Class and Home Problems (CHP) present scenarios that enhance the teaching of chemical engineering at the undergraduate or graduate level. Submissions must have clear learning objectives. CHP papers present new applications or adaptations that facilitate learning in specific ChE courses. Submit CHP papers through journals.flvc.org/cee, include CHP in the title, and specify CHP as the article type.

MOLECULAR DESIGN USING THE SIGNATURE MOLECULAR DESCRIPTOR

John A. Krew,[1] Donald P. Visco, Jr.,[2] and Phillip D. Bertke[2]
1. Revere High School • Richfield, OH 44286
2. The University of Akron • Akron, OH 44325

INTRODUCTION

Product design is an emerging area within the discipline of chemical engineering. While in the past product design was left to practicing engineers as well as research and development professionals, more recently, product design courses have started to appear within the chemical engineering curriculum, both as electives and required courses, though slowly. Two textbooks are available in this area,[1, 2] while another book discusses both product and process design.[3] One example of a course offering related to product design is at Rowan University where freshman have been introduced to the topic via a commercial beer project.[4] Additionally, Georgia Tech offers a course for students in chemical product design.[5] At Louisville, they have used a module to incorporate chemical product design into their design course.[6]

An important element of any product design, especially from a computational perspective, requires models that allow the user to predict certain properties of interest based on other attributes of the compound, typically the structure. Most of these models utilize an approach that deconstructs a molecule into smaller groups, with each of the groups contributing to the overall property of that molecule by a specific amount. Once these models are developed, they are then used in algorithms to help identify compounds with optimal predicted properties.

In this manuscript, we demonstrate the use of a relatively new approach to deconstructing molecules, the Signature molecular descriptor, as a means of designing molecules with desired properties. Signature has previously been used for product design of compounds such as foam blowing agents,[7] solvents,[8] surface tension reducing agents,[9] and protein inhibitors.[10]

INTRODUCTION TO QSAR/QSPR

The qualitative relationships that exist between the properties of a compound and the molecular structure of that compound have been documented since the 1800s. In 1868, Crum Brown and Fraser first stated that it "is obvious that there must exist a relation between the chemical constitution and the physiological action of a substance."[11] Crum Brown and Fraser showed that with the use of a chemical addition reaction, various toxic or dangerous substances could be made more benign.[11] For example, iodides and sulfates of

John Krew is a recent graduate of Revere High School and will be majoring in Chemical and Biomolecular Engineering at Johns Hopkins University starting in Fall 2019. He worked as Dr. Visco's research assistant during the summer of 2018.

Donald P. Visco, Jr. is a Professor of Chemical Engineering at The University of Akron. His first faculty position started in 1999 and most recently served as Dean of the College of Engineering. He has also held leadership positions within the Chemical Engineering Division of ASEE and the Education Division of AIChE. His research interests focus on computer-aided moleculer design in a variety of areas within the pharmaceutical, chemical process, and construction material industries.

morphine and nicotine made through this approach are less toxic to their host than the original substance.[11]

While it is useful to have a *qualitative* understanding of structure-property relationships (e.g. the normal boiling point of n-alkanes increases as n increases), more applicable for a chemical engineer are *quantitative* structure-property relationships, or QSPRs. Cros, in 1863, helped provide the groundwork within the QSPR field by examining the relationship of the water solubility of primary alcohols, whose structure-property relationship was understood, to that of the toxicity of the primary alcohols.[12] Likewise, several years later, Richet in 1893, Meyer in 1899 and Overton in 1901 all independently determined relationships between the oil-water partition coefficient and various biological effects.[12] More applicable to chemical engineers and industrial chemists was the 1937 work of Hammett, who proposed a relationship between the reactivity of benzoic acid derivatives and the various substituents that could be added.[13] That work provided the basis for Hansch and Fujita who, using hydrophobic parameters and electronic constants from Hammett's equation, introduced the Hansch equation that predicts biological activity of molecules.[13] Additionally, Kruhlak showed the utility of QSPRs for screening the toxicity of pharmaceutical impurities and various other FDA-regulated products.[14] (Note that when a biological activity is featured as the property of interest, the models are more popularly known as Quantitative Structure-Activity Relationships, or QSARs). Kruhlak discussed the FDA's Computational Toxicology Consulting Service (CTCS) in which computational toxicology software utilizes a QSAR to screen pharmaceutical drugs.[14] There has even been interest in QSPRs for physico-chemical properties when it comes to the Globally Harmonized System for regulatory purposes. A 2016 paper by Fayet examines various existing QSPRs created to predict properties such as flash point.[15] Fayet also created a QSPR to predict impact sensitivity of nitroaliphatics to show proof of concept.[15]

Finally, as is well known by most chemical engineers, the book *The Properties of Gases and Liquids* contains approaches to estimate various properties of pure components and mixtures, often through the knowledge of the structures of the molecules, in order for chemical engineers to estimate equipment size when experimental data is unavailable.[16]

Group Contribution Approaches

One of the more popular approaches to develop QSPRs for use within the chemical process industry has been the group contribution approach. Group contribution methods use functional groups of molecules to predict physical and thermodynamic properties of pure components. The simplicity of group contribution methods and their ability to quickly estimate properties contribute to their popularity. Lydersen proposed one of the first group contribution methods using functional groups in 1955 to estimate the critical properties (critical temperature, critical pressure, and critical volume) of organic compounds.[17] One of the more well-known group contribution approaches within the chemical process industry is called UNIFAC (Universal Quasichemical Functional Group Activity Coefficients). Fredenslund and coworkers first published the UNIFAC model in 1975 for predicting equilibrium conditions, which itself was based in part on the Universal Quasichemical (UNIQUAC) Activity Coefficient model.[18] In 1984, Joback built on Lydersen's original work and expanded this analysis to additional functional groups and properties.[19] In particular, he proposed a group contribution method that not only improved correlative ability but explored many additional properties as well.[19] Later, Gani added to Joback's work by using second-order groups, which include additional structural information.[20]

Although valuable and popular, these group contribution methods have important drawbacks. One main argument against these approaches has been the quality of the QSPRs for properties that are important in processes involving separation, such as boiling point.[21] Another challenge is the need for prior functional group knowledge when choosing groups for the model.[22] Finally, the use of specific functional groups does not allow for a full range of atomic arrangements.[23]

The Signature Molecular Descriptor

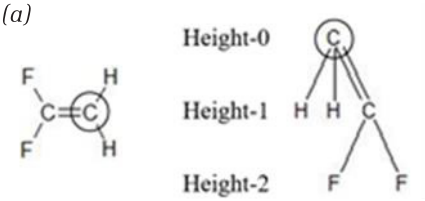
One useful and recent addition to the QSPR field is known as the Signature molecular descriptor. In 1994, Faulon first proposed Signature for use in structural elucidation studies.[24] Nearly a decade later, Faulon and co-workers further expanded on the use of Signature within QSPR studies and, soon after, as a descriptor within computer-aided molecular design approaches.[22, 25] In brief, the Signature molecular descriptor can be described as a tree of the connectivity within a molecule. Each atom (or root) of a molecule will have its own connectivity tree; thus, a compound of *n* atoms can be described by *n* atomic Signatures. The height of an atomic Signature indicates how far away from the root atom the tree will spread. An example of how Signature deconstructs a molecule can be seen in Figure 1. A height-0 atomic Signature is just the individual atom itself. A height-1 Signature includes the atom and its nearest neighboring atoms. A height-2 atomic Signature is the atom, its nearest and second-nearest neighbors (without backtracking).

Typically, but not always, height-1 has been chosen when using Signature, as it provides a good compromise between the generalization at height-0 (i.e. just the atoms and, accordingly, the molecular formula) and the specificity at height-2 (i.e. atoms, their nearest neighbors, and their second-nearest neighbors). Therefore, we choose height-1 in this work to describe any processes with Signature.

Signature in Computer-Aided Molecular Design

While QSPRs can be used for predicting properties, they can also be used in the solution of the inverse problem where structures are identified that are predicted to have a set of

(a)



(b)

Atomic Signatures

Height-0: C
Height-1: C(=CHH)
Height-2: C(=C(FF)HH)

(c)

Molecular Signatures

Height-0: 2C + 2F + 2H
Height-1: C(=CFF) + C(=CHH) + 2F(C) + 2H(C)
Height-2: C(=C(HH)FF) + C(=C(FF)HH)
2F(C(=CF)) + 2H(C(=CH))

Figure 1. *An example of the Signature of 1,1-Difluoroethene. (a) The spanning tree rooted at the circled carbon atom showing heights-0, 1 and 2. (b) The atomic Signatures for the circled carbon atom at heights-0, 1 and 2. Note that parentheses separate the atoms by distance from the root atom. (c) The molecular Signature for the molecule at heights-0, 1 and 2. Note the molecular Signature is the sum of the atomic Signatures of each atom at a given height.*

desired properties; this utilization of QSPRs is called inverse design or computer-aided molecular design (CAMD). The benefit of using QSPRs in an inverse design approach is that one can specify a range of desired properties and then find various compounds that would meet those properties. Gani and Brignole laid the groundwork for CAMD in 1983 when designing solvents for liquid-liquid extraction using their UNIFAC method.[26] Joback built off his original work to design refrigerants, polymers, solvents, and drugs with desired properties using his own Joback method, as well as other models, to predict properties.[27] While these CAMD approaches using group contribution methods are relatively simple, they sometimes restrict the search space and, thus, eliminate many innovative or novel designs.

More recently, the Signature molecular descriptor has been employed effectively in CAMD applications. Since 2003, Signature has been used in a wide variety of studies associated with molecular design. For example, Churchwell identified novel ICAM-1 inhibitors using CAMD whose results were later verified through experimentation.[10] Also, Weis used the inverse design process with Signature to find replacements for R-141b in polyurethane foam blowing agents with optimal properties.[7] Later, Weis utilized the GlaxoSmithKline solvent selection guide as the basis for identifying potential green solvents.[8] Chemmangattuvalappil also used Signature-based QSPRs for solvent selection.[23] Kayello's work in 2014 found novel surface tension reducing agents for use as chemical admixtures in concrete.[9] Signature has also been utilized for virtual high-throughput screening processes with successes in identifying compounds for pharmaceutical use.[28-31]

EXAMPLE

In order to demonstrate the utility of Signature in a CAMD framework, we propose a simple example. You are given a training set of 21 refrigerants spanning the class of hydrofluorocarbons, hydrofluoroolefins, and hydrofluoroethers and their experimental normal boiling points (see Table 1). The boiling points in this set span the range from 188.7 K to 337.0 K.[32] The goal of this CAMD problem is to identify different compounds outside the set of 21 that have predicted normal boiling points within the range of 320 – 330 K.

The first step in the process is to deconstruct the 2-D structures of the 21 compounds to their height-1 atomic Signatures; this resulted in 23 unique height-1 atomic Signatures (see Table 2).

The next step is to create a QSPR that can adequately correlate the experimental normal boiling point data. Here, for simplicity, we used a forward-stepping multiple linear regression approach (though other models can and have been used) [31] and arrived at a model with five height-1 atomic Signatures ($r^2 = 0.94$), as shown in Equation 1 (all units in K).

$$T_b = 39.34x_{23} + 8.765x_9 + 13.66x_{22} + 10.77x_{21} - 27.26x_{17} + 157.6 \quad (1)$$

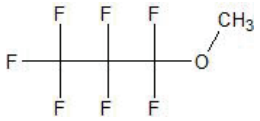
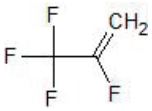
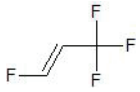
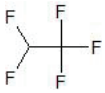
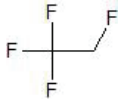
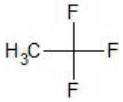
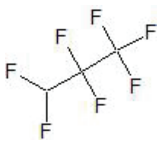
Now that the QSPR has been determined, we need to solve the CAMD problem utilizing all of the height-1 atomic Signatures available in Table 2. In order to do this, equations must be developed that describe how the various height-1 atomic Signatures can fit together through valence arguments. Constraint equations (in two forms: graphicality and consistency) are derived to ensure that each combination of atomic Signatures can generate a connected graph. The graphicality equation checks that the sum of the vertex degrees of a solution is even, as any solution that fails to satisfy this condition cannot produce a connected graph. There is a single graphicality equation for the entire set of the 23 unique atomic Signatures, obtained via Equation 2.

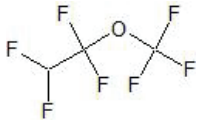
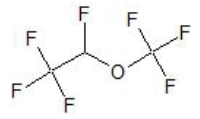
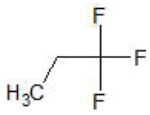
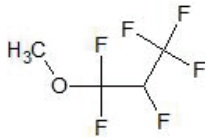
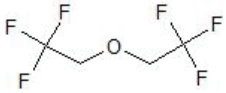
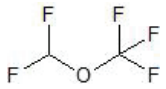
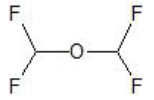
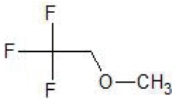
$$Modulus\left(\sum_{i=1}^{23} c_i x_i, 2\right) = 0 \quad (2)$$

Here, the values for c are equal to the vertex degree of the root of each atomic Signature, minus 2. For example, the atomic Signature x_{21} has its root (here, F) bonded to one atom and, thus, c_{21} = -1. Likewise, the atomic Signature x_6 has its root (here, C) bonded to four atoms and, thus, $c_6 = 2$. Application of Equation 2 to the atomic Signatures in Table 2 yields the following graphicality equation:

$$Modulus\left(x_1 + x_2 + x_3 + x_4 + x_5 + 2x_6 + 2x_7 + 2x_8 + 2x_9 + 2x_{10} + 2x_{11} + 2x_{12} + 2x_{13} + 2x_{14} + 2x_{15} + 2x_{16} + 2x_{17} + 2x_{18} + 2x_{19} + 2x_{20} - x_{21} - x_{22}, 2\right) = 0 \quad (3)$$

TABLE 1
Training Set of 21 Compounds

ASHRAE Number	Structure	T_{boil} (K)
HFE-7000		307.4
R-1132a		190.0
R-1141		201.0
R-1234yf		244.9
R-1234ze(E)		254.2
R-125		227.0
R-134a		246.6
R-143a		226.0
R-152a		248.4
R-161		236.0
R-227ca		257.2

ASHRAE Number	Structure	T_{boil} (K)
R-227ca2		261.0
R-227me		263.8
R-23		188.7
R-263		261.2
R-32		221.4
R-356mec		327.4
R-356mff2		337.0
R-E125		238.6
R-E134		277.8
R-E143a		304.8

The consistency equations guarantee that for each bond from a root atom A to a child atom B, there exists an atomic Signature in which root atom B is bound to a child atom A. The set of consistency equations contains one equation for each type of bonding that occurs in the training set. The coefficients within the consistency equations indicate the number of occurrences of that bond within a particular atomic Signature. More details about the consistency equations are provided elsewhere.[8] The five constraint equations provided in Equation 4 describe the specific bonding among the 23 atomic Signatures within the training set.

$$-2x_1 - x_2 - x_4 - 2x_6 - x_7 - 3x_9 - 2x_{10} - 2x_{11} - x_{12} - x_{13} - 3x_{16} - 3x_{17} - 2x_{18} - 2x_{19} + x_{21} = 0 \quad \text{C - F Bonding} \quad (4a)$$

$$-x_2 - 2x_3 - x_5 - x_7 - 2x_8 - x_{10} - 2x_{12} - x_{13} - 3x_{14} - 2x_{15} - x_{16} - 2x_{18} - x_{19} - 3x_{20} + x_{22} = 0 \quad \text{C - H Bonding} \quad (4b)$$

$$-x_{11} - x_{13} - x_{15} - x_{17} - x_{19} - x_{20} + x_{23} = 0 \quad \text{C - O Bonding} \quad (4c)$$

$$Modulus\left(x_4 + x_5 + 2x_6 + 2x_7 + 2x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15}, 2\right) = 0 \quad \text{C - C Bonding} \quad (4d)$$

$$Modulus\left(x_1 + x_2 + x_3 + x_4 + x_5, 2\right) = 0 \quad \text{C = C Bonding} \quad (4e)$$

The set of six equations (one from Equation 3 and five from Equation 4) are linear equations with integer coefficients and integer solutions. While there are known algorithms to solve such a system of Diophantine equations, we have utilized a brute-force method instead.[7] Since the system is underspecified (i.e. it has 23 unknowns and only 6 equations), we add additional constraints that require the number of occurrences of each atomic Signature in a solution to fall within the range of occurrences in the training set compounds (see Table 2). These constraints, when applied to this system, resulted in 1127 inverse solutions (or, in other words, height-1 molecular Signatures).

TABLE 2
Height-1 Atomic Signatures and Minimum/Maximum Number of Occurrences in Training Set

Variable	Height-1 Signature	Min/Max	Variable	Height-1 Signature	Min/Max
x_1	C(=CFF)	[0, 1]	x_{13}	C(CFHO)	[0, 1]
x_2	C(=CFH)	[0, 1]	x_{14}	C(CHHH)	[0, 1]
x_3	C(=CHH)	[0, 1]	x_{15}	C(CHHO)	[0, 2]
x_4	C(C=CF)	[0, 1]	x_{16}	C(FFFH)	[0, 1]
x_5	C(C=CH)	[0, 1]	x_{17}	C(FFFO)	[0, 1]
x_6	C(CCFF)	[0, 1]	x_{18}	C(FFHH)	[0, 1]
x_7	C(CCFH)	[0, 1]	x_{19}	C(FFHO)	[0, 2]
x_8	C(CCHH)	[0, 1]	x_{20}	C(HHHO)	[0, 1]
x_9	C(CFFF)	[0, 2]	x_{21}	F(C)	[1, 7]
x_{10}	C(CFFH)	[0, 1]	x_{22}	H(C)	[1, 5]
x_{11}	C(CFFO)	[0, 1]	x_{23}	O(CC)	[0, 1]
x_{12}	C(CFHH)	[0, 1]			

At this stage, we have our potential solutions but now must evaluate them for fitness as they relate to our desired normal boiling point constraints from the problem statement. Thus, we use the QSPR from Equation 1 to predict the normal boiling points of all 1127 solutions, reducing the solution set to 46 within the desired temperature range of 320 – 330 K (see Figure 2).

The next stage of the process converts the 46 height-1 molecular Signatures into 2-D structures. Each 2-D structure has a specific molecular Signature at a given height. However, depending on the height selected, each molecular Signature may have many

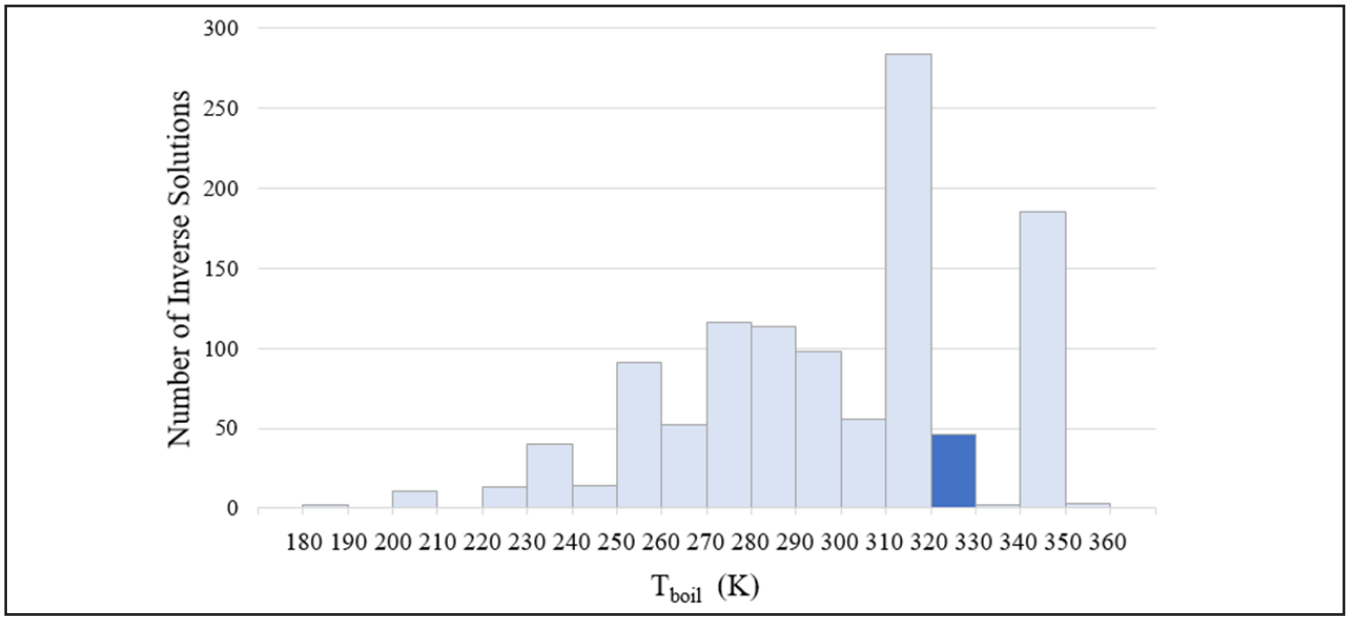


Figure 2. *Distribution of the Predicted T_{boil} for the 1127 inverse solutions; the shading indicates the 46 inverse solutions with predicted boiling points in the 320-330 K range.*

2-D structures. For example, a height-0 molecular Signature is just the molecular formula of a compound, which can result in an exponentially large number of structures depending on the molecular formula. As the height of the atomic Signature used increases, the number of structures associated with a particular molecular Signature also decreases. This degeneracy is all but eliminated by height-3.[25] Since the current work was performed at height-1, each molecular Signature will likely yield more than one structure. To generate structures, we use an available structure enumeration algorithm.[25] The algorithm begins with a graph of bare atoms, saturating all atoms with the same target atomic Signature simultaneously until the graph satisfies the desired molecular Signature. This process repeats multiple times per inverse solution, varying the order in which the atomic Signatures are saturated to enumerate all degenerate structures. In this case, we arrive at 70 structures from those 46 height-1 molecular Signatures. This set of 70 structures represents the pool of candidates for experimental verification to solve our computer-aided molecular design problem.

While we will not present all 70 structures in this demonstration example, we will highlight a few of the solutions. As one would expect, R-356mec (a training set compound with a normal boiling point of 327.4 K), was one of the 70 structures found. A more interesting candidate found by this CAMD approach was R-347pc-f, which was not in the original training set. For this compound, the predicted normal boiling point was 322.1 K, while the experimental boiling point was 329.8 K.[32] Thus, this CAMD approach with Signature identified a molecule that was within the required temperature range, but was not in the original training set.

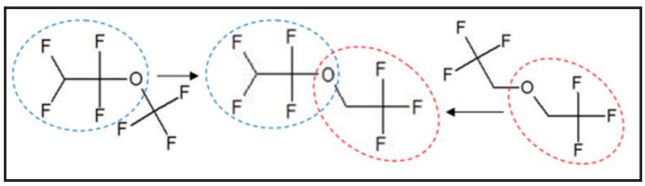


Figure 3. *R-347pc-f (center) was, in essence, formed by merging different parts of the training set compounds R-227ca2 (left) and R-356mff2 (right).*

An important aspect of CAMD with Signature is that it allows various structural motifs present in the training set to come together in beneficial ways to yield compounds with desired properties. For example, in the case of R-347pc-f, two features from two different compounds in the training set have been integrated via the algorithm to form a new compound, as described in Figure 3. Note that the approach correctly identified that combining certain features of the training set compounds R-227ca2 and R-356mff2 would produce a compound with a boiling point between those of its parent compounds (261.0 K and 337.0 K, respectively).

CONCLUSION

In this work, we demonstrated the use of a molecular descriptor called Signature in a simple computer-aided molecular design process. The approach identified 70 structures and demonstrated the discovery of compounds outside of the original training set. Additionally, we showed, in general, how various parts of other compounds within the training set come together to create these new compounds. It must be noted, however, that the example in this work using Signature was fixed within an initial set of compounds (21). The user could choose a dif

ferent dataset that includes more compounds and/or different properties. Of course, this focused approach using Signature requires deconstruction of the compounds within that dataset to identify atomic Signatures as well as the development of new QSPRs. Other, more established approaches (such as group contribution) have previously defined structural fragments as well as QSPRs regressed to a variety of properties.

Chemical product design is slowly becoming more integrated into the chemical engineering curriculum. As computer-aided molecular design approaches become more popular and accessible, especially within the curriculum, this will yield a pathway for chemical engineers to design compounds to meet specific property needs. The use of the Signature molecular descriptor is one approach to accomplish such design. Readers are welcome to contact Dr. Visco (dviscoj@uakron.edu) for specific information should questions arise during implementation.

REFERENCES

1. Cussler EL and Moggridge GD (2011) *Chemical Product Design*. 2nd ed. Cambridge University Press, Cambridge, United Kingdom.
2. Wei J (2007) *Product Engineering: Molecular Structure and Properties*. Oxford University Press, New York, NY.
3. Seider WD, Lewin DR, Seader JD, Widago S, Gani R and Ng KM (2016) *Product and Process Design Principles: Synthesis, Analysis, and Evaluation*. 4th ed. John Wiley & Sons, Hoboken, NJ.
4. Farrell S, Newell JA and Savelski MJ (2002) Introducing ChE students to product design through the investigation of commercial beer. *Chem. Eng. Ed.* 36(2): 108–113.
5. Georgia Institute of Technology (2018) Course catalog. http://www.catalog.gatech.edu/courses-undergrad/chbe/ accessed Sept. 15, 2018.
6. Watters JC (2008) A course module on chemical product design. *AIChE Annual Meeting*.
7. Weis D, Faulon JL, LeBorne R and Visco DP Jr. (2005) The Signature molecular descriptor. 5. The design of hydrofluoroether foam blowing agents using inverse-QSAR. *Industrial & Engineering Chemistry Research* 44(23): 8883-8891. 10.1021/ie050330y
8. Weis D and Visco DP Jr. (2010) Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. *Computers & Chemical Engineering* 34(7): 1018-1029. 10.1016/j.compchemeng.2009.10.017
9. Kayello H, Tadisina N, Shlonimskaya N, Biernacki J and Visco DP Jr. (2013) An application of computer-aided molecular design (CAMD) using the signature molecular descriptor—Part 1. Identification of surface tension reducing agents and the search for shrinkage reducing admixtures. *Journal of the American Ceramic Society* 97(2): 365-377. 10.1111/jace.12453
10. Churchwell CJ, Rintoul MD, Martin S, Visco DP Jr., Kotu A, Larson RS, Sillerud LO, Brown DC and Faulon JL (2004) The Signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. *Journal of Molecular Graphics and Modelling* 22(4): 263–273. 10.1016/j.jmgm.2003.10.002
11. Brown AC and Fraser TR (1868) On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Transactions of the Royal Society of Edinburgh* 25(1): 151–203. 10.1017/s0080456800028155
12. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T and Prachayasittikul V (2009) A practical overview of quantitative structure-activity relationship. *EXCLI Journal* 8: 74–88
13. Kubinyi H (1993) *QSAR: Hansch Analysis and Related Approaches*. VCH Publishers, New York, NY.
14. Kruhlak N, Contrera J, Benz R and Matthews E (2007) Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Advanced Drug Delivery Reviews* 59(1): 43-55. 10.1016/j.addr.2006.10.008
15. Fayet G and Rotureau P (2016) QSPR model for regulatory purpose: From development to integration into the QSAR toolbox. *Chemical Engineering Transactions* 48: 79–84.
16. Poling BE, Prausnitz JM and O'Connell JP (2001) *The Properties of Gases and Liquids*. 5th ed. McGraw-Hill Education, New York, NY.
17. Lydersen AL (1955) *Estimation of Critical Properties of Organic Compounds by the Method of Group Contributions*. University of Wisconsin, Madison, WI.
18. Fredenslund A, Gmehling J, Michelsen M, Rasmussen P and Prausnitz J (1977) Computerized design of multicomponent distillation columns using the UNIFAC group contribution method for calculation of activity coefficients. *Industrial & Engineering Chemistry Process Design and Development* 16(4): 450-462. 10.1021/i260064a004
19. Joback KG (1984) Master's thesis: *A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques*. Massachusetts Institute of Technology, Department of Chemical Engineering, Cambridge, MA.
20. Constantinou L and Gani R (1994) New group contribution method for estimating properties of pure compounds. *AIChE Journal* 40(10): 1697-1710. 10.1002/aic.690401011
21. Horvath A (1993) Critical evaluation of normal boiling points: Principle and example. *Chemosphere* 26(9): 1579-1594. 10.1016/0045-6535(93)90104-d
22. Faulon JL, Visco DP Jr. and Pophale R (2003) The Signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences* 43(3): 707-720. 10.1002/chin.200333232
23. Chemmangattuvalappil N, Solvason C, Bommareddy S and Eden M (2010) Reverse problem formulation approach to molecular design using property operators based on Signature descriptors. *Computers & Chemical Engineering* 34(12): 2062-2071. 10.1016/j.compchemeng.2010.07.009
24. Faulon JL (1994) Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules. *Journal of Chemical Information and Computer Sciences* 34(5): 1204-1218. 10.1021/ci00021a031
25. Faulon JL, Churchwell C and Visco DP Jr. (2003) The Signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *Journal of Chemical Information and Computer Sciences* 43(3): 721-734. 10.1002/chin.200333233
26. Gani R and Brignole E (1983) Molecular design of solvents for liquid extraction based on UNIFAC. *Fluid Phase Equilibria* 13: 331-340. 10.1016/0378-3812(83)80104-6
27. Joback KG (1989) Ph.D. Dissertation: *Designing Molecules Possessing Desired Physical Property Values*. Massachusetts Institute of Technology, Department of Chemical Engineering, Cambridge, MA.
28. Weis D, Visco DP Jr. and Faulon JL (2008) Data mining PubChem using a support vector machine with the Signature molecular descriptor: classification of factor XIa inhibitors. *Journal of Molecular Graphics and Modelling* 27(4): 466-475. 10.1016/j.jmgm.2008.08.004
29. Li H, Visco DP Jr. and Leipzig N (2014) Confirmation of predicted activity for factor XIa inhibitors from a virtual screening approach. *AIChE Journal* 60(8): 2741-2746. 10.1002/aic.14508
30. Chen J and Visco DP Jr. (2017) Developing an in silico pipeline for faster drug candidate discovery: Virtual high throughput screening with the Signature molecular descriptor using support vector machine models. *Chemical Engineering Science* 159: 31-42. 10.1016/j.ces.2016.02.037
31. Chen J, Schmucker LN and Visco DP Jr. (2018) Identifying new clotting factor XIa inhibitors in virtual high-throughput screens using PCA-GA-SVM models and Signature. *Biotechnology Progress* 34(6): 1553-1565. 10.1002/btpr.2693
32. Brown RL and Stein SE (2018) Boiling point data. Linstrom PJ and Mallard WG (eds.), *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology, Gaithersburg, MD. 10.18434/T4D303. ❐