New mathematical measures for apprehending complexity of chiral molecules using information entropy

Patrick Piras

New mathematical measures for apprehending complexity of chiral molecules using information entropy

Patrick Piras

Aix Marseille Univ, CNRS, Centrale Marseille, iSm2, Marseille, France

Correspondence
Patrick Piras, Aix Marseille Univ, CNRS, Centrale Marseille, iSm2, Marseille, France.
Email: patrick.piras@univ-amu.fr

Abstract

In this paper, we present several new theoretical measures based on information entropy that can be used to analyze the information content of a chiral molecule. Starting from a differentiation between "chiral" and "achiral" portions in a chiral molecule, we define a new concept that allows us to quantify the complexity of chiral constitutional 2D-isomers of C_{10} to C_{20} alkanes. Various new chiral and achiral information measures founded on joint entropy, mutual information, and conditional entropy are presented providing an access to a set of regression equations. Then, introducing a case-based measure of entropy, we demonstrate that the distribution of the chiral complexity in these molecules is mostly skewed-right: 60% of the chiral isomers follow a 60/40 distribution rule, which indicates a concentration of chiral complexity in a small number of topological features. Furthermore, by replacing 2D topological distances by 3D distances, the application of these new information measures goes from conformational to racemization and deracemization studies. Interestingly, when the geometrical distances between atoms and the chiral center(s) are taken into account when determining the chiral information entropy, one can observe a significative Pearson correlation coefficient ($R = 0.70$) between the chiral entropy of 3D molecules and the continuous chirality measure. Finally, we show that our approach is applicable to almost any type of chiral organic chemical structures if in the entropy equation, atoms are represented by their electrotopological state (E-state) index instead of connectivity.

KEYWORDS

chemical graph, chiral alkanes, information theory, molecular diversity, Shannon entropy

1 | INTRODUCTION

1.1 | Thermodynamic entropy

Entropy is the most enigmatic of the physical quantities such as temperature, pressure, or volume defining the state of a thermodynamic system. Entropy designates the inability of the energy contained in a system to provide work: the higher this quantity, the more the energy is dispersed, homogenized and therefore less usable. When a thermodynamic system is isolated, that is, without possible exchange with the outside, it can only evolve spontaneously toward the maximum of its entropy to tend to a state of definitive equilibrium (second law of thermodynamics), whereas its internal energy remains conserved (first law of thermodynamics). In thermodynamics,

entropy is a state function introduced in the middle of the 19th century by Rudolf Clausius as part of the second principle.[1] Clausius showed that the ratio Q/T (where Q is the quantity of heat exchanged by a system at the temperature T) corresponds, in classical thermodynamics, to the variation of a state function which he called entropy (S) and whose unit is the joule per kelvin (JK^{-1}).

Statistical thermodynamics then shed a new light on this abstract physical quantity: entropy measures the degree of disorder of a system at the microscopic level. The higher the entropy of the system, the less its elements are ordered, interrelated, capable of producing mechanical effects, and the greater the exchange of energy unused or used inconsistently. Boltzmann[2] formulated a mathematical expression of statistical entropy as a function of the number of microscopic states W defining the equilibrium state of a given system at the macroscopic level: $S = k\,log\,W$. In Boltzmann's formula, the entropy S of a system of N particles, distributed over i states, having $N1;N2;:::;Ni$ particles with energies $E_1;E_2;:::;E_i$, is related to the total number W of physical states of the system. For an isolated system, microscopic states are equiprobable ($p_i = 1/w$ with $\sum p_i = 1$).

This definition of entropy is not inconsistent with that of Clausius. The two expressions of entropy simply result from two different points of view, depending on whether one considers the thermodynamic system at the macroscopic level or at the microscopic level. Later, Gibbs extended Boltzmann's equation to non-equiprobable microstates: $S = -k\sum p_i \log(p_i)$ where p_i is the probability of occurrence of each microstate i.

1.2 | Information entropy

The Gibbs definition of entropy can be readily related to the entropy of information introduced by Shannon in 1948.[3] In information theory, entropy is a quantitative measure of information. Shannon developed this mathematical theory for quantifying the information loss in transmitting a message in a communication. According to this theory, the total entropy of information H of a system (message) composed of N elements is defined by the relation[4]:

$$H = N\,log_2 N - \sum_{i=1}^{n} N_i log_2 N_i \qquad (1)$$

where N_i is the number of elements in the ith group of elements and n is the number of different groups of elements.

By taking the average $\overline{H}$, we obtain the well-known equation of Shannon's theory:

$$\overline{H} = -\sum_{i=1}^{n} p_i log_2 p_i \qquad (2)$$

where $p_i = \frac{N_i}{N}$, $N = \sum N_i$ and $\sum p_i = 1$, p_i being the probability of the ith group of elements.

In Shannon's convention, logarithmic base 2 is used, and the unit of information is called a bit (binary digit). Let us note that the quantities p_i are determined by the chosen partitions.

In 1955, Rashevsky[5] was the first to apply entropy to quantify the information content of a molecular graph using topological properties and the symmetry of the graph defined by vertex orbits. One year later, Trucco[6] also used symmetry using the set of all edge automorphisms to determine the complexity of a graph. In 1968, entropy was generalized by Mowshovitz[7] to the quantity of the information content of any system of N elements partitioned into equivalent classes according to a criterion α based on the orbits of the automorphism group of a graph G:

$$I_\alpha(G) = -\sum_{i=1}^{k} \frac{|V_i|}{|V|} log \frac{|V_i|}{|V|} \qquad (3)$$

where $|V_i|$ is the cardinality of the ith orbit of G and k is the number of different orbits.

Bonchev and Trinajstic[8] introduced magnitude-based information indices extending Mowshovitz probability scheme to a weighted probability scheme which is a new generalization of the previous measures to any property of magnitude M.

Since then, many entropy measures have been applied to graphs. We refer the reader to more general articles on this topic.[7,9,10] Today, information theory covers a variety of processes in all areas of science with impressive applications in biological systems.[11]

1.3 | Our approach

In this article, we will deal with the idea of applying Shannon's theory to measure the structural information content of chiral molecules. In 1995, Collet et al.[12] highlighted the importance of thermodynamic aspects in the prevalence of heterochiral crystals versus their homochiral crystals (conglomerate). Collet proposed that the entropy of mixing of liquid enantiomers ($-Rlog2$) could be related to the cost of the phase separation which is not favoring the crystallization of a conglomerate. Let us remember that in Boltzman's theory, particles are independent and do not interact with each other. In these conditions, enantiomers are not distinguishable, and thus, homochiral, scalemic, or racemic mixtures should

have an equivalent entropy. In Gibbs's equation of entropy, the entropy S of a gas mixture is equal to the sum of the entropies of individual components:

$$S = m \times H + m \times c \log t + m \times a \log \frac{v}{m} \quad (4)$$

where t and v are the temperature and the volume of the system, m is the quantity of the gas and a, c, H are constants. Clearly, from Equation 4, chirality cannot be distinguished by the ideal gas equation of entropy because enantiomers share the same physical properties, that is, same mass, same kinetic energy and other physical characteristics. This problem can be related to the Gibbs paradox. When mixing two separated products A and B of equal volume V, each expands from V to $2V$. The Gibbs paradox postulates that if the two products are identical ($A = B$), entropy will not change ($\Delta S_{AA} = \Delta S_{BB} = 0$). But if A and B are different, mixing $A + B$ (1 mol of A and 1 mol of B) will give

$$\Delta S_{AB} = 2Rln2 = 11.53\,JK^{-1} \quad (5)$$

where R is the gas constant ($8.314\,JK^{-1}mol^{-1}$).

Gibbs's thought experiment adapted to enantiomers is illustrated in Figure 1, that is, when the two considered substances are two equal volumes of enantiomers R and S. For Gibbs, there will be no increase of entropy in the mixing of the two enantiomers because reinserting the partition would lead to the same thermodynamic state as the previous one from a macroscopic scale point of view. We cannot observe an increase of entropy in the mixing of enantiomers R and S because macroscopic properties have not changed. The thermodynamic entropy change ΔS_{SR} is equal to zero. But we know that for a chiral molecule the situation is not so simple. In Figure 1, one can see that in an actual (R,S) enantiomeric mixture, the system is more complex. Because of enthalpic effects, different homochiral and heterochiral interactions can occur on a microscopic level.

Later, Gibbs related the mixing process to the concept of impossibility of separating identical substances. We have seen that Collet found the entropy of mixing is equal and of opposite sign to the process of restoring the original microstate. Indubitably, enantiomers can be separated through chiral recognition processes. Hence, homochiral and heterochiral interactions make enantiomers distinguishable and an increase in entropy naturally can take place despite them being identical in the thermodynamic point of view. As recently raised by Ribó,[13] "the ultimate chemical definition of chemical chirality should be the existence of enantiomeric discrimination between enantiomers."

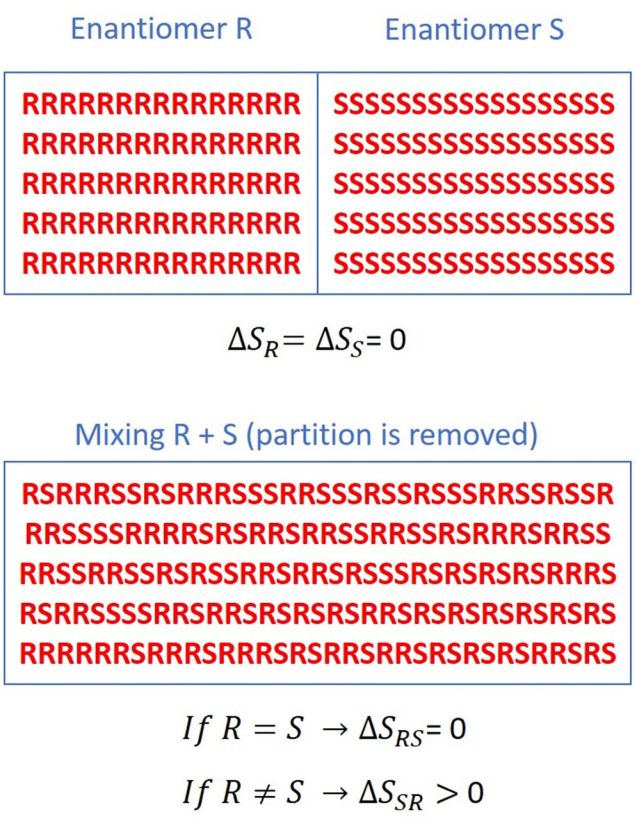

FIGURE 1 Application of the Gibbs paradox to a mixture of enantiomers

Here is where the concept of information entropy becomes important. In thermodynamics, only the energy level is considered and we have seen that chirality has nothing to do with energy. Clearly, chirality is a geometric property which can be translated into information.[14] This means that entropy of mixing of enantiomers, that is, entropy due to chiral interactions as proposed by Collet can be approached in term of the information loss characterized by Shannon's equation. This is the key idea of our approach: to analyze chiral molecules with regard to molecular information and not upon geometry. In what follows, we will show that Shannon entropy offers an original way to quantify the complexity of a chiral molecule possessing one or more stereogenic elements. More precisely, entropy measures will be built upon the partitioning of atom pairs which may be considered as functionally equivalent to two-point enantioselective or non-enantioselective interactions. Moreover, we will establish that the total information entropy H of chiral molecules also called joint entropy can be decomposed into a sum of chiral and achiral contributions. This new concept will allow us to access to various new 2D and 3D topological chiral and achiral information measures capable of capturing the complexity of a chiral molecule. The data set consists of all the chiral constitutional

isomers of C_{10} to C_{20} alkanes. The advantage of studying alkane isomers is that each C_{10} to C_{20} series is a set of all the possible isomers and thus covers an entire chemical feature space. This makes them particularly suitable for the discovery of mathematical rules.

2 | MATHEMATICAL PRELIMINARIES

Base 2 logs are generally used in information theory and information units are called "bit." Thus, any notation log stands for the logarithm in base 2 throughout this paper.

Definition 2.1. Undirected graph

An undirected graph $G=(V,E)$ consists of a set $V=V(G)$ of vertices and a set $E=E(G)$ of unordered pairs of vertices called edges. An edge $e\in E$ has the form $e=(u,v)$ for vertices $(u,v)\in V, u\neq v$; u and v are said to be adjacent in G.

Definition 2.2. Chemical graph

A chemical graph is a graph $G=(V,E)$, where V is an atom set and E is a chemical bond set.

Definition 2.3. Alkane

An alkane is an acyclic saturated hydrocarbon.

Definition 2.4. Adjacency matrix

The adjacency matrix (Table S5) of an n-vertex graph is an $(n X n)$ symmetric matrix $A=\left[a_{ij}\right]$ whose typical entry a_{ij} is defined as

$$a_{ij}=\begin{cases}1\ if\ i\neq j\\ 0\ if\ i=j\end{cases}$$

Definition 2.5. Shortest path

For a graph $G=(V,E)$, $d(u,v)$ represents the distance between u $\in$ V and $v\in V$ expressed as the minimum length of a path between u and v.

Definition 2.6. Degree of a vertex

The degree of a vertex v is the number of edges at v and is denoted by $\deg(v)$.

Definition 2.7. Distance matrix

The distance matrix (Table S6) of an n-vertex graph G is an $(n X n)$ symmetric matrix $D=\left[d_{ij}\right]$ whose typical entry d_{ij} is defined as

$$d_{ij}=\begin{cases}d(v_i,v_j)\ if\ i\neq j\\ 0\ if\ i=j\end{cases}$$

where $d(v_i,v_j)$ is the shortest path between $v_i\in G$ and $v_j\in G$.

Definition 2.8. Eigenvalue

Let A a $n\times n$ matrix with real entries, where n is a positive integer. A number λ is an eigenvalue of A if there exists a nonzero $n\times 1$ column vector x such that $Ax=\lambda x$. A vector $x\neq 0$ with this property is an eigenvector of A for the eigenvalue λ.

Definition 2.9. Graph Laplacian

The Laplacian L of a graph G is the operator

$$L(G)=D(G)-A(G) \tag{6}$$

where $D(G)=Diag(\deg(v_1);deg(v_2);:::;deg(v_n))$ is the degree matrix of G and $A(G)$ is the adjacency matrix.

$L(G)$ is a positive semidefinite and singular M-matrix. Thus, all the eigenvalues of $L(G)$ are real and usually arranged in a nonincreasing order

$$\lambda_n\geq\lambda_{n-1}\geq\ \cdots\geq\lambda_2\geq\lambda_1=0 \tag{7}$$

Definition 2.10. General definition of Shannon's entropy of a finite probability space

Let $X=\{x_1,\cdots,x_n\}$ be a nonempty finite set of elements with a discrete probability $\{p(x_1),\cdots,p(x_n)\}$ assigned to each, such that $p(x_i)\geq 0$ for all $1\leq i\leq n$ and $p(x_1)+\cdots+p(x_n)=1$. Then, the Shannon entropy H of the ensemble is the real number

$$H(X)=-\sum_{i=1}^{n}p(x_i)log_2p(x_i) \tag{8}$$

Remark:

- The function $H(X)$, defined on the set of all probability laws $P=(p(x_1),\cdots,p(x_n))$ on X, is strictly concave. Hence, it possesses a unique maximum and the maximum possible value of $H(X)$ is $log_2(n)$. This occurs when the distribution is uniform, that is, $p(x_i)=1/n$ for all i.
- In entropy calculations, we always use the following convention: $0\,log_2(0)=0$ according to $nlog_2(1/n)\to 0\,as\,n\to 0$.

Definition 2.11. Entropy of a graph

The entropy of a graph G is:

$$H_\propto(G)=-\sum_{i=1}^{k}p_i log_2 p_i=-\sum_{i=1}^{k}\frac{|X_i|}{|X|}log_2\frac{|X_i|}{|X|} \quad (9)$$

where X is a collection of graph invariants of G, α is the criterion that partitions X into k equivalent classes of cardinality $|X_i|$ and $p_i=\frac{|X_i|}{|X|}$ is the probability value of the ith partition.

Definition 2.12. Joint entropy

Let X,Y discrete random variables with joint distribution $p(x,y)$. The joint entropy $H(X,Y)$ is the real number

$$H(X,Y)=-\sum_x\sum_y p(x,y)log_2 p(x,y) \quad (10)$$

where x and y are particular values of X and Y and $p(x,y)$ is the probability of these values occurring together. If variables X and Y are independent, the joint entropy is the sum of individual entropies, that is, $H(X,Y)=H(X)+H(Y)$.

Definition 2.13. Conditional entropy

For two random variables X and Y, the conditional entropy of Y given X (or vice versa) is defined as

$$H(Y|X)=-\sum_x p(x)\sum_y p(y|x)log_2 p(y|x) \quad (11)$$

where $p(y|x)$ is the conditional probability. The conditional entropy is a measure of how much uncertainty remains about the random variable Y when we know the value of X.

3 | METHOD

In this article, we take advantage of the Shannon concept of self-information, that is, the information provided by a random process about itself.

3.1 | First theoretical experiment (isolated molecule)

In a first experiment, we will only consider intramolecular atom–atom topological features in the calculation of information entropy in an isolated chiral molecule. In this context, one can observe that a chiral molecule which possesses at least one stereogenic element can contain more or less large "achiral portions" that are isolated areas comprising no chiral center, axis or plane. This concept is illustrated in Figure 2. In these "achiral portions," there is no stereogenic element in the shortest path between two atoms. In return, we can also identify "chiral" atom pairs which fall within the surrounding of at less one chiral element of the molecule, that is, there is at least one stereogenic element in the shortest path

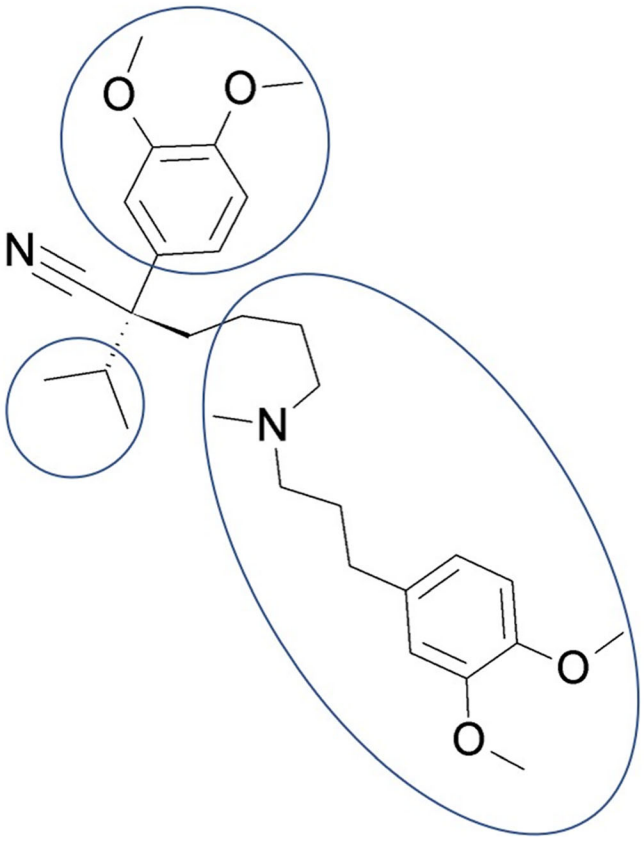

FIGURE 2 Example of "achiral portions" that can be isolated in a chiral molecule. Any atom pairs randomly selected in these regions have no stereogenic element in their shortest path

between the two atoms. Another way to visualize this concept is by considering the following chiral alkanes which possesses one asymmetric center at the end of the chain:

$$CH_3 - CH_2 - CH - (CH_3) - CH_2 - (CH_2)_n - CH_3$$

When the length of the alkyl chain increases, that is, $n \to +\infty$, the probability that any enantioselective interaction with a chiral target occurs relative to a non-enantioselective interaction decreases. So, one may claim that when $n \to +\infty$ the "achiral" characteristics of this chemical structure become more and more preponderant than its "chiral" characteristics.

In discrete probability, a sample space Ω, is the set of all possible outcomes of an experiment. In the atom pair space of a chiral compound, we will assume there exist two independent subsets: S_{chiral} and $S_{achiral}$. Accepting this concept and as entropy preserves the additivity of independent events, the joint entropy $H^*(S_{chiral}, S_{achiral})$ of a chiral molecule is the sum of individual S_{chiral} and $S_{achiral}$ entropies (Definition 2.12):

$$H^*(S_{chiral}, S_{achiral}) = H^*(S_{chiral}) + H^*(S_{achiral}) \tag{12}$$

where the asterisk means $H^*(S_{chiral})$ and $H^*(S_{achiral})$ are independent variables.

Using Shannon entropy (Definition 2.10), we obtain:

$$H^*(S_{chiral}, S_{achiral}) = -\left(\sum_{i=1}^{k} p_i^{chiral} \log p_i^{chiral} + \sum_{i=1}^{l} p_i^{achiral} \log p_i^{achiral}\right) \tag{13}$$

where k and l are respectively the number of chiral and achiral partitions.

$p_i = \frac{|X_i|}{|X|}$ is the probability value assigned to the ith chiral or achiral partitions.where X is a collection of all atom-atom pairs and $|X_i|$ is the cardinality of the ith equivalent class of atom pairs.

As X is a finite set, then

$$\sum_{i=1}^{k} p_i^{chiral} + \sum_{i=1}^{l} p_i^{achiral} = 1 \tag{14}$$

3.2 | Second theoretical experiment (molecule in interaction)

In the introduction, we have seen that chiral discrimination can emerge from the intermolecular interaction of two chiral compounds. Then, in a second experiment, we will suppose that the entropy measure depends on the interaction of a guest chiral molecule with a chiral host. This means that there is an effect of the interactions of the chiral molecules with its chiral anisotropic surroundings. In that situation, one can consider that there exist equivalent "chiral" and "achiral" atom pairs which can compete on the same binding sites of a chiral ligand or receptor, that is, the two subspaces S_{chiral} and $S_{achiral}$ share common structural features. Consequently, individual S_{chiral} and $S_{achiral}$ entropy contributions are not independent of one another. We have seen that entropy is only additive in systems without interaction. Since, in this second experiment, to measure the joint entropy $H(S_{chiral}, S_{achiral})$ we must subtract one overlap of entropy $I(S_{chiral} : S_{achiral})$ as visualized in the Venn diagram of Figure 3.

$$H(S_{chiral}, S_{achiral}) = H(S_{chiral}) + H(S_{achiral}) - I(S_{chiral} : S_{achiral}) \tag{15}$$

where $I(S_{chiral} : S_{achiral})$ is the mutual information which measures the information shared by S_{chiral} and $S_{achiral}$.

By reorganizing Equation 15, the mutual information $I(S_{chiral} : S_{achiral})$ can be defined as follows:

$$I(S_{chiral} : S_{achiral}) = H(S_{chiral}) + H(S_{achiral}) - H(S_{chiral}, S_{achiral}) \tag{16}$$

From Figure 3, the following relationships between the various information measures are straightforward but interesting to note:

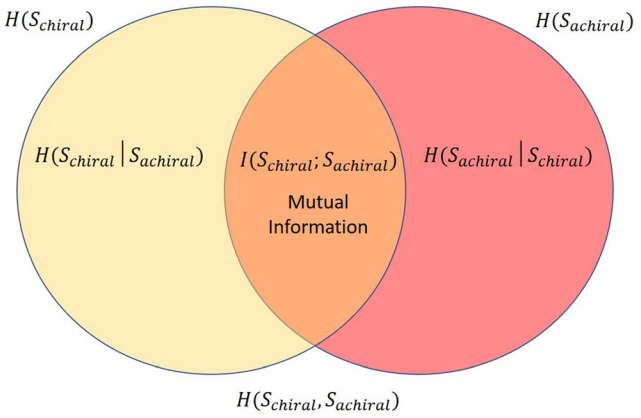

FIGURE 3 Venn diagram of the joint entropy $H(S_{chiral}, S_{achiral})$ for a chiral molecule in interaction with the environment. The area where both circles overlap is the mutual information $I(S_{chiral} : S_{achiral})$ of the S_{chiral} and $S_{achiral}$ subset distributions. In case of an isolated molecule, $H(S_{chiral})$ and $H(S_{achiral})$ are separated, that is, $I(S\_chiral : S\_achiral) = 0$

$$H(S_{chiral}, S_{achiral}) = H(S_{chiral}) + H(S_{achiral}|S_{chiral}) \quad (17)$$

$$H(S_{chiral}, S_{achiral}) \leq H(S_{chiral}) + H(S_{achiral}) \quad (18)$$

$$H(S_{chiral}|S_{achiral}) \leq H(S_{chiral}) \quad (19)$$

where $H(S_{achiral}|S_{chiral})$ and $H(S_{chiral}|S_{achiral})$ are the conditional entropies (Definition 2.13).

Equality holds in Equations 18 and 19 if and only if S_{chiral} and $S_{achiral}$ are independent as it was assumed in the theoretical experiment 1. It is also interesting to note that the inequality 18 provides a bound to the joint entropy of a chiral molecule.

Let us notice some remarks on interpreting these equations:

- Shannon entropy presents several advantages in comparison with other reported graph complexity measures: (1) It has an additive property which permits to distinguish between chiral and achiral structural feature contributions; (2) it has a physical meaning thanks to the relationship between thermodynamic entropy and information (e.g., Maxwell's demon[15]); (3) by taking the stereogenic element as a reference in entropy formula, it is able to take into account the symmetry of the structure (Figure S5).
- These definitions refer to chiral molecules which have stereogenic elements and those characterize the vast majority of chiral molecules. For Inherently chiral molecules such as helicenes or fullerenes where chirality only arises from the overall molecular arrangement, one can consider that $H(S_{achiral})$ is equal to 0 and thus the joint entropy is equal to $\mathrm{H}(S_{chiral})$.
- In the kinetic theory of gases, the notion of molecular chaos is historically attributed to Boltzmann: the higher the entropy, the greater the disorder. In information theory, a more chaotic system means more complexity and more richness in information. Accordingly, for a chiral molecule, a higher value of the entropy will be related to a lower symmetry, a higher molecular complexity and thus a greatest diversity.
- If one admits that "chiral" atom pairs are functionally equivalent to two-point enantioselective interactions, then a single additional interaction is sufficient to provide a chiral recognition.

This last remark suggests that a possible way to improve the recognition ability of a chiral selector or the substrate universality of a chiral catalyst may be to increase its joint entropy, that is, decrease the mutual information between S_{chiral} and $S_{achiral}$ (see Venn diagram). Indeed, by acting on the mutual information, one could reduce the redundancy. Redundancy means at the same time less information content and more molecular interaction competition between "chiral" and "achiral" atom pairs of a chemical structure.

One interesting problem that derives from the existence of the two different subsets S_{chiral} and $S_{achiral}$ is to see if one of the two subset distributions is far from the optimal and thus if the joint entropy can be maximized. Furthermore, in maximizing for instance the chiral entropy $H(S_{chiral})$, one expects minimizing the mutual information and at the same time increasing the performance of a chiral selector or catalyst. It will be shown in the next section that this problem of finding the maximum joint entropy distribution has a unique solution due to the concavity of Shannon entropy function.

3.3 | Maximum entropy approaches

The maximum entropy principle was introduced by Jaynes in 1957.[16] Jaynes stated that the most appropriate distribution to model a given set of data is the one with highest entropy. Jaynes formulates this problem as maximizing the entropy function in the presence of the following constraints:

$$\sum\nolimits_i p_i = 1 \text{ and } \langle f \rangle = \sum\nolimits_i^n p_i f(x_{i.})$$

where $\sum_i p_i$ is the normalization constraint and $\langle f \rangle$ is the first moment (average) of the distribution imposed by the knowledge one has about the data. To this end, Jaynes used the Lagrangian method.

In our first theoretical experiment, we have no information except the normalization constraint which is the sum $\sum_{i=1}^{k} p_i^{chiral} + \sum_{i=1}^{l} p_i^{achiral} = 1.0$. Our goal here is to estimate the maximum value of the chiral entropy $H^*(S_{chiral})$ under the hypothesis that there is no interaction between the "chiral" and the "achiral" parts of the molecule (first theoretical experiment). Using λ_0 as Lagrange multiplier, the Lagrangian $\mathtt{L}$ of $H^*(S_{chiral}, S_{achiral})$ is

$$\begin{aligned} &\mathtt{L}(S_{chiral}, S_{achiral}, \lambda_0) \\ &= -\left(\sum\nolimits_{i=1}^{k} p_i^{chiral} \log p_i^{chiral} + \sum\nolimits_{i=1}^{l} p_i^{achiral} \right. \\ &\quad \left. \log p_i^{achiral} \right) - \lambda_0 \left(\sum\nolimits_{i=1}^{k} p_i^{chiral} + \sum\nolimits_{i=1}^{l} p_i^{achiral} - 1 \right) \end{aligned} \quad (20)$$

Being interested in finding the value of entropy $H^*(S_{chiral})$ that maximizes the entropy $H^*(S_{chiral}S_{achiral})$, we take the partial derivative of the Lagrangian L with respect to p_i^{chiral} and setting it equal to zero

$$\frac{\partial}{\partial p_i^{chiral}}(\mathrm{L}(S_{chiral}, S_{achiral}, \lambda_0)) = 0 \tag{21}$$

yields

$$-\left(\log p_i^{chiral} + 1\right) - \lambda_0 = 0 \tag{22}$$

Rearranging gives:

$$\begin{gathered} p_i^{chiral} = e^{-\lambda_0 - 1} \\ \lambda_0 = const \text{ then } p_i^{chiral} = const \end{gathered} \tag{23}$$

This implies that $p_1^{chiral} = p_2^{chiral} ... = p_k^{chiral}$ and as a result, S_{chiral} distribution should be as uniform as possible to enable $H^*(S_{chiral})$ to reach a maximum entropy.

Then, using the normalization constraint we find:

$$\mathrm{k}e^{-\lambda_0 - 1} + \sum\nolimits_{i=1}^{l} p_i^{achiral} = 1 \tag{24}$$

Rearranging, we have:

$$e^{-\lambda_0 - 1} = \frac{1 - \sum_{i=1}^{l} p_i^{achiral}}{k} = \frac{\sum_{i=1}^{k} p_i^{chiral}}{k} \tag{25}$$

From Equation (25) one can now express the maximum chiral entropy $H^*_{max}(S_{chiral})$

$$H^*_{max}(S_{chiral}) = -\left(\sum\nolimits_{i=1}^{k} p_i^{chiral}\right) \log \frac{\sum_{i=1}^{k} p_i^{chiral}}{k} \tag{26}$$

The above expression allows us to estimate the maximum entropy of $H^*(S_{chiral})$ in the achiral environment fixed by $H^*(S_{achiral})$. An equivalent relationship is found for $H^*_{max}(S_{achiral})$ if we solve the partial derivative of the Lagrangian L with respect to $p_i^{achiral}$

$$H^*_{max}(S_{achiral}) = -\left(\sum\nolimits_{i=1}^{l} p_i^{achiral}\right) \log \frac{\sum_{i=1}^{l} p_i^{achiral}}{l} \tag{27}$$

According to Shannon entropy definition and from Brown and Martin,[17] to obtain a maximum $H^*(S_{chiral})$ joint entropy

$$p_i^{chiral} = p_i^{achiral} = e^{-\lambda_0 - 1} = \frac{1}{k+l} \tag{28}$$

From Equation (28) the following expressions can be derived

$$H^*_{max}(S_{chiral}, S_{achiral}) = (\lambda_0 + 1)e^{-\lambda_0 - 1}(k+l) = \lambda_0 + 1 \tag{29}$$

$$\begin{aligned} H^*_{max}(S_{chiral}, S_{achiral}) &= -\sum\nolimits_{1}^{k+l} \left(\frac{1}{k+l} \log \frac{1}{k+l}\right) \\ &= \log(k+l) \end{aligned} \tag{30}$$

$$H^*_{Max}(S_{chiral}) = \frac{k}{k+l} \log(k+l) \tag{31}$$

According to this last equation and as expected, the number of chiral partitions k should be as large as possible and the number of achiral partitions l as small as possible to tend toward a maximum $H^*(S_{chiral})$ entropy.

The expression 31 gives the maximum entropy of $H^*(S_{chiral})$ with no constraints except the normalization term. However, in an actual alkane graph, the maximum vertex degree is at most 4 and this constraint affects the full structure, that is, S_{chiral} and $S_{achiral}$. Consequently, an additional constraint $\sum_i p_i f(x_i)$, $\mathrm{i} \in S_{chiral}$ should be applied in the maximization of the entropy. Choosing an appropriate function f, solution to the new lagrangian equation gives the individual p_i^{chiral} for an optimal distribution

$$p_i^{chiral} = e^{-1-(\lambda_0 + \lambda_1 f(x_i))} \tag{32}$$

Clearly, there is a relationship between chiral probability distributions and the function f depending on x_i. This indicates that to find the optimal arrangement of the probability distribution p_i^{chiral}, one needs to include more information about the organization of the chemical structure. Since, as the additional constraint reduces the maximum entropy, the actual maximum entropy is lower than the estimated $H^*_{Max}(S_{chiral})$ upper bound given by the Equation (31), that is, $actual\, H^*_{Max}(S_{chiral}) < H^*_{Max}(S_{chiral})$.

Finally, if we apply the maximum entropy approach within the framework of the second theoretical experiment in which the mutual information $I(S_{chiral} : S_{achiral}) \neq 0$, the objective is now to maximize the joint entropy

$$H(S_{chiral}, S_{achiral}) = -\sum\nolimits_{x \in S_{chiral}} \sum\nolimits_{y \in S_{achiral}} p(x,y) log p(x,y) \tag{33}$$

subject to the two following constraints

$$\sum_{x \in S_{chiral}} \sum_{y \in S_{achiral}} p(x,y) = 1 \tag{34}$$

$$\sum_{i} p_i f_r(x_i,y_i), \mathrm{i} \in (S_{chiral}, S_{achiral}), 1 \le r \le m \tag{35}$$

Then, individual p_i^{chiral} for an optimal distribution is

$$p_i^{chiral} = e^{-1-\left(\lambda_0+\lambda_1 f_r\left(x_i,y_i\right)+\ldots+\lambda_m f_r\left(x_i,y_i\right)\right)} \tag{36}$$

4 | APPLICATION

4.1 | The data set

The chosen data set consists of all the chiral constitutional isomers of C_{10} to C_{20} alkanes. All the isomers of alkanes were exhaustively generated using Faulon's algorithm[18,19] providing 603,455 chiral alkanes and 14,582 achiral alkanes. Table 1 shows the progress of the number of chiral and achiral isomers when the number n of carbon atoms increases from 7 to 20. At first glance, one important remark is the exponential increase of chiral alkanes correlated with a less pronounced increase of achiral alkanes. A cross-over occurs very early when $n = 10$ (decane isomers) then the chiral alkane population rises exponentially until representing the vast majority of the molecules (98% of the C_{20} alkanes are chiral).

TABLE 1 Distribution of chiral and achiral C_n alkane constitutional isomers ($n = 7$ to 20)

n	Chiral	Achiral	Total	% of chiral
7	2	7	9	22
8	5	13	18	27
9	15	20	35	42
10	40	35	75	53
11	104	55	159	65
12	259	96	355	72
13	646	156	802	80
14	1591	267	1858	85
15	3909	438	4347	89
16	9612	747	10,359	92
17	23,655	1239	24,894	95
18	58,424	2099	60,523	96
19	144,786	3498	148,284	97
20	360,407	5912	366,319	98

Once all the chiral constitutional isomers have been enumerated, it is further possible to compare all the alkane families from the distribution of the number of chiral centers. In Figure 4, the number of isomers of each C_7 to C_{20} alkane category is arranged according to their number of chiral centers. One can note in this figure that alkane isomers form a regular spacial distribution. When the number n of carbons increases, an increase in the number of chiral centers is accompanied by an expansion of the number of isomers which behaves like a propagating wave. Another interesting finding is that we are able to predict from this figure the maximum number of chiral centers that can be found in a given C_n alkane. In Figure 5 is plotted the maximum number of chiral centers against the number n of carbon atoms for each C_n alkane. Let us start from $n = 7$, the smallest chiral alkane which presents two isomers with one center. One need to add one carbon atom to observe an increase of one chiral center and then three more carbon atoms to have another increase of one chiral center. Thus, as a general rule, the maximum number of chiral centers that can be reached by a given C_n alkane is given by the following equation:

$$N_{\max} = \left\lfloor \frac{n-4}{2} \right\rfloor \tag{37}$$

where n is the number of carbon atoms and the mathematical notation $\lfloor\rfloor$ means rounding down the result to the nearest integer.

4.2 | Graph information entropy measures of 2D chiral alkanes

We have seen that the measure of entropy is based on the partitions spawned by the structural features of a chemical graph. Partitions are then converted into probabilities as represented by the following scheme:

$$\text{Equivalent classes } C_1, C_2, \ldots, C_i$$

$$\text{Partitions } N_1, N_2, \ldots, N_i$$

$$\text{Probabilities } p_1, p_2, \ldots, p_i$$

For the purpose of simplicity and for a better clarity of the plots, only results related to C_{10} alkane isomers are shown throughout this paper but comparable findings were observed for the other C_{11} to C_{20} alkane data sets.

Entropy information measures were computed according to the equations explained in Sections 3.2 and 3.3 and using hydrogen depleted structures. For chiral

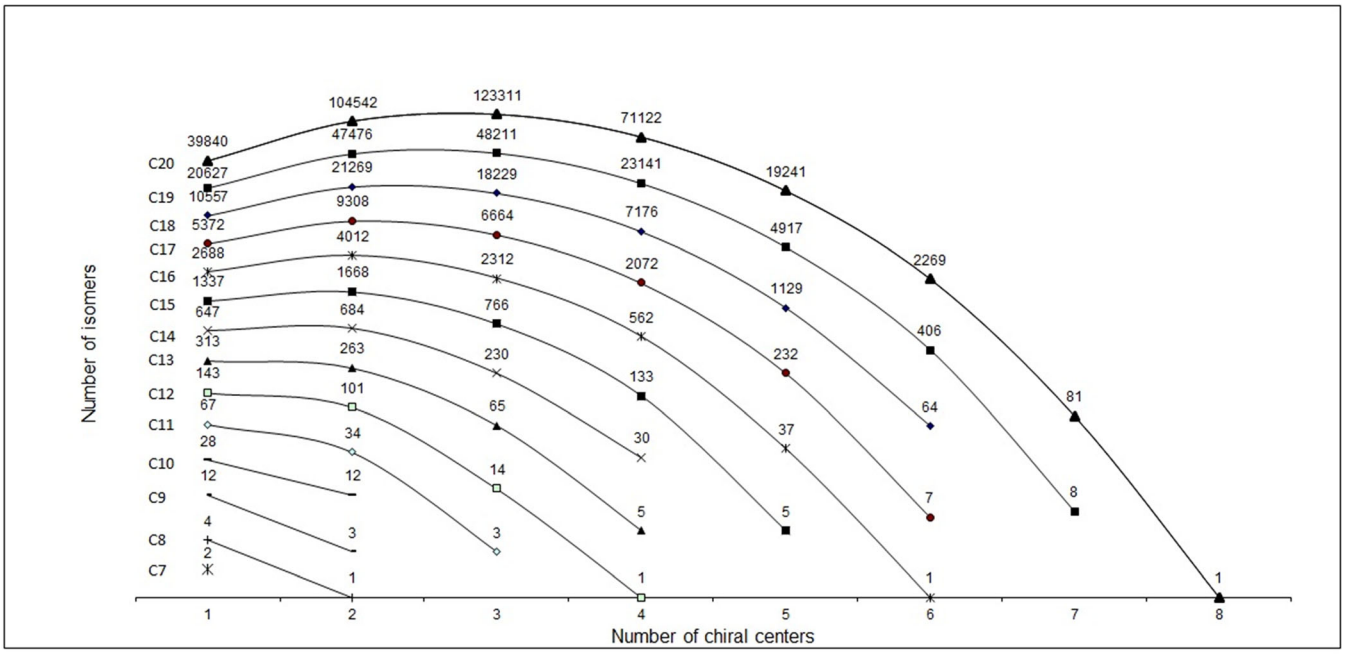

FIGURE 4 Distribution of chiral constitutional isomers of C_7 to C_{20} alkanes according to their number of chiral centers (plotted on a log scale)

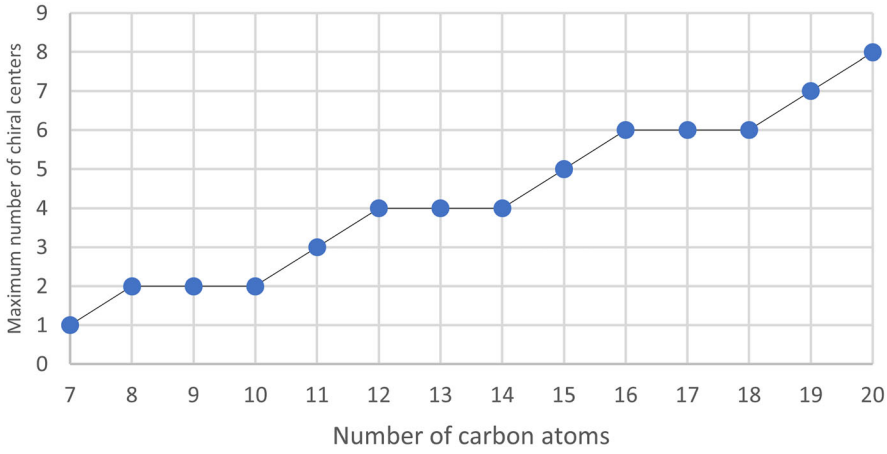

FIGURE 5 Plot of the number of carbon atoms n versus the maximum number of chiral centers that can be found in constitutional isomers of C_7 to C_{20} alkanes

alkanes, atom pairs are partitioned into equivalent classes by using graph invariants, for example, vertices, edges, degrees, and distances.

4.2.1 | 2D topological distance-based information entropy

As a pre-evaluation of our approach, we first examined if entropy could capture 2D topological distance information of chiral molecules. This is achieved by building the distance matrix of the chemical graph (Definition 2.7) and computing entropy using path distances between atom pairs. This means that atom pairs are partitioned into equivalent classes by collecting the shortest path lengths (Definition 2.5). For an isolated molecule (Section 3.1), we can write

$$H^*_{2D\_dist}(S_{chiral}, S_{achiral}) = H^*_{2D\_dist}(S_{chiral}) + H^*_{2D\_dist}(S_{achiral}) \tag{38}$$

where $H^*_{2D\_dist}(S_{chiral}) = -\sum_{i=1}^{k} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|}$ and $H^*_{2D\_dist}(S_{achiral}) = -\sum_{i=1}^{l} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|}$

X_i being the number of chiral or achiral atom pairs sharing the same shortest path length in partition i.

Since Laplacian eigenvalues are well-known to be related to the partitions of the graphs, we attempted to

find relationships between the Laplacian spectra of the chiral molecules and distance entropies.

Spectrum of a graph is a set of its eigenvalues and their multiplicities (see Definition 2.8). Let λ_i denote the eigenvalues of a graph Laplacian, λ_i can be arranged in a decreasing order:

$$\lambda_n \geq \lambda_{n-1} \geq \ldots \geq \lambda_2 \geq \lambda_1 = 0$$

This is a well-established property of Laplacian spectra theory for all connected graphs such as chemical graphs.

The nonzero smallest eigenvalue λ_2 of a Laplacian matrix was called by Fiedler the algebraic connectivity.[20] Fiedler introduced λ_2 as a quantitative measure of connectivity. The larger is λ_2 value, the more connected is the graph. Hence, measures of λ_2 determine "how well" graphs are connected. If a chemical graph exhibits a low λ_2, one may simply remove a few vertices or edges within the graph to identify regions that are different, that is, regions that can be easily isolated from the rest of the structure. In other words, λ_2 captures the local patterns and the larger λ_2 is, the more difficult it is to cut a graph into different elements. λ_2 can thus be interpreted as a measure of the regularity. From Figure 6, one can see that a good correlation exists between λ_2 and the distance-based joint entropy $H_{2D\_dist}(S_{chiral}, S_{achiral})$ computed from the distance matrix of chiral C_{10} alkanes. In this figure, high values of the distance-based joint entropies are associated with low λ_2 values, indicating a higher dissymmetry (less connectivity) and thus less homogeneity in the corresponding molecule graphs. This is a noteworthy result as to our knowledge, relationships between λ_2, eigenvalue and graph distance entropies have not yet been fully explored.

In order to determine the individual contributions of S_{chiral} and $S_{achiral}$ subsets to the distance variability, a regression analysis was performed between λ_2 and the two separated S_{chiral} and $S_{achiral}$ distance entropies. The following expression is obtained:

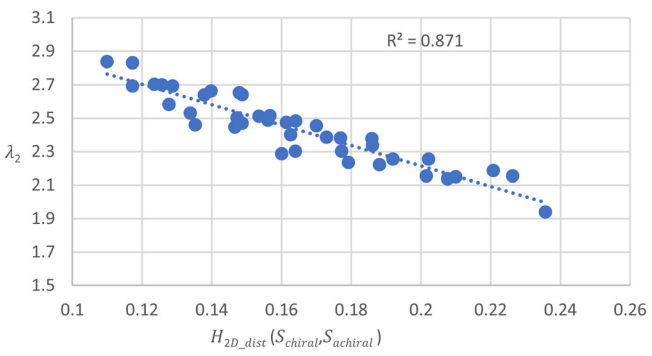

FIGURE 6 Correlation between eigenvalue λ_2 and the distance-based joint entropy $H^*_{2D\_dist}(S_{chiral}, S_{achiral})$

$$\lambda_2 = -0.136\, H^*_{2D_{dist}}(S_{chiral}) - 0.054\, H^*_{2D_{dist}}(S_{achiral}) + 0.462 \quad (R^2 = 0.79) \tag{39}$$

This equation shows that the distance variability of the chiral alkanes is mostly explained by the contribution of the S_{chiral} subset. On average, about 70% of the distance diversity can be explained by S_{chiral}.

4.2.2 | Combined 2D topological distance and connectivity-based information entropy

In the following sections, atom pairs are partitioned into equivalent classes according to two criteria: the shortest path and the connectivity of the carbon atoms.

Introduction of two new entropic indexes R_{chiral} and $D_{\max\_ent}$ for measuring chiral information complexity

In Table 2, we give the results obtained from the application of the various entropy formula seen above onto the fourteen chiral C_{10} alkane isomers. All the chemical structures are reported in Table S1.

In this table, the conditional entropies $H(S_{chiral}|S_{achiral})$ and $H(S_{achiral}|S_{achiral})$ are the first measures to consider in the situation of a molecule in interaction, that is, in the context of second theoretical experiment. Indeed, looking at the Venn diagram of Figure 3, maximizing the chiral diversity and complexity of a molecule involves maximizing $H(S_{chiral}|S_{achiral})$ and minimizing $H(S_{achiral}|S_{achiral})$ and the mutual information $I(S_{chiral} : S_{achiral})$ (Equation 15). This brings us to introduce a new index R_{chiral} that we will call chiral information richness

$$R_{chiral} = \frac{H(S_{chiral}|S_{achiral})}{H(S_{chiral}, S_{achiral})} \tag{40}$$

with $0 < R_{chiral} \leq 1.0$

R_{chiral} is a normalized measure. As seen in Table 1, its value can vary between 0 and 1.0 and a value near 1.0 means the molecule practically reaches the maximum chiral information richness. A combined plot of R_{chiral} against mutual information exhibited a highly significant negative relationship ($R^2 = 0.84$) indicating that high levels of chiral information richness are associated with a less competitive achiral environment (Figure S4).

Another way to evaluate the chiral information content of a molecule is to calculate the ratio of the chiral conditional entropy $H(S_{chiral}|S_{achiral})$ to the corresponding maximum chiral entropy value $H^*_{Max}(S_{chiral})$ (Equation 31). This new entropic index $D_{\max ent}$ is defined by

TABLE 2 Various chiral and achiral information measures of constitutional isomers of chiral C_{10} alkanes. $H^*(S_{chiral})$ and $H^*(S_{achiral})$ were measured according to first theoretical experiment (S_{chiral} and $S_{achiral}$ have no interaction)

No.	Name	$H^*(S_{chiral})$	$H^*(S_{achiral})$	$H(S_{chiral},S_{achiral})$	$H(S_{chiral}\|S_{achiral})$	$H(S_{achiral}\|S_{chiral})$	$I(S_{chiral}:S_{achiral})$	$H_{Max}(S_{chiral})$	$H_{Max}(S_{achiral})$	R_{chiral}
1	3-methylnonane	2.54	1.83	3.84	1.18	0.58	2.08	2.59	1.93	0.31
2	4-methylnonane	2.80	1.48	3.86	1.73	0.58	1.55	2.85	1.54	0.45
3	4-ethyloctane	2.69	1.12	3.64	2.16	0.83	0.65	2.84	1.16	0.59
4	2,6-dimethyloctane	2.54	2.02	4.17	1.70	1.18	1.29	2.59	2.06	0.41
5	2,5-dimethyloctane	2.84	1.57	4.16	2.37	1.04	0.75	2.91	1.61	0.57
6	2,4-dimethyloctane	2.96	1.48	4.19	2.44	0.89	0.86	3.06	1.53	0.58
7	2,3-dimethyloctane	2.74	1.62	4.13	2.00	1.19	0.94	2.85	1.68	0.48
8	3,5-dimethyloctane	3.31	0.62	3.79	2.91	0.44	0.44	3.45	0.63	0.77
9	4-ethyl-2-methylheptane	2.65	1.19	3.84	2.65	1.19	0.00	2.84	1.23	0.69
10	3,4-dimethyloctane	2.98	1.06	3.82	2.36	0.72	0.75	3.09	1.08	0.62
11	3-ethyl-2-methylheptane	2.81	1.27	3.92	2.27	0.98	0.66	3.01	1.29	0.58
12	4,5-dimethyloctane	2.99	0.82	3.66	2.72	0.54	0.40	3.09	0.82	0.74
13	4-ethyl-3-methylheptane	2.81	0.69	3.43	2.64	0.52	0.27	3.07	0.73	0.77
14	3-ethyl-4-methylheptane	2.59	1.57	3.81	1.71	0.82	1.28	2.70	1.61	0.45
15	3,6-dimethyloctane	3.22	0.44	3.66	3.22	0.44	0.00	3.35	0.45	0.88
16	3-ethyl-5-methylheptane	2.37	1.62	3.81	2.08	0.97	0.75	2.46	1.67	0.55
17	4-ethyl-2,3-dimethylhexane	2.47	1.59	3.76	1.77	0.98	1.00	2.54	1.61	0.47
18	3-ethyl-2,5-dimethylhexane	2.49	1.10	3.59	2.49	1.10	0.00	2.74	1.16	0.69
19	2,2,3-trimethylheptane	2.61	1.40	3.85	2.11	1.12	0.63	2.78	1.46	0.55
20	2,2,4-trimethylheptane	2.54	1.38	3.77	2.31	1.03	0.43	2.70	1.46	0.61
21	2,3,4-trimethylheptane	3.05	0.96	3.94	2.88	0.79	0.27	3.18	0.97	0.73
22	3,4,5-trimethylheptane	3.27	0.44	3.71	3.27	0.44	0.00	3.35	0.45	0.88
23	3,4,4-trimethylheptane	2.44	1.94	3.94	1.48	0.89	1.56	2.53	2.00	0.38
24	3-ethyl-2,3-dimethylhexane	3.00	0.89	3.82	2.77	0.75	0.30	3.23	0.91	0.73
25	3,3,4-trimethylheptane	2.73	1.57	4.04	2.21	0.89	0.94	2.85	1.61	0.55
26	3-ethyl-2,2-dimethylhexane	2.46	1.12	3.57	2.46	1.12	0.00	2.74	1.16	0.69
27	2,2,5-trimethylheptane	2.36	1.77	3.86	1.90	1.17	0.79	2.46	1.85	0.49
28	3,3,5-trimethylheptane	2.44	1.85	3.95	1.78	0.98	1.19	2.53	1.93	0.45
29	2,4,5-trimethylheptane	3.01	1.06	3.92	2.74	0.79	0.40	3.09	1.08	0.70

TABLE 2 (Continued)

No.	Name	$H^*(S_{chiral})$	$H^*(S_{achiral})$	$H(S_{chiral},S_{achiral})$	$H(S_{chiral}\|S_{achiral})$	$H(S_{achiral}\|S_{chiral})$	$I(S_{chiral}:S_{achiral})$	$H_{Max}(S_{chiral})$	$H_{Max}(S_{achiral})$	R_{chiral}
30	3-ethyl-3,4-dimethylhexane	2.22	1.89	3.62	1.10	0.82	1.70	2.32	1.93	0.30
31	2,3,5-trimethylheptane	3.30	0.62	3.91	3.30	0.62	0.00	3.45	0.63	0.84
32	3-ethyl-2,4-dimethylhexane	2.84	0.72	3.55	2.84	0.72	0.00	3.07	0.73	0.80
33	2,3,6-trimethylheptane	2.56	1.48	3.77	1.97	0.87	0.93	2.70	1.54	0.52
34	2,3,4,5-tetramethylhexane	2.36	0.67	3.04	2.36	0.67	0.00	2.48	0.69	0.78
35	2,3,4,4-tetramethylhexane	2.54	1.59	3.92	2.04	1.21	0.66	2.70	1.61	0.52
36	2,2,3,4-tetramethylhexane	2.57	0.86	3.43	2.57	0.86	0.00	2.80	0.90	0.75
37	2,3,3,4-tetramethylhexane	2.35	1.80	3.88	1.91	1.21	0.76	2.46	1.85	0.49
38	3-ethyl-2,2,4-trimethylpentane	2.00	1.03	3.03	2.00	1.03	0.00	2.37	1.07	0.66
39	2,2,3,5-tetramethylhexane	2.51	1.26	3.69	2.36	1.03	0.30	2.70	1.30	0.64
40	2,2,4,5-tetramethylhexane	2.44	1.32	3.68	2.30	1.03	0.35	2.62	1.37	0.63

$$D_{\max ent} = \frac{H(S_{chiral}|S_{achiral})}{H^*_{Max}(S_{chiral})} \tag{41}$$

$D_{\max ent}$ is also scaled to the range between 0 and 1 and can be seen as a measure of the proximity of the chiral information to its maximum entropy. Lower $D_{\max ent}$ means less information content and so a lower diversity of the structural elements. Examination of Figure 7 reveals that R_{chiral} and $D_{\max ent}$ are well correlated ($R^2 = 0.87$). This correlation indicates that the index R_{chiral} is indeed a good indicator of the richness and diversity of the chiral structural information contained in a chiral compound as exemplified by Figure 8.

Another question of interest is to determine the extent to which S_{chiral} and $S_{achiral}$ are contributing to the mutual information. An easy way to model this is to consider the two independent entropies $H^*(S_{chiral})$ and $H^*(S_{achiral})$. Then, the individual contributions to the mutual information $i(S_{chiral}:S_{achiral})$ and $i(S_{achiral}:S_{chiral})$ are obtained by subtracting their respective conditional entropies $H(S_{chiral}|S_{achiral})$ and $H(S_{achiral}|S_{chiral})$

$$i(S_{chiral}:S_{achiral}) = H^*(S_{chiral}) - H(S_{chiral}|S_{achiral}) \tag{42}$$

$$i(S_{achiral}:S_{chiral}) = H^*(S_{achiral}) - H(S_{achiral}|S_{chiral}) \tag{43}$$

A regression analysis between $i(S_{chiral}:S_{achiral})$ and $i(S_{achiral}:S_{chiral})$ was carried out and resulted in the following equation

$$i(S_{chiral}:S_{achiral}) = 0.9285\, i(S_{achiral}:S_{chiral}) \left(R^2 = 0.84\right) \tag{44}$$

This indicates that there is a linear dependence of the contributions of S_{chiral} and $S_{achiral}$ to the mutual information and according to the coefficient, the two contributions are comparable: for every unit increase of the contribution of S_{chiral}, there is a similar unit increase of $S_{achiral}$ (Figure 9).

A case-based entropy study for comparing distributions of chiral complexity

The index R_{chiral} is a global measure which gives information about the overall diversity of S_{chiral} in regards to $S_{achiral}$ subset but it does not provide any information about the diversity distribution across the different structural elements of the chiral molecule. Rajaram and Castellani have designed a solution to this problem by introducing a new complexity measure C_c called case-based entropy measure.[21] The method consists first to calculate the true diversity measure D which by taking

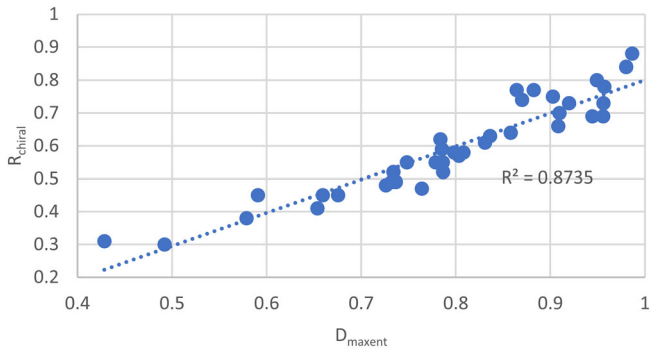

FIGURE 7 Relationship between the chiral information richness R_{chiral} and the distance to maximum chiral entropy $D_{\text{max}ent}$

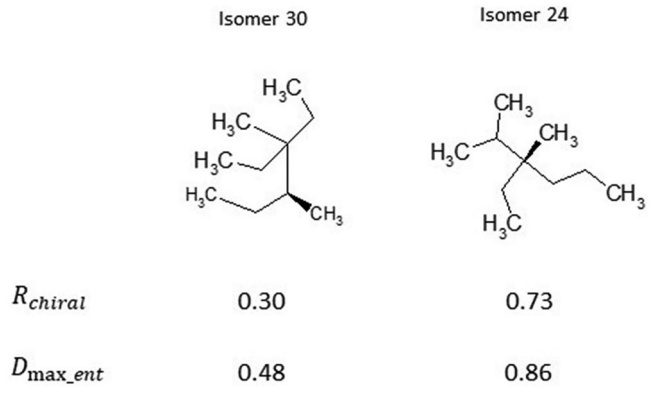

FIGURE 8 Comparison of two C_{10} alkane isomers in term of chiral information richness (R_{chiral}). Chiral complexity of isomer 30 is less important, that is, contains less diversity around the chiral center than isomer 24

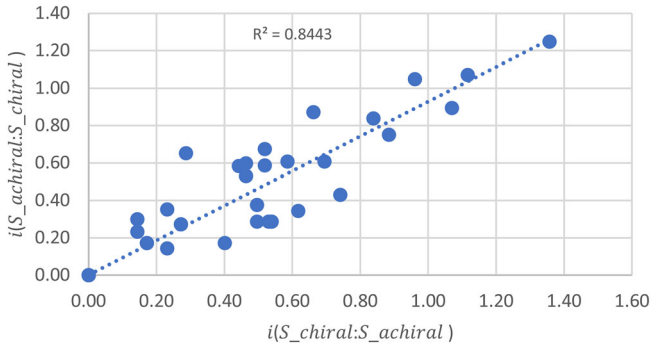

FIGURE 9 Relationship between $i(S_{chiral} : S_{achiral})$ and $i(S_{achiral} : S_{chiral})$ contributions to mutual information

the exponential of Shannon entropy, converts the additive property of entropy into a multiplicative one.[22] When applied to our independent S_{chiral} subset, the true diversity $D(S_{chiral})$ is calculated using the following expression:

$$D(S_{chiral}) = e^{H^*(S_{chiral})} \tag{45}$$

Then, in a second step, p_i^{chiral} is replaced by $\widehat{p}_i^{chiral}$ in $H^*(S_{chiral})$ entropy formula to give

$$H_c^*(S_{chiral}) = \sum_{i=1}^{k} \widehat{p}_i^{chiral} log_2 \widehat{p}_i^{chiral} \tag{46}$$

where $p_i^{chiral} = p_i^{chiral}/c$, c being the cumulative probability $\sum_{i=1}^{k} p_i^{chiral}$ so that $H_c^*(S_{chiral})$ becomes the chiral entropy calculated as if only the first k set is observed.

Therefore, using 46 and rearranging, the true diversity $D_c(S_{chiral})$ reduces to

$$D_c(S_{chiral}) = e^{H_c^*(S_{chiral})} = \prod_{i=1}^{k} \frac{1}{\widehat{p}_i^{chiral\,\widetilde{p}_i^{chiral}}} \tag{47}$$

Replacing $\widehat{p}_i^{chiral}$ by p_i^{chiral}/c we obtain

$$D_c(S_{chiral}) = \frac{c}{\prod_{i=1}^{k} p_i^{chiral\,\frac{p_i^{chiral}}{c}}} \tag{48}$$

Finally, the percentage diversity contribution $C_c(S_{chiral})$ is given by

$$C_c(S_{chiral}) = \frac{D_c(S_{chiral}) \times 100}{D(S_{chiral})} \tag{49}$$

The same approach is also applied to $S_{achiral}$ probability distribution to obtain

$$C_c(S_{achiral}) = \frac{D_c(S_{achiral}) \times 100}{D(S_{achiral})} \tag{50}$$

Next, the method consists to plot the diversity contribution C_c versus the cumulative probability c. As shown in Figure 10, the x-axis represents the diversity contribution and the y-axis the frequency of cases relative to the collection x. As both axes range between values from 0 to 1, different complexity distributions of C_{10} to C_{20} alkanes can be compared on the same graph. In Figure 10, the S_{chiral} and $S_{achiral}$ curves form a signature profile of the complexity of isomer 30. The maximum diversity is represented by the straight line, that is, the distribution is uniform ($p_i = 1/n$ for all i). Any deviation of S_{chiral} or $S_{achiral}$ complexity distribution from this line is associated with a lower diversity. Not surprisingly, S_{chiral} diversity of complexity of isomer 30 ($R_{chiral} = 0.30$) is found more restrained than $S_{achiral}$ diversity. Although the chiral information richness R_{chiral} gives us access to a global

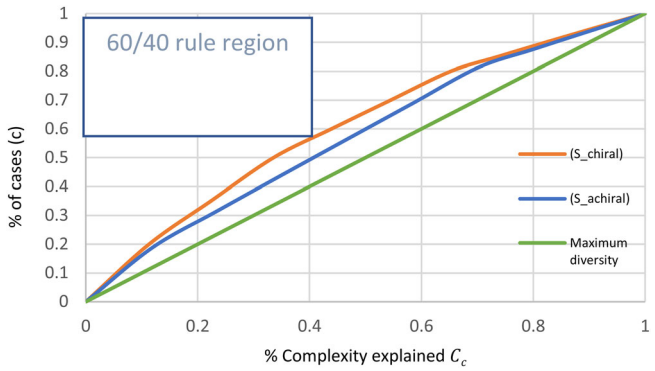

FIGURE 10 Diversity contribution C_c versus the cumulative frequency of cases c (isomer 30, $R_{chiral} = 0.30$). The straight line delineates the theoretical maximum diversity

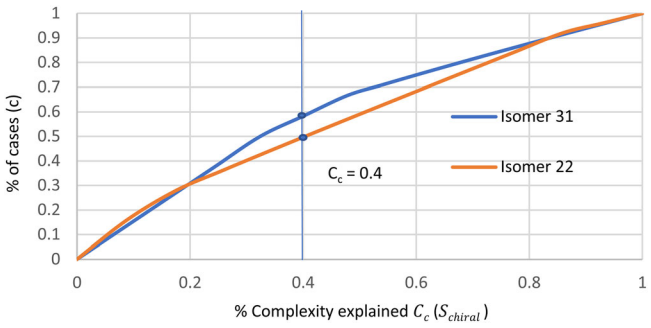

FIGURE 11 Diversity contribution C_c versus the cumulative frequency of cases c of S_{chiral} for isomers 31 and 22. The two isomers 31 and 22 have a similarly high level of chiral information diversity (R_{chiral} respectively equal to 0.84 and 0.88), but distribution of complexity C_c of isomer 31 is more restricted than isomer 22

complexity measure, the diversity contribution C_c tells us how is distributed the chiral information.

Rajaram and Castellani remarked that when a complexity distribution histogram is skewed-right, the system is governed by a law of "restricted diversity." Using C_c to measure the distribution of the diversity, authors revealed the universal existence of a 60/40 law from galaxies to genes[23]: a more or less majority of cases (≥60%) can explain a small percentage of the total diversity of complexity (≤40%). This is illustrated in Figure 10. The 60/40 rule is a delimited area which informs if there is or not a restricted distribution of complexity. It is defined on the graph by the region where the curve is above 60% of the cases and under 40% of the total complexity. Points of the curve located in this region indicate if the diversity of complexity is more or less restricted. When examining the S_{chiral} complexity distribution in this region of the curves generated from all chiral C_{10} alkanes, we observe that 60% of the isomers do obey the 60/40 rule (Table S4). In Figure 10, S_{chiral} complexity distribution of isomer 30 comes into contact with the 60/40 region whereas $S_{achiral}$ complexity shows a better distributed diversity.

Even more noteworthy is the comparison of two isomers as for example the isomers 31 and 22 of Figure 11 which both exhibit a high R_{chiral} value (respectively 0.84 and 0.88). One can see that the first isomer respects the 60/40 rule, whereas this rule is not followed by the second one. The two compounds have a similar level of chiral information complexity, but this complexity is more concentrated in certain topological elements of isomer 31 and more dispersed across the structural elements of isomer 22. More generally, as seen in Figure 12, it is especially interesting to note that there is no relationship between the chiral information richness R_{chiral} and the distribution of complexity. Indeed, R_{chiral} is a global measure that is not directly related to the distribution of the chiral complexity in chiral alkanes. From this

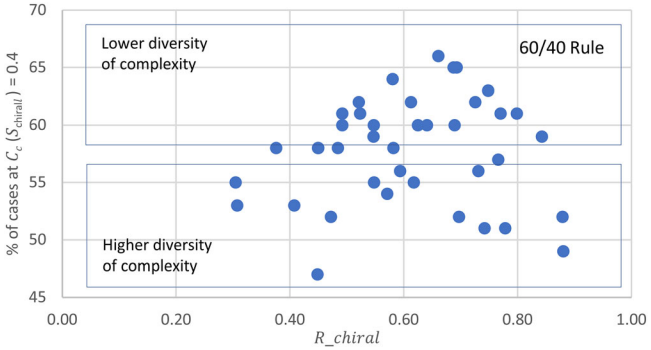

FIGURE 12 Chiral information richness (R_{chiral}) versus the cumulative frequency of cases at $C_c(S_{chiral}) = 0.4$

observation, an interesting question arises regarding to whether a better distribution of the chiral complexity may affect favorably the performance of a chiral molecule (a selector or a catalyst). For this purpose, a potential application of the case-based method would be to detect what local changes in molecule structure are required to achieve this objective.

4.3 | Graph information entropy measures of 3D chiral alkanes

To explain biological activities in a meaningful way, one often need 3D descriptors to analyze the variations in the 3D structures of chemical compounds. This is why we also investigated our entropy approach on chiral 3D molecules. To this end, we replaced the shortest path distances by the geometric distances in the dataset. Then, the fundamental idea is to partition the geometric distances into equivalent classes using a consensus over

multiple runs of a random K-means clustering. For comparing partitions, we used the very popular Adjusted Rand Index.[24]

One simple way to introduce geometric distances in the entropy measures is to build a geometric distance matrix of each 3D chiral alkane. The geometric distance matrix of a molecular graph (G) is a real symmetric $n X n$ matrix $D = [d_{ij}]$ where n represents the number of vertices in the chemical graph and each entry d_{ij} is defined as

$$d_{ij} = \begin{cases} d(v_i, v_j) \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}$$

where $d(v_i, v_j)$ is the Euclidean distance between $v_i \in G$ and $v_j \in G$

$$d(v_i, v_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (51)$$

The Cartesian coordinates for each vertex of the molecular graph were obtained from geometry optimizations utilizing $MMFF_{94}/MMFF_{94s}$ force fields as available within the open-source cheminformatics toolkit RDKit.[25]

In the following sections, when a given 3D C_{10} chiral alkane possesses two asymmetric carbon atoms, entropy measures were applied to one of its stereoisomers.

4.3.1 | Comparison of 2D and 3D structure-based entropy measures

The value of 3D descriptors is questioned in the literature because it is often shown that models built from 2D descriptors gives better results. On this subject, the well-known work of Brown and Martin is often cited. By comparing 2D and 3D molecular keys, 2D-descriptors were proven to be more efficient at separating biologically active molecules from inactive.[17] In a recent work, we also found that 2D-fingerprint descriptors were more powerful than 3D-descriptors to build models to find the most promising chiral selector to achieve the separation of a chiral compound.[26] It looks like meaningful 2D information is lost during the construction process of 3D-descriptors. We thus checked if this effect could be observed between $H^*_{2D}(S_{chiral})$ and $H^*_{3D}(S_{chiral})$ entropy measures. Interestingly, there is a good correlation between 2D and 3D structure-based entropy measures as illustrated in Figure 13. Such a correlation means that the 3D structure-based entropy shares some information content with the 2D entropy. The correlation factor ($R^2 = 0.77$) indicates that the values are well correlated but the entropy computed from the 3D coordinates of atoms also adds new information about the chiral molecule.

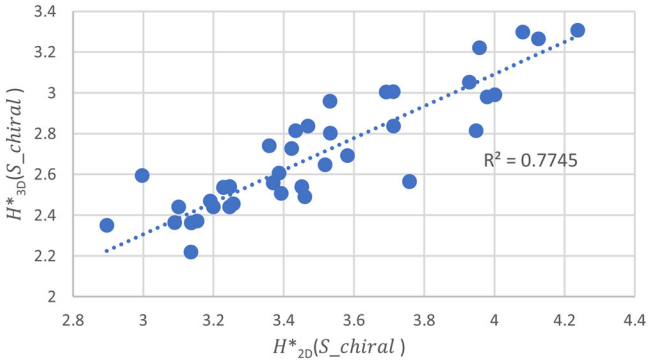

FIGURE 13 Relationship between $H^*_{2D}(S_{chiral})$ and $H^*_{3D}(S_{chiral})$

4.3.2 | 3D topological distance-based information entropy

In this section, the structural elements of the chiral molecule are partitioned into equivalent classes according to the geometrical distances. Then, the 3D distance-based joint entropy $H^*_{3D\_dist}(S_{chiral}, S_{achiral})$ for an isolated molecule is

$$H^*_{3D\_dist}(S_{chiral}, S_{achiral}) = H^*_{3D\_dist}(S_{chiral}) + H^*_{3D\_dist}(S_{achiral}) \quad (52)$$

where $H^*_{3D\_dist}(S_{chiral}) = -\sum_{i=1}^{k} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|}$ and $H^*_{3D\_dist}(S_{achiral}) = -\sum_{i=1}^{l} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|}$

X_i being the number of chiral or achiral atom pairs sharing the same geometrical distance in partition i.

Since, $H^*_{3D\_dist}(S_{chiral}, S_{achiral})$ is a kind of conformational entropy measure.

Typically, we find that for constrained 3D C_{10} alkane structures, molecular mechanics optimization produces when possible a more distant conformation (the less compact), that is, the joint entropy increases when the energy decreases. Thus, a higher geometrical distance diversity of atom pairs is observed in the lowest-energy conformations as for isomer 38 of Figure 14. As seen in Figure 15, the joint entropy of isomer 38 increases when the energy decreases and the lowest-energy conformations exhibit the higher entropy values. On the other hand, some isomers which have a better flexibility as isomer 40 of Figure 14 may produce more symmetric conformations when the energy decreases. In these cases, the lowest-energy conformations provide the lowest distance entropy values.

4.3.3 | Correlation between information entropy and degree of chirality

In this part of our article, we will be focusing on, to what extent, our new entropy measures encode the degree of chirality of the molecules.

The existence of chirality is widespread in the world of chemistry. This led the chemists to raise the question of "how chiral" is a given molecule. Various measures for quantifying chirality and symmetry have been proposed in the literature as it has been exhaustively reviewed by Petitjean.[27] One of the most important has been developed by Zabrodsky and Avnir who suggested a continuous chirality measure (CCM) to quantify the degree of chirality of a chiral molecule.[28,29] Their method is based on the minimal distances that the vertices of a shape have to be moved in order to reach the nearest achiral symmetry point group.

Our first attempts to correlate the various entropy measures of the C_{10} alkane isomers with the CCM degree of chirality were not successful. A reason of this bad result is probably that our entropy measures do not capture the position of the chiral center(s) and so these measures cannot incorporate information about the chirality strength. A better result was effectively achieved by replacing the geometrical distances between atoms by the geometric distances of atoms to the chiral center(s) in the partition data set of the chiral entropy $H^*(S_{chiral})$

$$H^*_{\chi}(S_{chiral}) = -\sum_{i=1}^{k} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|} \tag{53}$$

where χ is the criterion that partitions the collection X into k equivalent classes. χ is based on a collection of the geometric distances of atoms to the chiral centers and their connectivities. Using this new information entropy measure, we find a moderate relationship between $H^*_{\chi}(S_{chiral})$ and CCM values ($R^2 = 0.49$, Figure S1). The coefficient of determination R^2 characterizes the proportion of variation in CCM due to a linear relationship between $H^*_{\chi}(S_{chiral})$ and CCM. However, in our case, we are mostly interested in investigating how strongly the values of these variables are related to one another. This is the purpose of the Pearson's coefficient which is a normalized measurement of the covariance. Since, using

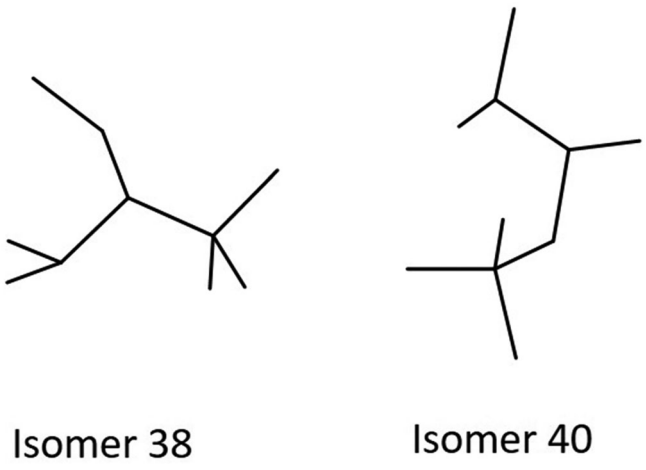

FIGURE 14 Lowest-energy conformation of alkane isomers 38 and 40. Isomer 38 presents a higher geometrical distance diversity than isomer 40 which contains more symmetries in the spatial arrangement

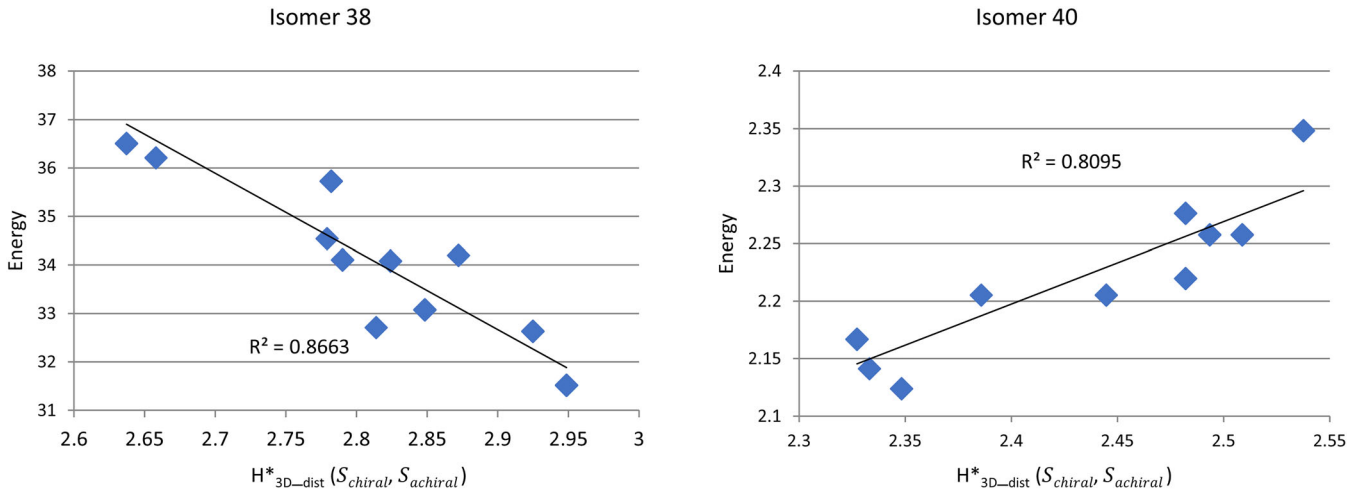

FIGURE 15 3D distance-based joint entropy versus conformation energy of alkane isomers 38 and 40

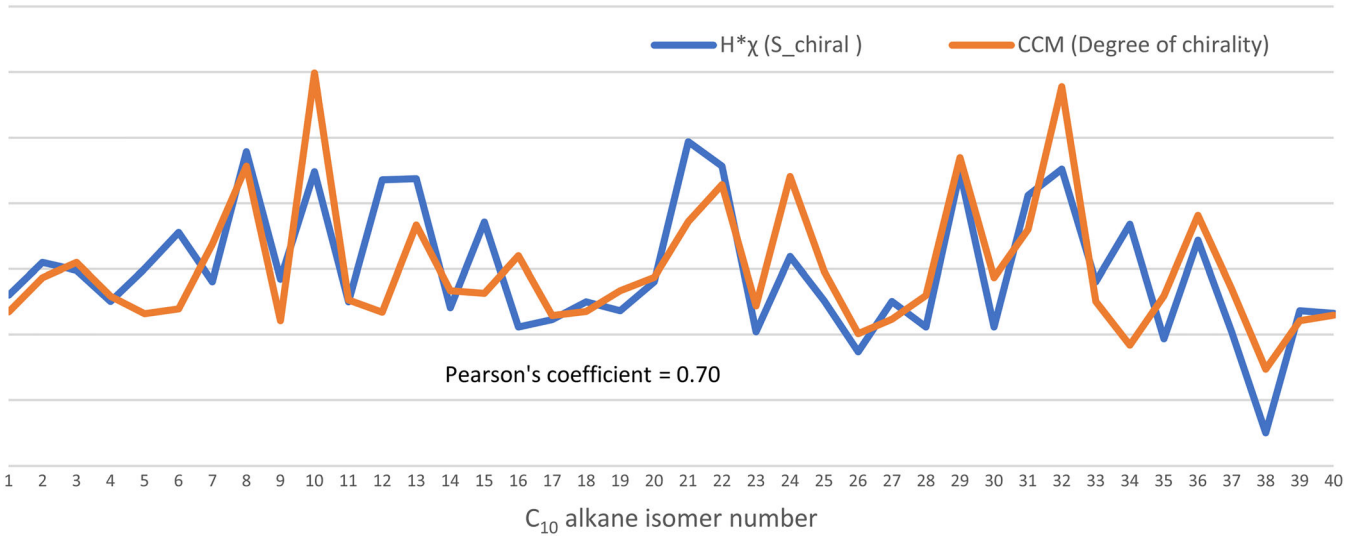

FIGURE 16 Comparison of the normalized distributions of $H^*_\chi(S_{chiral})$ and the CCM degree of chirality of 3D-optimized chiral C_{10} alkane isomers. Isomers are ranked on the *x*-axis according to Table 1 numbering

Pearson's coefficient, one can find the extent to which changes in the value of one variable are correlated to changes in the value of the other variable. In Figure 16, the correlation coefficient as measured in relation to $H^*_\chi(S_{chiral})$ and CCM, shows a significant relationship (Pearson's coefficient $R = 0.70$). $H^*_\chi(S_{chiral})$ and CCM are positively correlated, then as CCM increases, $H^*_\chi(S_{chiral})$ tends to increase, that is, $H^*_\chi(S_{chiral})$ encodes a significant amount of information about the degree of chirality. Finally, it is interesting to note that there is no relationship between the degree of chirality and the distribution of complexity evaluated in Section 4.2.2 (Figure S3).

4.3.4 | Information entropy of a racemization/deracemization process

When two enantiomers can interconvert, the equilibrium $(R) \rightarrow (S)$ is accompanied with no heat transfer ($\Delta H = 0$). Consequently, a racemization process is entropy driven, that is, $(R) \rightarrow (R,S)$ leads to a gain in entropy ($\Delta S > 0$), and is thermodynamically favored according to formula $\Delta G = \Delta H - T\Delta S$. In a last experiment, we will consider the following system:

$$(\mathrm{R}) \xrightarrow{\text{Racemization}} (\mathrm{R,S}) \xrightarrow{\text{Deracemization}} (\mathrm{S})$$

The joint entropy of each step of this process is calculated as follows

$$H_{\alpha(R,S)}(S_{chiral}, S_{achiral}) = H_{\alpha(R,S)}(S_{chiral}) + H_{\alpha(R,S)}(S_{achiral}) \quad (54)$$

where $\alpha(R,S)$ is the criterion that partitions the collection X into k equivalent classes. $\alpha(R,S)$ is based on a $(C_{(R)} + C_{(S)})$ mixture of geometric distances of atoms and their connectivities, $C_{(R)}$ and $C_{(S)}$ being the % amounts of each enantiomer.

In the partition procedure, throughout the racemization, any equivalent amount of (R) and (S) chiral topological elements are counted as diminishing chiral entropy and accordingly grouped with the achiral elements. This strategy allows us to obtain a decrease of the chiral elements during the racemization. In Figure 17 is plotted the joint entropy in function of the enantiomeric excess (ee) as obtained for 3D-optimized isomer 18. The growth of the joint entropy from the pure enantiomer to the racemate is nonlinear and is at its maximum when the mixture is racemic. Therefore, according to our information theory approach, racemization is a process of entropy increase and deracemization a process of decrease in entropy. It is worth mentioning that this result is consistent with the concept that the driving force for racemization is the increase in entropy.

From the definition of the joint entropy, we are then naturally led to compare the chiral and the achiral entropy evolution during the racemization process. This is illustrated by Figure 18 where we can see that $H_{\alpha(R,S)}(S_{chiral})$ decreases, whereas $H_{\alpha(R,S)}(S_{achiral})$ increases with decrease in ee. Interestingly, an

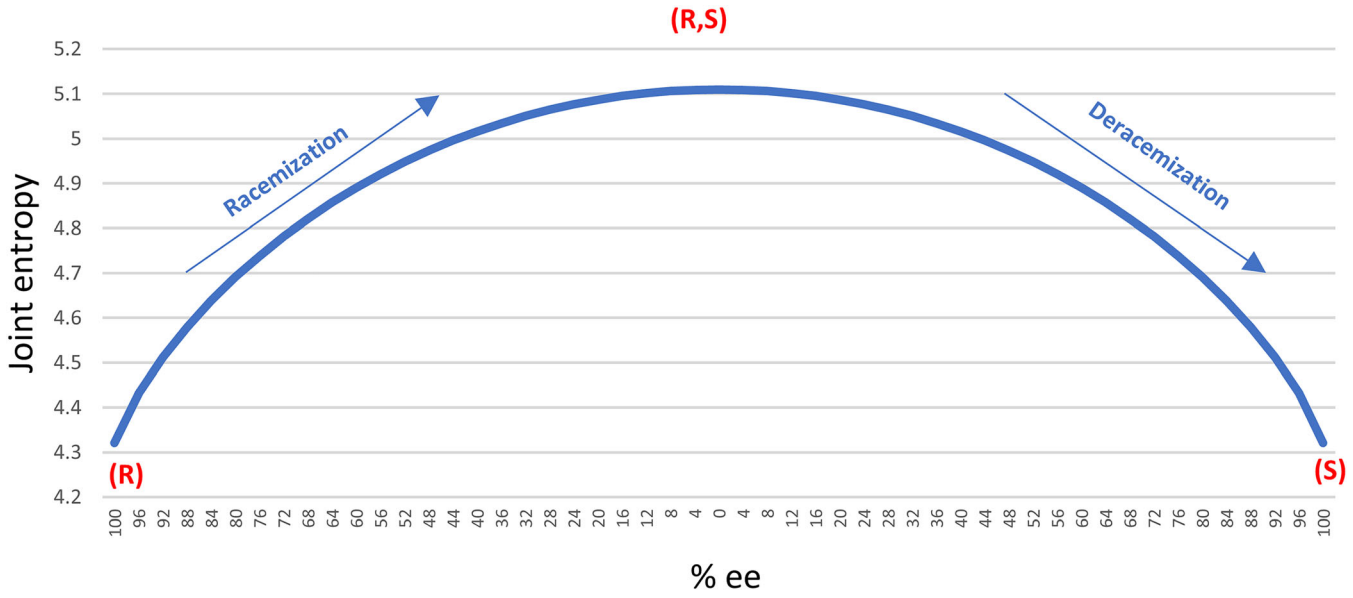

FIGURE 17 Common graph of the evolution of the joint entropy $H_{\alpha(R,S)}(S_{chiral}, S_{achiral})$ versus the ee (example shown is 3D-optimized isomer 18)

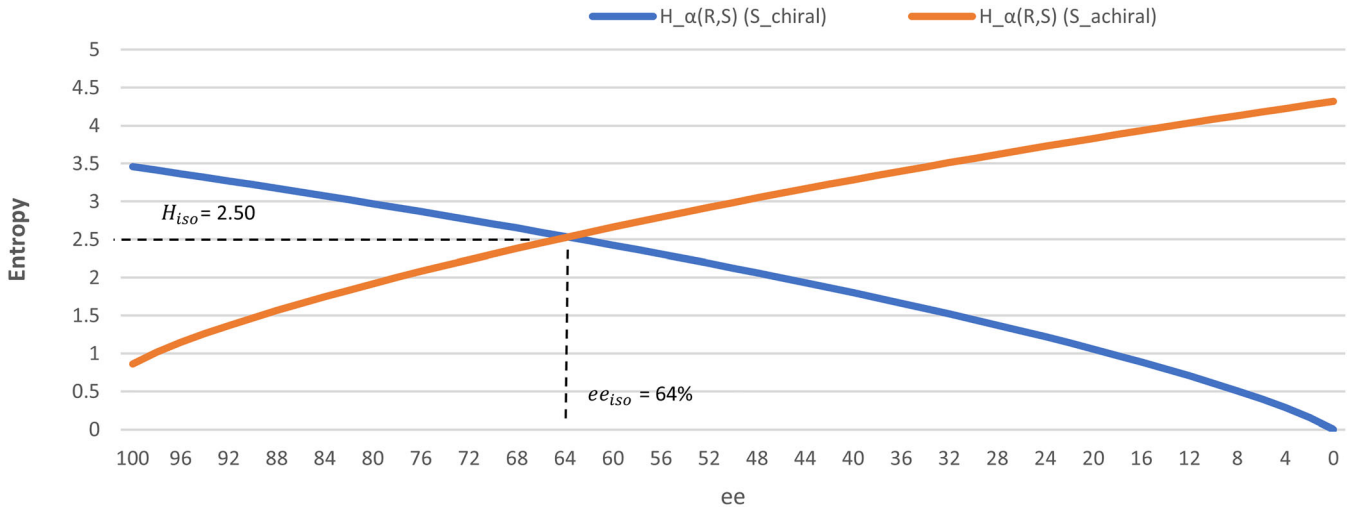

FIGURE 18 Common graph of the evolution of $H_{\alpha_{(R,S)}}(S_{chiral})$ and $H_{\alpha_{(R,S)}}(S_{achiral})$ versus the ee (example shown is 3D-optimized isomer 18)

isoentropic point is observed at the intersection points of the two curves. This isoentropic point (H_{iso}) and its corresponding isoentropic ee (ee_{iso}) define two new information measures for characterizing a given chiral molecule. One remarkable property of the ee_{iso} is its ability to differentiate chiral molecules which have a high degree of chirality from chiral molecules which have a lower value. Since, 90% of chiral molecules having a $ee_{iso} > 60$ exhibit the lowest degree of chirality (CCM < 7) and 77% of chiral molecules having a $ee_{iso} \leq 60$ exhibit the highest degree of chirality (CCM ≥ 7) (Figure S2). This is a result to note because it supports the idea that an enantiopure compound having a high degree of chirality needs a high amount of the other enantiomer to reach equality between S_{chiral} and $S_{achiral}$ entropies.

5 | GENERALIZATION OF THE METHOD TO ANY CHIRAL MOLECULE

In our information entropy approach derived from alkane graphs, atoms are represented by their connectivity and so the entropy measures cannot be applied to

molecule containing other atoms than carbons such as heteroatoms. In order to allow the generalization of these measures to any chiral molecule, atoms are no longer represented by their connectivity but by an information rich descriptor: the electrotopological state atom (E-state) index. This index developed by Hall and Kier[30] is a structural atomic descriptor encoding both the steric and electronic effects of the surrounding atoms. Consequently, the resulting combined 2D or 3D-topological distance and E-state-based information entropy is now applicable to almost any type of chiral organic structures. Preliminary results using Estate-based information entropy are encouraging. For example, by introducing our new chiral entropy measures in models, we were able to obtain a good prediction of the difference of biological activity between enantiomers in various biological data sets.[31] Those unpublished results show that in many cases, integrating chiral entropy improves prediction models. Other unpublished results concern the use of information entropy to investigate chiral recognition mechanisms. Any investigation of chiral ligand-binding complexes requires a knowledge of the non-covalent interactions that stabilize a given complex, that is, H-bond donor, H-bond acceptor, aromatic or lipophilicity. So, a last approach consisted to computationally decompose chiral entropy into the following entropy terms

$$H(S_{chiral}) = -\left(\sum_{i=1}^{k} p_i^{H_{donor}} \log p_i^{H_{donor}} + \sum_{i=1}^{l} p_i^{H_{acceptor}} \log p_i^{H_{acceptor}} + \sum_{i=1}^{m} p_i^{aromatic} \log p_i^{aromatic} + \sum_{i=1}^{n} p_i^{lip} \log p_i^{lip}\right)$$

where k, l, m, and n are the number of chiral partitions assigned to the different types of interactions.

Using this formula, we were able to reveal and quantify the individual entropy contributions of the different enantiospecific interactions occurring during chiral HPLC separations achieved on several commercially available chiral columns (Figures S7–10).

6 | CONCLUSION

In this work, we defined for the first time the concept of chiral information entropy and declined it in a number of ways. This concept has initiated an investigation of new information measures to capture the chiral complexity of chiral molecules. Different studies lead to a series of interesting conclusions:

- Chirality is usually presented as a qualitative concept: a molecule is chiral or not chiral. We have seen that in contrast with this binary concept, a different description of the chemical features may be useful to apprehend chirality not as a whole but through the contribution of chiral and achiral parts.
- Chiral information entropy measures can give an indirect access to the chiral complexity of a molecule according to topological criteria such as the connectivity or the electrotopological state of atoms. However, a certain subjectivity in the criteria choice is inevitable and thus any change in these criteria would allow to identify other different properties of chiral molecules. For example, one can imagine that chiral entropy measured according to the distribution of atom charges or lipophilicities can provide different insights on the measure of chiral complexity as well as chirality of molecule.
- Our approach revealed that distribution of chiral complexity is far from uniformity for a majority of the studied chiral molecules. Thus, a global measure of chiral complexity or an overall measure of chirality may not really reflect the distribution of complexity which can vary greatly even among similarly complex chiral compounds. For future research, it would be relevant to investigate if complexity distribution can have an influence on the enantioselective property of chiral molecules.

Although this article focused on chiral alkanes to provide the insights detailed above, this new conceptual framework highlights the potential of information theory to shed new light on the properties of a great number of chiral molecules. Undeniably, this leaves open a wide range of possibilities for the application of information theory to the field of chiral chemistry.

1. Clausius R. Uber einige fur anwendung bequeme formen der hauptgleichungen der nischen warmetheorie. *Annalen der Physik Und der Chemie*. 1865;125(7):353-400.
2. Boltzmann L. Weitere Studien uber das Warmegleichgewicht unter Gasmolekulen. *Sitzungsberichte Akademie der Wissenschaften*. 1872;66:275-370.
3. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379-423.
4. Brillouin L. *Science and Information Theory*. New York: Academic Press; 1956.
5. Rashevsky N. Life, information theory and topology. *Bull Math Biophys*. 1955;17(3):229-235.
6. Trucco E. A note on the information content of graphs. *Bull Math Biol*. 1956;18(2):129-135.
7. Mowshowitz A. Entropy and the complexity of graphs. I. An index of the relative complexity of a graph. *Bull Math Biophys*. 1968;30(1):175-204.
8. Bonchev D, Trinajstic N. Information theory, distance matrix and molecular branching. *J Chem Phys*. 1977;67(10):4517-4533.
9. Bonchev D. Information theoretic complexity measures. In: Meyers R, ed. *Encyclopedia of Complexity and Systems Science*. Vol. 5. Springer; 2009:4820-4838.
10. Dehmer M. Information-theoretic concepts for the analysis of complex networks. *Appl Artif Intell*. 2008;22(7–8):684-706.
11. Schneider TD. A brief review of molecular information theory. *Nano Commun Netw*. 2010;1(3):173-180.
12. Collet A, Ziminski L, Garcia C, Vigné-Maeder F. Chiral discrimination in crystalline enantiomer systems: facts, interpretations, and speculations. In: Siegel JS, ed. *Supramolecular Stereochemistry. NATO ASI Series (Series C: Mathematical and Physical Sciences)*. Vol. 473. Dordrecht: Springer; 1995:91-110.
13. Ribo JM. Chirality: the backbone of chemistry as a natural science. *Symmetry*. 2020;12(12):1982.
14. Loscrí V, Vegni AM. Enabling molecular communication through chirality of enantiomers. *ITU J-FET*. 2021;2(3):25-32.
15. Parrondo JMR, Horowitz JM, Sagawa T. Thermodynamics of information. *Nat Phys*. 2015;11(2):131-139.
16. Jaynes ET. Information theory and statistical mechanics. *Phys Rev*. 1957;106(4):620-630.
17. Brown RD, Martin YC. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci*. 1996;36(3): 572-584.
18. Faulon JL, Visco DP, Roe D. Enumerating molecules. *Rev Comput Chem*. 2005;21:209-286.
19. Koch M, Duigou T, Carbonell P, Faulon JL. Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0. *J Cheminform*. 2017;9(64):1-17.
20. Fiedler M. Algebraic connectivity of graphs. *Czechoslov Math J*. 1973;23(2):298-305.
21. Rajaram R, Castellani B. An entropy based measure for comparing distributions of complexity. *Phys A Stat Mech Appl*. 2016;453:35-43.
22. Jost L. Entropy and diversity. *Oikos*. 2006;113(2):363-375.
23. Castellani B, Rajaram R. Past the power law: complex systems and the limiting law of restricted diversity. *Complexity*. 2016; 21(S2):99-112.
24. Hubert L, Arabie P. Comparing partitions. *J Class*. 1985;2(1): 193-218.
25. Tosco P, Stief N, Landrum G. Bringing the MMFF force field to the RDKit: implementation and validation. *J Chem*. 2014;6(1): 37-40.
26. Piras P, Sheridan R, Sherer E, Schafer W, Welch C, Roussel C. Modeling and predicting chiral stationary phase enantioselectivity: an efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy. *J Sep Sci*. 2018;41(5):1365-1375.
27. Petitjean M. Chirality and symmetry measures: a Transdisciplinary Review. *Entropy*. 2003;5(3):271-312.
28. Zayit A, Pinsky M, Elgavi H, Dryzun C, Avnir D. A web site for calculating the degree of chirality. *Chirality*. 2011;23(1):17-23.
29. Zabrodsky H, Avnir D. Continuous symmetry measures. 4. Chirality. *J Am Chem Soc*. 1995;117(1):462-473.
30. Kier LB, Hall LH, Frazer JW. An index of electrotopological state for atoms in molecules. *J Math Chem*. 1991;7(1):229-241.
31. Schneider N, Lewis R, Fechner N, Ertl P. Chiral cliffs: investigating the influence of chirality on binding affinity. *ChemMedChem*. 2018;13(13):1315-1324.