PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

rsc.li/pccp

View Article Online
DOI: 10.1039/C9CP02358E

ARTICLE

A molecular contact theory for simulating polarization: application to dielectric constant prediction

Théophile Gaudin,*[a] and Haibo Ma*[a]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Microscopic polarization in liquids, challenging to account for intuitively and quantitatively, can impact the behavior of liquids in numerous ways and thus is ubiquitous in a broad range of domains and applications. To contribute overcoming this challenge, in this work, a molecular contact theory is proposed as a proxy to simulate microscopic polarization in liquids. In particular, molecular surfaces from implicit solvation models are used to predict both dipole moment of individual molecules and mutual orientations arising from contacts between molecules. Then, the calculated dipole moments and orientations are combined in an analytical coupling which allows predicting the effective (polarized) dipole moments of all distinct species in the liquid. As a proof-of-concept, the model focuses in predicting the dielectric constant. Tested on 420 pure liquids, 269 binary organic mixtures (3792 individual compositions) and 46 aqueous mixtures (704 individual compositions), the model proves flexible enough to reach an unprecedented satisfactory mean relative error of about 16-22% and a classification accuracy of 84-90% within four meaningful classes of weak, low average, high average and strong dielectric constants. The method also proves computationally very efficient, with calculation times ranging from a few seconds to about ten minutes on a personal computer with a single CPU. This success demonstrates that much of microscopic polarization can be satisfactorily described based on a simple molecular contact theory. Moreover, the new model for dielectric constant provides a useful alternative to computationally expensive molecular dynamics simulations for large scale virtual screenings in chemical engineering and material sciences.

Introduction

Polarization, in particular within liquids, is a determining phenomenon for a number of their macroscopic properties[1] which is still challenging to rationalize and efficiently quantify based on molecular structures alone. In particular, the polarization within the liquids has a high impact on the conductivity of liquids (in turn tightly related to the dielectric constant, or relative permittivity ε_r, of the medium). Knowledge of conductivity for liquids is central in economic, industrial and environmentally relevant applications of liquids such as batteries, super-capacitors, fuel cells or solar cells where they can play the role of electrolytes[2, 3]. As efficient rationalization and prediction of the microscopic polarization of liquids would help speed up the improvement of such devices, allowing easier access to this property alone would suffice to make it an important bottleneck to overcome. Still, even beyond electric conductivity, microscopic polarization is also involved in transport processes, thus impacting the viscosity of liquids, strongly determines phase equilibrium phenomena (to the point that researchers developed a number of dedicated equations of state for strongly polar liquids[1]), or influences outcomes of chemical reactions[4]. Therefore, understanding and predicting microscopic polarization is of broad interest across chemical disciplines.

Molecular dipoles, at the basis of microscopic polarization, depend on the electron distribution within individual molecules: molecules in which electron-rich and electron-poor zones are well separated are dipolar (or have high dipole moments). Therefore, molecular dipoles are determined by electron density and molecular geometry and can be modeled using quantum chemical methods of any cost levels depending on the appropriate tradeoff between efficiency and accuracy. However, to determine microscopic polarization from molecular dipoles, intermolecular interactions have to be considered. Thus, polarization is usually simulated using molecular dynamics[5, 6] (MD). However, MD suffers from two difficulties in simulating polarization. First, MD requires to simulate the interactions between a large number of particles, and therefore is computationally expensive. Moreover, the force fields may not be transferable from one system to another, so MD faces limitations if one seeks to screen an important set of simple systems according to an application-relevant property. To the end, the direct simulation of polarized systems may not provide a vivid and intuitive picture of the phenomenon of microscopic polarization. So, a simpler concept of polarization can both clarify its microscopic origins and allow access to a cheaper prediction of all properties related to polarization. Moreover, by making simple, homogeneous systems describable by said intuitive concepts, MD simulations,

[a] Key Laboratory of Mesoscopic Chemistry of MOE, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023 China. E-mail: gaudin.theophile@gmail.com; haibo@nju.edu.cn

† Electronic Supplementary Information (ESI) available: technical details about the simulation of molecular contacts and dimers. See DOI: 10.1039/x0xx00000x

which require a high amount of computational resources, can be focused on studying more complex and inhomogeneous systems of which the essential behavior cannot be described based on a simpler model picture.

Besides MD simulation, condensed phase effects can also be described based on solvent-accessible molecular surfaces. Solvent-accessible molecular surfaces are modeling tools designed to represent the limit between the molecule and the external environment (e. g., the solvent). Interest in molecular surfaces began in 1971 when Lee and Richards[7] used it to investigate the structure-body function relationship of proteins. Since then, molecular surfaces have gained prominence in the treatment of solvation effects through the calculation of Apparent Surface Charges (ASC)[8]. In particular, one of these ASC methods, the COnductor-like Screening MOdel (COSMO)[9], treats solvation effects based on the deviation from an ideal conductor, and this approach led to an efficient solution of the so-called outlying charge problem[10] which had led to errors in solvation energy predictions in the past. The advantages provided by using molecular surfaces to model intermolecular interactions is twofold: first, they are based on quantum chemistry, which has the potential to be accurate and highly transferable, and secondly, their computational cost is much lower compared to MD simulation.

However, at that time, only solute properties such as solvation energy, rather than properties of complete condensed systems, could be modeled since the statistical thermodynamic self-consistency between solutes and solvents was not taken into account. This was addressed in 1995 based on evaluating the interactions between COSMO molecular surface elements[11, 12] and resulted in COSMO-RS (with RS standing for Realistic Solvation), which met a considerable amount of success in helping chemical engineers and pharmaceutical chemists to design new compounds and processes[13-15].

Compared to MD, this approach is computationally efficient because if the molecular surface representation of the molecule is not already available in a database[16] or from previous calculation, the prediction only requires one geometric optimization at the ground state of a single molecule in a standard quantum chemistry level like BP86/TZVP for each distinct structure relevant to represent the composition of the system of interest. For example, a flexible molecule could be represented by a set of only 10 suitably selected conformations, the weights of which are self-consistently optimized as part of the calculation.

The demonstrated potential of molecular surfaces to predict equilibrium thermodynamic properties encourage to evaluate whether other properties of molecular surfaces could be exploited to facilitate predictions in other areas of research and applications. In this work, a molecular contact theory is proposed to simplify understanding and prediction of microscopic polarization exploiting advantageous properties of molecular surfaces. This theory combines local polarization from mutual orientations of dipoles with COSMO-RS statistical mechanics of molecular contacts.

As stated in the beginning of this article, predicting polarization at a microscopic level represents a challenge of broad academic and industrial interest. The dielectric constant ε_r, which is a macroscopic outcome of microscopic polarization, is a relevant target-property because thousands of data[17, 18] are already available as test cases, and also because it has numerous applications, so that the developed models can readily be used beyond the theoretical benefit of a proof-of-concept. So, ε_r is chosen as a target property in the present work.

ε_r represents the decrease of an electric force experienced by a charge due to its surrounding medium. This property is a characteristic of the medium and thus, in principle, is completely determined by its microscopic structure and experimental conditions (temperature and pressure).

In the case of liquid mixtures, the numerous applications derived from knowledge of ε_r can broadly be classified as direct and indirect. Because ε_r directly describes the ability of a medium to screen a charge from an electric field, direct applications are related to ion solvation description[19]. Rationalizing chemical reactions which often involve ions dissolved in solvents, and their yields[20], is a first large field of applications of ion solvation. In analytical chemistry, the proper choice of solvents to analyze ions in High-Performance Liquid Chromatography[21] or electroanalytical measurements[20] (like potentiometry or voltammetry) can be guided by ion solvation energy prediction. ε_r can also directly help in industrial applications, such as optimization of battery electrolytes[22], hydrometallurgy or nuclear waste disposal[20].

Indirect applications of ε_r exploit its correlation with other relevant phenomena. For example, ε_r is often used as a single number-indicator of the polarity of a medium[23] where qualitative knowledge of polarity is required, like in solvent extraction[24] or to elucidate the kinetics and thermodynamics of reactions[25, 26]. Since it is sensitive to the 3D arrangement of compounds, ε_r can also provide experimental indications about the kinds of complexes formed in solution[27] (also called liquid structure). To the end, ε_r is required as a parameter of many equations of state, notably when the goal is to model behavior of ions in solutions[28] and is frequently used as a target property to optimize force fields in MD simulations[29].

As can be seen in the above paragraphs, ε_r has fundamental significance for a large span of academic and industrial, theoretical and experimental applications. These motivated researchers to propose a large set of formulas relating the microscopic structure of materials with ε_r since the 1930s, first phenomenological, and then based on fundamental electrostatic principles (i. e., Maxwell equations)[30, 31]. A great body of literature was focused on clarifying the derivation of such formulas and their underlying assumptions throughout the 20th century. The details of these investigations are not of direct interest for the present work, and a pedagogical summary is provided in a book from Raju[32] for the interested reader. These developments were later combined with statistical mechanics[33] to generate an equation (generally called the Kirkwood-Fröhlich equation) suitable for predictive purposes (eq. 1).

View Article Online
DOI: 10.1039/C9CP02758B

$$\frac{(2\varepsilon_r+1)(\varepsilon_r-1)}{9\varepsilon_r}=\frac{4\pi}{3}\frac{\langle M^2\rangle}{Vk_BT} \quad (1)$$

where $\langle M^2\rangle$ is the expectation value of the squared dipole moment of the simulation box (representing dipole fluctuations), V is the volume of the spherical simulation box, k_B is the Boltzmann constant and T is the temperature.

This equation is nowadays used in MD simulations. Note that the equation could be written in several other forms where the constants 4π or ε_0 (the vacuum permittivity) are either explicitly written or implicitly included in the variables through the choice of units. Neumann chose to write 4π explicitly and to include ε_0 in the units of dipole moment in his paper, perhaps due to his derivation approach. Variants of eq. 1 are available for different box shapes (e. g. square box).

The Kirkwood-Fröhlich equation allows to understand that ε_r of a liquid increases with dipole moment fluctuations and decreases with the molecular volume (or increases with density). Thus, any modeling technique that would yield reliable estimates of these two properties would, in principle, allow to access ε_r. In MD simulations, the bottleneck is to obtain an accurate estimate of $\langle M^2\rangle$, and the volume of the box is fixed. This requires to properly choose a force field that reproduces the experimental dipole moments and molecular interactions, and to simulate a sufficiently large box to get reliable $\langle M^2\rangle$.

In parallel, Quantitative Structure-Property Relationship (QSPR) models were proposed to predict ε_r of liquids. QSPR models are empirical relationships between numbers derived from the molecular structure (the descriptors) and a target property (here ε_r). To the best of our knowledge, only one was developed using a diverse set of organic liquids[34], while some others were developed for particular classes of compounds[35, 36]. These models have the advantage of giving instantaneous results once molecular descriptors are calculated, which can also be computationally cheap depending on which descriptor is required. However, purely empirical, these models are intrinsically limited in their applications to molecules similar to the ones used to fit the model. Moreover, obtained purely empirically, QSPR models do not provide a direct physical interpretation of the results, though the meaning of the selected descriptors can be interpreted according to the targeted property.

It can also be mentioned that techniques based on integral equation theory, such as the hypernetted chain approximation of the molecular Orstein-Zernike theory[37], have been applied to yield predictions of the dielectric constant. However, these techniques require substantial supervision for each molecule since, like MD simulations, they require force-field parameters to be applied. Moreover, their mathematical formalism is generally quite complex and calculations may take some time, though less than MD simulations. For these reasons, they are generally not used for large scale screening but rather to gain detailed insights on a single system or a handful of specific systems.

Evaluating ε_r from molecular surfaces allows to remove the necessity of explicit simulation of a large number of identical interacting structures and provides means to limit the reliance on fitting, by striving in keeping as much generality and physical significance during the conception of the model. Moreover, they can provide a means to screen a large number of molecules with limited user supervision for each molecule. In this study, explore how molecular surfaces and their interactions can lead to ε_r predictions.

1. Theory

In this section, the theory developed to predict ε_r based on local polarization on molecular surfaces is presented. We begin by describing how to compute contact probabilities from COSMO and COSMO-RS. To the best of our knowledge, no framework was developed to directly deal with fluctuations of dipole moments as represented by arbitrary molecular surfaces in contact with each other without assuming some surrounding box. Thus, we continue with deriving a Kirkwood-Fröhlich-like equation that we believe to be appropriate for such a situation from Maxwell equations. Then, we develop a framework to deal with polarization of the medium by a given molecule in the system, as represented by its charged molecular surface. We continue by some phenomenological considerations to model the hydrogen-bonding effect within this framework, which leads to the need of two fitting parameters for a fully predictive model in the present proof-of-concept approach.

1. 1. COSMO, COSMO-RS and Contact probabilities

This study relies on modelling the 3D geometry of compounds through molecular surfaces, and intermolecular interactions in liquids through contacts between molecular surfaces. To the best of our knowledge, the best approach available in the literature for this endeavor is COSMO-RS. In this section, we briefly describe the relevant features of the COSMO molecular surface and the COSMO-RS model based on COSMO molecular surfaces. Note that this work uses COSMO-RS as a workhorse to calculate contact probabilities with the aim to simulate local polarization and does not modify COSMO-RS itself. The calculated contact probabilities by COSMO-RS will be used later on to evaluate coupling between dipoles and therefore the dielectric constant by our theory.

COSMO is an implicit solvation method in quantum chemistry based on a surface on which the reaction of the solvent is projected as a surface charge density. The screening ability of solvents is represented by their dielectric constant ε_r (Note that in such context, ε_r is an external parameter of the calculation and not a target of the modelling). The surface charge density allows to calculate a solvation contribution to the Hamiltonian, which in turns can simulate the effect of the solvent on the geometry of the molecule. For more details about the method, the reader is referred to the original publication[9] and a more recent pedagogical review[14]. Notable outcomes of a COSMO calculation are the total volume enclosed by the COSMO surface (which can yield a rough estimate of the density when combined with the molar mass),

and the COSMO surface itself in the form of a set of surface segments, each having a position vector, a surface area, a surface charge density and an underlying atom associated to them.

COSMO-RS[11, 14] is based on COSMO molecular surfaces obtained in an ideal conductor (i. e. with $\varepsilon_r = \infty$). From a simple formula for the interaction energy density between two segments $e_{int,\eta\nu}$ based on their COSMO polarization charge densities and underlying elements, as well as a fitted effective contact area a_{eff}, COSMO-RS theory allows to self-consistently calculate the chemical potential of each surface segment for each structure in a homogeneous liquid mixture, $\mu^{\eta}{}_{M}$, and the conformer weights of each structure if one compound is present in several conformers.

The original goal of COSMO-RS was the prediction of the chemical potential of molecules in liquids, leading to the prediction of thermodynamic properties. Contact probabilities $p_{\nu|\eta}$ can also be derived as an auxiliary property of a COSMO-RS calculation, as:

$$p_{\nu|\eta} = \frac{p_\nu \exp\left(-a_{\text{eff}} \frac{e_{\text{int},\eta\nu} - \mu_\nu}{k_B T}\right)}{\sum_\nu p_\nu \exp\left(-a_{\text{eff}} \frac{e_{\text{int},\eta\nu} - \mu_\nu}{k_B T}\right)} \quad (2)$$

where p_ν is the probability of a segment to be the one labeled by ν (based on the ratio of its surface area to the surface area of an average molecule in the mixture), μ_ν is the chemical potential of segment ν, k_B is the Boltzmann constant and T is the temperature.

Note that segments can be grouped into clusters[12] as a function of their polarization charge densities and underlying elements. This is used to speed-up the solution of the COSMO-RS equations. Then, for individual segments, the probabilities are derived from their portion in the total area of their respective clusters. Also, it can be noticed that any method leading to a calculation of contact probabilities between two molecular surfaces could be used, even though, here, we use COSMO-RS theory.

The present work aims at simulating microscopic polarization from molecular surfaces. Towards this goal, the COSMO-RS contact probabilities will be used to quantitatively evaluate the coupling between microscopic dipoles. In order to validate the theory, we chose ε_r, which is physically caused by microscopic polarization, as an experimental target for prediction. In the next section, the relevant relationship between coupled microscopic dipoles and ε_r is derived.

1. 2. Relationship between microscopic dipole moment and macroscopic dielectric constant

First, from Maxwell equations applied to electrostatics (cf. Supporting Information, part 2 for a detailed derivation), we can derive a relationship between the polarization **P** and the electric field **E** which is mediated by the dielectric constant ε_r:

$$\mathbf{P} = (\varepsilon_r - 1)\varepsilon_0 \mathbf{E} \quad (3)$$

where ε_0 is the permittivity of the vacuum.

Now we are interested in a set of instantaneous configurations where a molecule is fixed and measure its polarization effect on the surroundings. We assume that our system is represented by this set of instantaneous configurations. We also focus our analysis on polarization and therefore neglect the polarizability effects, due to instantaneous rearrangement of electrons, for the time being. We re-introduce them later in a simplified manner, though this effect is not the focus of the present work. Indeed, this work is focused on advancing towards a simple and efficient description of local polarization, which is generally the main contribution to ε_r at moderate to large values, more than to reach a high-resolution account of all other potentially relevant effects for ε_r.

At low fields, based on these considerations, we can derive a relationship between **E** and the average dipole moment of a random sample at zero field $\langle \mathbf{m} \rangle^0$ based on a Mc Laurin series (cf. Supporting Information, part 3 for a detailed derivation):

$$\langle \mathbf{m} \rangle^0 \approx \frac{\langle m_{\text{eff}}{}^2 \rangle}{3k_B T} \mathbf{E} \quad (4)$$

where $\langle m_{eff}{}^2 \rangle$ is the average of the squared dipole moments of a molecule embedded in the medium under consideration (or effective dipole moments), which include the gas phase dipole moment and the additional dipole moment induced along the dipole vector of the molecule by polarization of the medium.

Meanwhile, by definition, the polarization of the medium, at zero field, is:

$$\mathbf{P} = \frac{\langle \mathbf{m} \rangle^0}{V} \quad (5)$$

In our system, the volume V is the average volume of a molecule in this system (or average molecular volume). Combining equations 4-6, we reach:

$$\varepsilon_r = 1 + \frac{\langle m_{\text{eff}}{}^2 \rangle}{3k_B T V \varepsilon_0} \quad (6)$$

This equation is valid for low fields and nonpolarizable molecules. In practice, the polarizability of molecules leads to ε_r larger than 1 even for completely nonpolar compounds such as alkanes. Since, in general, this value is about 2[17], we considered that this effect would be described well enough by considering $\varepsilon_r = 2$ for molecules with no dipole moments. This led to:

$$\varepsilon_r = 2 + \frac{\langle m_{\text{eff}}{}^2 \rangle}{3k_B T V \varepsilon_0} \quad (7)$$

Eq. 7 is used as a predictive device in this work.

View Article Online
DOI: 10.1039/C9CP02758E

In the next subsection, we present our approach to quantitatively evaluate the coupling between the microscopic (molecular) dipoles.

1. 3. Evaluation of the coupling between microscopic dipoles

Molecules polarize other molecules which are themselves polarized by other molecules. At equilibrium, the following equation describes such a coupling:

$$m_{\mathrm{eff,s}} = m_{\mathrm{s}} + \sum_{\mathrm{t}} x_{\mathrm{st}} \left\langle \cos\theta \right\rangle_{\mathrm{st}} m_{\mathrm{eff,t}} \qquad (8)$$

where $m_{\mathrm{eff,s}}$ is the effective dipole moment of s-th structure in the mixture, m_{s} is the s-th molecule gas phase dipole moment, x_{st} is the fraction of t-th structure around s-th structure in the mixture (which can be approximated as the mole fraction of t-th structure in the mixture) and $\langle\cos\theta\rangle_{\mathrm{st}}$ is the average cosine of the angle θ between s-th and t-th structures. The summation runs over all distinct structures in the mixture. For a pure solvent, it may be all distinct conformers of the molecule.

Eq. 8 can be rewritten as a matrix equation:

$$\mathbf{m}_{\mathrm{eff}} = \mathbf{C} \cdot \mathbf{m} \qquad (9)$$

where $\mathbf{m}_{\mathrm{eff}}$ is the vector of effective dipole moments, **m** is the vector of gas phase dipole moments, and **C** is a coupling matrix defined by:

$$\mathbf{C} \equiv \left(\mathbf{1} - \mathbf{x} \cdot \mathbf{cos\ \theta}\right)^{-1} \qquad (10)$$

where 1 is the unitary matrix, **x** is the matrix of mole fractions of structures around all other structures, **cos θ** is the matrix of average cosine of θ between all structure pairs. With eq. 8 written this way, a simple matrix inversion can achieve prediction of m_{eff} for every structure in the mixture once mutual average orientations and local compositions are properly estimated.

In the next subsection, we describe our approach to evaluate the elements of the **cos θ** matrix.

1. 4. Contactwise averaging of angles between dipoles

The expectation value of the cosine of the angles between the dipole moments of each structure with respect to the considered structure $\langle\cos\theta\rangle_{\mathrm{st}}$ can be defined as:

$$\left\langle \cos\theta \right\rangle_{\mathrm{st}} = V_{\mathrm{s}}^{-1} \int_{\mathrm{s}} \left\langle \cos\theta \right\rangle_{\mathrm{t}} \left(\mathbf{r}_{\mathrm{s}}\right) d\mathbf{r}_{\mathrm{s}} \qquad (11)$$

with

$$\left\langle \cos\theta \right\rangle_{\mathrm{t}} \left(\mathbf{r}_{\mathrm{s}}\right) = \int_{\mathrm{t}} p'\left(\mathbf{r}_{\mathrm{t}} \mid \mathbf{r}_{\mathrm{s}}\right) \left\langle \cos\theta \right\rangle \left(\mathbf{r}_{\mathrm{s}}, \mathbf{r}_{\mathrm{t}}\right) d\mathbf{r}_{\mathrm{t}} \qquad (12)$$

where V_{s} is the total volume in which contacts can occur around s-th structure, $\langle\cos\theta\rangle_{\mathrm{t}}(\mathbf{r}_{\mathrm{s}})$ is the expectation value of the cosine of the angle between the dipoles of the t-th structure and the s-th structure at coordinate $\mathbf{r}_{\mathrm{s}}$ in the system of the s-th structure, $p'(\mathbf{r}_{\mathrm{t}} | \mathbf{r}_{\mathrm{s}})$ is the probability distribution of a contact between a point in $\mathbf{r}_{\mathrm{t}}$ coordinate of t-th structure and a point in $\mathbf{r}_{\mathrm{s}}$ coordinate of s-th structure, and $\langle\cos\theta\rangle(\mathbf{r}_{\mathrm{s}}, \mathbf{r}_{\mathrm{t}})$ is the expectation value of the cosine of the angle between the t-th and s-th structures for this specific contact (averaging over all possible rotations about this contact point).

Note that to preserve generality, eq. 11 does not assume the presence of a molecular surface, which is a modelling object. It considers that contacts between molecules can occur in every point in the coordinate systems of molecules. However, in reality, contacts will only occur close to the molecular surface, which has been constructed for this very reason. So, eq. 11 can be particularized in terms of surface elements rather than volume elements, which leads to:

$$\left\langle \cos\theta \right\rangle_{\mathrm{st}} = A_{\mathrm{s}}^{-1} \int_{\mathrm{s}} \left\langle \cos\theta \right\rangle_{\mathrm{t}} \left(\mathbf{y}_{\mathrm{s}}\right) d\mathbf{y}_{\mathrm{s}} \qquad (13)$$

with

$$\left\langle \cos\theta \right\rangle_{\mathrm{t}} \left(\mathbf{y}_{\mathrm{s}}\right) = \int_{\mathrm{t}} p'\left(\mathbf{y}_{\mathrm{t}} \mid \mathbf{y}_{\mathrm{s}}\right) \left\langle \cos\theta \right\rangle \left(\mathbf{y}_{\mathrm{s}}, \mathbf{y}_{\mathrm{t}}\right) d\mathbf{y}_{\mathrm{t}} \qquad (14)$$

where, this time, **y** is a surface coordinate rather than a 3D coordinate.

The remaining problem is to find a workable, computable model to test this theory. In the next section, a rough model which requires some parameterization is built for this. This model is tested in the remainder of the article. The attention of the reader is turned to the fact that this article is mainly a proof of concept that such formulation opens the way to a new general predictive perspective microscopic polarization; the detailed model can still be improved later on to reach better accuracy for ε_{r} in particular with fewer, or even no parameter.

1. 5. Rotational averaging of angles between dipoles at a contact point

As a first approximation, overall mole fractions of each structure in the mixture can be taken as estimates of local composition. In addition, for the present purposes, there is a need for a scheme adapted to molecular surfaces and their contacts to estimate $\langle\cos\theta\rangle_{\mathrm{st}}$. In this study, for the sake of simplification, we assume free rotation around contact points between molecular surfaces as an ideal starting point, except for the fraction of these contacts which are locked by a dimerizing interaction (like, for example, in the case of acetic acid immersed in a nonpolar environment[11]). The numerical approach chosen in this work to simulate molecular contacts based on COSMO molecular surface segments, focused on H-bonding which is the most relevant phenomenon that impacts coupling between dipoles in water and organic liquids, is provided in Supporting Information, section 1.

Let us view molecular surfaces as sets of segments, all in contact with each other in the mixture at equilibrium. First, under the free rotation assumption, the 3D law of cosines gives (cf. Supporting Information, part 4 for a detailed derivation), for a given contact between two molecular surface segments (labelled η and ν), the following relationship for an angle $\theta_{\eta\nu}$ between the dipole moments of the two molecules:

$$\cos\theta_{\eta\nu} = -\cos\theta_{\eta}\cos\theta_{\nu} \quad (15)$$

where θ_η is the angle between the surface normal at η and the dipole moment vector of the underlying molecule associated with η (called dipole angle in this article), and accordingly, θ_ν is the dipole angle at surface segment ν.

Guided by preliminary tests under the free rotation assumption, we assumed that the free rotation assumption alone leads to an underestimation. This indicates that compounds, especially big ones, tend to favour chain-like HB network for sterical hindrance reasons. Thus, we conceded two adjustable parameters in the form of a function which, for a given structure, decreases the effective angle between dipoles at a contact to simulate this deviation from free rotation. These parameters are meant to complement the proposed molecular contact theory which neglects the steric/entropic constraints arising for chains of molecules in a discrete solvent. This 2-parameter correction writes:

$$\theta_{\mathrm{s,corr}} = x_{\mathrm{M}}^{\mathrm{HB}} \max\left(k_1 V_{\mathrm{s}}, k_2\right) \quad (16)$$

$$\theta_{\eta\nu} = \max\left(0, \cos^{-1}\left(\cos\theta_{\eta}\cos\theta_{\nu}\right) - \theta_{\mathrm{s}\ni\eta,\mathrm{corr}}\right) \quad (17)$$

where $x^{\mathrm{HB}}{}_{\mathrm{M}}$ is the fraction of structures which can both give and receive HB (thus participating in potential HB chains), V_s is the molecular volume of the s-th structure (i. e. a representation of its size). k_1 is a parameter which quantifies the proportionality of this correction to the molecular volume (i. e., the deviation from free rotation about HB is expected to be lower for smaller molecules), and k_2 is a parameter which quantifies the maximum possible correction from free rotation. This correction was only applied for molecules which can both give and receive HB as it seems reasonable to assume that molecules which can only receive HB are less likely to be in chains of approximately parallel dipole moments. We assume that the probability of molecules to form chains is proportional to the number of molecules that form chains in the liquid, which led to penalize the correction by $x^{\mathrm{HB}}{}_{\mathrm{M}}$. To the best of our knowledge, there is no obvious way to calculate the deviation from free rotation of contacts from first principles. This has to be considered as the theoretical bottleneck of the present method.

Secondly, in the case of dimerization (our algorithm for automatic anticipation of dimerizing behavior on a per-segment basis is provided in Supporting Information, part 1), we assume that for a dimerizing contact, the two molecules are locked in a 180° angle between each other (by definition), which leads, in this case, to:

$$\cos\theta_{\eta\nu}^{\mathrm{d}} = -1 \quad (18)$$

2. Computational details

2. 1. Dielectric constant calculation procedure

In the following, the steps used to apply the theory proposed in the present work to predict ε_r are summarized.

1. COSMO surfaces are generated from any available method for the relevant conformers of the compounds in the mixture
2. COSMO-RS equations are solved for the system, yielding both conformer weights and contact probabilities (eq. 2) between segments. Fractions of each structure are calculated by multiplying the mole fraction by the COSMO-RS obtained conformer weight
3. Normal vectors are calculated for each segment (cf. Supporting Information, section 5)
4. Dipole moment vector **m** of each structure in the mixture is estimated based on COSMO polarization charge densities σ_η, areas a_η and positions $\mathbf{r}_\eta$ of each segment of the COSMO molecular surface of structures, as $\mathbf{m} = \sum_\eta \sigma_\eta a_\eta \mathbf{r}_\eta$. Note that the magnitude of dipole moment vector (i. e. the dipole moment itself), is stored as well for later use
5. Dipole angles are calculated for each segment using basic geometry for angle between the dipole moment of the structure and the normal vector of the segment
6. Dimerizing contacts are identified based on the procedure outlined in Supporting Information, section 1
7. Rotational averages of angles between dipoles for each HB-contact (identified based on the COSMO-RS threshold for hydrogen bonding) are evaluated using eqs. 15-18
8. Contactwise averages of angles between dipoles $\langle\cos\theta\rangle_{st}$ are calculated using eq. 14, and then eq. 13
9. Eqs 9-10 are used to predict the effective dipole moments of each structure in the system from the $\langle\cos\theta\rangle_{st}$, fractions of each structure, and dipole moments obtained based on the COSMO-surface
10. Square of calculated effective dipole moments and volumes of the COSMO cavities are averaged based on the fractions of each modeled structure
11. These averages are used in eq. 7 as $\langle m_{eff}^2\rangle$ and V, respectively, to predict ε_r.

2. 2. Experimental data

A dataset of 433 dielectric constants of various pure liquids was extracted from the CRC Handbook of Chemistry and Physics[17]. The COSMO surfaces of 421 of these liquids were available in the COSMObase 2017 database[38]. One of these compounds (selenium oxychloride) contained selenium (Se), for which no

DOI: 10.1039/C9CP02754E

COSMO-RS dispersion parameter could be found, and thus this compound was discarded. The study was focused on the 420 remaining structures optimized using DFT (with BP86 functional and TZVP basis set). Each compound was represented with one to ten conformations.

For mixtures, the data were extracted from the database built by Wohlfarth[18]. Since the focus of this article is not on the temperature dependency, only the data at ambient temperature (298.15K) were selected. Remaining duplicates were eliminated as follows: the data covering the largest range of composition (e. g. from 0 to 1 mole fraction of the 2nd component) and containing the largest amount of data were kept whereas the other data were not considered. After this data curation, 704 data points for 46 aqueous mixtures and 3792 data points for 269 organic mixtures between 111 compounds remained.

2. 3. Analysis of results

Errors were estimated from the determination coefficient R^2, the mean absolute error MAE ($|\varepsilon_{r,exp} - \varepsilon_{r,calc}|$) and the mean relative error MRE (MAE/$\varepsilon_{r,exp}$). Besides, the ability of the model to determine relevant solvent's ε_r category was estimated by a scale inspired by Griffiths[39]. In this article, on the ground that solvents with similar dielectric constants show similar ability to dissolve compounds, Griffiths proposed four solvent classes: Hydrocarbon solvents (ε_r = 0-2.5), electron-donor solvents (ε_r = 3.5-10), hydroxylic solvents (ε_r = 10-35), and dipolar aprotic solvents (17 and more with $\varepsilon_r m$ above 50, with m, the dipole moment, in debye). Based on this view, the classification in Table 1 was considered meaningful.

class label	dielectric constant range
weak	1-3.5
low average	3.5-10
high average	10-35
strong	35+

Table 1 Dielectric constant classes considered in evaluating the performances of the model.

We calculated confusion matrix statistics[40] from Table 1 classification (which will be termed Griffiths' classification in the remainder of the article). The overall accuracy was calculated as the sum of correctly predicted Griffiths' classes divided by the total amount of data, and predictivities were evaluated for each class as the ratio of correct prediction to total amount of predictions for a given class.

2. 4. Implementation

All calculations were performed on a personal computer with i7 processor and 16GB of RAM. The COSMO-RS theory was implemented following Pye et al. [41] approach. All COSMO-RS parameters given by Pye et al.[41] were used (as an average over the five sets of proposed parameters). The temperature dependency coefficients and the binary dispersion parameters for C, H and O were not present in the Pye et al. study, and thus the ones of the Andersson et al.[42] study were used. To the end, dispersion constants for F, Br, I, Si, P and S, which were also not present in Pye et al. study, were extracted from Klamt's book[43]. The implemented COSMO-RS theory was used as a workhorse to predict contact probabilities required to apply the new molecular contact theory of this paper. Both COSMO-RS theory and the new molecular contact theory were combined in the same Python script, which was then applied to calculate dielectric constants of pure liquids and binary mixtures.

3. Results and discussion

3.1. Parameterization

k_1 and k_2 parameters in eq. 16 were fitted on 20 compounds part of the CRC dataset (cf. Fig. 2). To obtain these parameters, many tens of pairs of values were manually attempted until a satisfactory solution was found and no further apparent improvement in R^2 could be obtained. The final values (also reported in Table 2) were k_1 = 0.4375 °Å^{-3} and k_2 = 35°, with $\theta_{s,corr}$ in degrees (°) and the volume in cube angstroms (Å^3). The compounds used for the parameterisation (cf. Table 3) were selected to cover the most expected behaviors in terms of polarization: no significant polarization (octane, propanone, dimethylformamide, acetonitrile), HB alignment (hydrogen cyanide, methylformamide, methanol, 1-butanol), weak HB (1-butylamine, n-propylamine, dipropylamine), HB 3D network (water, ethylene glycol, glycerol, diethanolamine), and dimerization (acetic acid, butyric acid, but also formic acid, 2-pyrrolidon and formamide).

parameter	value	unit
k_1	0.4375	°/Å^3
k_2	35	°

Table 2 Fitted parameters for eq. 16

The two highlighted compounds of Fig. 2 (a) (diethanolamine and ethylene glycol) show relatively high errors, both in terms of relative and absolute error, despite the fitting. In the case of ethylene glycol, the molecule has only ten distinct conformers (accounting for degeneracy) which are not all sampled in the COSMObase. In the case of diethanolamine, one of the ten considered conformers ends up having a high weight, and the molecule has many rotatable bonds. It is likely that this sampling misses significant low-dipole moment conformers. More detailed conformational analyses are reported in the next subsection for these two compounds.

3.2. Conformational analysis (glycol and diethanolamine)

Glycol and diethanolamine yielded poor predictions despite being part of the parameterization. Since both compounds are conformationally flexible (cf. Fig. 1), a possible explanation is that the conformational space is not properly spanned in the original conformations of the COSMObase. In order to explore the conformational space of these two molecules, we considered the staggered configurations of the various rotatable bonds as represented in Fig. 1.

View Article Online
DOI: 10.1039/C9CP02758F

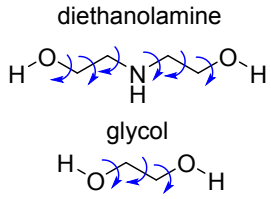

Fig. 1 Rotatable bonds (in blue) explored for glycol and diethanolamine in the present work.

We included the rotations about the C-O bonds in our treatment as the orientation of hydrogens is critical both for the dipole moment and for the relative orientation of hydrogen-bonded conformations.

Fortunately, since glycol is relatively small as well as highly symmetrical, the full conformational space could be covered with just 10 conformations, including the degeneracies arising from their enantiomers. The more realistic option of using all these conformations and considering the associated degeneracies led to a calculated dielectric constant of 41.0, which is in excellent agreement with experimental data.

molecule	ε_r exp	ε_r calc
hcn	114.9	136.6
h2o	80.1	89.2
methanol	33.0	39.5
formamide	111.0	106.7
methylformamide	189.0	148.2
dimethylformamide	38.3	37.8
1-butanol	17.8	18.6
propanone	21.0	23.8
glycol	41.1	15.2
diethanolamine	25.8	61.3
glycerol	46.5	38.5
hydrazine	51.7	52.7
formicacid	51.1	50.5
2-pyrrolidon	28.2	39.2
butyricacid	3.0	4.6
aceticacid	6.2	5.5
1-butylamine	4.7	5.3
n-propylamine	5.1	6.5
acetonitrile	36.6	46.7
dipropylamine	2.9	3.3

Table 3 Experimental vs. calculated relative permittivities of the 20 compounds used for the fitting.

A full rotamer search using the CONFAB software as implemented in Openbabel led to 263 different conformations for diethanolamine (discarding conformations leading to interpenetrating atoms). We wanted to explore whether a relevant choice of 10 conformations could already represent an improvement. Since we had spanned the full rotamer space for diethanolamine, we carried out the calculation on the first 10 conformations. Note that the difference in COSMO energies between the most favorable and least favorable conformer in the first 10 conformations was only -1.5 kcal/mol, thus, they were likely not representing the full conformational diversity of diethanolamine.

Furthermore, all their dipole moments were between 4.37 and 5.57 D, which represents a low dipole moment diversity. Considering this restricted ensemble led to one dominating conformer (with a weight of 41%), and a too high prediction (69.4), like for the default database.

So, ranking conformations by COSMO energy, we decided to mix series of ten conformations from low energies to high energies with this dominating conformer and kept the ones which represented weights above 10% at the end of each COSMO-RS calculations until we had 8 alternative conformers. We included the lowest COSMO energy conformer, the dominating conformer in the initial calculation and these 8 alternative conformers. The conformers obtained this way spanned geometries allowing for a larger diversity of dipole moments (1.44 to 5.38 D), thus probably more representative of the dipole moment distribution in the real solvent. This allowed us to obtain ε_r = 23.8, in much closer agreement with experiment. In this calculation, most of the represented conformers were represented in significant weights, above 5%.

These conformational analyses strongly suggest that the high errors for the two molecules were due to inadequately sampled conformers and point out this as a potential interpretation for high discrepancies with experimental data in future calculations.

Moreover, the results highlight the importance of conformational sampling on predicted ε_r, which will be addressed in further detail in a separate study. Since we did not modify the k_1 and k_2 parameters for this analysis, they also confirm our parameterization choice, as shown in Fig. 2 (b).

3.3. Pure compounds

With k_1 = 0.4375 Å^{-3} and k_2 = 35°, the model was applied to the full set of 420 molecules. Overall, the performances are better than the ones obtained from molecular dynamics, and on par with those obtained for QSPR models which are based on empiricism rather than an attempt to directly reflect the physical phenomenon in the model (cf. Table 4).

Regarding molecular dynamics in particular, recent developments using the thermalized Drude oscillator model[43] yielded good results for water and ethanol in minutes (in the case of ethanol, about two minutes without polarizable force fields and 6 minutes or more using polarizable force fields) on a parallelized 12-processor workstation. This indicates that with quality implementations and modern devices, computational times of molecular dynamics may become more tractable. However, semi-implicit methods such as the one proposed in this paper still have the inherent advantage of not requiring computation of explicit interactions between all molecules in a sample of the system. Already with the rough implementation of the present theory in a python script, reasonable values can be obtained for water and ethanol within a few seconds. Therefore, it makes no doubt that with an optimized implementation, the present method could be used for large-scale screening of relatively simple systems with low supervision from the user, which remains an inherent challenge with molecular dynamics simulations. Since the systems

View Article Online
DOI: 10.1039/C9CP02758B

simulated with molecular dynamics take similar computational times no matter the complexity of the intermolecular and mesoscopic simulated arrangement for a given box size, the present semi-implicit method allows to focus the molecular dynamics efforts for dielectric behavior prediction in the tasks where they have a unique advantage, i. e. where explicit interactions of most MD simulated molecules are relevant due to specific arrangement, such as in microporous materials or polyelectrolytes with complex dielectric response functions.

Though the performance is good overall, there is still some remaining scattering (cf. Fig. 2 (c)). Some patterns emerged upon the analysis of the largest errors in prediction. First of all, a major part of the largest errors were obtained for flexible compounds (e. g. glyme with $\varepsilon_{r,exp}$ = 7.2 and $\varepsilon_{r,calc}$ = 3.4, 53% underestimation, or 1,3-butanediol with with $\varepsilon_{r,exp}$ = 28.8 and $\varepsilon_{r,calc}$ = 45.6, 58% overestimation), which is in line with errors observed during the parameterization and may be addressed by further conformational analyses.

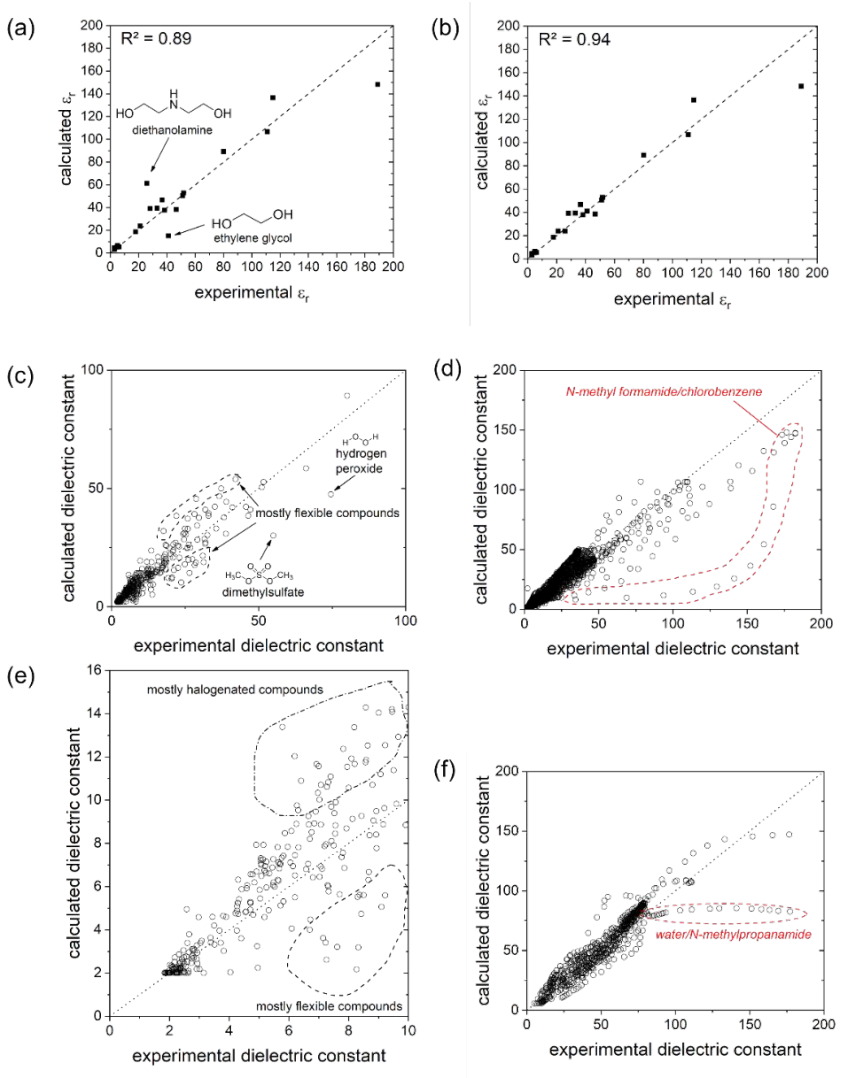

Fig. 2 Scatter plots of experimental vs. calculated εr for: (a) compounds of the parameterization; (b) compounds of the parameterization after complementary conformational analyses; (c) pure liquids; (d) organic mixtures; (e) pure liquids with $\varepsilon_r < 10$; (f) aqueous mixtures.

set	Griffith's accuracy	MRE	MAE	R^2	ref
pure compounds	84%	21%	2.5	0.91	present work
Sild, 2002 (QSPR pure compounds)	-	27%	-	-	[34]
Caleman, 2012 (MD pure compounds)	-	35-43%	-	<0.60	[44]
organic binary mixtures	85%	20%	2.8	0.88	present work
aqueous binary mixtures	90%	16%	7.6	0.84, 0.92*	present work

ARTICLE View Article Online
DOI: 10.1039/C9CP02358E Journal Name

*without water/N-methyl propanamide mixture

Table 4 Validation metrics of literature methods and present work for ε_r prediction.

For low (<10) dielectric constants (cf. Fig. 2 (e)), overestimations were mainly associated with simple halogenated compounds, without any accounting for HB (e. g. 3-bromopropene, with $\varepsilon_{r,exp}$ = 7 and $\varepsilon_{r,calc}$ = 11.6, 65% overestimation). Possible causes are an inadequate parameterization of halogens, leading to overpolarization compared to experiment, or a non-HB antiparallel association effect related to halogen bonding.

Among the overestimations, some contained a terminal H-C=O moiety (aldehyde or formate). For example, ethyl formate has $\varepsilon_{r,exp}$ = 8.6 and $\varepsilon_{r,calc}$ = 14.3, which is a 67% overestimation. Despite this moiety not forming HB, a 6-ring dimerization between such moieties may occur substantially, which cannot be detected by the present treatment that only relies on HB.

To the end, two specific compounds were significantly underpredicted among compounds with a high dielectric constant, dimethylsulfate and hydrogen peroxide. In the case of dimethylsulfate ($\varepsilon_{r,exp}$ = 55.0 and $\varepsilon_{r,calc}$ = 30.0, 45% underestimation), a possible cause may be a rare form of associating behavior among non-HB compounds that cannot be detected by the present method. Indeed, the multiple oxygens on the sulfur may substantially attract the electrons of the methyl moieties (cf. Fig. 3), making them polar not to the point of generating hydrogen bonds, but enough to generate significant linear association that would then result in an increased dielectric constant.

Hydrogen peroxide ($\varepsilon_{r,exp}$ = 74.6 and $\varepsilon_{r,calc}$ = 47.5, 38% underestimation) is a small molecule which predominantly interacts through hydrogen bonding in its own liquid phase. In such a case, the approximation used implicitly by the method that molecules are frozen (i. e. atomic polarization due to internal vibrations is not taken into account) might not be sufficient.

The case of hydrogen peroxide also suggests that the decrease of the deviation from the ideal free rotation for very small molecules used in the present parameterization might encounter success due to the parallel increase of role played by atomic polarization at low sizes that makes a constant correction not a sufficient approximation at these scales. It seems that the success of this approximation scheme could stem from the fact that atomic polarization could be largely independent of external interactions and more related to internal vibrations of molecules, thus generally decreasing the electric screening effect of molecules. In such case, hydrogen peroxide would be the exception to the rule, with its random vibrations actually increasing somewhat the polarization in the liquid. This hypothesis is supported by a molecular dynamics simulation of hydrogen peroxide[45] which concluded that the geometry and charge distribution of hydrogen peroxide, when compared to the ideal gas-phase conformation (also used in the present work), undergo significant fluctuations in the liquid phase that enhance the dipole moment of the molecule. In the future, a possible way to overcome such an issue would be to model mathematically the spreading of dipoles due to vibrations.

The overall Griffith's accuracy (84%) is good when compared to a random 4-class accuracy (25%). All errors of classification come from classifying the molecule to a neighboring class, which is the least problematic classification error. When looking at the individual classes (cf. Fig. 4 (a)), the most common error is misclassification of a molecule as having an average high ε_r rather than an average low ε_r. As previously noted, these errors are generally associated with a systematic overprediction of ε_r for halogenated compounds such as 3-bromopropene. A prediction of strong ε_r has 33% of chance to be issued for a molecule which has only a high-average ε_r, which is mostly associated with conformationally flexible compounds such as 1,3-butanediol. This emphasizes the need of careful consideration of significant conformers with both low and high dipole moments for these molecules.

Nevertheless, the method as it stands already gives a quite reliable estimation of the order of magnitude of dielectric constant of molecules which can find use, for example, in predicting yields of reactions in solvents. The patterns appearing in the errors can guide the user in interpreting predictions depending on the compound.

3. 4. Binary mixtures

3. 4. 1. Organic mixtures

The method was used with no modification for the 3792 organic mixtures. As Table 4 shows, the performances were comparable with those for pure compounds, indicating that the method can be used, on the whole, as reliably for pure compounds and for mixtures. The scatter plot (Fig. 2 (d)) allows to identify a single mixture for which the present model failed: N-methyl formamide/chlorobenzene.

Upon investigating the original publication[46], an inconsistency is spotted in the published data. Indeed, the excess dielectric constants published by the authors are not consistent with the ones that can be calculated from the published dielectric constants. The formula used by the authors to calculate the excess dielectric constant ε^E is:

$$\varepsilon^{E} = \left(\varepsilon_{r,m} - \varepsilon_{\infty}\right) - \left[\left(\varepsilon_{r,CBZ} - \varepsilon_{\infty}\right)x_{CBZ} + \left(\varepsilon_{r,NMF} - \varepsilon_{\infty}\right)x_{NMF}\right] \quad (19)$$

where $\varepsilon_{r,m}$ is the dielectric constant of the mixture, ε_∞ is chosen as 2 by the authors, $\varepsilon_{r,CBZ}$ is the dielectric constant of pure chlorobenzene (5.54), x_{CBZ} is the mole fraction of chlorobenzene, $\varepsilon_{r,NMF}$ is the dielectric constant of N-methyl formamide (176.54), and x_{NMF} is the mole fraction of N-methyl formamide.

View Article Online
DOI: 10.1039/C9CP02358E

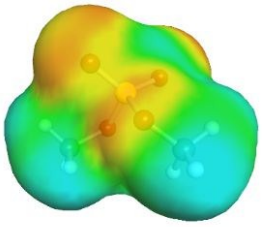

Fig. 3 COSMO surface of dimethyl sulfate.

Using eq. 19, we calculated the ε^E from Pawar et al.[46] data and compared them to the values provided graphically in the paper (cf. Table 5). The values are widely different. This inconsistency prevents us from elucidating the cause of the large errors in the present work, as it is uncertain whether they can be explained from an insufficiency of the model presented in this paper or from errors in experimental data. New experimental analyses are recommended for chlorobenzene/N-methyl formamide mixture to clarify the inconsistency.

In terms of classification, the method again reveals as efficient as for pure compounds, with an accuracy of 85%. Only two of the 577 classification errors were more than a neighbor classification error, thus the erroneously classified molecules can reasonably be assumed to belong to neighbor classes. Classification accuracy is relatively homogeneous among the various classes (74-90%, cf. Fig. 4 (b)). The large majority of the erroneously classified "low average ε_r" were experimentally "high average ε_r" molecules, probably due to a slight overall underestimation of dipole correlation in systems composed of different species. This may be due to the neglect of any correlation arising from interactions other than hydrogen bonding and may be corrected in later refinements of the method.

x_{NMF}	ε^E (recalculated)	ε^E (published[46])
0.0553	19.9	-84.0
0.1163	43.8	-100.6
0.1841	57.0	-72.7
0.2599	63.7	-22.3
0.3450	65.3	27.2
0.4413	60.2	54.2
0.5513	52.9	65.7
0.6781	39.6	88.5
0.8258	20.9	139.3

Table 5 Comparison of published and recalculated excess permittivities for chlorobenzene/N-methyl formamide mixture from Pawar et al.[46]

Nevertheless, the success met in the prediction of ε_r of mixtures already allows potential use of them as a screening tool. Moreover, the calculated dielectric constants could be used to parameterize the implicit solvation model in quantum chemical calculations so that the effect of mixed solvents can be simulated in any calculation where implicit solvation is relevant, such as for lanthanide/actinide separation through complex formation[47].

3. 4. 2. Aqueous mixtures

Absolute errors for aqueous mixtures should be interpreted differently from the metrics for pure compounds and organic mixtures. Indeed, most (73%) aqueous mixtures of the database have strong (>35) ε_r. So, larger absolute errors are expected for a given relative error. With that in mind, the performances in modelling these aqueous mixtures are good and in line with the performances achieved for pure compounds and organic mixtures (with MRE of 16%). One of the mixtures, with N-methyl propanamide (cf. Fig. 2 (f)), substantially decreases the R^2, from 0.92 to 0.84. This is due to the underestimation of ε_r of N-methyl propanamide. Indeed, its ε_r is close to the one of N-methyl formamide, despite having a greater volume and a similar dipole moment, which indicates that N-methyl propanamide aligns even more than N-methyl formamide in the liquid phase, a feature that cannot be predicted by the current model and would deserve further theoretical investigations.

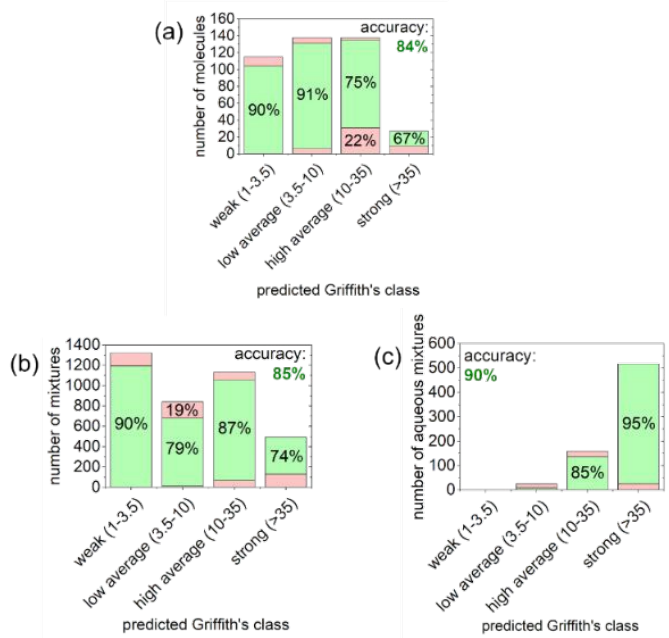

Fig. 4 Griffith's accuracy for each class, (a) for pure liquids, (b) for organic mixtures, (c) for aqueous mixtures.

The classification accuracy is high (90%), but most molecules fall in either "high average" or "strong" class (cf. Fig. 4 (c)). Only 10 of the 704 aqueous mixtures of the dataset belong to the "low average" class, and 65% of the molecules predicted in belonging to "low average" class were in fact in the "high average" class. Thus, prediction of "low average" classes are likely to be erroneous for aqueous mixtures in general. As previously for organic mixtures, this error may be due to the neglect of alignment interactions which are not related to conventional HB. Nevertheless, these errors only affect a minority of cases and the overall performance remains as good for aqueous mixtures as for other types of systems.

Conclusions

This work demonstrates that molecular surfaces and their interactions (here modeled using COSMO-RS theory) are a

useful proxy to characterize the polarization of molecules. The concepts presented in this article allow for an original description, both qualitative and quantitative, of the microscopic polarization in associated liquids. This perspective could be summarized as the combination of a local evaluation of mutual orientations based on molecular surface contacts with an analytical treatment of effective dipole coupling.

When applied to estimate the dielectric constant, the approach proves a good predictive power, better than from molecular simulation or QSPR models for such a large diversity of liquids, with a low computational time, in the order of seconds to minutes on a single CPU laptop computer. A mean relative error of about 20% is obtained for 420 pure compounds, 269 organic mixtures and 46 aqueous mixtures. This predictive model can already be used to quickly screen solvents (both pure and mixtures) whenever a given range of dielectric constants is targeted for a particular application.

Though the first results obtained are already encouraging, this work opens a whole area of investigation for refinements. In particular, the importance of sampling conformations that represent the diversity of dipole moments of a compound in its liquid phase should be characterized. Dielectric constants below 10 should be investigated in more detail and prediction accuracy could be improved, as the actual value of ε_r becomes important in this range to predict the electrostatic part of the ion solvation energy, which is important in many applications. Halogenated compounds seem to be systematically overestimated for an unknown cause which is still to be elucidated. Moreover, specific systems that could not be accounted for by the approximation scheme presented in this work have been emphasized: dimethyl sulfate, hydrogen peroxide, N-methyl propanamide, and the N-methyl formamide/chlorobenzene mixture. These systems seem to adopt behaviors that should be clarified, potentially bringing a more accurate understanding of intermolecular interactions in liquids.

In terms of applications, the method could be tested (and potentially refined) for many more industry-relevant and challenging systems such as ionic liquids or deep-eutectic solvents. For example, in the specific case of ionic liquids[48], a recurring challenge is to find an ionic liquid which would allow good dissolution of a given water-insoluble material. In that context, our framework could allow to evaluate how cations and anions associate geometrically, leading to more accurate solubility predictions.

Another future improvement of the present theory could be towards polyelectrolytes, which are promising components in the design of more efficient electrochemical devices in the energy sector[49]. Indeed, one could imagine representing block copolymers in terms of connectivity relations between their monomers, and using the coupling equations 8-10, their mutual polarization could be estimated, potentially leading to cost-efficient prediction of dielectric response functions in the context of ion solvation in polyelectrolytes.

To the end, the free rotation assumption about contacts between molecular surfaces leads to an underestimation and the deviation from free rotation had to be evaluated from two fitted parameters in this work. In the future, it would be of interest to derive a first-principle formula to quickly predict the deviation from free-rotation, thus releasing the need of fitted parameters which decrease the generality of the method.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Dr F. Eckert for his advice to implement an in-house COSMO-RS script to test the concepts presented in this manuscript. The work was supported by the National Natural Science Foundation of China [Grant Nos: 21722302 , 21673109] and the Fundamental Research Funds for the Central Universities [Grant No: 020414380126].

References

1. B. E. Poling, J. M. Prausnitz and J. P. O'Connell, *The Properties of Gases and Liquids*, Mc Graw-Hill, New York, 5 edn., 2000.
2. M. Paluch, *Dielectric Properties of Ionic Liquids*, Springer, New York, 2016.
3. T. R. J. K. Xu, O. Borodin and M. Ue, *Electrolytes for Lithium and Lithium-Ion Batteries*, Springer, New York, 2014.
4. C. Reichardt, *Solvents Effects in Organic Chemistry*, Wiley, Marburg, 1990.
5. M. L. Klein and I. R. McDonald, *J. Chem. Phys.*, 1979, **71**, 298-308.
6. Y. Wang, M. J. M. Mazack, D. G. Truhlar and J. Gao, in *Many-Body Effects and Electrostatics in Biomolecules*, CRC Press, Boca Raton, 2016, pp. 33-64.
7. B. Lee and F. M. Richards, *J. Mol. Biol.*, 1971, **55**, 379-IN374.
8. J. Tomasi, B. Mennucci and R. Cammi, *Chemical Reviews*, 2005, **105**, 2999-3094.
9. A. Klamt and G. Schuurmann, *J. Chem. Soc., Perkin Trans. 2.*, 1993, DOI: 10.1039/P29930000799, 799-805.
10. J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999-3093.
11. A. Klamt, *J. Phys. Chem.*, 1995, **99**, 2224-2235.
12. A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074-5085.
13. A. Klamt, *COSMO-RS, From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, Elsevier Science Ltd., Amsterdam, The Netherlands, 2005.
14. A. Klamt, F. Eckert and W. Arlt, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 101-122.
15. A. Klamt, *Wiley Interdiscip. Rev. Comput. Mol. Sci*, 2018, **8**, e1338.
16. F. Eckert, *COSMOtherm Reference Manual, version C3.0 Release 17.01*, COSMOlogic GmbH & Co KG, Leverkusen, Germany, 2016.
17. *Handbook of Chemistry and Physics*, CRC Press, Boca Raton, Internet Version edn., 2005.
18. C. Wohlfarth, *Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures*, Springer, New York City, 2008.

View Article Online

19. A. Jouyban and S. Soltanpour, *J. Chem. Eng. Data*, 2010, **55**, 2951-2963.
20. Y. Marcus, *Ions in Solution and Their Solvation*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2015.
21. S. C. Moldoveanu and V. David, *Essentials in modern HPLC separations*, Elsevier, Amsterdam, The Netherlands, 2013.
22. A. Ponrouch, E. Marchante, M. Courty, J.-M. Tarascon and M. R. Palacín, *Energy Environ. Sci.*, 2012, **5**, 8572-8583.
23. A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn and D. A. Dobchev, *Chem. Rev.*, 2010, **110**, 5714-5789.
24. J. Liu, Ph. D., Brigham Young University, 2011.
25. M. O. Iwunze, *Phys. Chem. Liq.*, 2005, **43**, 195-203.
26. E. Hladký, M. Kučera and K. Majerová, *Polymer*, 1966, **7**, 587-594.
27. C.-J. Hsieh, J.-M. Chen and M.-H. Li, *J. Chem. Eng. Data*, 2007, **52**, 619-623.
28. B. Maribo-Mogensen, G. M. Kontogeorgis and K. Thomsen, *J. Phys. Chem. B*, 2013, **117**, 10523-10533.
29. J. Tomasi, B. Mennucci and C. Capelli, in *Handbook of Solvents*, ed. G. Wypich, ChemTec Publishing, Toronto, 2001, ch. 8, pp. 419-504.
30. J. N. Wilson, *Chem. Rev.*, 1939, **25**, 377-406.
31. W. M. Latimer, *Chem. Rev.*, 1949, **44**, 59-67.
32. G. G. Raju, in *Dielectrics in electric fields*, Marcel Dekker, Inc., New York, 2003, pp. 35-95.
33. M. Neumann, *Mol. Phys.*, 1983, **50**, 841-858.
34. S. Sild and M. Karelson, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 360-367.
35. P. G. R. Achary, *SAR QSAR Environ. Res.*, 2014, **25**, 507-526.
36. A. Liu, X. Wang, L. Wang, H. Wang and H. Wang, *Eur. Polym. J.*, 2007, **43**, 989-995.
37. J. Richardi, P. H. Fries and C. Millot, *J. Mol. Liq.*, 2005, **117**, 3-16.
38. *Journal*.
39. T. R. Griffiths and D. C. Pugh, *Coord. Chem. Rev.*, 1979, **29**, 129-211.
40. T. Gaudin, G. Fayet, P. Rotureau and I. Pezron, *Journal of Surfactants and Detergents*, 2018, **21**, 835-843.
41. C. C. Pye, T. Ziegler, E. van Lenthe and J. N. Louwen, *Canadian Journal of Chemistry*, 2009, **87**, 790-797.
42. M. P. Andersson, F. Eckert, J. Reinisch and A. Klamt, *Fluid Phase Equilib.*, 2017, DOI: https://doi.org/10.1016/j.fluid.2017.06.005.
43. A. Dequidt, J. Devémy and A. A. H. Pádua, *J. Chem. Inf. Model.*, 2016, **56**, 260-268.
44. C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa and D. van der Spoel, *Journal of Chemical Theory and Computation*, 2012, **8**, 61-74.
45. C.-Y. Yu and Z.-Z. Yang, *J. Phys. Chem. A*, 2011, **115**, 2615-2626.
46. V. P. Pawar, A. R. Patil and S. C. Mehrotra, *J. Mol. Liq.*, 2005, **121**, 88-93.
47. B. Sadhu, M. Sundararajan and T. Bandyopadhyay, *Inorg. Chem.*, 2016, **55**, 598-609.
48. I. Nakamura, C. J. Shock, L. Eggart and T. Gao, *Isr. J. Chem.*, 2018, **0**.
49. I. Nakamura, *Soft Matter*, 2014, **10**, 9596-9600.

Footnote

† Electronic supplementary information (ESI) available: derivations of formula presented in the Theory section, numerical treatment of contacts and dimers from molecular surfaces, calculation of normal vectors, experimental and calculated relative permittivity values and Griffith's classes for all studied systems.

DOI: 10.1039/C9CP02358F