Generation of a Novel Equation for Molecular Boiling Point Estimation Using a Neural Network

—Development of a Computer Aided Molecular Design System: MolWorks—

Sumie TAJIMA^a and Umpei NAGASHIMA^{b*}

^aBeyond Computing Co Ltd., 1-4-1 Takezono, Tsukuba, 305-0032, Japan

^bNational Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, 305-8562, Japan

*e-mail: u.nagashima@aist.go.jp

(Received: August 9, 2009; Accepted for publication: October 30, 2009; Advance publication: March 28, 2010)

We have developed a novel property estimation equation with a group contribution scheme for molecular properties (boiling points), in the standard condition using a three layers perceptron type neural network and are equipped MolWorks™ with it. 142 groups are newly defined as a set to reproduce the differences of isomers and to realize more accurate predictions than are available with usual methods. 765 data of molecular boiling points are selected for training of the neural network. 953 data were applied to evaluate the efficiency of the equation. The correlation of observed and predicted molecular boiling points by this work is better than the values obtained by Joback's equation. The equation is applicable to estimate a wide thermal range, including high and low temperature regions. Furthermore, the equation well reproduces the differences of boiling points for not only *ortho*-, *meta*-, and *para*- isomers but also for *cis*- and *trans*-isomers.

Keywords: Neural network, Boiling point, Property estimation

1 Introduction

Several types of information: chemical properties, physical, thermodynamic, electric and magnetic properties, optical spectra, etc., are very important for designing and developing new materials. To get such information, we have to use several software programs and/or databases. Concretely, molecular modeling programs, computational chemistry programs, network computing programs, chemical engineering programs, database programs, property estimation programs and multiple variable analysis programs have been used on many different types of computers, such as PC's, workstations, supercomputers and massively parallel computers.

The authors have developed a bench side computing environment on PC, MolWorks™ [1], for computer aided material design. Overview of MolWorks™ is depicted in Figure 1. MolWorks™ is a combined system of many different programs on small PCs and network environment. MolWorks™ is programmed with Java™ of Sun Microsystems, Inc., which does not depend on the user's machine types.

MolWorks™ can perform computer-aided material

design. For example, MolWorks™ has a semi-empirical calculation engine, which can support major molecular orbital calculation programs and can estimate several physical and chemical properties by using classical property estimation schemes (additivity schemes of group contribution), neural networks and genetic algorithms. MolWorks™ can thus treat the properties of not only molecules but also chemical substances. A free version was opened to the world from 1998 via the Web.

In this paper, we focus our discussion on a new property estimation function of MolWorks™. It is for Boiling Point, which is an important basic property to predict many thermal properties. The group contribution method is useful and is a major component in several property prediction methods, whereas there are also some different methods for boiling point estimation [2, 3]. Joback's equation [4] is a famous boiling point estimation method by group contribution method and is already equipped on the free version of MolWorks™.

A number of investigators have demonstrated that neural network systems can be an effective tool for estimating several properties [5]. The neural network approach is especially suitable for mapping complex nonlinear relationships that may exist between output (i.e., physicochemical properties) and input (i.e.,

Figure 1. Overview of program structure of MolWorks

molecular descriptors). We have attempted to construct a novel Boiling Point estimation equation using a three-layer perceptron neural network to construct more accurate prediction equation.

2 Method

2.1 Joback's equation

Joback's equation has 40 fragments consisting of carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine, chlorine, bromine and iodine. MolWorksTM can calculate boiling point, melting point, critical temperature, critical pressure, critical volume using Joback's equation by only building or one loading molecule and clicking one button. MolWorksTM automatically divides the molecule into fragments. Joback's 40 fragments can not express *ortho*-, *meta*-, and *para*- isomers and *cis*- and *trans*-isomers because no such descriptor is included in the fragments.

2.2 Perceptron Type Neural Network

For developing a novel neural network boiling point estimation scheme, we use the modified PSDD (Perceptron Simulator for Drug Design) [6] program. PSDD has a multi-layer perceptron neural network system and a back-propagation algorithm.

For generation of equations for molecular boiling

point estimation using a neural network, we have collected 1718 boiling point data and molecular geometry data from NIST [7] and ChemExper [8]. All 1718 molecules were consisted of fragments which are supported by Joback's method. Namely these molecules include only carbon, hydrogen, nitrogen, oxygen, sulfur and halogen atoms. The number of carbon atoms is less than 35 in the molecules.

The 1718 molecules include 624 chain molecules, 516 aromatic molecules and 578 ring molecules. 765 data of molecular boiling points including 251 chain molecules, 272 aromatic molecules and 242 ring molecules were selected for training of the neural network (training data). 953 data including 373 chain molecules, 244 aromatic molecules and 336 ring molecules were applied to evaluate the efficiency of the equation (test data). Deviations of training and test boiling point data along with molecular weights are shown in Figure 2.

We have decided 142 groups for estimating 1718 organic molecule boiling points as shown in Figure 3, where No.1 and No. 142 are molecular weights to grow in importance. In Figure 3, "Ar" and "Ri" mean aromatic ring and ring, respectively. We add some groups as independent groups (No. 111-139) to describe the *cis-trans* and the *ortho-meta-para* differences. Input data for the neural network are vectors consisting of 143 elements, where element is a number of descriptor in a molecule except for molecular weight (No.1 and 142). Twice counting of molecular weights: No.1 and 142 decreased error by almost 12%.

The neural network structure has three layers: one input layer with 143 input nodes (142 + 1 bias term), one hidden layer with 10 nodes, and one output layer with one output. The number of nodes in the hidden layer is optimized from 300 by using reconstruction learning method [6]. The reconstruction learning method has been developed as a model for solving over-learning problems.

3 Results and Discussion

The mean absolute error of training data set is 0.60%. The maximum relative error is 26.63%. The mean

Figure 2. Deviation of training and test boiling point data along molecular weights.

1	2	3	4	5	6	7	8	9	10	11	12
Molecular Weight	-OH(alcohol)	-OH(phenol)	-O-	Ar-O-	-CO-	Ar-CO-	-CHO	Ar-CH=O	-COOH	Ar-COOH	-O-CHO
13	14	15	16	17	18	19	20	21	22	23	24
-CO-O-	Ar-CO-O-	-NO ₂	Ar-NO ₂	Ri-CO-Ri	Ri-COO-Ri	Ri-O-Ri	-CF ₃	Ar-CF ₃	-CCl ₃	Ar-CCl ₃	-CBr ₃
25	26	27	28	29	30	31	32	33	34	35	36
-F	Ar-F	-Cl	Ar-Cl	-Br	Ar-Br	-I	Ar-I	-CF ₂ -	Ar-CF ₂ -	-CCl ₂ -	-CBr ₂ -
37	38	39	40	41	42	43	44	45	46	47	48
-CHCl-	-CHBr-	-CF<	-CCK	-CHF ₂	-CHCl ₂	-CHBr ₂	-CH ₃	Ar-CH ₃	-CH ₂ -CH ₃	Ar-CH ₂ -CH ₃	-CH(CH ₃) ₂
49	50	51	52	53	54	55	56	57	58	59	60
Ar-CH(CH ₃) ₂	-C(CH ₃) ₃	Ar-C(CH ₃) ₃	-CH ₂ -	Ar-CH ₂ -	-CH<	Ar-CH<	>C<	Ar>C<	≡CH	-C≡	Ar-C≡
61	62	63	64	65	66	67	68	69	70	71	72
>C=	Ar>C=	-CH=	Ar-CH=	=CH ₂	=C=	Ar=CH-Ar	Ar=C<Ar	Ar=C<Ar, Ar	-NH ₂	Ar-NH ₂	-NH-
73	74	75	76	77	78	79	80	81	82	83	84
Ar-NH-	-NK	Ar-NK	Ar-N=Ar	-CN	Ar-CN	-SH	Ar-SH	-S-	Ar-S-	Ri-CH ₂ -Ri	Ri-CH(Ri)
85	86	87	88	89	90	91	92	93	94	95	96
Ri-CH(Ri) with one electron acceptor	Ri>C<Ri	Ri>C<Ri with one electron acceptor	Ri>C<Ri with two electron acceptors	Ri,Ri>CH-Ri	Ri,Ri>C<Ri	Ri-(C=)-Ri	Ri,Ri>C<Ri, Ri	Ri-NH-Ri	Ri-NK(Ri)	Ri-NK(Ri), Ri	Ri-S-Ri
97	98	99	100	101	102	103	104	105	106	107	108
Ri>CH=	Ri>C=	Ri-N=	electron donor group	electron acceptor group	atomatic ring	3 ring	4 ring	5 ring	6 ring	7 ring
109	110	111	112	113	114	115	116	117	118	119	120
hetero atoms in ring	substitution group to ring	cis	trans
121	122	123	124	125	126	127	128	129	130	131	132

133	134	135	136	137	138	139	140	141	142
R>C=C<R	R>C=N-R	Molecular Weight		

Figure 3. Descriptors for boiling point estimation using neural network.No.120–141 are number of bonds shown by heavy line in a ring. ○ in No. 130–136 is a ligand. X in No. 137–139 is a hetero atom.

absolute error of the test data set is 2.56%. The maximum relative error is 15.61%. Correlation of observed and estimated values for the test data of the neural network is shown in Figure 4. That of Joback's method is shown in Figure 5 [9]. The correlation of observed and predicted molecular boiling points by this work is better than the values by Joback's equation. The quality of the

estimation was excellent because the correlations R^2 of our method and traditional Joback's method are 0.9808 and 0.8945, respectively. The equation generated by the neural network is efficiently applicable to estimate a wide thermal range, namely at high and low temperature region.

Figure 6 shows observed and estimated boiling points

Figure 4. Correlation of observed values and predicted values of molecular boiling point by neural network.

Figure 5. Correlation of observed values and predicted values of molecular boiling point by Joback's method [9].

Figure 6. Boiling points of alkane molecules (C_nH_{2n+2} , $n=1-35$). There are no observation data at $n=27, 31, 33$ and 34 .

Table 1. Examples of boiling point prediction for *cis-trans*, *ortho-meta-para*- isomers. “Training” is used for training. “Test” is used for evaluation of prediction quality.

	Observed	Joback	Relative error (%)	Neural network	Relative error (%)	training/test
321.0	324.2	1.0	324.8	1.2	training
333.0	324.2	2.6	330.5	0.8	training
418.8	415.5	0.8	412.1	1.6	test
424.4	415.5	2.1	419.1	1.2	test
424.6	415.5	2.1	425.2	0.1	test
518.2	519.6	0.3	524.5	1.2	training
550.0	519.6	5.5	551.9	0.3	training
558.2	519.6	6.9	555.7	0.4	training
348.6	371.1	6.5	348.7	0.0	test
362.0	371.1	2.5	353.3	2.4	test
368.0	371.1	0.9	366.5	0.4	test

of alkane molecules (C_nH_{2n+2} , $n=1-35$). Diamond, square and triangle indicate observed values, Joback prediction and neural network prediction, respectively. The values predicted by neural network show excellent agreement with the observed values.

The estimated boiling points of Joback's equation can not express the observed ones because of the linear equation. Differences between Joback's values and observed values become bigger in larger molecules.

Some examples of observed boiling point and estimated values are shown in Table 1. “Training” is used for training of neural network. “Test” is used for quality evaluation of prediction. Our equation generated by the neural network well reproduces the differences of boiling points not only for *ortho*-, *meta*-, and *para*-isomers but also *cis*- and *trans*-isomers.

The better results given by neural network was caused by larger number of descriptors and fitting coefficients in neural network. The numbers of descriptors for Joback's and neural network are 40 and 142, respectively. And those of fitting coefficients are 40 for Joback's and 1430 for neural network (1420 between input and hidden layers and 10 between hidden and output layers).

4 Conclusion

We have developed a novel boiling point estimation equation with the group contribution scheme using a three layer perceptron type neural network. 141 groups are newly defined as a set to reproduce the difference of isomers and to realize more accurate prediction than usual methods.

The correlation of observed and predicted molecular boiling points by this work is better than the values obtained by Joback's equation. This is because no descriptor to express *ortho*-, *meta*-, and *para*- isomers and *cis*- and *trans*-isomers is included in Joback's 40 fragments.

The equation is applicable to estimate a wide thermal range, namely both high and low temperature region. Furthermore, our equation well reproduces the differences of boiling points not only for *ortho*-, *meta*-, and *para*- isomers but also for *cis*- and *trans*-isomers.

This work was partially supported by Advanced Corporate Networking project (Business Starting of Novel Material and Drug Discovery System: 3-20-136) of the Organization for Small & Medium Enterprises and Regional Innovation, JAPAN (SMRJ).

References

- [1] MolWorks: <http://www.molworks.com>
- [2] There are another methods for BP estimation for example, Y. Nannoolal, J. Rarey, D. Ramjugernath, W. Cordes, *Fluid Phase Equilibria*, **226**, 45-63, 2004 [doi:10.1016/j.fluid.2004.09.001](https://doi.org/10.1016/j.fluid.2004.09.001), Y. Nannoolal, J. Rarey, J. Ramjugernath, *Fluid Phase Equilibria*, **252**, 1-27, 2007. [doi:10.1016/j.fluid.2006.11.014](https://doi.org/10.1016/j.fluid.2006.11.014)
- [3] There are another methods using a neural network for example, J. Taskinen, J. Yliruusi, *J. Advanced Drug Delivery Reviews*, **55**, 1163-1183, 2003. [doi:10.1016/S0169-409X\(03\)00117-0](https://doi.org/10.1016/S0169-409X(03)00117-0)
- [4] K. G. Joback, “*M.S. thesis in chemical engineering*”, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1984.
- [5] J. Gastiger and T. Engel ed., *Cheminformatics A Textbook*, Wiley-Vch Verlag GmbH & Co. KGaA, Weinheim, 2003.
- [6] H. Ichikawa, “*Perceptron type Neural network*” –an application for analysis of nonlinear phenomena-, Kyoritsu Publisher Co Ltd., 1993 (in Japanese).
- [7] NIST: <http://webbook.nist.gov/chemistry/>
- [8] ChemExper: <http://www.chemexper.com/>
- [9] S. Tajima, T. Yagi, T. Fukuda, U. Nagashima, *J. Comp. Chem., Jpn.*, **5**, 23-28 (2006). [doi:10.2477/jccj.5.23](https://doi.org/10.2477/jccj.5.23)