear regression fit of $\gamma_{\min}(p)$ vs. p, where $\gamma_{\min}(p)$ is the value of γ which minimizes the RMSE at a given p. R^2 is the Pearson correlation coefficient value for the linear fit. Numbers in brackets indicate the 95% confidence interval for calculated values. RMSE_{exp} is the solvation energy error in units of $RT = 0.592$ kcal mol^{-1} when using experimental values[10] for $p = 0.0248$ kcal mol^{-1} Å^{-3} and surface tension $\gamma = 0.103$ kcal mol^{-1} Å^{-2}. RMSE_{min} is the error found when scanning the space of (p, γ) parameters and choosing the $p_{\min}$ and $\gamma_{\min}$ values which minimize the error.

Set	Alkanes	SAMPL0 (OPLS)	SAMPL0 (GAFF)	SAMPL2	Pooled
R^2	0.99	0.99	0.99	0.98	0.99
Slope	−0.74	−0.75	−0.74	−0.77	−0.78
(Å)	[−0.80, −0.68]	[−0.80, −0.69]	[−0.80,−0.68]	[−0.84, −0.70]	[−0.83,−0.72]
Intercept	0.120	0.114	0.118	0.121	0.119
(kcal mol^{-1} Å^{-2})	[0.118, 0.122]	[0.112, 0.116]	[0.116, 0.120]	[0.119, 0.123]	[0.117, 0.121]
$p_{\min}$	0.040	0.030	0.025	0.055	0.045
(kcal mol^{-1} Å^{-3})					
$\gamma_{\min}$	0.090	0.090	0.100	0.080	0.085
(kcal mol^{-1} Å^{-2})					
RMSE_{exp}	1.03	4.69	4.12	3.61	3.72
(RT)					
RMSE_{min}	0.32	3.47	4.12	3.27	3.21
(RT)					

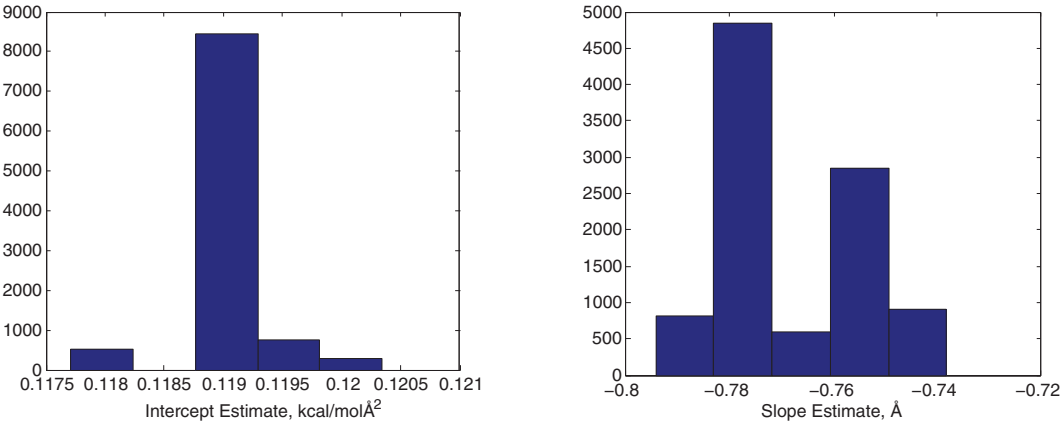

FIG. 2. Histograms of the intercept (left panel) and slope (right panel) for linear fits of the optimal surface tension $\gamma_{\min}$ and p based on 10 000 random sets of 53 small molecule compounds drawn randomly and without replacement from the set of 58 compounds.

that the geometric flow results are robust to minor variations in force field. More detailed cross-validations of the optimal pressure and surface tension values can also be found in the supplementary material.[34]

B. Performance of the geometric flow solvation model for individual molecules

While the RMSE differences between experimental (p, γ) and the global minimum are small, Table S1 of the supplementary material[34] shows that for individual molecules, RMSE difference between experimental (p, γ) and the global minimum averages 2.96 *RT* and can be as high as 10 *RT* for *N*, *N*-4-trimethylbenzamide and *N*, *N*-dimethyl-*p*-methoxybenzamide. Many molecules with RMSE differences above 3 *RT* are nitrogen-rich compounds, including imidazole, uracils, caffeine, cyanuric acid, and benzamides. One unusual property of such molecules is that they form very strong hydrogen bonds with water; this may be poorly approximated by geometric flow and warrant further investigation. In another study,[23] it was shown that errors for benzamides can be reduced with different charged assignments obtained from the density functional theory on a different set of atomic coordinates. Molecules with ether linkages (e.g., diethoxyethane) also tend to perform poorly, with the exception of dimethoxymethane.

C. Scaling relationships between small molecule volumes and areas

Figures 3(a) and S1 of the supplementary material[34] illustrate the relationships between volumes and areas calculated using the geometric flow models described above. Below, we offer two complementary interpretations of the observed strong correlation between volume and area.

A linear model with no intercept fits poorly to the data with a slope of 0.92 ± 0.01 Å, RMSD residuals of 74 Å^3, and Pearson correlation coefficient $R^2 = 0.997$. With a floating intercept, the model fits much better with a slope of 1.07 ± 0.01 Å, RMSD = 23 Å^3, and Pearson correlation coefficient $R^2 = 0.990$; however, the intercept obtained is −27 Å^3, or more than 10% of the median volume for the unified set. Additionally, a strictly "spherical" model ($V \propto A^{3/2}$ with no intercept) also performs badly (RMSD = 219 Å^3). Thus, we perform a nonlinear least-squares fit with a floating exponent, fitting to the model $V = \alpha A^{\beta}$ and obtain $\alpha = 0.383 \pm 0.030$

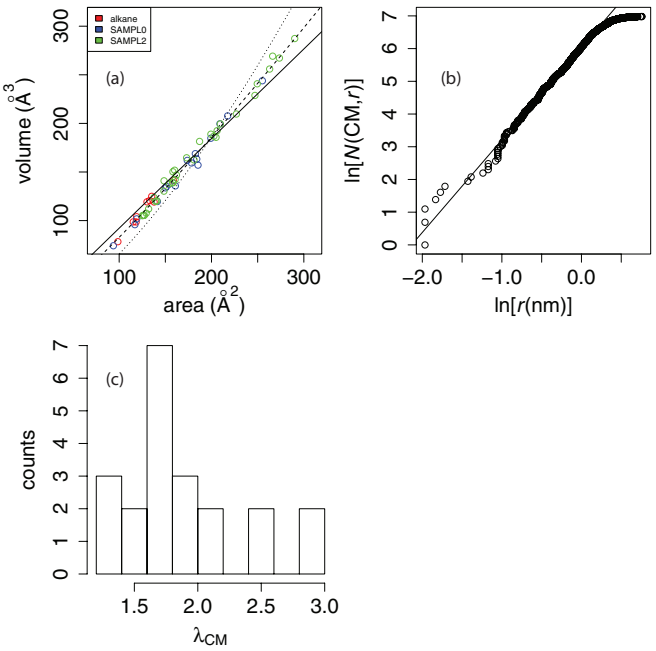

FIG. 3. Area/volume relationship for small molecule test sets. In panel (a), the solid line indicates a linear least-squares fit with a slope of 1.1 ± 0.014 Å, an intercept of −27.3 ± 2.44 Å^3 and a Pearson correlation coefficient of $R^2 = 0.99$. The dotted line indicates a nonlinear least-squares "spherical" fit ($V = \alpha A^{3/2}$), where $\alpha = 0.066 \pm 0.001$, and the dashed line indicates a "free exponent" fit $V = \alpha A^{\beta}$ where $\alpha = 0.38 \pm 0.03$ and $\beta = 1.17$ ± 0.02. The nonlinear least-squares fits were performed with the nls function in R (www.R-project.org). Panel (b) shows the natural log of the radial counting function about the center of mass vs. log (r) for the protein villin. The first 2/3 of the points, representing the "interior volume," can be fit with a slope of 2.83, which is the fractal density dimension λ_{CM}.[38] Panel (c) shows the distribution of fractal density dimensions λ_{CM} for small molecules in the pooled set for which the correlation coefficient of log atom count vs. log (r) is greater than 0.9. 38 of the 58 molecules in the unified set met this criterion.

and $\beta = 1.17 \pm 0.015$ with a resulting RMSD of 22 Å^3. Analysis of variance (ANOVA) shows that the floating-exponent model outperforms both the linear and spherical models with p-values of 10^{-15} or less (see Table S2 of the supplementary material[34]). Furthermore, the 95% confidence interval around β is [1.14,1.20], indicating that the exponent is statistically distinct from either 1 or 1.5.

In addition, we examine the area/volume relationship more broadly and its force field dependence using data from Gong and Yang[36] (see Figure S4 of the supplementary material[34]). Four other types of molecular area/volume calculations (molecular face, van der Waals area/volume, solvent-accessible surface area and volume, and solvent-excluded area/volume) show a scaling behavior of $V \propto A^{1.2}$, suggesting this shape behavior is general across different small molecules and different molecular area/volume calculation methods.

1. Geometric interpretation of volume-area correlation

In Figures 3(b) and 3(c), we use the concept of density dimension to further explore the origin of the volume-area scaling relationships. For a given molecule, the "fractal" density dimension $\lambda_{CM}(x)$ about a point x is the best-fit slope of $\log(N(x, r))$ vs. $\log(r)$ according to a linear least squares regression, where $N(x, r)$ is the "radial counting function"; i.e., the number of atoms within radius r of x.[37,38] If a molecule has an "interior volume," then its radial counting function should scale with approximately r^3 except for the rough surface region. We use the protein villin, which is large enough to have an interior volume, as a reference case. To exclude contributions from the non-flat surface, we perform a linear regression of $\log(N(x, r))$ vs. $\log(r)$ over the inner 2/3 of atoms to estimate λ_{CM} of 2.83 (Figure 3(b)), but that beyond $r \approx 1$ nm, the radial counting function, and thus V, scales with a smaller power of r. By comparison, we fit $\log(N(x, r))$ vs. $\log(r)$ over the inner 2/3 of atoms for the small molecules in this work, which reveal density dimensions averaging about 1.84 and ranging from 1.05 to 3.71 (Figure 3(c)).

For the interior volume (V) of an idealized spherical molecule, $V \propto r^3$ and area $A \propto r^2$, thus $V \propto A^{3/2}$; i.e., the "density dimension" is 3.[37] In practice, molecules such as proteins have dimensions closer to 2.9 for the interior volume due to imperfect packing.[37,38] However, the surface region of a large molecule does not behave as a three-dimensional object since rather than being flat-surfaced like a sphere, the surface region has many crevices and protrusions.[37] Similarly, in a small molecule, the surface is only a few atomic diameters from the center of mass and there may not be a proper "interior volume"; this likely explains why we estimate an average fractal density dimension of about 1.67 for small molecules. Thus, most small molecules behave more like the protein surface than the protein interior, with the exception of d-xylose ($\lambda_{CM} = 2.98$) and diethyl propanedioate ($\lambda_{CM} = 3.71$). Since glucose also has a relatively high $\lambda_{CM} = 2.58$, this may be a property of sugars due to their compact ring structure. By contrast, the molecules with the lowest λ_{CM} like pentachloronitrobenzene and the parabens are primarily aromatic and thus flat, so that the radial counting function will only scale in two dimensions with r. Given the large differences in V/A scaling between small and large molecules, our results suggest that for broad applicability, both γA and pV terms are important.

2. Thermodynamic interpretation of volume-area correlation

The observed volume-area correlation can also be interpreted based on thermodynamic arguments.[39] For simplicity, we will focus only on the nonpolar contribution, assuming that this is the energetic contribution primarily associated with the anti-correlation between p and γ, without loss of generality. Consider a small nonpolar solute inserted into a solvent where a differential Gibbs energy (dG) can be described by

$$dG = -SdT + VdP + \sum_i \mu_i dN_i + \gamma dA + \rho_0 \int_{\Omega_s} U_{\mathrm{vdW}}^{\mathrm{att}} d\mathbf{x}, \tag{10}$$

where T, S, and V denote the temperature, entropy, and volume of the solvent, respectively, and Ω_s is the region of space in the solvent outside of the solute. Here, μ_i and N_i are the chemical potential and number of moles for ith component of the solvent. Consider a cavity in a homogeneous solvent at constant temperature and assume that the solute-solvent van der Waals interactions give a negligible contribution to the overall energy. Under this approximation, Eq. (10) becomes $dG = \gamma dA + Vdp$ and a simple Maxwell relationship gives

$$\left(\frac{\partial \gamma}{\partial p}\right)_A = \left(\frac{\partial V}{\partial A}\right)_p. \tag{11}$$

The total amount of volume resulting from both solvent (V) and cavity (V_m) is $V_{\mathrm{total}} = V_m + V$ and the change due to cavity insertion is $dV = -dV_m$. Furthermore, the created surface area in the solvent due to insertion is $dA = dA_m$ if there is no significant deformation of the cavity upon the insertion. Given the assumptions above, the Maxwell relationship can be rewritten as

$$\left(\frac{\partial \gamma}{\partial p}\right)_{A_s} = -\left(\frac{\partial V_m}{\partial A_m}\right)_p, \tag{12}$$

which provides a simple relationship between the variation of surface tension with respect to pressure and that of solute volume with respect to solute surface area. To test the applicability of Eq. (12), we examined the relation between $(\frac{\partial \gamma_{\min}}{\partial p})_{A_s}$ and $(\frac{\partial V_m}{\partial A_m})_p$ over all three sets with van der Waals energetics set to zero to be consistent with the conditions for Eq. (12) (see Figure S5 of the supplementary material for details[34]). The two derivatives were linearly correlated with a slope of -1.00 ± 0.08, intercept of -0.22 ± 0.08 Å, and a Pearson correlation coefficient of $R^2 = 0.72$. This relationship implies that our data are qualitatively consistent with Eq. (12). The assumptions of constant area and pressure in the two respective derivatives are approximately justified since the calculated areas and volumes vary from the mean by less than 3% and 9%, respectively, for each set of molecules over the entire (p, γ) space examined in this study.

Furthermore, our results suggest that $(\frac{\partial\gamma}{\partial p})_{A_s}$ varies over a small range from −0.9 to −0.7 Å. This small variation in the rate of change is supported by past work which investigated the pressure dependence of the interfacial tension between two immiscible fluid phases with a planar interface[39] and concluded that the dependence comes from the coupling of the pressure to differences in the partial molar volumes of two fluids between the two phases. The study also showed that the dependence also varies slightly for the interface of several hydrocarbon molecules with experimental measurements of $(\frac{\partial\gamma}{\partial p})_A$ in the range of approximately −0.7 to 0.3 Å. This range is similar to the range we obtained in our analysis of *optimal computed* surface tensions and pressures using the geometric flow method.

IV. CONCLUSIONS

The geometric flow approach provides a physically realistic solvation model without considering explicit solvent and has previously compared well to experimental data in limited tests. In this work, calculated solvation energies for multiple sets of small molecules with the OPLS-AA force field and showed that the geometric flow model has good accuracy for most molecules. More importantly, we demonstrated that experimental values can be used the solvent internal pressure and surface tension model parameters, thus eliminating the need for additional *ad hoc* parameterization of the model. With a set of 58 molecules and a solute dielectric constant of $\varepsilon_m = 1.8$, we find that the experimental parameters for the air-water interface, pressure $p = 172$ MPa (0.0248 kcal mol^{-1} Å^{-3}) and surface tension $\gamma = 72$ mN/m (0.103 kcal mol^{-1} Å^{-2}), produce solvation free energies within 1 *RT* of the global minimum root mean square deviation over the set. Thus, it is possible that the previously reported need to use a different "microscopic" surface tension closer to 0.03 kcal mol^{-1} Å^{-2} for small molecules[12,13] may result not from the curvature of small molecules,[40] but rather from the neglect of pressure-volume work and of a correct definition of internal pressure.[10] Future work investigating geometric flow solvation predictions for a wider size range of small molecules is required for a detailed test of this hypothesis. The ability of geometric flow to make reasonable predictions of solvation free energy with experimentally derived parameters argues for the physical relevance of the model and its broad applicability. The reduction in the number of free parameters will also facilitate the extension of geometric flow to multi-conformational systems, proteins, and other more complicated cases where solvation is important to function. This adds to the existing benefit that the geometric flow formulation allows for simultaneous optimization of the polar and nonpolar components of solvation free energy.

In our previous work,[8] we found that the optimal values for γ and p are strongly anti-correlated for all molecules. Thomas *et al.* rationalized this anti-correlation based on the fact that γ increases with stronger water/water interactions, while water/water interactions become weaker as p increases.[8] While the p and γ terms of the solvation model are linearly correlated, our data on the interior volumes of proteins and small molecules suggest that this correlation only holds over a small range of molecular sizes, and that these terms are thus not redundant.

In summary, the geometric flow approach not only provides unambiguous coupled development of nonpolar and polar free energy functionals but also provides excellent results using experimental values for p and γ. This reduction in the number of free parameters will also facilitate the extension of geometric flow to blind predictions of solvation free energy and its use as a complement for interpreting related experiments. Future work should investigate the scalability of the geometric flow model to larger systems such as host-guest or protein-ligand binding energies where a broader range of solvation phenomena, including cavity de-wetting, influences the energetics of the system.

ACKNOWLEDGMENTS

We thank David Mobley for help compiling the SAMPL0 set parameters and for helpful discussion, and Julie Mitchell for providing guidance on how to interpret our observed volume/area relationships in small molecules and proteins. Funding for this work was provided by NIH Grant Nos. R01 GM069702 and R01 GM090208.

[1] H. Tjong and H.-X. Zhou, J. Chem. Theory Comput. **4**, 507 (2008).
[2] M. L. Connolly, Science **221**, 709 (1983).
[3] J. Dzubiella, J. M. Swanson, and J. A. McCammon, Phys. Rev. Lett. **96**, 087802 (2006).
[4] J. Dzubiella, J. M. Swanson, and J. A. McCammon, J. Chem. Phys. **124**, 084905 (2006).
[5] P. W. Bates, G. W. Wei, and S. Zhao, J. Comput. Chem. **29**, 380 (2008).
[6] Z. Chen, N. A. Baker, and G. W. Wei, J. Comput. Phys. **229**, 8231 (2010).
[7] Z. Chen, N. A. Baker, and G. W. Wei, J. Math. Biol. **63**, 1139 (2011).
[8] D. G. Thomas, J. Chun, Z. Chen, G. W. Wei, and N. A. Baker, J. Comput. Chem. **34**, 687 (2013).
[9] I. Leontyev and A. Stuchebrukhov, Phys. Chem. Chem. Phys. **13**, 2613 (2011).
[10] Y. Marcus, Chem. Rev. **113**, 6536 (2013).
[11] *C.R.C. Handbook of Chemistry and Physics*, 58th ed., edited by R. C. Weast (CRC Press, 1977).
[12] C. Chothia, Nature (London) **248**, 338 (1974).
[13] D. Eisenberg and A. D. McLachlan, Nature (London) **319**, 199 (1986).
[14] R. M. Levy, L. Y. Zhang, E. Gallicchio, and A. K. Felts, J. Am. Chem. Soc. **125**, 9523 (2003).
[15] J. A. Wagoner and N. A. Baker, Proc. Natl. Acad. Sci. U.S.A. **103**, 8331 (2006).
[16] F. Stillinger, J. Solution Chem. **2**, 141 (1973).
[17] R. A. Pierotti, Chem. Rev. **76**, 717 (1976).
[18] K. Lum, D. Chandler, and J. D. Weeks, J. Phys. Chem. B **103**, 4570 (1999).
[19] G. Hummer, S. Garde, A. E. Garcia, and L. R. Pratt, Chem. Phys. **258**, 349 (2000).
[20] S. Rajamani, T. M. Truskett, and S. Garde, Proc. Natl. Acad. Sci. U.S.A. **102**, 9475 (2005).
[21] J. Wagoner and N. A. Baker, J. Comput. Chem. **25**, 1623 (2004).
[22] M. S. Lee and M. A. Olson, J. Chem. Phys. **139**, 044119 (2013).
[23] Z. Chen and G. W. Wei, J. Chem. Phys. **135**, 194108 (2011).
[24] Z. Chen, S. Zhao, J. Chun, D. Thomas, N. A. Baker, P. Bates, and G. W. Wei, J. Chem. Phys. **137**, 084101 (2012).
[25] J. D. Weeks, D. Chandler, and H. C. Andersen, J. Chem. Phys. **54**, 5237 (1971).
[26] A. Nicholls, D. Mobley, P. Guthrie, J. Chodera, C. Bayly, M. Cooper, and V. Pande, J. Med. Chem. **51**, 769 (2008).
[27] E. Gallicchio, M. M. Kubo, and R. M. Levy, J. Phys. Chem. B **104**, 6271 (2000).

[28]P. V. Klimovich and D. L. Mobley, J. Comput.-Aided Mol. Des. **24**, 307 (2010).
[29]W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
[30]J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).
[31]J. Wang, W. Wang, P. A. Kollman, and D. A. Case, J. Mol. Graphics Modell. **25**, 247 (2006).
[32]A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, J. Comput. Chem. **21**, 132 (2000).
[33]A. Jakalian, D. B. Jack, and C. I. Bayly, J. Comput. Chem. **23**, 1623 (2002).
[34]See supplementary material at http://dx.doi.org/10.1063/1.4832900 for linear regression fitting parameters of optimal coefficients, volume-area curve-fitting relationships, and pressure-surface tension derivative plots.
[35]C. Tan, Y.-H. Tan, and R. Luo, J. Phys. Chem. B **111**, 12263 (2007).
[36]L.-D. Gong and Z.-Z. Yang, J. Comput. Chem. **31**, 2098 (2010).
[37]J. C. Mitchell, R. Kerr, and L. F. Ten Eyck, J. Mol. Graphics Modell. **19**, 325 (2001).
[38]L. A. Kuhn, M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, and J. A. Tainer, J. Mol. Biol. **228**, 13 (1992).
[39]L. A. Turkevich and J. A. Mann, Langmuir **6**, 445 (1990).
[40]K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig, Science **252**, 106 (1991).