JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 479–485

pubs.acs.org/JCTC

Calculation of Diffusion Coefficients through Coarse-Grained Simulations Using the Automated-Fragmentation-Parametrization Method and the Recovery of Wilke–Chang Statistical Correlation

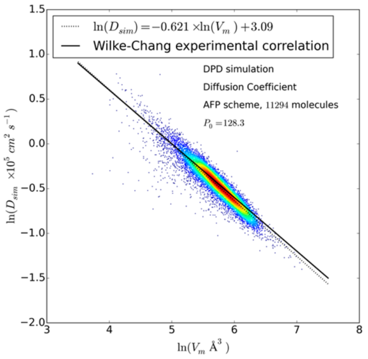
Johannes G. E. M. Fraaije,[*,†,‡] Jan van Male,[‡] Paul Becherer,[‡] and Rubèn Serral Gracià[‡]

[†]Leiden Institute of Chemistry, Leiden University, P. O. Box 9500, 2300 RA Leiden, The Netherlands

[‡]CULGI BV, Galileiweg 8, 2333 BD Leiden, The Netherlands

Supporting Information

ABSTRACT: We introduce a model for the calculation of diffusion coefficients using dissipative particle dynamics coarse-grained molecular simulations. We validate the model on experimental diffusion data of small organics and drug-like molecules in water. The new model relies on our automated-fragmentation-parametrization protocol for cutting molecules into fragments, which are calibrated using the COSMO-RS thermodynamic model (*J. Chem. Inf. Model.* **2016**, *56* (12), 2361–2377, DOI: 10.1021/acs.jcim.6b00003). By simulations over the entire CULGI database of more than 11000 molecules, we recover the decades-old empirical Wilke–Chang correlation between diffusion coefficient and molar volume. We believe this is the first demonstration of the correlation by simulation or theory. From a comparison of simulated and experimental diffusion coefficients, we find that one full time unit of coarse-grained simulation equals 64 ± 13 ps real time.



1. INTRODUCTION

Coarse-grained simulations could be a great aid in the design and analysis of complex soft materials, whether of synthetic or of biological origin. Much effort has gone into the modeling of force fields. The returning question has been, to what extent can coarse-grained interactions mimic the thermodynamics of real systems? In a recent work, we introduced the automated-fragmentation-parametrization protocol (AFP),[1] which aims to automate the cumbersome process of finding proper coarse-grained interaction parameters for a given set of arbitrary (organic) molecules. In short, AFP combines two steps: (i) the rule-based cutting of molecules into smaller pieces (referred to as fragments or beads) and (ii) the subsequent calibration of interactions through comparison with thermodynamic data. Both the rule-based cutting and calibration are nontrivial methods that both also require parametrization on a meta-level. The AFP contribution has an extensive appendix with all technical details, which we assume here as background material. Also, the AFP work has a more complete list of references and discussion of coarse-grained force fields that we refer to without repeating here.

So far, the attention of coarse-grained simulation on thermodynamics has left the diffusion coefficients untouched. The recent review on systematic coarse-graining from van der Vegt[2] does not discuss the time scale at all. Transport coefficients have been modeled in the closely related field of atomistically detailed simulations (molecular dynamics (MD)).[3−6] Especially the diffusion coefficient calculations of Wang and Hou[6] using molecular dynamics (AMBER/GAFF) are illustrative, with useful references to earlier molecular simulations. Wang and Hou point out that simulations of solute diffusion coefficients are rare because of required extensive sampling times (as opposed to slightly more common solvent self-diffusion coefficient simulations). For coarse-grained simulations, the number of studies is even smaller. In fact, for dissipative particle dynamics (DPD),[7−9] the simulation method of our choice here, there are no works that we know of, although there are a few studies on transport coefficients (viscosity) of simplified systems (Lennard-Jones atoms) on a theoretical level[10,11] through a bottom-up approach. The lack of comparison is somewhat surprising since DPD was meant to reproduce hydrodynamic interactions more accurately than competing methods. DPD is very efficient in generating complex structures, such as all kinds of self-assembly systems in so-called soft matter (polymers, micelles, and bilayers, etc.). We surmise that a next step should be the actual calculation of diffusion coefficients, with the great advantage being that the coefficients could be obtained for sluggish complex inhomogeneous materials, where very few if any simulation methods are available.

From the coarse-grained molecular dynamics community, there are no systematic contributions on this same topic of solutes diffusion, as far as we know. There are some works in the biomolecular area, for example discussing rate constants in protein ligand binding. Rate calculations are closely related to the current study, in the sense that one needs an accurate time scale to calibrate the simulations. One such example is from

Received: October 29, 2017
Published: December 22, 2017

DOI: 10.1021/acs.jctc.7b01093

Negami et al.[12] using the Martini force field in regular molecular dynamics. In the Martini model, one takes the approach to identify the time scale in coarse-grained molecular dynamics as "the speed up factor in the diffusional dynamics of CG water compared to real water"[13] (see also the Martini perspective work and references therein of ref 14). We find that such an approach, although practical, is potentially inadequate, since it does not readily justify extension to other systems. Apart from the comments in the original Martini work,[13] the Martini time-scale factor has been discussed but very briefly in the perspectives work[14] and a thesis.[15] A more theoretical study is clearly needed.

In the spirit of AFP, we calibrate the simulation method in top-down database fashion, over a set of wide chemical diversity, given essentially semiempirical laws based on phenomenological consideration. In other words, we do not attempt to model diffusion coefficients by some mapping of molecular dynamics trajectories[10,11] (that would anyway almost be impossible given the large data set we are exploring) but instead follow a more practical engineering approach for the thermodynamics interactions.

The actual diffusion simulation for a given molecule takes a few minutes using a single core on an Intel Xeon E5-2670 processor. While quantitative structure−property relations (QSPR) are still much faster, timings of a few minutes per molecule could bring coarse-grained simulations to practical chemical problems.

An important result is that by comparing experimental and simulated diffusion coefficients, we find a time scaling for the DPD coarse-grained simulations of about 64±13 ps real time per DPD unit of time. We refer to this scale as the dissipative time scale. The time scaling is dependent on the level of coarse-graining and the value of the solvent prefactor for the friction. Despite the fact that the mass enters into the kinetic time scale, the mass is irrelevant for calibration on diffusion.

A second result is that, by predictive simulations of all molecules in the CULGI database (more than 11000 molecules), we recover the decades-old empirical Wilke−Chang[16] correlation $D \propto V^{-0.6}$ between diffusion coefficient and molar volume V. The original Wilke−Chang publication is from 1955 and has been cited more than 3000 times since then (according to Web of Science). The correlation has found its way in many engineering transport models, but as far as we know, there never has been a satisfactory explanation from theory or simulation. We present here the first demonstration of the correlation by simulation, which is not only relevant for the transport modeling community but also suggests our methodology is correct.

While the proposed method is novel on a fundamental level, on a practical level one could ask what the benefit is of a simulation method (atomistic or coarse-grained) for diffusion coefficients when statistical regression methods such as Wilke−Chang seem to work fine. The answer is the same as we have given before in the AFP work: it is especially the extension to (inhomogeneous) systems where correlative data are difficult to obtain, or not publicly available, where coarse-grained simulations could be useful. The advantage of DPD with the AFP protocol is that the method is in principle, without adjustment, applicable to (much) more complex problems in biology or materials science. We cannot hope to address the much more complex inhomogeneous system if the method cannot reproduce behaviors of more simple homogeneous systems, hence the current study.

This work is organized as follows. In Theory, we briefly discuss the parametrization strategy. In Results, we discuss a comparison of experimental and theoretical diffusion coefficients in water and the predictive calculation of the diffusion coefficients of molecules extracted from the CULGI database. All simulation algorithms are presented in Methods.

2. THEORY

It is illustrative to assess briefly the pros and cons of the different particle simulation technologies,[17] to understand what type of simulation is best suited for the study of dynamics. Obviously, the qualifier "best" is a trade-off between practical results and fundamental rigor, and to a certain extent arbitrary. One notices that in molecular dynamics, whether atomistic or coarse-grained, there is no friction and noise term. In principle, the reference velocity is determined by the masses and temperature through the equipartition theorem. In this case, one must rely on the quality of the force field that must both get diffusion and thermodynamics correct, which is not trivial to say the least. In Brownian dynamics, one has an additional parameter: the friction, with dissipation determined by the temperature through the fluctuation−dissipation theorem. The additional parameter is helpful, since then one has an additional handle to set the scale, with the understanding that the additional parameter makes the simulations less "ab initio", and one does need experimental calibration. The disadvantage of Brownian dynamics is that it is not Galilean invariant and, therefore, does not include hydrodynamic interactions (unless added ad hoc through an additional friction tensor model). In DPD, the friction is determined by hydrodynamic flow generated by the simulation itself (as opposed to Brownian dynamics), and noise is again by fluctuation dissipation theorem (same as in Brownian dynamics). Thus, within the framework of DPD one has the unique (or best) possibility of both including hydrodynamic interactions *automatically* and having an additional parameter space, the friction coefficients, to calibrate the time scale.

The documentation on the DPD simulation technique is extensive. An excellent recent review is available that we refer to for all details.[9] We briefly recall the essentials in the Supporting Information.

DPD coarse-grained simulations have three sets of parameters: the masses of the beads, the parameters in the conservative forces, and the friction coefficients in the dissipative forces. We focus in the simulations exclusively on results in the dissipative regime. We set all masses of all beads to the same value (arbitrarily chosen as unity). The approach is customary in DPD, but potentially confusing for the molecular modeling community. In our case, we will demonstrate that the mass is, in fact, an irrelevant parameter, at best to be interpreted as a numerical setting. The simplication on mass leaves two sets of parameters: one set for the calculation of the conservative force and a second set for the calculation of the friction coefficients. The two sets are to a considerable extent (but not completely) orthogonal. Each set separately can be matched to experiments, thermodynamics, and transport coefficients, respectively. In this work, we make the further simplification to keep all friction coefficients the same.

3. METHODS

Simulation. We used the CULGI software for all calculations, installed on a Dell Precision T7610 PC equipped with an Intel Xeon E5-2670 dual processor. The DPD time step was set to 0.01, the friction coefficient (see Supporting Information) to $\gamma = 5$, and the number of steps in the diffusion calculation to 10^5, with box size $L_x = L_y = L_z = 5$ (DPD units; cell size, $r_c = 7.65$ Å).

We calculated the diffusion coefficient using the standard mean-squared displacement (MSD) method through the Einstein relation $\langle r^2 \rangle = 6D\tau$.[18] Details are in the Supporting Information.

Since the usual interest in diffusion coefficient predicition is an estimate for the expected *relative* error, we use a scoring function that minimizes the error in $D_{\text{exp}}/D_{\text{sim}}$. The scoring function is conveniently expressed in naural logarithms as

$$\text{RMSD}(S) = \sqrt{\overline{\ln^2\left(\frac{D_{\text{exp}}}{D_{\text{sim}}}\right)}}$$
$$D_{\text{sim}} = SD_{\text{DPD}} \tag{1}$$

where RMSD is the root-mean-squared deviation, D_{DPD} the unscaled diffusion coefficient from the simulations, S the scale factor, and D_{exp} the experimental value. The overbar indicates the average over all molecules. Minimization of the given RMSD function is trivially equivalent to minimizing the difference in logarithms. The relative error $\Delta S/S$ in the *estimate* of the optimal scaling is calculated from the RMSD through

$$\frac{\Delta S}{S} = \mathrm{e}^{\text{RMSD}} - 1 \tag{2}$$

The minimization with respect to the scale can be carried out analytically and leads to the optimal scale value:

$$\ln S = \overline{\ln\left(\frac{D_{\text{exp}}}{D_{\text{DPD}}}\right)} \tag{3}$$

For a given molecule, a complete diffusion calculation consists of three or four steps. First, we do a quantum calculation of the COSMO charge envelope. Most molecules in the diffusion data set are quite common and already included in the CULGI database (see below). For 28 molecules (mostly steroids and nucleobases), the COSMO information was missing, and we did a full quantum calculation first, using the same settings as in the CULGI database quantum calculations. The second step is then the decomposition in fragments (a few seconds calculation time at most), followed by the thermodynamics calibration through COSMO-RS (less than a minute) and the actual diffusion calculation (a few minutes).

The molar volume was calculated from finite difference calculation of the volume between two systems differing in one molecule only while keeping the pressure constant $V_m \equiv (\partial V/\partial n)_{P,T} \simeq V(n+1,p,T) - V(n,p,T)$.

AFP Method. The AFP method is extensively documented.[1] Molecules are fragmented according to a scoring function, through a simulated annealing function that cuts through bonds. The optimal bond scission pattern is preserved, and the fragments are stored. The scoring function is

$$\text{fragment score} = \left(1 - \frac{V}{V_0}\right)^2 \tag{4}$$

with V being the volume of the fragment and $V_0 = 67.7$ Å^3 the reference volume. The reference volume is the volume of a cluster of three water molecules. As we have discussed before, we use the molecule-unique fragmentation in order to preserve as much as possible the properties of the mother molecule. This means that the fragments are not database-unique, as is customary in coarse-grained simulations, but specific to a given molecule. The thermodynamic calibration is through a semiempirical rule for the DPD a parameter, which we use here without any modification:

$$a_{ij} = \alpha_{\text{EV}}\nu_i\nu_j + \alpha_{\text{res}}\sqrt{\nu_i\nu_j}\,\beta\Delta G_{\text{res},ij}$$
$$\nu_i \equiv \frac{V_i}{V_0} \tag{5}$$

with $\beta = 1/k_BT$ and $\Delta G_{\text{res},ij}$ the excess Gibbs energy of mixing of two fragments i and j. The parameters α_{EV} and α_{res} are global parameters, determined by optimization on thermodynamics. We use $\alpha_{\text{EV}} = 50.0$ commensurate with a background pressure $P_0 = 128.3$ and dimensionless density of water $\rho = 5$. For the residual interaction, we used the same value as in the AFP work, $\alpha_{\text{res}} = 6.1$. As before, we calculate the Gibbs energy of mixing through COSMO-RS calculations,[19,20] using the σ profile of the fragments.

Database. We use diffusion coefficients from the compilation by Hills et al.,[21] and data from Song et al.[22] and Seki et al.[23] All reported values are at 25 °C. The list of molecules including experimental and simulated diffusion coefficients is copied in the Supporting Information. By comparison to the Hills data set we have excluded molecules with COSMO volume smaller than 40 Å^3, since coarse-graining does not make much sense for very small molecules. The data set mostly deals with quite common chemicals (industrial solvents, for example) but also includes a few tens of steroids (from the Seki data set) and a few nucleobases (from Song). The σ profiles were calculated from accurate density functional quantum-chemical calculations (NWChem[24]), from geometry-optimized molecules in a vacuum, using the def2-TZVPD basis set and BP86 exchange functional. In all of the diffusion coefficient calculations, by far most of the time is taken by the expensive σ-profile calculation, which can easily go into a few days of CPU time for even modestly sized drug-like molecules, such as the steroids in the data set. In contrast, as we have mentioned above, the coarse-grained simulation itself takes only a few minutes.

4. RESULTS

Diffusion Coefficients. The calculated diffusion coefficients are reported in Figure 1. To convert the dimensionless DPD values to dimensioned values (in units 10^{-5} cm^2 s^{-1}), we used the optimal scale value found by fitting $S \pm \Delta S = 9.2 \pm 1.9$. A few things are noteworthy.

First, the number of data points $n = 164$ is rather small, while the recent Hills data set itself is already a collection of older sources. Although there could be quite some scattered isolated data that we did not find, and therefore did not include, as far as we know we did not miss a large collection somewhere. Compared to thermodynamic data, the tininess of the diffusion data set is apparent. For example, to calibrate log P predictions, the cornerstone of many a QSPR model, one could have access to tens of thousands of data points, if not more. The relative scarcity of diffusion data has an important implication: QSPR models for diffusion with even a few descriptors are difficult to assess, or simply not trustworthy at all, because of the small size of the training set. Our approach through coarse-grained

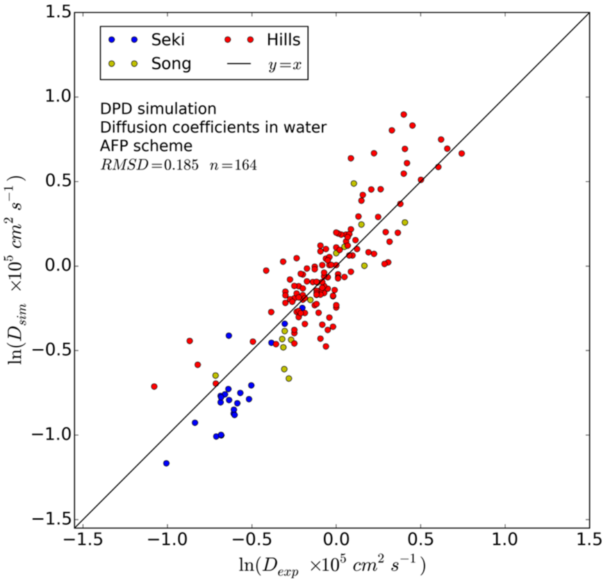


Figure 1. Calculated diffusion coefficients versus experimental values. Sources of data indicated: Seki,[23] Song,[22] and Hills.[21]

simulation, which needs only one additional time-unit parameter (the scale factor) over the original thermodynamic calibration, could be quite helpful.

A second observation is that the spread in values is not large. The range from smallest to largest value on a natural logarithmic scale runs from −1 to +1. The small range corresponds to a factor of 7.3; i.e., the relative difference between the largest and smallest diffusion coefficient is less than an order of magnitude.

A third noteworthy observation is that the difference between simulation and experiment is also minor. We find RMSD = 0.19, which corresponds to a relative error in diffusion coefficient prediction of 21%. While this is amazingly accurate compared to the typical accuracy in thermodynamics property prediction (where one could already be happy with an accuracy of one 10 log unit), such a high accuracy is also typical for quantitative structure−activity correlations (see the summary of models in the Hills et al. work). The reason for the accurate prediction is relatively simple: by far the dominant factor that determines the diffusion coefficient is the volume of the molecule since the dissipation is controlled by the hydrodynamic interactions all around the molecule. Thermodynamics has an important, but a secondary, influence. In the molecular dynamics study of Wang and Hou,[6] one finds a relative error of 12.6%, which is smaller than what we find, but then the Wang study involved only five solutes (acetic acid, expt 1.290/calc 0.963; acetonitrile, 1.260/1.333; cyclohexane, 0.840/0.903; diethylamine, 0.970/0.913; phenol, 0.89/1.054). Of course, in the case of molecular dynamics, the prediction is absolute, whereas we need a scale factor.

In our calculation of the relative error in the estimate of the optimal scaling, $\Delta S/S$, we assume implicitly that all errors are due to the model, and not due to experiments. The logic is that if the experiments are exact, and one attributes all errors to incorrect theory, then one would have an optimal simulation scaling per molecular system $S_1 ... S_n$, as opposed to one scale for all molecules, and hence in such case the difference between average and molecular-specific scale is an indication of the overall spread in scale values. Unfortunately, without additional sources, it is impossible to judge whether and to what extent the reported experimental values are correct. Overall the difference in behavior between the different experimental data sets is minor. It does seem as if the performance of the model over the Seki data is less good than those of Hills and Song. We double-checked a few noticeable outliers, such as anthracene and benzanthracene (see the Supporting Information)—these were also reported as outliers by Hills et al. We notice that the experiments are not trivial, especially in the case of badly soluble polycyclic aromatics. The source of the Hills data for the polycyclics is Gustafson and Dickhut,[25] but while the present analysis deals with infinite dilution, the experiments are (for the anthracenes) at 50% saturation. One could easily imagine that a small amount of aggregation leads to a lower apparent diffusion coefficient. The data sets unfortunately also contained typographical errors. Hills et al. report a value for diox*in* in their table, but this should have been diox*ane*, a very different molecule.

The optimal scale immediately leads to an assessment of the time scale t/τ in the DPD simulations. From matching the diffusion between real time and dimensionless time, we find the time scale

$$\frac{t}{\tau} = \frac{\langle r^2 \rangle}{\langle {r_{\mathrm{DPD}}}^2 \rangle} \frac{D_{\mathrm{DPD}}}{D_{\mathrm{exp}}} = \frac{\langle r^2 \rangle}{\langle {r_{\mathrm{DPD}}}^2 \rangle} \frac{1}{S} \tag{6}$$

and with the length scale r/r_{DPD} = 7.65 Å; this leads to

$$\frac{t}{\tau} = \frac{7.65^2}{(9.2 \pm 1.9) \times 10^{11}}\ \mathrm{s} = 64 \pm 13\ \mathrm{ps} \tag{7}$$

In other words, to get the time scaling of diffusion correct, we need that one full time unit of coarse-grained simulation equals 64 ps real time. A typical DPD simulation time step of 0.01 corresponds to 640 fs. This can be compared to molecular dynamics in two ways.

First, a typical atomistically detailed molecular dynamics time step is 1 fs.[17] Hence, DPD is roughly 600 times as *time* efficient. We notice that the *computational* efficiency is yet much higher, since per time step one needs to evaluate fewer forces in DPD than in MD. The simulation is in total over 10^5 steps. With a time step of 0.01 this corresponds to 1000 DPD time units. Using the optimized time scaling of 64 ps per time unit, the total simulation time is therefore 64 ns. Notice this is achieved in just a few minutes calculation time on a single CPU core.

Second, in molecular dynamics the time scale is locked by the mass and temperature through the equipartition theorem ($m{v_{\mathrm{T}}}^2 = 3kT$ where v_{T} is the thermal velocity; see the table of units in the Supporting Information). But here, we did we did not use the mass at all. Instead, we found the scaling by matching the dissipation to experiment. It is, in fact, the other way around, in that we can estimate a *fictitious* or numerical mass of the beads by imposing that the kinematic time, denoted as τ_{K}, corresponds to the dissipative time, τ_{D}, calculated by the fitting of diffusion (at 300 K):

$$\tau_{\mathrm{K}} \equiv \frac{r_{\mathrm{c}}}{v_{\mathrm{T}}} = \tau_{\mathrm{D}} = 64\ \mathrm{ps} \quad \rightarrow \quad m = 3kT\left(\frac{\tau_{\mathrm{D}}}{r_{\mathrm{c}}}\right)^2 \triangleq 52 \times 10^3\ \mathrm{Da} \tag{8}$$

One expects a mass of $O(50)$ dalton for the physical mass of a bead consisting of three united atoms. Remarkably, to match the kinetic and dissipative time scale, we need the fictitious mass to be $O(10^3)$ larger than that. We note therefore that in these DPD simulations mass, perhaps contrary to expectations,

DOI: 10.1021/acs.jctc.7b01093

is not determined by the physical system. Rather, mass should be considered as a numerical parameter. It is for this reason that we need not worry about fragment-specific mass values, and one nonspecific value for the entire system suffices. The significant difference between fictitious and physical mass makes one wonder why one would need such a parameter at all. In fact, it could be possible to reduce the DPD simulation system further, by removing the inertial forces altogether from the set of equations and render the model in the overdamped limit. There could be a potential for a further speed-up. But here, we merely point to such a reduction and leave further discussion to following publications.

A second refinement could be to include fragment-specific friction coefficient. One could imagine that a polar fragment which is hydrogen-bonded to water molecules, experiences a larger friction than an inert apolar fragment, simply because such a hydrogen-bonded fragment would be more closely coupled to a sluggish fluctuation network of interactions. Given the discussion on the experimental accuracy above, one could equally wonder if the remaining deviation from experiment is just due to experimental noise, in which case any additional model would, of course, be overkill. In Hills et al.,[21] there is an ample discussion of adapting the Abraham descriptors to correlating with diffusion. The discussion points to the relevance of volume, as expected, and the Abraham hydrogen-bonding parameters (basicity and acidity). While there is some correlation with the Abraham descriptors, in our case the thermodynamics (including hydrogen-bonding propensity) is already included in the model for the DPD α parameters, and apparently, such an effect is already enough to capture almost all of the relevant molecular properties.

Wilke–Chang Scaling. Next, we turn to the prediction of diffusion coefficients over the entire CULGI data set. The prediction includes more than 11000 molecules. Except for those molecules that are also present in the experimental data sets discussed above, the prediction is purely speculative. The properties of the CULGI database are extensively discussed in the AFP work. While on average the molecules are smaller than that of a typical drug bank, we believe the data set is a reasonable representation of an arbitrary molecular distribution. The diffusion coefficients were calculated, using the established optimal time scale as discussed above. No further fitting or adjustment of parameters was used. The results are in Figure 2.

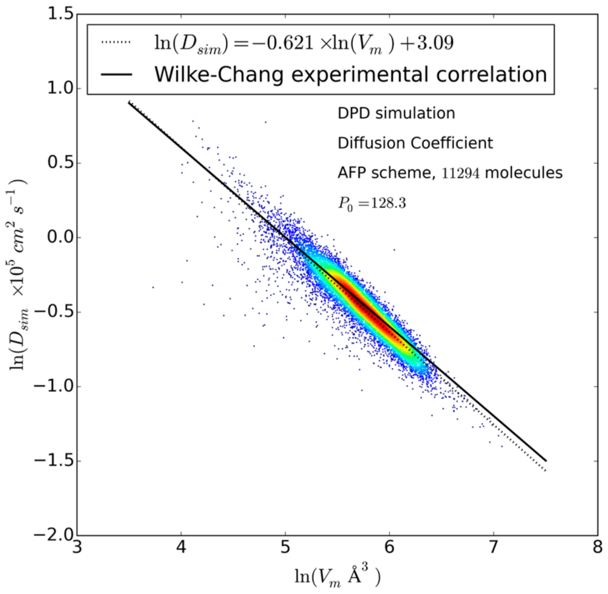


Figure 2. Prediction of diffusion coefficients for all molecules in CULGI database. Regression (dotted) and Wilke–Chang correlation indicated. Since the plot has very many points, we overlaid the scatter diagram with a color scheme akin to a heat map, to indicate the local density (red, high overlap; blue, isolated points).

We have also included the scaling according to the Wilke–Chang empirical law $D = 20.15V^{-0.6}$. The law was trivially adjusted from the original to the current dimension system, with diffusion measured in 10^{-5} cm^2 s^{-1} and molar volume in Å^3 per molecule. Compared to the original Wilke–Chang correlation we made one minor modification. The original Wilke–Chang law uses the molar volume at *boiling point*, whereas we use the molar volume at room temperature as obtained from the AFP DPD simulations. Apart from that, no further modifications were made. For details see the Supporting Information. We did not make any adjustments in the coarse-grained simulations, given the optimal time-scaling value obtained from the calibration on the experimental data. We find that the agreement between prediction and Wilke–Chang correlation is almost perfect. The Wilke–Chang correlation was discovered more than 60 years ago (published in 1955),[16] by analyzing diffusion and viscosity data of a range of organic liquids, and has been the cornerstone of engineering transport models since then. Since that date, more than 3000 papers have cited the Wilke–Chang work, but, as far as we know, there has never been a theoretical explanation nor a demonstration by simulation of the scaling behavior.

The details of the simulations offer a clue to the background of the Wilke–Chang correlation. If one picks a random distribution of molecules, some will be more elongated, some more spherical, some others more flexible, yet others more rigid and compact, each such a shape with its volume-scaling law. The net scaling over all molecules is then an average over all those differently shaped objects. In other words, if our CULGI database were to consist exclusively of spherical molecules, the scaling would have resembled the Stokes law with $D \propto V^{-1/3}$; if by chance all the molecules would be more elongated, the scaling would be more like Rouse (with independent and additive contributions from all fragments), $D \propto V^{-1}$. From the diffusion theory of oblate objects,[26] one can calculate that the scaling law for thin disk-like molecules would be $D \propto V^{-0.45}$. In our case, the database is a random collection of *real* molecules, with arbitrary shape and volume, and on top of that, none of the coarse-grained molecules is rigid. All molecules are flexible (to a certain degree set by the bonds) and can fluctuate in the solvent, depending on the thermodynamic interactions. It is possible to find the scaling by hydrodynamic simulation on a diverse data set, as we have shown here. But there can, in principle, never be a *simple* fundamental theory for Wilke–Chang, since the scaling is all determined by the *distribution* in the sample, as opposed to scaling derived from a geometrical or homologous series of chemicals. Apparently the CULGI database is sufficiently universal to warrant a realistic distribution of arbitrary chemicals.

Our conclusion here is that the fact that we find Wilke–Chang back in a highly abstract coarse-grained DPD simulation model demonstrates, foremost, that DPD simulations are capable of capturing hydrodynamic flow patterns around complex molecular structures, of whatever shape and flexibility. There have been many attempts, since the original Wilke–Chang findings, of improving the correlation by adding other

DOI: 10.1021/acs.jctc.7b01093

descriptors to the volume, from the shape or some interaction model (see the textbook[26] and Hills et al.[21]), but no such model has been derived from hydrodynamics.

Figure 2 also shows a few tens of outliers. While they represent only a minor fraction of the total data set of more than 11000 molecules, they could be highly relevant for further improving the AFP scheme. We have found that some modified triazoles and diazoles are exceptionally difficult to model with the current version of AFP. One outlier is hydroflumethiazide (Figure 3), with ln $D_{\text{regression}}/D_{\text{sim}}$ = 1.4 (this is the biggest outlier in Figure 2). Or, in other words, the simulated value is only 25% of what the regression would predict. The explanation is in the strong interactions between some very polar groups and water, which the DPD soft repulsion model cannot handle well. These outliers are very helpful in pointing to an improvement of the AFP scheme, for example by incorporating hydration layers.

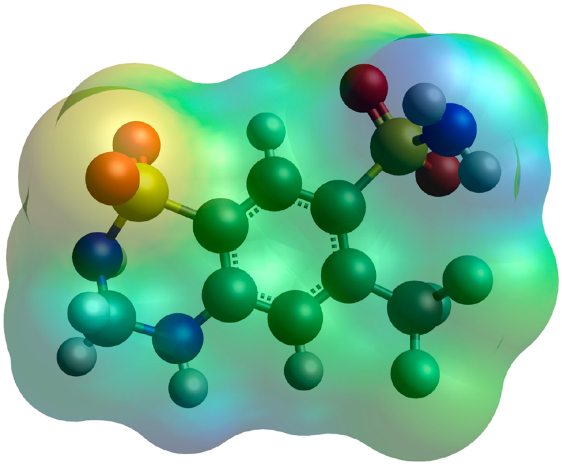


Figure 3. Hydroflumethiazide. The COSMO charge distribution is indicated.

On the other extreme are molecules that do very well. The molecule with the lowest diffusion coefficient is tetracontane (Figure 4), which is almost on top with the Wilke−Chang curve. The molecule is obviously apolar and flexible, and one expects that in dilute aqueous solution the molecule is folded into a compact but fluctuating object. We have included a snapshot from the coarse-grained molecule in solution, taken from the diffusion trajectory simulation that clearly shows the expected compaction. Correspondingly, the diffusion of tetracontane is not that of Rouse (small flexible polymer in good solvent, $D \propto V^{-1}$), neither that of a rigid sphere $D \propto V^{-1/3}$ but something in between.

In the simulations, we keep the fragmentation pattern constant. While this is an implicit assumption in the AFP method, such approximation seems an absolute necessity to make coarse-grained simulations at all possible. For the calculation of the way each molecule is cut into pieces, we use only one conformation—the conformation as it is in the CULGI database (which is in the end obtained through PubChem). The fragmentation pattern relies on the molecular COSMO volume and is therefore *in principle* dependent on the original conformation, but we have found that the dependency of the fragmentation pattern on conformation is only very mild. Even though the fragmentation pattern is a constant, the coarse-grained simulations themselves include all conformations. As we have demonstrated, when the coarse-grained molecule is flexible (as a hydrocarbon) it folds completely on itself into a compact fluctuating object. The more rigid molecules (as drug-like molecules, for example) do not fluctuate extensively, not even on coarse-grained level, as expected. But the method makes no presupposition about any of that given the fragmentation; all conformations are sampled as in agreement with statistical thermodynamics. The calculated diffusion coefficient is always an average over conformations, just as the experimental diffusion coefficient is from an average.

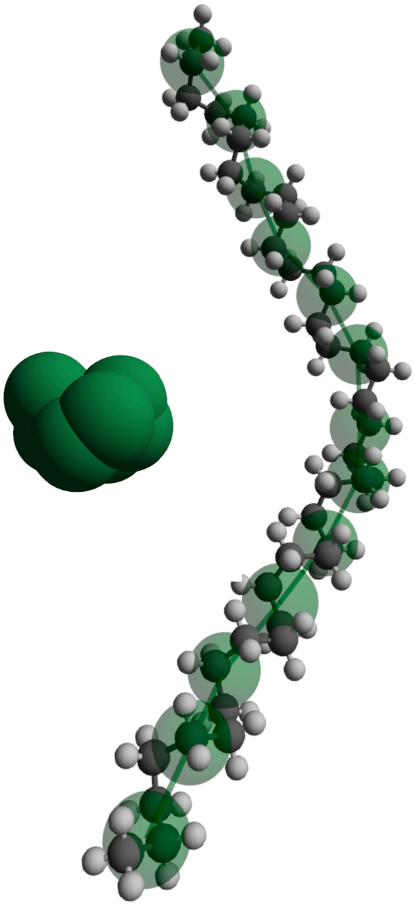


Figure 4. Tetracontane. Right: molecular structure in CULGI Database, overlaid with bead structure. Left: a collapsed state in solution. The molecule is coarse-grained to 13 fragments.

5. CONCLUSION

We have shown that coarse-grained DPD simulations based on using the automated-fragmentation-parametrization scheme can capture the diffusion of a variety of organic molecules, with high accuracy. A major result is the recovery of the classical Wilke−Chang statistical correlation between diffusion coefficient and molar volume, on a database of more than 11000 molecules. Since the simulations do not assume a solute or solution structure, we suggest one could extend the protocol to more complex systems from biology and materials science. From a comparison of simulated and experimental diffusion coefficients, we find that one full time unit of coarse-grained simulation equals 64 ± 13 ps real time. The coarse-graining and simulation together take only a few minutes of calculation time on a single CPU core. In all of the diffusion coefficient calculation, by far most of the time is taken by the expensive σ-

profile calculation. We point to several possible refinements of the proposed simulations: (i) by reduction of the simulations to the overdamped limit, (ii) by introduction of fragment-specific friction parameters, (iii) by simplification of the σ-profile calculations, and (iv) by adjusting the AFP scheme to very polar molecular groups by the inclusion of hydration layers.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b01093.

> Descriptions of DPD, of the diffusion sampling algorithm, and of Wilke−Chang scaling and table with experimental and calculated diffusion coefficients (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: j.fraaije@chem.leidenuniv.nl.

ORCID

Johannes G. E. M. Fraaije: 0000-0002-3856-3621

Funding

We thank the members of the CULGI industrial consortium for continued support.

Notes

The authors declare the following competing financial interest(s): The first author is founder and owner of CULGI BV.

■ REFERENCES

(1) Fraaije, J. G. E. M.; Van Male, J.; Becherer, P.; Serral Gracià, R. Coarse-Grained Models for Automated Fragmentation and Parametrization of Molecular Databases. *J. Chem. Inf. Model.* **2016**, *56* (12), 2361−2377.

(2) Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodriguez-Ropero, F.; van der Vegt, N. F. A. Systematic coarse-graining methods for soft matter simulations - a review. *Soft Matter* **2013**, *9* (7), 2108−2119.

(3) Hess, B. Determining the shear viscosity of model liquids from Molecular Dynamics simulations. *J. Chem. Phys.* **2002**, *116* (1), 209−217.

(4) Wensink, E. J. W.; Hoffmann, A. C.; van Maaren, P. J.; van der Spoel, D. Dynamic properties of water/alcohol mixtures studied by computer simulation. *J. Chem. Phys.* **2003**, *119* (14), 7308−7317.

(5) Guevara-Carrion, G.; Vrabec, J.; Hasse, H. Prediction of self-diffusion coefficient and shear viscosity of water and its binary mixtures with methanol and ethanol by molecular simulation. *J. Chem. Phys.* **2011**, *134* (7), 074508.

(6) Wang, J. M.; Hou, T. J. Application of Molecular Dynamics Simulations in Molecular Property Prediction II: Diffusion Coefficient. *J. Comput. Chem.* **2011**, *32* (16), 3505−3519.

(7) Groot, R. D.; Warren, P. B. Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **1997**, *107* (11), 4423−4435.

(8) Hoogerbrugge, P. J.; Koelman, J. Simulating Microscopic Hydrodynamic Phenomena With Dissipative Particle Dynamics. *Europhys. Lett.* **1992**, *19* (3), 155−160.

(9) Espanol, P.; Warren, P. B. Perspective: Dissipative particle dynamics. *J. Chem. Phys.* **2017**, *146* (15), 150901.

(10) Fu, C.-C.; Kulkarni, P. M.; Shell, M. S.; Leal, L. G. A test of systematic coarse-graining of molecular dynamics simulations: Transport properties. *J. Chem. Phys.* **2013**, *139* (9), 094107.

(11) Lei, H.; Caswell, B.; Karniadakis, G. E. Direct construction of mesoscopic models from microscopic simulations. *Phys. Rev. E* **2010**, *81* (2), 026704.

(12) Negami, T.; Shimizu, K.; Terada, T. Coarse-Grained Molecular Dynamics Simulations of Protein-Ligand Binding. *J. Comput. Chem.* **2014**, *35* (25), 1835−1845.

(13) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812−7824.

(14) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801−6822.

(15) de Jong, D. H. A molecular view on the organizational complexity of proteins in membranes. Ph.D. Thesis; University of Groningen: Groningen, The Netherlands, 2013.

(16) Wilke, C. R.; Chang, P. Correlation Of Diffusion Coefficients In Dilute Solutions. *AIChE J.* **1955**, *1* (2), 264−270.

(17) Berendsen, H. J. C. *Simulating the physical world: Hierarchical modeling from quantum mechanics to fluid dynamics*; Cambridge University Press: Cambridge, U.K., 2007.

(18) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press: San Diego, CA, USA, 1996.

(19) Diedenhofen, M.; Klamt, A. COSMO-RS as a tool for property prediction of IL mixtures—A review. *Fluid Phase Equilib.* **2010**, *294* (1−2), 31−38.

(20) Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Amsterdam, 2005.

(21) Hills, E. E.; Abraham, M. H.; Hersey, A.; Bevan, C. D. Diffusion coefficients in ethanol and in water at 298 K: Linear free energy relationships. *Fluid Phase Equilib.* **2011**, *303* (1), 45−55.

(22) Song, H. Y.; Vanderheyden, Y.; Adams, E.; Desmet, G.; Cabooter, D. Extensive database of liquid phase diffusion coefficients of some frequently used test molecules in reversed-phase liquid chromatography and hydrophilic interaction liquid chromatography. *Journal of Chromatography A* **2016**, *1455*, 102−112.

(23) Seki, T.; Mochida, J.; Okamoto, M.; Hosoya, O.; Juni, K.; Morimoto, K. Measurement of diffusion coefficients of parabens and steroids in water and 1-octanol. *Chem. Pharm. Bull.* **2003**, *51* (6), 734−736.

(24) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181* (9), 1477−1489.

(25) Gustafson, K. E.; Dickhut, R. M. Molecular Diffusivity Of Polycyclic Aromatic-Hydrocarbons In Aqueous-Solution. *J. Chem. Eng. Data* **1994**, *39* (2), 281−285.

(26) Cussler, E. L. *Diffusion, mass transfer in fluid systems*, 3rd ed.; Cambridge University Press: Cambridge, U.K., 2009.