

Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry

Tharun Kumar Allu[†] and Tudor I. Oprea^{*‡}

Department of Computer Science and Division of Biocomputing, University of New Mexico School of Medicine, MSC 11 6145, 1 University of New Mexico, Albuquerque, New Mexico 87131

Received April 19, 2005

Methods that rapidly evaluate molecular complexity and synthetic feasibility are becoming increasingly important for in silico chemistry. We propose a new metric based on relative atomic electronegativities and bond parameters that evaluate both synthetic and molecular complexity (SMCM) starting from chemical structures. Against molecular weight, SMCM has the lowest fraction of adjusted variance ($R^2=0.535$) on a series of 261 048 diverse compounds, when compared to the complexity metric of Baron and Chanon ($R^2=0.777$; *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272) and Rücker ($R^2=0.895$ for log complexity values; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386), respectively. These metrics are in general agreement when the metabolic synthesis of cholesterol from S-3-hydroxy-3-methyl-glutaryl coenzyme A is monitored, indicating that SMCM can be useful in discerning increases in complexity. Because the presence of substructure patterns can be directly incorporated into this scheme, SMCM is relatively straightforward and can be easily tailored to rapidly evaluate virtual (combinatorial) libraries and high throughput screening hits.

INTRODUCTION

As in silico chemistry is gaining importance in drug discovery, so is the rapid evaluation of virtual chemistry. Synthetic feasibility, or the world of ‘could-be-made’, as opposed to ‘difficult-to-make’ or ‘impossible’ chemicals, needs rapid computer-based evaluation tools that can potentially be time- and resource-saving for the end user. Efforts for computer-assisted organic synthesis (CAOS)^{1,2} and evaluation of chemical feasibility have reached a mature stage in, e.g., LHASA (logic and heuristics applied to synthetic analysis),^{3,4} WODCA,^{5,6} and CAESA (computer assisted estimation of synthetic accessibility).⁷ Designed for detailed organic chemistry evaluation, these schemes may not be easily converted for massive in silico efforts such as virtual screening or library design and enumeration. General models to evaluate molecular complexity^{2,8} are needed for both in silico chemistry⁹ as well as for structure–activity relationships.¹⁰ A recent evaluation of LHASA’s ‘strategic bonds’¹¹ in view of the mathematical rigor of topological molecular complexity¹² indicates a good degree of correspondence between heuristic rules (LHASA) and mathematical models.¹² In the absence of tools for rigorous evaluation of molecular complexity, molecular weight (MW) is often used as a rapid, albeit crude indicator for complexity. This is an oversimplified measure, since molecules with heavy atoms such as bromine and iodine might still be simple, yet would have a significantly higher MW compared to their, e.g., hydrogen or fluoro substituted analogues.

We propose a simple, empirical Synthetic and Molecular Complexity Metric (SMCM) that departs from MW in a consistent and reproducible manner and is intended as an

Table 1. Atoms and Their Respective Sanderson Electronegativity Relative to Carbon¹⁵

atom	electronegativity	atom	electronegativity
B	0.851	P	1.086
C	1.000	S	1.235
N	1.149	Cl	1.384
O	1.297	Br	1.244
F	1.446	I	1.103

Table 2. Bond Parameters¹⁵ Computed with Atomic Number

atom	atom	single	double	triple	aromatic
C	C	1.000	0.500	0.333	0.667
C	N	0.857	0.429	0.286	0.571
C	O	0.750	0.375		
C	F	0.667			
C	P	0.400			
C	S	0.375	0.188		0.250
C	Cl	0.353			
C	Br	0.171			
C	I	0.113			
N	N	0.735	0.367		0.490
N	O	0.643	0.321		0.423
O	S	0.281	0.141		

intuitive ‘plug-in’ for rapid chemical feasibility and complexity evaluation. Starting from the number of different atom types and bonds, SMCM takes into account a hybridization state at the simple ‘aliphatic’ or ‘aromatic’ level and other structural features such as the number of (vicinal) chiral centers, geminal substitutions, and the number of rings and spiro carbons as well as the type of (fused) ring systems per molecule. We also present unique features as an illustration of how particular chemotypes can be empirically added into SMCM and how certain functional groups and repeating structures can yield a lower (perceived) SMCM value for a molecule.

Using information theory, Bertz proposed a rigorous mathematical approach for calculating the complexity of

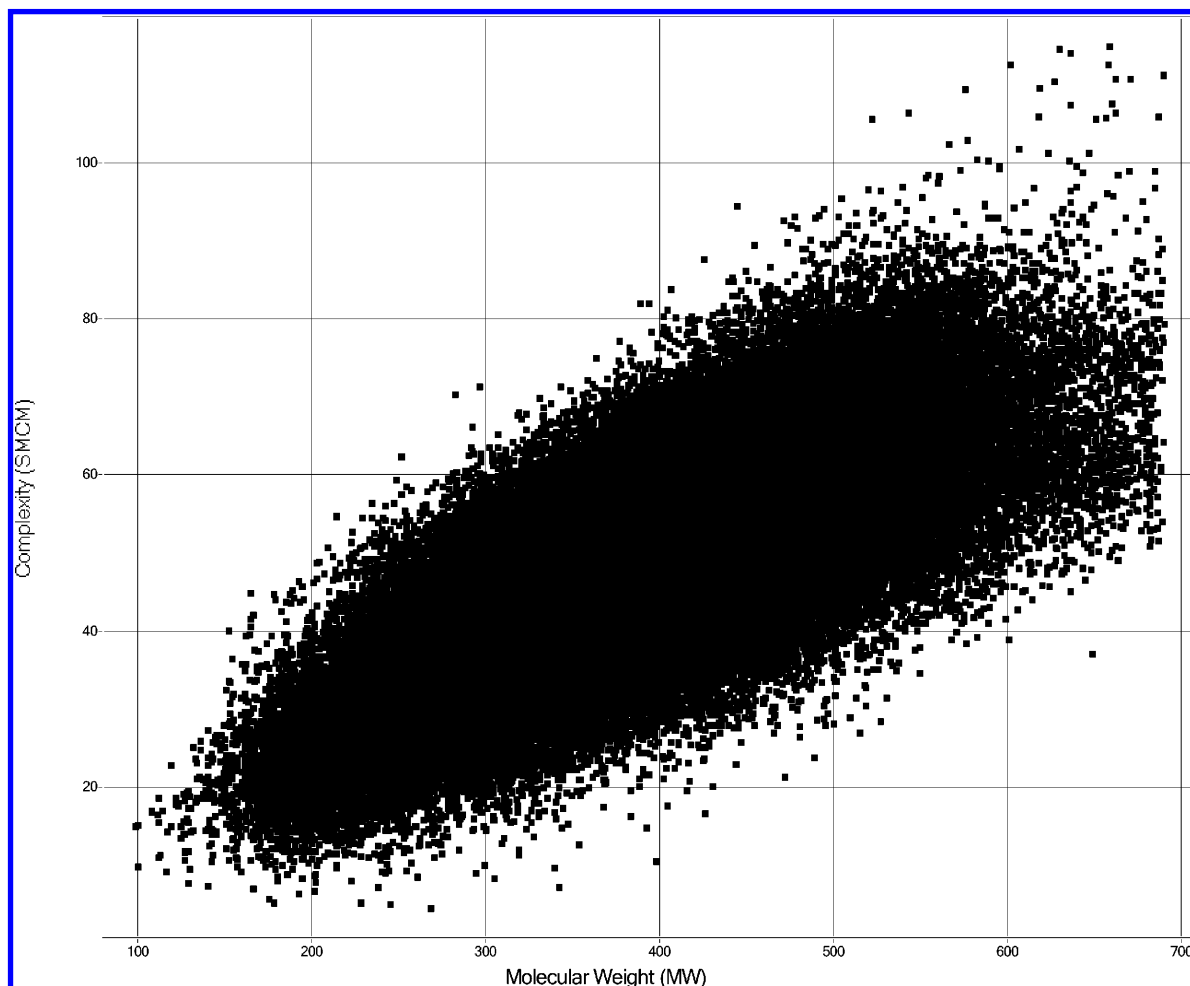
* Corresponding author phone: +1 5052723694; fax: +1 5052728738; e-mail: toprea@salud.unm.edu.

[†] Department of Computer Science.

[‡] Division of Biocomputing.

Table 3. Examples of SMARTS Patterns for Chemical Substructures Captured in the Final SMCM Score

SMARTS	description
<chem>[\$([C!D1]-@C(=O)-@[N!a])]</chem>	peptides, cyclic
<chem>[\$([C!D1]-!@C(=O)-!@[N!a])]</chem>	peptides, acyclic
<chem>[\$([#6]=,[*R!C!c]=,[#6R]~@[*R!C!c])]</chem>	C attached to Hetero in ring
<chem>[\$([A!C!H](-!@[#6])~!@[#6])]</chem>	hetero attached to 2 carbons not in ring
<chem>[\$([#7]=[C!\$*a])~!@[N!H0])]</chem>	amidine, guanidine
<chem>[\$([#8]=[#6H0]~!@[N!H2])!\$(NC!#6))]</chem>	nonterminal amide
<chem>[\$([#8]=[#6](-!@[NH1,NH0])~!@[N!H2])]</chem>	nonterminal urea
<chem>[\$(O=[CD3]([#6])([#6])!\$(!#6)=CC=O))]</chem>	ketone, not diketo
<chem>[\$([#16X2v2])([#6])([#6])]</chem>	thio ether
<chem>[\$([#8X2H0])([#6])([#8D1])!\$(O~C(~O)~O))]</chem>	carboxyl ester, not carbonate
<chem>[\$([#8X2v2])([#6])([#6])]</chem>	ether oxygen
<chem>C1:c-@C-@N-@C-@C-@O-@C-@c:c-1</chem>	SMARTS from ICCB's Diversity-Oriented Synthesis approach ¹⁹
<chem>[N,O,C]C(=O)C1=C[C@H](*)C[C@H](O)O1</chem>	
<chem>C[C@H]1O[C@H]~3O[C@H]~C~2~C~C[C@H]([C@@H]~1)[C@@H]~23</chem>	
<chem>C[C@H]1O[C@H]~2O[C@H][C@H][C@@H]~3~C~C~C~1[C@@H]~23</chem>	
<chem>C[C@H]2C[C@]14~C~C~C~C[C@H]1Oc3cccc(CN2)c34</chem>	
<chem>[\$([#6][C@H]1[C@@H]([#6])O[C@@H](-a)O1</chem>	
<chem>C2=CC[C@@H]1C(=O)~*~*~C(=O)[C@@H]1[C@@H]2-a</chem>	
<chem>C2=CC[C@@H]1C(=O)~*~*~C(=O)[C@@H]1[C@@H]2-a</chem>	
<chem>a-[CH,CH2;R0;0*]</chem>	from ref 17
<chem>[R;0*]-[CH2R0,NHR0,OR0;0*]-[R]</chem>	
<chem>*-[CD3H,ND2;R0;0*](-a)-a</chem>	
<chem>[a]-&!@[a]</chem>	
<chem>[NR;0*]-[CD3R0;0*](=O)-[R]</chem>	
<chem>[NR;0*]-[CD2R0;0*]-[R]</chem>	
<chem>[NR;0*]-[CD2R0;0*]-[CD2,CD3,OD2,ND2,ND3,aD2,aD3]</chem>	
<chem>a-[NHR0]-[CR0;0*](=O)-[OR0,NR0;0*]</chem>	
<chem>[CR,NR]=[CR]-&!@[a]</chem>	
<chem>[\$([#6](-!@[#7])~!@[#7!H0])]</chem>	NHCN not in ring
<chem>[\$([#6](@[#7])@[#7!H0])]</chem>	NHCN in ring
<chem>[\$([A!C!H](-@[#6])~@[#6])]</chem>	hetero attached to 2 carbons in ring
<chem>[\$([#6](=[#8])~[#6])([#6])]</chem>	diketo, keto-Ryl
<chem>[\$([#16X3v4,#16X4v6])([#6])([A])]</chem>	hypervalent sulfur
<chem>[\$(C(~O)(~O)~O)]</chem>	carbonate

**Figure 1.** SMCM versus MW; $R^2 = 0.535$, $N = 261\,048$.

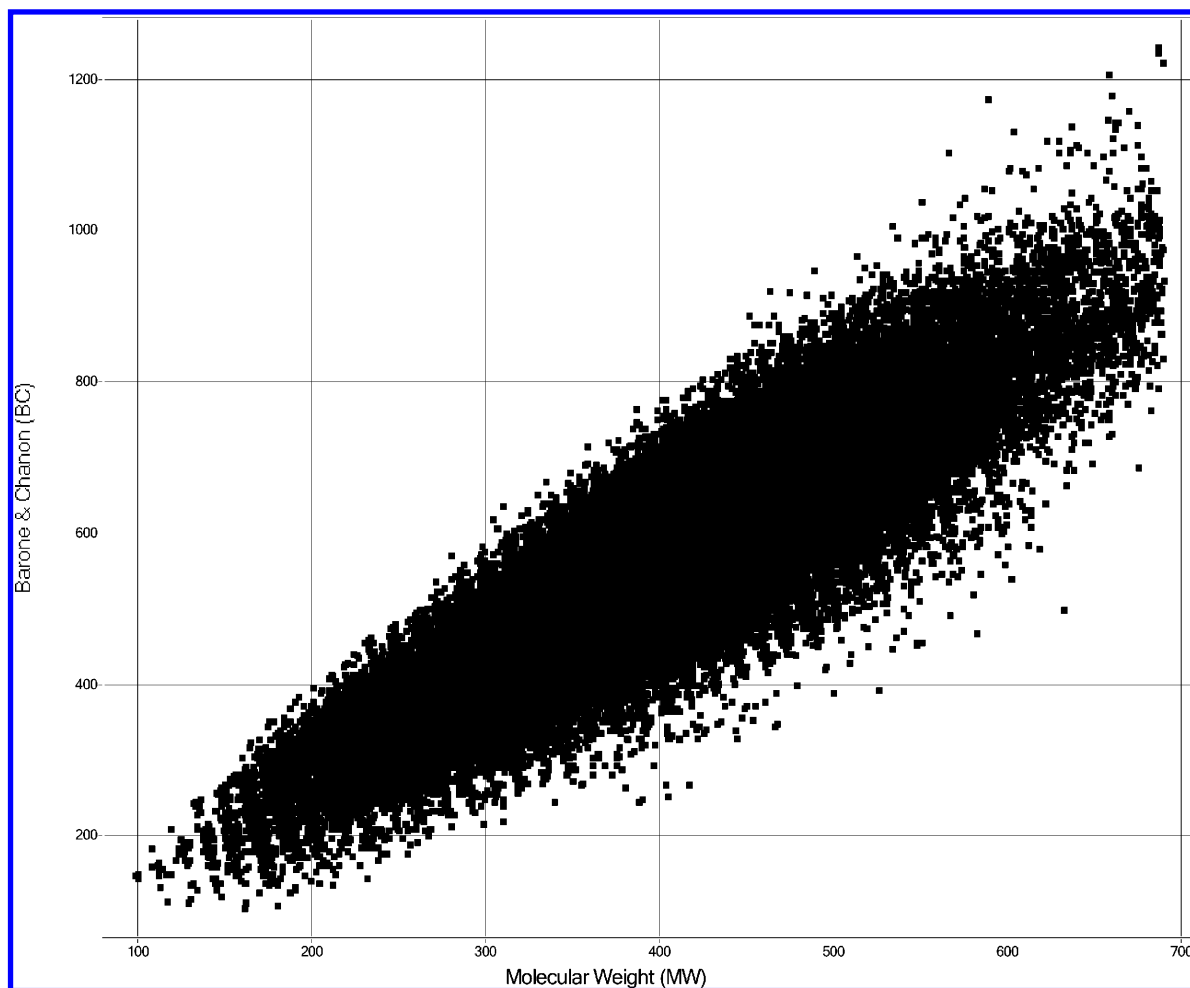


Figure 2. BC versus MW: $R^2 = 0.777$, $N = 261\,048$.

chemical structures.⁸ Whitlock's intuitive approach sums 4 times the number of rings, 2 times the number of nonaromatic unsaturations, 1 times the number of heteroatoms, and 2 times the number of chiral centers.¹³ Barone and Chanon¹⁴ ameliorated this scheme by considering only the connectivity of an atom in a molecule and the size of the rings; all other atoms other than carbon contribute equally to the complexity value. Neither the total walk count (TWC) from Rücker¹² nor Barone and Chanon's complexity¹⁴ (BC) consider fused rings, chiral centers, and chemical substructures or functional groups whose presence could alter the synthetic feasibility of a molecule. SMCM attempts to account for such features, e.g., fused ring systems and substructures that reduce the complexity of a molecule: For example, most chemists consider that vicinal (1–2) chiral centers are simpler to introduce compared to 1–3 chiral centers. Empirically, SMCM does not score non-carbon atoms equally, as the presence of each element (and hybridization state), based on different electronegativity¹⁵ and reactivity, has a different influence on the synthetic feasibility of a molecule.

By definition electronegativity is the ability of an atom or a molecule to attract shared electrons to itself. This is an important property in chemical reactions. By using electronegativity of the atoms for the base SMCM calculations we are trying to remove subjective assignment of scores as in the other complexity measures discussed above. Although the main goal of this method is to depart correlation with MW, it should be noted that SMCM takes into consideration

the chemical nature and properties of the atoms. The other parameters used for calculating SMCM as shown in the method section are assigned based on previous medicinal chemistry experience and are pure integer values depicting the difficulty of synthesis when these are present. This is not a definitive solution to the problem, but it is designed to be fast and is used to estimate synthetic accessibility of large libraries of real or virtual compounds.

METHOD

Intuitive and empirical, SMCM is designed with simplicity and speed in mind. Table 1 shows all the atoms and their relative electronegativity (EN) values.¹⁵ We scan for these atoms in the molecule which are in the SMILES¹⁶ format and multiply the number of atoms found with their EN value. Table 2 represents all the bond parameters computed with the atomic number.¹⁵ We scan for these bonds in our molecule and multiply the number of occurrences of these with the values given in the table. We also consider chiral centers and add two times the total number of chiral centers and one time the number of adjacent chiral centers to the resulting SMCM value. We also consider rings and ring systems: As rings of size 3, 4, 7, 8, and 9 are relatively harder to synthesize compared to rings of size 5 and 6 (A. T. Balaban, personal communication) this is taken into account by assigning them a relative score of 2 and 1, respectively. Rings of size 10 or higher are also given a score of 1. Each ring system is given a score of 1, fused rings get

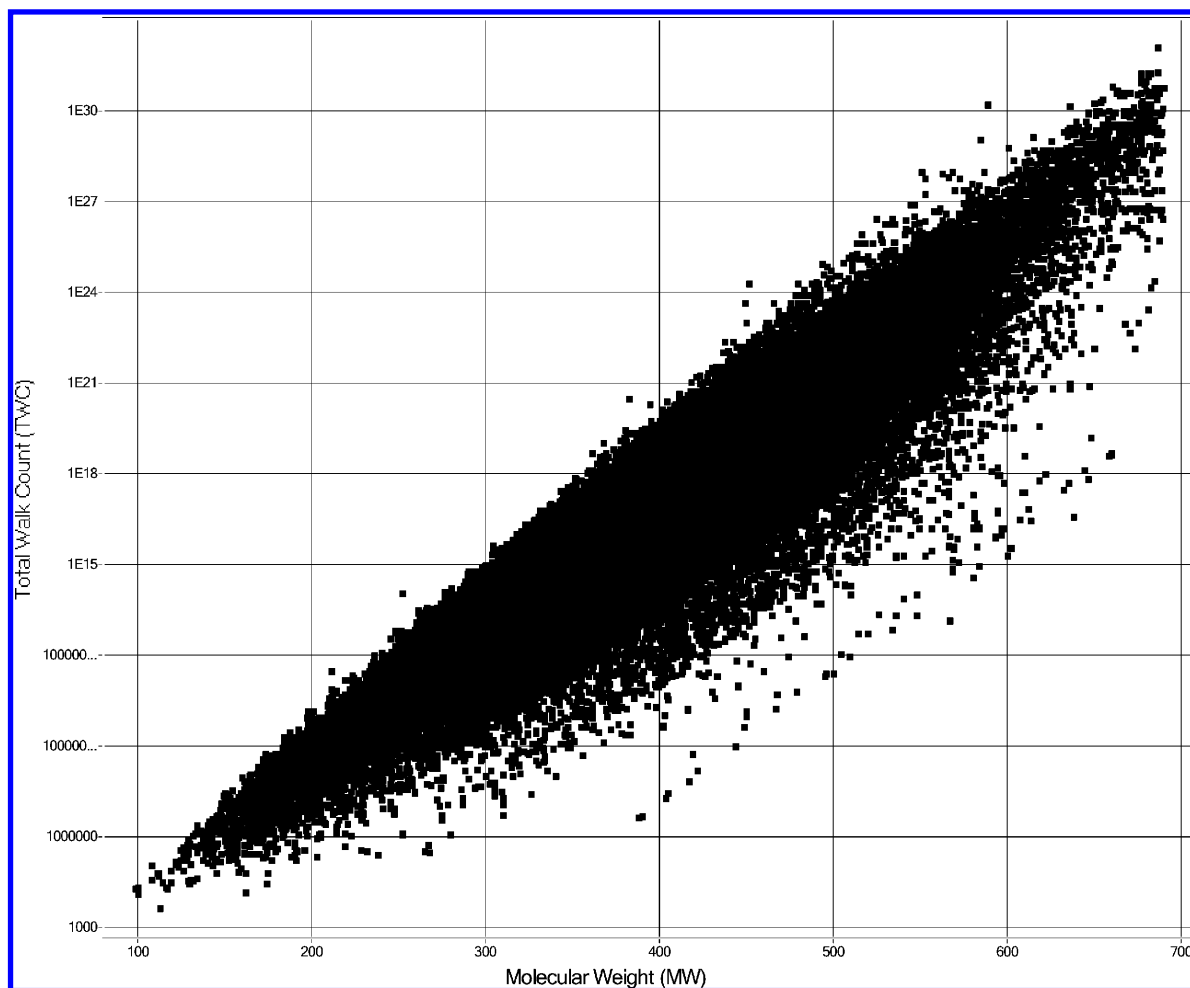


Figure 3. TWC (log10 values) versus MW: $R^2 = 0.895$, $N = 261\,048$.

a score of 2, spiro carbon gets a score of 3, and bridged atoms get a score of 4. All the above features contribute to the positive score of the SMCM value. As some moieties in molecules may be peptides, ketones, carbonate, ether oxygen, etc., as shown in Table 3, their presence is likely to reduce the complexity of the molecule. Therefore, we subtract two times the number of occurrences of these features from SMCM. This is a key element of SMCM scoring as these features play an important role in deciding which molecules are easier to synthesize even though they are of similar molecular weight (MW). This makes SMCM very useful for chemists as it does not correlate much to molecular weight.

For the sake of illustration, we include scaffolds used in the diversity-oriented synthesis (DOS) approach¹⁸ from the ChEMBL database¹⁹ (Table 3). All the SMARTS shown in Table 3 are assigned negative values in the total SMCM score, as they are expected to improve the chemical accessibility of a molecule. While the presence of all these features in a molecule is expected to make its synthesis easier, one cannot determine exactly by what degree the complexity is reduced, as it will surely not drop to zero. Without additional information, we consider these features to be additive, i.e., each feature that is included with the final SMCM score has the same weight.

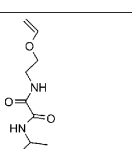
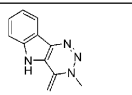
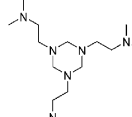
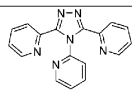
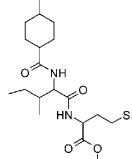
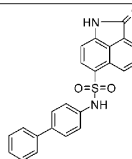
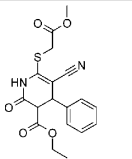
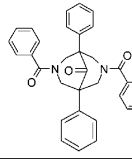
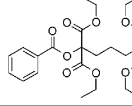
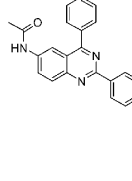
APPLICATION

Lead discovery requires an increasingly large number of virtual libraries and high throughput screening (HTS) hits

to be evaluated *in silico*. Besides biological activities (in HTS), the parameters used to prioritize compounds for reevaluation or synthesis are reagent availability, the cost of starting materials, the possibility of identifying few-steps, high-yield products, purity, and chemical space neighborhood. Since molecular similarity/diversity metrics have been used to evaluate virtual libraries and HTS hits, we propose this metric for the rapid evaluation of both molecular complexity and synthetic accessibility. The need to judge which molecules are easier to synthesize has been discussed since the early days of combinatorial chemistry. Alone or in combination with the octanol/water partition coefficient (logP)²⁰ or polarizability^{21,22} molecular weight has been used by most chemists for this task. SMCM is designed to reduce the dependence on MW, while taking into consideration simple substructural features of molecules. Randić et al.²³ considered simple molecular graphs which are not weighted (graphs whose edges or vertices have the same properties) as their complexity measure. The main disadvantage of this method is having a complexity metric that ignores heteroatoms or other chemical features. SMCM does not depend on the molecular graph; it is based on chemical features and the properties of certain atom types and their electronegativity.¹⁵

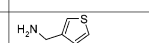
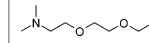
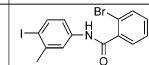
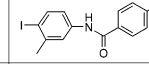
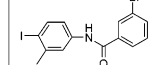
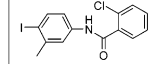
We wrote programs that calculated BC complexity, TWC, and SMCM based on their respective algorithms. We applied the SMCM score on 261 048 small molecules (MW < 700) from Chemical Diversity Labs (CDL).²⁴ Figure 1 shows a

Table 4. Pairwise Examples Illustrating Molecules with the Same MW but Different Complexity Values

Structure	SMCM	BC	TWC	Log ₁₀ (TWC)	MW
	10.867	183	61,303,200	7.7875	200.235
	37.65	345	5,419,020,480	9.7339	200.197
	26.749	261	513,343,743,957	11.710	300.487
	47.447	561	1,029,834,328,797,804	15.013	300.318
	41.524	369	542,126,741,142,169	14.734	400.577
	65.612	657	2,779,412,998,382,050,704	18.444	400.451
	39.998	519	71,646,805,748,866,224	16.855	500.309
	89.464	831	3,270,287,731,653,259,379,200	21.515	500.587
	54.818	723	154,050,521,569,869,325,336,576	23.188	600.61
	84.722	1080	65,166,920,647,493,496,443,508,883,456	28.814	600.668

plot between the SMCM values and MW (<700) for these 261 048 CDL molecules ($R^2 = 0.535$). The spread of the graph indicates some molecules with equal MW have rather different SMCM score values. While the R^2 value is significant, SMCM is less dependent on MW compared to BC complexity (Figure 2) ($R^2 = 0.777$). The total walk count¹² (TWC) proposed by Rücker as a general index for complexity yields a rather low correlation, $R^2 = 0.05$ for the same data set. A high TWC does not necessarily mean that a compound is difficult to synthesize, nor does

Table 5. Examples Illustrating Different Values of Molecular Complexity

S. No.	Structure	SMCM	TWC	Log ₁₀ (TWC)	BC
1		11.409	4348	3.63829	132
2		17.034	11288150	7.052623	138
3		32.029	16688312416	10.22241	387
4		32.029	15548172166	10.19168	387
5		32.029	15700125956	10.1959	387
6		32.169	16688312416	10.22241	387

a low TWC say that a compound is easy to synthesize (C. Rücker, personal communication). However, we found that TWC correlates with MW if one plots log values of TWC against MW (Figure 3) ($R^2 = 0.895$). The data on which the calculations are carried out are present at <http://pimento.health.unm.edu/complexity/smcmm.data.tab>.

Our own effort to depart from MW based on a simple, empirical scheme is countered by the natural trend of higher MW molecules to have an increasingly higher number of atoms, which is paralleled by higher complexity values. We note that the DOS-related SMARTS patterns included in Table 3 do not match the current data set, nor did we explicitly include CDL-related SMARTS patterns into the final SMCM score, as their presence would have biased our own scoring scheme. We anticipate that including such patterns would further decrease the correlation between SMCM and MW.

Several examples (Table 4) illustrate why moving away from MW can be useful: Five pairs of molecules chosen from the extremes of the Figure 1 graph have almost equal MW and significantly different SMCM scores. These examples are best evaluated by visual inspection, as more complex molecules appear to be less amenable to easy synthesis. Examining these structures, one can conclude that, while MW may be a crude measure for complexity, it is far from informative with respect to complexity and synthetic feasibility. Therefore, the effort to rapidly evaluate both properties while reducing the correlation with MW appears to be justified—in particular for massive evaluation of in silico libraries (e.g., over 10^5 compounds).

One possible advantage of SMCM over general complexity measures is shown in Table 5: Examining structures 1 and 2, it becomes apparent that structure 2 can be easily synthesized due to its symmetry, compared to structure 1. Considering structures 3–6, we find that TWC scores differ based on the aromatic substitution position of the halogen. When heteroatoms are present, TWC places a value at position (i, i)—where **i** is the atom (node) number in the molecular graph). The value given for heteroatoms in the adjacency matrix, as implemented in the original work,¹² assigns 2 for nitrogen and 1 for oxygen. We implemented a value of 1 for all other atoms. Thus, a change in the substituent position yields different TWC scores. We note that both SMCM and BC values do not differ for molecules 3–6.

Table 6. Different Values of Molecular Complexity SMCM, BC, and log (TWC) for the Cholesterol Synthesis Pathway from S-3-Hydroxy-3-methyl-glutaryl-CoA to Cholesterol²⁵

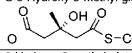
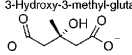
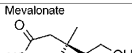
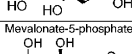
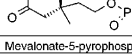
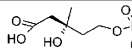
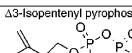
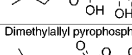
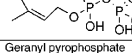
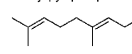
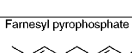
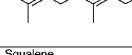
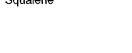

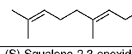


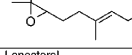
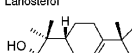
Molecule Name	SMCM	BC	Log ₁₀ (TWC)	MW
S-3-Hydroxy-3-methyl-glutaryl CoA 	89.76	990	39.4669	911.611
3-Hydroxy-3-methyl-glutarate 	21.485	156	5.5843	160.125
Mevalonate 	19.813	126	4.849	148.157
Mevalonate-5-phosphate 	24.79	165	7.88441	228.137
Mevalonate-5-pyrophosphate 	28.767	204	10.6753	308.117
Δ^3 -Isopentenyl pyrophosphate 	20.001	147	8.33604	246.092
Dimethylallyl pyrophosphate 	21.001	150	8.34187	246.092
Geranyl pyrophosphate 	30.001	216	11.2627	314.209
Farnesyl pyrophosphate 	39.001	282	14.1831	382.326
Squalene 	54	390	13.9893	410.718
(S)-Squalene-2,3-epoxide 	57.797	420	14.5134	426.717
Lanosterol 	70.047	597	17.631	426.717
14-Demethylsterol 	68.047	579	17.0171	412.691
$\Delta^8,24$ -Cholestadien-3 β -ol 	64.047	570	16.6701	384.638
$\Delta^7,24$ -Cholestadien-3 β -ol 	63.047	585	17.0402	384.638
Δ^7 -Cholesten-3 β -ol (Lathosterol) 	64.047	567	16.8104	386.654
7-dehydro-Cholesterol 	64.047	585	17.4786	384.638
$\Delta^5,24$ -Cholestadien-3 β -ol (Demosterol) 	63.047	585	16.8682	384.638
Cholesterol 	64.047	567	16.8602	386.654

Table 6 shows the SMCM, BC, and log (TWC) values for the pathway from S-3-hydroxy-3-methyl-glutaryl-CoA to cholesterol. Looking at the SMCM and other complexity values we observe that there is general agreement and that SMCM parallels the increase in complexity, as computed by BC and TWC (given in log form) during the biochemical synthesis of cholesterol (Table 6). All three methods agree that (S)-squalene-2,3-epoxide is simpler compared to lanosterol, although they have the same MW, and that lanosterol has the highest complexity value. However, TWC attributes higher complexity to farnesyl pyrophosphate, compared to squalene, in contrast to SMCM and BC. In turn, SMCM outputs the same value (64.047) for 4 steroids: $\Delta^8,24$ -cholestadien-3 β -ol, lathosterol, 7-dehydro-cholesterol, and cholesterol. Minor differences are noted in the other two metrics, BC and TWC. These four molecules differ mostly by the presence of double bonds, e.g., at the B-ring in 7-dehydro-cholesterol, or at C24 in the side-chain, e.g., in $\Delta^8,24$ -cholestadien-3 β -ol. Such subtle variations are obviously not picked up by SMCM. Therefore, for fine-tuning of complexity values, multiple metrics are recommended.

CONCLUSION

The issue of evaluating molecular complexity while incorporating synthetic feasibility is not easily tractable: Most chemists are familiar with certain reactions, but even within a given pair of reagents, e.g., aliphatic monoamines and monocarboxylates, some will yield amides easier than others. In this work, we proposed a simple and flexible metric to calculate molecular complexity, which is aimed at lowering the correlation with molecular weight. Results based on over 260 000 molecules indicate that the SMCM score is the least correlated to MW, when compared to two other known complexity measures. This metric, by means of SMARTS pattern recognition, incorporates chemical tractability in a user-defined manner. The increase in complexity during metabolic synthesis of cholesterol indicates that SMCM is in general agreement with two other methods but is less likely to discern relatively small differences in the chemical structure. However, SMCM can become useful when evaluating large sets of virtual (combinatorial) libraries, as encoded SMARTS particular to the product(s) of the (virtual) chemical reactions can be incorporated into the complexity metric evaluation scheme. We recommend the use of SMCM in prioritizing virtual screening results, HTS hits, and compound libraries for synthesis. However, for in-depth analysis of complexity, the use of multiple metrics is advised.

ACKNOWLEDGMENT

This work was supported by New Mexico Tobacco Settlement funds. We thank Prof. Dr. Alexandru T. Balaban (Texas A&M University at Galveston) and Dr. Cristian Bologa (UNM Division of Biocomputing) for helpful comments. Metaphorics LLC of Santa Fe, NM is gratefully acknowledged for access to the EMPATH database.

REFERENCES AND NOTES

- (1) Gasteiger, J. A case study in computer-assisted organic synthesis design. *Chim. Oggi*. **1989**, 7, 65–72.

- (2) Dengler, A.; Fontain, E.; Knauer, M.; Stein, N.; Ugi, I. Competing concepts in CAOS. *Recl. Trav. Chim. Pays-Bas*. **1992**, *111*, 262–269.
- (3) <http://lhasa.harvard.edu/and> Pensak, D. A.; Corey, E. J. LHASA – logic and heuristics applied to synthetic analysis. *ACS Symp. Ser., Comput.-Assisted Org. Synth., Symp. Cent. Meet. Am. Chem. Soc.* **1977**, *61*, 1–32.
- (4) Corey, E. J. Computer-assisted analysis of complex synthetic problems. *Quart. Rev., Chem. Soc.* **1971**, *25*, 455–482.
- (5) <http://www2.chemie.uni-erlangen.de/software/wodca/index.html> and Ihlenfeldt, W.; Gasteiger, J. Computer-assisted planning of organic syntheses: the second generation of programs. *Angew. Chem., Int. Ed.* **1996**, *34*, 2613–2633.
- (6) Gasteiger, J.; Pfortner, M.; Sitzmann, M.; Hollering, R.; Sacher, O.; Kostka, T.; Karg, N. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspect. Drug Discovery Des.* **2000**, *20*, 245–264.
- (7) http://www.chem.leeds.ac.uk/ICAMS/new_web/CAESA/caesa.htm and Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34–50.
- (8) Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.
- (9) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.
- (10) Bonchev, D. Overall connectivity and topological complexity: A new tool for QSPR/QSAR. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach Science Publishers: Amsterdam, 1999; pp 361–401.
- (11) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General methods of synthetic analysis. Strategic bond disconnections for bridged polycyclic structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116–6124.
- (12) Rücker C.; Rücker G.; Bertz S. H. Organic synthesis—art or science? *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386.
- (13) Whitlock, H. W. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982–7989.
- (14) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272.
- (15) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Design of Topological Indices. Part 10.¹ Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395–401.
- (16) <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
- (17) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224–232.
- (18) Burke, M. D.; Schreiber, S. L. A planning strategy for diversity-oriented synthesis. *Angew. Chem., Int. Ed.* **2004**, *43*, 46–58.
- (19) <http://chembank.med.harvard.edu/>.
- (20) Leo, A. J. Calculating log P oct from structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
- (21) Glen, R. C. A fast empirical method for the calculation of molecular polarizability. *J. Comput. Mol. Des.* **1994**, *8*, 457–466.
- (22) Leo, A.; Weininger, D. CMR3. Daylight Chemical Information Systems, Santa Fe, New Mexico, <http://www.daylight.com/>, 1995.
- (23) Randic, M.; Plavsic, D. Characterization of Molecular Complexity. *Int. J. Quantum. Chem.* **2003**, *91*, 20–31.
- (24) ChemNavigator, Inc., San Diego, CA, <http://www.chemnavigator.com/>.
- (25) EMPATH Metabolic Pathways from <http://cabinet.metaphorics.com/> and Povolna, V.; Dixon, S. D.; Weininger D. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, 2005; Chapter 10, pp 241–269.

CI0501387