doi.org/10.26434/chemrxiv.12671027.v1

Understanding Conformational Entropy in Small Molecules

Lucian Chan, Garrett Morris, Geoffrey Hutchison

Submitted date: 17/07/2020 • Posted date: 20/07/2020

Citation information: Chan, Lucian; Morris, Garrett; Hutchison, Geoffrey (2020): Understanding Conformational Entropy in Small Molecules. ChemRxiv. Preprint.
https://doi.org/10.26434/chemrxiv.12671027.v1

The calculation of the entropy of flexible molecules can be challenging, since the number of possible conformers grows exponentially with molecule size and many low-energy conformers may be thermally accessible. Different methods have been proposed to approximate the contribution of conformational entropy to the molecular standard entropy, including performing thermochemistry calculations with all possible stable conformations, and developing empirical corrections from experimental data. We have performed conformer sampling on over 120,000 small molecules generating some 12 million conformers, to develop models to predict conformational entropy across a wide range of molecules. Using insight into the nature of conformational disorder, our cross-validated physically-motivated statistical model can outperform common machine learning and deep learning methods, with a mean absolute error ≈4.8 J/mol•K, or under 0.4 kcal/mol at 300 K. Beyond predicting molecular entropies and free energies, the model implies a high degree of correlation between torsions in most molecules, often as- sumed to be independent. While individual dihedral rotations may have low energetic barriers, the shape and chemical functionality of most molecules necessarily correlate their torsional degrees of freedom, and hence restrict the number of low-energy conformations immensely. Our simple models capture these correlations, and advance our understanding of small molecule conformational entropy.

File list (2)

Understanding Conformational Entropy in Small Molecules

Lucian Chan,[†] Garrett M. Morris,[†] and Geoffrey R. Hutchison[*,‡,¶]

†*Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, U.K.*

‡*Department of Chemistry, 219 Parkman Avenue, Pittsburgh, PA 15260, U.S.A.*

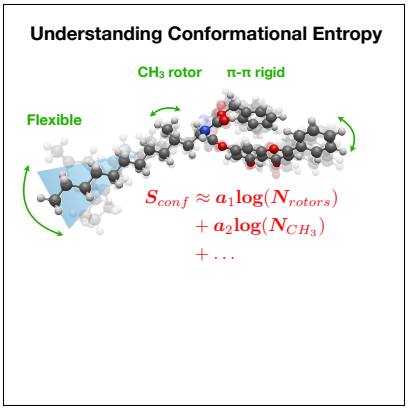
¶*Department of Chemical and Petroleum Engineering, Pittsburgh, PA 15260, U.S.A.*

E-mail: geoffh@pitt.edu

Abstract

The calculation of the entropy of flexible molecules can be challenging, since the number of possible conformers grows exponentially with molecule size and many low-energy conformers may be thermally accessible. Different methods have been proposed to approximate the contribution of conformational entropy to the molecular standard entropy, including performing thermochemistry calculations with all possible stable conformations, and developing empirical corrections from experimental data. We have performed conformer sampling on over 120,000 small molecules generating some 12 million conformers, to develop models to predict conformational entropy across a wide range of molecules. Using insight into the nature of conformational disorder, our cross-validated physically-motivated statistical model can outperform common machine learning and deep learning methods, with a mean absolute error $\approx$4.8 J/mol $\cdot$ K, or under 0.4 kcal/mol at 300 K. Beyond predicting molecular entropies and free energies, the model implies a high degree of correlation between torsions in most molecules, often assumed to be independent. While individual dihedral rotations may have low energetic barriers, the shape and chemical functionality of most molecules necessarily correlate their torsional degrees of freedom, and hence restrict the number of low-energy conformations immensely. Our simple models capture these correlations, and advance our understanding of small molecule conformational entropy.

Graphical TOC Entry



Significance

Conformational entropy is a key quantity in understanding the stability of small molecules and macromolecules, and how tightly they bind. However, its calculation becomes more challenging as the number of degrees of freedom increases, and the number of potential conformers may increase exponentially. Across a set of over 120,000 molecules, we show that the number of low energy conformers increases at a modest, sub-exponential rate with the number of degrees of freedom, implying only a logarithmic increase in conformational entropy. Using a physically-motivated statistical model and new descriptors, we gain new understanding of conformational entropy, correlations between molecular torsions, and the contributions of flexible rings and intramolecular interactions.

While entropy is a major driving force in many chemical changes and is a key component of the free energy of a molecule, it can be challenging to calculate with standard quantum thermochemical methods. Proper consideration in flexible molecules, even within a rigid rotor approximation, requires not just the calculation of the translational, rotational, and vibrational partition functions, but sampling all thermally-accessible conformational degrees of freedom. Several previous efforts have focused on both exhaustive quantum mechanical evaluations of multiple conformers[1–5] and empirical estimates of the entropy from multiple thermally-accessible conformers.[6] Other efforts have used molecular dynamics with varying force fields, which may not yield the same accuracy as modern quantum chemical methods.[7,8]

In principle, the number of possible conformers increases exponentially with the size of a molecule, or more accurately, the number of torsionally rotatable bonds since each of these free or partially hindered rotors should be independent. In solution or gas phase, many bonds have low torsional energy barriers (*e.g.*, sp^3-sp^3 single bonds) even if in the solid state, matrix effects may restrict free torsional motion. Thus, it is common practice in conformer generation to focus on sampling hundreds or thousands of geometrically diverse conformers,[9–11] and using fast molecular mechanics force fields for energy evaluations – even if they do not always correlate well with more accurate electronic structure methods.[12–14]

Recent improvements in density functional tight-binding approximations[15–18] and in availability of computational resources have enabled the work we present here: an evaluation of conformer ensembles and the corresponding entropies of over 120,000 small molecules with up to twenty rotatable bonds, and comprising over 12 million conformers. We have previously noted that the GFN2 method is a relatively fast approximate quantum method with a high degree of correlation with more accurate DLPNO-CCSD(T)[19] single-point energies.[13] Since the GFN2 method is applicable to a wide range of elements, compounds were drawn from the Crystallographic Open Database (COD)[20,21] as well as more complex organic macrocycles from the ZINC database.[22] Most of our analysis focuses on ∼93,000 molecules comprising 9.9

million conformers, with the remainder used as validation sets for statistical and machine-learning prediction models. The set includes a wide range of molecular sizes, with up to 128 atoms, up to 181 bonds, and up to twenty rotatable bonds (see Appendix A, Figure S1).

In all cases, molecular geometries were optimized using the GFN2 method,[15,16] followed by conformer sampling using the iterative metadynamic sampling and genetic crossover (iMTD-GC) method implemented in the CREST program,[17,18] including additional geometry optimization of the final conformer ensemble. The lowest energy conformer was selected for calculating the vibrational modes to evaluate standard rigid rotor harmonic oscillator vibrational, translational, and rotational entropies.[23] We can therefore compare the magnitudes and relative distributions of the GFN2-calculated component entropies (see Figure 1).

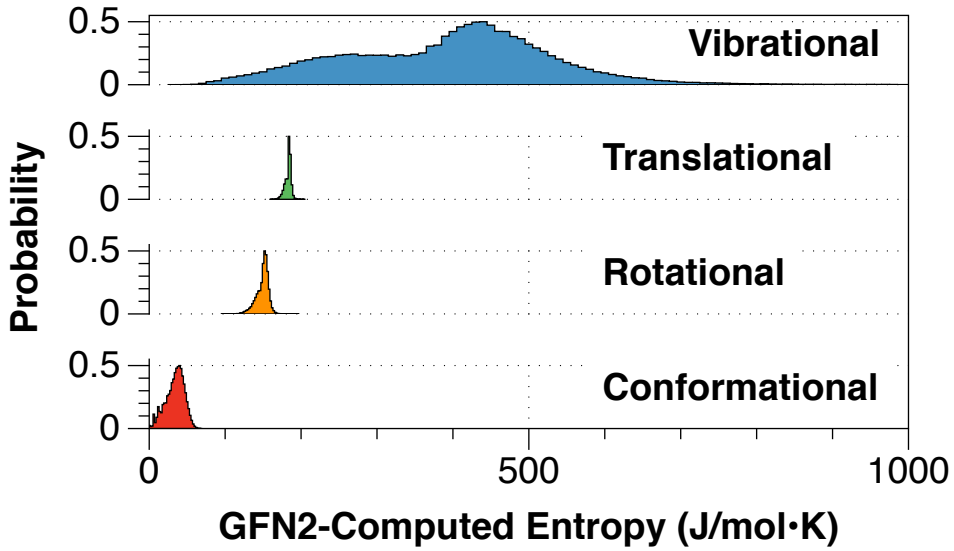


Figure 1: Distributions of GFN2-calculated vibrational, translational, rotational, and conformational entropies across the molecules studied.

The component entropies are not additive, since rotational entropies depend on the moment of inertia, which will be conformer-dependent, and low-energy torsional modes for the thermally-accessible conformers should be removed from the vibrational entropy.[3,4,24–26] Still, vibrational entropies generally contribute the greatest fraction of the total molecular entropy, followed by translational and rotational entropies, respectively. The median confor-

mational entropy comprises $36.3\,\mathrm{J/mol \cdot K}$, or $\approx$2.6 kcal/mol at 300 K and while relatively small, should not be neglected.

We find a reasonable linear correlation between the number of atoms and vibrational entropy, since the vibrational energy is expected to depend on the thermal occupation of low-energy breathing modes (see Appendix A, Figure S5). Thus, it can be easily predicted with linear models with descriptors such as number of atoms, number of bonds, and molecular weight. Higher accuracy predictions may be obtained by optimization and vibrational calculations with other density functional and *ab initio* methods. Similarly, the rotational and translational entropies may be calculated analytically, given the mass, symmetry of the molecule, and the moments of inertia of a particular geometry. While the vibrational entropy may be the largest in magnitude, it can also be calculated efficiently from an optimized geometry (*i.e.* median time of 16 seconds with the GFN2 method using a dual-core job; see Appendix A, Figure S6).

In contrast, few studies have considered conformational entropy across a wide range of small molecules. The time required is 200-300 times longer than the vibrational calculations, with a median time of 1.01 hours per compound, and an average of 2.08 hours per compound for a dual-core job using the GFN2 method on the same hardware (see Appendix A, Figure S6).

As mentioned above, conformer generation for small molecules often involves sampling dihedral angles from a set of defined "rotatable bonds", specified from a set of patterns for acyclic bonds with low rotational barriers (*e.g.*, sp^3–sp^3 non-ring single bonds).[10] The number of conformers should therefore increase with the number of rotatable bonds, as will the conformational entropy. Since we use an approximate density functional method, GFN2, we seek to build physical understanding of the components of conformational entropy through a statistical model across our entire set of molecules, using separate validation data to consider the generality of the trends and avoid overfitting.

An unbranched n-alkane, C_nH_{n+2}, is the simplest type of acyclic saturated hydrocarbon.

The low torsional energy barrier of carbon-carbon single bonds enables all bonds to freely rotate and result in different conformations. In principle, with low torsional barriers and all bonds being equal, the number of conformers should increase exponentially with the count of rotatable bonds ($\approx 3^{n-3}$), assuming three possible local minima per rotatable bond. However, symmetry, correlated dihedral angles, and excluded volume often reduce the number of thermally accessible conformers.[2,3,5,27]

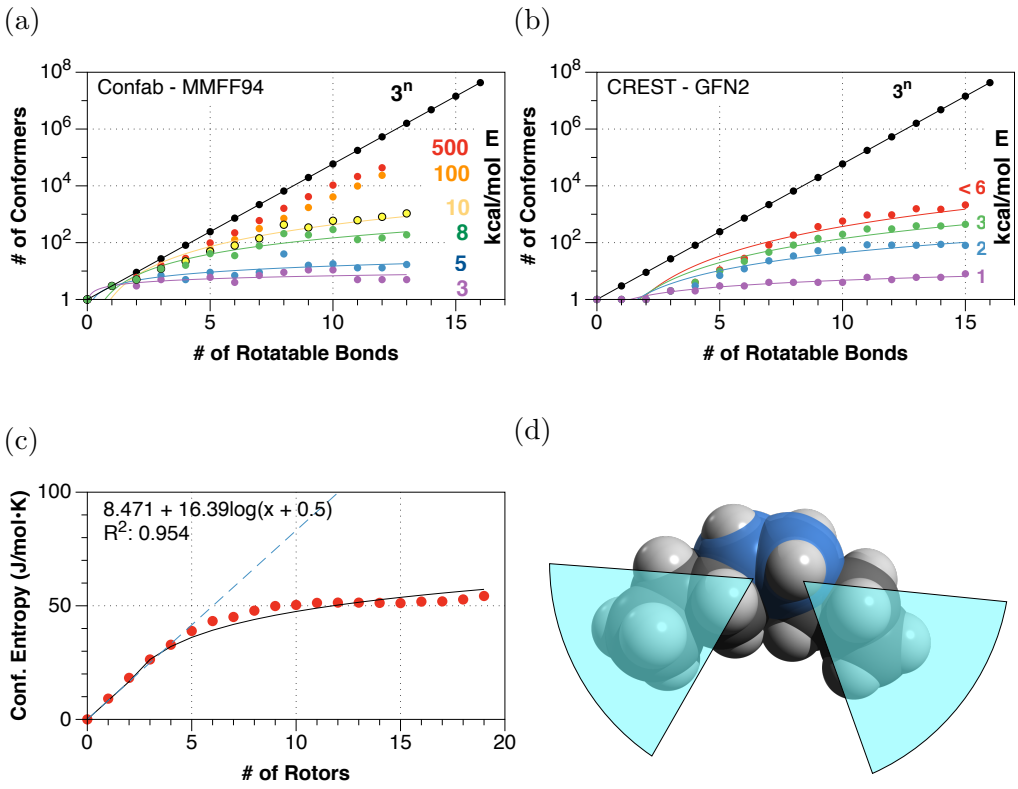


Figure 2: Conformational entropies for increasing lengths of n-unbranched alkanes, C_nH_{2n+2}. The counts of conformations in (a) and (b) are shown on a logarithmic scale. (a) Number of alkane conformers within a given energy range (in kcal/mol) of the global minimum (*i.e.*, within 3, 5, 8, 10, 100, and 500 kcal/mol) using Confab exhaustive sampling with the MMFF94 force field. (b) Number of conformers within a given energy window in kcal/mol of the global minimum using CREST sampling and the GFN2 method. (c) Conformational entropies for calculated for n-alkanes using CREST / GFN2. Note that for smaller hydrocarbons ($n < 4$ carbons) the scaling is approximately linear, and beyond $n = 8-10$ carbons, the conformational entropies are roughly constant. (d) Schematic of central torsion in octane C_8H_{18} indicating potential steric bumping (clashing carbons shown in blue) between the two molecular ends.

Rather than an exponential number of possible conformers in the linear alkanes, the number of *low-energy conformers* increases sub-linearly on a logarithmic scale, when evaluated either with exhaustive systematic conformer enumeration (Confab)[9] using a standard molecular force field (MMFF94),[28] or using CREST conformer generation with the GFN2 method, as illustrated in Figures 2a and 2b respectively. The curves fit roughly to a power-law function, with exponents $\approx 1.5 - 2.6$, depending on the method and the energy window.

Since the number of low-energy conformers increases relatively slowly (*i.e.* sub-exponentially) with the number of rotatable bonds, the conformational entropy will therefore increase logarithmically, as found by the computed CREST / GFN2 entropies. For short alkane chains ($n < 4$ carbon atoms), the increase in conformational entropy is approximately linear, and approximately logarithmic or perhaps close to constant for long chains (see Figure 2c). One can understand that in long chains, dihedral motion in the center of the molecule will inherently restrict otherwise free rotations to avoid steric clashes—a concept known as *excluded volume* in polymer theory. These results match previous detailed quantum chemical calculations of conformational entropy in linear alkanes.[2,3,5,27]

Figure 3a shows the conformer populations across the set of ∼93,000 molecules at different GFN2-computed energy cutoffs (shown in different colors up to 6 kcal/mol), and the number of conformers within 6 kcal/mol of the global minimum grows at a logarithmic rate, reaching $\sim 10^3$ conformers for molecules with twenty rotatable bonds. Across the set, this still suggests the number of rotatable bonds is a useful predictor of the number of thermally-accessible conformers, and thus the conformational entropy — even if in larger molecules, the degrees of freedom are inherently correlated.

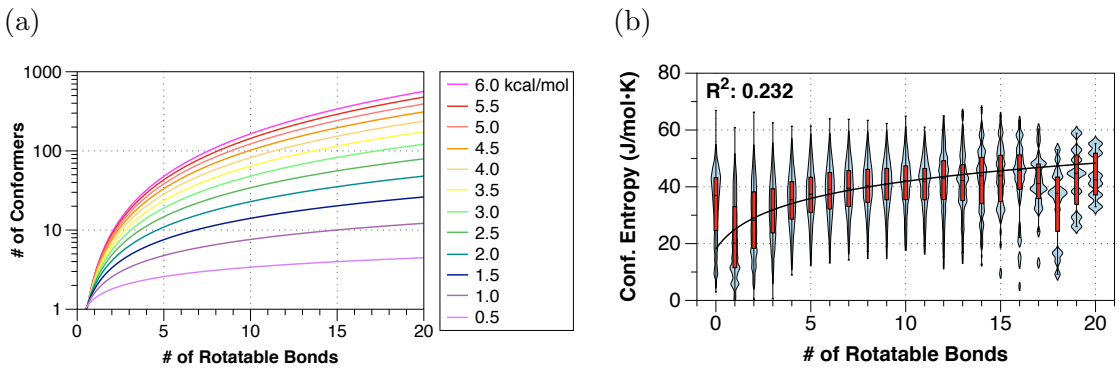


Figure 3: (a) Scaling of the number of conformers across the ∼93,000 molecules in the training set, on a logarithmic scale, within a given energy threshold, as a function of the number of rotatable bonds; and (b) correlation between the number of rotatable bonds, N_{rotor}, and GFN2-calculated conformational entropies, shown as violin plots for each rotatable bond bin. The line indicates a logarithmic best fit, *i.e.* $a + b\log(N_{\text{rotor}} + 1)$, with a coefficient of determination of 0.232; this highlights the need for better predictors than simply the number of rotatable bonds.

Beyond simple linear alkanes, branched alkanes and cycloalkanes can be used as models to understand other components of the conformational entropy. Both propylene chains and highly branched alkanes exhibit logarithmic increases in CREST-computed conformational entropy, based on the number of terminal CH_3 groups (see Figures 4a and 4b). Note that methyl groups are known to increase entropy as hindered rotors.[29,30] The magnitude of the methyl rotor entropies are higher from the CREST/GFN2 ensembles than previous quantum chemical estimates (*i.e.* $9.1\,\text{J/mol}\cdot\text{K}$ from CREST/GFN2 vs. $6.8\,\text{J/mol}\cdot\text{K}$ from HF/6-31G(d) using a hindered rotor model),[29] but reflect that beyond iso-pentane, correlations between multiple CH_3 groups slow the increase in conformational entropy to logarithmic. Similarly, while cycloalkanes have fewer torsional degrees of freedom (*i.e.* $N - 3$ for an N-membered ring), the CREST-computed conformational entropy increases logarithmically with the ring size (see Figure 4c).

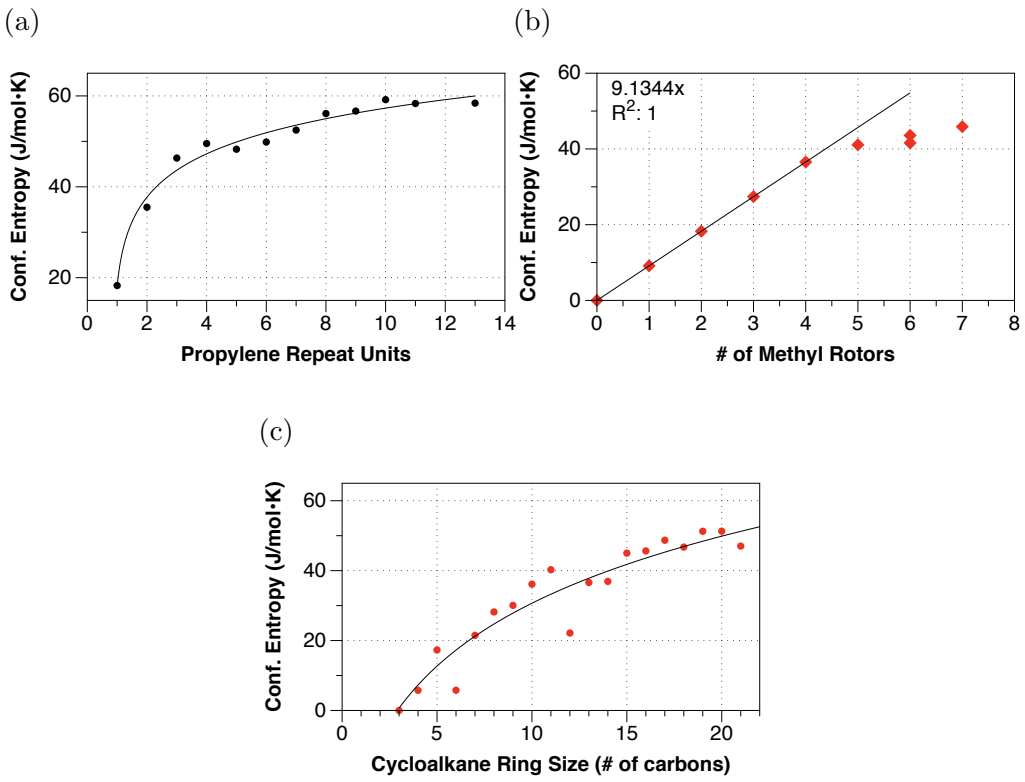


Figure 4: Conformational entropies for alkanes. (a) Conformational entropies calculated for increasing lengths of polypropylene chains, as a function of the number of repeat units, illustrating approximately logarithmic increase; and (b) branched alkane chains as a function of the number of terminal methyl rotors. (c) Conformational entropies calculated for increasing ring size of cycloalkanes, n-C_nH_{2n}. The conformational entropies tends to grow logarithmically with ring size.

Building from the simple alkanes, we can understand that the conformational entropy has multiple components based on the torsional degrees of freedom, including rotatable bonds, terminal CH_3 groups, and correlated motions in flexible rings, such as the cycloalkanes.

Rings can be fused together, forming bicyclic and polycyclic rings. These rings can share one atom (spirocyclic), two adjacent atoms (fused), and three or more atoms (bridged). These three modes of ring junction impose different steric constraints on the molecule, and give distinct low energy conformations. The number of degrees of freedom of these complex rings cannot be simply explained by the ring size. Thus, we introduce a new descriptor, namely *Total Ring Flexibility*, R_f^{Total}, to better understand the flexibility and thus conformational

entropy of rings. We apply the concept of unique ring families (URFs)[31] to decompose the ring systems into subfamilies, and calculate the number of degrees of freedom (or ring flexibility) for each subfamily. The Total Ring Flexibility is simply the sum of the ring flexibility of all subfamilies, minus any penalties from constraints imposed in the ring(s), such as endocyclic double bonds, shared aromatic bonds and/or different ring junction types (see Appendix C, Table S5). Adding the Total Ring Flexibility measure to our statistical model shows good correlation with the CREST-computed conformational entropy values, with a Pearson correlation coefficient, $R^2 = 0.7$ (see Appendix C, Figure S11).

In addition to the additive effects of the number of rotatable bonds, terminal methyl groups, and ring flexibility, some types of molecular functional groups *reduce* the conformational degrees of freedom. For example, in our previous work, BOKEI, we found correlated neighboring dihedral torsions due to steric clashes and intramolecular interactions.[32] Consequently, the conformational entropy decreases due to various intramolecular functional group interactions, for example hydrogen bonds and π-π stacking.

It is well-known that five- to eight-membered ring systems have a high propensity to form intramolecular hydrogen bonds (see definition in Appendix C, Table S3 and S4), and some motifs are commonly used in drug design.[33,34] Their geometries, including bond angles and distance, have been studied by others.[33,34] In this work, we further characterise the shortest path between donor and acceptor atoms, and show that the atom types and their positions along the shortest path are highly conserved for a given acceptor atom type (see Appendix C, Figure S13). For example, an alkoxyl hydrogen bond acceptor has a high propensity to form hydrogen bonds with hydroxyl donors, and the atoms along the shortest path tend to be ring atoms. Our analysis provides insights into the chemical characteristics of the shortest path between intramolecular hydrogen bonds, and this information can be used to identify potential intramolecular hydrogen bonds without knowing 3D structures of the molecule.

Intramolecular π-π interactions usually require longer paths than intramolecular hydrogen

bonds, and the formation of π-π interactions usually leads to a so-called 'folded' structure. The "foldability" of the path depends on the path length, and intramolecular hydrogen bonds generally assist the formation of long π-π stacking. We analyzed the simulated molecular structures from ZINC and COD, and found that six functional groups (amide, ester, ketone, ether, urea and carbamate group) play an important role in supporting such small molecule 'folding'. Figure S15 in Appendix C shows that the position of these functional groups are highly conserved. Amide and ester are commonly observed in our datasets and amide has higher variations in position than other functional groups.

Further, the partial double bond character in amides, thioamides and esters increases the rotational energy barrier, and reduce the conformer population size, and thus the conformational entropy. To estimate the effect of the intramolecular interactions and the delocalization of electrons, we introduce new descriptors, namely the *foldability*, F, and functional group counts, N_{SG}, to count the number of rotatable bonds involved in the shortest path between the terminal heavy atoms involved in intramolecular hydrogen bonds and π-π interactions, and the number of specified functional groups in the molecule respectively.

As discussed above, calculation of conformational entropy is time-consuming, thus it is beneficial to build models that can rapidly estimate conformational entropy. Using the physical understanding of the logarithmic functional form and contributions to the number of low-energy conformers and thus conformational entropy, we compared different statistical and machine learning models, including linear regression, LASSO, ridge regression, kernel ridge regression (KRR), and neural networks (NN). The new descriptors mentioned above were used in linear regression. Standard molecular fingerprints (ECFP6)[35] were used as inputs for various machine learning models.

The conformational entropy, S_{conf}, depends on the number of low-energy thermally-accessible conformations, which increases sub-exponentially as seen in Figure 3a, and can be approxi-

mated by Equation (1) for some constant, C.

$$S_{\text{conf}} = \log(n_{\text{states}}) \approx C \log(N_{\text{rotor}}) \tag{1}$$

As illustrated in Figure 3b, there is a weak *linear* correlation between the computed conformational entropies and the number of rotatable bonds, N_{rotor} ($R^2 = 0.23$), because the number of terminal CH_3 groups, and contributions of ring flexibility including of molecules with no "rotatable bonds", are not considered. For example, molecules with no rotatable bonds are calculated to have a median conformational entropy of 40 J/mol K, indicating most have some flexibility. The residual plots in Figure S17 in Appendix D show that the model tends to be underestimated for small fitted values and be overestimated for high fitted values, suggesting a more sophisticated multivariate model is needed.

Using the contributions discussed above in model systems, we may approximate the conformational entropy with following linear model (LM-Best):

$$\begin{aligned} S_{conf} \approx \beta_0 + \beta_1 \log(N_{\text{rotor}} + 1) + \beta_2 \log(N_{\text{Methyl}} + 1) + \beta_3 \log(N_{\text{SG}} + 1) \\ \beta_4 \log(F_{\text{HBond}} + 1) + \beta_5 \log(F_{\pi-\pi} + 1) + \beta_6 \log(R_f^{Total} + 1) \end{aligned} \tag{2}$$

where N_{rotor}, N_{Methyl} and N_{SG} are the number of rotatable bonds, number of methyl (CH_3) groups and number of specified functional groups (amide, ester and thioamide) respectively. R_f^{Total} is the total ring flexibility, F_{HBond} and $F_{\pi-\pi}$ are the foldability scores for intramolecular hydrogen bonds and π-π stacking within a molecule (see Appendix C).

This model is in close agreement with the GFN2-calculated conformational entropy, with a coefficient of determination, $R^2 = 0.715$. The error is smaller than the one-variable model in Eq.(1). Negative values of the parameters associated with the number of specified functional groups, foldability with intramolecular hydrogen bonds, and $\pi - \pi$ stacking, suggests that the conformational entropy decreases as these variables increase, which matches our expec-

tations. Small p-values from the t-tests indicate these parameters are significantly different from zero (see Appendix D). Surprisingly, the parameters associated with the ring flexibility are slightly negative, which is not consistent with observations in cycloalkanes (Figure 4c) and the small cyclic molecules subset shown in Appendix C Figure S11. This indicates our proposed descriptor, ring flexibility, may not fully capture the conformational entropy of complex rings. Taking the substituent information and the intramolecular interactions within rings into account remains areas for future work.

To assess the predictive power of the model, we calculated the mean absolute error between the model-predicted and GFN2-computed conformational entropies for two independent test sets, ZINC-I and the peptides set. The ZINC-I set contains diverse small molecules selected from ZINC, and has no overlap with the training data. The peptides set contains 8,861 cyclic tetrapeptides (CTPs) composed of fourteen different naturally-occurring amino acids (see Appendix A, Table S1). Our proposed linear model outperforms machine learning models, and gives a mean absolute error of 4.77 and 4.65 $\mathrm{J/mol \cdot K}$ respectively (see Table 1). We argue that the ECFP6 fingerprints only consider local information about the corresponding atom, and the global topological information including long range intramolecular interactions therefore cannot be encapsulated in such representations. This limits the predictive power of the models. The kernel approach fails to obtain good predictions in peptides, as the cyclic peptides are likely too dissimilar from molecules in training data (see Appendix A, Figure S3).

Table 1: Model performance. Comparison of the mean absolute error (MAE) between the model-predicted and GFN2-computed conformational entropies, in $J/mol \cdot K$, for training set and both test sets. LR-1 is a single-variable linear model, with number of rotatable bonds as the sole explanatory variable. LR-Best gives the lowest MAE in both test sets.

Model	Training (MAE)	ZINC-I Set (MAE)	CTPs Set (MAE)
LR-1	8.67	8.83	9.00
LR-Best	5.16	**4.77**	**4.65**
LASSO	5.55	5.47	6.76
Ridge	4.95	5.29	5.83
KRR	5.90	5.87	8.79
DNN	7.32	7.22	6.76

In summary, our analysis shows that the conformational entropy increases logarithmically with the number of degrees of freedom in the small molecules. Despite the possible number of conformers increasing exponentially, inherent correlation between multiple rotatable bonds and terminal CH_3 groups restricts the number of thermally-accessible conformations greatly. Intramolecular interactions such as π-π stacking and intramolecular hydrogen bonds further reduce the number of thermally-accessible conformers, and decrease the conformational entropy as a result. Such effects, here in small molecules, relate to Levinthal's paradox and energy landscapes found in protein folding.[36–38] The contribution of ring entropy from flexible rings has to be assessed carefully. Our new descriptors consider the intramolecular functional groups that decrease conformational flexibility, *ring flexibility* and *foldability*, and thus improve the prediction of the conformational entropy component of standard molecular entropy.

The resulting statistical model, based on a physical understanding of the various contributions to conformational entropy, outperforms current machine learning methods, and gives a mean absolute error of $4.8\,J/mol \cdot K$, or $\approx 0.34\,kcal/mol$ at $300\,K$. Our approach facilitates the calculation of thermodynamic properties and provides insights into the effect of intramolecular interactions on conformational preferences and intrinsic correlation between molecular torsional motion. This work can also be extended to predict the change in solva-

tion entropy as well as ligand conformational entropy upon protein-ligand binding, and thus provide better estimates of binding free energies for drug discovery.

Methods and Materials

Data

Molecules with twenty or fewer rotatable bonds from the Crystallography Open Database (COD)[20,21] and ZINC[22] were used to construct the training and testing sets; details are given in Appendix A. We also constructed an additional test set consisting of cyclic tetrapeptides composed of all combinations of four out of fourteen different amino acids (see Appendix A, Table S1). CREST[17,18] was used to simulate the lowest energy conformations of all molecules, and GFN2[15,16] was used for energy evaluation after geometry optimization. Note that the CREST calculation may break molecules into fragments; those molecules that were fragmented in the final output were excluded from our analysis.

Models

Extended circular fingerprints with diameter 6, ECFP6,[35] as computed by RDKit,[39] were used as model inputs for all machine learning models. We used the Scikit-Learn[40] implementations of LASSO regression, ridge regression, kernel ridge regression, using the Tanimoto (a.k.a. Jaccard) kernel,[41] and cross-validation. Keras[42] was used to train the neural network. The implementation details of these models are described in detail in Appendix E.

Molecular descriptors such as the number of rotatable bonds, number of methyl groups, counts of functional groups (amide, ester and thioamide), plus our own descriptors, namely total ring flexibility, and foldability, were used in the linear models. RDKit was used to compute these counts. RingDecomposerLib[43] was used for the ring decomposition and calculation of ring flexibility. The construction of the foldability score and total ring flexibility

are discussed in detail in Appendix C.

Acknowledgement

GRH thanks the National Science Foundation (CHE-1800435) for support. GMM thanks the EPSRC and MRC for financial support under grant number EP/L016044/1. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work, the University of Pittsburgh Center for Research Computing through the computational resources provided, and using resources provided by the Open Science Grid,[44,45] which is supported by the National Science Foundation award 1148698, and the U.S. Department of Energy's Office of Science. G.R.H. thanks Lillian Chong and Jennifer Laaser for helpful discussions. L.C. thanks Catherine Wong for helpful discussions.

Supporting Information Available

The data and code can be found in GitHub `https://github.com/ghutchis/molecular-entropies`.

References

(1) Zheng, J.; Yu, T.; Papajak, E.; Alecu, I. M.; Mielke, S. L.; Truhlar, D. G. Practical methods for including torsional anharmonicity in thermochemical calculations on complex molecules: The internal-coordinate multi-structural approximation. *Physical Chemistry Chemical Physics* **2011**, *13*, 10885–10907.

(2) Speybroeck, V. V.; Vansteenkiste, P.; Neck, D. V.; Waroquier, M. Why does the uncoupled hindered rotor model work well for the thermodynamics of n-alkanes? *Chemical Physics Letters* **2005**, *402*, 479 – 484.

(3) Ellingson, B. A.; Lynch, V. A.; Mielke, S. L.; Truhlar, D. G. Statistical thermodynamics of bond torsional modes: Tests of separable, almost-separable, and improved Pitzer–Gwinn approximations. *The Journal of Chemical Physics* **2006**, *125*, 084305.

(4) Simón-Carballido, L.; Bao, J. L.; Alves, T. V.; Meana-Pañeda, R.; Truhlar, D. G.; Fernández-Ramos, A. Anharmonicity of Coupled Torsions: The Extended Two-Dimensional Torsion Method and Its Use To Assess More Approximate Methods. *Journal of Chemical Theory and Computation* **2017**, *13*, 3478–3492.

(5) Wu, J.; Ning, H.; Xu, X.; Ren, W. Accurate entropy calculation for large flexible hydrocarbons using a multi-structural 2-dimensional torsion method. *Physical Chemistry Chemical Physics* **2019**, *21*, 10003–10010.

(6) Ghahremanpour, M. M.; van Maaren, P. J.; Ditz, J. C.; Lindh, R.; van der Spoel, D. Large-scale calculations of gas phase thermochemistry: Enthalpy of formation, standard entropy, and heat capacity. *The Journal of Chemical Physics* **2016**, *145*, 114305.

(7) Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. Estimating entropies from molecular dynamics simulations. *The Journal of Chemical Physics* **2004**, *120*, 2652–2661.

(8) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **2011**, *7*, 2638–2653.

(9) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *Journal of Cheminformatics* **2011**, *3*, 8.

(10) Hawkins, P. C. D. Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling* **2017**, *57*, 1747–1756.

(11) Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J. et al. PubChem3D: a new resource for scientists. *Journal of Cheminformatics* **2011**, *3*, 32–32.

(12) Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A sobering assessment of small-molecule force field methods for low energy conformer predictions. *International Journal of Quantum Chemistry* **2018**, *118*, e25512.

(13) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies using Electronic Structure and Machine Learning Methods. *International Journal of Quantum Chemistry* **2020**,

(14) Rai, B. K.; Sresht, V.; Yang, Q.; Unwalla, R.; Tu, M.; Mathiowetz, A. M.; Bakken, G. A. Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein-Ligand Complexes using ab Initio Calculations. *Journal Of Chemical Information And Modeling* **2019**, *59*, 4195–4208.

(15) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(16) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009.

(17) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *Journal of Chemical Theory and Computation* **2019**, *15*, 2847–2862.

(18) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemi-

cal space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(19) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics* **2018**, *148*, 011101.

(20) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography* **2009**, *42*, 726–729.

(21) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **2012**, *40*, D420–D427.

(22) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

(23) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chemistry – A European Journal* **2012**, *18*, 9955–9964.

(24) Pfaendtner, J.; Yu, X.; Broadbelt, L. J. The 1-D hindered rotor approximation. *Theoretical Chemistry Accounts* **2007**, *118*, 881–898.

(25) Ayala, P. Y.; Schlegel, H. B. Identification and treatment of internal rotation in normal mode vibrational analysis. *The Journal of Chemical Physics* **1998**, *108*, 2314–2325.

(26) Vansteenkiste, P.; Van Neck, D.; Van Speybroeck, V.; Waroquier, M. An extended hindered-rotor model with incorporation of Coriolis and vibrational-rotational coupling

for calculating partition functions and derived quantities. *The Journal of Chemical Physics* **2006**, *124*, 044314.

(27) Vansteenkiste, P.; Van Speybroeck, V.; Marin, G. B.; Waroquier, M. Ab Initio Calculation of Entropy and Heat Capacity of Gas-Phasen-Alkanes Using Internal Rotations. *The Journal of Physical Chemistry A* **2003**, *107*, 3139–3145.

(28) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 587–615.

(29) Irikura, K. K. How much does a methyl rotor (internal rotation) contribute to the entropy? 2020; `https://cccbdb.nist.gov/methylrotor.asp`.

(30) Irikura, K. K. Appendix B: Essential Statistical Thermodynamics. *Computational Thermochemistry* **1998**, 402–418.

(31) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling* **2012**, *52*, 2013–2021.

(32) Chan, L.; Hutchison, G. R.; Morris, G. M. BOKEI: Bayesian optimization using knowledge of correlated torsions and expected improvement for conformer generation. *Physical Chemistry Chemical Physics* **2020**, *22*, 5211–5219.

(33) Kuhn, B.; Mohr, P.; Stahl, M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2010**, *53*, 2601–2611.

(34) Bilton, C.; Allen, F. H.; Shields, G. P.; Howard, J. A. K. Intramolecular hydrogen bonds: common motifs, probabilities of formation and implications for supramolecular organization. *Acta Crystallographica Section B* **2000**, *56*, 849–856.

(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(36) Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology* **1997**, *4*, 10–19.

(37) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 20–22.

(38) Levinthal, C. Are there pathways for protein folding? *Journal de chimie physique* **1968**, *65*, 44–45.

(39) Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org`, 2019.

(40) 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. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(41) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093 – 1110.

(42) Chollet, F. Keras. `https://github.com/fchollet/keras`.

(43) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *Journal of Chemical Information and Modeling* **2017**, *57*, 122–126.

(44) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F. et al. The Open Science Grid. Journal of Physics: Conference Series. 2007; p 012057.

(45) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Wurthwein, F. The Pilot Way to Grid Resources Using glideinWMS. 2009 WRI World Congress on Computer Science and Information Engineering. 2009; pp 428–432.

Supplementary Information: Understanding Conformational Entropy in Small Molecules

Lucian Chan,[†] Garrett M. Morris,[†] and Geoffrey R. Hutchison[*,‡,¶]

†*Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK*

‡*Department of Chemistry, 219 Parkman Avenue, Pittsburgh, PA 15260, U.S.A.*

¶*Department of Chemical and Petroleum Engineering, Pittsburgh, PA 15260, U.S.A.*

E-mail: geoffh@pitt.edu

Appendix A: Data Summary

Molecules with twenty or fewer rotatable bonds from the Crystallography Open Database[S1,S2] and ZINC[S3] were selected for the study, and consisted of any of the following elements: hydrogen, boron, carbon, nitrogen, oxygen, fluorine, silicon, phosphorus, sulfur, chlorine, bromine and iodine. In our analysis, we split the data into two sets: (i) a training set (93021 molecules), and (ii) a testing set (ZINC-I) (15547 molecules). Figure S1 shows the atom counts, bonds counts, number of methyl groups, rings counts, and the number of rotatable bonds in the training dataset. RDKit[S4] was used to compute these counts. Note that RDKit's strict definition of rotatable bonds does not count bonds like amides and esters as rotatable.

The ZINC-I testing set consists of diverse small molecules taken from ZINC only. The InChI key of each molecule in both the training set and testing set was computed to ensure there was no overlap between the two datasets. We also constructed an additional testing set, the cyclic tetrapeptides (CTPs) set, to assess model performance. The peptides set contains only head-to-tail cyclic peptides, *i.e.* cyclization from the N-terminus to the C-terminus, and thus gives a set of 12-membered rings. They are composed of four out of fourteen amino acids (see Table S1). Figure S2 shows the distribution of the rotatable bonds in training set, ZINC-I testing set, and CTPs testing set. The Tanimoto similarity[S5] of the ECFP6[S6] fingerprints of the molecules between the training and the testing sets was also calculated, as illustrated in Figure S3.

Table S1: Amino acids that are included in the cyclic tetrapeptides (CTPs) test set.

Type	**Amino Acids**
Special	Cysteine, Glycine
Charged	Histidine, Lysine, Aspartic Acid, Glutamic Acid
Polar Uncharged	Serine, Threonine
Hydrophobic	Alanine, Valine, Leucine, Phenylalanine, Tyrosine, Tryptophan

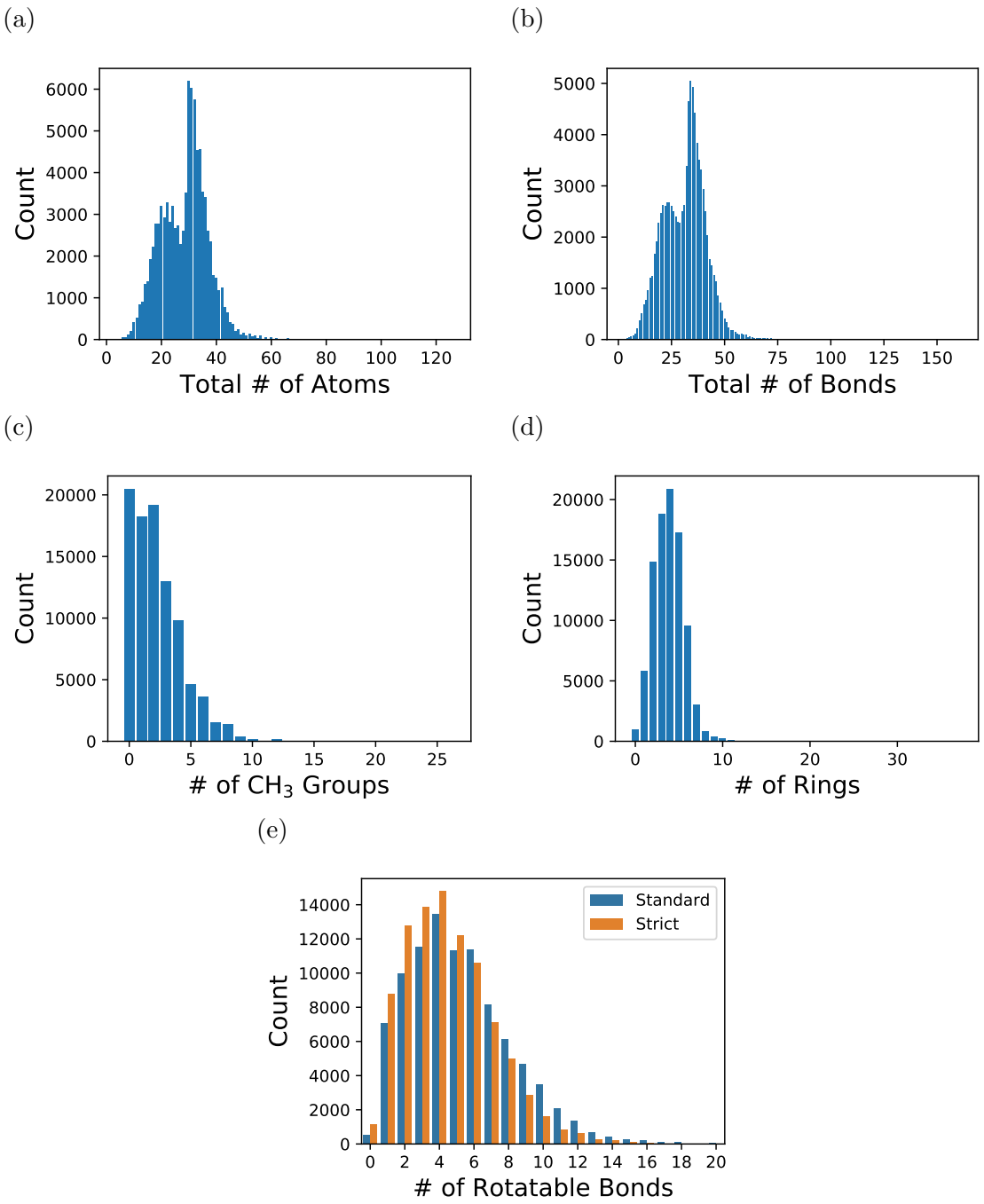


Figure S1: Histograms showing the number of molecules as a function of various descriptors, including: (a) the number of atoms including hydrogens, (b) number of bonds, (c) number of methyl, CH_3, groups, (d) number of rings, and (e) the number of rotatable bonds as calculated by RDKit including (standard, in blue) and excluding (strict, in orange) amides, esters, and other rotationally-restricted bonds.

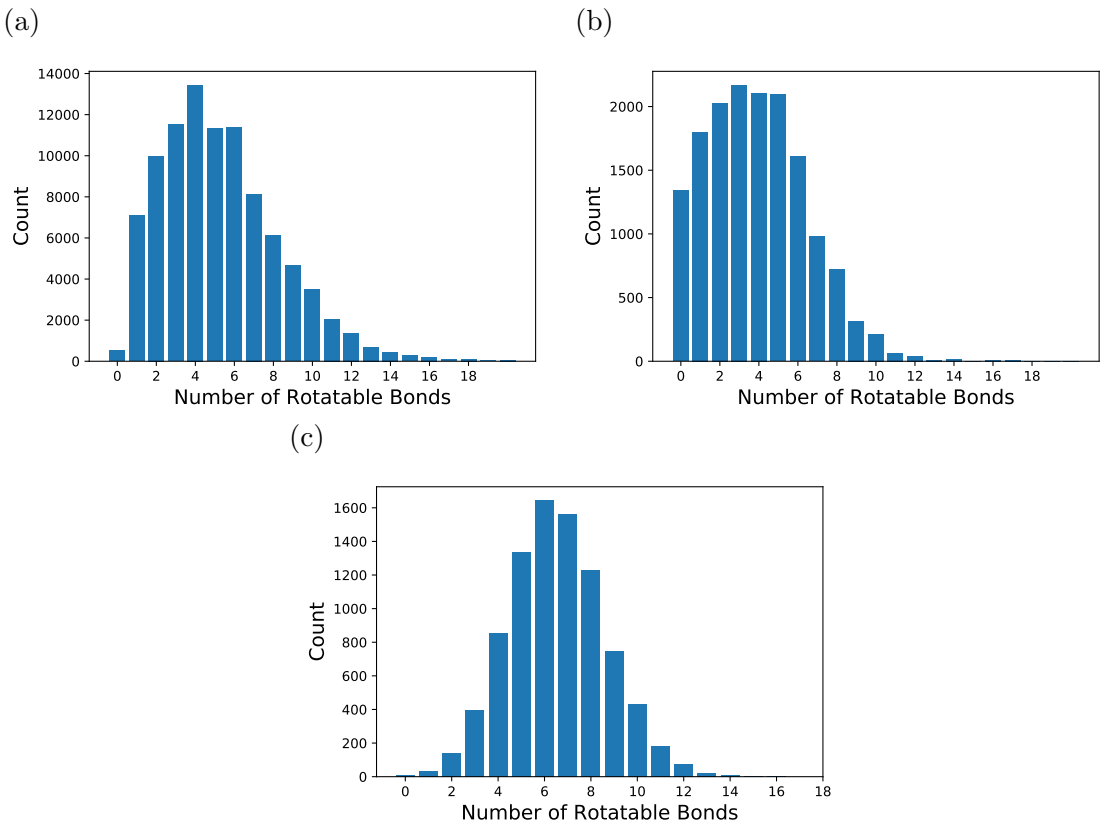


Figure S2: Number of rotatable bonds in training and testing sets. (a) Training set. (b) ZINC-I set. (c) CTPs set.

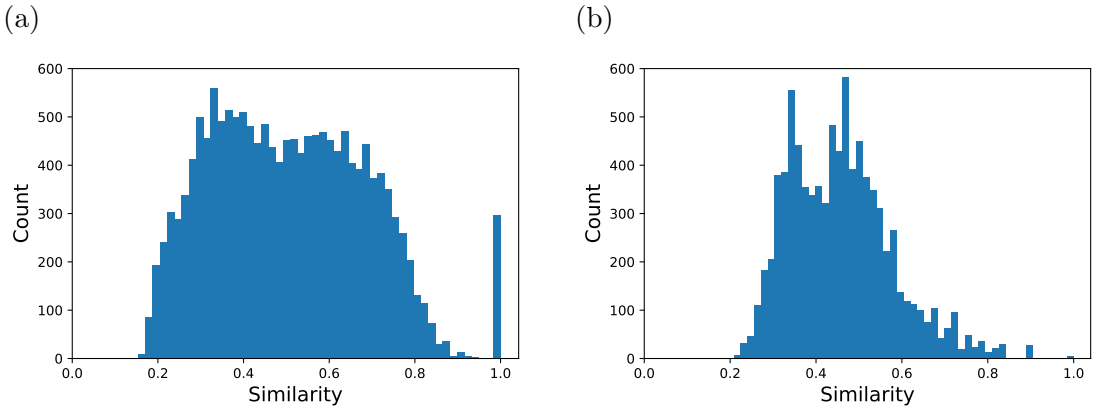


Figure S3: Similarity between training set and testing sets. (a) Training vs ZINC-I. (b) Training vs CTPs. Both testing sets contain molecules that are dissimilar with the one in training set.

Appendix B: Vibrational Entropy and Conformational Entropy

With proper consideration, the total entropy can break down into different components, including vibrartional entropy, translational entropy, rotational entropy and conformational entropy. Figure S4 shows the contributions of vibrational, translational, rotational and conformational entropies to the CREST-computed entropies. Vibrational entropy has the largest contribution, while the conformational entropy has the lowest contribution.

Vibrational entropy derives from thermal population of low-energy modes, often from "breathing" long-axis motions. The vibrational entropy should thus increase in proportion to molecular size, based on the number of atoms, number of bonds, and/or molecular weight. As illustrated in Figure S5, there is indeed a high correlation (R^2 of 0.897) between the number of atoms and the vibrational entropy.

Figure S6 shows that the vibrational entropies can be computed efficiently, with median 16 seconds per compound, while the median time for conformational entropy calculation is 1.01 hours per compound. Both calculation times show a power-law scaling with the number of atoms (see Figure S7).

Figure S8 reveals the conformer population size at different energy thresholds from the lowest energy, and it also exhibits a power-law fit with number of rotatable bonds. The coefficients and exponents of the power law fits are given in Table S2, and the best fit lines are shown in Figure S9.

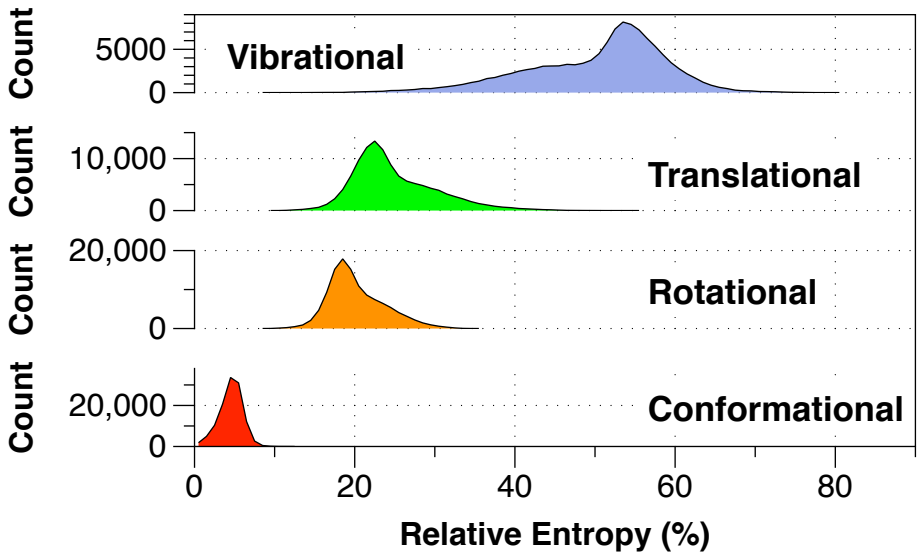


Figure S4: Distributions of CREST-computed entropies, as a percentage of the total entropy, assuming additivity, with vibrational entropy in blue, translational entropy in green, rotational entropy in orange, and conformational entropy in red.

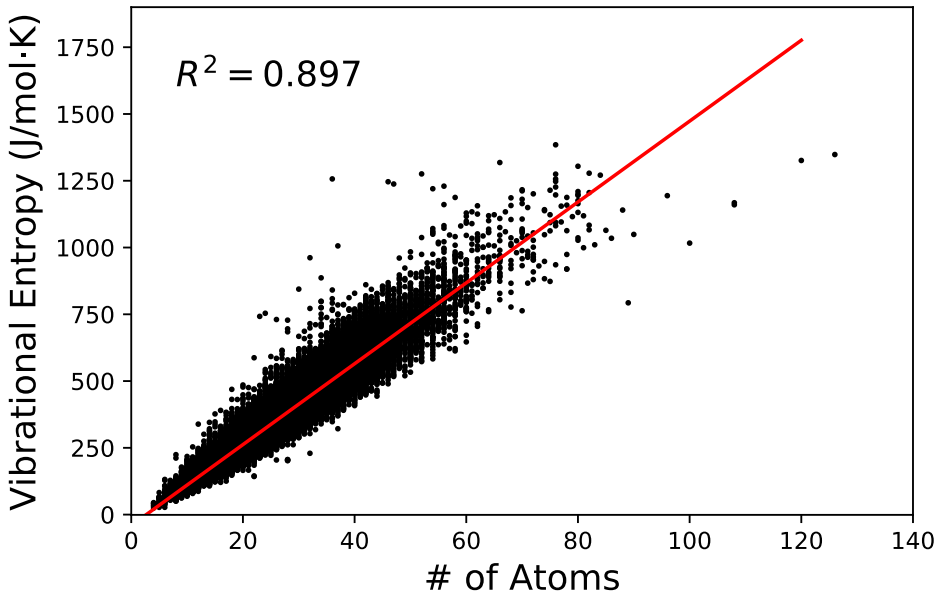


Figure S5: Correlation between number of atoms and GFN2-calculated vibrational entropies.

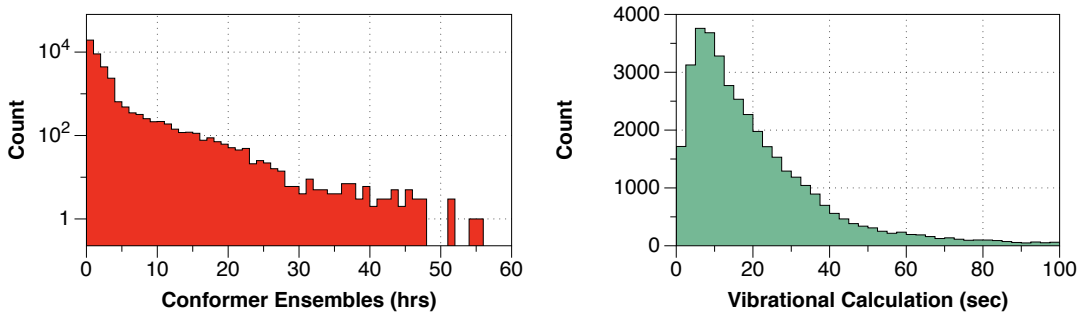


Figure S6: Histograms showing the counts of CREST dual-core run times (in hours, left) on a logarithmic scale; and GFN2 dual-core vibrational run times (in seconds, right).

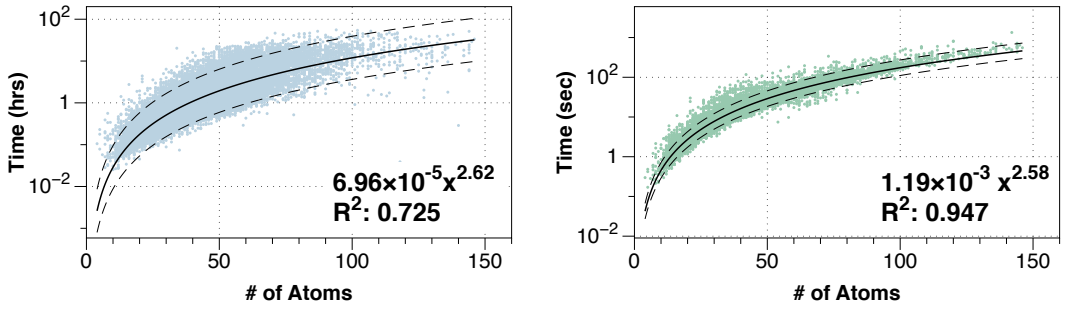


Figure S7: Variation in CREST dual-core run times (in hours, left) as a function of the number of atoms; and variation in GFN2 vibrational dual-core run times (in seconds, right) with the number of atoms, across a set of across 39,254 molecules run on the same hardware. Trendlines indicate power-law best fits with dashed lines indicating 95% confidence bands.

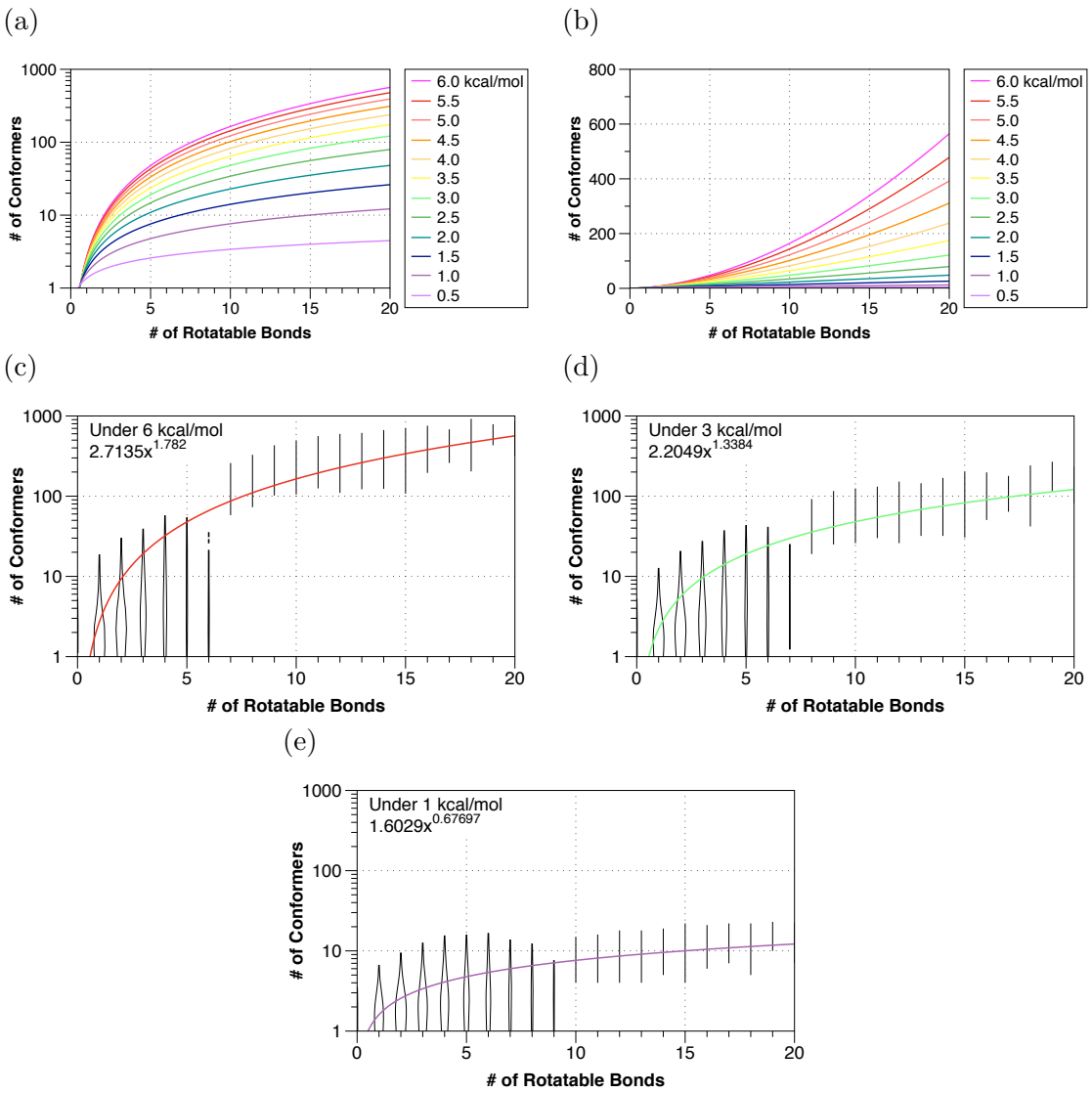


Figure S8: Predictions of the number of conformers as a function of the number of rotatable bonds and the energy difference from the minimum energy for that molecule, on (a) logarithmic, and (b) linear scales; and again, showing violin plots for conformers with minimum energy differences under (c) 6 kcal/mol, (d) 3 kcal/mol, and (e) 1 kcal/mol, respectively.

Table S2: Coefficients and exponents of power-law fits for predicting the number of conformers within a given energy threshold (in kcal/mol) of the minimum-energy conformer.

Threshold (kcal/mol)	**Coefficient**	**Exponent**
0.5	1.367	0.396
1.0	1.603	0.677
1.5	1.790	0.890
2.0	1.949	1.070
2.5	2.085	1.215
3.0	2.205	1.338
3.5	2.309	1.444
4.0	2.407	1.534
4.5	2.494	1.611
5.0	2.577	1.677
5.5	2.650	1.734
6.0	2.714	1.782

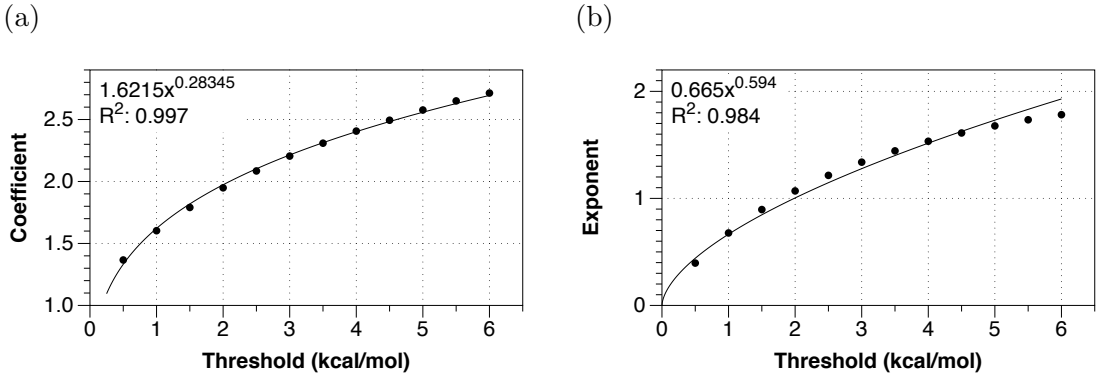


Figure S9: Coefficients and exponents for the number of thermally accessible conformers from Table S2 indicating best-fit lines, as a function of the energy threshold in kcal/mol.

Appendix C: Ring Flexibility and Foldability

We will introduce two new descriptors, namely *Total Ring Flexibility* and *Foldability*, for conformational entropy prediction. The definition of rotatable bond and the list of functional groups that used to construct these descriptors are listed in Table S3. The distance and angle constraints for the intramolecular hydrogen bonds and π-π stacking are listed in Table S4.

Table S3: Feature definitions used in Ring Flexibility and Foldability calculation. The definition of donors and acceptors are adapted from Gobbi and Poppinger.[S7]

Feature	SMARTS
Rotatable Bond	[!$(*#*)&!D1]-!@[!$(*#*)&!D1]
Donor	[$([N;!H0;v3,v4&+1]),$([O,S;H1;+0]),n&H1&+0]
Acceptor	[$([O,S;H1;v2;!$(*-*=[O,N,P,S])]),$([O,S;H0;v2]), $([O,S;-]),$([N;v3;!$(N-*=[O,N,P,S])]), n&H0&+0,$([o,s;+0;!$([o,s]:n);!$([o,s]:c:n)])]
Linear Amide	[NX3]!@;-[CX3]=[OX1]
Linear Ester	[O]!@;-[CX3]=[OX1]
Linear Thioamide	[NX3]!@;-[CX3]=[SX1]
Cyclic Amide	[NX3]@[CX3]=[OX1]
Cyclic Ester	[O]@[C](=O)
Cyclic Thioamide	[NX3]@[CX3]=[SX1]

Table S4: Distance and angle constraints used to determine the intramolecular interactions. The distance and angle constraints are adapted from Schrödinger.[S8]

Interactions	Distance and Angle Constraints
π-π stacking	i. Maximum distance between ring centroids of two rings is 4.4 Å. ii. Angles between the ring planes is less than 30°.
Hydrogen bonds	i. Maximum distance is 2.5 Å. ii. Minimum donor angle is 120°. iii. Minimum acceptor angle is 90°.

Ring Flexibility

The calculation of Ring Flexibility (R_f) requires the concept of unique ring families (URFs),[S9] which is used to describe the ring topologies in molecular graphs. The calculation of URFs comprises two parts: (i) calculation of Relevant Cycles (RCs), where RCs are defined as

the union of all minimum cycle bases; (ii) pairing of RCs if they are URF-pair-related (see Definition 1).

Definition 1. *Let C_1 and C_2 be two RCs in a molecular graph, G; then C_1 and C_2 are URF-pair related if and only if all of the following conditions hold:*

1. *$|C_1| = |C_2|$, i.e. the number atoms in each ring is the same;*
2. *$E(C_1) \bigcap E(C_2)$, i.e. the two rings share one or more bonds; and*
3. *There exists a set, S, of strictly smaller rings, c, in G such that $C_1 \bigoplus (\bigoplus_{c \in S} c) = C_2$.*

After the identification of all URFs, a subgroup ring flexibility score, R_f^{Group}, can be computed using Equation (1). Cremer and Pople[S10] showed the general conformation of a monocyclic N-membered ring can be fully described by $N-3$ parameters. This provides the basis of the calculation. Simple summation of the number of degrees of freedom of unique ring families (URFs) overestimate the flexibility of the ring systems, as the ring junctions in bicyclic and polycyclic rings impose additional constraints on the system. The number of degrees of freedom decrease as the number of ring junctions increases. We therefore introduce a penalty on the complexity of the ring systems. The penalties take ring junction types and bond orders into account (see Table S5). Figure S11 shows that the ring flexibility score has a good correlation with the conformational entropies of simple ring systems, giving a R^2 of 0.70.

The subgroup ring flexibility is defined as follows:

$$R_f^{\text{S}} = \frac{1}{R_S} \sum_{r=1}^{R_S} max(N_r - 3 - p, 0) \tag{1}$$

where r is the index of the ring in the subgroup S, and R_S is the total number of rings in the subgroup S. N_r is number of atoms in the ring r, and p is the penalty term.

The *Total Ring Flexibility* is defined as follows:

$$R_f^{\text{Total}} = \sum_{t=1}^{T} R_f^t \tag{2}$$

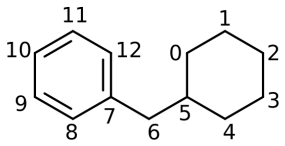
where t is the index of the subfamilies, and T is the total number of subfamilies in the molecule.

Note that *cis-trans* isomers in fused ring systems (*e.g.* *cis*-decalin and *trans*-decalin) will give different conformational entropy values, as the steric interactions in *cis* and *trans* configurations may vary. We do not penalize on *cis-trans* isomerism in our analysis.

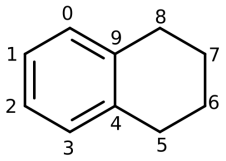
Table S5: Ring Penalties. A fused ring penalty is only applied to fused rings with shared atoms linked by a single bond. The definition of cyclic amides, esters and thioamides can be found in Table S3.

Condition	Penalty, p
Non-single bond	Count of non-single bonds
Spiro	Count of spiro atoms
Fused	Count of fused single bonds
Bridge	Count of bridgehead atoms
Polycyclic	Count of polycyclic atoms
Cyclic amide	Count of cyclic amides
Cyclic ester	Count of cyclic esters
Cyclic thioamide	Count of cyclic thioamides

(a)



(b)



(c)

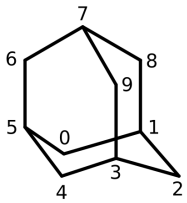


Figure S10: Example calculations of the total ring flexibility: (a) cyclohexylmethylbenzene; (b) 1,2,3,4-tetrahydronaphthalene; (c) adamantane. In (a), there are two URFs: URF0, (0,1,2,3,4,5); and URF1, (7,8,9,10,11,12). The subgroup ring flexibility scores are $R_f^0 = 3$ and $R_f^1 = 0$. The total ring flexibility score, R_f^{Total}, is 3. In (b) there are two URFs: URF0, (0,1,2,3,4,9); and URF1, (4,5,6,7,8,9). The corresponding subgroup ring flexibility scores are $R_f^0 = 0$ and $R_f^1 = 2$. The total ring flexibility score, R_f^{Total}, is 2. In (c), there are four URFs: URF0 (0, 5, 6, 7, 8, 1), URF1 (0, 1, 2, 3, 4, 5), URF2 (3, 4, 5, 6, 7, 9), and URF3 (1, 2, 3, 7, 8, 9). The total ring flexibility score, R_f^{Total}, is 0.

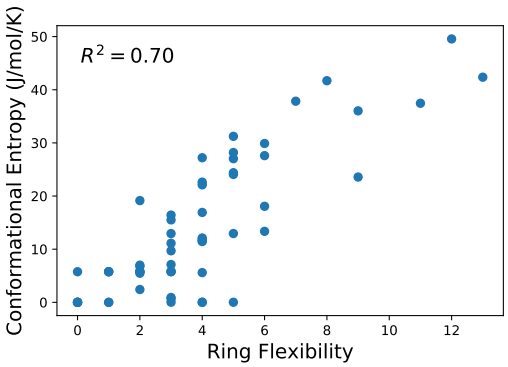


Figure S11: Correlation between conformational entropy and ring flexibility. Conformational entropies of 60 simple molecules, including unsubstituted monocyclic heterocycles, fused rings, bridged rings, spiros, and rings with simple substituents (*e.g.*, hydroxyl, carbonyl) are calculated. Note that these molecules do not contain any rotatable bonds and methyl groups. The ring flexibility shows a good correlation with conformational entropy, with $R^2 = 0.70$.

Foldability

Kuhn et al.[S11] systematically studied the topologies prone to intramolecular hydrogen bond formation, and suggested a number of motifs that are commonly observed in the X-ray crystal structures found in the Cambridge Structural Database (CSD) and Protein Data Bank (PDB). The distribution of distance and torsion angles were investigated in their work. Here, we attempted to generalise path topologies and provide insights into the small molecule equivalent of "folding". Two types of intramolecular interactions, hydrogen bonds and π-π stacking, were considered in our analysis. We applied the junction analysis to characterise the shortest path of these intramolecular interactions. Junction analysis has been carried out for antibody structures and their amino acid sequences.[S12] Given a path in a molecular graph, the path is aligned such that the terminal nodes are the atoms with desired intramolecular interaction features. For instance, we can treat the first atom as the hydrogen bond donor and the last atom as the hydrogen bond acceptor. Gaps are inserted in between when the path is shorter than maximum path length. This is not unlike protein loops of variable length between common secondary structural elements in proteins, in particular antibodies. This allows us to investigate the positional preferences of a functional group or particular atom types with this standardized atomic numbering system.

Hydrogen Bonds

There are approximately 36% of the data containing intramolecular hydrogen bonds. The definition hydrogen bond donor and acceptor in Table S3, and the distance and angle constraints for hydrogen bond in Table S4 are used in the analysis. Figure S12 shows that path length of four and five are commonly observed in our training dataset, and the minimum path length is three. Note that only heavy atoms are considered in the path analysis. Hence bond between donor atom and hydrogen is not counted. Long hydrogen bonds path is uncommon, and is used to assist the formation of other intramolecular interactions, including but not limited to CH-π interaction and π-π interaction. In this analysis, we considered the

path length up to nine (equivalently ten atoms in the path). We categorized the hydrogen bonds into six groups, based on the acceptor types. It includes heterocycle nitrogen, acyclic nitrogen, alkoxy group, terminal hydroxyl group, carbonyl (C=O) and sulfone (S=O). The six acceptor groups cover more than 90% of the observed intramolecular hydrgoen bonds, and carbonyl group is the most commonly observed.

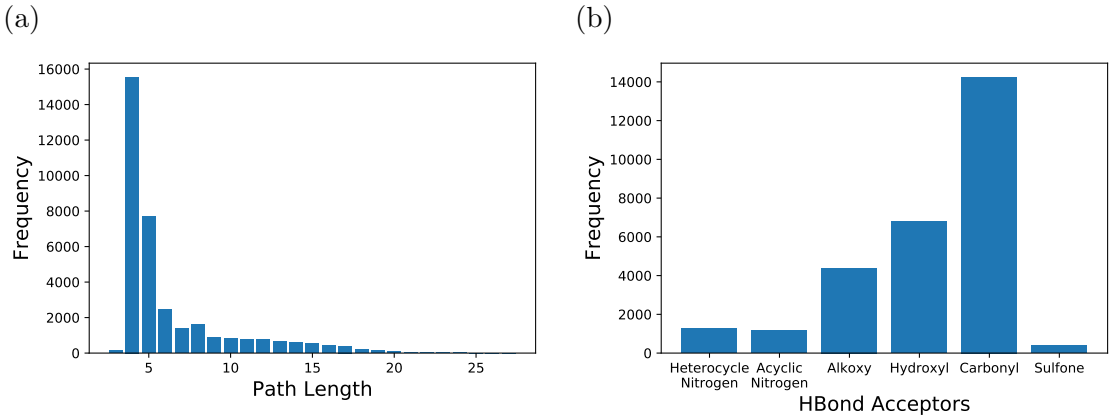


Figure S12: (a) Distribution of shortest path length between hydrogen bond donors and acceptors. Length of four and five are commonly occured in both COD and ZINC database. (b) Frequency of occurrence of six different hydrogen bond acceptors in the dataset. Carbonyl group is the most commonly observed, followed by hydroxyl and alkoxy group.

We applied the junction analysis to characterise the path between hydrogen bond donors and hydrogen bond acceptors. We considered the hybridization of the atoms and identified whether the atom in a ring. The hybridization of the atom is inherited from the definition in RDKit. Note that the oxygen in amide, and ether oxygen in ester group are classified as sp^2 hybridized. The path direction was unified such that the path begins with donor atom and ends with acceptor atom.

Carbonyl. Figure S13a shows that acyclic sp^2 and sp^3 hybridized oxygen and sp^2 hybridized nitrogen donors are commonly observed. It also shows a wide range of atom types along the path.

Hydroxyl. Figure S13b shows that hydroxyl group (sp^3 hybridized oxygen) primarily forms intramolecular hydrogen bond with another hydroxyl groups. It also shows a widely-varying

atom types along the path.

Alkoxy. The alkoxy receptor takes various forms, with one-third observed in ring form. Three-fourths of the oxygen in alkoxy group is sp^3 hybridized. Figure S13c shows that the donor type is primarily sp^3 oxygen, *i.e.* hydroxyl group. It also shows that the first, second and the penultimate atoms in the path are primarily carbon.

Acyclic Nitrogen. The path length for acyclic nitrogen receptor is shorter than that in other receptors, such as carbonyl, hydroxyl and alkoxy. The atom types along the path have less variation, and are primarily carbon atom.

Heterocycle Nitrogen. We categorised the heterocycle nitrogen into three classes, (i) sp^3 hybridized nitrogen acceptor with hydroxyl donor, (ii) sp^3 hybridized nitrogen acceptor with amine donor and (iii) sp^2 hybridized nitrogen acceptor. The variation in the paths for sp^3 nitrogen acceptor is small.

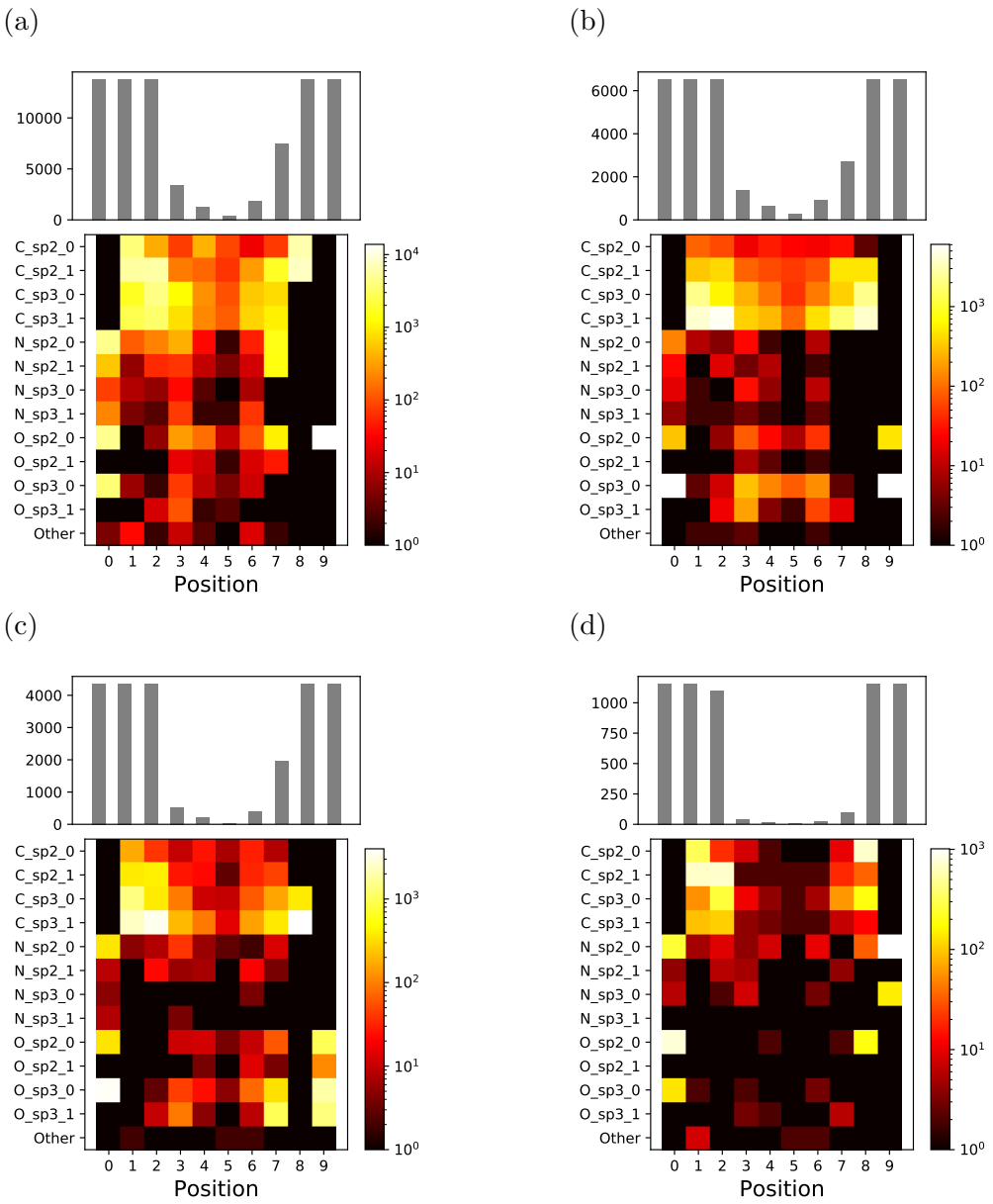


Figure S13: Positional analysis for a given intramolecular hydrogen-bond acceptor type, showing counts of molecules with each type of atom. Counts of molecules containing each kind of intramolecular hydrogen bond are shown as histograms above each heat map, where black represents zero, and color spectrum from dark red, red, orange, yellow to white maps to the maximum number of molecules in each category. Plots are shown for each type of H-bond acceptor: (a) carbonyl; (b) hydroxyl; (c) alkoxy; (d) acyclic nitrogen (e) heterocyclic sp^3 nitrogen (with nitrogen donor), (e) heterocyclic sp^3 nitrogen (with oxygen donor); (f) heterocycle sp^2 nitrogen. Positions 0 and 9 represent the donor and acceptor atoms respectively. The atom types are defined by their elements, hybridization and whether the atom is in ring, *e.g.,* "C_sp3_0" is an sp^3 carbon and is not in a ring. In general, the shortest path between the intramolecular hydrogen bonding heavy atoms for heterocyclic sp^3 nitrogen acceptors (e and f) is shorter than other acceptors, and the atoms types along the path are highly restricted. In contrast, we observe a widely-varying range of atoms types along the path between carbonyl and hydroxyl acceptors (a and b).

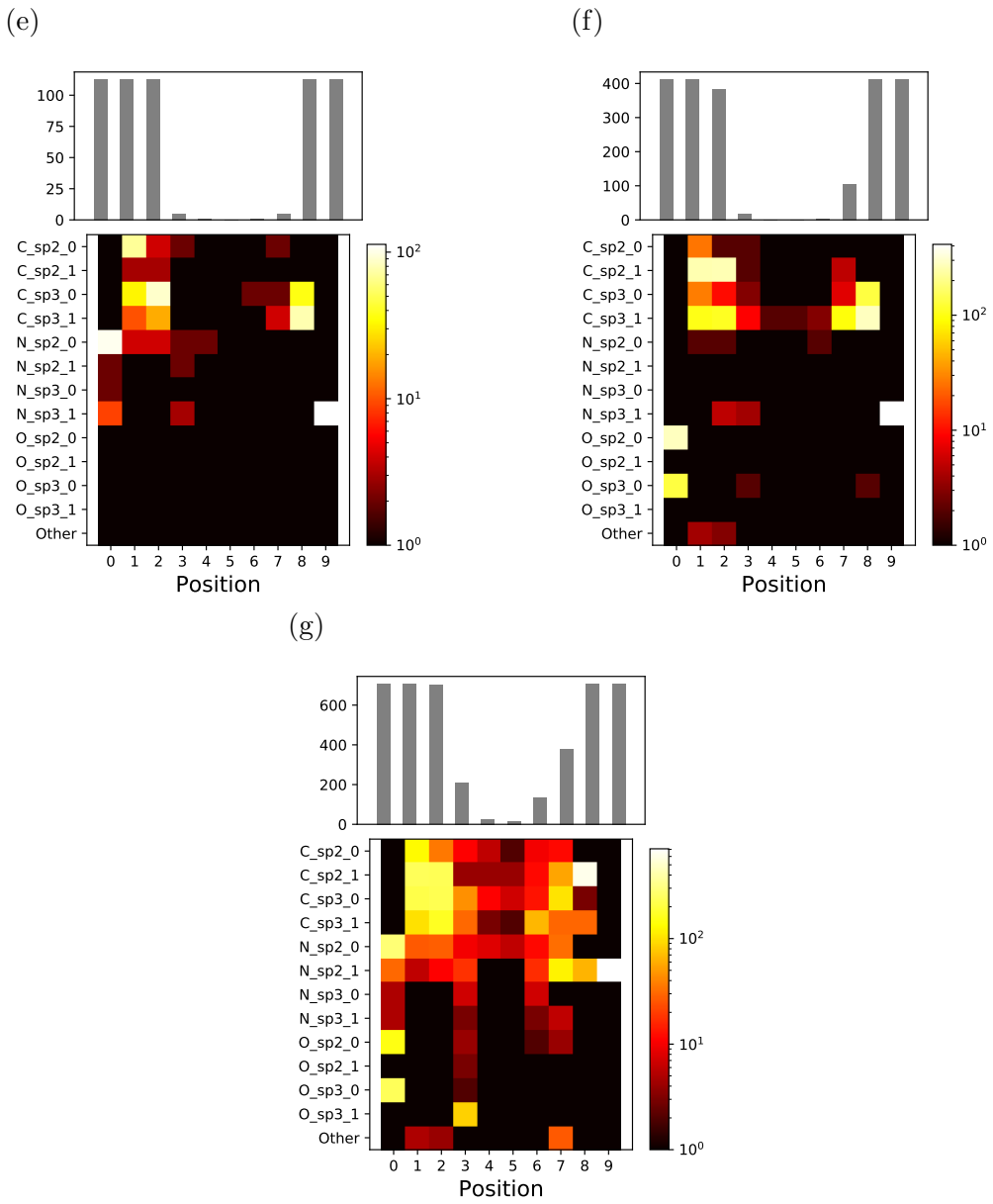


Figure S13: Positional analysis for a given intramolecular hydrogen-bond acceptor type, showing counts of molecules with each type of atom. Counts of molecules containing each kind of intramolecular hydrogen bond are shown as histograms above each heat map, where black represents zero, and color spectrum from dark red, red, orange, yellow to white maps to the maximum number of molecules in each category. Plots are shown for each type of H-bond acceptor: (a) carbonyl; (b) hydroxyl; (c) alkoxy; (d) acyclic nitrogen (e) heterocyclic sp^3 nitrogen (with nitrogen donor), (e) heterocyclic sp^3 nitrogen (with oxygen donor); (f) heterocycle sp^2 nitrogen. Positions 0 and 9 represent the donor and acceptor atoms respectively. The atom types are defined by their elements, hybridization and whether the atom is in ring, *e.g.,* "C_sp3_0" is an sp^3 carbon and is not in a ring. In general, the shortest path between the intramolecular hydrogen bonding heavy atoms for heterocyclic sp^3 nitrogen acceptors (e and f) is shorter than other acceptors, and the atoms types along the path are highly restricted. In contrast, we observe a widely-varying range of atoms types along the path between carbonyl and hydroxyl acceptors (a and b).

Face-to-Face and Parallel π-π stacking

Face-to-Face and parallel π-π stacking are commonly observed in our database. That said, T-shaped π-stacking is less frequently observed in our database, and therefore is not included in our analysis. Approximately 6.8% of the molecules in our training set exhibit π-π stacking. Figure S14a shows a widely-varying path length (number of bonds), ranging from three to thirty bonds. The median path length is six. In this analysis, we considered path lengths up to 15 (*i.e.* 16 atoms in the path). Six functional groups were considered in π-π stacking, including carbamate, urea, amide, ester, ketone and ether, which contributed 70% of the observed π-π stacking. Figure S14b shows that amide groups are the most commonly observed functional group, followed by ether and ester groups. We discuss the role each functional group plays in "folding" below:

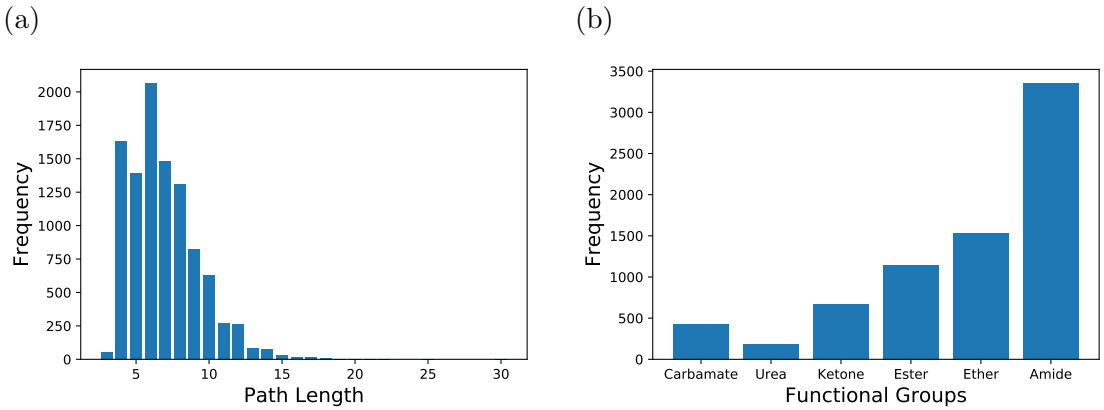


Figure S14: (a) Distribution of shortest path length between two aromatic rings. The path length varies widely, ranging from three to thirty. The median path length is six. (b) Frequency of occurrence of six different functional groups in the shortest path. Amide is the most commonly observed, followed by ether and ester groups.

Carbamate. Figure S15a shows that the position of carbamate carbon and the direction of the C−O bond are conserved. Its paired functional group, ester, is commonly observed near the other end, and the orientation of the C−O bond is also conserved.

Urea. We considered three forms of urea, including cyclic ureas, ureas with one nitrogen in a ring, and acyclic ureas. Figure S15b shows that the position of the cyclic urea is conserved.

Cyclic ureas and ureas with one nitrogen in a ring are usually paired with an amide (see Figures S15b and S15c). The variation in the position of amide carbon is small and the relative orientation of the C−N bond is conserved. The peaks at position 2 and position 13 in Figure S15d suggests that acyclic ureas are occasionally found in pairs.

Ketone. Ketone groups are commonly found in a short path (*i.e.* path with five to seven atoms). Figure S15e shows that the ketone carbon typically comes after the aromatic rings. For longer paths, it tends to be found in the middle of the path and form intramolecular hydrogen bonds, to support longer range π-π stacking and other intramolecular interactions. It pairs with other functional groups, including ether and ester. The ether oxygen is normally found next to the terminal aromatic rings.

Ester. Figure S15f shows that the ester carbon is found near the terminal aromatic rings, however the orientation of C−O bond depends on the position. It normally pairs with sulfonamide groups, and the sulfur atom is linked to the terminal aromatic rings. It can also be paired with ether, however the location of ether oxygen varies.

Ether. The ether oxygen is often observed from first to third position of the path, and it pairs with amide. The orientation of C−N bond is conserved.

Amide. We considered three forms of amide, including acyclic amides, amides with a nitrogen in a ring and cyclic amides. We divided the path into two types, path with and without ring bonds. Figure S15h to S15k shows that the amide carbon position is predominantly found at the first, second or third positions, however, the orientation of C−N bond is position dependent.

To assess the generalizability of these positional preferences, we compared the positional preference of amide carbon in the training set and also in the peptides set. Figure S16 shows that the amide position is under-represented in the training set, and this suggests that different positional model may be required for peptides.

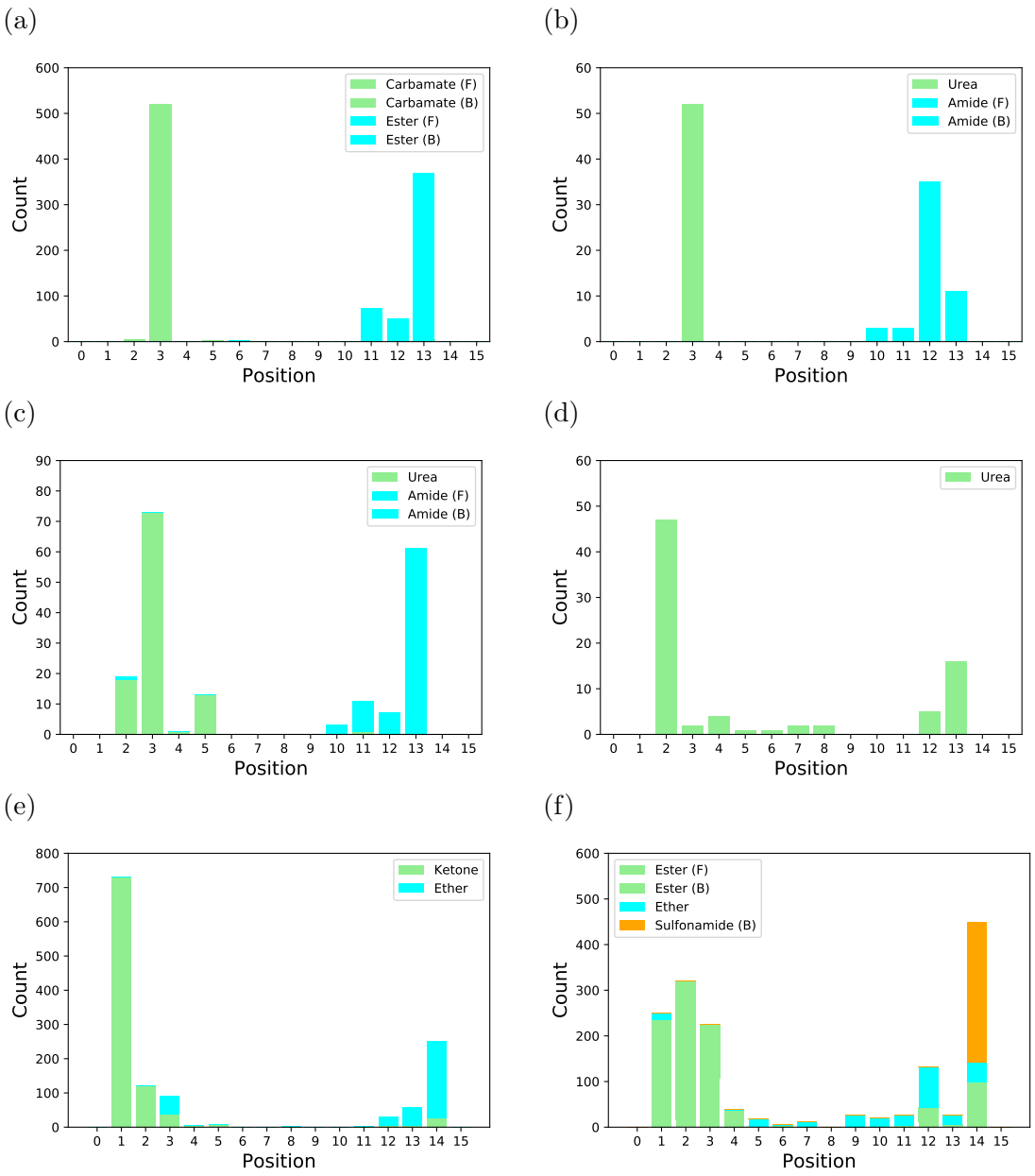


Figure S15: Counts of functional groups by position in the training set. Atoms 0 and 15 belong to the aromatic rings. The functional groups are: (a) carbamate; (b-d) urea; (e) ketone; (f) ester; (g) ether; and (h-k) amide. The orientations of the C−O bond in carbamates and esters, the S−N bond in sulfonamides, and the C−N bond in amides are specified thus: (F) indicates that the oxygen or nitrogen atoms are *followed* by the carbon or sulfur atom, while (B) indicates that the oxygen or nitrogen atoms comes *before* the carbon or sulfur atom. Three forms of urea are considered: (i) a urea in a ring; (ii) one nitrogen in a ring; and (iii) acyclic ureas. Four forms of amide are considered: type 0: acyclic amides with no ring bond along the path; type 1: cyclic amides; type 2: nitrogen in a ring; and type 3: acyclic amides with one or more ring bonds along the shortest connecting path. The positions of carbamate, urea, and ketone groups tend to be conserved, while amide positions show a huge variation.

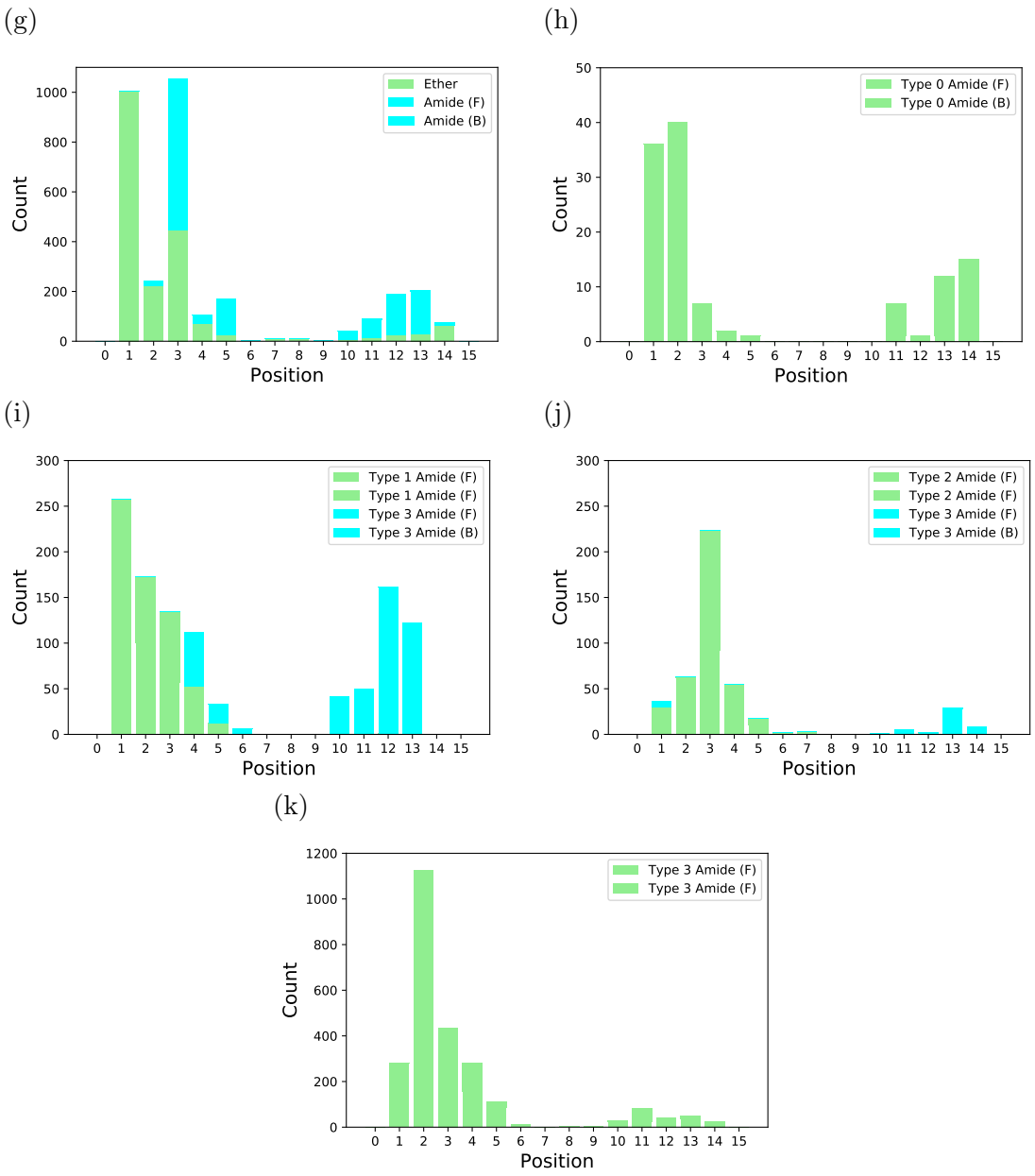


Figure S15: Counts of functional groups by position in the training set. Atoms 0 and 15 belong to the aromatic rings. The functional groups are: (a) carbamate; (b-d) urea; (e) ketone; (f) ester; (g) ether; and (h-k) amide. The orientations of the C−O bond in carbamates and esters, the S−N bond in sulfonamides, and the C−N bond in amides are specified thus: (F) indicates that the oxygen or nitrogen atoms are *followed* by the carbon or sulfur atom, while (B) indicates that the oxygen or nitrogen atoms comes *before* the carbon or sulfur atom. Three forms of urea are considered: (i) a urea in a ring; (ii) one nitrogen in a ring; and (iii) acyclic ureas. Four forms of amide are considered: type 0: acyclic amides with no ring bond along the path; type 1: cyclic amides; type 2: nitrogen in a ring; and type 3: acyclic amides with one or more ring bonds along the shortest connecting path. The positions of carbamate, urea, and ketone groups tend to be conserved, while amide positions show a significant variation.

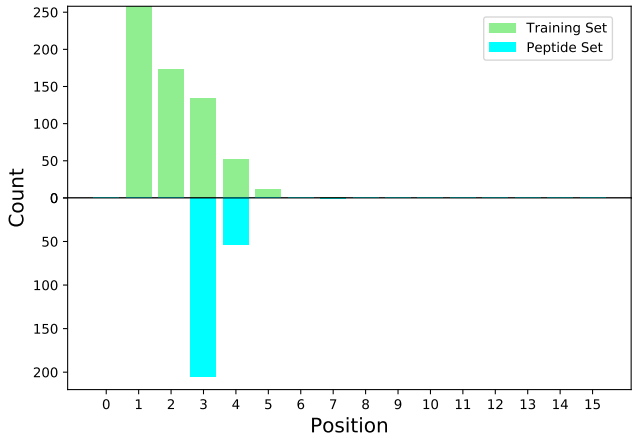


Figure S16: Variation in the amide positional preferences in the training set and peptide set.

This analysis provides insights into the topology of the intramolecular hydrogen bond and π-π stacking. We set thresholds (see Table S6) for all functional groups and atom types at given positions, and the over-represented groups (above the threshold) are then used to identify potential interactions in the unseen dataset. To compute the foldability score (F), we calculated the sum of the number of rotatable bonds involved in all potential paths. To overcome the existence of multiple hydrogen bond paths, we calculated a score, W, for each rotatable bonds in the molecule. It measures how likely the rotatable bond is "locked" by the intramolecular hydrogen bonds. The foldability score for intramolecular hydrogen bonds is simply the sum of the scores of all rotatable bonds (see Equation 4). Similarly, we compute the sum of the number of rotatable bonds involved in all π-π interaction paths (see Equation 5).

Table S6: Position and Count Thresholds

Intramolecular Interactions	**Thresholds**
Hydrogen Bond	(i) >= 10% of the total count at a given position. (ii) minimum 100 observations at a given position.
π-π Stacking	(i) >= 10% of the total count at a given position (ii) minimum 50 observations at a given position.

$$W_r = \sum_{i=1}^{K} \sum_{j=1}^{L_i} w_r^{ij} \mathbb{1}_{r \in ij} = \sum_{i=1}^{K} \sum_{j=1}^{L_i} \frac{1}{L_i} \mathbb{1}_{r \in ij} \tag{3}$$

$$F_{HBond} = \sum_{r=1}^{R} \min(W_r, 1) \tag{4}$$

$$F_{\pi-\pi} = \sum_{p=1}^{p} N_{rotor}^{P} \tag{5}$$

where K, L_i and R are the number of donors in the molecule, the number of possible acceptors interact with a fixed donor i, and the number of rotatable bonds respectively. W_r is the score for rotatable bonds r. w_r^{ij} is the weight of rotatable bond r in the path from donor i to acceptor j. $\mathbb{1}_{r \in ij}$ is a indicator function, and it is 1 if the rotatable bond r is in the path from donor i to acceptor j. N_{rotor}^{p} is the number of rotatable bonds found in the shortest path between two aromatic rings that form π-π stacking. P is the total number of possible π-π interactions in the molecule.

Appendix D: Model Diagnostic and Comparison

The explanatory variables are denoted as follows: number of rotatable bonds: N_{rotor}; number of methyl groups: N_{Methyl}; counts of specified functional groups (amide, ester and thioamide): N_{SG}; total ring flexibility: R_f^{Total}; and foldability: F. The linear regression parameters are represented by β, and S_{conf} is the conformational entropy. We denote the error term by ϵ, and $\epsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$, where σ^2 is a constant.

Model Diagnostics

Model 1

$$S_{conf} = \beta_0 + \beta_1 \log(N_{\text{rotor}} + 1) + \epsilon \tag{6}$$

From Figure S17a, we have no evidence that the normality assumption is violated. However, Figure S17b shows that the residuals tend to be positive at small fitted values, and negative at high fitted values. This suggests that the assumption of homoscedasticity does not hold. Since the model assumptions do not hold, we do not comment on the model and the results of the hypothesis tests are shown in Table S7.

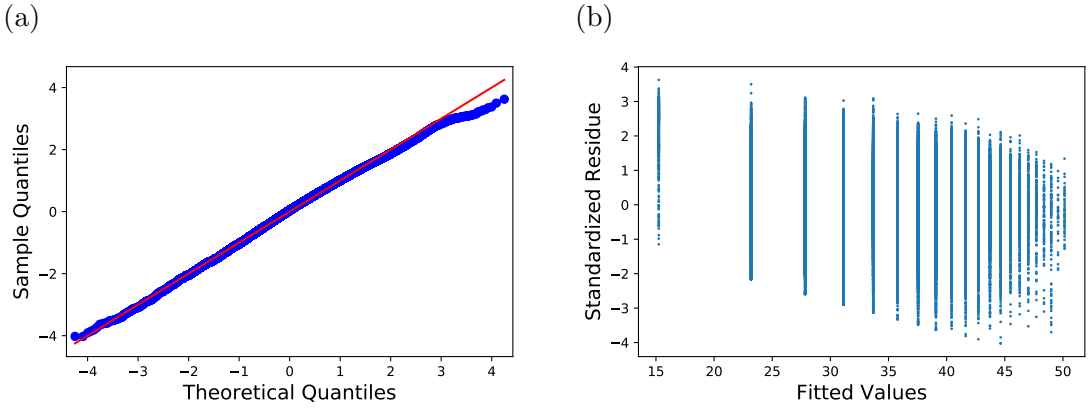


Figure S17: Model diagnostics: (a) Q-Q plot; (b) residual plot. The Q-Q plot shows a good fit in general. Figure (b) shows that the residuals tend to be positive at small fitted values, and negative at high fitted values.

Table S7: Model Summary

	Estimate	std. Error	t	Pr(> | t |)
Intercept	15.24	0.12	124.82	0.000
$\log(N_{rotor}+1)$	11.46	0.07	167.83	0.000
$R^2 = 0.232$				

Model 2

$$S_{conf} = \beta_0 + \beta_1 \log(N_{rotor}+1) + \beta_2 \log(N_{Methyl}+1) + \epsilon \tag{7}$$

From the Q-Q plot in Figure S18a, we do not have strong evidence that the normality of data is violated. Figure S18b shows a wider spread of residuals at small fitted values. This suggests that the homoscedasticity assumption may not hold. Similarly, we do not comment on the model and results of the hypothesis test shown in Table S8 as the model assumptions do not hold.

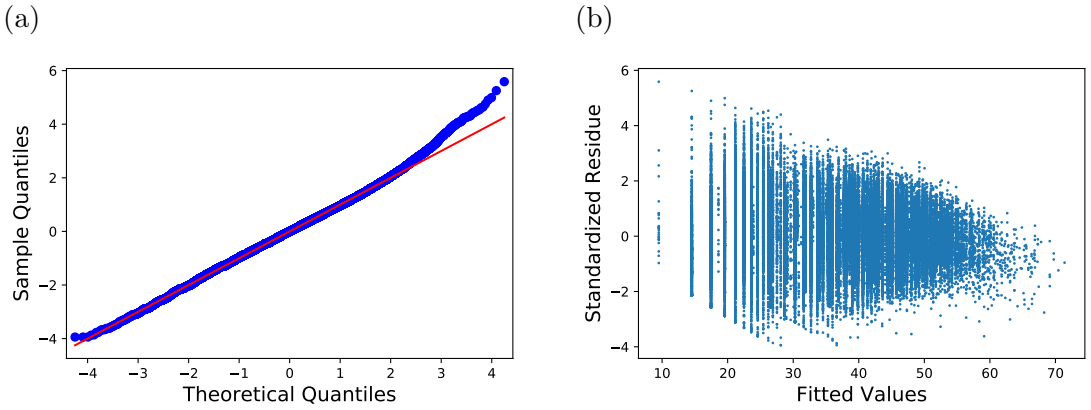


Figure S18: Model Diagnostics: (a) Q-Q plot; (b) residual plot. The Q-Q plot shows a good fit in general, with a short light tail on the right; panel (b) shows that the variation in residuals is larger at small fitted values.

Table S8: Model Summary

	Estimate	std. Error	t	Pr(> | t |)
Intercept	9.49	0.08	120.99	0.00
$\log(N_{\text{rotor}}+1)$	7.26	0.04	163.18	0.00
$\log(N_{\text{Methyl}}+1)$	13.13	0.04	375.88	0.00
$R^2 = 0.695$				

Model 3

$$S_{conf} = \beta_0 + \beta_1 \log(N_{rotor} + 1) + \beta_2 \log(N_{Methyl} + 1) + \beta_3 \log(N_{SG} + 1) + \epsilon \tag{8}$$

From the Q-Q plot in Figure S19a, we do not have strong evidence the normality of data is violated. Figure S19b shows a wider spread of residuals at small fitted values, and the variation decreases as the fitted value increases. This suggests that the homoscedasticity assumption may not hold. Similarly, we do not comment on the model and results of the hypothesis tests shown in Table S9 as the model assumptions do not hold.

(a) (b)

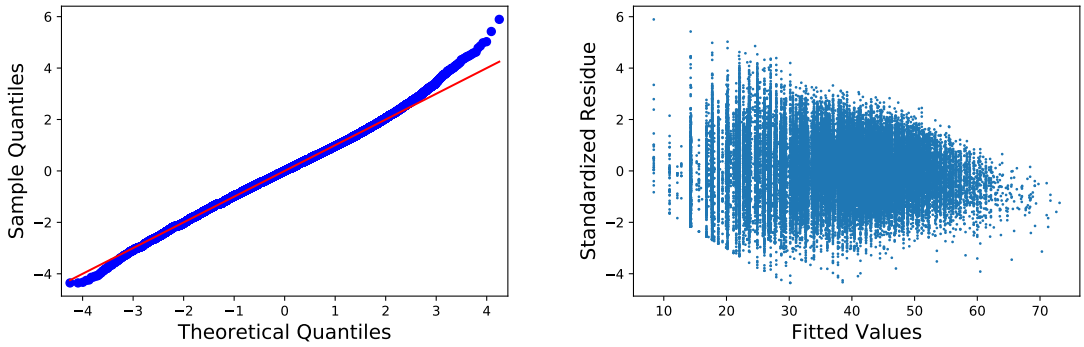


Figure S19: Model Diagnostic: (a) Q-Q plot. (b) residual plot. The Q-Q plot shows a goodfit in general. Figure (b) shows wider spread of the residuals at small fitted values.

Table S9: Model Summary

	Estimate	std. Error	t	Pr(> | t |)
Intercept	8.40	0.08	107.56	0.000
$\log(N_{\text{rotor}} + 1)$	8.49	0.05	181.07	0.000
$\log(N_{\text{Methyl}} + 1)$	13.03	0.03	382.06	0.000
$\log(N_{\text{SG}} + 1)$	-3.06	0.04	-69.21	0.000
$R^2 = 0.710$				

Model 4

$$S_{conf} = \beta_0 + \beta_1 \log(N_{rotor} + 1) + \beta_2 \log(N_{Methyl} + 1) + \beta_3 \log(N_{SG} + 1) + \beta_4 \log(F_{HBond} + 1) + \beta_5 \log(F_{\pi-\pi}) + \epsilon \tag{9}$$

From the Q-Q plot in Figure S20a, we do not have strong evidence that the normality of data is violated. Figure S20b shows an even spread of the residuals, with some outliers lying above. In general, we do not have strong evidence that the homoscedasticity assumption is violated. As expected, the parameters associated with the functional groups counts and the intramolecular interactions are negative, and the small p-values suggests that these parameters are significantly different from zero. This model also shows a high coefficient of determination ($R^2 = 0.715$), which indicates a good fit of the data.

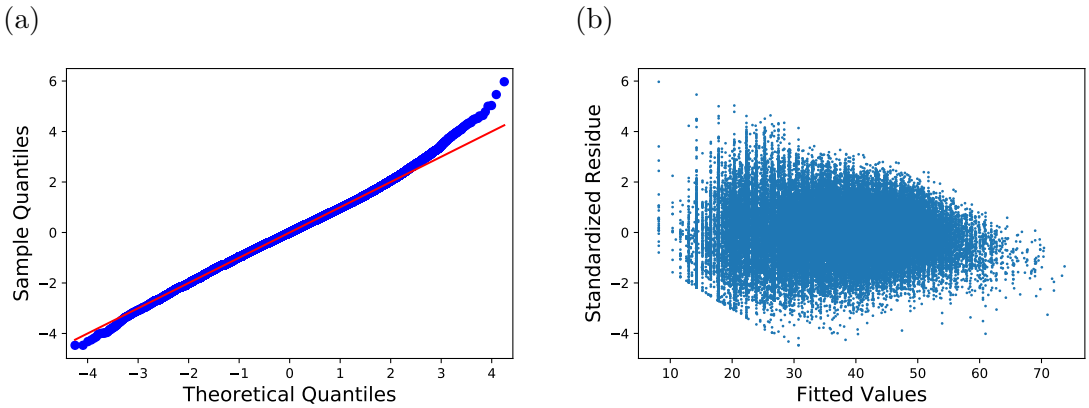


Figure S20: Model Diagnostic: (a) Q-Q plot. (b) residual plot. The Q-Q plot shows a goodfit in general. Figure (b) shows even spread of the residuals.

Table S10: Model Summary

	Estimate	std. Error	t	Pr(> | t |)
Intercept	7.97	0.08	101.61	0.000
$\log(N_{\text{rotor}} + 1)$	9.06	0.05	182.92	0.000
$\log(N_{\text{Methyl}} + 1)$	12.98	0.03	381.26	0.000
$\log(N_{\text{SG}} + 1)$	-2.42	0.05	-50.38	0.000
$\log(F_{\text{HBond}} + 1)$	-1.77	0.05	-33.79	0.000
$\log(F_{\pi-\pi} + 1)$	-0.62	0.03	-18.52	0.000
$R^2 = 0.715$				

Model 5

$$S_{conf} = \beta_0 + \beta_1 \log(N_{rotor} + 1) + \beta_2 \log(N_{Methyl} + 1) + \beta_3 \log(N_{SG} + 1) + \beta_4 \log(F_{HBond} + 1) + \beta_5 \log(F_{\pi-\pi}) + \beta_6 \log(R_f^{Total} + 1) + \epsilon \quad (10)$$

From the Q-Q plot in Figure S20a, we do not have strong evidence that the normality of data is violated. Figure S20b shows even spread of the residuals, and we do not have strong evidence that the homoscedasticity assumption is violated. As expected, the parameters associated with the functional group counts and the intramolecular interactions are negative, and small p-values in Table S11 provide strong evidence that the parameters significantly differ from zero. The parameter associated with the ring flexibility shows a negative value, which indicates that the output (*i.e.*, conformational entropy) decreases as the ring flexibility increases. This is inconsistent with the empirical study shown in S11. It suggests that the ring flexibility descriptor has to be reassessed carefully.

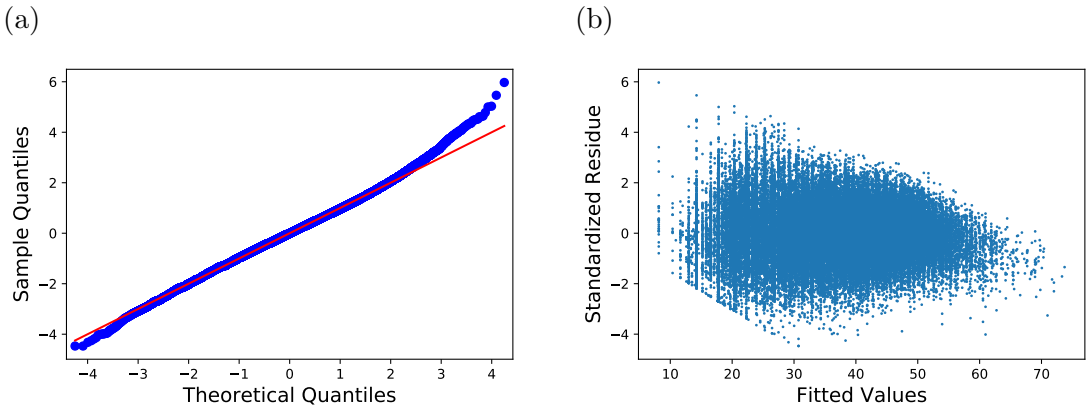


Figure S21: Model Diagnostic: (a) Q-Q plot. (b) residual plot. The Q-Q plot shows a goodfit in general. Figure (b) shows even spread of the residuals.

Table S11: Model Summary

	Estimate	std. Error	t	Pr($>\mid t \mid$)
Intercept	8.15	0.08	99.44	0.000
$\log(N_{\text{rotor}}+1)$	9.02	0.05	180.93	0.000
$\log(N_{\text{Methyl}}+1)$	13.03	0.04	375.39	0.000
$\log(N_{\text{SG}}+1)$	-2.44	0.05	-50.77	0.000
$\log(F_{\text{HBond}}+1)$	-1.72	0.05	-32.47	0.000
$\log(F_{\pi-\pi}+1)$	-0.59	0.03	-17.34	0.000
$\log(R_f^{Total}+1)$	-0.23	0.03	-7.51	0.000
$R^2 = 0.715$				

Models Performance

Table S12: Model Performance. Mean absolute error (MAE) between model-predicted conformational entropies and CREST/GFN2-computed conformational entropies on the training set and two independent test sets, the ZINC-I test set and the cyclic tetrapeptides (CTPs) test set, are reported. Note that the training and testing errors of the linear models (LR1-3) are included for reference.

	LR1	LR2	LR3	LR4	LR5(LR-Best)	Lasso	Ridge	KRR	DNN
Training Set	8.665	5.337	5.193	5.159	5.155	5.550	4.949	5.904	7.324
ZINC-I Set	8.832	4.887	4.793	4.758	4.767	5.472	5.291	5.865	7.221
CTPs Set	9.004	4.623	5.049	4.704	4.654	6.764	5.826	8.786	6.762

Appendix E: Model Implementation

Linear Regression

Number of rotatable bonds (see Table S3), number of methyl groups, functional groups (amide, ester and thioamide) counts, total ring flexibility and foldability are the variables for the linear regression models. Five models (LR1-5) are constructed with different number of variables. The simplest model (LR1) is a single-variable model, with the number of rotatable bonds as the sole explanatory variable. The full model (LR5) includes all the descriptors mentioned before. We use statsmodels[S13] to estimate parameters and analyze the models.

LASSO and Ridge Regression

ECFP6 with 4096 bits is used as model inputs. The hyperparameter (penalty), α, is optimized by 3-fold cross-validation with grid-search over a parameter grid, $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0\}$.

The implementation of ECFP6 in RDKit is used. The implementation of LASSO (least absolute shrinkage and selection operator), ridge regression and cross-validation in Scikit-learn[S14] are used.

Kernel Ridge Regression (KRR)

ECFP6 with 4096 bits is used as the model's inputs. The Tanimoto kernel (Jaccard kernel)[S5] is used in KRR. Nyström approximation[S15] is used to approximate the feature map. The hyperparameter α is optimized by 3-fold cross-validation with a grid search over a grid $\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$.

The implementation of ECFP6 in RDKit is used. The implementation of cross-validation in scikit-learn is used.

Neural Network

ECFP6 with 4096 bits is used as the model's inputs. ReLu activation functions are used in the hidden layers of our neural network model. A linear activation function is used for the final layer. ADAM optimizer is used for the optimization. We also add elastic net regularization in the hidden layers, with $L1 = 1e-4$ and $L2 = 1e-4$. The hyperparameters, *i.e.* number of hidden nodes, number of layers and dropout rate are optimized by cross-validation with grid search over a parameter grid: $N_{hidden} \in \{32, 64\}$, $N_{Layer} \in \{3, 4, 5\}$) and dropout rate $\in \{0.5, 0.6\}$.

The implementation of ECFP6 in RDKit is used. Keras[S16] is used to construct the neural networks.

Acknowledgement

GRH thanks the National Science Foundation (CHE-1800435) for support. GMM thanks the EPSRC and MRC for financial support under grant number EP/L016044/1. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work, the University of Pittsburgh Center for Research Computing through the computational resources provided, and using resources provided by the Open Science Grid,[S17,S18] which is supported by the National Science Foundation award 1148698, and the U.S. Department of Energy's Office of Science. G.R.H. thanks Lillian Chong and Jennifer Laaser for helpful discussions. L.C. thanks Catherine Wong for helpful discussions.

References

(S1) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography* **2009**, *42*, 726–729.

(S2) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **2012**, *40*, D420–D427.

(S3) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

(S4) Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org`, 2019.

(S5) Ralaivola, L.; Swamidass, S.; Saigo, H.; Baldi, P. *Neural Networks* **2005**, *18*, 1093–1110.

(S6) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(S7) Gobbi, A.; Poppinger, D. Genetic optimization of combinatorial libraries. *Biotechnology and Bioengineering* **1998**, *61*, 47–54.

(S8) Schrodinger, Schrodinger. `https://www.schrodinger.com/kb/1556`.

(S9) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *Journal of Chemical Information and Modeling* **2012**, *52*, 2013–2021.

(S10) Cremer, D.; Pople, J. A. General definition of ring puckering coordinates. *Journal of the American Chemical Society* **1975**, *97*, 1354–1358.

(S11) Kuhn, B.; Mohr, P.; Stahl, M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2010**, *53*, 2601–2611.

(S12) Monod, M. Y.; Giudicelli, V.; Chaume, D.; Lefranc, M.-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics* **2004**, *20*, i379–i385.

(S13) Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference. 2010.

(S14) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(S15) Williams, C. K. I.; Seeger, M. *Advances in Neural Information Processing Systems 13*; MIT Press, 2001; pp 682–688.

(S16) Chollet, F. Keras. `https://keras.io`, 2015.

(S17) Pordes, R. et al. The Open Science Grid. Journal of Physics: Conference Series. 2007; p 012057.

(S18) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Wurthwein, F. The Pilot Way to Grid Resources Using glideinWMS. 2009 WRI World Congress on Computer Science and Information Engineering. 2009; pp 428–432.