Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

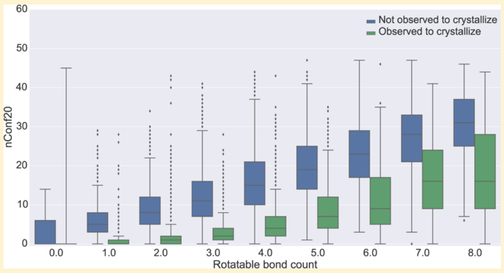
Beyond Rotatable Bond Counts: Capturing 3D Conformational Flexibility in a Single Descriptor

Jerome G. P. Wicker and Richard I. Cooper*

Chemical Crystallography, University of Oxford, Oxford OX1 3TA, U.K.

Supporting Information

ABSTRACT: A new molecular descriptor, $nConf_{20}$, based on chemical connectivity, is presented which captures the accessible conformational space of a molecule. Currently the best available two-dimensional descriptors for quantifying the flexibility of a particular molecule are the rotatable bond count (RBC) and the Kier flexibility index. We present a descriptor which captures this information by sampling the conformational space of a molecule using the RDKit conformer generator. Flexibility has previously been identified as a key feature in determining whether a molecule is likely to crystallize or not. For this application, $nConf_{20}$ significantly outperforms previously reported single-variable classifiers and also assists rule-based analysis of black-box machine learning classification algorithms.



INTRODUCTION

One of the key steps in approaching a cheminformatics problem is the definition of the "chemical space" used to describe the problem.[1] The set of numerical descriptors chosen to capture the characteristics of molecules defines the basis vectors of this space, with n linearly independent descriptors giving rise to an n-dimensional space, in which the coordinates of a particular molecule are given by the values of the descriptors for that molecule. Careful selection of these descriptors provides a useful chemical space for data visualization, similarity measures, and classification or clustering algorithms.

Descriptors can be broadly categorized according to their "dimensionality", based on the type of molecular representation used to calculate them.[2] Zero-dimensional descriptors can be calculated directly from the molecular formula, e.g. molecular weight, while one-dimensional descriptors are bulk properties of the molecule, e.g. calculated solubility.[3] Two-dimensional descriptors, such as connectivity indices and properties of the molecular bond graph, are calculated from a traditional two-dimensional representation of the molecule. Three-dimensional descriptors are computed from a known conformation of a molecule and capture features of a molecule such as shape, distribution of charges, and radius of gyration.[4]

It has previously been shown that increased flexibility can reduce the crystallization tendency of a molecule[5] and that rotatable bond count (RBC), a 2D descriptor based on a set of SMARTS pattern matching rules, was an important feature of molecules for the determination of how easily a molecular material can be crystallized.[6] This may be due to the reduced effective concentration of the "correct" crystallizing conformer in solution for a molecule with more rotatable bonds.[5] However, the exact mechanisms of nucleation and growth of crystals and the influence of molecular conformation are still not fully understood.[7,8]

RBC is quite a crude approximation of molecular flexibility; Kier devised a way of encoding this attribute based on the chemical graph, but this uses descriptors which are also based solely on two-dimensional information.[9]

Other studies of conformational flexibility have been computationally expensive because they attempt to evaluate the entire potential energy hypersurface,[10,11] which is impractical for a large number of molecules. They also either do not yield a single value which can be correlated with a physical property of the molecule or are only appropriate for use with a specific subset of chemical space, such as alkanes.[12,13]

A more direct measure could be obtained by designing a descriptor based on a sampling of the energetically accessible conformers of a given molecule. This approach is relatively quick to compute as it only involves finding the minima of the potential energy surface, and it falls into the 2D category of descriptors outlined above since its value depends only on the chemical connectivity of a molecule but will capture 3D information about the number of low energy conformers of that molecule. This eliminates the computation of the barriers to interconversion and assumes that all conformers are energetically accessible on the time scales of crystallization. Implementation of this new "3D from 2D" descriptor and its application to the problem of predicting if a molecule will be observed to crystallize is described herein.

METHODS

Conformer Generation. Molecules were provided to the conformer-generation step as SMILES strings to ensure no

Received: September 19, 2016
Published: November 23, 2016

DOI: 10.1021/acs.jcim.6b00565

residual conformational information was retained, and explicit hydrogen atoms were added to the skeleton. RDKit cheminformatics toolkit[14] functions were used to generate 50 random molecular conformations, while retaining the starting stereochemistry. RDKit was chosen over other open-source conformer generation tools like BALLOON, CONFAB, and FROG2 and commercial platforms such as MOE, due to speed and the ability to generate conformers which are structurally similar to experimentally determined structures.[15] A knowledge-based conformer generator which uses experimental observations of torsional angle distributions is available in the latest release of the Cambridge Structural Database (CSD) tools.[16] These alternatives have not been explored in this work but could potentially be used to sample conformational space in a similar manner to RDKit.

Each randomly generated conformer was optimized using the Merck Molecular Force Field (MMFF94).[17] MMFF94 is a general purpose parametrized force field comprised of several well-defined contributions to the total potential energy of a molecule, including bond stretching energy, bond torsion, and electrostatic and van der Waals interaction energies. The force field parametrization is determined by training on a large set of computational data derived from *ab initio* calculations on a diverse range of organic and bioorganic structures and has been implemented within the RDKit.[18] Some other force fields suitable for organic molecules include Amber, Gaff, and CHARMM.[19] The Universal Force Field (UFF) can be used to compute energies and gradients of molecules containing almost any element and may therefore prove useful if extending this work to metal−organic complexes or inorganic molecular materials. MMFF94 has been shown to reproduce gas-phase conformer energies more accurately than these other widely available force fields[20] and was chosen for its significantly shorter computational time compared to a more accurate molecular dynamics calculation including solvent effects.

If the optimization did not converge to a stable minimum the conformer was removed. The force field is then used to calculate the energy of each conformer; its energy relative to the lowest-energy conformer found is stored. The lowest energy conformer is retained, and for each other conformer the alignment of all permutations of matching atom orders with the other conformers is checked, to account for symmetry. Any duplicate conformers with a heavy atom root-mean-square (RMS) distance of less than 1.0 Å to any other conformer are removed.

For the small minority (0.05%) of molecules where the MMFF optimization failed, the molecule was removed from the study.

The entire calculation of the energies takes around 0.2 s for molecules with fewer than two rotatable bonds, 1−2 s for molecules with 4 or 5 rotatable bonds, and up to 5 s for molecules with 8 rotatable bonds.

Predictive Model. CSD molecules were obtained from crystal structures in the 2016 CSD release (version 5.37), while lists of commercially available molecules were obtained from ZINC15 downloaded in May 2016. 177 molecular descriptors were calculated using the RDKit cheminformatics toolkit,[14] version Q1 2016. Machine learning algorithms and performance metrics were implemented using version 17.0 of the scikit-learn package.[21] The descriptor definitions and an example of the method used to train a model and output a predictive accuracy from a set of training and test molecules with known labels are given in the Supporting Information of Wicker and Cooper (see http://www.rsc.org/suppdata/ce/c4/c4ce01912a/c4ce01912a1.pdf).[6]

Training and test molecules were selected as in a previous study[6] using information extracted from the ZINC[22] database and the CSD.[16] In this instance, no drug-like filter was applied, to include all organic molecules, which resulted in a set of 48112 commercially available molecules of which 36083 were used for training and 12029 were reserved for a test set. Half of each set consisted of "observed to crystallize" molecules (found in both ZINC and the CSD) and the other half consisted of "not observed to crystallize" molecules (found only in ZINC).

Support vector machines (SVMs) were used as the machine learning algorithm to create the predictive model using the molecular descriptors, having previously been found to give the best performance for this classification problem.[6]

Rule Extraction. The "black-box" nature of the nonlinear SVM predictive model prevents direct determination of the most important descriptors used in performing the classification.[23] Two methods were used to identify these descriptors.

Single Variable Classifiers. The descriptors which are calculated by RDKit were each used in turn to create a single variable classifier built in order to find the descriptor which gave the best predictive accuracy and therefore the most effective classification.[24] The accuracy was assessed both by 5-fold cross-validation on the training set and by prediction on an external validation set. This approach can be extended to two (or more) variable classifiers.

Decision Tree. Rule extraction techniques can be used to mimic the SVM model as closely as possible in order to infer how it is performing the classification.[25] This approach has been used to extract simple rules from machine-learning models which were trained to classify reaction outcomes.[26] The SVM algorithm is trained as usual on the training data set, and the resulting model is then used to obtain the predicted labels for the training data. A conventional decision tree classifier is then trained on the predicted labels to represent the SVM predictive model in terms of a rule-based decision tree.

Descriptor. A new single value descriptor was developed based on the distribution of relative conformer energies. The new descriptor is a count of additional conformers (not including the lowest energy conformer) with energies between selected relative energy thresholds and is designed to approximate the number of energetically accessible conformations of a molecule.

In order to find the optimal energy thresholds for the descriptor, a 5-fold cross-validation was carried out on the training set using the descriptor to create a single variable classifier. Figure 1 shows the distribution of accuracies, which has a broad maximum between an upper threshold of 16 to 20 kcal/mol and a lower energy threshold of 0 to 1 kcal/mol, with no significant difference between the predictive accuracies. This led to a choice of 0 as the lower threshold and 20 as the upper threshold. An example of calculating this descriptor using a 20 kcal/mol cutoff is given in Table 1.

■ RESULTS AND DISCUSSION

Figure 2 shows that rotatable bond count and nConf_{20} capture similar but slightly different information. There is a positive correlation of 0.75 between the two features, but the spread of values of nConf_{20} for each value of RBC is significantly different for those molecules observed to crystallize compared to those which are not. Histograms of the distribution of nConf_{20} values

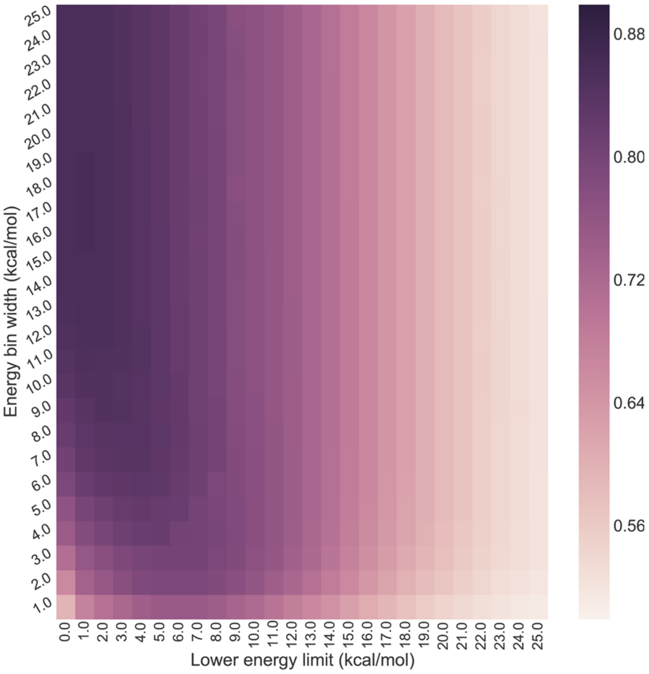
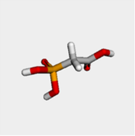
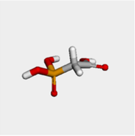
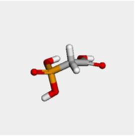
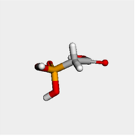
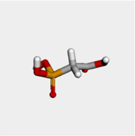
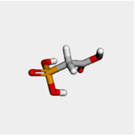


Figure 1. Predictive accuracies for the conformer energy descriptor with varying limits, as determined by 5-fold cross-validation.

Table 1. Example $nConf_{20}$ Calculation for CSD Refcode TERLUX[a]

Conformer ID	Energy (kcal/mol)	Relative energy (kcal/mol)	Conformer
6	-171.3	0.0	
14	-163.3	8.0	
4	-162.6	8.7	
2	-157.3	14.0	
1	-152.5	18.8	
41	-145.9	25.4	

[a] The lowest energy conformer is not counted, and conformer 41 is above the energy threshold, giving an $nConf_{20}$ value of 4.

in each class are plotted in Figure 3, and the distribution of the RBC descriptor is shown in Figure 4. Those molecules which are not observed to crystallize tend to have a larger value of $nConf_{20}$ than those with the same RBC which are observed to crystallize, indicating that $nConf_{20}$ provides better discrimination between the two classes than RBC. Table 2 shows an example molecule where RBC and the new descriptor differ significantly in their estimation of the flexibility of the molecule. Some rotatable bonds cause no change to the molecule, especially when there is symmetry present, information which is captured by $nConf_{20}$.

When $nConf_{20}$ is used to make a single variable classifier of molecules observed and not observed to crystallize, the predictive accuracy on the external validation set is 86.1%, 7.7 percentage points better than any other single variable (Table 3). The new descriptor therefore captures more information than any other single 2D descriptor about the likelihood of a molecule being observed to crystallize.

$nConf_{20}$ was then combined with every other descriptor in turn to create a set of two variable classifiers, and their accuracy was assessed by cross-validation on the training set and prediction on the external test set. In combination with the SMR VSA3 descriptor it produces the best two-descriptor model with a predictive accuracy of 89.2%, as shown in Figure 5. SMR VSA3 is a subdivided surface area descriptor which encodes information about the van der Waals surface area of the molecule with a molar refractivity in the range 0.26−0.35 and has a strong positive correlation with the number of cyclic nitrogen atoms (0.84). The heatmap shows that while the molecules which are not observed to crystallize have a spread of values for both descriptors, the vast majority of molecules observed to crystallize have a value of 0 for both descriptors. This appears to imply that molecules with no additional conformers and no cyclic nitrogen atoms are likely to be observed to crystallize. The black dotted line denoting the SVM decision boundary between the two classes shows an effective separation, and the predictive accuracy is an increase of 4.4 percentage points on any other two-variable classifier of crystallization propensity.

When the algorithm was trained with $nConf_{20}$ and all 177 original descriptors, the predictive accuracy improves by only 0.1% to 96.1% relative to the model with the 177 descriptors without $nConf_{20}$, suggesting that this descriptor provides information to the model that is already indirectly captured by the other original descriptors. However, the new descriptor captures this flexibility information more directly, as demonstrated by the high predictive accuracy when used in a single variable classifier. This is important for unpicking and understanding the decisions made by the machine learning process and will also allow it to be used more easily in linear machine learning classifiers and decision trees, which can become very complicated if a combination of variables is required to predict the output.

The rule extraction analysis further supports the high importance of this flexibility descriptor in performing the classification, as shown in Figure 6. The first node in the tree (which mimics the labels provided by the predictive model for the training data set) provides the best initial split of the data and therefore indicates the most important classification feature. In this case, $nConf_{20}$ is the most important feature; the decision tree shows that the best single−decision approximation of the SVM can be obtained by assuming that the majority of molecules with fewer than 6 low energy

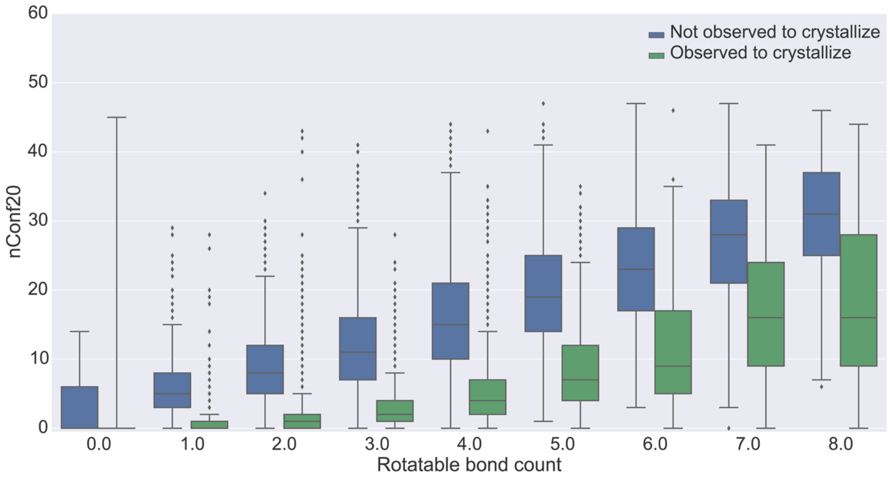


Figure 2. Boxplot of the distribution of $nConf_{20}$ for each value of rotatable bond count, split by class. The central line in the box shows the median of $nConf_{20}$ for that value of RBC. The bottom and top of the box denote the 25th and 75th quartiles, respectively. The whiskers extend to 1.5 times the interquartile range, and any points outside this are plotted as outliers.

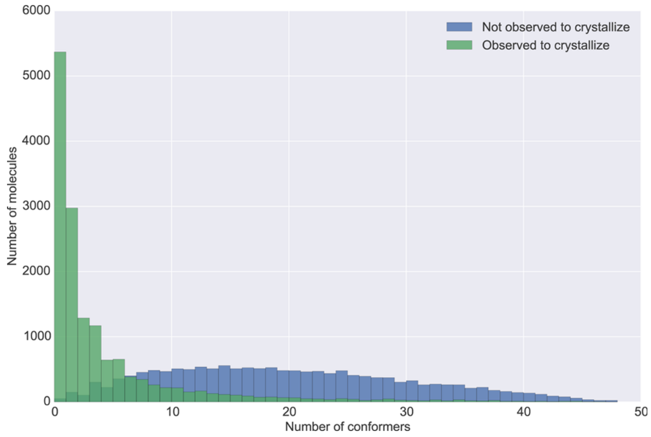


Figure 3. Histogram of $nConf_{20}$ for each of the two classes.

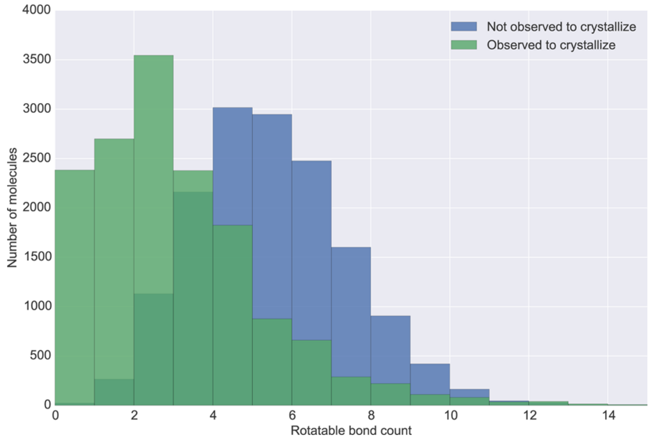


Figure 4. Histogram of rotatable bond count for each of the two classes.

conformers are observed to crystallize, while most of those above this cutoff are assumed to not be observed to crystallize. This agrees with the distribution shown in the histograms in Figure 3. The leaves below this node show that a single $nConf_{20}$ decision alone reproduces the SVM predictive model with an accuracy of 92% on the crystalline leaf and 83% on the noncrystalline leaf (an overall accuracy of 87%).

Table 2. Example Rotatable Bond Counts and $nConf_{20}$ Values

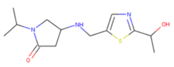
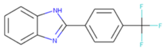
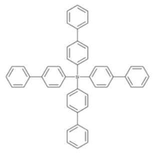
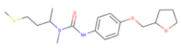
Name	RBC	$nConf_{20}$	Observed to crystallize	Molecule
ZINC000290539224	5	10	No	
ZINC000001235036	1	0	Yes	
ZINC000169816555	8	0	Yes	
ZINC000133698543	8	35	No	

Table 3. Best-Performing Single Variable Classifiers by 5-Fold Cross-Validation and Prediction on an External Test Set[a]

descriptor	cross-validation accuracy (%)	external validation accuracy (%)
$nConf_{20}$	85.9	86.1
valence electron count	78.1	78.4
Kappa1	78.3	78.3
Chi0n	78.3	77.9
SlogP VSA2	77.7	77.8
Kappa2	77.8	77.8
Kier molecular flexibility index	77.3	77.6
path length flexibility index	75.2	75.2
rotatable bond count	73.9	74.8

[a] Existing flexibility descriptors[27] are included for comparison.

■ CONCLUSIONS

We have created and optimized a new descriptor, $nConf_{20}$, which captures the conformational flexibility of a particular molecule based on its 2D chemical connectivity. The descriptor

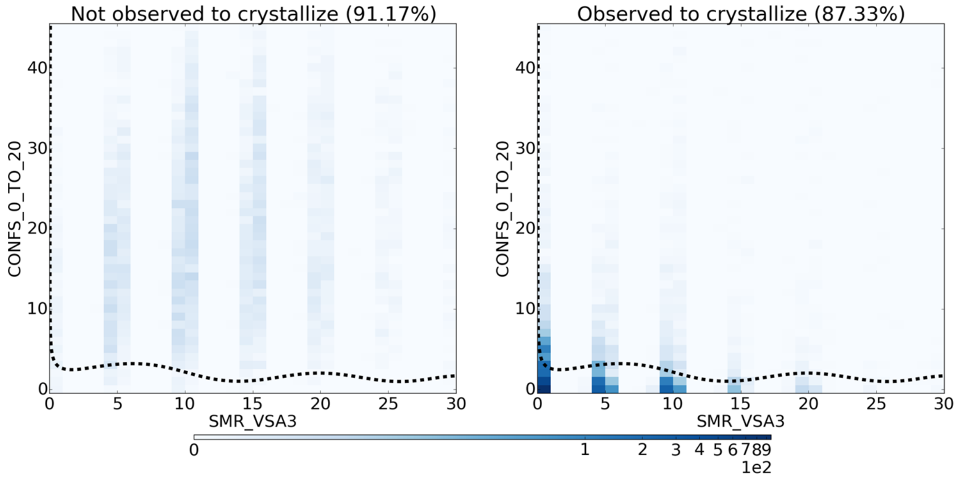


Figure 5. 2D histogram of $nConf_{20}$ against SMR VSA3 for all test molecules color-coded by density of molecules. The dashed line shows the boundary between the crystalline and noncrystalline regions as predicted by the SVM algorithm using RBF kernel.

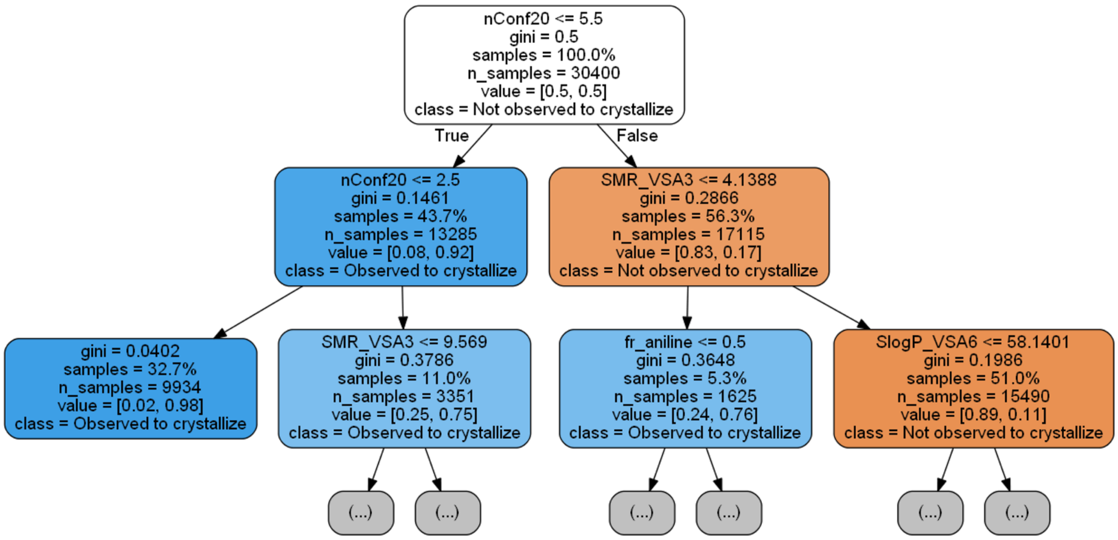


Figure 6. Decision tree used for rule extraction (top 3 levels shown). The gini coefficient is a measure of the impurity of the node. "Samples" indicates the percentage of the total data set present at that node, and "value" is the proportion of "not observed to crystallize" (orange leaves) and "observed to crystallize" (blue leaves) molecules at the node. Each node has been assigned an overall class based on these proportions.

improves on rotatable bond count by taking account of both molecular symmetry and relative energies of conformations, in a manner that is correlated with the crystallization propensity of the molecule. The descriptor encodes relevant information about the 3D shape and flexibility of a molecule from a 2D representation, without the need to consider the interconversion energies, as we have assumed that all contributing conformers are energetically accessible on the time scale of crystallization. We have shown that, of the descriptors tested, this one is the most relevant for predicting crystallization propensity of organic molecules, using both a single variable classifier approach and rule extraction analysis. The overall predictive accuracy of a full-descriptor model including this descriptor is slightly increased, suggesting that the descriptor captures similar information to the other descriptors in a more direct manner; however, use of this descriptor in rule-based classification methods will reduce the complexity of the resulting model. This descriptor has the potential to be applied to other chemical problems where flexibility is a key factor, such as QSAR studies or the prediction of polymorphism. Further improvements could be made, at the expense of computational speed, by incorporating information from molecular dynamics calculations, to take account of solvent effects which may influence how the conformers behave in solution.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00565.

> Python code for conformer generation and subsequent descriptor calculation from the conformer energies, names and descriptor values for training and test molecules (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: richard.cooper@chem.ox.ac.uk.

ORCID

Richard I. Cooper: 0000-0001-9651-6308

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Greg Landrum, author and maintainer of RDKit, the authors of scikit-learn, and William I. F. David for discussions. J.G.P.W. is currently supported by a joint EPSRC doctoral training partnership grant EP/L50503/1 and an STFC postgraduate studentship.

■ REFERENCES

(1) Wong, Y.-S. Exploring Chemical Space. *Methods Mol. Biol.* **2012**, *800*, 11−23.

(2) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(3) Leszczynski, J.; Shukla, M. *Practical Aspects of Computational Chemistry: Methods, Concepts and Applications*; Springer: Heidelberg, 2009; p 203.

(4) Kombo, D. C.; Tallapragada, K.; Jain, R.; Chewning, J.; Mazurov, A. A.; Speake, J. D.; Hauser, T. A.; Toler, S. 3D Molecular Descriptors Important for Clinical Success. *J. Chem. Inf. Model.* **2013**, *53*, 327−342.

(5) Yu, L.; Reutzel-Edens, S. M.; Mitchell, C. A. Crystallization and Polymorphism of Conformationally Flexible Molecules: Problems, Patterns, and Strategies. *Org. Process Res. Dev.* **2000**, *4*, 396−402.

(6) Wicker, J.; Cooper, R. Will It Crystallise? Predicting Crystallinity of Molecular Materials. *CrystEngComm* **2015**, *17*, 1927−1934.

(7) Back, K. R.; Davey, R. J.; Grecu, T.; Hunter, C. A.; Taylor, L. S. Molecular Conformation and Crystallization: The Case of Ethenzamide. *Cryst. Growth Des.* **2012**, *12*, 6110−6117.

(8) Cruz-Cabeza, A. J.; Bernstein, J. Conformational Polymorphism. *Chem. Rev.* **2014**, *114*, 2170−2191.

(9) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221−224.

(10) Koca, J.; Carlsen, P. H. J. DAISY, A Computational Method: A Novel Tool for the Study of the Conformational Behavior of Flexible Molecules. *J. Mol. Struct.: THEOCHEM* **1992**, *257*, 105−130.

(11) Koca, J. Travelling Through Conformational Space: An Approach for Analyzing the Conformational Behavior of Flexible Molecules. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 137−173.

(12) Luisi, P. L. Molecular Conformational Rigidity: An Approach to Quantification. *Naturwissenschaften* **1977**, *64*, 569−574.

(13) Dervarics, M.; Ötvös, F.; Martinek, T. A. Development of a Chirality-Sensitive Flexibility Descriptor for 3 + 3D-QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 1431−1438.

(14) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org/ (accessed Nov 29, 2016).

(15) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

(16) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *B72*, 171−179.

(17) Halgren, T. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17*, 490−519.

(18) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J. Cheminf.* **2014**, *6*, 1−4.

(19) González, M. A. Force fields and molecular dynamics simulations. *Collection SFN* **2011**, *12*, 169−200.

(20) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular Interaction Energies and Geometries. *J. Comput. Chem.* **1999**, *20*, 730−748.

(21) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(22) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(23) Martens, D.; Huysmans, J.; Setiono, R.; Vanthienen, J.; Baesens, B. Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. *Studies in Computational Intelligence* **2008**, *80*, 33−63.

(24) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(25) Barakat, N.; Diederich, J. Eclectic Rule-Extraction from Support Vector Machines. *Int. J. Comput. Intell.* **2005**, *2*, 59−62.

(26) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73−76.

(27) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, 2000; pp 178−179.