



Research Journal of Pharmaceutical, Biological and Chemical Sciences

Prediction of Refractive index of Organic Compounds using Structure-Property studies

Krishnaraj S¹ and Neelamegam P²

¹Department of Physics, PRIST University, Thanjavur-613403, Tamil Nadu, India,

²School of Electrical and Electronics Engineering, SASTRA University, Thirumalaisamudram, Thanjavur - 613401, Tamil Nadu, India, Tel: 9443077327.

ABSTRACT

This paper describes Quantitative Structure Property Relationship (QSPR) method for prediction of refractive index values of organic compounds based on molecular descriptors derived from molecular structures. A genetic algorithm based Unsupervised forward selection method is used to select the most statistically effective molecular descriptors computed with E-DRAGON software for the present study. Associative Neural Network (ASNN) and Polynomial Neural Network (PNN) methods are used to construct the Non-linear prediction models. The selected descriptors are used as input data for training and testing the Associative Neural Network and Polynomial Neural Network. Predicted results are in good agreement with the experimental refractive index of organic compounds. The results are cross-validated by Leave-one-out (LOO) cross-validation procedure. Comparison of ASNN and PNN reveals that ASNN predicts refractive indices with better accuracy with $R^2=0.962$. The results of this study indicate that it is possible to estimate the refractive index of organic compounds from their theoretically derived molecular descriptors.

Keywords: QSPR model, Refractive index, Descriptors, Associative Neural network (ASNN), Polynomial Neural Network (PNN).

**Corresponding author*



INTRODUCTION

The refractive index (n) plays a vital role in many branches of physics, biology and chemistry. The refractive index is a unit less parameter which is one of the most significant optical properties that is frequently employed to characterize organic compounds [1]. It is defined as the ratio of the speed of light in vacuum to the speed of light in the tested compound. It has been used as an indicator of the purity of organic compounds, but the relationship of refractive index to other optical, electrical, and magnetic properties has more significance. Since refractive index is a fundamental physical property of a substance, it is often used to identify a particular substance, or measure its concentration. Refractive index is used to measure solids, liquids, and gases. Most commonly, it is used to measure the concentration of a solute in an aqueous solution [2, 3].

Quantitative Structure Property Relationship (QSPR)

The chemical and physical properties of a compound are a function of its molecular structure. Structure-property relationships are developed by finding one or more molecular descriptors derived from structure that explains variations in the physical or chemical properties of a group of congeners/analogs. While some descriptors can be determined experimentally, deriving them from either the two-dimensional (2-D) or three-dimensional (3-D) molecular structure is generally more convenient and practical. A relationship, once established, can be used to estimate the properties of other molecules simply from their structures and without the need for experimental determination or synthesis. This has resulted in the development of quantitative structure property relationships (QSPRs) as an important tool in chemical, biological, and environmental research. When a structure-property relationship is found, it may also provide insight into which aspect of the molecular structure influences the property. Such insight can facilitate a systematic approach to the design of new molecules with more desirable properties. QSPR development requires three basic steps (i) a property data set, measured experimentally (ii) Molecular Descriptors, which are the Quantitative descriptions of structural properties, and (iii) Statistical method or Neural network approach to establish the relationship between molecular descriptors and properties [4]. One of the important problems in QSPR is the description of molecular structures using molecular descriptors, which can include structural information as much as possible. Theoretical descriptors such as constitutional descriptors and topological indices have found the major popularity in QSPR studies for several reasons such as

- a) Their calculation is simple and fast,
- b) They do not need information about three dimensional structure of molecules,
- c) They are exact number without uncertainty and
- d) They represent high correlation with many physico-chemical properties [5]

The most familiar standard approaches to QSPR are based on statistical methods such as Multiple Linear Regression, Cluster analysis, Principal component analysis and Partial least

square regression [6]. For the prediction of physical properties, high-quality models are obtained based on predictive equations using linear regression techniques, are used to correlate structure related descriptors with observed properties. The models developed on Multiple linear regression requiring a priori assumption of the linear form of the mathematical correlation model. Such models do not consider the non-linearity that may exist among the input descriptors and calculated property. The above deficiencies have been addressed in the past using Neural Networks. Currently, neural networks are used with encouraging success in development of various QSPR models. An artificial neural network (ANN) represents non-linear methods and well suited to describe structure-property models. Moreover, ANN is able to consider not only particular structure characteristics, but also interrelations and interdependencies between mutually influencing structural features. Therefore, they can be easily adapted for processing large data set formed by a set of descriptors [7-9].

In the literature, most of the papers are reported for prediction of refractive index using Multivariate regression. Ivanciuc et al. published a paper which employs Multiple linear regression as a tool for prediction of refractive index of 134 alkanes [10]. Their model involves Wiener polynomial descriptors yielded a correlation coefficient(R) of 0.98. Katrizky et al.(1998) have designed a CODESSA software based 5 descriptor model [11] for a data set of 125 organic compounds having a R^2 of 0.945. Xihua and Juguan published a paper which employs Multiple linear regression as a tool for prediction of refractive index of 64 alkanes [12]. Their model involves a new structure information autocorrelation index yielded a correlation coefficient(R) of 0.98. Only few works are available in literature for prediction of refractive index of organic compounds using neural network. Hence, this paper explains Associative neural network (ASNN) and Polynomial Neural Network (PNN) based prediction of refractive index of organic compounds based on 8 descriptors provided by E-DRAGON [13] having specific physical meaning corresponding to different molecular interactions.

MATERIALS AND METHODS

Data

The source of experimental refractive index values of 148 compounds at 20⁰C (wavelength 589nm) are collected from [11, 14].

Molecular descriptors

The chemical structures of the 148 organic compounds are drawn with MarvinSketch(Chemaxon) [15] and exported as SMILES notation. Next, organic compounds represented by SMILES format are used as input for the online software, E-Dragon (vcclab) which converts the molecules from SMILES notation into 3-dimensional structures, and then it calculates various types of descriptors.



Selection of Molecular descriptors

The challenge in developing the QSPR model is the selection of molecular descriptors from the pool of available descriptors that strongly correlate with desired physical property. The use of all available descriptors in the model development causes dimensionality problems. Further, the use of redundant or irrelevant descriptors decreases the capability of prediction the performance of a QSPR model, especially when non-linear algorithms are used in model development. The descriptor selection process involves the identification of the most relevant set of descriptors for model development and is the most important step in all QSPR development efforts. Several different methods are described in literature for descriptor selection. The most widely used techniques are Principal component analysis (PCA), Partial least squares (PLS) and Unsupervised Forward selection.

Unsupervised Forward Selection (UFS) is a data selection process by deleting redundant or irrelevant variables that selects from a data matrix a maximal linearly independent set of columns with a minimal amount of multiple correlations [16]. UFS was designed for use in the development of QSPR models, where the m by n data matrix contains the values of n variables (typically molecular properties) for m objects (typically compounds). QSPR data sets often contain redundancy (exact linear dependencies between subsets of the variables), and multicollinearity (high multiple correlations between subsets of the variables). Both of these features decrease the development of QSPR models with the ability to generalize successfully to new objects. Continuum regression, an algorithm encompassing ordinary least squares regression, regression on principal components, and partial least squares regression, are used to construct models from the selected variables. UFS produces a reduced data set that contains no redundancy and a minimal amount of multicollinearity. The variable selection routine is shown to produce simple, robust, and easily interpreted models for the chosen data sets. The freeware for employing Unsupervised Forward selection is available online at [17].

Neural network

A neural network is a efficient data modeling tool that is able to capture and represent complex non-linear input/output relationships. Neural network technology performs "intelligent" tasks similar to those performed by the human brain. A neural network may simply be viewed as a highly parallel computational device typically used when there are a large number of observations and when the problem is not understood well enough to write a procedural program or expert system. The ANN are trained to perform a particular function by adjusting the values of the connections, or weights, between elements until a particular input leads to a specific output. The ANN consists of three layers: input, hidden and output layers. These three layers are connected with each other. The input layer receives the input data from computer user and sends them to the hidden layer. The hidden layer contains interconnected neurons for the pattern recognition and the relevant information interpretation for adjusting the weights on the connections. Afterwards, the results from the hidden layer are sent to the output layer for the outputs. The neurons contain several functions and variables including

weights, non-linear transfer functions, methods to add up all inputs and bias values. The sum of all products of all the inputs multiplied with the weights and the bias values passes through a non-linear transfer function as the output of each neuron [18]. A neural network is thus a mathematical model to represent a non-linear hypersurface. The increasing interest and availability of neural network has prompted several groups to apply this technology in QSPR studies for prediction of physical and chemical properties effectively [19].

Associative neural network

The traditional artificial feed forward neural network is a memory-less approach. This means that after training is complete, all information about the input patterns is stored in the neural network weights and input data are no longer needed, i.e. there is no explicit storage of any presented example in the system. In Contrast, ASNN is a method with improved predictive abilities including combination of memory-based and memory-less method. It offers an elegant approach to incorporate “on the fly” the user’s data [20]. The ASNN is an extension of the committee of machines that goes beyond a simple/weighted average of different models. An ASNN is a combination of an ensemble of feed forward neural networks (memory-less) and the K – nearest neighbour technique (memory-based). It uses the correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbour techniques. An associative neural network has a memory that can agree with training set. If new data is available the network improves its predictive ability and gives a good approximation of unknown function without a need to retrain the neural network ensemble. This method dramatically enhances its predictive ability over traditional neural network and K–nearest neighbour techniques [21].

The ASNN models are selected based on selection processes that include the algorithm, the number of neurons and hidden layers, and the iterations and number of ensembles. The early stopping over ensemble (ESE) method was used for training the neural networks). In ESE, initial training sets were randomly constructed with equal size learning and validation sets for each neural network in the ensemble. Thus, each neural network had its own learning and validation sets. The learning set was used for adjusting neural network weights. The training is stopped when a minimum error for the validation set is calculated (‘early stopping point’). Following ensemble learning, a simple average of all networks is used for predicting the test patterns. The developed algorithm of ASNN is available online at the Vcclab website [22].

Polynomial Neural network

Polynomial Neural Network (PNN) represents a promising method for applications in QSPR and QSAR studies. It provides the model in parametric form as an equation that can be easily interpreted by the users. It is a iterational heuristic algorithm of Group method of Data Handling (GMDH). GMDH was originally proposed by Prof. A.G. Ivakhnenko in Late 1960s for identifying non-linear relation between input and output variables. PNN (also known as GMDH) represents a group of inductive algorithms for computer-assisted mathematical modeling of

multi-parametric datasets that features fully automatic structural and parametric optimization of models. Inductive algorithms provide the possibility to find automatically interrelations in data, to select optimal structure of model or network and to decrease the error of existing algorithms. PNN is applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition. PNN calculates analytical non-linear models between descriptors of organic compounds and the target physico-chemical property and provides a clear interpretation of the detected relations. The structure of PNN is similar to that of a feed forward neural network. A neuron of feed forward network is replaced by a neuron of PNN [23]. The developed software for PNN is available online at [24].

RESULTS AND DISCUSSION

E-DRAGON software is used to compute more than 1600 descriptors for each compound and all the descriptors are not relevant to the property (Refractive index) considered. Therefore, Unsupervised forward selection (UFS) method is employed for descriptor selection in the present study. The Descriptors selected for present study must not be highly correlated. Only those descriptors having intercorrelation co-efficient below 0.91 are considered for the present study. The selected descriptors involved in the present QSPR model are:

- | | | |
|--------|------|--|
| (i) | C001 | : Atom centered fragments/ CH ₃ R |
| (ii) | nX | :Number of Halogen atoms |
| (iii) | RBF | :Rotatable Bond Fraction |
| (iv) | AMW | :Average Molecular Weight |
| (v) | X2Av | :Average valence connectivity index chi-2 |
| (vi) | nNr | :Relative Number of Nitrogen atoms. |
| (vii) | Mp | :Mean Atomic Polarizability(Scaled on Carbon atom) |
| (viii) | Sp | :Sum of Atomic Polarizabilities(Scaled on Carbon atom) |

C-001 is a atom centered fragment descriptor defined by counting first neighbours of carbon atoms (CH₃R), where R is the presence of heteroatoms. nX is a constitutional descriptor simply computed by sum of counts of Halogen atoms. The RBF is a constitutional descriptor which represents fraction of rotatable bonds. The average molecular weight (AMW) is obtained by dividing molecular weight with total number of atoms. nNr is the relative number of Nitrogen atoms obtained by dividing the number of nitrogen atoms with Number of atoms. X2Av is a topological descriptor encodes presence of heteroatom, double and triple bonds calculated from hydrogen suppressed graph. Mp is a constitutional descriptor calculated by dividing sum of atomic polarizabilities by Number of atoms. Sp is also a constitutional descriptor calculated by sum of atomic polarizabilities.

The selected descriptors and the experimental refractive index values are listed in Table 1. The data set is randomly divided into two subsets one for training set and the other for testing. The data set in Table1 is used for training and it is used to build models using ASNN and PNN. During the Training process, the network involves eight neurons (eight descriptors) in the

input layer, seven neurons in the hidden layer and one neuron (Refractive index (n)) in the output layer. The network is trained using the LevenBerg Marquardt algorithm. The input and output data are normalized in the range of 0.1 to 0.9 and logistic activation function is used for all neurons. Number of hidden neuron is decided by training and predicting the 'training data' by varying the number of hidden neurons in the hidden layer. A sufficient training level is not reached with smaller number of neurons and overfitting exists with a larger number of neurons in the hidden layer. Out of the different configuration tested, a hidden layer with 7 hidden neurons give the optimum result for prediction of refractive index of organic compounds. The seed number is used in to start sequence of random numbers for neural network weights initialization and partition of initial training set data on Training /test sets. During training process the seed number is adjusted until best model is obtained. The architecture of the final model is shown in Table 2.

Table 1: Molecular descriptors, Experimental and Computed Refractive index values for Training set using ASNN and PNN.

S.No	Compound Name	C-001	nX	RBF	AMW	X2Av	nNr	Mp	Sp	Experimental	Predicted by ASNN	Residual	Predicted by PNN	Residual
1	Ethylbromide	1	1	0	13.62	1.389	0	0.7	5.64	1.42	1.42	0	1.42	0
2	Propane	2	0	0	4.01	0.707	0	0.55	6.05	1.29	1.3	-0.01	1.31	-0.02
3	1-Chloropropane	1	1	0	7.14	0.533	0	0.63	6.9	1.39	1.39	0	1.39	0
4	2-Methylbutane	3	0	0.063	4.25	0.451	0	0.56	9.57	1.36	1.36	0	1.36	0
5	2,2,3-Trimethylbutane	5	0	0.045	4.36	0.391	0	0.57	13.1	1.39	1.39	0	1.38	0.01
6	Cyclobutane	2	0	0.077	4.15	0.5	0	0.56	7.81	1.37	1.36	0.01	1.35	0.02
7	1-Chlorobutane	1	1	0.077	6.61	0.474	0	0.62	8.66	1.4	1.4	0	1.4	0
8	2-Chloro butane	2	1	0.077	6.61	0.484	0	0.62	8.66	1.4	1.39	0.01	1.38	0.02
9	1,3-Dibromobutane	1	2	0.077	15.42	0.723	0	0.75	10.5	1.51	1.5	0.01	1.51	0
10	n-Butylbromide	1	1	0.077	9.79	0.612	0	0.65	9.16	1.44	1.44	0	1.46	-0.02
11	Pentane	2	0	0.125	4.25	0.451	0	0.56	9.57	1.36	1.37	-0.01	1.36	0
12	3-MethylPentane	3	0	0.105	4.31	0.384	0	0.57	11.3	1.38	1.38	0	1.38	0
13	2,4 Dimethyl pentane	4	0	0.091	4.36	0.432	0	0.57	13.1	1.38	1.39	-0.01	1.38	0
14	2-Methyl-3 ethyl pentane	4	0	0.12	4.39	0.353	0	0.57	14.9	1.4	1.4	0	1.4	0
15	2,2,4 Trimethyl pentane	5	0	0.08	4.39	0.416	0	0.57	14.9	1.39	1.39	0	1.39	0
16	Propylcyclopentane	1	0	0.083	4.68	0.327	0	0.59	14.1	1.43	1.44	-0.01	1.45	-0.02
17	Butylcyclopentane	1	0	0.111	4.68	0.329	0	0.59	15.9	1.43	1.44	-0.01	1.46	-0.03
18	Hexylcyclopentane	1	0	0.152	4.68	0.333	0	0.59	19.4	1.44	1.45	-0.01	1.47	-0.03
19	Hexane	2	0	0.158	4.31	0.427	0	0.57	11.3	1.38	1.38	0	1.38	0
20	2,3 Dimethyl Hexane	4	0	0.12	4.39	0.376	0	0.57	14.9	1.41	1.4	0.01	1.4	0.01
21	2,4 Dimethyl hexane	4	0	0.12	4.39	0.393	0	0.57	14.9	1.4	1.4	0	1.4	0
22	2-Methyl 3 ethyl hexane	4	0	0.143	4.42	0.356	0	0.57	16.6	1.41	1.41	0	1.41	0
23	Cyclohexane	0	0	0	4.68	0.354	0	0.59	10.6	1.43	1.43	0	1.44	-0.01
24	4-Methylheptane	3	0	0.16	4.39	0.383	0	0.57	14.9	1.4	1.4	0	1.4	0



25	2,4-dimethyl heptane	4	0	0.143	4.42	0.391	0	0.57	16.6	1.4	1.41	-0.01	1.4	0
26	2,3-Dimethyl-5 Ethylheptane	5	0	0.147	4.47	0.345	0	0.58	20.1	1.42	1.42	0	1.43	-0.01
27	Cycloheptane	0	0	0	4.68	0.354	0	0.59	12.3	1.45	1.45	0	1.46	-0.01
28	2-Methyl octane	3	0	0.179	4.42	0.405	0	0.57	16.6	1.4	1.41	-0.01	1.41	-0.01
29	2,3 Dimethyloctane	4	0	0.161	4.45	0.372	0	0.57	18.4	1.41	1.41	0	1.41	0
30	2,6-Dimethyl octane	4	0	0.161	4.45	0.384	0	0.57	18.4	1.41	1.41	0	1.41	0
31	2,2,3-Trimethyl octane	5	0	0.147	4.47	0.366	0	0.58	20.1	1.42	1.42	0	1.42	0
32	2,3,5-Trimethyl octane	5	0	0.147	4.47	0.363	0	0.58	20.1	1.42	1.42	0	1.42	0
33	Nonane	2	0	0.214	4.42	0.395	0	0.57	16.6	1.41	1.41	0	1.41	0
34	2-Methyl Nonane	3	0	0.194	4.45	0.4	0	0.57	18.4	1.41	1.41	0	1.41	0
35	Decane	2	0	0.226	4.45	0.39	0	0.57	18.4	1.41	1.42	-0.01	1.41	0
36	4-Methyldecane	3	0	0.206	4.47	0.374	0	0.58	20.1	1.42	1.42	0	1.42	0
37	Undecane	2	0	0.235	4.47	0.386	0	0.58	20.1	1.44	1.43	0.01	1.43	0.01
38	n-Dodecane	2	0	0.243	4.48	0.383	0	0.58	21.9	1.42	1.43	-0.01	1.43	-0.01
39	n-Pentadecane	2	0	0.261	4.52	0.376	0	0.58	27.2	1.43	1.43	0	1.43	0
40	n-Heptadecane	2	0	0.269	4.54	0.373	0	0.58	30.7	1.44	1.44	0	1.42	0.02
41	Dimethoxymethane	0	0	0.167	5.85	0.232	0	0.53	6.95	1.35	1.37	-0.02	1.38	-0.03
42	Dimethoxy ethane	1	0	0.133	5.63	0.208	0	0.54	8.72	1.38	1.39	-0.01	1.39	-0.01
43	Diethoxy methane	2	0	0.222	5.48	0.221	0	0.55	10.5	1.37	1.37	0	1.37	0
44	1,1-Diethoxyethane	3	0	0.19	5.37	0.211	0	0.56	12.2	1.38	1.39	-0.01	1.39	-0.01
45	1-Methoxypropane	1	0	0.143	4.94	0.331	0	0.55	8.26	1.36	1.37	-0.01	1.37	-0.01
46	2-Methoxypropane	2	0	0.071	4.94	0.321	0	0.55	8.26	1.36	1.37	-0.01	1.37	-0.01
47	1-Ethoxypropane	2	0	0.176	4.9	0.299	0	0.56	10	1.37	1.37	0	1.37	0
48	2-Ethoxy ethyl ether	2	0	0.286	5.6	0.223	0	0.56	16.2	1.41	1.41	0	1.4	0.01
49	Benzyl ethyl ether	1	0	0.136	6.19	0.193	0	0.64	14	1.5	1.49	0.01	1.5	0
50	Propyl ether	2	0	0.2	4.87	0.322	0	0.56	11.8	1.38	1.39	-0.01	1.38	0
51	Iso propyl ether	4	0	0.1	4.87	0.319	0	0.56	11.8	1.37	1.38	-0.01	1.38	-0.01
52	n-Amyl ether	2	0	0.241	4.81	0.334	0	0.57	17.1	1.41	1.41	0	1.41	0
53	Ethyl octyl ether	2	0	0.25	4.8	0.329	0	0.57	18.8	1.41	1.42	-0.01	1.42	-0.01
54	2-Chloro ethyl ether	0	2	0.143	9.53	0.349	0	0.67	9.98	1.46	1.45	0.01	1.45	0.01
55	Methanol	0	0	0	5.34	0	0	0.5	2.98	1.33	1.33	0	1.34	-0.01
56	Ethanol	1	0	0	5.12	0.316	0	0.53	4.74	1.36	1.35	0.01	1.35	0.01
57	2-Iodoethanol	0	1	0	19.11	0.746	0	0.82	7.4	1.57	1.56	0.01	1.57	0
58	2-Propanol	2	0	0	5.01	0.365	0	0.54	6.5	1.38	1.37	0.01	1.35	0.03
59	2-Pentanol	2	0	0.118	4.9	0.328	0	0.56	10	1.41	1.39	0.02	1.39	0.02
60	Allyl alcohol	0	0	0.111	5.81	0.236	0	0.57	5.74	1.41	1.4	0.01	1.39	0.02
61	Benzyl alcohol	0	0	0.045	5.19	0.299	0	0.58	12.8	1.54	1.49	0.05	1.47	0.07
62	1-Butanol	1	0	0.143	4.94	0.359	0	0.55	8.26	1.4	1.38	0.02	1.37	0.03
63	tert-Butyl alcohol	3	0	0	4.94	0.362	0	0.55	8.26	1.38	1.37	0.01	1.36	0.02
64	1-Butoxy-2-propanol	2	0	0.208	5.29	0.285	0	0.56	14	1.42	1.41	0.01	1.41	0.01
65	1,2 Ethane diol	0	0	0.111	6.21	0.224	0	0.52	5.19	1.43	1.4	0.03	1.38	0.05
66	1-Hexanol	1	0	0.2	4.87	0.357	0	0.56	11.8	1.42	1.4	0.02	1.4	0.02
67	1-Heptanol	1	0	0.217	4.84	0.356	0	0.56	13.6	1.42	1.41	0.01	1.41	0.01
68	1-Octanol	1	0	0.231	4.82	0.356	0	0.57	15.3	1.43	1.42	0.01	1.42	0.01
69	Cyclopentanol	0	0	0	5.38	0.277	0	0.58	9.26	1.45	1.44	0.01	1.44	0.01



70	Methyl formate	0	0	0.143	7.51	0.166	0	0.55	4.43	1.34	1.36	-0.02	1.38	-0.04
71	Ethyl formate	1	0	0.2	6.74	0.184	0	0.56	6.19	1.36	1.36	0	1.37	-0.01
72	Ethyl acetate	2	0	0.154	6.29	0.185	0	0.57	7.95	1.37	1.38	-0.01	1.38	-0.01
73	Propyl formate	1	0	0.231	6.29	0.242	0	0.57	7.95	1.38	1.37	0.01	1.37	0.01
74	Methyl propanoate	1	0	0.154	6.29	0.186	0	0.57	7.95	1.38	1.39	-0.01	1.4	-0.02
75	Butyl formate	1	0	0.25	6.01	0.264	0	0.57	9.72	1.39	1.39	0	1.38	0.01
76	3-Methylbutyl acetate	3	0	0.182	5.66	0.28	0	0.58	13.2	1.4	1.41	-0.01	1.41	-0.01
77	Propyl butanoate	2	0	0.227	5.66	0.246	0	0.58	13.2	1.4	1.41	-0.01	1.41	-0.01
78	Allyl acetate	1	0	0.214	6.68	0.182	0	0.6	8.95	1.4	1.41	-0.01	1.41	-0.01
79	Benzyl acetate	1	0	0.143	7.15	0.174	0	0.65	13.7	1.52	1.51	0.01	1.52	0
80	Octyl acetate	2	0	0.258	5.38	0.283	0	0.58	18.5	1.42	1.42	0	1.43	-0.01
81	Phenyl acetate	1	0	0.111	7.56	0.159	0	0.66	12	1.5	1.51	-0.01	1.51	-0.01
82	Ethyl trifluoroacetate	1	3	0.154	10.15	0.111	0	0.55	7.77	1.31	1.31	0	1.31	0
83	Ethyl fluoroacetate	1	1	0.154	7.58	0.156	0	0.56	7.89	1.38	1.38	0	1.37	0.01
84	Acetaldehyde	1	0	0	6.29	0.236	0	0.57	3.98	1.33	1.34	-0.01	1.36	-0.03
85	Benzaldehyde	0	0	0.071	7.58	0.17	0	0.7	9.74	1.55	1.54	0.01	1.53	0.02
86	Butyraldehyde	1	0	0.167	5.55	0.318	0	0.58	7.5	1.38	1.38	0	1.38	0
87	Diethylamine	2	0	0.133	4.57	0.319	0.1	0.55	8.81	1.39	1.39	0	1.37	0.02
88	Triethylamine	3	0	0.143	4.6	0.27	0	0.56	12.3	1.4	1.4	0	1.4	0
89	Propylamine	1	0	0.083	4.55	0.394	0.1	0.54	7.05	1.39	1.39	0	1.39	0
90	Dipropylamine	2	0	0.19	4.6	0.35	0	0.56	12.3	1.4	1.41	-0.01	1.41	-0.01
91	Butylamine	1	0	0.133	4.57	0.381	0.1	0.55	8.81	1.4	1.41	-0.01	1.4	0
92	tert-butyl amine	3	0	0	4.57	0.394	0.1	0.55	8.81	1.37	1.37	0	1.38	-0.01
93	Isobutylamine	2	0	0.067	4.57	0.407	0.1	0.55	8.81	1.4	1.4	0	1.39	0.01
94	Diisobutylamine	4	0	0.148	4.62	0.383	0	0.57	15.9	1.41	1.41	0	1.42	-0.01
95	Allylamine	0	0	0.1	5.19	0.262	0.1	0.57	6.29	1.42	1.42	0	1.43	-0.01
96	Benzylamine	0	0	0.043	4.92	0.304	0	0.58	13.3	1.54	1.53	0.01	1.54	0
97	Aniline	0	0	0	6.65	0.176	0.1	0.66	9.29	1.59	1.57	0.02	1.58	0.01
98	Ethyl Aniline	1	0	0.1	6.06	0.19	0.1	0.64	12.8	1.56	1.54	0.02	1.55	0.01
99	p-Fluoroaniline	0	1	0	7.94	0.155	0.1	0.66	9.23	1.52	1.51	0.01	1.52	0
100	Pentyl amine	1	0	0.167	4.59	0.374	0.1	0.56	10.6	1.45	1.43	0.02	1.42	0.03
101	Heptyl amine	1	0	0.208	4.61	0.367	0	0.56	14.1	1.42	1.43	-0.01	1.45	-0.03
102	Acetone	2	0	0	5.81	0.303	0	0.57	5.74	1.36	1.36	0	1.36	0
103	2-Heptanone	2	0	0.19	5.19	0.308	0	0.58	12.8	1.41	1.41	0	1.41	0
104	3-Hexanone	2	0	0.167	5.27	0.274	0	0.58	11	1.4	1.4	0	1.4	0
105	2-Octanone	2	0	0.208	5.13	0.314	0	0.58	14.6	1.42	1.41	0.01	1.42	0
106	Cyclohexanone	0	0	0	5.77	0.262	0	0.6	10.3	1.45	1.46	-0.01	1.46	-0.01
107	Cyclo heptanone	0	0	0	5.61	0.273	0	0.6	12	1.46	1.47	-0.01	1.48	-0.02
108	Iso propyl acetone	3	0	0.111	5.27	0.329	0	0.58	11	1.4	1.4	0	1.4	0
109	Methyl vinyl ketone	1	0	0.1	6.37	0.204	0	0.61	6.74	1.41	1.41	0	1.41	0
110	Diethyl ketone	2	0	0.133	5.38	0.249	0	0.58	9.26	1.39	1.39	0	1.39	0
111	Methyl propyl ketone	2	0	0.133	5.38	0.29	0	0.58	9.26	1.39	1.39	0	1.39	0
112	Ethyl phenyl ketone	1	0	0.1	6.71	0.179	0	0.66	13.3	1.53	1.51	0.02	1.52	0.01
113	Methyl isopropyl ketone	3	0	0.067	5.38	0.294	0	0.58	9.26	1.39	1.39	0	1.39	0
114	Ethyl propyl ketone	2	0	0.167	5.27	0.274	0	0.58	11	1.4	1.4	0	1.4	0
115	Dipropyl ketone	2	0	0.19	5.19	0.291	0	0.58	12.8	1.41	1.41	0	1.41	0

116	Acetylacetone	2	0	0.143	6.68	0.226	0	0.6	8.95	1.45	1.42	0.03	1.41	0.04
117	1-pentene	1	0	0.143	4.68	0.359	0	0.59	8.81	1.37	1.38	-0.01	1.38	-0.01
118	3-Ethyl-2-pentene	3	0	0.1	4.68	0.276	0	0.59	12.3	1.41	1.41	0	1.41	0
119	1-Hexene	1	0	0.176	4.68	0.358	0	0.59	10.6	1.38	1.39	-0.01	1.4	-0.02
120	Cyclo hexane	0	0	0	5.14	0.293	0	0.61	9.81	1.45	1.44	0.01	1.45	0

Table 2: Architecture and Specification of the generated ASNN

No. of nodes in the input layer	8
No. Of nodes in the hidden layer	7
No. of nodes in the output layer	1
Seed value	78
Number of KNN	10
Activation function	Logistic $1/(1+\exp(-x))$

Internal Validation

Cross-validation is a statistical method to evaluate the stability of developed models. In this validation technique, a number of modified data sets are created by deleting, one compound. For each reduced data set, the model is calculated and responses for the deleted compounds are predicted from the model. In this study, the predictive power of the models is checked by leave-one-out (LOO) cross-validation and the square of the cross-validated correlation coefficient (q^2) is used to measure the models predictivity. A good correlation is obtained with LOO correlation co-efficient $q^2 = 0.961$ for training and 0.939 for testing. So the predictive power of the ASNN model is very significant.

External validation

After the Cross-validation process, predictive ability of the model is estimated from an external test set of compound not included in the training set. The test set included 28 compounds with diverse set of chemical compounds. The predicted refractive index for 28 compounds using ASNN is given in Table 3. The quality of prediction is evaluated by using two parameters: squared correlation co-efficient (R^2) and Root mean square error (RMSE). The high value of R^2 and low value of RMSE indicated a more stable model. The statistical performance of the Associative Neural Network QSPR model for refractive index estimation is summarized in Table 4. The Root mean square errors of ASNN model for training and testing are 0.01 and 0.0098 respectively. Figure 1 shows scatter plot of the ASNN predicted versus experimental values of refractive indices for training and test set. Squared correlation co-efficient (R^2) of 0.962 for training and 0.9527 for testing confirms the suitability of the ASNN model and shows a good agreement of ASNN predicted values with experimental one. Analyzing the residual values obtained for training and testing, it can be concluded that 130 out of 148 cases, the model described by ASNN resulted in better values (from 0.00 to 0.01), This confirms the refractive index of organic compounds are strongly dependent on selected descriptors. The residual of the ASNN predicted values are plotted against their experimental values shown in

Figure 2. The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of ASNN. The predicted values have an average absolute deviation (%AAD) of 0.48% for training set and 0.38 % for testing set compared with experimental values. Therefore this QSPR relationship can be used for the prediction of refractive index values with a high degree of confidence.

Table 3: Predicted refractive index values using ASNN and PNN for tested compounds that have not been included in training set.

S.No	Compound Name	Experimental	Predicted by ASNN	Residual	Predicted by PNN	Residual
1	Propylbromide	1.43	1.45	-0.02	1.45	-0.02
2	2,2 Dimethyl butane	1.37	1.38	-0.01	1.37	0
3	2-MethylPentane	1.37	1.38	-0.01	1.38	-0.01
4	Nonylcyclopentane	1.45	1.46	-0.01	1.48	-0.03
5	4-Ethyl Heptane	1.41	1.41	0	1.41	0
6	Octane	1.4	1.4	0	1.4	0
7	2,7-Dimethyl octane	1.41	1.41	0	1.41	0
8	2,4-Dimethyl Nonane	1.42	1.42	0	1.42	0
9	Tetradecane	1.45	1.43	0.02	1.43	0.02
10	Diethyl ether	1.35	1.36	-0.01	1.36	-0.01
11	Ethyl pentyl ether	1.39	1.39	0	1.4	-0.01
12	Ethyl isobutyl ether	1.37	1.38	-0.01	1.38	-0.01
13	Methyl acetate	1.36	1.39	-0.03	1.39	-0.03
14	2-Methylpropyl acetate	1.4	1.4	0	1.4	0
15	Propyl acetate	1.38	1.39	-0.01	1.39	-0.01
16	2-Fluoroethanol	1.37	1.37	0	1.34	0.03
17	1-Decanol	1.44	1.44	0	1.43	0.01
18	1-Nonanol	1.43	1.43	0	1.43	0
19	Propanoic acid	1.39	1.39	0	1.39	0
20	n-Butylbenzene	1.49	1.49	0	1.5	-0.01
21	Dibutylamine	1.42	1.43	-0.01	1.44	-0.02
22	Triisobutylamine	1.43	1.43	0	1.43	0
23	2-Hexanone	1.4	1.4	0	1.4	0
24	2-Pentanone	1.39	1.39	0	1.39	0
25	Cyclo pentanone	1.44	1.45	-0.01	1.45	-0.01
26	Methyl ethyl ketone	1.38	1.38	0	1.38	0
27	Acetophenone	1.53	1.53	0	1.52	0.01
28	Methyl isobutyl ketone	1.4	1.4	0	1.4	0

The data set used in ASNN is also applied to Polynomial Neural Network (PNN) for comparative study. The PNN represents a new challenge for development of physicochemical data prediction methods. It should be noted that, predictive ability of the PNN model is

enhanced by varying number of variables in the polynomials. Numbers of iterations are increased until convergence is achieved. Degree of polynomial represents number of terms in the polynomial equation that will be used in model selection. The present work involves second order polynomial equation which includes 22 square and cross terms of input descriptors produce the best result. This result is achieved in 50 iterations. The polynomial equation for the best model had the following formula:

$$\text{Refractive Index (n)} = \left\{ \begin{array}{l} -0.00418 * (\text{AMW}^2) - 0.797 * (\text{RBF}^2) + 0.0307 * \text{Mp} * \text{Sp} - 0.0339 * \\ \text{AMW} * \text{Mp} - 3.49 * \text{RBF} * \text{nNr} - 3.34 \text{e}^{-04} * (\text{Sp}^2) - 0.0206 * (\text{C}-001) \\ * \text{X2Av} - 0.0268 * \text{nX} * \text{nNr} - 0.0359 * \text{nX} * \text{X2Av} - 0.379 * \text{X2Av} * \text{nNr} - \\ 0.0820 * (\text{X2Av}^2) + 1.64 - 0.320 * (\text{C}-001) * \text{nNr} + 0.320 * \text{nX} * \text{RBF} \\ + 0.0220 * (\text{C}-001) * \text{nX} + 0.151 * \text{nNr} * \text{Sp} + 0.0166 * \text{AMW} * \text{X2Av} \\ + 0.00127 * (\text{C}-001^2) + 0.171 * \text{AMW} * \text{Mp} - 0.00254 * (\text{C}-001) * \text{AMW} \\ - 0.0824 * \text{nX} \end{array} \right.$$

Table 1 & 3 gives the calculated refractive index values for training and test set using the best fitted polynomial equation. Figure 3 shows experimental refractive index values versus calculated for both Training and Test set. The statistical performance of PNN analysis is included in Table 4. Squared correlation coefficient of 0.91 for training and 0.901 for testing confirms the predictive ability of PNN but accuracy is less compared to ASNN predicted model.

Table 4: Statistical Comparison of QSPR models obtained using ASNN and PNN.

Data set	ASNN			PNN		
	R ²	q ²	RMSE	R ²	q ²	RMSE
Training	0.962	0.961	0.01	0.910	0.901	0.0125
Testing	0.9527	0.939	0.0098	0.901	0.898	0.013

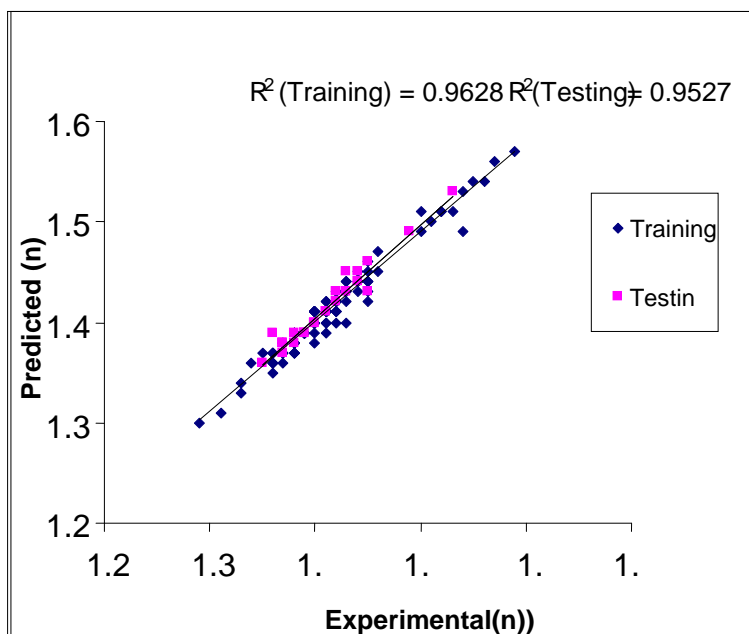


Fig 1: Scatter plot of Experimental versus Predicted Refractive indices for Training and Testing Set (ASNN)

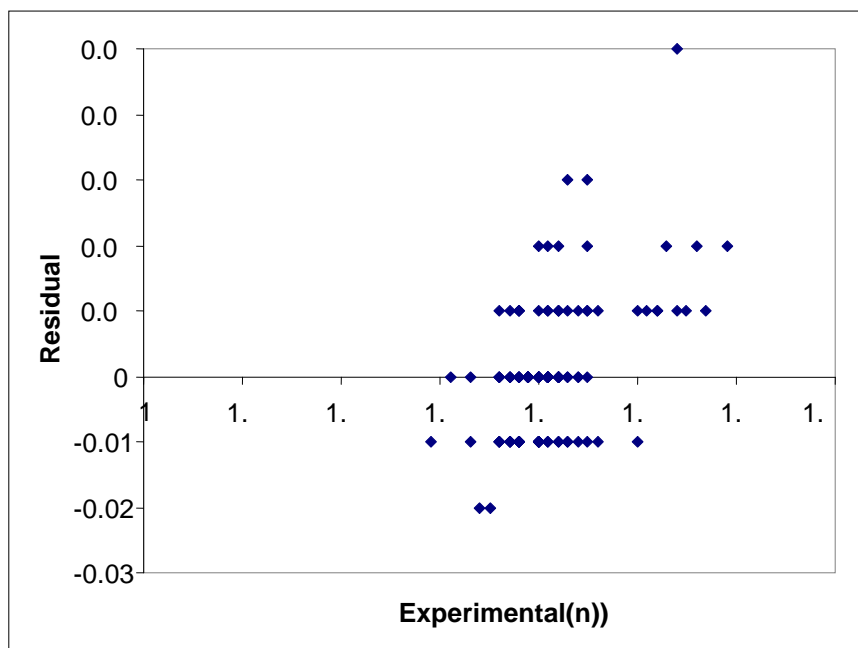


Fig 2: Experimental versus residual values for ASNN predicted values

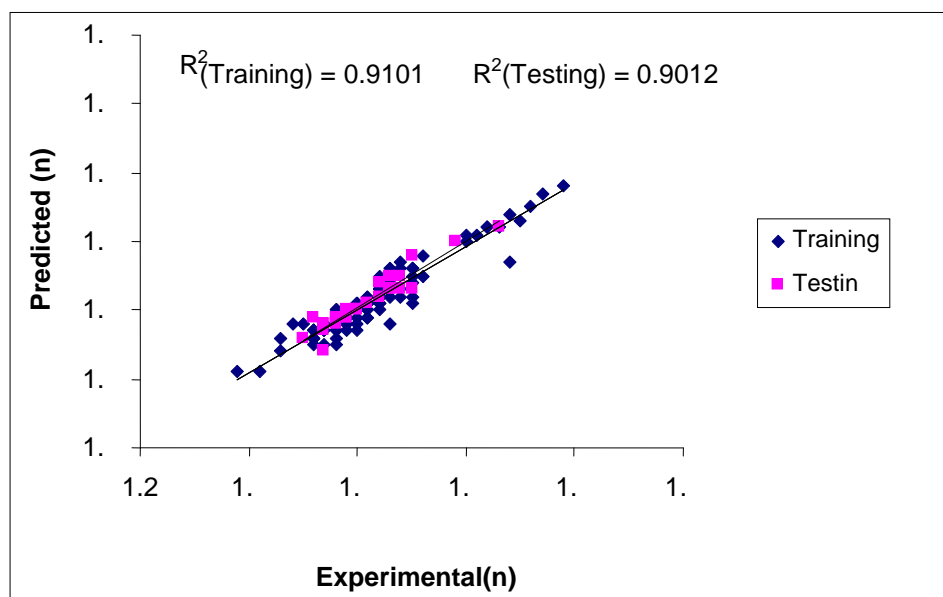


Fig 3: Scatter plot of Experimental versus Predicted Refractive indices for Training and Testing Set (PNN)

Interpretation of descriptors

It is well accepted that refractive index (n) is governed by London dispersion forces in organic compounds. It should be noted that Molecular interaction force is directly proportional to polarizability [25]. Therefore descriptors M_p and S_p directly encode information related to molecular polarizability. The atom-centered fragment descriptor (C-001) is used to differentiate the isomers within same group of compounds. The descriptor n_{Nr} relates a polar interaction among the molecules in the bulk liquid. It is well known that molecular polarizability is directly proportional to number of electrons in the molecule. When the size (number of electrons) of the molecule increases, intermolecular interaction is stronger, therefore the higher the refractive index will be. Compounds containing more polarizable groups (Halogen atoms) will normally have higher refractive indices than compounds containing less polarizable groups (oxygen atoms). Hence the descriptor n_X represents a measure of molecular polarizability. The descriptor RBF represents Molecular flexibility which increases with the number of flexible bonds in the molecule and the importance associated with flexible bond might be owing to the fact that they play an important role to identify different groups of organic compounds with similar property. This descriptor is also used to differentiate cyclic compounds with other type of compounds. The descriptor X_{2Av} is used for heteroatom differentiation. It can be concluded that the descriptors in the present QSPR model has definite chemical meaning and these can account the structural features that affect on the refractive index of the organic compounds. Since density of most organic compounds is roughly proportional to refractive index [26], therefore, descriptor Average molecular weight (AMW) indirectly related to refractive index of organic compounds.

CONCLUSION

The results reported in this paper clearly show the prediction of refractive indices of wider variety of organic compounds with better statistics than other models reported in literature. Eight significant descriptors are selected by employing Unsupervised forward selection procedure. By using ASNN, a statistically significant QSPR model with the squared correlation co-efficient values for training and for testing are 0.962 and 0.9527 respectively. On the same data set, another QSPR model is developed on PNN which predict refractive index values accurately with Squared correlation co-efficient $R^2 = 0.9101$ for training and $R^2 = 0.901$ for test set. The obtained results in this paper suggest that the ASNN predicts refractive index of organic compounds very well compared with PNN. The QSPR models developed in this study can provide a useful tool to predict the refractive index of new compounds. All descriptors are solely derived from the chemical structure of compounds. The descriptors involved in the present study reveal several interaction mechanisms are important for the refractive index of organic compounds. Specifically, they include the polarizability of the molecule and molecular size dependent effects in the molecules.

REFERENCES

- [1] Mathews JH, Williams JH, Bender P, Alberty RA. Experimental Physical Chemistry 5th Ed, Mc-Graw Hill: New York 1956.



- [2] Owens JC. Appl Opt 1967; 6: 51.
- [3] Bicerano J. Prediction of Polymer Properties. 2nd ed., Marcel Dekker Inc: New York 1996.
- [4] Mon J, Flury M, Harsh JB. J Hydrol 2006; 316: 84.
- [5] Todeschini R, Consonni V. Handbook of Molecular Descriptors, Wiley-VCH Verlag: Weinheim 2000.
- [6] Hansch LC, Leo C. Exploring QSAR: Fundamentals and applications in chemistry and biology. American Chemical Society: Washington, DC 1996.
- [7] Yaffe DL. A Neural Network Approach for Estimating Physicochemical Properties Using Quantitative Structure Property Relationships (QSPRs), University of California, Los Angeles, Ph.D. Thesis 2001.
- [8] Peixum Liu, Wei King. Int J Mol Sci Rev 2009; 10: 1978-1998.
- [9] Koziol J. J Mol Des 2003; 2:315.
- [10] Ivanciuc O, Ivanciuc T, Klien DJ. SAR and QSAR Environ Res 2001; 12:1.
- [11] Katritzky AR, Sild S, Karelson M. J Chem Inf Comput Sci 1998; 38: 840.
- [12] Du Xihua, Cu Juguan. Chin J Chem Phys 2005; 18: 211.
- [13] E-Dragon 1.0, On-line software; Virtual Computational Chemistry Laboratory (vcclab), <http://146.107.217.178/lab/edragon/index.html>.
- [14] Lide DR. CRC Hand book of Chemistry and Physics. 75th Ed., CRC Press, Boca Raton 1994.
- [15] <http://www.chemaxon.com/Marvin/Sketch/index.php>.
- [16] Whitley DC, Ford MG, Livingstone DJ. J Chem Inf Comput Sci 2000,40,1160.
- [17] Vcclab website: <http://www.vcclab.org/lab/ufs/>.
- [18] Zupan J, Gasteiger J. Neural Networks for Chemists: An Introduction: VCH; Weinheim 1993.
- [19] Agrafiotis DK, Cedeno W, Lobanov VS. J Chem Inf Comput Sci 2002; 42: 903-911.
- [20] Tetko IV. Neural Processing Letters 2002; 16: 187-199.
- [21] Tetko IV. J Chem Inf Comput Sci 2002; 42:717.
- [22] Vcclab Website : <http://www.vcclab.org/asnn/>
- [23] Aksyonova TI, Tetko IV. SAMS 2003; 43: 1331.
- [24] Vcclab website : <http://www.vcclab.org/pnn/>
- [25] London F. Trans Faraday Soc 1937; 33: 8-26.
- [26] Eremenko LT, Korolev M. RUSS CHEM B+ 1972; 21:172-174.