Article

Improving the Interpretation of Small Molecule Diffusion Coefficients

Robert Evans, Guilherme Dal Poggetto, Mathias Nilsson, and Gareth A. Morris

Just Accepted

Improving the Interpretation of Small Molecule Diffusion Coefficients

Robert Evans*[a], Guilherme Dal Poggetto[b], Mathias Nilsson[b] and Gareth A. Morris[b]

a. Aston Institute of Materials Research, School of Engineering and Applied Science, Aston University, Birmingham, B4 7ET, UK.

b. School of Chemistry, University of Manchester, Manchester M13 9PL, UK.

Diffusion-ordered NMR spectroscopy (DOSY) is increasingly widely used for the analysis of mixtures by NMR spectroscopy, dispersing the signals of different species according to their diffusion coefficients. DOSY is used primarily to distinguish between the signals of different species, with the interpretation of the diffusion coefficients observed usually being purely qualitative, for example to deduce whether one species is bigger or smaller than another. In principle the actual values of diffusion coefficient obtained carry important information about the sizes of different species and on interactions between species, but the relationship between diffusion coefficient and molecular mass is in general a very complex one. Here a recently-proposed analytical relationship between diffusion coefficient and molecular mass for the restricted case of small organic molecules is tested against a wide range of data from the scientific literature, and generalised to cover a range of solvents and temperatures.

Molecular self-diffusion in a liquid originates from the random, thermal motion of the molecules present. Chemical information, such as the molecular mass of an unknown species, its aggregation, or its association with other species, can in principle be obtained from measurements of diffusion coefficients. However in diffusion-ordered NMR spectroscopy (DOSY), an analytical tool that disperses NMR signals according to diffusion coefficient, such data are normally only interpreted qualitatively. Quantitative interpretation of such spectra in terms of the sizes of species gives insight into molecular weights and association, including the formation of dimers, trimers, and higher oligomers. However, while there is a rough inverse correlation between molecular mass and the speed at which a species moves through solution, the wide range of possible molecular shapes, mobilities and solute-solvent interactions, and some fundamental problems with theories of diffusion, make quantitative interpretation difficult.

One way forward is to use more or less empirical correlations to determine relationships between diffusion coefficient and molecular mass for chemically cognate systems, for example homologous series in a particular solvent. This has been very successful, particularly in organometallic chemistry where diffusion NMR has been used to identify organolithium reactive intermediates[1] and alkali metal cyclopentadienides[2], but is of little use when dealing with the analysis of a mixture of unknown species. The complementary approach that is examined here is to try to find a less accurate but much more general relationship that can be used without prior knowledge of the chemistry involved. If the range of chemical space to be spanned is limited to small to medium-sized organic molecules (< *ca.* 1 kDa), much of the uncertainty introduced by molecular flexibility and composition is avoided. It has been shown[3] that, given these limitations, it becomes possible to derive a general correlation between molecular mass and diffusion coefficient that works over a wide range of chemistry and in multiple solvents. The method is restricted to dilute solutions (or pure solvents), because of the uncertainties introduced by obstruction effects. Here we test that method, which was originally based on a set of 109 different combinations of 44 different solutes in 5 different deuteriated solvents, against a wide range of diffusion measurements drawn from the chemical literature, extend it to provide diffusion-molecular mass relationships for a range of temperatures, and provide software tools in Excel and Matlab for estimating diffusion coefficients from molecular masses and *vice versa*.

Any discussion of the relationship between diffusion and molecular size starts with the Stokes-Einstein equation[4] (Eqn. 1). This equation assumes that the solute acts as a hard sphere with hydrodynamic radius, r_H, moving randomly through a continuum solution in response to random buffeting from the species around it. The thermal driving force at a temperature T, $k_B T$, is balanced by the frictional force $6 f \pi \eta r_H$ that acts on a solute particle of effective radius r_H moving through a fluid with viscosity η, giving the relationship

$$D = \frac{k_B T}{6 \pi f \eta r_H} \qquad (1)$$

where the shape correction factor f is equal to 1 for a spherical particle. While the Stokes-Einstein equation is simple and can give good estimates for the diffusion coefficients of large species, it is not a good guide to small molecule diffusion. There are systematic failures in the model as real molecules are not hard spheres and no solution is a continuum. As a result, attempts to predict diffusion coefficients using the Stokes-Einstein equation, with hydrodynamic radii estimated from densities, tend to severely underestimate the diffusion coefficients of molecules smaller than *ca.* 1 kDa[3].

There are clear systematic trends in the ways in which the Stokes-Einstein equation fails in dilute solution. Four main reasons for its failure can be identified: the continuum approximation fails, as real solvents are made of discrete molecules that are not negligibly small compared to solutes; molecules are not spherical, and can adopt a wide range of possible geometries; molecules are not static, and have widely differing ranges of conformational freedom; and molecules can interact, and bind, with the solvent itself. A further obstacle to establishing a relationship between diffusion coefficient and molecular mass is that species containing heavy atoms will on balance be denser, and hence show smaller hydrodynamic radii, than those containing only light atoms. Any attempt to generate an accurate universal law would need to take all of these complications into account and is likely to be doomed to failure: molecules are simply too varied.

However, analytical theories do exist that can successfully account for some of the sources of failure. First of all, molecules are not spheres. Perrin[5] analysed the effects of molecular shape, considering molecules as ellipsoids, and analytical equations do exist for the effect of increasing aspect ratios in ellipsoidal shapes. For molecules that are not long thin rods or wide thin disks, the effects are typically much less than 10 % and can often be safely ignored.

More significantly, a liquid is not a continuum fluid. It consists of small molecules moving randomly, tumbling as they collide with other molecules in the liquid. The effect of non-negligible solvent particle size is to change the friction term acting on the solute molecules. This can be represented as a change in the correction factor f, and was modelled by Gierer and Wirtz[6] as Eqn. 2.a

$$f_{GW} = \left(\frac{3\alpha}{2} + \frac{1}{1+\alpha}\right)^{-1} \qquad (2.a)$$

where α is the ratio of the radius of the solute to that of the solvent. Chen and Chen[7] adapted this, using an empirical expression (Eqn. 2.b) fitted to experimental data.

$$f_{Chen} = (1 + 0.695\alpha^{2.234})^{-1} \qquad (2.b)$$

This approach is effectively a hybrid between the Gierer and Wirtz model (Eqn. 2.a) and power-law fitting, using two parameters. Using a test set of crown ethers in methanol, a nearly homologous series of structurally very similar molecules, it gave good results. The choice of crown ethers, which are highly flexible and hence are expected to show inflated hydrodynamic radii[8], was perhaps an unfortunate one, and their attempts to extend the study to crown ethers and alkanes in ethanol and butan-1-ol found the model lacking.

The general approach of using empirically-obtained power laws has found wide use in the study of macromolecules, in particular of polymers[9] and peptides[10]. The diffusion coefficient or, often, the relative diffusivity (the measured diffusion coefficient normalised by an internal reference) can be expressed in terms of the mass of the species, M, raised to some empirical power, α, as in Equation 3.

$$D \propto M^{-\alpha} \qquad (3)$$

This use of a power law is reminiscent of the Flory equation[8] where the radius of gyration of a polymer, R_g, is related to its mass through a parameter, δ, known as the Flory exponent or the inverse of the fractal dimension, d_F, of the molecular chain.

This gives two potential general methods for the estimation of small molecule diffusion coefficients. One is specific, producing power-law relationships for homologous series of similar species. This is well-established for polymers,[9,11-14], proteins,[10,11,15] and large macrocycles,[16] although a two-parameter fit is required for each structurally distinct polymer type in each different solvent. While there have been attempts to link the exponent in Eqn. 3 to the fractal dimension contained within the Flory equation, these have not always proved successful. For spherical species, the fractal dimension is 3 and $\alpha = 1/3$. In one of the first attempts to relate diffusion coefficients to molecular size, Polson[17] found that this cube root relationship held for some large species, but was much less useful for smaller molecules.

With the use of internal and external references, methods based on power-law relationships have also found application in small molecules, for example in the work of Williard and Li[18] on the characterisation of organolithium compounds. Diffusion measurements of unknown organolithium compounds were acquired in the presence of a number of internal references. An internal calibration curve based on diffusion measurements of reference compounds within the sample is then produced for each experiment. The method was subsequently extended to include ^2H, ^6Li and ^{31}P DOSY experiments.[19-21] The internal references used in these studies were chosen for a number of properties including inertness and high solubility, and their chemical shifts were typically far from those of the lithium-containing species studied. These compounds spanned a range of different chemical functionalities including, but not limited to, aromatic compounds, olefins and TMS. It is not always possible to find suitable internal reference materials, so a similar approach used an external calibration curve to rationalise relative diffusion coefficients D/D_{ref}.[22] The effect of shape was considered, with different calibration curves used for three distinct classes of shapes: spheres, "dissipated spheres", and ellipsoids and discs. This method has now been extended to a range of common NMR solvents (DMSO-d_6, C_6D_{12}, C_6D_6, CDCl$_3$, and CD$_2$Cl$_2$).[23] Shape-specific calibration curves have been shown to be very accurate for chemically cognate species; a degree of generality can be introduced, at the expense of accuracy, with the use of 'merged' calibration curves.

The approach investigated here is both more general, and necessarily more approximate. Starting again with the Stokes-Einstein equation, the largest source of error that can be treated analytically is the breakdown in the continuum model. Here, the Gierer-Wirtz equation for the correction factor f (Eqn. 2.a) is used. This leaves the effects of shape, flexibility, solvation and composition, none of which can be treated analytically without prior chemical knowledge. A model is therefore sought that approximates species as hard spheres, with an average effective density ρ_{eff} that is a variable parameter, that obey the Gierer-Wirtz modification of the Stokes-Einstein law. Restricting the scope of the approach to small to medium-

2

sized (up to *ca*. 1 kDa) organic molecules with no heavy atoms limits the impact of composition (excluding atoms heavier than chlorine, putting an upper limit on the effective density) and of flexibility (excluding flexible high polymers, which can adopt extended conformations with inflated hydrodynamic radii[8]). The residual effects of flexibility, and the effects of shape and of solvation, all tend to increase hydrodynamic radius, and hence to decrease effective density; different compositions can increase or decrease the density, but only to a limited extent. The Gierer-Wirtz equation requires knowledge of the ratio of the solute and the solvent radius, but since the solute radius is being estimated using the hard sphere approximation with an effective density, it is reasonable to apply the same logic to estimating the solvent radius. (Of course the same limitations then apply to the solvent as to the solute – systematic bias will result if the solvent is, for example, a flexible high polymer). This then yields the expression in Eqn. 4, which links the diffusion coefficient for a given temperature and solvent viscosity to the solute and solvent molecular masses MW and MW_S, through a single adjustable parameter, the effective density ρ_{eff}.

$$D = \frac{k_B T \left(\frac{3\alpha}{2} + \frac{1}{1+\alpha} \right)}{6 \pi \eta \sqrt[3]{\frac{3MW}{4 \pi \rho_{eff} N_A}}}$$

$$\alpha = \frac{r_S}{r} = \sqrt[3]{\frac{MW_S}{MW}} \tag{4}$$

The problem of deriving a usable relationship between molecular mass and diffusion coefficient then reduces to that of finding an optimum value for the effective density ρ_{eff}. Using a set of experimental diffusion coefficients, D, all measured at 25 °C, for 109 combinations of 44 solutes and 5 common solvents, gave an effective density of 619 kg m^{-3}.[3] As expected, this is much smaller than the typical densities of solid or liquid organic materials, because of the effects (in roughly decreasing order of importance) of solvation, flexibility and shape. The root-mean-square difference between estimated and experimental diffusion coefficients for the set of measurements was 14.6 %. The model has subsequently found widespread use, in a range of different areas[24-30].

Here the Stokes-Einstein Gierer-Wirtz Estimation (SEGWE) method is tested against a range of measurements on dilute systems drawn from the literature that span a much wider chemical space than the original data set (formally, the training set), increasing the range of compound masses up to *ca*. 1.5 kDa, removing the temperature restriction of 25 °C, and widening the range of solvents covered. The motivation for deriving Eqn. 4 was to facilitate the quantitative interpretation of results obtained with diffusion-ordered NMR spectroscopy (DOSY),[31,32], but diffusion coefficients can of course be determined using a variety techniques, for example tracer measurements[33,34] and chronoamperometry.[35,36] The diffusion data studied here are not limited to those measured by NMR in deuteriated solvents, but include a range of protiated liquids, widening the range of applicability of SEGWE. The estima-

tion method has been implemented as an Excel spreadsheet and as a Matlab package, as detailed in SI.1; both are available for free download from http://nmr.chemistry.manchester.ac.uk/ and also at doi:10.17632/cxt99xf2d2.2.

METHOD

Diffusion Coefficients. In order to test SEGWE against a wider range of samples, containing different solvents and solutes, a comprehensive review of literature for a wide range of diffusion coefficients from a wide range of sources was undertaken. In addition to the original data set in the paper detailing the method,[3] a further 23 papers have been used,[14,18,21-23,25,26,28,37-51] to create a literature data set containing a further 558 measurements of diffusion coefficients of various species in various solvents, spanning a range of molecular weights from 18 to 1273 g mol^{-1} in 23 different solvents, both deuteriated and protiated. The single biggest source of additional data is that of Crutchfield and Harris[37], with 200 diffusion coefficients in two solvents at two different temperatures. A further five papers[41,44,45,47,48] contained variable temperature data, contributing 86 measurements between them. The rest of data were acquired at 25 °C. While most of the diffusion data were acquired using NMR techniques, data acquired where tracer diffusion methods were used as a calibrating standard in two references,[47,49] while a third reference[51] used an electrochemical method.

Not all possible sources of diffusion data were used in this study. A set of five criteria were used to determine whether literature diffusion coefficients were to be used in this study.

Scope. Solutes with molecular weights greater than 1.5 kDa were not used. Likewise, long flexible polymeric species, which tend to adopt extended conformations with inflated hydrodynamic radii, were excluded from the literature data set.

Systematic miscalibration. There are a number of possible sources of systematic errors in diffusion NMR experiments.[44] If the relevant paper contained evidence of lack of, or miscalibration, the data were excluded from the literature data set. In particular, sets of experimentally acquired diffusion coefficients with systematically large deviations from prediction in every measurement were excluded.

Inconsistent diffusion coefficients. Related to criterion 2, if a repeated measurement of the same species had inconsistent diffusion coefficients reported within a single experimental report, it is highly likely that there were significant problems with the experiments. The data were excluded from the literature data set.

Evidence of convection. The presence of convection in a sample is likely to lead to higher diffusion coefficients than expected.[52] Variable temperature studies, or those in solvents particularly prone to convection, that showed systematically higher diffusion coefficients than predicted were excluded from the literature data set.

Evidence of aggregation. Species that form aggregates diffuse more slowly than single molecules. Systems that might be expected to aggregate and that had lower diffusion coefficients than predicted were therefore excluded from the literature data set.

It is instructive to look at two diffusion measurements that failed these tests. In the original data set, [3] trimesic acid (benzene-1, 3, 5-tricarboxylic acid) in DMSO-d_6, exhibited a diffusion coefficient *ca.* 60 % smaller than predicted. Trimesic acid is known to form extended self-assembled structures in both the solid state[53] and at liquid-solid interfaces,[54] producing extended hexagonal networks with either 'chicken wire' or 'flower' structures. This diffusion coefficient was, therefore, removed from the original data set.

Experimental diffusion coefficients much higher than predicted indicate the likely presence of convection. Reference 45 contains a variable temperature study of water, 2-ethoxyethanol and caffeine in D_2O.[45] At 304.4 K, the measured diffusion coefficients for water and caffeine were respectively 22 and 23 % higher than predicted. This increased to over 30 % for all three species at 309.8 K. This indicates the probable onset of convection, and hence the diffusion coefficient measurements in this reference made at higher temperatures were all removed from the literature data set.

The references used in this study (as well as the solvents used for the samples, the nature of the experiments (variable temperature, non-NMR) and whether any measurements of diffusion within a given reference were excluded) are summarised in SI.2. The diffusion and solvent viscosity data collected from the literature and carefully organized here, although not covering all possible reported data on diffusion coefficients, will aid future investigations on the prediction of molecular weight from diffusion coefficients.

Temperature Dependence of Diffusion Coefficients. The original investigation used only data acquired at 25 °C. There are two ways in which temperature affects diffusion coefficients. First, in the numerator of the Stokes-Einstein equation, the thermal energy driving diffusion is given by $k_B T$. The higher the temperature, the more energy the solute and solvent molecules have and the faster they move through the solution. Second, and much more importantly, in the denominator, the solvent viscosity η depends strongly on temperature. Over the range of temperatures likely to be encountered in NMR diffusion measurements this temperature dependence is well represented by an Arrhenius-like equation (Eqn. 5), with two variable parameters a and b.

$$\eta = a\, e^{\frac{b}{T}} \qquad (5)$$

In the original study, the viscosities used for the deuteriated solvents were estimated from those of protiated solvents (by multiplying the value for the protio- solvent by the ratio of the deuterio- and protio- molecular masses), at a single temperature (298 K). Here, literature data are used for the viscosities of all solvents, whether protiated or deuteriated. Reported values at different temperatures are fitted to obtain Arrhenius parameters a and b, from which viscosities at different temperatures are calculated. Of all the solvents used, only one had no reported measurements of its viscosity at temperatures away from 298 K. Frustratingly this was the solvent for which the largest number of diffusion measurements are available, CDCl$_3$. In order to estimate the temperature dependence of the

viscosity of CDCl$_3$, the value of b for CHCl$_3$ was used, with the value of a determined by the estimated viscosity of CDCl$_3$ at 298 K, taken from previous work[3] (5.44×10^{-4} kg m^{-1} s^{-1}). The experimental viscosity data available for toluene-d_8 are also limited, with only a few low-temperature measurements in the literature. Literature data for all the solvent viscosities differ slightly from the estimates used previously; however, using the new, temperature dependent, values for viscosity has little effect on the root mean square deviation for the original data set. A comparison of the viscosities at 298 K used in the original study with those calculated using the literature data and Eqn. 5 is included in SI.3. The references used for both deuteriated[55-60] and protiated[55,60-83] solvents are detailed in SI.4. Table SI.4 summarises relevant information for each of the solvents used. Protiated solvents which had a deuteriated analogue used in the study are also included. SI.4 also contains all data used to calculate a and b for each solvent used in this work.

RESULTS

Testing SEGWE against the Literature. Figure 1 shows the results of plotting experimental *versus* predicted diffusion coefficients for the set of 558 literature measurements of diffusion coefficients used, as blue diamonds, alongside the same comparison for 108 of the original measurements, plotted as red squares. (As noted above, the original data set had the outlier point, trimesic acid in DMSO-d_6, removed, the effective density has been re-optimised, to give a new value of 627 kg m^{-3}).
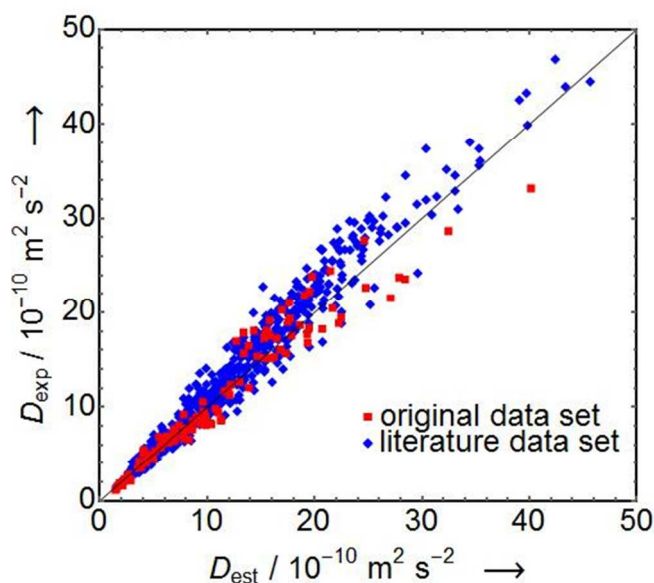

Figure 1. Measured diffusion coefficient plotted against diffusion coefficient calculated using Eqn. 4 for 108 samples of 44 small molecules in five deuteriated solvents (original data set, filled red squares) and 558 samples in 23 solvents, both deuteriated and protiated (literature data set, filled blue diamonds), with a solid line of unit slope.

4

With the approximations made, the model cannot hope to predict the diffusion coefficients of all the compounds used in this study, but it performs well over 600 individual experimentally acquired diffusion coefficients.

Gratifyingly, there is no immediately apparent difference between the original data set and the larger literature data set: the latter provides a strong validation of the method. The quality of fit can be represented and quantified in a number of ways. Plots of difference in diffusion coefficient, as a percentage of the calculated diffusion coefficient, against both solute molecular weight (Fig. 2.upper) and measured solute diffusion coefficient (Fig 2.lower) give an immediate graphical indication.
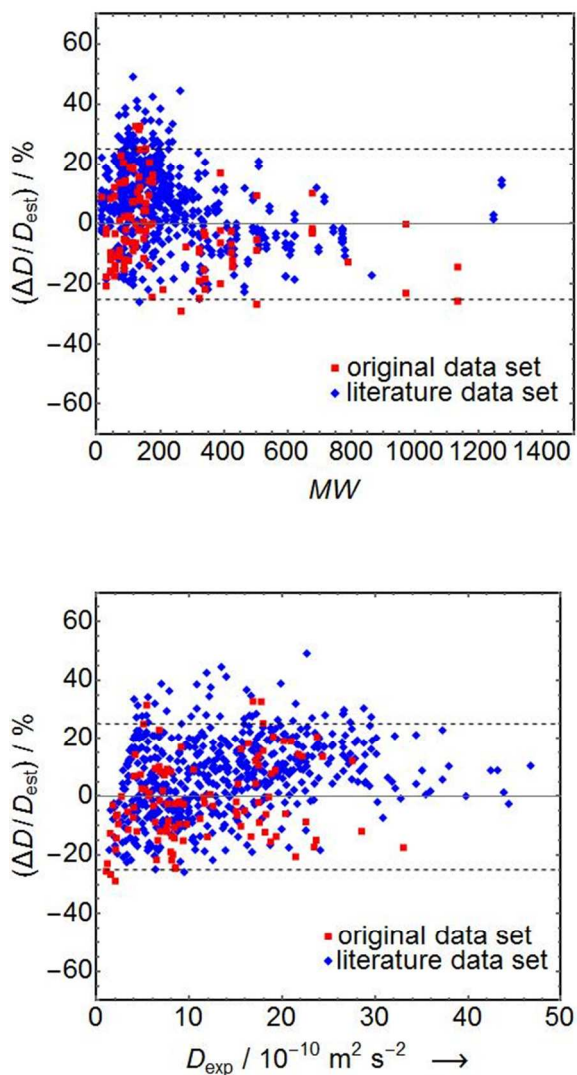

Figure 2. Difference between calculated and measured diffusion coefficients, expressed as a percentage of calculated diffusion coefficient, plotted against (upper) molecular mass and (lower) experimental diffusion coefficient for both the original data set (filled red squares) and the literature data set (filled blue diamonds). Dashed lines indicate errors of $\pm$ 25 %.

Data Set	# diffusion coefficients	RMS error (%)
Original	108	14.0
Literature	558	14.9
Combined	666	14.8

Table 1. Comparison of RMS error for original, literature and combined data sets.

For this large set of experimentally measured diffusion coefficients, the vast majority lie within 25 % of the value predicted. Figure 2 shows that there is a slight trend present within the literature data set, with the experimentally measured diffusion coefficients being on average slightly higher than expected, particularly for lighter molecules and/or higher diffusion coefficients. The diffusion coefficient measurements lying above and below the dashed lines are listed in SI. 5.

The RMS error for SEGWE applied to the larger, literature data set can be calculated and compared with that obtained for the original data set, as shown in Table 1. When all of the valid data presented here are included, the RMS error is 14.8 %, comparable with that originally reported for the original set of data.

It is instructive to compare histograms, shown in Figure SI.6, of the original and literature data that summarise the information contained in Figures 1 and 2. They show substantial overlap, but with a clear tendency for the literature experimental diffusion coefficients to be slightly (between 10 and 20 %) underestimated by Eqn. 4. While the original data set was acquired with strict temperature control and samples acquired in thick walled tubes, such precautions were not always taken in the literature measurements. Convection is conventionally seen as an example of a critical phenomenon. NMR tubes are long, narrow tubes made out of thin glass. If a large enough negative temperature gradient forms between two ends of the tube, then Rayleigh-Bernard convection will spontaneously form, with the warmer fluid flowing upwards, displacing the colder fluid above. However, recent studies of convective flow in NMR experiments have revealed that some convective flow is almost always present in typical diffusion NMR experiments. The non-critical phenomena of Hadley convection occurs when transverse temperature gradients form around the tube. In a temperature-regulated NMR probe, the air flow around the sample is disrupted and transient transverse temperature gradients form, making convective flow of the sample, and higher than expected diffusion coefficients, almost certain.[52,84]

Certain subsets of the data can be studied to obtain further insight into the performance of the estimation method over a range of different experimental parameters. Of the eight deuteriated solvents used across the literature data, seven have sufficient diffusion coefficients reported to make plotting the individual fits worthwhile. Figures SI.7.a to SI.7.h plot the experimental diffusion coefficients acquired in eight common deuteriated solvents against diffusion coefficients estimated using Eqn. 4. It is noteworthy that the solvent for which few viscosity measurements are available in the literature, toluene-

5

d_8, shows the most obvious systematic deviation between experimental and estimated diffusion coefficients. Other solvents show less marked trends, but overall it is perhaps surprising how well the single compromise approach of Eqn. 4 represents diffusion measurements in a wide range of very different solvents. The RMS errors for the four subsets of diffusion data corresponding to common deuterated NMR solvents are collected in Table 2.

Solvent	# diffusion coefficients	RMS error (%)
CDCl3	213	12.4
D2O	107	11.1
DMSO-d6	60	17.8
Toluene-d8	54	16.2

Table 2. Comparison of RMS error for subsets of the combined data set corresponding to the common NMR solvents CDCl₃, D₂O, DMSO-d₆ and toluene-d₈. See also Figures SI.7.

Extending SEGWE to Different Temperatures. In order to estimate diffusion coefficients at different temperatures, the Arrhenius model of Eqn. 5 for solvent viscosity has been used with Eqn. 4. As explained earlier, two parameters were used to fit viscosities as an exponential function of temperature for all the solvents used in this study. An assessment of the effectiveness of this extension was made using the data contained in reference 37, a compilation of 200 diffusion coefficients measured in two different solvents at two different temperatures. Figure 3 compares measured diffusion coefficients with those estimated with Eqn. 4 for both solvents and both temperatures.
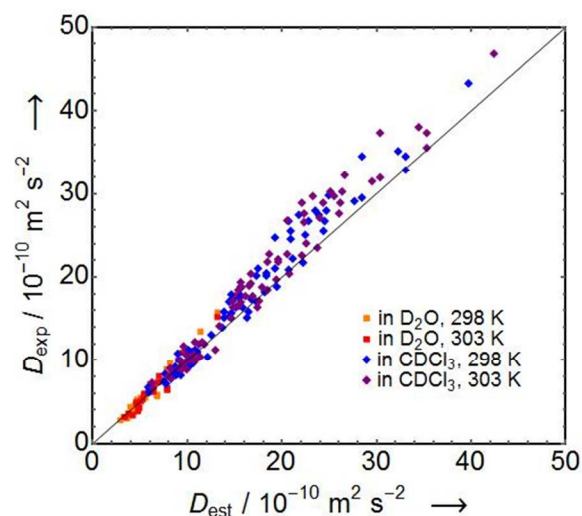

Figure 3. Measured diffusion coefficient plotted against diffusion coefficient calculated using Eqn. 4 for 24 small molecules in D₂O at 298 K (filled yellow squares), 24 small molecules in D₂O at 303 K (filled red squares), 76 small molecules in CDCl₃ at 298 K (filled blue diamonds) and 76 small molecules

in CDCl₃ at 303 K (filled purple diamonds). Data are drawn from reference 37.

As with the data in Figure 1, the quality of the fit can be represented and quantified in a number of ways. SI.8 depicts analogous plots to those in Fig.2 , plotting ΔD (as a percentage of experimental diffusion coefficient) against both solute diffusion coefficient (Fig SI.8.a) and solute molecular weight (Fig. SI.8.b) give an immediate graphical indication of the quality of the fit and also show how little the scatter in results changes with increasing temperature. While almost all of the experimental data are correctly predicted to within 25 %, there is a slight upwards trend in the average deviation between estimation and experiment as diffusion coefficient increases, just as in Figure 2.

DISCUSSION

The small but clear systematic trends in Figures 2 and SI.8 reflect the difficulty of using a single model, with only one adjustable parameter, for all solvents. One source of such a trend is clear from Figures SI.7: in CDCl₃, toluene-d₈, benzene-d₆ and cyclohexane-d₁₂, there is a small tendency to overestimate D in small species. This is far from surprising: the same factors that lead to a *ca*. 15% RMS uncertainty in the diffusion coefficients of individual analytes, and by implication in the estimation of their hydrodynamic radii r, apply to the estimates of the hydrodynamic radii of the solvents. It might be expected that the overall fit of the data could be improved significantly by making the hydrodynamic radii of the different solvents into adjustable parameters. Such an approach can be found applied to the original data set in SI.9. While the quality of the fit is improved, the improvement is only from 14.0% to 12.9 %. This method sacrifices the simplicity and universality of the SEGWE approach and the improvement in reliability from increasing the number of adjustable parameters from one to six is small, because the approach does not address the other fundamental sources of uncertainty set out in the Introduction. It would require a much larger evidence base to give robust results.

A further reason for caution in seeking to improve the prediction accuracy of this, and of other methods for correlating molecular masses with diffusion coefficients, is, as noted earlier, that experimental measurements of the latter are subject to a source of systematic error, sample convection.[52,84] This has historically been underestimated. There are a number of ways in which the effects of convection can be reduced or compensated for. The thermal conductivity of sapphire is approximately 25 times that of borosilicate glass[85], so the use of sapphire NMR tubes will greatly reduce the temperature gradients that drive convection. Narrower bore NMR tube reduce the convective flow that does occur. Convection-compensated diffusion NMR pulse sequences are designed to cancel the effects of laminar flow in the sample[86]. However, many of the measurements in the literature data set did not use convection compensation. Since the effect of convection is to increase the apparent diffusion coefficient, there will be a small but unquantifiable bias in the great majority of the measurements used here (and, by extension, in many of the measurements that Eqn. 4 is likely to be called upon to assist in interpreting). As noted earlier, comparing the original and literature results in the histograms of Figure SI.6, there is a small but clear shift

6

between the two centroids that may be attributable to the slight bias introduced by convection. It is noteworthy, for example, that many of the measurements exhibiting higher diffusion coefficients than predicted in Figs. SI.8.a and SI.8.b are those made in chloroform. This is not unexpected; the relative ease with which a solvent convects is a function of the parameter $\chi = \beta \eta \kappa$, where β is the volumetric thermal expansion coefficient in K^{-1}, η is the dynamic viscosity in Pa s, and κ is the thermal conductivity in $W\ m^{-1}\ K^{-1}$.[52] For the two solvents compared in reference 37, χ is nearly 12.5 times larger for $CDCl_3$ than for D_2O.

Instead of using the literature data to validate the result of reference 3, it is tempting to take advantage of the much larger evidence base available from the literature to re-optimise the parameter ρ_{eff} in Eqn. 4. If done, this results in an increase in ρ_{eff} from 627 to 744 kg m^{-3}, but the RMS error decreases only slightly, from 14.8 to 12.9%. Any marginal benefit is outweighed by the extra uncertainty that is introduced by the fact that there is a known but unquantifiable bias in the literature data set caused by sample convection. The accompanying software therefore uses a default value 627 kg m^{-3} for ρ_{eff}, which is straightforward for the user to change if wished.

CONCLUSIONS

NMR measurements of diffusion coefficient, such as in diffusion-ordered spectroscopy (DOSY) experiments, are typically used in a qualitative manner only, separating out signals in a manner akin to chromatography. By making pragmatic decisions about the assumptions underlying the Stokes-Einstein equation, a general method for the prediction of small molecule diffusion coefficients can be constructed for extracting approximate information on molecular mass from such data. For a known molecular mass, a diffusion coefficient can be estimated (or vice-versa, albeit with much greater uncertainty). This, in turn, allows for deductions about whether molecules are associating, or oligomeric species are present, delivering chemical insight from measurements that are often left uninterpreted. The method has been tested against a large body of literature data on diffusion coefficients, showing a high level of consistency, has been extended to cover both a wide range of NMR solvents and a wider range of temperatures, and has been implemented in freely-available software, including the very recent General NMR Analysis Toolbox[87].

SUPPORTING INFORMATION

The Supporting Information is available free of charge on the ACS Publications website. This contains summaries of all references, solvents used and viscosity measurements, as well as figures illustrating subsets of data, as described in text. Further supporting material can be found at doi:10.17632/cxt99xf2d2.2. This contains Excel spreadsheets containing all diffusion coefficients used in this work and Mathematica notebooks to reproduce every calculation and every image in both manuscript and Supporting Information, all solvent data used, as well as Matlab GUI and Excel sheets for the prediction of diffusion coefficient from MW and *vice versa*.

AUTHOR INFORMATION

Corresponding Author

*Email: r.evans2@aston.ac.uk. Phone +44 121 204 5382

REFERENCES

(1) Li, D.; Keresztes, I.; Hopson, R.; Williard, P. G. *Acc. Chem. Res.* **2008**, *42*, 270–280.
(2) Bachmann, S.; Gernert, B.; Stalke, D. *Chem. Commun.* **2016**, *52*, 12861–12864.
(3) Evans, R.; Deng, Z.; Rogerson, A. K.; McLachlan, A. S.; Richards, J. J.; Nilsson, M.; Morris, G. A. *Angew. Chem. Int. Ed.* **2013**, *52*, 3199–3202.
(4) Einstein, A. *Ann. Phys.* (*Berl.*) **1905**, *17*, 549–560.
(5) Perrin, F. *J. Phys. Radium.* **1936**, *7*, 1–11.
(6) Gierer, A.; Wirtz, K. *Z. Naturforsch.* **1953**, *8*, 522–532.
(7) Chen, H. C.; Chen, S. H. *J. Phys. Chem.* **1984**, *88*, 5118–5121.
(8) Flory, P. J. *Principles of Polymer Chemistry*; 1st Ed, Cornell University Press: Ithaca, NY, **1953**.
(9) Chen, A.; Wu, D. H.; Johnson, C. S. *J. Am. Chem. Soc.* **1995**, *117*, 7965–7970.
(10) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. *J. Biochem.* **1999**, *38*, 16424–16431.
(11) Augé, S.; Schmit, P.-O.; Crutchfield, C. A.; Islam, M. T.; Harris, D. J.; Durand, E.; Clemancey, M.; Quoineaud, A.-A.; Lancelin, J.-M.; Prigent, Y.; Taulelle, F.; Delsuc, M.-A. *J. Phys. Chem. B* **2009**, *113*, 1914–1918.
(12) Assemat, O.; Coutouly, M.-A.; Hajjar, R.; Delsuc, M.-A. *C. R. Chim.* **2010**, *13*, 412–415.
(13) Viéville, J.; Tanty, M.; Delsuc, M. A. *J. Magn. Reson.* **2011**, *212*, 169–173.
(14) Arrabal-Campos, F. M.; Oña-Burgos, P.; Fernández, I. *Polym. Chem.* **2016**, *7*, 4326–4329.
(15) Jones, J. A.; Wilkins, D. K.; Smith, L. J.; Dobson, C. M. *J. Biomol. NMR* **1997**, *10*, 199–203.
(16) Bogdan, A. R.; Davies, N. L.; James, K. *Org. Biomol. Chem.* **2011**, *9*, 7727–7733.
(17) Polson, A. *J. Phys. Colloid Chem.* **1950**, *54*, 649–652.
(18) Li, D.; Kagan, G.; Hopson, R.; Williard, P. G. *J. Am. Chem. Soc.* **2009**, *131*, 5627–5634.
(19) Kagan, G.; Li, W.; Hopson, R.; Williard, P. G. *Org. Lett.* **2009**, *11*, 4818–4821.
(20) Kagan, G.; Li, W.; Hopson, R.; Williard, P. G. *Org. Lett.* **2010**, *12*, 520–523.
(21) Guang, J.; Hopson, R.; Williard, P. G. *J. Org. Chem.* **2015**, *80*, 9102–9107.
(22) Neufeld, R.; Stalke, D. *Chem. Sci.* **2015**, *6*, 3354–3364.
(23) Bachmann, S.; Neufeld, R.; Dzemski, M.; Stalke, D. *Chem. Eur. J.* **2016**, *22*, 8462–8465.
(24) Abet, V.; Evans, R.; Guibbal, F.; Caldarelli, S.; Rodriguez, R. *Angew. Chem. Int. Ed.* **2014**, *53*, 4862–4866.

7

(25) Poveda, A.; Alonso, I.; Fernández-Ibáñez, M. Á. *Chem. Sci.* **2014**, *5*, 3873–3882.

(26) Giuffrida, M. L.; Rizzarelli, E.; Tomaselli, G. A.; Satriano, C.; Trusso Sfrazzetto, G. *Chem. Commun.* **2014**, *50*, 9835–9838.

(27) Kennedy, S. R.; Miquelot, A.; Aguilar, J. A.; Steed, J. W. *Chem. Commun.* **2016**, *52*, 11846–11849.

(28) Maugeri, L.; Asencio-Hernández, J.; Lébl, T.; Cordes, D. B.; Slawin, A. M. Z.; Delsuc, M. A.; Philp, D. *Chem. Sci.* **2016**, *7*, 6422–6428.

(29) Claridge, T. D. W. *High-resolution NMR techniques in Organic Chemistry*; 3rd Ed, Elsevier: Amsterdam, **2016**.

(30) Rideau, E.; You, H.; Sidera, M.; Claridge, T. D.; Fletcher, S. P. *J. Am. Chem. Soc.* **2017**, *139*, 5614–5624.

(31) Johnson, C. S. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 203–256.

(32) "Diffusion-Ordered Spectroscopy" Morris, G. A. In *Encyclopedia of Nuclear Magnetic Resonance*, Grant, D. M.; Harris, R. K., Ed.; John Wiley & Sons Ltd.: Chichester, **2002**, pp 34–44.

(33) Mills, R. *J. Phys. Chem.* **1973**, *77*, 685–688.

(34) Evans, D. F.; Tominaga, T.; Davis, H. T. *J. Chem. Phys.* **1981**, *74*, 1298–1305.

(35) Denuault, G.; Mirkin, M. V.; Bard, A. J. *J. Electroanal. Chem. Interfacial. Electrochem.* **1991**, *308*, 27–38.

(36) Baur, J. E.; Wightman, R. M. *J. Electroanal. Chem. Interfacial. Electrochem.* **1991**, *305*, 73–81.

(37) Crutchfield, C. A.; Harris, D. J. *J. Magn. Reson.* **2007**, *185*, 179–182.

(38) Lucas, L. H.; Otto, W. H.; Larive, C. K. *J. Magn. Reson.* **2002**, *156*, 138–145.

(39) Barjat, H.; Morris, G. A.; Swanson, A. G. *J. Magn. Reson.* **1998**, *131*, 131–138.

(40) Nilsson, M.; Gil, A. M.; Delgadillo, I.; Morris, G. A. *Anal. Chem.* **2004**, *76*, 5418–5422.

(41) Cohen, Y.; Avram, L.; Frish, L. *Angew. Chem. Int. Ed.* **2005**, *44*, 520–554.

(42) Colbourne, A. A.; Meier, S.; Morris, G. A.; Nilsson, M. *Chem. Commun.* **2013**, *49*, 10510–10512.

(43) Nilsson, M.; Gil, A. M.; Delgadillo, I.; Morris, G. A. *Chem. Commun.* **2005**, 1737–1739.

(44) Antalek, B. *Concepts Magn. Reson.* **2002**, *14*, 225–258.

(45) Harmon, J.; Coffman, C.; Villarrial, S.; Chabolla, S.; Heisel, K. A.; Krishnan, V. V. *J. Chem. Educ.* **2012**, *89*, 780–783.

(46) Lin, M.; Shapiro, M. J. *J. Org. Chem.* **1996**, *61*, 7617–7619.

(47) Holz, M.; Heil, S. R.; Sacco, A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740–4742.

(48) Schulze, B. M.; Watkins, D. L.; Zhang, J.; Ghiviriga, I.; Castellano, R. K. *Org. Biomol. Chem.* **2014**, *12*, 7932–7936.

(49) Holz, M.; Weingärtner, H. *J. Magn. Reson.* **1991**, *92*, 115–125.

(50) Hoffman, R. E.; Shabtai, E.; Rabinovitz, M.; Iyer, V. S.; Mullen, K.; Rai, A. K.; Bayrd, E.; Scott, L. T. *J. Chem. Soc.*, *Perkin Trans. 2* **1998**, 1659–1664.

(51) Valencia, D. P.; González, F. J. *J. Electroanal. Chem.* **2012**, *681*, 121–126.

(52) Swan, I.; Reid, M.; Howe, P. W.; Connell, M. A.; Nilsson, M.; Moore, M. A.; Morris, G. A. *J. Magn. Reson.* **2015**, *252*, 120–129.

(53) Kolotuchin, S. V.; Thiessen, P. A.; Fenlon, E. E.; Wilson, S. R.; Loweth, C. J.; Zimmerman, S. C. *Chem. Eur. J.* **1999**, *5*, 2537–2547.

(54) Lackinger, M.; Griessl, S.; Heckl, W. A.; Hietschold, M.; Flynn, G. W. *Langmuir* **2005**, *21*, 4984–4988.

(55) Holz, M.; Mao, X. a.; Seiferling, D.; Sacco, A. *J. Chem. Phys.* **1996**, *104*, 669–679.

(56) Dixon, J. A.; Schiessler, R. W. *J. Phys. Chem.* **1954**, *58*, 430–432.

(57) Weingärtner, H.; Holz, M.; Sacco, A.; Trotta, M. *J. Chem. Phys.* **1989**, *91*, 2568–2574.

(58) Artaki, I.; Jonas, J. *J. Chem. Phys.* **1985**, *82*, 3360–3370.

(59) Harris, K. R.; Woolf, L. A. *J. Chem. Eng. Data* **2004**, *49*, 1064–1069.

(60) Hardy, R. C.; Cottington, R. L. *J. Res. Natl. Bureau Standards*, **1949**, 42, 573–578.

(61) Exarchos, N. C.; Tasioula-Margari, M.; Demetropoulos, I. N. *J. Chem. Eng. Data* **1995**, *40*, 567–571.

(62) Sovilj, M. N. *J. Chem. Eng. Data* **1995**, *40*, 1058–1061.

(63) Mekhtiev, S. I.; Mamedov, A. A.; Khalilov, S. K.; Aleskerov, M. A. *Izv. Vyssh. Uchebn. Zaved. Neft Gaz* **1975**, *3*, 64–100.

(64) Zakurenov, V. M.; Konyakhin, V. P.; Nozdrev, V. F. *Zh. Fiz. Khim.* **1975**, *49*, 548–549.

(65) Knapstad, B.; Skjølsvik, P. A.; Øye, H. A. *J. Chem. Eng. Data* **1989**, *34*, 37–43.

(66) Assael, M. J.; Dalaouti, N. K. *High Temp. - High Press.* **2000**, *32*, 179–184.

(67) Harris, K. R.; Newitt, P. J.; Woolf, L. A. *J. Chem. Eng. Data* **2004**, *49*, 138–142.

(68) Marchetti, A.; Preti, C.; Tagliazucchi, M.; Tassi, L.; Tosi, G. *J. Chem. Eng. Data* **1991**, *36*, 360–365.

(69) Bernal-García, J. M.; Guzmán-López, A.; Cabrales-Torres, A.; Estrada-Baltazar, A.; Iglesias-Silva, G. A. *J. Chem. Eng. Data* **2008**, *53*, 1024–1027.

(70) Akhtar, S.; Omar Faruk, A. N. M.; Saleh, M. A. *Phys. Chem. Liq.* **2001**, *39*, 383–399.

(71) Carmen Grande, M. d.; Juliá, J. A.; García, M.; Marschoff, C. M. *J. Chem. Thermodyn.* **2007**, *39*, 1049–1056.

(72) Geddes, J. A. *J. Am. Chem. Soc.* **1933**, *55*, 4832–4837.

(73) Caudwell, D. R.; Trusler, J. P. M.; Vesovic, V.; Wakeham, W. A. *Int. J. Thermophys.* **2004**, *25*, 1339–1352.

(74) Khattab, I. S.; Bandarkar, F.; Fakhree, M. A. A.; Jouyban, A. *Korean J. Chem. Eng.* **2012**, *29*, 812–817.

(75) Rathnam, M. V.; Bhanushali, K. R.; Sayed, R. T.; Kumar, M. S. S. *Eur. Chem. Bull.* **2013**, *2*, 434–444.

(76) Mikhail, S. Z.; Kimel, W. R. *J. Chem. Eng. Data* **1961**, *6*, 533–537.

(77) Domanska, U.; Laskowska, M. *J. Solution Chem.* **2009**, *38*, 779–799.

(78) Mutalik, V.; Manjeshwar, L. S.; Sairam, M.; Aminabhavi, T. M. *J. Chem. Thermodyn.* **2006**, *38*, 1062–1071.

(79) Dymond, J.; Øye, H. *J. Phys. Chem. Ref. Data* **1994**, *23*, 41–53.

(80) Carvajal, C.; Tölle, K. J.; Smid, J.; Szwarc, M. *J. Am. Chem. Soc.* **1965**, *87*, 5548–5553.

(81) Metz, D. J.; Glines, A. *J. Phys. Chem.* **1967**, *71*, 1158.

(82) Byers, C. H.; Williams, D. F. *J. Chem. Eng. Data* **1987**, *32*, 344–348.

(83) Santos, F. J. V.; Nieto de Castro, C. A.; Dymond, J. H.; Dalaouti, N. K.; Assael, M. J.; Nagashima, A. *J. Phys. Chem. Ref. Data* **2006**, *35*, 1–8.

(84) Barbosa, T. M.; Rittner, R.; Tormena, C. F.; Morris, G. A.; Nilsson, M. *RSC Adv.* **2016**, *6*, 95173–95176.

(85) Haynes, W. M. *CRC Handbook of Chemistry and Physics*; CRC press: Boca Raton, FL, 2014.

(86) Jerschow, A.; Muller, N. *J. Magn. Reson.* **1998**, *132*, 13-18.

(87) Castañar, L.; Dal Poggetto, G.; Colbourne, A.A.; Morris, G.A.; Nilsson, M. *Magn. Reson. Chem.*, in press. DOI: 10.1002/mrc.4717.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Estimation of Diffusion Coefficients

Stokes-Einstein
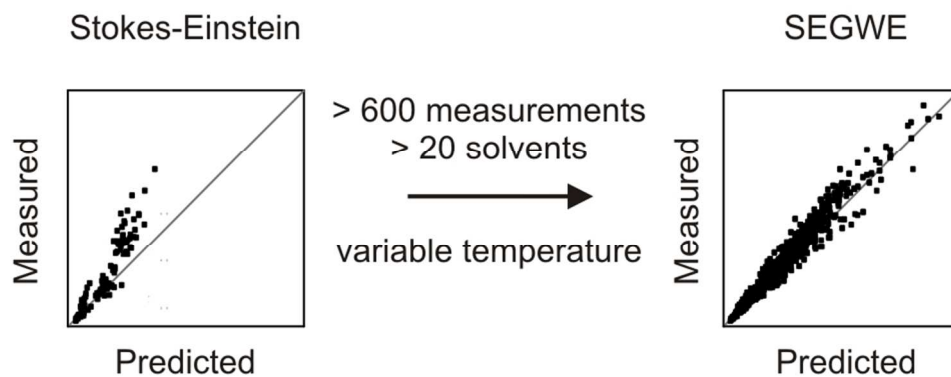
SEGWE

Measured

Predicted

> 600 measurements
> 20 solvents

→

variable temperature

Measured

Predicted

Table of Contents Graphic