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 gives 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.

1 Introduction

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.[7–11]

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,[12–14] and using fast molecular mechanics force fields for energy evaluations – even if they do not always correlate well with more accurate electronic structure methods.[15–17]

Recent improvements in density functional tight-binding approximations[18–21] 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)[22] single-point energies.[16] Since the GFN2 method is applicable to a wide range of elements, compounds were drawn from the Crystallographic Open Database (COD)[23,24] as well as more complex organic macrocycles from the ZINC database.[25] 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).

2 Computational Methods

2.1 Data

Molecules with twenty or fewer rotatable bonds from the Crystallography Open Database (COD)[23,24] and ZINC[25] 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).

In all cases, molecular geometries were optimized using the GFN2 method,[18,19] followed by conformer sampling using the iterative metadynamic sampling and genetic crossover (iMTD-GC) method implemented in the CREST program,[20,21] including additional geometry optimization of the final conformer ensemble. The concept is similar to other efforts to sample conformational minima.[11] Note that the CREST calculation may break molecules into fragments; those molecules that were fragmented in the final output were excluded from our analysis.

The lowest energy conformer was selected for calculating the vibrational modes to evaluate standard rigid rotor harmonic oscillator vibrational, translational, and rotational entropies.[26]

All entropies were computed at fixed temperature $T = 298.15$ K in this analysis. We can therefore compare the magnitudes and relative distributions of the GFN2-calculated component entropies (see Figure 1).

The component entropies (translation, rotational, vibrational, conformational) in Figure 1 are not strictly 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.[3,4,27–29] Still, vibrational entropies generally contribute the greatest fraction of the total molecular entropy, followed by translational and rotational entropies, respectively. The median conformational entropy comprises $36.3\,\mathrm{J/mol\cdot K}$, or $\approx 2.6\,\mathrm{kcal/mol}$ at $300\,\mathrm{K}$ and while relatively small, should not be neglected.

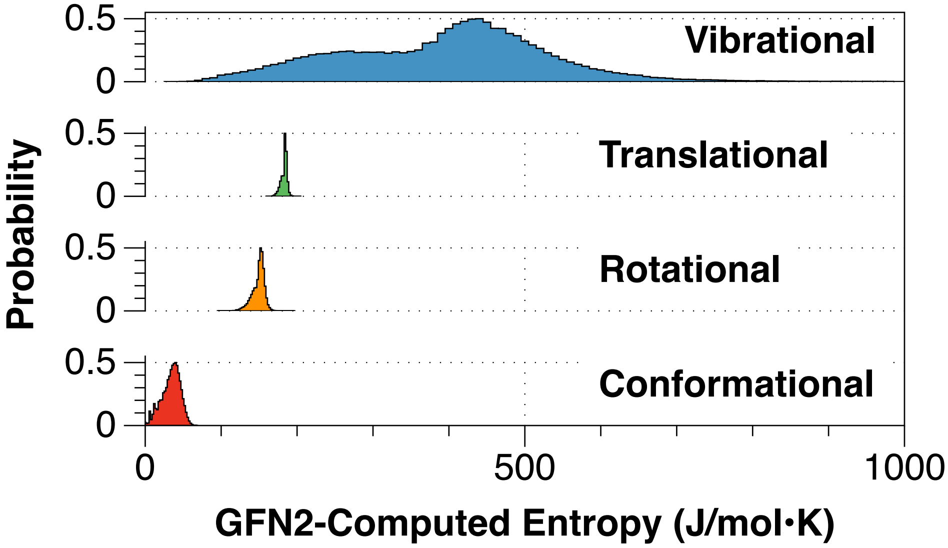


Figure 1: Distributions of GFN2-calculated vibrational, translational, rotational, and conformational entropies across the molecules studied.

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).

2.2 Models

We developed machine learning models using Extended Circular Fingerprints with a diameter 6 (ECFP6),[30] as well as continuous data-driven descriptors (CDDD).[31] The implementation of ECFP6 in RDKit[32] was used. The pretrained model as implemented by Winter et al. was used to generate the CDDD features. Note that we did not apply any preprocessing step on SMILES strings when generating the CDDD features. We used the Scikit-Learn[33] implementations of LASSO regression, ridge regression, kernel ridge regression, and cross-validation. Keras[34] was used to train the neural network. We also included an end-to-end molecular graph convolutional neural network[35] as implemented in DeepChem[36] for comparison. 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[37] 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.

3 Results and Discussion

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).[13] 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.

3.1 Rotatable Bond Dihedrals and Numbers of Conformers

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.[2,3,5,38]

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)[12] using a standard molecular force field (MMFF94),[39] or using CREST conformer generation with the GFN2 method, as

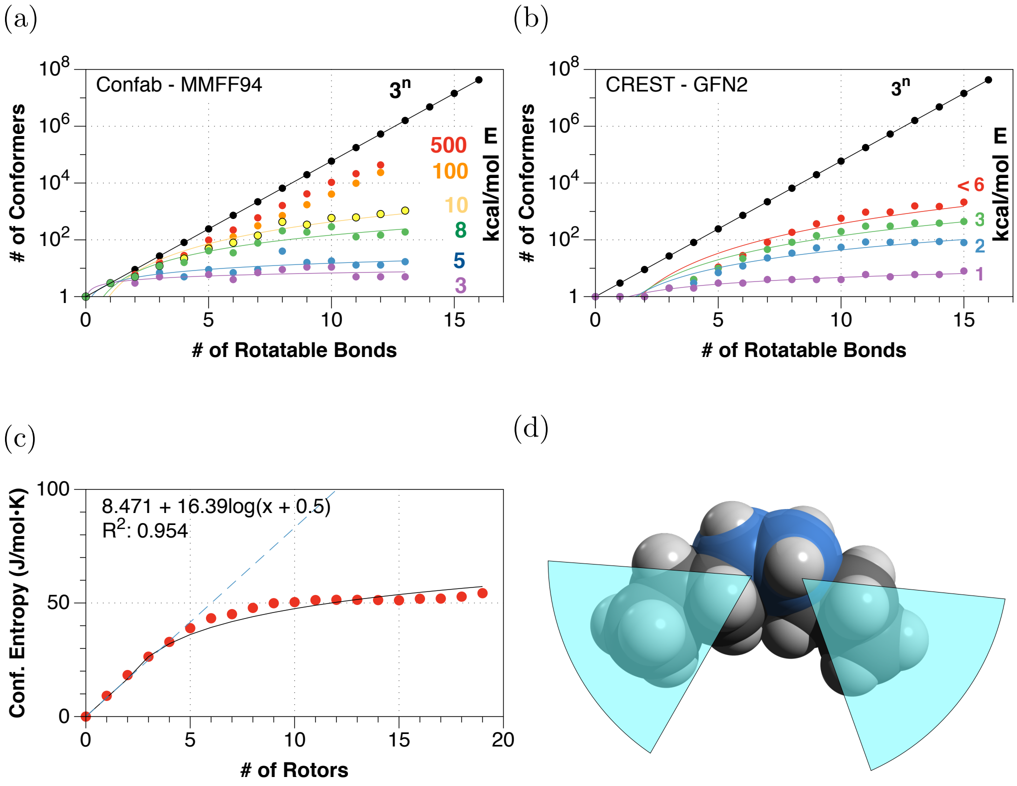


Figure 2: Conformational entropies for increasing lengths of n-unbranched alkanes, C_nH_{2n+2}. The counts of conformations in (a) and (b) are shown on a logarithmic scale. (a) Number of alkane conformers within a given energy range (in kcal/mol) of the global minimum (*i.e.*, within 3, 5, 8, 10, 100, 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.

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.[2,3,5,38]

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.

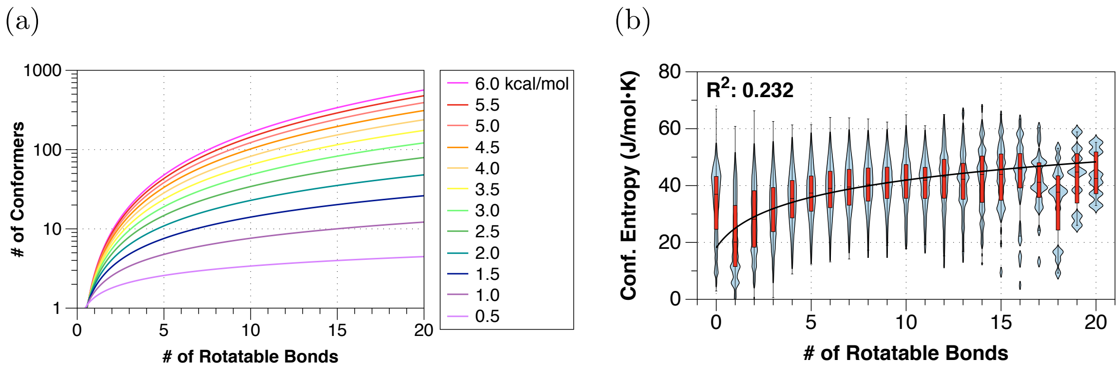


Figure 3: (a) Scaling of the number of conformers across the ∼93,000 molecules in the training set, on a logarithmic scale, within a given energy threshold, as a function of the number of rotatable bonds; and (b) correlation between the number of rotatable bonds, N_{rotor}, and GFN2-calculated conformational 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.

3.1.1 Branching and Terminal Methyl Rotors

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.[40,41] The magnitude of the methyl rotor entropies are higher from the CREST/GFN2 ensembles than previous quantum chemical estimates (*i.e.* 9.1 J/mol · K from CREST/GFN2 vs. 6.8 J/mol · K from HF/6-31G(d) using a hindered rotor model),[40] 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).

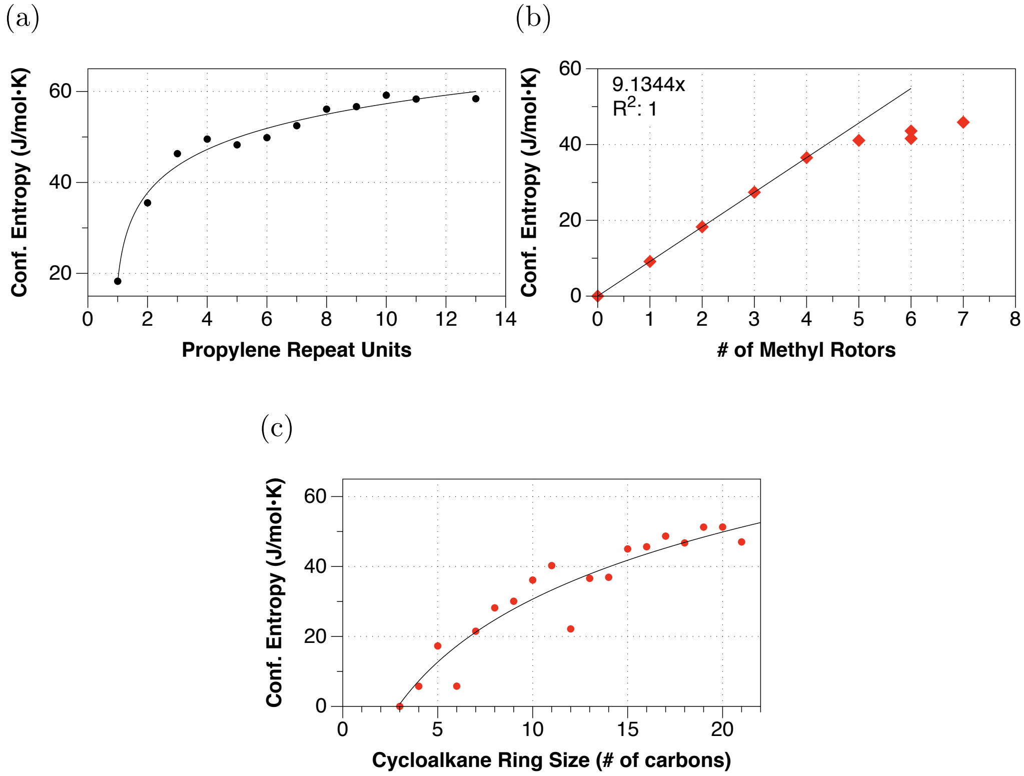


Figure 4: Conformational entropies for alkanes. (a) Conformational entropies calculated for increasing lengths of polypropylene chains, as a function of the number of repeat units, illustrating approximately logarithmic increase; and (b) branched alkane chains as a function of the number of terminal 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.

3.2 Ring Conformations

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.[42]

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)[43] 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).

3.3 Intramolecular Functional Group Effects

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.[44] 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.[45,46] Their geometries, including bond angles and distance, have been studied by others.[45,46] In this work, we further characterise the shortest path between donor and acceptor atoms, and show that the atom features at given positions are highly conserved for a given acceptor type (see Appendix C, Figure S14). The path with acceptor in a ring shows contrasting features to the path with acceptor not in a ring, indicating a strong influence of the ring geometry on the interaction. 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.

The formation of intramolecular 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 studied the interaction path containing the following functional groups: amide ester, ketone, ether, urea and carbamate group. These functional groups play an important role in supporting such small molecule 'folding'. Figure S16 in Appendix C shows that the atom features at given positions are highly conserved, especially for the path containing carbamate and urea groups. The paths containing amide or ether show varying atom features along the path.

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.

3.4 Statistical Models

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)[30] and continuous and data driven descriptors[31] were used separately as inputs for various machine learning models. We also included an end-to-end graph convolution networks for comparison.

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 approximated 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 plot in Figure S17 in Appendix D shows 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):

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

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, $R^2 = 0.72$. 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 expectations. 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.

3.5 Validation and Model Comparison

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 with ECFP6 as model inputs, and gives a mean absolute error of 4.79 and 4.46 $\mathrm{J/mol \cdot K}$ and Pearson correlation coefficient between predicted and computed conformational entropies $R^2 = 0.74$ and 0.65 respectively (see Table 1, and Appendix E Table S12). 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).

The kernel ridge regression and deep neural network models that are fed with continuous, data-driven descriptors (CDDD) somewhat outperform the linear regression by capturing additional nonlinearity (see Table 1). This highlights that the data driven representation discussed in this work encapsulates long range dependences of conformational entropy in the molecular graph, and therefore gives better predictive performance than local fingerprints such as ECFP6 alone.

Table 1: Model performance. Comparison of the mean absolute error (MAE) between the model's predicted and GFN2-computed conformational entropies, in J/mol · 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 is the proposed model. DNN with CDDD gives the lowest MAE in train and test sets, while KRR with CDDD gives the lowest MAE in CTPs sets.

Model	Training (MAE)	ZINC-I Set (MAE)	CTPs Set (MAE)
LR-1	8.67	8.83	9.00
LR-Best	5.08	4.79	4.46
LASSO (ECFP6)	5.55	5.47	6.76
Ridge (ECFP6)	4.95	5.29	5.83
KRR (ECFP6)	6.04	6.01	8.25
DNN (ECFP6)	5.22	5.27	6.99
LASSO (CDDD)	4.34	4.30	4.85
Ridge (CDDD)	4.27	4.26	4.54
KRR (CDDD)	4.14	4.14	**4.17**
DNN (CDDD)	**3.66**	**3.83**	4.32
GCN	5.18	4.86	5.46

4 Limitations and Future Work

The Gibbs-Shannon entropy formulation is used in the conformational entropy calculation, which means the conformer population (probability) is determined by GFN2-computed conformer energy. Hence the resulting conformational entropy of a given molecule may differ when other energy functions based on density functional theory (DFT) and force fields (FF) are used.[47] Moreover, an error may occur when the molecular dynamic simulations produce duplicated structures, as the calculation takes the unique conformations as input.

In addition, the temperature effect on entropy was not considered in our analysis, limiting the applicability of the model beyond 300 K to more general temperatures. We can introduce additional interaction terms in our model to take the temperature effect into account, as discussed in Appendix F. Figure S22 in Appendix F shows that the conformational entropies of unbranched alkanes at different temperatures ($T = 298.15$ K, 320 K, 360 K, 400 K) and the temperature has a significant effect on very flexible alkane molecules, in which the conformational entropies increase by approximately 5-6 J/mol · K at 400 K. The model with interaction terms is thus able to learn the underlying relationship and give accurate predictions.

Furthermore, as shown in Appendix A Figure S1, the training data predominantly consists of small and medium-sized organic molecules, *i.e.* number of heavy atoms $N_{\text{atoms}} < 70$. Hence the model is not readily applicable to large molecules with more than 100 atoms. To extend the capability of the model to a wider range of molecules, future work is required to (i) increase sampling of large ($N_{\text{atoms}} > 70$) and highly flexible ($N_{\text{rotor}} > 20$) molecules, and (ii) introduce additional descriptors to capture the effect of other long range intramoleculr forces in flexible molecules. Additional work is underway to consider similar conformational entropy effects in inorganic and organometallic molecules, including sampling a larger subset of databases such as PubChem.[48] Work to expand to larger and more flexible molecules is challenging to ensure full sampling of all thermally-accessible minima despite the stochastic nature of most conformer search methods.

5 Conclusion

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.[49–51] 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 $4.8\,\mathrm{J/mol\cdot K}$, or $\approx 0.34\,\mathrm{kcal/mol}$ at $300\,\mathrm{K}$. 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 solvation entropy as well as ligand conformational entropy upon protein-ligand binding, and thus provide better estimates of binding free energies for drug discovery.[11,52] With recent work on calculating accurate absolute entropies,[47] we believe similar efforts can include treatment of rotational, translational, and vibrational entropy contributions as well.

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,[53,54] 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

A summary and analysis of the molecules in the training and validation data sets, analysis of vibrational entropies, including relative timings, predictions of the number of conformers as a function of relative energies, details of the total ring flexibility descriptor, foldability descriptor, intramolecular functional group analysis, model diagnostics figures and analysis, model implementation details, including temperature effects. This information is available free of charge via the Internet at `https://pubs.acs.org`.

Additional Information Available

Data and code can be found in GitHub `https://github.com/hutchisonlab/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. *Phys. Chem. Chem. Phys.* **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? *Chem. Phys. Lett.* **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. *J. Chem. Phys.* **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. *J. Chem. Theory Comput.* **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. *Phys. Chem. Chem. Phys.* **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. *J. Chem. Phys.* **2016**, *145*, 114305.

(7) Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. Estimating entropies from molecular dynamics simulations. *J. Chem. Phys.* **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. *J. Chem. Theory Comput.* **2011**, *7*.

(9) Chang, C.-E.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017–1028.

(10) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5*, 3150–3160.

(11) Head, M. S.; Given, J. A.; Gilson, M. K. “Mining Minima”: Direct Computation of Conformational Free Energy. *J. Phys. Chem. A* **1997**, *101*, 1609–1618.

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

(13) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

(14) Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem3D: a new resource for scientists. *J. Cheminf.* **2011**, *3*, 32–32.

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

(16) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2020**, e26381.

(17) 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. *J. Chem. Inf. Model.* **2019**, *59*, 4195–4208.

(18) 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. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(19) 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). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

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

(21) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(22) 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)]. *J. Chem. Phys.* **2018**, *148*, 011101.

(23) 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. *J. Appl. Crystallogr.* **2009**, *42*, 726–729.

(24) 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 Res.* **2012**, *40*, D420–D427.

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

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

(27) Pfaendtner, J.; Yu, X.; Broadbelt, L. J. The 1-D hindered rotor approximation. *Theor. Chem. Acc.* **2007**, *118*, 881–898.

(28) Ayala, P. Y.; Schlegel, H. B. Identification and treatment of internal rotation in normal mode vibrational analysis. *J. Chem. Phys.* **1998**, *108*, 2314–2325.

(29) 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. *J. Chem. Phys.* **2006**, *124*, 044314.

(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(31) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692–1701.

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

(33) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825–2830.

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

(35) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Cambridge, MA, USA, 2015; p 2224–2232.

(36) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(37) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *J. Chem. Inf. Model.* **2017**, *57*, 122–126.

(38) 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. *J. Phys. Chem. A* **2003**, *107*, 3139–3145.

(39) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J. Comput. Chem.* **1996**, *17*, 587–615.

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

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

(42) Chan, L.; Hutchison, G. R.; Morris, G. M. Understanding Ring Puckering in Small Molecules and Cyclic Peptides. *J. Chem. Inf. Model.* **2021**, *61*, 743–755.

(43) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013–2021.

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

(45) Kuhn, B.; Mohr, P.; Stahl, M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. *J. Med. Chem.* **2010**, *53*, 2601–2611.

(46) 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 Crystallogr., Sect. B* **2000**, *56*, 849–856.

(47) Pracht, P.; Grimme, S. Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple. 2021; `https://chemrxiv.org/articles/preprint/Calculation_of_Absolute_Molecular_Entropies_and_Heat_Capacities_Made_Simple/13626083/1`.

(48) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2020**, *49*, D1388–D1395.

(49) Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.* **1997**, *4*, 10–19.

(50) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 20–22.

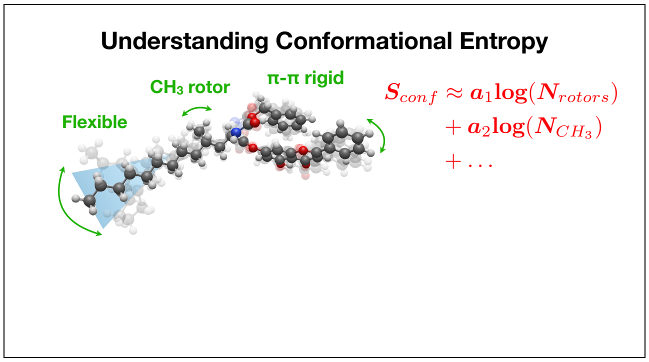
(51) Levinthal, C. Are there pathways for protein folding? *J. Chim. Phys* **1968**, *65*, 44–45.

(52) e. A. Chang, C.; Chen, W.; Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1534–1539.

(53) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F.; Foster, I.; Gardner, R.; Wilde, M.; Blatecky, A.; McGee, J.; Quick, R. The Open Science Grid. J. Phys.: Conf. Ser. 2007; p 012057.

(54) Sfiligoi, I.; Bradley, D. C.; Holzman, B.; Mhashilkar, P.; Padhi, S.; Wurthwein, F. The Pilot Way to Grid Resources Using glideinWMS. 2009 Proc. WRI World Congr. Comput. Sci. Inf. Eng. 2009; pp 428–432.

Graphical TOC Entry



Understanding Conformational Entropy

$$\boldsymbol{S_{conf}} \approx \boldsymbol{a_1}\mathbf{log}(\boldsymbol{N_{rotors}}) + \boldsymbol{a_2}\mathbf{log}(\boldsymbol{N_{CH_3}}) + \ldots$$