Composing Homologous Series on Demand [1,2]

Vladimir Diky[3], *Ala Bazyleva, Andrei Kazakov, Angela Li*[*]

Applied Chemicals and Materials Division, National Institute of Standards and Technology,

Boulder, Colorado 80305-3337, USA

[*]Peak to Peak Charter School, 800 Merlin Dr, Lafayette, CO 80026. Presently: Department of Bioengineering, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX, 75080, USA

Abstract: Homologous series (or compound series with repeated incremental structural changes) are frequently used for analysis and prediction of properties of chemical substances. They are usually constructed by adding CH_2 or other functional groups to the parent molecule and form one-dimensional spaces where the molecule or substance properties are a function of one variable, the number of added groups. Analysis of the property changes in such a series can help to identify anomalies. Interpolation and limited extrapolation can also be used for property prediction. A simple method is proposed for an automated generation of such series. It shows all possible changes in molecular structure leading to the construction of series within a given collection of compounds, including non-intuitive combinations of changes. The series constructed using this algorithm may include sets of isomers, the properties of which are expected to be similar, thus increasing the coverage in sparsely populated collections of chemical compounds.

Keywords: thermodynamic properties; compound families; homologous series; molecular structure.

INTRODUCTION

Homologous series is a widely used chemical concept. Although no definition of this concept exists in the IUPAC Gold Book [1], they are usually considered to be sequences of compounds constructed by the repeated addition of the same structural unit (functional group) to the initial molecule. It is frequently restricted to just a -CH_2- group; however, in the most general sense, it can be applied to any polymer-like sequence, either organic or inorganic (e.g., [2]). Discrete structural increments effectively define a one-dimensional space. Properties of compounds typically vary systematically along this "chemical coordinate", although some notable exceptions are well-documented (*e.g.*, properties of the smallest molecule in a series, odd-even variation of properties in the crystalline state [3], crystal properties associated with polymorphism [4], and the cases when structure modifications change the polarization behavior, as will be discussed later for one of the examples involving partially fluorinated and perfluoro-groups). Reduction of the dimensionality of molecular structure variations to a single coordinate provides

---

[1] This contribution of the National Institute of Standards and Technology is not subject to copyright in the United States.

[2] Trade names are used for information purpose and do not imply endorsement by the National Institute of Standards and Technology. Similar products by other manufacturers may work as well or better.

[3] Corresponding author. E-mail: vladimir.diky@nist.gov

the foundation for important practical applications such as planning of systematic chemical studies, trend visualization, critical evaluation of experimental data with detection of outliers, and development of predictive correlations [5]. Evaluation of existing data and validation of new data are of particular interest for this work [6]; however, this use is limited to the situations when one can create a homologous series from compounds with available experimental property data. The goal of the present work was to develop an efficient automated method for composing homologous series from large collections of compounds with property data in databases such as the NIST/TRC SOURCE database [7]. This automation should reduce labor and inevitable human errors. The algorithm should be able to find which modifications of a molecular structure produce a compound series with the number of members sufficient for property correlation (3 or more).

Algorithms and methods revealing repeated differences between molecular structures do exist, for example, [8] and references therein. However, the method described in [8] is designed to perform matched molecular pair analysis (MMPA) [9], where pairs of molecules with the same changes in their structure are identified. It is not directly applicable to the search for series. Furthermore, direct analysis of the serialized form of molecular structures typically used for database storage is complex. Both of these factors motivated the search for other solutions, and the procedure described here is based on the representation of molecular structures by vectors corresponding to group-contribution property prediction methods. Examples of well-known prediction methods are those of Benson [10] and Joback [11]. Necessary definitions of vector-based methods are as follows (the examples are given in Tables 1-2):

Chemical space description: a given chemical space is defined by a set of unique structural groups and features. A structural group is a fragment of a molecule, which may consist of one or more atoms, *e.g.*, CH_2. A feature is a combination of one or more structural groups, such as the existence of two substituents in ortho-position in a benzene ring. Each method has a specific algorithm for structure decomposition into these groups/features. To facilitate automation, a unique integer identifier (ID) is assigned to each group and feature defined within a selected method (for example: ID=1 for CH_3 group, 2 for CH_2 group, *etc.*).

Individual compound description: a compound is represented by a vector of structural groups/features and their counts (feature vector) after applying a method-specific structure-decomposition algorithm.

Molecular structure modification description: the difference between two compounds is represented by the difference between the corresponding feature vectors of the two compounds, which, in turn, also forms a vector (increment vector). The feature vector along a homologous series encounters a repeated change, which can be mathematically expressed as a repeated addition or subtraction of the same increment vector.

Vector size: a feature vector and a corresponding increment vector should have the dimension of the entire chemical space of defined structural groups and features. Such a dimension may be very large, and the majority of vector elements would usually be zeros. Hence, a compact form of vectors is used containing only non-zero vector elements (present groups/features) accompanied by the group/feature identifiers, symbolic or numeric, as shown in Table 1.

It is important that only one feature vector for a selected method corresponds to each molecule (uniqueness). The opposite does not have to be true. A set of isomers may have the same feature vector, for example, 2-, 3-, and 4-chloroheptanes for the Joback method. Thus, compound series built this manner may include multiple isomers of the compounds generated by the same structural change. However, generation of such series is still very helpful. First, properties of all compounds corresponding to the same feature vector value are typically close to each other. This

is the basis of the group-contribution methods, which predict the same property value for all isomers with the same feature vector. Second, if the chemical space (the space of possible molecular structures) is sparsely populated with property data, this approach may vastly increase the size of any series by including the isomers matching the same feature vectors while retaining the one-dimensional nature and all the related advantages of the corresponding trend analysis. Third, any unambiguous series generated with strict definitions of structural changes will be a subset of such an ambiguous series discussed above. Based on the first two considerations, we also propose to extend the concept of homologous series by adopting the chemical space approach from the group-contribution methods. Within this framework, we define *Expanded Homologous Series* (EHS) where each point may be populated by multiple compounds corresponding to the same feature vector.

ALGORITHM

The basic idea of the algorithm is to use a library of molecular structures and a method for building feature vectors representing molecular structures in a chemical space of functional groups, which may be defined by a property prediction method (such as the Joback method [11]). A method is deemed appropriate if any compound corresponds to only one feature vector, and a repeated modification of the molecular structure causes repeated modification of the corresponding feature vector. The algorithm used to construct the series is depicted in Fig. 1 and its example application is discussed below. Each series is based on the initial compound (seed), a pool of the remaining candidates (database), and a second compound randomly selected from the pool that defines a direction in the chemical space (direction compound). The increment vector for a selected method (the difference between the feature vectors of the seed and the direction compounds) is normalized dividing by the greatest common divisor of group/feature counts to avoid excluding possible intermediate structures. The resulting (normalized) increment vector is then applied to the seed feature vector in both directions until further fragment elimination is impossible as manifested by a negative group count or the molecule size exceeding the maximum size available in the pool (the maximum size should be determined beforehand). Compounds identified as EHS members are stored and excluded from the pool (*i.e.*, not considered further as direction compounds or members for other series). The procedure is repeated for all compounds remaining in the pool (in a random order), until the compound pool is exhausted, and a complete set of all possible series within the database is produced. The order in which the direction compounds are selected has no bearing on the final result. To accelerate the search for series associated with a specific property, the pool can be limited to compounds with data available for the needed property. In the implementation used by ThermoData Engine (TDE) software [12], the group decompositions for several prediction methods (NIST-modified UNIFAC [13], Joback [11], Benson [10], etc.) are pre-generated for all compounds in the SOURCE database [7]. The TDE interface allows users to select the molecular structure decomposition method from a list and initiate automated construction of possible series.

The application of the algorithm is demonstrated in Table 1. The seed is octane, and the direction compound is 2-methylheptane (randomly selected). The Joback method [11] with the corresponding structural groups and their IDs defined in Table 2 was chosen for building feature vectors. The numeric representation of the feature and increment vectors is as follows: the first number in each pair separated by semicolon is the count of the structural groups of the type identified by the second number in each pair, an ID listed in Table 2. Specifically:

seed feature vector: (2,1; 6,2);

direction compound feature vector: (3,1; 4,2; 1,3);

increment vector obtained as a difference between them: (1,1; -2,2; 1,3).

Application of the increment vector to the seed for the first time gives the following:

factor -1: the resulting feature vector (1,1; 8,2; -1,3) – a negative count for >CH- groups; hence, the limit is reached in this direction;

factor +1: the resulting feature vector (3,1; 4,2; 1,3) – 2-methylheptane (direction compound) and other isomeric methylheptanes (including enantiomers) and 3-ethylhexane.

Application of the increment vector to the seed for the second time gives the following:

factor +2: the resulting feature vector (4,1; 2,2; 2,3) – isomeric dimethylhexanes (including enantiomers) and 3-ethyl-2-methylpentane.

Application of the increment vector to the seed for the third time gives the following:

factor +3: the resulting feature vector (5,1; 3,3) – 2,3,4-trimethylpentane.

Application of the increment vector to the seed for the fourth time gives the following:

factor +4: the resulting feature vector (6,1; -2,2; 4,3) – a negative count for CH_2 groups; hence, the limit is reached in this direction, the search procedure is terminated for the selected direction compound.

The next direction compound is selected from the pool, and the procedure is repeated for the new increment vector. Of note: all compound identified above are excluded from the pool.

Here, for the clarity of the discussion to follow, we also define a chemical coordinate for a series member numerically as a factor – a number of increments applied to the seed feature vector that produces the feature vector for this specific series member. As defined, this coordinate can be negative or positive, depending on the direction. For the example given in Table 1, the chemical coordinate varies from 0 to 3.

As seen, the resulting series are not intuitive from a human perspective (if compared with, *e.g.*, the simple n-alkane series) as each modification of the structure may involve several structural groups simultaneously, but the algorithmic implementation is able to find all EHS in a systematic manner. Availability of these additional series may be very helpful for sparsely populated databases. If multiple sets of compounds are generated, they can be ranked by the size and/or nature of the structure variation. Analysis of changes in non-trivial series should be done cautiously in order to distinguish erroneous values and actual irregularities, which can be justified.

EHS for binary mixtures can also be constructed by combining one component (chemical compound) with an EHS for the second component or by synchronous variation of both components.

MORE EXAMPLES

If 2,2-difluoroethanol is taken as the seed, the procedure automatically generates several EHS using the SOURCE database [7]. Four representative cases are shown in Fig. 2. The corresponding feature vectors are given in Table 3, and the details regarding the members of the series are provided in Table 4. The codes assigned to the Joback groups are given in Table 2. To involve more compounds in series, another definition of structural groups (fragment families) is also shown there. Each of those is an umbrella for a set of Joback's groups (similar to UNIFAC main groups being umbrellas for the sets of corresponding subgroups [13]). Like the UNIFAC main groups, the fragment family 1f combines all Joback groups formed by sp^3 carbon atoms and surrounding hydrogen atoms (CH_3, CH_2, CH, and C). In contrast to the UNIFAC main groups, families of Joback's fragments do not contain combined entities such as CH_3CO. Fig. 3 shows the

variation of the normal boiling point (NBP) in each series of compounds. A smooth (in fact, nearly linear) behavior is observed for each series, except for the first one, where the special behavior can be explained by polarization of the C-H bond by the neighboring F atoms and increasing intermolecular attractive forces. That polarization does not occur for CF_3 and CH_3 groups present in the terminal members of that series.

Vapor pressures generated by the TDE software [12] via evaluation and regression of experimental data available in the database for a series of compounds involving hexafluorobenzene as the seed (Table 5) are shown in Fig. 4 in a logarithmic form as function of reciprocal temperature $1/T$. Fig. 5 shows the corresponding Waring functions [14] derived from these vapor pressures, which are equal to the enthalpy of vaporization derived from vapor pressure by the Clapeyron-Clausius equation, ignoring liquid volume and gas non-ideality (*i.e.*, $d \ln(P)/d (1/T)$). Compounds having identical feature vectors exhibit very similar properties, and variation of the vapor pressure along the chemical coordinate in a series is regular.

As mentioned, the proposed automatic procedure will also identify the regular homologous series. For example, if 1-butanol is used as the seed with the Joback groups, the conventional family of 1-alkanols (*i.e.*, methanol, ethanol, 1-propanol, 1-butanol, 1-pentanole, *etc.*) is among 50 generated EHSs. In addition, non-trivial series are also generated, such as the one composed of 1,3-propanediol, 1-butanol, and pentane (Table 5), based on replacement of OH group with CH_3 (increment vector (1, 1; -1, 20)).

Liquid-liquid equilibrium (LLE) data for an automatically generated series of binary mixtures formed by polychlorophenols (Fig. 6) with water are shown in a composition-stretched representation [15] in Fig. 7. LLE compositions correlate (monotonously change) with application of the chosen increment vector, which represents substitution of an H atom by halogen. The figure shows that LLE compositions are very similar for subsets of isomeric polychlorophenols with the same molecular formula. An exception is *o*-chlorophenol, an endpoint of the family affected by strong intramolecular interactions.

Compound series should be used for property prediction with caution if any extrapolation is involved, or if interpolation is not smooth. Shown in Fig. 3 is an example for the normal boiling point of 2,2-difluoroethanol. The series “a” is not smooth and, hence, cannot be used for the property prediction of interest by interpolation. At the same time, using series “c” for the prediction would involve extrapolation, but this fact does not immediately exclude it from the consideration. Using linear regressions involving the data for the compounds other than 2,2-difluoroethanol, the predictions gave the results summarized in Table 6: 389.2 K (series “b”), 370.0 K (series “c”), 362.0 K (series “d”), where the second value is obtained by the extrapolation mentioned above. The experimental value for 2,2-difluoroethanol is 367.9 K [12]. In this specific case, the extrapolated value appears to be the closest to the experimental result, but this is clearly not always the case – many factors should be considered, such as the size of series, their smoothness, quality of the involved experimental data, *etc*.

In general, property predictions using homological series is a separate research area, where multiple contributing factors should be considered. The proposed series construction algorithm can provide a substantial assistance in these efforts by quickly providing multiple series alternatives for further in-depth consideration. One of the best uses of this algorithm is outlier detection, when the actual prediction is not needed, but any outliers would be clearly visible throughout several homological series.

CONCLUSIONS

The proposed method automates the construction of homologous series of chemical compounds for the analysis of trends in their properties. The procedure is comprehensive, revealing all existing series within the chosen metric, and can identify compound series that are not immediately evident, *i.e.*, beyond the traditional homologous series formed by addition or removal of CH_2 or another single functional group. The concept of Expanded Homologous Series based on the variation of vectors of structural features has been proposed, which can be efficient for chemical spaces sparsely populated with property data.

ACKNOWLEDGMENT
The authors thank Dr. E. Paulechka for helpful suggestions to improve the presentation.

Table 1. Composing an EHS from octane (seed) and 2-methylheptane (direction compound) using the Joback method. The increment vector is (1 CH_3, -2 CH_2, 1 CH) in the symbolic representation or (1,1; -2,2; 1,3) in the numeric representation according to the definitions in Table 2.

Compound	Chemical coordinate	Feature vectors[b]	
		Symbolic	Numeric
octane	0	2 CH_3, 6 CH_2	2,1; 6,2
2-methylheptane	1	3 CH_3, 4 CH_2, 1 CH	3,1; 4,2; 1,3
3-methylheptane (R-, S-, and RS-)[a]	1	3 CH_3, 4 CH_2, 1 CH	3,1; 4,2; 1,3
4-methylheptane	1	3 CH_3, 4 CH_2, 1 CH	3,1; 4,2; 1,3
3-ethylhexane	1	3 CH_3, 4 CH_2, 1 CH	3,1; 4,2; 1,3
3,4-dimethylhexane (RR-, SS-, and R*S*-)[a]	2	4 CH_3, 2 CH_2, 2 CH	4,1; 2,2; 2,3
2,3-dimethylhexane (R-, S-, and RS-)[a]	2	4 CH_3, 2 CH_2, 2 CH	4,1; 2,2; 2,3
2,4-dimethylhexane (R-, S-, and RS-)[a]	2	4 CH_3, 2 CH_2, 2 CH	4,1; 2,2; 2,3
2,5-dimethylhexane	2	4 CH_3, 2 CH_2, 2 CH	4,1; 2,2; 2,3
3-ethyl-2-methylpentane	2	4 CH_3, 2 CH_2, 2 CH	4,1; 2,2; 2,3
2,3,4-trimethylpentane	3	5 CH_3, 3 CH	5,1; 3,3

[a]Possible optical isomers are shown in the parentheses; [b]Structural group IDs are defined in Table 2

Table 2. The Joback groups [11] and fragment families as implemented in TDE software [12]

Structural fragment	ID	Fragment family ID [a]		Structural fragment	ID	Fragment family ID
–CH_3	1	1f		–O–	22	22f
–CH_2-	2	1f		–O– (ring)	23	23f
>CH-	3	1f		>C=O	24	24f
>C<	4	1f		>C=O (ring)	25	25f
=CH_2	5	5f		–CH=O (aldehyde)	26	24f
=CH–	6	5f		–COOH	27	27f
=C<	7	5f		–COO– (ester)	28	28f
=C=	8	8f		=O (except as above)	29	29f
≡CH	9	9f		–NH_2	30	30f
≡C–	10	9f		>NH	31	30f
–CH_2– (ring)	11	11f		>NH (ring)	32	30f
>CH– (ring)	12	11f		>N–	33	33f
>C< (ring)	13	11f		=N-	34	34f
=CH– (ring)	14	14f		=N- (ring)	35	35f
=C< (ring)	15	14f		=NH	36	36f
–F	16	16f		–CN	37	37f
–Cl	17	17f		–NO_2	38	38f
–Br	18	18f		–SH	39	39f
–I	19	19f		–S–	40	39f
–OH (alcohol)	20	20f		–S– (ring)	41	41f
–OH (phenol)	21	20f				

[a] Letter "f" is added to the Fragment family IDs to distinguish the Joback fragment families from the Joback groups.

Table 3. Changes in the feature vectors for the compound series for 2,2-difluoroethanol as the seed shown in Fig. 2 with the structural fragments from Table 2.

Set	Direction	Method	Change	Increment vector (symbolic and numeric) [a]
a	2,2,2-trifluoroethanol	Fragment families	+F	1 F (1,16f)
b	1,2,4-butanetriol	Joback	+CH_2OH -F	1 CH_2, -1 F, 1 OH (1,2; -1,16; 1,20)
c	2,2,3,3-tetrafluotobutane-1-ol	Joback	+CF_2	1 C, 2 F (1, 4; 2, 16)
d	1,1,2,2-tetrafluoroethane	Joback	+CHF_2 -CH_2OH	-1 CH_2, 1 CH, 2 F, -1 OH (-1,2; 1,3; 2,16; -1,20)

[a] Letter "f" is added to the Fragment family IDs to distinguish the Joback fragment families from the Joback groups.

Table 4. Selected compound series for 2,2-difluoroethanol from Fig. 2 and Table 3 with the structural fragments from Table 2.

Chemical coordinate	Compound	Feature vectors	
		Symbolic	Numeric [b]
Compound series "a" from Table 3; increment vector (1 F), or (1,16f)			
-2	ethanol	2 C(sp^3), 1 OH	2,1f; 1,20f
-1	2-fluoroethanol	2 C(sp^3), 1 F, 1 OH	2,1f; 1,16f; 1,20f
0	2,2-difluoroethanol	2 C(sp^3), 2 F, 1 OH	2,1f; 2,16f; 1,20f
1	2,2,2-trifluoroethanol	2 C(sp^3), 3 F, 1 OH	2,1f; 3,16f; 1,20f
Compound series "b" from Table 3; increment vector (1 CH_2, -1 F, 1 OH), or (1,2; -1,16; 1,20)			
-1	trifluoromethane	1 CH, 3 F	1,3; 3,16
0	2,2-difluoroethanol	1 CH_2, 1 CH, 2 F, 1 OH	1,2; 1,3; 2,16; 1,20
2	1,2,4-butanetriol	3 CH_2, 1 CH, 3 OH	3,2; 1,3; 3,20
Compound series "c" from Table 3; increment vector (1 C, 2 F), or (1,4; 2,16)			
0	2,2-difluoroethanol	1 CH_2, 1 CH, 2 F, 1 OH	1,2; 1,3; 2,16; 1,20
1	2,2,3,3-tetrafluoro-1-propanol	1 CH_2, 1 CH, 1 C, 4 F, 1 OH	1,2; 1,3; 1,4; 4,16; 1,20
1	1,2,2,3-tetrafluoro-1-propanol	1 CH_2, 1 CH, 1 C, 4 F, 1 OH	1,2; 1,3; 1,4; 4,16; 1,20
3	2,2,3,3,4,4,5,5-octafluoro-1-pentanol	1 CH_2, 1 CH, 3 C, 8 F, 1 OH	1,2; 1,3; 3,4; 8,16; 1,20
3	1,2,2,3,3,4,4,5-octafluoro-1-pentanol	1 CH_2, 1 CH, 3 C, 8 F, 1 OH	1,2; 1,3; 3,4; 8,16; 1,20
5	2,2,3,3,4,4,5,5,6,6,7,7-dodecafluoro-1-heptanol	1 CH_2, 1 CH, 5 C, 12 F, 1 OH	1,2; 1,3; 5,4; 12,16; 1,20
5	1,2,2,3,3,4,4,5,5,6,6,7-dodecafluoro-1-heptanol	1 CH_2, 1 CH, 5 C, 12 F, 1 OH	1,2; 1,3; 5,4; 12,16; 1,20
7	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9-hexadecafluoro-1-nonanol	1 CH_2, 1 CH, 7 C, 16 F, 1 OH	1,2; 1,3; 7,4; 16,16; 1,20
9	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11-eicosafluoro-1-undecanol	1 CH_2, 1 CH, 9 C, 20 F, 1 OH	1,2; 1,3; 9,4; 20,16; 1,20
Compound series "d" from Table 3; increment vector (-1 CH_2, 1 CH, 2 F, -1 OH), or (-1,2; 1,3; 2,16; -1,20)			
-1	1,2-ethanediol	2 CH_2, 2 OH	2,2; 2,20
0	2,2-difluoroethanol	1 CH_2, 1 CH, 2 F, 1 OH	1,2; 1,3; 2,16; 1,20
1	1,1,2,2-tetrafluoroethane	2 CH, 4 F	2,3; 4,16

[a]Letter "f" is added to the Fragment family IDs to distinguish the Joback fragment families from the Joback groups.

Table 5. Selected compound series for hexafluorobenzene, 1-butanol, and phenol as the seeds with the structural fragments from Table 2.

Chemical coordinate	Compound	Feature vectors	
		Symbolic	Numeric
Compound series for hexafluorobenzene; increment vector (2 =CH– (ring), -2 =C< (ring), -3 F, 1 Cl), or (2,14; -2,15; -3,16; 1,17)			
0	hexafluorobenzene	6 =C< (ring), 6 F	6,15; 6,16
1	1-chloro-2,4,6-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
1	1-chloro-2,3,5-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
1	1-chloro-2,4,5-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
1	1-chloro-2,3,4-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
1	1-chloro-3,4,5-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
1	1-chloro-2,3,6-trifluorobenzene	2 =CH– (ring), 4 =C< (ring), 3 F, 1 Cl	2,14; 4,15; 3,16; 1,17
2	1,4-dichlorobenzene	4 =CH– (ring), 2 =C< (ring), 2 Cl	4,14; 2,15; 2,17
2	1,3-dichlorobenzene	4 =CH– (ring), 2 =C< (ring), 2 Cl	4,14; 2,15; 2,17
2	1,2-dichlorobenzene	4 =CH– (ring), 2 =C< (ring), 2 Cl	4,14; 2,15; 2,17
Compound series for 1-butanol; increment vector (1 CH_3, -1 OH), or (1,1; -1,20)			
-1	1,3-propanediol	3 CH_2, 2 OH	3,2; 2,20
0	1-butanol	1 CH_3, 3 CH_2, 1 OH	1,1; 3,2; 1,20
1	pentane	2 CH_3, 3 CH_2	2,1; 3,2
Compound series for phenol; increment vector (-1 =CH– (ring), 1 =C< (ring), 1 Cl), or (-1,14; 1,15; 1,17)			
0	phenol	5 =CH– (ring), 1 =C< (ring), 1 OH (phenol)	5,14; 1,15; 1,21
1	2-chlorophenol	4 =CH– (ring), 2 =C< (ring), 1 Cl, 1 OH (phenol)	4,14; 2,15; 1,17; 1,21
1	3-chlorophenol	4 =CH– (ring), 2 =C< (ring), 1 Cl, 1 OH (phenol)	4,14; 2,15; 1,17; 1,21
1	4-chlorophenol	4 =CH– (ring), 2 =C< (ring), 1 Cl, 1 OH (phenol)	4,14; 2,15; 1,17; 1,21
2	2,4-dichlorophenol	3 =CH– (ring), 3 =C< (ring), 2 Cl, 1 OH (phenol)	3,14; 3,15; 2,17; 1,21
3	2,4,6-trichlorophenol	2 =CH– (ring), 4 =C< (ring), 3 Cl, 1 OH (phenol)	2,14; 4,15; 3,17; 1,21

Table 6. Prediction of the normal boiling point (NBP) for 2,2-difluoroethanol using linear regressions based on different automatically generated homological series (Table 4, Fig. 3). "Evaluated experimental" values were produced by the TDE software [12] by evaluation and regression of experimental data available in the database.

Chemical coordinate	Compound	NBP (K)	
		Evaluated experimental	Predicted
Compound series "a" from Table 3; increment vector (1, 16f)			
-2	ethanol	351.4	
-1	2-fluoroethanol	358.0	
0	2,2-difluoroethanol	367.9 (not used in regression)	impossible
1	2,2,2-trifluoroethanol	347.0	
Compound series "b" from Table 3; increment vector (1, 2; -1, 16; 1, 20)			
-1	trifluoromethane	191.1	
0	2,2-difluoroethanol	367.9 (not used in regression)	389.2 (interpolated)
2	1,2,4-butanetriol	587.2	
Compound series "c" from Table 3; increment vector (1, 4; 2, 16)			
0	2,2-difluoroethanol	367.9 (not used in regression)	370.0 (extrapolated)
1	2,2,3,3-tetrafluoro-1-propanol	382.2	
1	1,2,2,3-tetrafluoro-1-propanol	387.0	
3	2,2,3,3,4,4,5,5-octafluoro-1-pentanol	413.8	
3	1,2,2,3,3,4,4,5-octafluoro-1-pentanol	413.2	
5	2,2,3,3,4,4,5,5,6,6,7,7-dodecafluoro-1-heptanol	444.6	
5	1,2,2,3,3,4,4,5,5,6,6,7-dodecafluoro-1-heptanol	444.4	
7	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9-hexadecafluoro-1-nonanol	473.5	
9	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11-eicosafluoro-1-undecanol	501.2	
Compound series "d" from Table 3; increment vector (-1, 2; 1, 3; 2, 16; -1 20)			
-1	1,2-ethanediol	470.5	
0	2,2-difluoroethanol	367.9 (not used in regression)	362.0 (interpolated)
1	1,1,2,2-tetrafluoroethane	253.4	

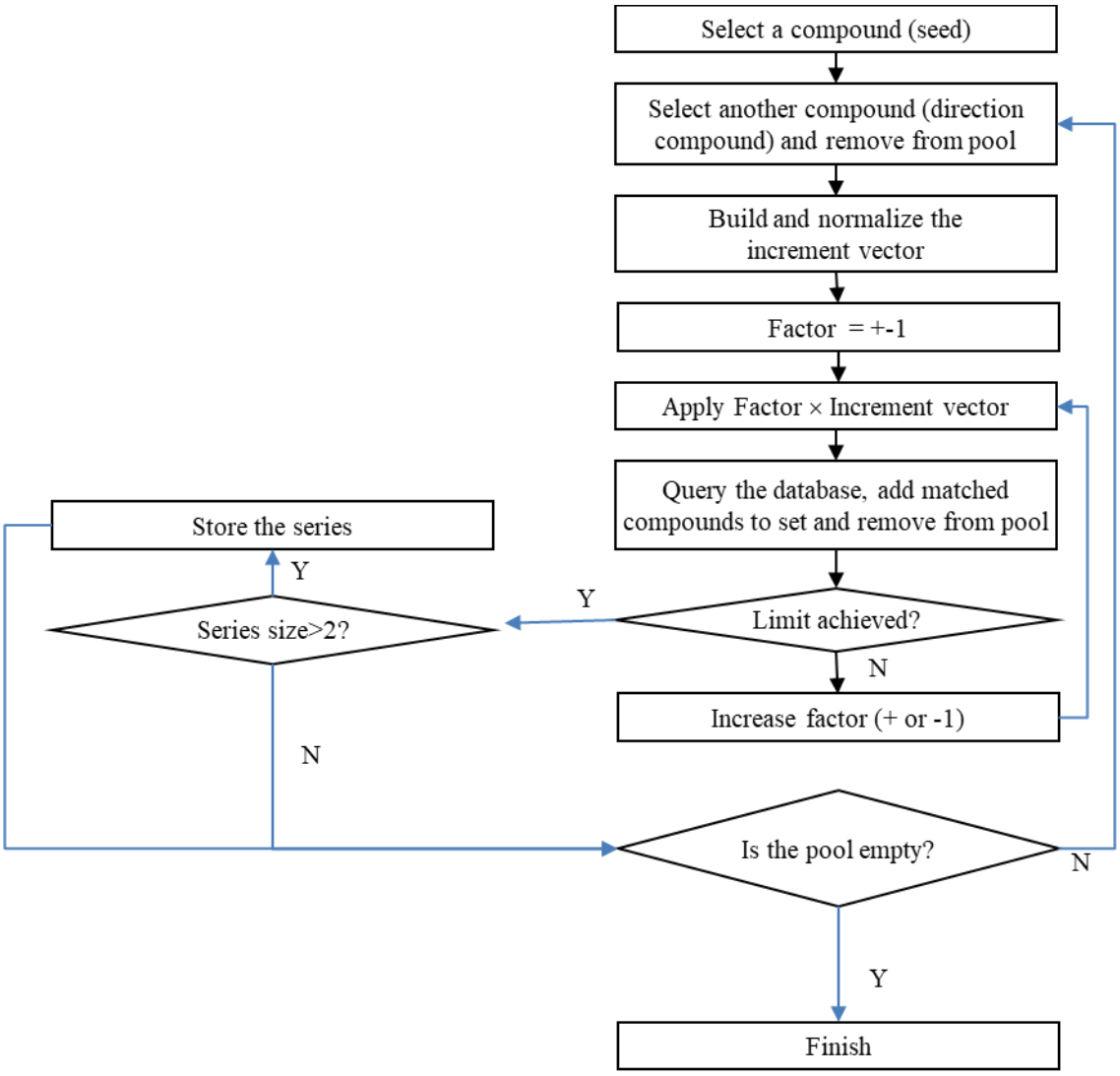
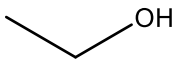
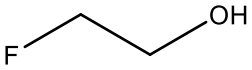
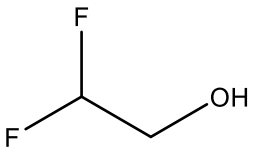


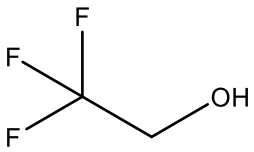
Figure 1. General algorithm for constructing compound series. The next unmarked compound is taken as the direction compound for another series based on the same seed. The limit is achieved when a negative number appears, or the molecular size exceeds the maximum size of molecules stored in the database used.

Compound series “a”

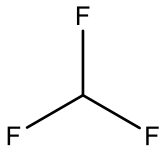


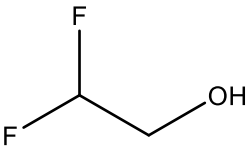


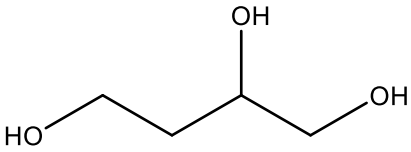




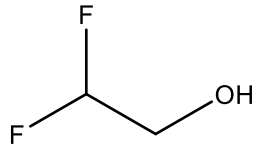
Compound series “b”

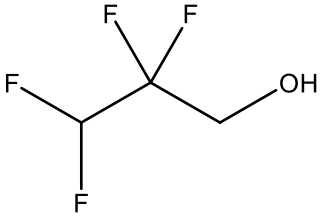


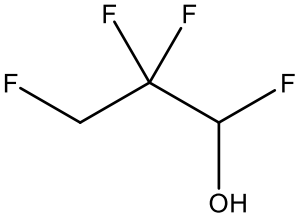


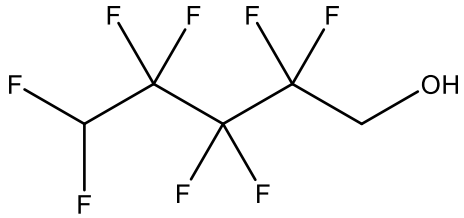


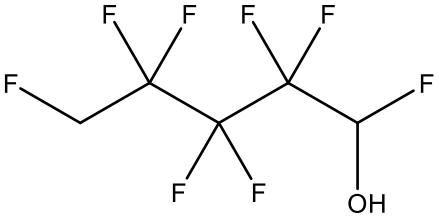
Compound series “c”

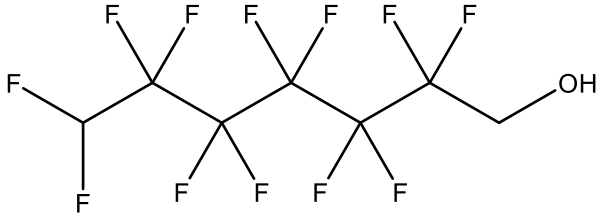


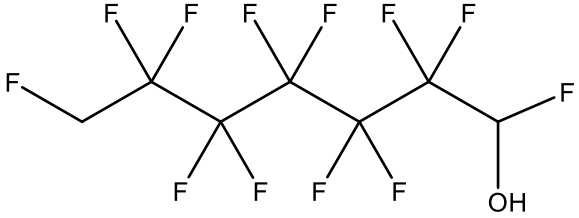


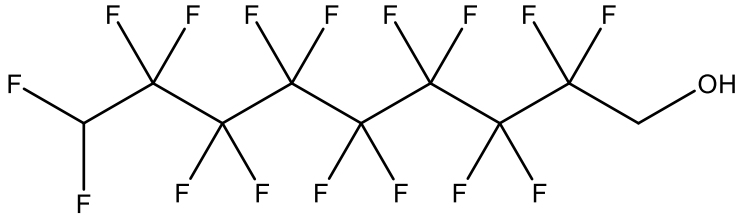


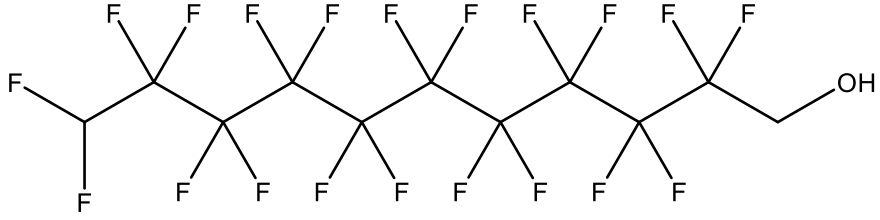












Compound series “d”

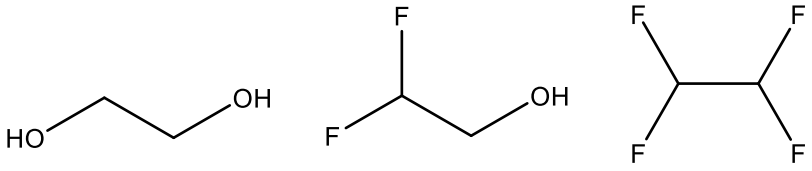
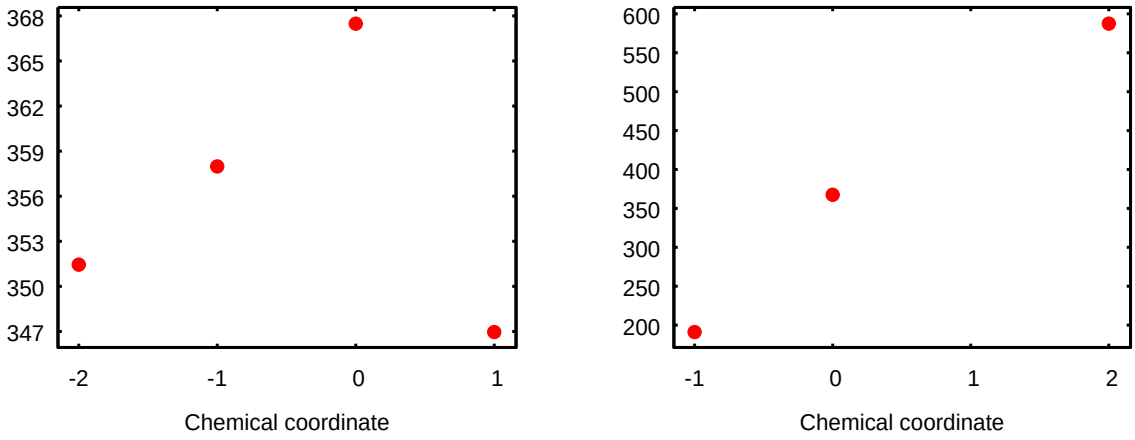
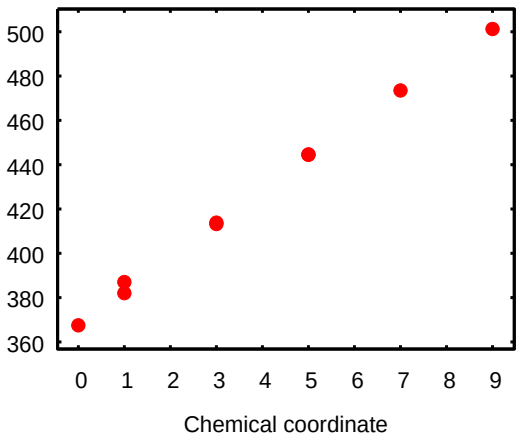


Figure 2. Automatically constructed series of compounds involving 2,2-difluoroethanol.

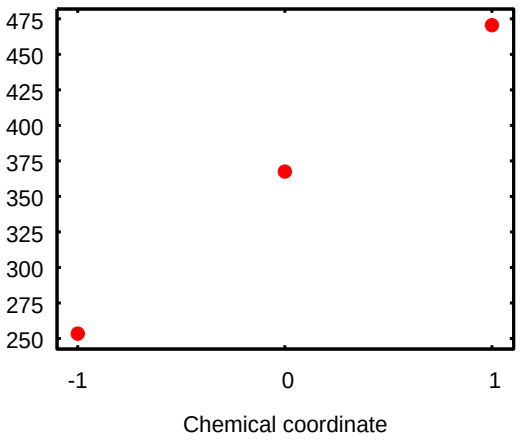


Compound series “a”

Compound series “b”



Compound series “c”



Compound series “d”

Figure 3. Normal boiling point (NBP) variations in compound series shown in Table 4 involving 2,2-difluoroethanol. The NBP values were produced by the TDE software [12] by evaluation and regression of experimental data available in the database.

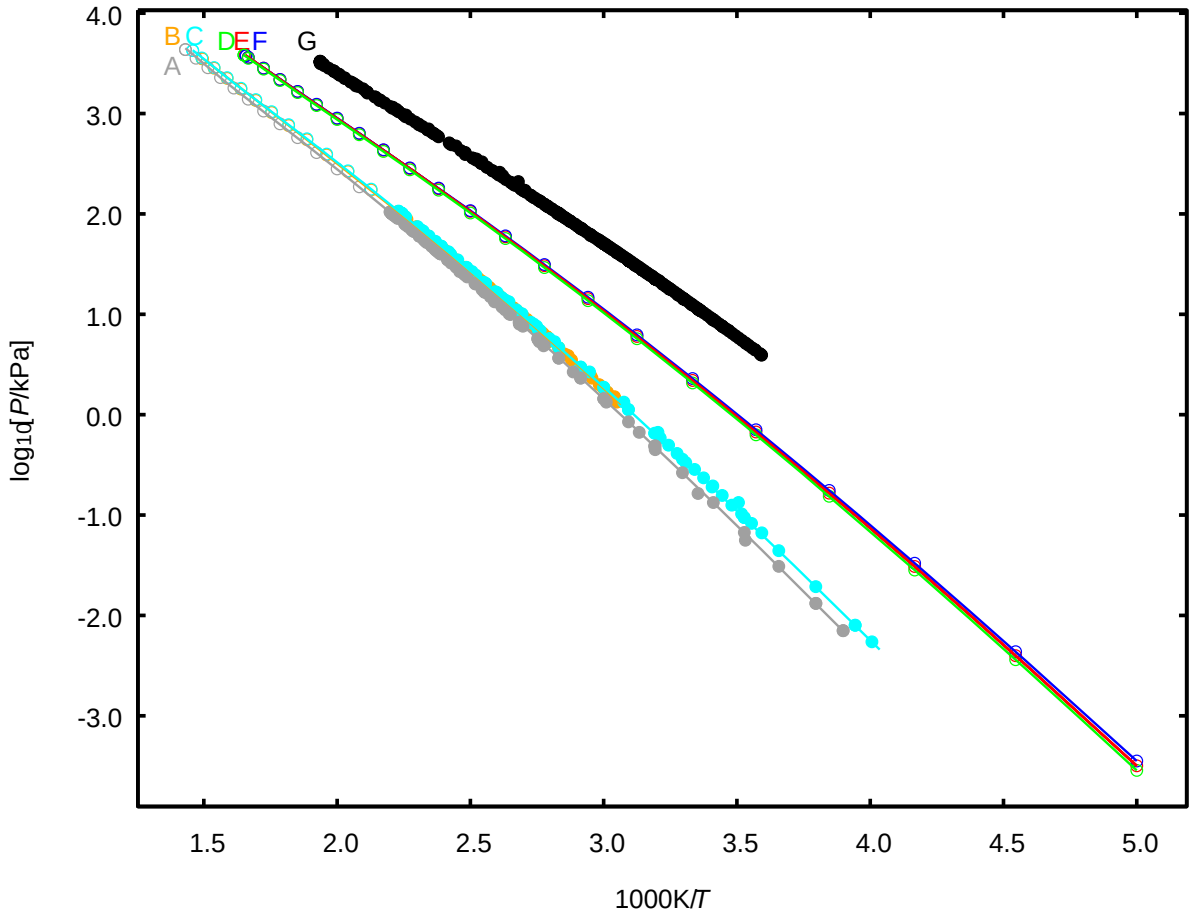


Figure 4. Logarithmic curves for saturated vapor pressures over liquid for one automatically generated series involving hexafluorobenzene as the seed obtained with the use of the Joback groups (increment (2, 14; -2, 15; -3, 16; 1, 17) according to Table 5). Chemical coordinate "2": (A) 1,2-dichlorobenzene; (B) 1,4-dichlorobenzene; (C) 1,3-dichlorobenzene. Chemical coordinate "1": (D) 1-chloro-2,3,6-trifluorobenzene; (E) 1-chloro-2,4,6-trifluorobenzene; (F) 1-chloro-2,3,4-trifluorobenzene. Chemical coordinate "0": (G) hexafluorobenzene. Experimental data are shown with filled circles, and predicted data (by the Ambrose-Walton method [16]) are shown with open circles. Lines represent the evaluation generated by the TDE software [12] with the use of experimental data and predictions. The curves for 1-chloro-2,3,5-trifluorobenzene, 1-chloro-2,4,5-trifluorobenzene, and 1-chloro-3,4,5-trifluorobenzene coincide with (E) and are not shown.

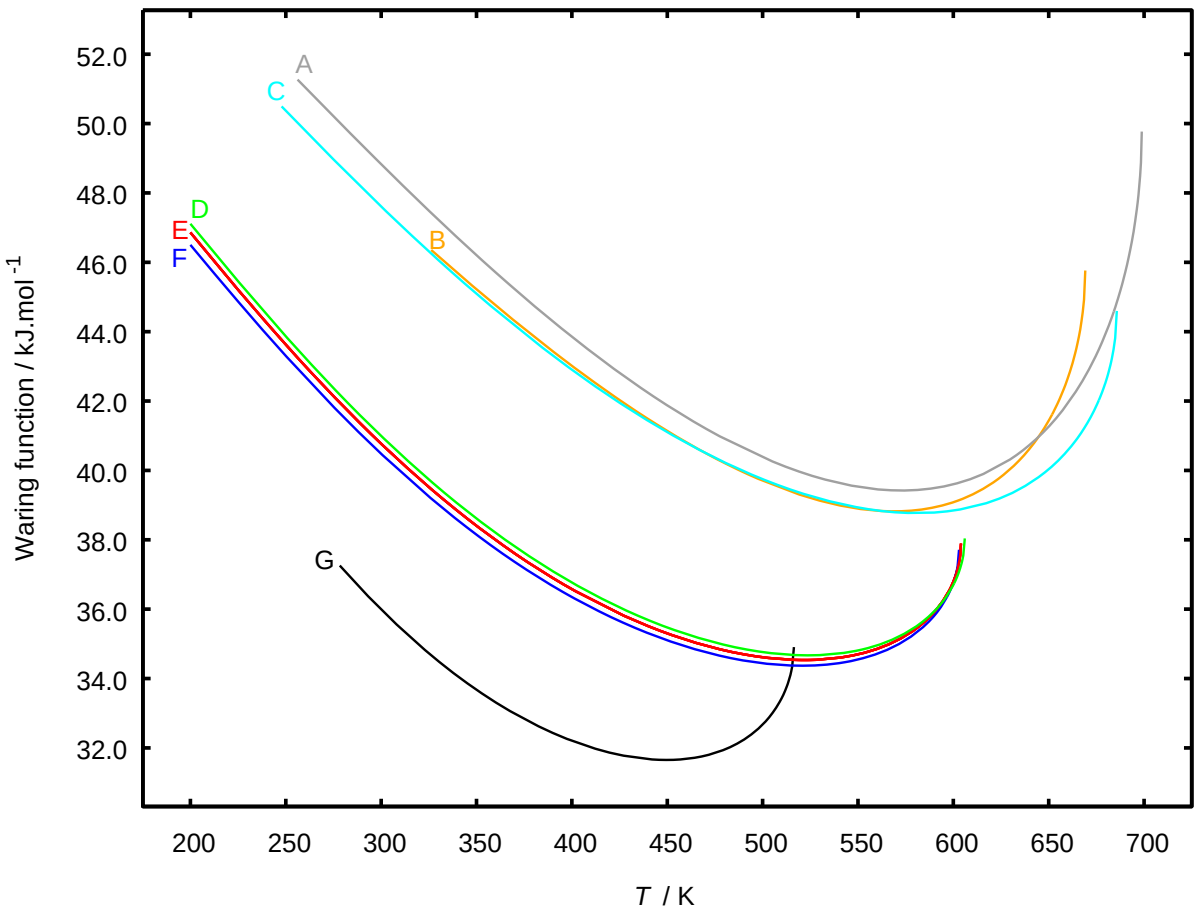


Figure 5. Waring curves for saturated vapor pressures over liquid for one automatically generated series involving hexafluorobenzene as the seed obtained with the use of the Joback groups (increment (2, 14; -2, 15; -3, 16; 1, 17) according to Table 5). Chemical coordinate "2": (A) 1,2-dichlorobenzene; (B) 1,4-dichlorobenzene; (C) 1,3-dichlorobenzene. Chemical coordinate "1": (D) 1-chloro-2,3,6-trifluorobenzene; (E) 1-chloro-2,4,6-trifluorobenzene; (F) 1-chloro-2,3,4-trifluorobenzene. Chemical coordinate "0": (G) hexafluorobenzene. The curves have been derived by the TDE software [12] from vapor pressure equations based on available experimental data and predictions. The curves for 1-chloro-2,3,5-trifluorobenzene, 1-chloro-2,4,5-trifluorobenzene, and 1-chloro-3,4,5-trifluorobenzene coincide with (E) and are not shown. (Waring function d ln(P)/d (1/T) [14] is equal to the enthalpy of vaporization derived from vapor pressure by the Clapeyron-Clausius equation ignoring liquid volume and gas non-ideality.)

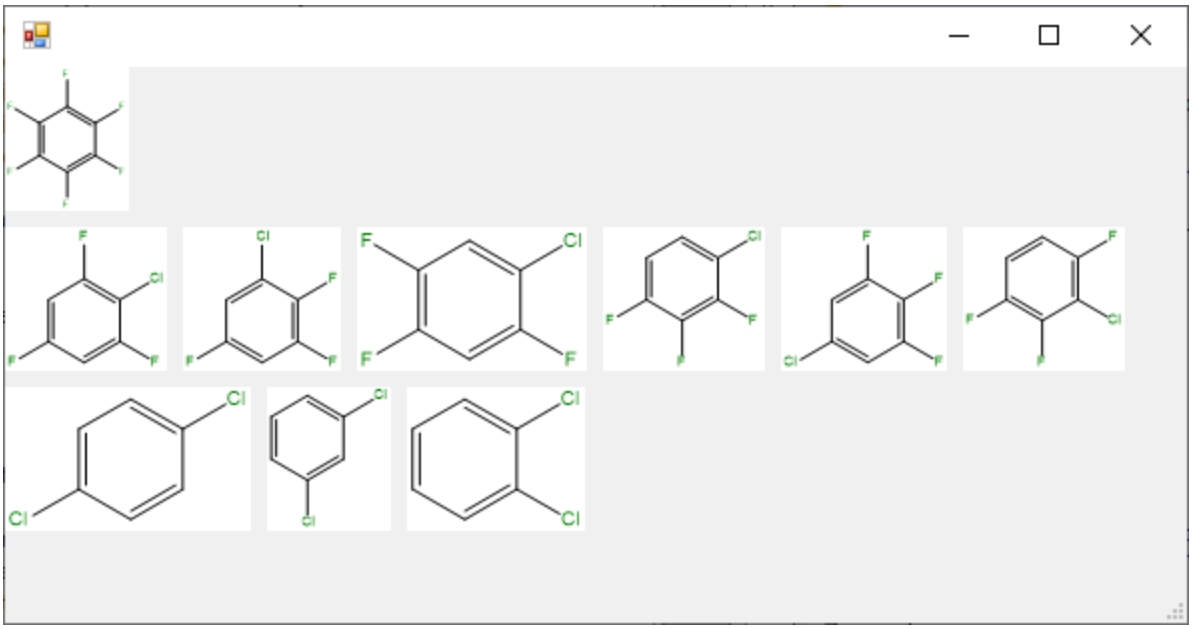


Figure 6. A view of a series for Figures 4-5 in the TDE software [12]. Each row corresponds to compounds with the same feature vector.

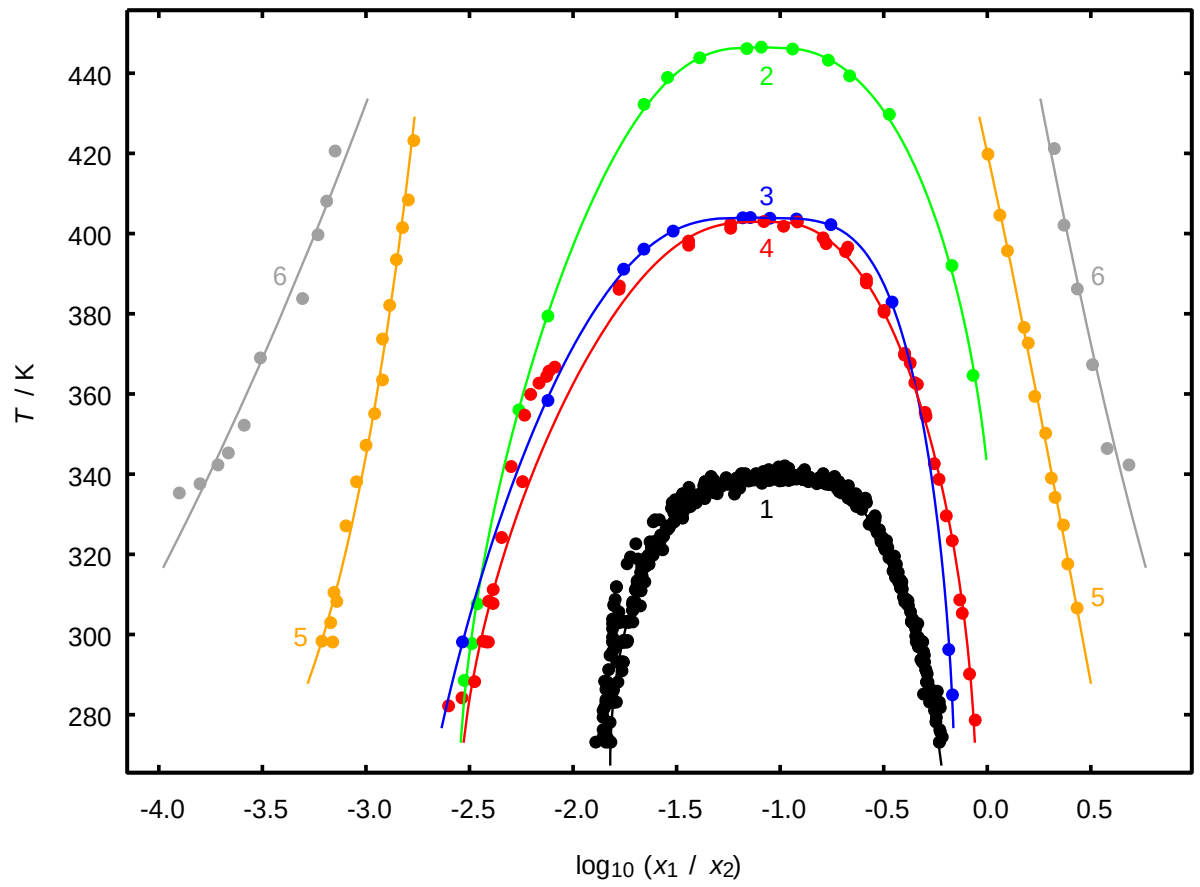


Figure 7. LLE for binary mixtures of water with a series of polychlorophenols automatically generated by the proposed algorithm. 1: phenol, 2: 2-chlorophenol, 3: 3-chlorophenol, 4: 4-chlorophenol, 5: 2,4-dichlorophenol, 6: 2,4,6-trichlorophenol. Filled circles are available experimental data; curves are smoothing equations produced by the TDE software [12]. The higher value of the upper consolute temperature for 2-chlorophenol (2) in comparison to 3- and 4-chlorophenols (3 and 4, respectively) can be explained by intramolecular electrostatic interactions reducing solvation by water.

References:

[1] Compendium of Chemical Terminology: IUPAC Gold Book. https://goldbook.iupac.org/ Accessed on 03/31/2022.

[2] Seko, A.; Togo, A.; Oba, F.; Tanaka, I. Structure and Stability of a Homologous Series of Tin Oxides. *Phys. Rev. Lett.* **2008**, *100*, 045702. https://doi.org/10.1103/PhysRevLett.100.045702

[3] Pradeilles, J. A.; Zhong, S.; Baglyas, M.; Tarczay, G.; Butts, C. P.; Myers, E. L.; Aggarwal, V. K. Odd–even alternations in helical propensity of a homologous series of hydrocarbons. *Nature Chemistry* **2020**, *12*, 475–480. https://doi.org/10.1038/s41557-020-0429-0

[4] Lohani, S.; Grant, D. J. W. Thermodynamics of Polymorphs. *In*: Polymorphism: in the Pharmaceutical Industry, Hilfiker, R. (Ed.); Wiley-VCH, 2006, pp. 21–42 https://doi.org/10.1002/3527607889.ch2

[5] Bloxham, J. C.; Hill, D.; Knotts IV, T. A.; Giles, N. F.; Wilding, W. V. Liquid Heat Capacity Measurements of the Linear Dicarboxylic Acid Family via Modulated Differential Scanning Calorimetry. *J. Chem. Eng. Data* **2020**, *65*, 591–597. https://doi.org/10.1021/acs.jced.9b00789

[6] Diky, V.; Bazyleva, A.; Paulechka, E.; Magee, J. W.; Martinez, V.; Riccardi, D.; Kroenlein, K. Validation of thermophysical data for scientific and engineering applications. *J. Chem. Thermodyn*, **2019**, *133*, 208–222. https://doi.org/10.1016/j.jct.2019.01.029

[7] Frenkel, M.; Dong, Q.; Wilhoit, R. C.; Hall, K. R. TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations. *Int. J. Thermophys.* **2001**, *22*, 215–226. https://doi.org/10.1023/A:1006720022161

[8] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348. https://doi.org/10.1021/ci900450m

[9] Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In: Chemoinformatics in Drug Discovery, Oprea, T. I. (Ed.); Wiley-VCH, 2005, pp. 271-285. https://doi.org/10.1002/3527603743.ch11

[10] Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev*. **1969**, *69*, 279–324. https://doi.org/10.1021/cr60259a002

[11] Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Comm.* **1987**, 57, 233–243. https://doi.org/10.1080/00986448708960487

[12] Diky, V.; Muzny, C. D.; Kazakov, A.; Paulechka, E.; Lemmon, E. W.; Bazyleva, A.; Townsend, S.; Renken, T.; Smolyanitsky, A. Y.; Chirico, R. D.; Frenkel, M.; Magee, J. W.; Kroenlein, K. NIST ThermoData Engine, NIST Standard Reference Database 103b, version 10.4.5, National Institute of Standards and Technology, USA (2023) https://www.nist.gov/mml/acmd/trc/thermodata-engine/srd-nist-tde-103b

[13] Kang, J. W.; Diky, V.; Frenkel, M. New modified UNIFAC parameters using critically evaluated phase equilibrium data. *Fluid Phase Equilibria* **2015**, *388*, 128–141. http://dx.doi.org/10.1016/j.fluid.2014.12.042

[14] Waring, W. Form of a Wide-Range Vapor Pressure Equation. *Ind. Eng. Chem.* **1954**, *46*, 762–763. https://doi.org/10.1021/ie50532a042

[15] Diky, V. An Efficient Way of Visualization of Mutual Solubility Data in the Whole Range of Compositions. *J. Chem. Eng. Data* **2017**, *62*, 2920–2926. https://doi.org/10.1021/acs.jced.7b00174

[16] Ambrose. D.; Walton, J. Vapor-Pressures up to Their Critical-Temperatures of Normal Alkanes and 1-Alkanols. *Pure Appl. Chem.* **1989**, *61*, 1395–1403. https://doi.org/10.1351/pac198961081395