QSPR Correlation of Melting Point for Drug Compounds Based on Different Sources of Molecular Descriptors

Hassan Modarresi,[†] John C. Dearden,*[,‡] and Hamid Modarress[†]

Department of Chemical Engineering, Amirkabir University of Technology (Tehran Polytechnic), 424 Hafez Avenue, Tehran, Iran, and School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, U.K.

Received August 1, 2005

Five linear QSPR models for melting points (MP) of drug-like compounds are developed based on three different packages for molecular descriptor generation and a combined set of all descriptors. A data set of 323 gaseous, liquid, and solid compounds was used for this study. Two models from the combined set of descriptors based on stepwise regression and genetic algorithm (GA) descriptor selection methods have acceptable prediction abilities. The statistical results of these models are $r^2 = 0.673$ and root-mean-square error (RMSE) of 40.4 °C for stepwise regression-based quantitative structure−property relationships (QSPRs) and $r^2 = 0.660$ and RMSE of 41.1 °C for GA-based QSPRs. Interpretation of descriptors of all models showed a strong correlation of hydrogen bonding and molecular complexity with melting points of drug-like compounds.

1. INTRODUCTION

Melting occurs when the forces of thermal agitation overcome the interactions holding the solid crystal together; these interactions can include ionic, polar, dispersion, and hydrogen bonding as enthalpic forces and positional, expansion, rotational, and conformational flexibility effects as entropic forces. Crystal packing also plays a great part, with symmetrical molecules showing higher melting points.[1,2] Melting point can be used for rough and rapid determination of the purity of a substance and also for prediction of aqueous solubility[3−5] and liquid viscosity.[6]

Mainly, prediction methods for melting point can be categorized as property−property relationship (PPR),[7] group contribution,[8] and quantitative structure−property relationship (QSPR).[9−12] Comprehensive reviews of the subject[1,2,13] reveal that many studies involved hydrocarbons and homologous compounds. This is because of the difficulty of melting point prediction for complex organic compounds, since the numerous factors that control it are not easy to quantify. Also, melting point is dependent upon the arrangement of the molecules in the crystal lattice as well as the strength of the pairwise group interactions, which are difficult to deal with using simple molecular descriptors. For organic compounds, it is understood that the dominant factor affecting the melting point is intermolecular hydrogen bonding. Compounds with intramolecular hydrogen bonding normally exert less intermolecular attraction and, therefore, have a lower melting point than their intermolecularly hydrogen-bonded analogues.[14]

Despite the enormous number of available melting point data, few useful guidelines exist for understanding the relationship between the melting point of a compound and its chemical structure.[15] Techniques for the estimation of the melting point of organic compounds would significantly assist medicinal chemists in designing new drugs within a specified range of melting point and solubility, since melting point is a controlling factor in solubility. Adequate aqueous solubility is necessary for a compound to be transferred to the active site within an organism.[16] Melting point also affects the toxicity of a compound, through its solubility. If a compound is only poorly soluble, its concentration in the aqueous environment may be too low for it to exert a toxic effect.[14]

Available molecular descriptors from commercial packages do not satisfactorily describe the many-body crystal packing effects and intermolecular forces in condensed media,[17] and suggested models for complex organic compounds such as drug-like compounds have large errors.[11] Therefore, the aim of this study was the evaluation of the performances of different QSPR models of melting point for drug-like compounds from different commercial descriptor generators and from a combined set of descriptors from all of the packages. In addition, we have tried to develop guidelines for explaining the relation between melting points of drug-like compounds and their molecular structures.

2. MATERIALS AND METHODS

(a) Data set. The data set used in this study comprises melting points (°C) of a set of 323 drug-like organic compounds. The data, collected from different sources,[11,18] are mostly of compounds that are solid at room temperature but also include some liquids and gases. Figure 1 shows the histogram of the data set for melting point and molecular weight, which has a mean melting point of 146 °C. The melting point values are spread between −118 and 345 °C. The data set was divided into a training set of 278 compounds and a test set of 45 compounds for multilinear regression

* Corresponding author phone: +44-(0)151-231-2066; fax: +44-(0)151-231-2170; e-mail: j.c.dearden@ljmu.ac.uk.
[†] Amirkabir University of Technology (Tehran Polytechnic).
[‡] Liverpool John Moores University.

10.1021/ci050307n CCC: $33.50
Published on Web 01/27/2006

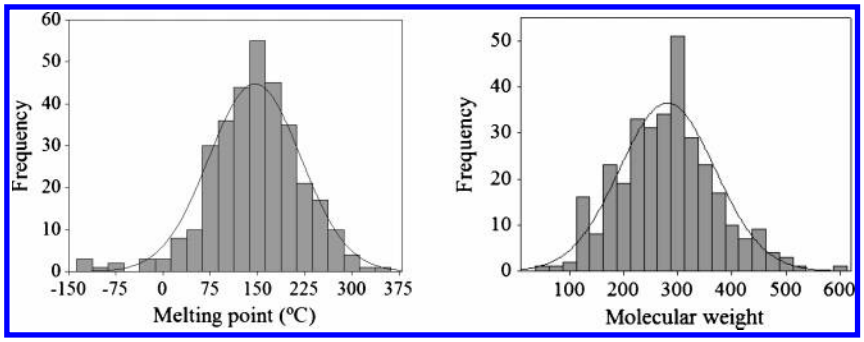

Figure 1. Histogram of melting point and molecular weight of compounds.

(MLR) analysis. The data-set division was carried out randomly with a check to ensure that the test data set covered all states of matter.

(b) Molecular Optimization and Descriptor Generation. The SMILES strings of all compounds were entered into the Tsar[19] software to generate the 3-D structures of molecules. Then the *mol* files of compounds were exported from Tsar to AMPAC software[20], where the molecular structures were optimized using the AM1 Hamiltonian in vacuo until the root-mean-square gradient was 0.1. All calculations were carried out at restricted Hartree−Fock level with no configuration interaction. The optimized geometries were transferred into CODESSA,[21] Dragon,[22] and Tsar packages to calculate molecular descriptors.

3. RESULTS

In the search for the best descriptor subset by stepwise regression from a large set of the descriptors, a major problem is connected with the mutual collinearity of descriptors, which leads to instability of the regression coefficients, overestimated standard errors, and a critical loss of predictive information. In addition, although inclusion of some descriptors with low correlation coefficient with melting points in a QSPR model may lead to better statistical results, these descriptors do not describe well the behavior of the property, which for external prediction may lead to large errors. Therefore, MATLAB codes were developed for excluding those descriptors that had <5% correlation with the melting point data and those descriptors that had >70% pairwise collinearity. Of the intercorrelated descriptors, that one was excluded that had the lower correlation with melting point. After the refining phase, the total number of descriptors was 244, 599, and 101 for CODESSA, Dragon, and Tsar, respectively, for each compound. A combined set of 727 descriptors from all descriptor sets was established for another QSPR model, using the above-mentioned refining criteria.

The Minitab[23] stepwise regression (forward−backward) and best-subset-selection method were jointly used for selecting the best MLR models. The Fisher *F*-criterion value of 4 was applied for entering and removing a descriptor from a QSPR model in the stepwise regression. Different linear models with different numbers of descriptors were developed based on three descriptor sets from Tsar, CODESSA, and Dragon packages and a combined set of all generated descriptors, which showed the superiority of the combined set as a descriptor source for a QSPR model. Figure 2 shows the plot of r^2 values and RMSE of training, test, and leave-one-out (LOO) cross-validation data versus the number of descriptors in the QSPR models from the combined set of descriptors. The data for this plot are given in Table S1. Figure 2 reveals that most of the increase in the r^2 value and of the decrease in RMSE were obtained in a model with two descriptors; the same phenomenon has been observed for other models. Therefore, these two descriptors play an important role in the melting process. There were no significant changes in model performance with more than eight descriptors, and also, a minimum RMSE of test data was achieved in a model with eight descriptors. Therefore, four QSPR models were developed with eight descriptors from four descriptor sets. The descriptors and their coefficients, *t*-values, and standard errors of coefficients (SE) in the QSPR models I, II, III, and IV based on Tsar, CODESSA, Dragon, and combined descriptor sets, respectively, are listed in Tables 1−4. In addition, the percentage of correlation values (CO%) of descriptors with melting point are given

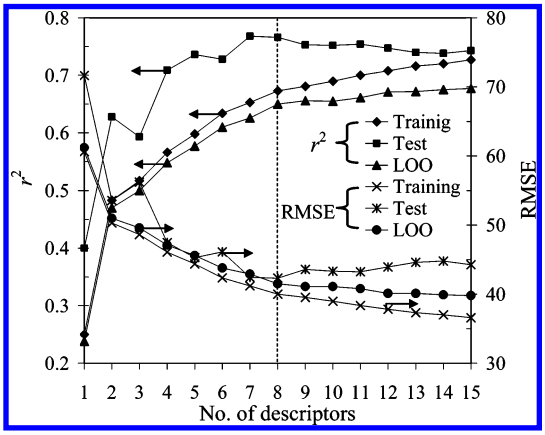

Figure 2. Effect of number of descriptors on the performances of QSPR models: dashed vertical line indicates the optimum descriptor numbers.

Table 1. Descriptors and Their Statistical Values in Model I

no.[a]	descriptor symbol	coefficient	*t*-value	SE	CO%
	intercept	243.85	4.99	48.91	
1	ΣE_{state}	3.05	7.72	0.40	42.0
2	n_{rb}	−10.49	−7.35	1.43	−12.2
3	n_{HD}	17.38	4.96	3.50	31.6
4	ϵ_{LUMO}	−31.13	−5.68	5.48	−40.2
5	n_F	−25.24	−3.23	7.81	14.0
6	n_{HA}	−9.37	−2.65	3.54	32.9
7	IP	−18.78	−3.86	4.86	16.6
8	n_{phenyl}	−19.94	−3.56	5.60	13.0

[a] 1, Sum of E-state indices; 2, number of rotatable bonds; 3, number of hydrogen-bond donors; 4, lowest unoccupied molecular orbital energy; 5, number of fluorine atoms; 6, number of hydrogen-bond acceptors; 7, ionization potential; 8, group count of phenyl.

Table 2. Descriptors and Their Statistical Values in Model II

no.[a]	descriptor symbol	coefficient	*t*-value	SE	CO%
	intercept	112.11	2.90	38.66	
1	$\epsilon_{LUMO-HOMO}$	−11.02	−4.00	2.76	−35.6
2	FHASA [QC−PC][b]	222.86	5.95	37.47	50.0
3	n_{rings}	20.09	7.95	2.53	29.7
4	HA_HDCA-1/TMSA [QC−PC]	3033.75	5.93	511.60	46.1
5	P^C_{max}	42.75	3.80	11.24	36.6
6	RNCG [Zefirov's PC][c]	−174.69	−4.91	35.57	−45.9
7	HA_HDCA-2 [QC−PC]	−22.06	−4.65	4.75	47.5
8	μ	5.08	3.32	1.53	35.3

[a] 1, HOMO−LUMO energy gap; 2, hydrogen-bonding acceptor ability of the molecule/total molecular surface area; 3, number of rings; 4, hydrogen-acceptor dependent hydrogen-bonding donor ability of the molecule/total molecular surface area; 5, maximum bond order of a carbon atom; 6, relative negative charge; 7, hydrogen-acceptor dependent area-weighted surface charge of hydrogen-bonding donor atoms based on quantum chemical partial charge; 8, dipole moment of the molecule. [b] On the basis of quantum chemical partial charges. [c] On the basis of Zefirov's partial charges.

Table 3. Descriptors and Their Statistical Values in Model III

no.[a]	descriptor symbol	coefficient	*t*-value	SE	CO%
	intercept	−369.28	−12.55	29.4	
1	IC_1	165.17	10.65	15.5	57.7
2	RT(m)	−6.95	−5.96	1.2	16.0
3	IC_2	−60.14	−4.81	12.5	49.1
4	VEA_1	58.87	5.39	10.9	40.4
5	JGI_1	256.41	5.25	48.9	19.0
6	R_{ww}	−2.15	−5.83	0.4	5.4
7	Mor19v	53.59	4.73	11.3	6.6
8	UI	22.30	4.58	4.9	34.9

[a] 1, Information content index (neighborhood symmetry of order 1); 2, R total index/weighted by atomic masses; 3, information content index (neighborhood symmetry of order 2); 4, eigenvector coefficient sum from adjacency matrix; 5, mean topological charge index of order 1; 6, reciprocal hyper-detour index; 7, 3D-MoRSE-signal 19/weighted by atomic van der Waals volumes; 8, unsaturation index (an empirical descriptor).

Table 4. Descriptors and Their Statistical Values in Model IV

no.[a]	descriptor symbol (source)	coefficient	*t*-value	SE	CO%
	intercept	23.86	0.78	30.4	
1	SPI (Dragon)	−0.000534	−3.99	0.00013	5.2
2	PW_5 (Dragon)	1101.83	8.07	136.5	41.0
3	IC_1 (Dragon)	50.04	5.73	8.7	57.7
4	GGI_1 (Dragon)	13.53	8.16	1.7	41.3
5	R_{1e} (Dragon)	−99.60	−6.74	14.8	28.4
6	h_f (CODESSA)	8.67	8.17	1.1	12.7
7	RPCS [QC−PC][b] (CODESSA)	−4.92	−4.03	1.2	−15.5
8	FHASA [QC−PC] (CODESSA)	312.05	9.37	33.3	50.0

[a] 1, Superpendentic index; 2, path/walk 5 − Randić shape index; 3, information content index (neighborhood symmetry of order 1); 4, topological charge index of order 1; 5, R autocorrelation of lag 1/weighted by atomic Sanderson electronegativities; 6, heat of formation per atom; 7, relative positive charged surface area; 8, hydrogen-bonding acceptor ability of the molecule/total molecular surface area. [b] On the basis of quantum chemical partial charges.

in these tables. The prediction results of all models for training and test data sets are provided in Table S2 as Supporting Information, and the related scatter diagrams against experimental melting points are represented in Figure 3. This figure shows that model I gives the highest scatter in both training and test data. Models II, III, and IV have

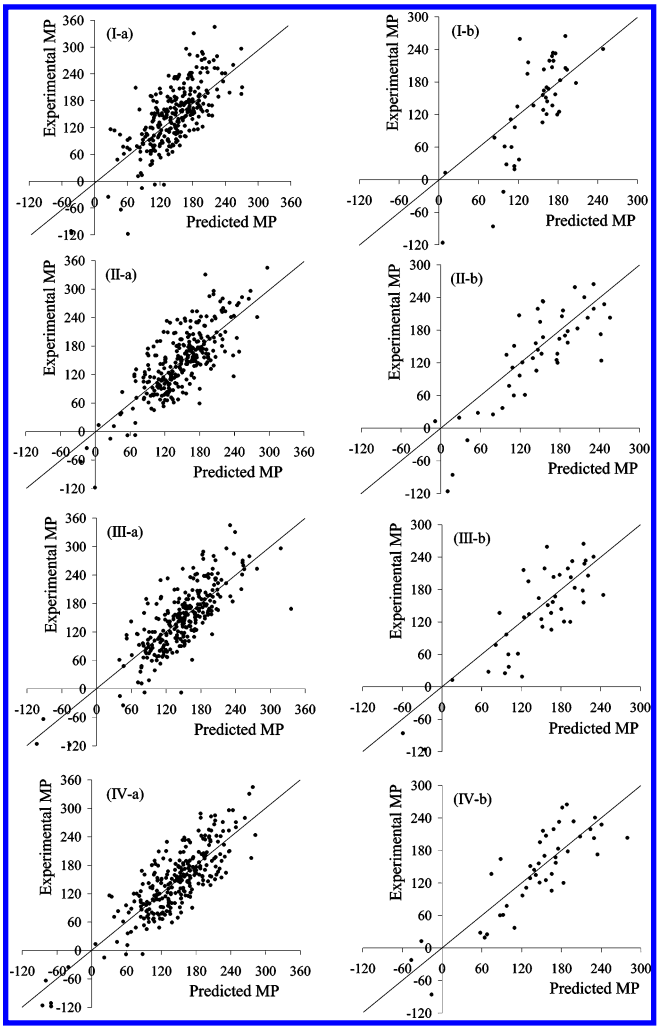

Figure 3. Scatter diagrams of the QSPR models (I, II, III, and IV) for (a) training and (b) test data.

comparable performances for training data, but the prediction results of model IV for test data are better than those of the others. Hence, the selection routine of genetic algorithm−partial least-squares (GA−PLS)[24−27] was applied on the combined descriptor set to select the eight best descriptors. The population, crossover probability, and mutation probability of GA were set to 30 chromosomes, 50%, and 1%, respectively. There was not much difference in results by increasing the number of chromosomes; however, it slowed the evolutionary process. The performance of PLS over a leave-group-out cross-validation procedure (with four elements out in each epoch of the cross-validation phase) was used as a fitness measure of GA, followed by the breeding (crossover and mutation) process. After 10 000 evolutions, the surviving frequencies of descriptors in all generations were used to select the eight best descriptors for the QSPR model. The selected descriptors for this model (model V) and their coefficients, *t*-values, and standard errors of coefficients (SE) are given in Table 5. Quantitative comparison of the performances of all models is presented in Table 6. Although model V has almost the same performance as model IV, some descriptors in model IV and also in models I, II, and III, but not in model V, are related to melting point in an incorrect way. Specifically, the signs of the *t*-value, coefficient, and correlation coefficient of some descriptors are not consistent. There are some descriptors in

Table 5. Descriptors and Their Statistical Values in Model V

no.[a]	descriptor symbol (source)	coefficient	*t*-value	SE	CO%
	intercept	1934.50	3.25	596.1	
1	IC_1 (Dragon)	44.56	4.57	9.7	57.7
2	FHASA [QC−PC][b] (CODESSA)	119.40	3.25	36.7	50.0
3	P^C_{ave} (CODESSA)	188.61	3.11	60.7	44.5
4	n_{rb} (Tsar)	−6.75	−5.93	1.1	−12.2
5	BEH_{e3} (Dragon)	85.87	6.99	12.3	37.2
6	V^H_{max} (CODESSA)	−2454.00	−4.19	585.9	−20.2
7	ATS_{3v} (Dragon)	0.93	4.75	0.2	24.7
8	HA_HDSA-1/TMSA [QC−PC] (CODESSA)	140.95	2.69	52.3	45.0

[a] 1, information content index (neighborhood symmetry of order 1); 2, hydrogen-bonding acceptor ability of the molecule/total molecular surface area; 3, average bond order of a carbon atom; 4, number of rotatable bonds; 5, highest eigenvalue no. 3 of Burden matrix/weighted by atomic Sanderson electronegativities; 6, maximum valency of a hydrogen atom; 7, Broto-Moreau autocorrelation of topological structure−lag 3/weighted by atomic van der Waals volumes; 8, hydrogen-acceptor dependent hydrogen bonding donor ability of the molecule/total molecular surface area. [b] On the basis of quantum chemical partial charges.

Table 6. Statistical Results of Six QSPR Models

	r^2 (%)			RMSE (°C)				Fisher
model	training	LOO[a]	test[b]	training	LOO	test	total[c]	statistic
I	0.481	0.426	0.458	50.5	53.1	65.8	52.9	31.2
II	0.618	0.589	0.681	43.3	44.9	47.9	44.0	54.3
III	0.626	0.595	0.624	42.9	44.6	55.9	44.9	56.2
IV	0.673	0.650	0.766	40.1	41.5	42.3	40.4	66.3
V	0.660	0.631	0.789	40.9	42.6	42.0	41.1	65.2
VI				38.1		42.0	38.6	

[a] Leave-one-out cross-validation over 278 data points. [b] External test over 45 data points. [c] Total RMSE of 323 training and test data points.

models I−IV that show contrary behavior against melting point from the real behavior, which is deduced from the sign of the correlation coefficients. For example, the coefficient and *t*-value of n_{HA} in model I have negative signs, which suggests a decrease in melting point with an increase in n_{HA}, while the correlation of this descriptor with melting point is +32.9%, i.e., an increase in melting point with an increase in n_{HA} (Table 1). In contrast, all signs of descriptor coefficients in model V are consistent with their real effect on melting point (Table 5).

To try to increase the prediction accuracy of the melting point of drug-like compounds, a consensus model (model VI) was developed by averaging the prediction results of models II, III, and IV. The r^2 value and RMSE of this model for all 323 data points are 0.724 and 38.6 °C, respectively (Table 6). The prediction results of models V and VI are given in Table S2. In addition, the collinearity tables of descriptors for each model are shown in Table S3.

4. DISCUSSION

In the melting process, a hypothetical partial melt as a transition state between solid and liquid is postulated in the theory of melting, which can be divided into four independent sub-processes:[28] (1) *expansional*, a sharp increase in the average distance between molecules that usually occurs on melting and is evidenced by an increase in volume; (2) *positional*, the change from the ordered arrangement of molecular centers of gravity in the crystal to the randomized arrangement; (3) *rotational*, the change from the ordered arrangement of major axes of crystalline molecules to a randomly oriented arrangement (for nonspherical molecules); and (4) *internal*, the change from the uniform conformation of flexible molecules of the crystal to random conformations of such molecules (for nonrigid molecules). From the above, it would be expected that the melting process is controlled by structural features and behavior of molecules in the crystal lattice and by intermolecular interactions during the melting process rather than simply by individual features of molecules. Most probably this is the reason a comprehensive model for melting points of a diverse range of organic chemicals has not yet been achieved, especially for compounds with complex structures.

However, an extended set of molecular descriptors which covers the diverse features of the molecules, and also a sophisticated searching method for the best descriptors from this space, can lead to a satisfactory QSPR model of melting point (Table 6).

(a) Model I. Although the Tsar package is a good source of descriptors for quantitative structure−activity relationship (QSAR) modeling of many biological activities and molecular properties, it failed in the case of melting points of drug-like chemicals. Model I has a low training r^2 of 0.481 and a high total RMSE of 52.9 °C. Four out of eight descriptors have sign inconsistency of the correlation coefficient with the descriptor coefficient and *t*-value. However, it reveals some effective molecular features relating to melting point, which have been reported by previous workers.[11,12] The most important descriptors of model I are ΣE_{state} and n_{rb}. The E_{state} index gives information related to the electronic and topological state of an atom in the molecule. In other words, it is a measure of the electronic accessibility of the atom and can be interpreted as a probability of interaction with other molecules. However, the index cannot be considered a pure electronic descriptor; in fact, ΣE_{state} is a descriptor of atom polarity and steric accessibility.

A complex molecule with rotatable bonds is freer to change its conformational structure in the semimelted phase; hence, it causes an increase in internal entropy, ΔS_{int} (fourth subprocess in hypothetical partial melt) and, consequently, a decrease in melting point. The sign of n_{rb} (number of rotatable bonds) in the model I is consistent with this assumption and can be approved with the following rigorous thermodynamic relationship:[2,28]

$$MP = \frac{\Delta H_m}{\Delta S_m} = \frac{\Delta H_m}{\Delta S_{exp} + \Delta S_{pos} + \Delta S_{rot}\Delta S_{int}} \qquad (1)$$

where ΔS_{exp}, ΔS_{pos}, ΔS_{rot}, and ΔS_m are the entropies of

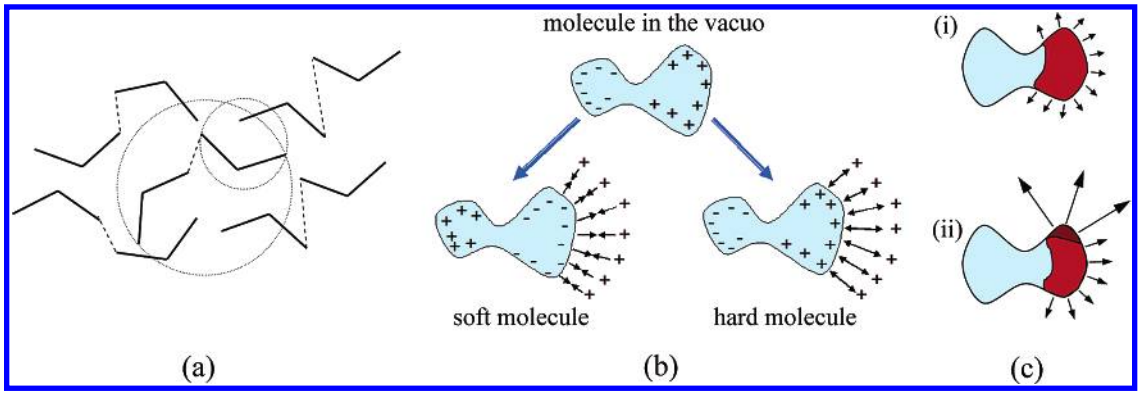

Figure 4. Molecular descriptor effects on melting point in semimelted phase: (a) hydrogen-bonding effect; (b) hardness and softness effect; and (c) relative negative (or positive) charge, (i) uniform charge distribution; (ii) nonuniform charge distribution.

expansion, position, rotation, and melting, respectively, and ΔH_m is the enthalpy of melting.

Hydrogen-bonding ability influences the melting point of the organic compounds. Figure 4a schematically illustrates the semimelted phase of a compound. Solid lines and dashed lines represent the structure of molecules and intermolecular hydrogen bonds, respectively. Obviously, molecules without hydrogen bonding need a smaller space (small dotted-line circle) for rotation, while hydrogen-bonded molecules require a larger space (large dotted-line circle). Therefore, hydrogen bonds lower the entropy of rotation and increase the melting point (eq 1). The number of hydrogen bond acceptors (n_{HA}) and donors (n_{HD}) are included in model I, but the former has an incorrect sign.

The energy of the lowest unoccupied molecular orbital (ϵ_{LUMO}) plays a major role in determining electronic band gaps in solids; it is directly related to the electron affinity and characterizes the susceptibility of the molecule toward attack by nucleophiles.[29] According to Koopman's theorem, the frontier orbital energy of LUMO is directly related to electron affinity ($\epsilon_{LUMO} = -A$). Thus, hard electrophiles have a high-energy LUMO (low electron affinity) and soft electrophiles have a low-energy LUMO (high electron affinity). Hence, an increase in ϵ_{LUMO} decreases the strength of intermolecular covalent bonds in the crystal lattice and, consequently, lowers melting point, as represented in model I with the negative coefficient and *t*-value.

(b) Model II. This model is based on CODESSA descriptors and has an acceptable performance (Table 6). Just one of eight descriptors has inconsistency in its signs of coefficient, *t*-value, and correlation coefficient. The number of rings (n_{rings}) is the most important descriptor in this model. The effect of rings is mentioned by Dearden[2] as being related to lattice packing of a solid compound, which influences the strength of intermolecular interactions (for example, compare the melting points of benzene (5.5 °C), naphthalene (80.2 °C), and anthracene (215 °C)). FHASA, HA_HDCA-1/TMSA, and HA_HDCA-2, as electrostatic descriptors, are related directly to hydrogen-bonding ability and show the importance of hydrogen bonding on melting point. HOMO (highest occupied molecular orbital)−LUMO gap ($\epsilon_{LUMO-HOMO}$) is related to the hardness of a molecule, with hard molecules having a large HOMO−LUMO gap and soft molecules having a small HOMO−LUMO gap.[30] Soft molecules can easily change both their number of electrons and the distribution of charge within the molecule by the introduction of an external electrostatic field (Figure 4b). Here, this field is introduced by other molecules, so that two such molecules can attract each other in several ways. Conversely, hard molecules cannot change their charge distribution easily (Figure 4b), and in the hypothetical partial melt this causes an increase in molecular disorder, entropy, and lattice instability. These conditions lower the melting point. The sign of $\epsilon_{LUMO-HOMO}$ in model II is consistent with this theory, as soft molecules (with low $\epsilon_{LUMO-HOMO}$) are more diverse in attraction sites and should have high melting points.

Relative negative charge (RNCG) and its sign indicate that, although the charged areas of molecules give rise to stronger attraction in a lattice, the sharp spots or undistributed negative (or positive) charges on the molecules cause intensive repulsion forces on other like charges, as well as intensive attraction forces on the opposite charges, and therefore an instability in the hypothetical partial melt phase, leading to instability of semimelted lattice nodes and low melting point (Figure 4c). The dipole moment (μ) is a vector quantity and encodes displacement with respect to the center of gravity of positive and negative charges in a molecule and is a rough criterion of electrostatic bond strength. Molecules with high dipole moments have strong intermolecular attractions and, hence, high melting points. The signs of coefficient, *t*-value, and correlation coefficient of this descriptor (Table 2) are in agreement with the above. Finally, the maximum bond order of a carbon atom (P^C_{max}) is a quantum-chemical descriptor which is related to electronic density of a molecule and is a measure of the extent of electron sharing between atoms.

(c) Model III. After the refining process on the descriptor pools, the biggest descriptor set was obtained from the Dragon software with 599 descriptors. Three of eight descriptors have inconsistent signs of coefficient, *t*-value, and correlation coefficient (Table 3). The most important descriptor of model III is first-order neighborhood information content (IC_1). It is calculated based on a hydrogen-depleted molecular graph and represents a measure of structural complexity per vertex. The positive sign of the descriptor coefficient in the model confirms that complex molecular structures with a diverse set of atoms in addition to carbon, such as nitrogen, oxygen, and halogens which can establish covalent bonds in the lattice, have a high melting point.

R total index (RT(m)) is derived from R-GETAWAY (R matrix-geometry, topology and atom-weights assembly)[31,32] descriptors. These descriptors combine the information from the molecular influence matrix (MIM), which is based on

molecular M-matrix (Cartesian spatial coordinates of atoms of a molecule in a specific conformation), with geometric interatomic distances in the molecule. The RT(m) is a measure of molecular complexity.[31] The eigenvector coefficient sum from adjacency matrix (VEA_1)[33] is a topological index based on adjacency matrix (an $n \times n$ matrix for a 2D-structural graph with n vertices) which encodes some information about molecular complexity and branching. Mean topological charge index of order 1 (JGI_1)[34] is a measure of charge distribution within a molecule. Reciprocal hyperdetour index (R_{ww})[35] is a complex topological descriptor based on reciprocal distance matrix. MoRSE (molecular representation of structures based on electron diffraction)-signal 19/weighted by atomic van der Waals volumes (Mor19v)[36] encodes structural features such as mass and amount of branching as evidenced by an investigation of monosubstituted benzene derivatives. Unsaturation index (UI)[37] gives information about bonds and atoms of a molecule. Clearly, most of the descriptors in model III are topological and geometrical descriptors; only one descriptor (JGI_1) encodes electrostatic information. All descriptors are positively correlated with melting point. The above indicates that complexity and branching of the molecules of complex drug compounds are much more important than the symmetry effect, which strongly influences the melting points of simple hydrocarbon compounds.[2]

(d) Models IV and V. The models from combined sets of descriptors have the best melting point prediction ability (Table 6). However, model IV, which is developed by the stepwise regression method, has two descriptors with inconsistent signs of coefficient, *t*-value, and correlation coefficient (Table 4). Therefore, the behavioral interpretation of all descriptors will lead to a misunderstanding of the melting process. Also, model IV may poorly predict the melting point of very complex structures that were not used in developing the model, as some descriptors are not correlated properly with melting point in the model. On the other hand, model V, which was developed by a sophisticated method of descriptor selection (genetic algorithm), has no incorrect correlation sign in the model and is, therefore, more reliable. SPI, PW_5, IC_1, and R_{1e} in model IV are topological indices from Dragon software that encode molecular complexity and branching. GGI_1 is a topological charge index similar to JGI_1, which encodes the charge distribution in a molecule. Two descriptors from the CODESSA software, RPCS and FHASA, encode the molecular electrostatic and hydrogen-bonding ability features, respectively. Also, heat of formation (per atom), h_f, is a measure of molecular stability, which is weakly correlated with melting point (12.7%).

Model V involves three topological descriptors from Dragon software, two hydrogen-bonding, one quantum-chemical, and one topological descriptor from CODESSA and a topological descriptor from Tsar. It thus appears that topological descriptors are more important than other descriptors in governing the melting point of drugs, as they have a strong contribution in all QSPR models. However, some hydrogen-bonding and electrostatic descriptors appeared in the models, which indicates that the melting point of drugs is controlled by molecular complexity, branching, hydrogen bonding, and electrostatic intermolecular interactions.

5. CONCLUSIONS

Eight-descriptor QSPR models with reasonably good prediction abilities have been developed for the prediction of melting points of drug-like compounds. The descriptor set of the Tsar package has no adequate descriptors to describe precisely the melting point features of drug-like compounds. CODESSA and Dragon packages have good electrostatic, quantum-chemical, geometrical, and topological descriptors that can model the melting point more accurately (models II and III). These packages may also model other properties of chemicals better than Tsar software, since Tsar lacks a diverse set of electrostatic and quantum chemical descriptors. The best models for the melting point of drug-like compounds were achieved with the combined set of descriptors (combination of Tsar, CODESSA, and Dragon descriptors). At first sight, it appears that genetic algorithm descriptor selection in this study has no benefit over stepwise regression. Comparison of the performances of models IV and V, obtained from the same descriptor source, shows almost the same prediction ability and errors for both models. However, model IV includes two descriptors with incorrect signs of coefficient, *t*-value, and correlation coefficient. This situation may lead to large prediction errors for very complex drug-like compounds that are not used in the development of the model. On the other hand, model V, based on the genetic algorithm selection method, has no such incorrect descriptors. Thus, model V can be considered to be more reliable for predicting the melting points of the external data, since there are no incorrectly correlated descriptors with melting point.

Model V, with eight descriptors and based on 323 gaseous, liquid, and solid drug-like compounds, has reasonable statistical results in comparison with a previous similar model[11] which has nine descriptors and is based on 277 solid drug-like compounds. Model V has an r^2 value and training and test RMSEs of 0.66, 40.9 °C, and 42.0 °C (Table 6), respectively, while the previously published model[11] has the r^2 value and training and test RMSEs of 0.57, 36.6 °C, and 49.8 °C, respectively. The small difference between the values of the training and test RMSEs in model V confirms that this model is better able to predict the melting points of additional drug-like compounds.

The molecular descriptors in the QSPR models cover the features of solid and semimelted phases that control the melting phenomenon. Hydrogen-bonding descriptors explain the strength of intermolecular bonds in the crystal lattice of a compound. Also, hydrogen bonds cause the melting points of drugs to increase by constraining the rotational entropy of molecules in the semimelted phase. Most of the topological descriptors, such as IC_1, are implicitly related to molecular bonding strength in the crystal lattice, since atomic diversity is included in the calculation of these descriptors. Also, they explicitly relate to molecular branching and complexity, which control the entropy of the system in the semimelted phase.

This study shows how difficult is the modeling of the melting point of drug-like compounds because of their complexity. The prediction results for the data of this study from a well-known commercial package of melting point prediction confirms this difficulty. The implemented group contribution method of melting point in the EPISuite

(MPBPWIN v1.41)[38] for the 323 compounds of our data set resulted in an RMSE of 67.9 °C and an r^2 value of 0.231. The predicted values are listed in Table S2. These poor results reveal the impact of entropic effects of complex molecular structures, although simple hydrocarbon data sets (e.g., normal paraffins) can be modeled accurately with the group contribution method or with just a few simple and well-known molecular descriptors in a QSPR model.[2] In contrast to the simple case of hydrocarbons, interpretation of effective molecular features of complex compounds on melting point is a difficult task, because of the various entropic parameters involved in the melting process. Analysis of the coefficients, *t*-values, and correlation coefficients of model V, as an acceptable model of this study, illustrates the pronounced effect of molecular complexity and hydrogen bonding on the melting point. Furthermore, the signs of the descriptor coefficients of the model can be utilized for a qualitative explanation of the relation between melting point and the structural features of compounds.

ACKNOWLEDGMENT

The authors thank the Iranian Ministry of Science, Research, and Technology and the British Council for supporting this work under a Partial Scholarship to H. Modarresi.

Supporting Information Available: Table S1 represents the performances of the QSPR model from the combined descriptor set according to the number of descriptors. Table S2 lists the compound names, experimental melting points, and predicted melting points from models I−VI and also the EPISuite group contribution model. In addition, the collinearity tables of descriptors for each model are shown in Table S3. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Dearden, J. C. Quantitative structure−property relationships for prediction of boiling point, vapor pressure, and melting point. *Environ. Toxicol. Chem.* **2003**, *22* (8), 1696−1709.
(2) Dearden, J. C. The prediction of melting point. In *Advances in Quantitative Structure Property Relationship*; Charton, I., Charton, M., Eds.; JAI Press Inc.: Stamford, CT, 1999; Vol. 2, pp 127−175.
(3) Yalkowsky, S. H.; Banerjee, S. C. *Aqueous Solubility: Methods of Estimation for Organic Compounds*; Marcel Dekker: New York, 1992.
(4) Meylan, W. H.; Howard, P. H.; Boethling, R. S. Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.
(5) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.
(6) Nimko, J.; Kukkonen, J.; Riikonen, K. A model for evaluating physicochemical substance properties required by consequence analysis models. *J. Hazard. Mater.* **2002**, *91*, 43−61.
(7) Mackay, D.; Shiu, W. T.; Bobra, A.; Billington, J.; Chan, E.; Yeun, A.; Ng, C.; Szeto, F. U.S. Environmental Agency Report PB 82-230939; U.S. Environmental Agency: Athens, GA, 1982.
(8) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. Boiling point and melting point prediction for aliphatic, non-hydrogen-bonding compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530−2535.
(9) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. Prediction of melting points for substituted benzenes: A QSPR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913−919.
(10) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71−74.
(11) Bergström, C. A. S.; Norinder, Ulf; Luthman, K.; Artursson, P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177−1185.
(12) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45*, 581−590.
(13) Tesconi, M.; Yalkowsky, S. H. Melting point. In *Handbook of Property Estimation Methods for Chemicals*; Boethling, R. S., Mackay, D., Eds.; Lewis: Boca Raton, FL, 2000; pp 3−27.
(14) Dearden, J. C. The QSAR prediction of melting point, a property of environmental relevance. *Sci. Total Environ.* **1991**, *109/110*, 59−68.
(15) Abramowitz, R.; Yalkowsky, S. H. Melting point, boiling point, and symmetry. *Pharm. Res.* **1990**, *7*, 942−947.
(16) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. Perspective on the relationship between melting points and chemical structure. *Cryst. Growth Des.* **2001**, *1* (4), 261−265.
(17) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24* (4), 279−287.
(18) *The Merck Index*, 12th ed.; Budavari, S., Ed.; Merck & Co, Inc: Whitehouse Station, NJ, 1996.
(19) *http://www.accelrys.com*.
(20) *AMPAC 8*; Semichem, Inc.: PO Box 1649, Shawnee, KS 66222, 1992−2004.
(21) *http://www.semichem.com*.
(22) *http://www.disat.unimib.it/chm/*.
(23) *www.Minitab.com*.
(24) *http://www.models.kvl.dk/source/GAPLS/index.asp*.
(25) Leardi, R.; Gonzalez, A. L. Generic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.
(26) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection *J. Chemom.* **1992**, *6*, 267−281.
(27) Leardi, R. Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. *J. Chemom.* **1994**, *8*, 65−79.
(28) Yalkowsky, S. H. Estimation of entropies of fusion of organic compounds. *Ind. Eng. Chem. Fundam.* **1979**, *18* (2), 108−111.
(29) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.
(30) Pearson, R. G. Absolute electronegativity and hardness: Applications to organic chemistry. *J. Org. Chem.* **1989**, *54*, 1423−1430.
(31) Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.
(32) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693−705.
(33) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517−523.
(34) Gálvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520−525.
(35) Diudea, M. V. Indices of reciprocal properties or Harary indices. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 292−299.
(36) Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure−spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.
(37) Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley−VCH Verlag GmbH: Weinheim, Germany, 2000; Vol. 11.
(38) *http://www.epa.gov/opptintr/exposure/docs/episuite.htm*.

CI050307N