JOURNAL OF CHEMICAL INFORMATION AND MODELING

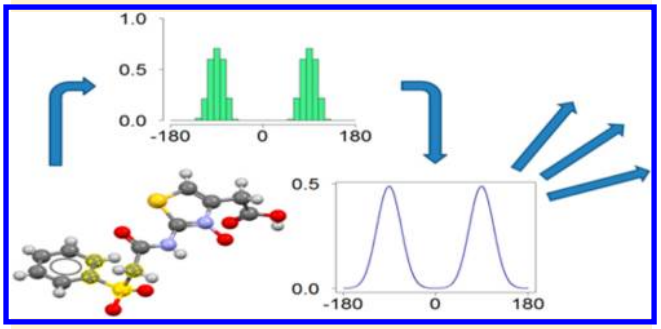
Letter

pubs.acs.org/jcim

Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions

Patrick McCabe,* Oliver Korb, and Jason Cole

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom



ABSTRACT: We describe the method of kernel density estimation (KDE) and apply it to molecular structure data. KDE is a quite general nonparametric statistical method suitable even for multimodal data. The method generates smooth probability density function (PDF) representations and finds application in diverse fields such as signal processing and econometrics. KDE appears to have been under-utilized as a method in molecular geometry analysis, chemo-informatics, and molecular structure optimization. The resulting probability densities have advantages over histograms and, importantly, are also suitable for gradient-based optimization. To illustrate KDE, we describe its application to chemical bond length, bond valence angle, and torsion angle distributions and show the ability of the method to model arbitrary torsion angle distributions.

INTRODUCTION

Given observations on structural chemistry data, typically, bond length, bond valence angle, and torsion angle data, it is common to present these data graphically as histograms. This is a straightforward but essential aspect of any exploratory data analysis. From this, the scientist can form some appreciation of the data distribution. Histograms however have well-known disadvantages as a representation, most particularly they are not smooth and the shape depends on both the width and end points of the bins. One approach to estimating the underlying distribution is to connect midpoints of adjacent histogram bins. However, this causes three problems: (i) Information will be discarded because all data within the bin is represented by the single midpoint. (ii) Only a piece-wise smooth function is generated, which again is nondifferentiable and a poor representation of the underlying smooth density. (iii) The technique is highly sensitive to the arbitrary bin width. Kernel density estimation was developed to overcome these problems.[1]

Parametric density estimators assume a fixed functional form for the distribution and rely on the appropriate distribution parameters such as mean and standard deviation. Nonparametric kernel density estimators in contrast do not assume a functional form for the data distribution and depend on every observed data point to reach an estimate. In addition, kernel density estimators do not bin data but instead center a kernel function at each data point thus removing the dependence on the end points of the bins. Using a smooth kernel function generates a smooth density estimate. Thus, two of the drawbacks, namely, nondifferentiability and bin starting point, of using histograms can be overcome using kernel density estimators. The problem of how wide, i.e., how smooth, to make the kernel functions does remain, but this will be discussed below.

With structural chemical data, we encounter two classes of data, i.e., linear and circular. Bond length and bond valence angle data (despite by definition being angular) can both be treated as linear data, whereas torsion angle data is fundamentally circular in nature, and the usual methods that apply to linear data are inappropriate in this case. A pitfall for the unwary is that simple linear descriptive statistics formulas, e.g., for the mean, should not be applied to circular data. The same is true in kernel density estimation, and kernels appropriate for linear data are not necessarily appropriate for circular data. Therefore, we describe the kernel function we have used for describing linear data and the kernel function we have used for circular data and describe how their parameters have been obtained.

Although other knowledge-based force fields exist in the literature for predicting protein structure or function and assessing protein−ligand complexes from docking calculations,[2−5] these do not employ kernel density estimates. With a view to molecular geometry optimization, the kernel density approach can provide objective functions based entirely on observations of structural data. In addition, the objective functions so generated are differentiable. Differentiability is extremely important as the efficiency of structural optimization can be greatly enhanced by the application of gradient-based optimization techniques, and a prerequisite for the gradient calculation is differentiable objective functions. Thus, such knowledge-based objective functions could in principle be used to optimize the geometry of molecules according to the data contained in chemical databases containing 3D information.

The kernel density technique has been used previously.[6,7] However, showing how the technique can be applied to different molecular structural parameters, such as bond lengths, bond valence angles, and torsion angles, and comparing against appropriate histograms was not the focus of that research. Also, some research has used Gaussian kernels for circular data, and

dx.doi.org/10.1021/ci500156d | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

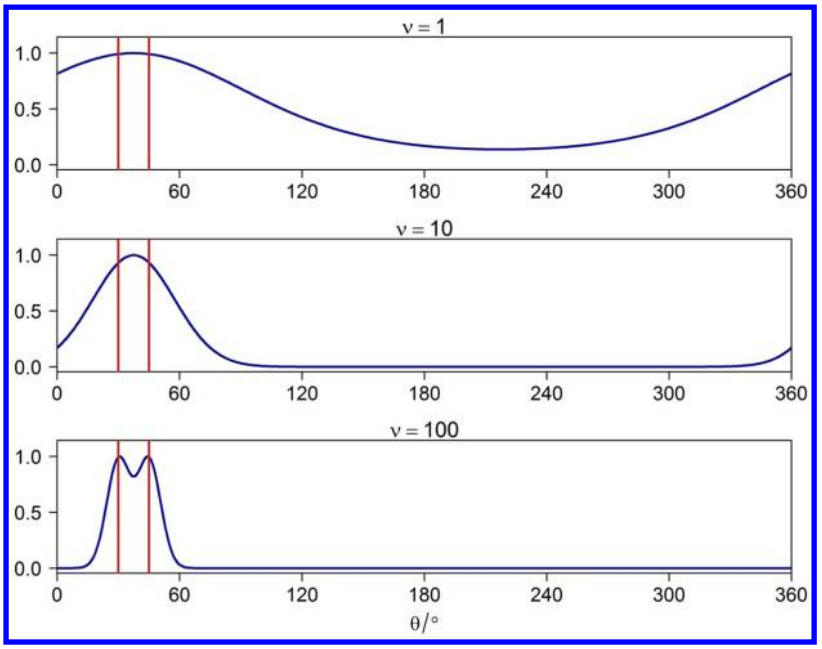


Figure 1. Variation of the von Mises kernel density estimated probability density function (dark blue lines) for a sample consisting of two observations at 30 ° and 45° (red vertical lines), with the value of ν. The top curve is for $\nu = 1$, which shows that this kernel density approximation estimates a substantial nonzero density across the entire angular range and represents a case of over-smoothing. The bottom curve is for $\nu = 100$, and this resolves the two nearby observations into two separate peaks, which represents under-smoothing. The middle curve for $\nu = 10$ produces a single relatively narrow smooth peak near 30° and 45° and generates very low probabilities of observations for larger deviations from these values, consistent with the given data.

the challenges arising from the circular nature of the data was overcome by using shifted copies of the data.[7] The focus of this work is showing how to apply the method itself. All relevant formulas are presented, both linear and circular data are addressed, without using shifted copies of the circular data, and suitable kernels described. Bandwidth estimates are provided clearly and concisely, and comparisons against the corresponding histograms for specific molecular structures are presented.

METHODS

In this work, the structural data is extracted from the Cambridge Structural Database.[8] Given a sample of size n of data (e.g., the data may be bond lengths observations for a particular bond over a range of structures) $x_1,\cdots,x_n$ from some unknown density $f(x)$, we make use of the nonparametric kernel density estimate

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_i(x) \tag{1}$$

where the kernel function $K_i(x)$ possesses the property

$$\int K_i(x)\mathrm{d}x = 1 \tag{2}$$

and the integral is over the entire domain of the variable x.

For linear data, we employ Gaussian kernel functions

$$K_i(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{h}e^{-1/2\left(\frac{x-x_i}{h}\right)^2} \tag{3}$$

with a smoothing (bandwidth) parameter h. A problem for nonparametric density estimation is making a good choice for the value of the smoothing parameter, which has both sound theoretical properties and is also good in practice. This is similar to choosing good values for the bin width when generating histograms. The computation of the bandwidth parameter for the Gaussian kernel follows the Silverman[1] rule-of-thumb with the modification of Scott[9]

$$h = 1.06\ \min\left\{\sigma, \frac{\mathrm{iqr}}{1.34}\right\} n^{-1/5} \tag{4}$$

In eq 4, σ is the standard deviation, and iqr is the interquartile range of the data.

For circular data, we employ the von Mises kernel[10−12]

$$K_i(\theta) = \frac{1}{2\pi I_0(\nu)}\exp\{\nu\ \cos(\theta - \theta_i)\} \tag{5}$$

where $I_r(\nu)$ is the modified Bessel function of the first kind of order r, and ν plays the role of the smoothing parameter. The von Mises distribution was introduced to study deviations of measured atomic weights from integral values.[10] This distribution is unimodal and symmetrical around the mode θ_i. The larger the value of ν is, the greater the clustering around the mode is, and for $\nu = 0$, the distribution reduces to the uniform distribution. The importance of the von Mises distribution as a standard distribution for circular data is comparable to that of the normal distribution for linear data, but it appears to have been seldom employed in the chemical literature.

For the smoothing parameter, in this work, we use the expression[12]

$$\nu = [3n\hat{\kappa}^2 I_2(2\hat{\kappa})\{4\pi^{1/2} I_0(\hat{\kappa})^2\}^{-1}]^{2/5} \tag{6}$$

In this eq 6, $\hat{\kappa}$ is the so-called concentration parameter, and $I_r(\hat{\kappa})$ is again the modified Bessel function of the first kind of order r. To calculate $\hat{\kappa}$, an estimate of the von Mises concentration parameter, we follow Fisher[13] and give the

dx.doi.org/10.1021/ci500156d | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

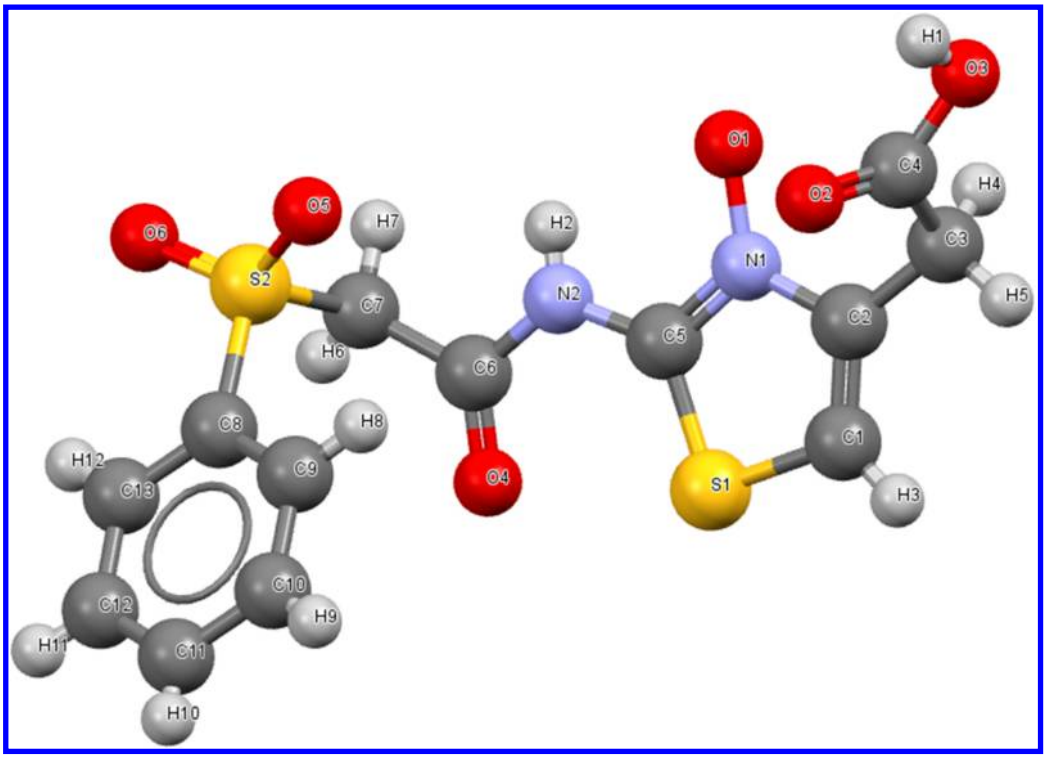


Figure 2. CSD entry (3-oxido-2-(((phenylsulfonyl) acetyl) amino)-1,3-thiazol-4-yl) acetic acid, EFOCUY.

most important formulas here. We calculate the mean resultant length $\bar{R}$ of the circular observations

$$\bar{R} = \frac{(C^2 + S^2)^{1/2}}{n} \tag{7}$$

where

$$C = \sum_{i=1}^{n} \cos(\theta_i) \tag{8}$$

$$S = \sum_{i=1}^{n} \sin(\theta_i) \tag{9}$$

and then

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R}^3 + 5\bar{R}^5/6 \\ -0.4 + 1.39\bar{R} + 0.43/(1 - \bar{R}) \\ (\bar{R}^3 - 4\bar{R}^2 + 3\bar{R})^{-1} \end{cases} \tag{10}$$

for

$$\bar{R} < 0.53$$
$$0.53 \le \bar{R} \le 0.85$$
$$\bar{R} > 0.85$$

respectively.

The modified Bessel functions $I_0(x)$ and $I_1(x)$ are computed using standard polynomial approximations,[14] and those for higher order, e.g., $I_2(x)$ can be calculated by recursion.[14]

To understand the behavior of the kernel density estimate of a probability density function with the von Mises kernel, in Figure 1 we show plots of a theoretical sample of two angles, namely, 30°and 45°. Note that both the value and also the slope of the kernel density estimates agree at 0° and 360° as they should for a circular distribution with period 360°.

RESULTS

The kernel density method has been applied to structural data for many structures in the CSD, but to illustrate typical results, we concentrate on the entry with CSD refcode EFOCUY,[15] Figure 2, and present four typical distributions from each of the main structural parameters of interest, namely, bond length, bond valence angles, and torsion angles.

True histograms in the sense of approximating the density, rather than the frequency of observations, and the kernel density approximation of the underlying density are presented in the figures.

In Figure 3, we show typical results for bond length distributions.

Figure 4 shows results for four bond valence angle distributions. In some instances, the underlying raw observations may not be available. This may arise for example when the number of observations are simply too numerous, and for unimodal distributions, a mean value and standard deviation for the parameter are supplied instead. In this case, the mean and standard deviation are used to generate a parametric normal probability density function. In other instances, there may be no mean value available. In this case, it is natural to fall back to the single value of the structural parameter of the input structure along with a heuristic value for the standard deviation, e.g., 0.05 Å for the bond lengths. This at first sight may appear quite large; however, it prevents the distributions becoming too narrow and too sharply peaked. This is important because it would not be reasonable to claim the single observed value represented the only allowed value that will ever be seen for that bond, i.e., sufficient data variation should be allowed for. Also, from the perspective of structural optimization, peaks that

dx.doi.org/10.1021/ci500156d | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

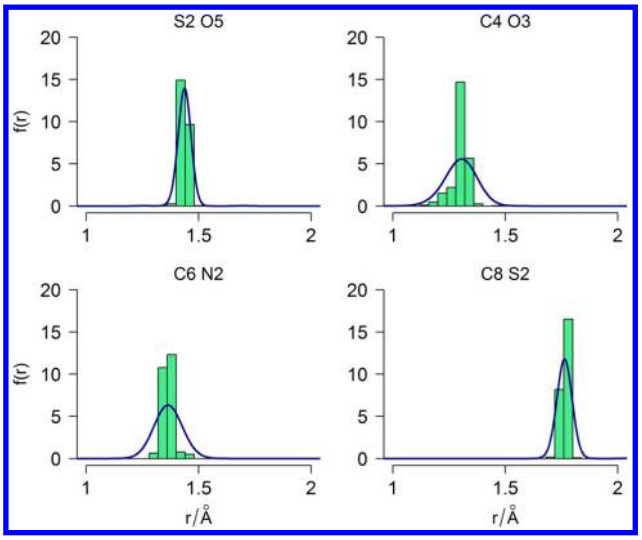


Figure 3. Histograms (light green bars) and Gaussian kernel density approximations (dark blue line) for four bond length distributions in the structure EFOCUY taken from the CSD. It is irrelevant whether bonds are shorter or longer or that the distributions are skewed left or right.

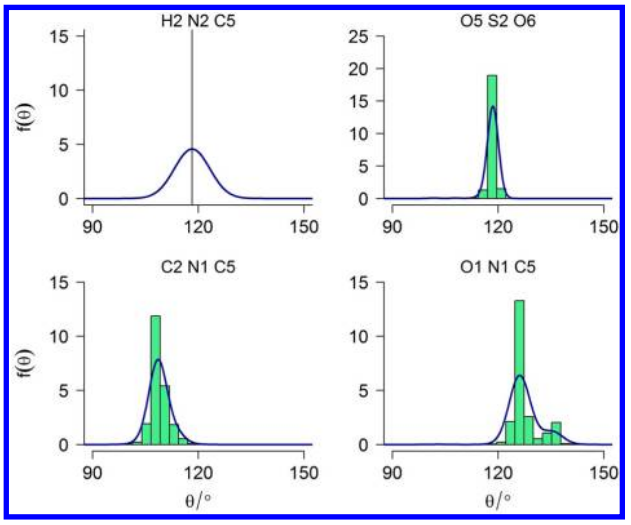


Figure 4. Histograms (light green bars) and Gaussian kernel density approximations (dark blue lines) for four bond angle distributions in the structure EFOCUY from the CSD, with smoothing parameter scaled by a factor of 5 to produce smoother densities. The first plot is based on the mean and standard deviation of the distribution, so no histogram is displayed. Instead, the mean bond angle is shown along with a single normal distribution representing it.

are very sharp and very narrow can represent somewhat difficult objective functions to optimize.

The histogram of the final plot of this sequence in the lower right-hand corner is included as it suggests some bimodality in the bond angle distribution for the valence angle defined by the atoms O1 N1 C5, and this highlights an issue to consider when using the density estimates in, for example, a molecular optimization framework. The kernel density estimate in this plot approximately reflects this bimodal behavior but does not exactly reproduce the bimodality in this instance due to a practical trade-off between the desire to avoid too many maxima, which reduces the utility of gradient-based, e.g., L-BFGS (limited memory Broyden–Fletcher–Goldfarb–Shanno),[16] local minimization routines and the desire to reproduce the variation of the underlying density accurately. The problem of too many maxima can be avoided by introducing a degree of additional smoothing. This is achieved by increasing the smoothing parameters as calculated according to eqs 4 and 6. The degree of smoothing applied depends on the type of parameter under consideration. For bond lengths, we increase the computed smoothing parameter by a factor of 10 and for bond angles by a factor of 5. For torsion angles, no additional smoothing was required. However, in the case of torsion angles, we found it useful to impose the range $10 \leq \nu \leq 500$ on the smoothing parameter computed by eq 6. The lower bound prevents the kernel density estimate becoming excessively broad (see for example the first plot in Figure 1), and the upper bound prevents overflows in $I_0(\nu)$, while still allowing the peaks in the density estimate to become sufficiently narrow to represent well those distributions with sharper maxima.

To illustrate the problem arising when no additional smoothing is included, in Figure 5, we show the results

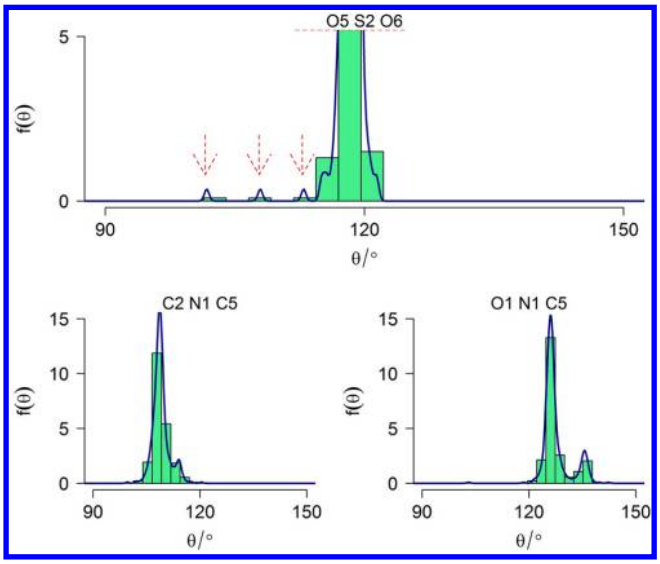


Figure 5. Histograms (light green bars) and Gaussian kernel density approximations (dark blue lines) for the bond angle distributions shown in Figure 4 for the structure EFOCUY from the CSD. The KDE curves now use unadjusted smoothing parameters. The individual plots shown in this figure correspond to those in Figure 4 (omitting the upper left-hand plot from Figure 4, which is neither KDE based nor affected by smoothing.) The scale has also been changed on the upper plot to highlight the regions of interest, indicated by dashed red arrows.

obtained for bond valence angles when the unadjusted smoothing parameters are used directly. The agreement between the histograms and the kernel density approximation is improved, and isolated observations are now clearly visible in the KDE approximation to the PDF. However, this is not desirable from a molecular structure optimization perspective.

In Figure 6 we show typical results for torsion distributions. Here, we see an interesting range of behaviors including bimodal and trimodal distributions. We also see some relatively broad distributions along with some very narrow distributions. Although a common model for torsion distributions is that of a simple cosine dependence, it is hard to see how such a simple model could reproduce the observed distributions in all except

dx.doi.org/10.1021/ci500156d | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

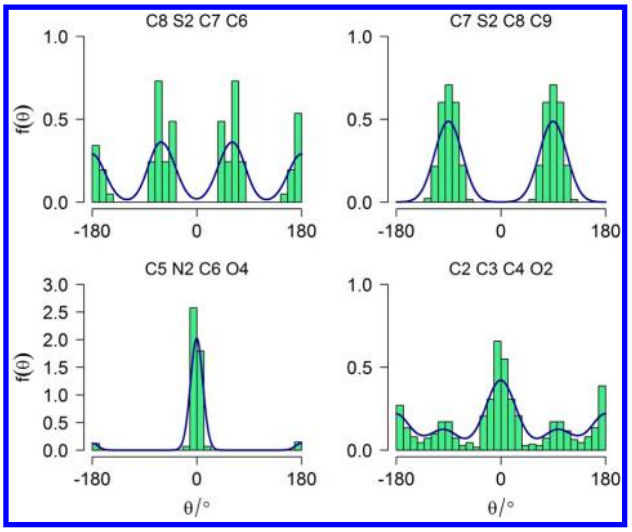


Figure 6. Histograms (light green bars) and von Mises kernel density approximation (dark blue lines) for torsion distributions for four rotatable bonds in the structure EFOCUY from the CSD.

perhaps that in the simplest case, whereas the kernel density approach works well in each case.

Finally, we note that the pdfs can be used to obtain probabilities that parameters lie within certain intervals. For example, for a particular bond whose length is given by r and whose pdf is $f(x)$, the probability r lies in the range $a < r < b$ is given by

$$P(a < r < b) = \int_a^b f(x)\mathrm{d}x \tag{11}$$

In this study, all data was extracted from the CSD, which is composed of small molecule crystal structures, and any probabilities calculated from this data must be interpreted in this context.

CONCLUSION

We have shown how KDE techniques can be applied to the underlying probability distributions generated from molecular geometry parameters. This demonstrates how the disadvantages of histograms for representing these distributions can be overcome by using a kernel density estimate approach. The KDE approach yields smooth probability density functions with the important advantage over histograms of being applicable in gradient-based optimization techniques. In addition, once the PDF is available, one is able to answer further questions relating to probability, such as the probability that a particular bond has a length in a given interval (by integrating the PDF over that interval) or the most likely bond length or the probability of finding a longer (or shorter) bond and so on. We have shown the ability of the method to model arbitrary torsion angle distributions.

A natural application of this KDE approach is a "knowledge-based force field" for use in molecular geometry optimization.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mccabe@ccdc.cam.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Robin Taylor for working on the code providing access to the structural data and Dr. Colin Groom for carefully reading the manuscript.

ABBREVIATIONS

PDF, probability density function; KDE, kernel density estimate; CSD, Cambridge Structural Database; DOI, digital object identifier

REFERENCES

(1) Silverman, B. W. *Density Estimation for Statistics and Data Analysis*; 1st ed.; Monographs on Statistics & Applied Probability; Chapman and Hall: London, 1986.

(2) Godzik, A. Knowledge-based potentials for protein folding: What can we learn from known protein structures? *Structure* **1996**, *4*, 363−366.

(3) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498−2503.

(4) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.

(5) Neudert, G.; Klebe, G. DSX: A knowledge-based scoring function for the assessment of protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2731−2745.

(6) Shapovalov, M. V.; Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844−858.

(7) Bermejo, G. A.; Clore, G. M.; Schwieters, C. D. Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures. *Protein Sci.* **2012**, *21*, 1824−1836.

(8) Allen, F. H. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.

(9) Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons, Inc.: New York, 1992.

(10) Von Mises, R. Uber die "Ganzzahligkeit" der atomgewichte und verwandte fragen. *Phys. Z.* **1918**, *19*, 490−500.

(11) Best, D. J.; Fisher, N. I. Efficient simulation of the von Mises distribution. *J. R. Stat. Soc., Ser. C: Appl. Stat.* **1979**, *28*, 152−157.

(12) Taylor, C. C. Automatic bandwidth selection for circular density estimation. *Comput. Stat. Data Anal.* **2008**, *52*, 3493−3500.

(13) Fisher, N. I. Smoothing a sample of circular data. *J. Struct. Geol.* **1989**, *11*, 775−778.

(14) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*; Dover Books on Mathematics; Dover Publications, Inc.: New York, 1965.

(15) Tegley, C. M.; Viswanadhan, V. N.; Biswas, K.; Frohn, M. J.; Peterkin, T. A. N.; Chang, C.; Bürli, R. W.; Dao, J. H.; Veith, H.; Rogers, N.; Yoder, S. C.; Biddlecome, G.; Tagari, P.; Allen, J. R.; Hungate, R. W. Discovery of novel hydroxy-thiazoles as HIF-alpha prolyl hydroxylase inhibitors: SAR, synthesis, and modeling evaluation. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3925−3928.

(16) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503−528.