ISSN 2220-637X

Вісник
Харківського
Національного
Університету
імені В. Н. Каразіна

СЕРІЯ «ХІМІЯ»
Вип. 40 (63)

Kharkiv University Bulletin
Chemical series. Issue 40 (63)

Заснований 1935 року як “Труди інституту хемії при Харківському державному університеті”

Published since 1935; initially under the title "Proceedings of the Institute of Chemistry at Kharkiv State University"

Харків   Kharkiv
2023

Вісник містить статті, присвячені різним аспектам теоретичної хімії, хімічного аналізу, органічної хімії, спектроскопії, фізико-хімії розчинів та поверхневих явищ, електрохімії, хімічного матеріалознавства.

Для науковців і фахівців. Видання є фаховим в галузі хімічних наук. (Наказ Міністерства освіти і науки України № 1643 від 28.12.2019 року)

Затверджено до друку рішенням Вченої ради Харківського національного університету імені В.Н. Каразіна (протокол № 17 від 02 жовтня 2023 р.)

Адреса редакційної колегії: Україна, 61022, Харків, майдан Свободи, 4,
ХНУ імені В.Н. Каразіна, хімічний факультет; тел.: +38 057 707 51 29.
E-mail: chembull@karazin.ua a.korobov@karazin.ua

Статті пройшли внутрішнє та зовнішнє рецензування.

Свідоцтво про державну реєстрацію КВ № 21563-11463Р від 27.07.2015.

The Chemical Series publishes papers devoted to various aspects of theoretical chemistry, chemical analysis, organic chemistry, inorganic chemistry, physical chemistry of solutions and surface phenomena, electrochemistry, materials chemistry. The bulletin is officially authorized by the Highest Attestation Commission of Ukraine to publish results of research submitted for PhD and ScD degrees.

Publication of this issue is approved by the Academic Council of V.N. Karazin Kharkiv National University (protocol № 17 from 02.10.2023).

Editor-in-chief

Alexander Korobov	ScD, Prof., V.N. Karazin Kharkiv National University

Editors

Andrey Doroshenko	ScD, Prof., V.N. Karazin Kharkiv National University
Nikolay Mchedlov-Petrossyan	ScD, Prof., V.N. Karazin Kharkiv National University, corresponding member of NAS of Ukraine

Managing editor

Anton Zakharov	PhD, V.N. Karazin Kharkiv National University

Editorial board

Vladimir Ivanov	ScD, Prof., V.N. Karazin Kharkiv National University
Vladislav Cheranovsky	ScD, Prof., V.N. Karazin Kharkiv National University
Sergey Shapovalov	ScD, Prof., V.N. Karazin Kharkiv National University
Oleg Yurchenko	ScD, Prof., V.N. Karazin Kharkiv National University

International advisory board

Pál Jedlovszky	Dr. Prof., Eszterházy Károly University, Hungary
Abdenacer Idrissi	Prof., Universite Lille, France
Oleg Kalugin	PhD, Prof., V.N. Karazin Kharkiv National University
Alexander Nazarenko	PhD, Prof., Buffalo State College, USA
Oleg Prezhdo	PhD, Prof., University of Southern California, USA
Valentyn Chebanov	ScD, Prof., SSI "Institute for Single Crystals", Ukraine corresponding member of NAS of Ukraine

Consulting editors

Ivan Vyunnik	ScD, Prof., V.N. Karazin Kharkiv National University
Vasyl Larin	ScD, Prof., V.N. Karazin Kharkiv National University
Valentyn Lebed	ScD, Prof., V.N. Karazin Kharkiv National University

Address of editorial team: 4 Svobody Sq., Kharkiv, 61022, Ukraine, V.N. Karazin
Kharkiv National University, School of Chemistry; tel.: +38 057 707 51 29.
E-mail: chembull@karazin.ua a.korobov@karazin.ua

All articles have been reviewed.

Certificate of state registration KB № 21563-11463P from 27.07.2015.

© V.N. Karazin Kharkiv National University,
Design, 2023

ЗМІСТ

CONTENTS

Вісник Харківського національного університету, серія "Хімія", вип. 40 (63), 2023
https://doi.org/10.26565/2220-637X-2023-40-01 ISSN 2220-637X
УДК: 539.192

ENERGY SPECTRUM AND MAGNETIC PROPERTIES OF THE DECORATED SPIN LADDER MODELS OF NANOMAGNETS ON THE BASE OF POLYMERIC TRANSITION METAL COMPOUNDS

V.O. Cheranovskii[a], V.V. Mukhomodiarova[b]

V. N. Karazin Kharkiv National University, 4 Svobody sq., Kharkiv, 61022, Ukraine

a) ✉ *cheranovskii@karazin.ua* *https://orcid.org/0000-0002-2255-400X*
b) ✉ *vicky123886@gmail.com* *https://orcid.org/0009-0007-8802-4128*

The work is devoted to the theoretical study of the energy spectrum and low-temperature magnetic properties of the decorated spin-ladder model with the polyacene topology and the three types of the site spins. On the base of cluster expansion technique an approximate analytical treatment of lowest part of the energy spectra of two isomeric ladder structures was given. It is shown that the ladder model with singlet ground state is more stable than its isomeric analog with the macroscopic ground state spin. In addition, the numerical study of field dependence of low-temperature magnetization of 8- spin clusters of both ladder models was performed by means of exact diagonalization method. On the base of these results, it was shown the presence of an intermediate plateau in low-temperature magnetization profile of the above spin ladder models.

Keywords: *mixed spin ladder model, intermediate magnetization plateau.*

Introduction

Low-dimensional quantum spin systems have become the subject of immense interest during the last few years, since their properties are strongly affected by quantum fluctuations. The design of highly ordered systems of paramagnetic metal centers is a current subject with the aim of providing magnetic materials exhibiting spontaneous magnetization. The most known mixed spin systems correspond to rather big family of so-called bimetallic magnets like $NiCu(pbaOH)(H_2O)_3{\cdot}2H_2O$, where pba is 1,3-propylenebis(oxamato) [1]. There are also trimetalic polymeric systems like Prussian blue analog $(Ni_x\ Mn_{1-x})_{1.5}[Cr(CN)_6]\ nH_2O$ ($x{\sim}0.4$). The exchange interaction of localized spin moments of metal centers in polymeric complexes is mediated by ligands [1-3]. Hence, the chemical modification of the ligands effects on the magnetic properties of these complexes. It open the way for targeted design of new magnetic materials for different technological applications. On the other side, numerous experimental realizations of mixed spin compounds motivate their study by means of different methods of solid state physics and quantum chemistry.

In this study the main attention will be given to the consideration of two isomeric decorated spin ladder systems formed by three different site spins s=1/2, 1 and 3/2. These spin systems have the topology of polyacene lattice described by effective Heisenberg spin Hamiltonians – spin formalism of well- known Valence Bond method of quantum chemistry. The exact energy spectra of these systems are unknown. For approximate evaluation of the lowest part of the energy spectra of our ladder models with the infinite number of unit cells we use simple two block cluster expansion technique [4]. In addition, in order to get information about the magnetic properties of the above spin systems, we use the exact diagonalization study for the Hamiltonians of small lattice clusters at different values of spin coupling parameters.

Heisenberg spin lattice model with three different types of the site spins

Let us start our consideration with two simple lattice systems formed by three different site spins s=1/2, 1 and 3/2 with uniform antiferromagnetic coupling of neighboring spins (Fig.1)

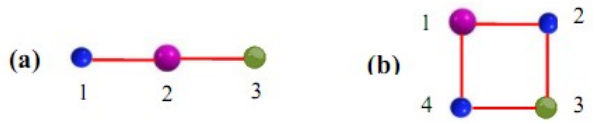


Figure 1. Three and four spin systems formed by three different types of site spins. Here crimson, green and blue balls correspond to the site spins *s*=3/2, 1 and 1/2, respectively.

In nearest neighbor approximation these two systems can be described by the following Heisenberg spin Hamiltonians with coupling constant ***J***=1:

$$\mathbf{H}_a = \mathbf{S}_1\mathbf{S}_2 + \mathbf{S}_2\mathbf{S}_3, \qquad \mathbf{H}_b = \mathbf{S}_1\mathbf{S}_2 + \mathbf{S}_2\mathbf{S}_3 + \mathbf{S}_3\mathbf{S}_4 + \mathbf{S}_1\mathbf{S}_4 \tag{1}$$

These Hamiltonians can be rewritten in the scalar product of two new spin operators with obvious energy spectra:

$$\mathbf{H}_a = \mathbf{S}_2\mathbf{S}_A, \quad \mathbf{S}_A = (\mathbf{S}_1 + \mathbf{S}_3), \qquad \mathbf{H}_b = \mathbf{S}_A\mathbf{S}_B, \quad \mathbf{S}_A = (\mathbf{S}_1 + \mathbf{S}_3), \mathbf{S}_B = (\mathbf{S}_1 + \mathbf{S}_3)$$

It can be easily shown that the ground state of $\mathbf{H}_a$ is singlet (nonmagnetic) with the energy E_0=-15/4. For the Hamiltonian $\mathbf{H}_b$ the ground state spin is S_0=3/2 and the corresponding energy is E_0=-7/2. These results are in agreement with the extended Lieb theorem [5-7] for bipartite spin lattice systems.

Let us now consider more complicated two-leg spin-1/2 ladder model with the legs decorated by two different site-spins *s*=1 and *s*=3/2, as it is shown below on Fig.2.

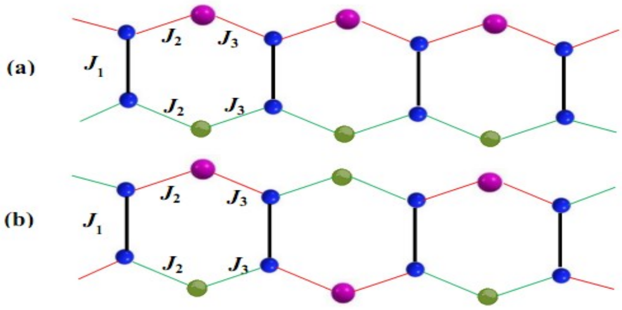


Figure 2. Two isomeric mixed spin ladder systems.

The above two spin ladder systems are described by Heisenberg spin Hamiltonian with antiferromagnetic coupling of neighbor site spins (J_1 - J_3>0). For example, the Hamiltonian of the system 2(a) with periodic boundaries has the form

$$\mathbf{H}_a = \sum_{l=1}^{L}\left[J_1\mathbf{S}_{2,l}\mathbf{S}_{3,l} + \mathbf{S}_{1,l}\left(J_2\mathbf{S}_{2,l} + J_3\mathbf{S}_{2,l+1}\right) + \mathbf{S}_{4,l}\left(J_2\mathbf{S}_{3,l} + J_3\mathbf{S}_{3,l+1}\right)\right] \tag{2}$$

where *L* is the total number of 4-spin unit cells and all spins are enumerated along the unit cells; $\mathbf{S}_{2,i}$-is the spin-1/2 operator, located on *l*-th unit cell of the ladder; $\mathbf{S}_{1,i}$ and $\mathbf{S}_{3,i}$ - are the operators of spin *s*=3/2 and *s*=1, respectively.

The spin ladder systems 2(a) and 2(b) have bipartite symmetry and, according to the generalized Lieb theorem, the corresponding Heisenberg Hamiltonian have nondegenerate ground state with total spin S_0=*L*/2 for model 2(a), and S_0=0 for model 2(b) with even number of 4-spin structural units (unit cells in the case of model 1(a)). In addition, according to [7, 8] the ground state of the model 2(a) should have ferrimagnetic spin ordering. Note also, that for the model 1(a) similar to the perturbative treatment from [9, 10], we may expect an appearance of intermediate plateau in field dependence of magnetization at least in the case of the weak interaction between unit cells(J_3<<J_2).

In order to get more information about the lowest energy states of the above spin ladder models 2(a) and 2(b), we performed numerical calculations of the exact energy spectra of finite ladder clusters formed by 8 site spins (two unit cells for model 2(a)) at some values of model parameters. For this purpose we used basis of spin configurations (3) having the form of direct products of eigenfunctions of the site spins. It can be easily shown, that these functions are the eigenfunctions of the operator of z-projection of the ladder total spin M:

$$\Phi_{\{s,m\}}\left(M\right)=\prod_{k=1}^{N}\Omega\left(s_k,m_k\right) \tag{3}$$

where N is a total number of site spins; multiindex $\{s, m\}$ enumerates all possible combinations of site quantum numbers (s_k, m_k). $\Omega(s_k, m_k)$ is the eigenfunction of the site spin operator $\mathbf{S}_k^z$ with the specified values of spin s and its z-projection m.

For example:

$$\Omega\left(1/2,1/2\right)=\alpha \qquad \Omega\left(3/2,1/2\right)=\frac{1}{\sqrt{3}}\left(\alpha\alpha\beta+\alpha\beta\alpha+\beta\alpha\alpha\right)$$

where for site spin s=1/2 $\mathbf{S}_k^z\alpha=1/2\alpha$, $\mathbf{S}_k^Z\beta=-1/2\beta$.

The exact diagonalization study was performed for all fixed values of quantum number M separately. Due to the scalar character of the ladder Hamiltonians, the comparison of the energy levels for each subspace with specified value of M permits us to obtain the lowest energy levels with fixed value of total spin S. We also used the result of our analytical consideration of the three-spin clusters (Fig.1) for testing of the above numerical scheme.

Some results of the exact diagonalization study of the ladder clusters for model 2(a) with J_1=J_2=1 are presented on Table 1.

Table 1. The lowest energies of the 8-spin clusters of the ladder models 2(a).

	S=0	S=1	S=2	S=3	S=4
J_3=1	-5.385	-5.959	-5.854	-5.664	-4.681
J_3=0.5	-4.889	-5.238	-5.130	-4.943	-3.911
J_3=0.1	-4.781	-4.856	-4.753	-4.617	-3.491

According this study the ground state of the cluster 2(a) corresponds to the total spin S=1 in agreement with the extended Lieb theorem. Let $E_{\min}(1)$ and $E_{\min}(0)$ are the corresponding exact energies of the triplet ground state and lowest singlet excited state. Similar to [9, 10], we can suppose that these two lowest energy states can be described by the following effective spin s=1/2 Hamiltonian:

$$\mathbf{H}_2=J_{eff}\mathbf{S}_1\mathbf{S}_2+R+2\varepsilon_0 \tag{4}$$

where ε_0 is the ground state energy of the 4-spin unit cell. The parameters J_{eff} and R can be estimated from the obvious system of linear equations

$$\begin{cases} E_{\min}\left(1\right)=2\varepsilon_0+R+J_{eff}/4 \\ E_{\min}\left(0\right)=2\varepsilon_0+R-3J_{eff}/4 \end{cases} \tag{5}$$

In the result, we obtain:

$$J_{eff}=E_{min}\left(1\right)-E_{min}\left(0\right)<0\ ,\quad R=(3E_{min}\left(1\right)+E_{min}\left(0\right)-8\varepsilon_0)/4 \tag{6}$$

The results of the corresponding exact diagonalization study for 8 –spin cluster of the ladder model 2(b) at J_1=J_2=1 are presented below on Table 2.

According to these results, J_{eff} = $E_{\min}(1)$ - $E_{\min}(0)$ > 0, which is in accordance with the extended Lieb theorem. Similar to the above analysis it can be shown that the lowest part of the energy spectrum of ladder model 2(b) can be described by the effective spin s=1/2 Hamiltonian of the form (6) with antiferromagnetic coupling.

This means that the mixed spin ladder model 2(a) and 2(b) should have gapless energy spectrum in the thermodynamic limit $L\rightarrow\infty$ and different types of the ground states with S=L and S=0, respectively.

Table 2. The lowest energies of the 8-spin clusters of the ladder models 2(b)

	S=0	S=1	S=2	S=3	S=4
$\boldsymbol{J_3}$=1	-5.986	-5.938	-5.833	-5.645	-4.640
$\boldsymbol{J_3}$=0.5	-5.259	-5.211	-5.106	-4.925	-3.873
$\boldsymbol{J_3}$=0.1	-4.867	-4.836	-4.742	-4.609	-3.478

It is of interest, that the ground state energy of the triplet cluster (ladder 2(a)) is bigger than the ground state energy of singlet cluster (ladder 2(b)) at all the values of parameter $\boldsymbol{J_3}$ studied. This result may be treated as a bigger chemical stability of the singlet ladder cluster in comparison with isomeric triplet cluster of our mixed spin ladder. Moreover we can estimate relative stability of the infinite spin ladder systems 2(a) and 2(b) using formulas (5) and (6). For this purpose let us estimate the difference in the corresponding ground state energies E_0^a and E_0^b per 4-spin structural unit. After simple manipulation we have:

$$\Delta E = E_0^a - E_0^b = E_{\min}^a(1) - E_{\min}^b(1) + \left(E_{\min}^b(1) - E_{\min}^b(0)\right)\ln 2 \qquad (7)$$

where $E_{\min}^a(1)$, $E_{\min}^b(1)$, $E_{\min}^b(0)$ are the lowest triplet and singlet energies of 8-spin clusters of the spin ladders 2(a) and 2(b), respectively. These energies are presented in Table1 and Table2. In the result, for example, for $J_1=J_2=J_3=1$ $\Delta E\sim 0.0123>0$. This means that for this set of coupling parameters the infinite spin ladder (2b) is more stable than the isomeric ladder structure (2a).

We also used exact diagonalization approach for the study of low-temperature magnetization profiles of 8- spin clusters of the ladder models 2(a) and 2(b). For this purpose we used simplified version of the models with equal *g*-factors for each type of site-spins. In the result, for each energy state of the ladder, the projection of total spin on the direction of the external magnetic field is a good quantum number. Therefore, external magnetic field change the energy of the state with specified value of z-projection of total spin M by the quantity $\Delta E=-hM$, where h is a strength of magnetic field in energy units. Using standard Boltzmann distribution we have the following expression for ladder magnetization per 4-spin structural unit:

$$\bar{M}(h,T) = \sum_i M_i \exp(-(E_i - hM_i)/k_BT) / \sum_i \exp(-(E_i - hM_i)/k_BT) \qquad (8)$$

where k_B –is a Boltzmann constant.

The results of this numerical simulation for model 2(a) are presented below (Fig.3).

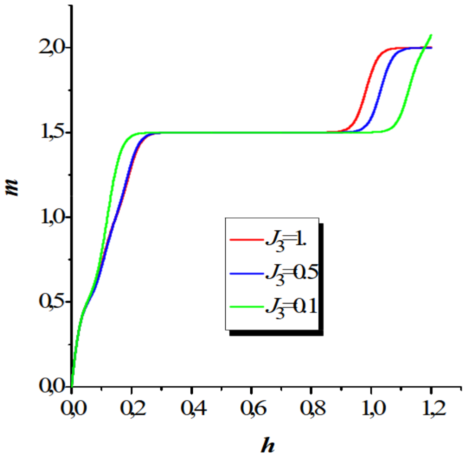


Figure 3. Field dependence of magnetization per 4 spin unit cell m of the cluster 2(a) at $J_1=J_2=1$, k_BT=0.02.

According to these calculations the low temperature magnetization profile of 8-spin cluster of model 2(a) has intermediate magnetization plateau at m=1.5. The appearance of this plateau can be explained by relatively big energy gap between states with S=3 and lowest state with S=4. The size of this plateau is increased with the decrease of the interaction between unit cells (coupling constant J_3). We have also small peculiarity in behavior of the magnetization profile near the point m=0.5 which can be explained by the triplet ground state of the ladder cluster studied.

Finally, in the result of our approximate analytical consideration and numerical simulation we may suppose that the ladder model 2(a) has intermediate plateau in low-temperature magnetization profile at arbitrary values of coupling parameter J_3 from the interval (0, J_2).

The results of similar numerical study of magnetization profile of the ladder cluster of the model 2(b) are presented below on Fig.4.

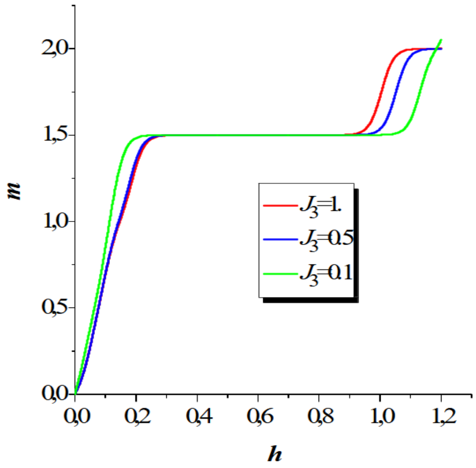


Figure 4. Field dependence of magnetization m per 4 spin structural unit of the cluster 2(b) at J_1= J_2=1, k_BT=0.1.

According to this simulation, the magnetization profile of the ladder cluster 2(b) has also intermediate magnetization plateau at m=1.5. In contrast to the cluster 2(a) there are not magnetization peculiarities at m=0.5. On the other side, according to our two block cluster expansion analyses, the initial part of the low-temperature magnetization profile of infinite ladder model 2(b) should be close to the magnetization profile of antiferromagnetic Heisenberg spin-1/2 chain without magnetization plateau.

Conclusions

The exact diagonalization study of the energy spectra and low –temperature magnetic properties of the two isomeric mixed tree-spin ladder models of polyacene topology was performed for 8-spin lattice clusters. On the base of this study we derived simple two block cluster expansion technique and showed that that the ladder model with the singlet ground state is more stable than its isomeric analog with the macroscopic ground state spin. We also found the presence of an intermediate plateaus in the low temperature magnetization profiles of the above spin ladder models.

Список літератури

1. Koningsbruggen van, P. J., Kahn O., Nakatani K., Pei Yu, Renard J. P., Drillon M., and Legoll P.. Magnetism of ACu^{II} Bimetallic Chain Compounds (A = Fe, Co, Ni): One- and Three-Dimensional Behaviors. Inorg.Chem. 1990, 29, 3325-3331. https://doi.org/10.1021/ic00343a014.
2. Hao Xy., et al. Cyanide-bridged polynuclear and one-dimensional Fe-III-Mn-III/II bimetallic complexes based-on pentacyanoferrite (III) building block: synthesis, crystal structures, and magnetic properties. *Acta Chim. Slov.* **2020**, 67, 916-926. https://doi.org/10.17344/acsi.2020.5909.
3. Egan L., et.al. Pressure-induced sequential magnetic pole inversion and antiferromagnetic-ferromagnetic crossover in a trimetallic Prussian blue analoge. JACS communications **2006**, 128, 6034-8035. https://doi.org/10.1021/ja061514m.
4. Cheranovskii V.O.,.Ezerskaya E.V, Klein D.J., Kravchenko A. A.. Magnetic properties of model non-carbon nanotubes with macroscopic value of ground state spin. *J. Magn. Magn. Mater.* **2011**, *323*, 1636–1642. https://doi.org/10.1016/j.jmmm.2011.01.027.
5. Lieb E.H., Mattis D.C.. Ordering energy levels of interacting spin systems. *J. Math. Phys.* **1962**, *3*, 749-751. https://doi.org/10.1063/1.1724276.

6. Ovchinnikov A.A. Multiplicity of the ground state of large alternant organic molecules with conjugated bonds - (Do Organic Ferromagnetics Exist? *Theor. Chim. Acta*. **1978**, *47*, 297-304. https://doi.org/10.1007/BF00549259
7. Klein D.J. Ground state features for Heisenberg models. *Chem. Phys.* **1982**, 77, 3098-3100. https://doi.org/10.1063/1.444232
8. Ovchinnikov A.A., Cheranovskii V.O. Perturbation theory in the spin Hamiltonian method. *Theor. Experim. Chemistry*, **1980**, 16, 119-124. https://doi.org/10.1007/BF00515986.
9. Cheranovskii V.O., Slavin V.V. Ezerskaya E.V. Effective low-energy spin model for narrow zigzag graphene nanoribbons. *Low. Temp.Phys*. **2021**, 46, 683-687. https://doi.org/10.1063/10.0001364
10. Cheranovskii V.O., Ezerskaya E.V., Kabatova A.O. The energy spectrum and low-temperature magnetic properties of the decorated two-leg mixed spin ladder. *Low. Temp.Phys*. **2023**, 49, 438-442. https://doi.org/10.1063/10.0017584.

Received 21.04.2023 *Accepted 11.09.2023*

В.О. Черановський, В.В. Мухомодярова. Енергетичний спектр і магнітні властивості декорованих спінових сходових моделей наномагнетиків на основі полімерних сполук перехідних металів.

Харківський національний університет імені В.Н. Каразіна, хімічний факультет, майдан Свободи, 4, Харків, 61022, Україна

Робота присвячена теоретичному дослідженню енергетичного спектру та низькотемпературних магнітних властивостей декорованої спін-сходової моделі з поліаценовою топологією та трьома типами вузлових спінів. На основі методу кластерного розширення проведено наближену аналітичну обробку нижньої частини енергетичних спектрів двох ізомерних сходових структур. Показано, що сходова модель із синглетним основним станом більш стабільна, ніж її ізомерний аналог із макроскопічним спіном основного стану. Крім того, методом точної діагоналізації проведено чисельне дослідження польової залежності низькотемпературної намагніченості 8-спінових кластерів обох сходових моделей. На основі цих результатів було показано наявність проміжного плато в низькотемпературному профілі намагніченості наведених вище моделей спінових сходів.

Ключові слова: змішана спінова сходова модель, проміжне плато намагніченості.

Надіслано до редакції 21.04.2023 *Прийнято до друку 11.09.2023*

Kharkiv University Bulletin. Chemical Series. Issue 40 (63), 2023

https://doi.org/10.26565/2220-637X-2023-40-02 ISSN 2220-637X
УДК: 544.169+519.237.5

INTERNAL VALIDATION PARAMETERS OF LINEAR REGRESSION EQUATIONS IN QSAR PROBLEM

I. V. Khristenko[a], V. V. Ivanov[b]

V. N. Karazin Kharkiv National University, School of Chemistry, 4 Svobody sqr., Kharkiv, 61022 Ukraine

a) ✉ khristenko@karazin.ua https://orcid.org/0000-0001-7227-8333
b) ✉ vivanov@karazin.ua https://orcid.org/0000-0003-2297-9048

The article discusses a set of internal validation parameters that are (or can be) used to describe the quality of regression models in quantitative structure-activity relationship problems. Among these parameters there are well known determination coefficient, root mean square deviation, mean absolute error, etc. Also the indices based at Kullback-Leibler divergence as a measure of distance between two sets have been investigated. All the parameters (indices) were calculated for several regression models which describe boiling point of saturated hydrocarbons (alkanes). Regression models include a four-component additive scheme and equations describing the property as a function of topological indices. The two types of regressions based on these indices are linear dependencies on only one topological index and linear dependencies on topological index and the number of carbon atoms in the hydrocarbon. Various linear regression equations have been described with internal validation parameters that evaluate the quality of the equations from different perspectives. It is shown that a wide set of test parameters is not only an additional yet alternative description of regression models, but also provides the most complete description of the predictive characteristics and quality of the obtained regression model.

Keywords: *Quantitative Structure-Activity Relationships (QSAR), regression models, internal validation, topological descriptors*

Introduction

It is easier to calculate a regression equation than to prove its predictive ability. This sentence is especially true for QSAR (*Quantitative Structure-Activity Relationships*) linear regression models. Necessity of proper investigation of obtained equations has been recognized during the last years. It has been demonstrated that poorly validated regression equations can be misleading when evaluating molecular activity/property. Several important articles discuss typical situations and difficulties in description of the predictive ability of regression models. The provocative titles of the articles – “The importance of being earnest[1]…” [1] “Beware of q2!” [2], “Beware of R^2….” [3] call for attention to this problem. In the presented paper, we consider the problem of validating of QSAR regression equations from a somewhat specific point of view.

First of all, we note that, for common practice, QSAR studies involve dividing the primary data into two data sets. These sets are the *training set* that is used to generate the corresponding QSAR model, and the *test set* is the data for validation of the resulting models (equations). The parameters characterizing the description of the training set by the obtained equations are considered as internal validation, while the parameters characterizing the quality of the description of the test set are external validation. In recent years, significant attention has been paid to external validation, which can be considered as a model for the practical use of the obtained equations. Regarding the content of external validation, several important issues should be noted. The primary set must be divided in a certain ratio between the training and test sets. What is this ratio? How to specify the separation of systems (points) between two sets? How to prove the correctness of the division? And, in the end, will such a division lead to a decrease in the predictive ability of equations due to a decrease in the size of the training sample? So we see that external validation leads to additional questions for which there are no general answers yet (see, however, the article and references therein discussing this problem [4-6]). Hence

---

[1] *The quote from the famous play by Oscar Wilde emphasizes the main idea of the authors of the article – "first, validate, and then explore".*

such a procedure is not computationally well defined. In contrast to external validation, the internal validation does not need dividing the input data into two subsets.

Without denying the necessity for an external validation procedure, in the present paper we propose to take a closer look at the internal validation (goodness-of-fit) of QSAR regression. Usually the restricted set of parameters used for the internal validation. Among these, the most important are the determination coefficient (R^2) and the standard deviation (root mean square deviation, *RMSD*). Such parameters cannot be considered as those that give a complete description of the training sample and the corresponding regression equation. Also these parameters usually demonstrate a low sensitivity to variation of the model. An extremal example is the classic Anscombe paper, where the very different data unexpectedly fit the same equation, with the same R^2 and the same *RMSD* [7,8].

In this paper, we analyze a wide set of known internal validation parameters and a few new parameters that we have proposed. As an example we describe boiling points (BP, C°) of saturated hydrocarbons (alkanes). QSAR-models of these properties include additive scheme and graph theory approaches based on known topological indices. It should be noted that the interest to topological indices has been stable for a long time up to the present day. For instance, in the paper [9], new graph theory model for description of boiling points of alkanes is discussed. Also graph theory approaches currently used for description of anti-cancer activity [10] and even for description of potential anti-COVID-19 substances [11,12].

All the calculations were performed by using Python3 script language. RDKit package was used for manipulations with chemical structures and calculations of molecular descriptors. [13] The experimental data for BP of the alkanes were obtained from [14]. When information about the physicochemical properties of alkanes includes several values, we used the average values. In total, the training set contains information on 39 different saturated hydrocarbons with 1 to 9 carbon atoms.

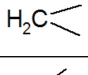
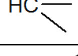
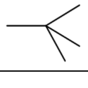
Linear regression models and internal validation parameters

For the physical-chemistry property of alkanes we consider three types of linear regression models. The first one is correspond to simple additive scheme.

$$Y = n_1x_1 + n_2x_2 + n_3x_3 + n_4x_4 \tag{1}$$

Where Y is the dependent variable – the physicochemical property of alkanes is a function of four parameters (n_1, n_2, n_3, n_4) that describe the molecular structure. The partial values (increments) x_1, x_2, x_3, x_4 are contributions from elements of the molecular structure (Table 1).

Table 1. Parametrization of additive scheme for alkane molecules

Fragment	Number of Fragments in the molecule	Increments
H_3C-	n_1	x_1
	n_2	x_2
	n_3	x_3
	n_4	x_4

Also we consider *two* regression models based on graph theory. The topological indices $X = \left\{\chi^{(1)}, ZM_1, ZM_2, ZM_2, IC_1, InfD\right\}$ (see Table 2) were used in the calculations as the molecular descriptors. For a detailed descriptions of the indexes presented in the Table 2, see for example [15,16].

The first graph theory based model is single-parameter equations:

$$Y = a_0 + a_1X \tag{2}$$

where X is the descriptor from Table 2. The second equation includes descriptor X and the number of carbon atoms (N_C) in the hydrocarbon:

$$Y = a_0 + a_1 N_C + a_2 X \tag{3}$$

The regression coefficients (a_0, a_1, a_2) as well as partial values for additive scheme (x_1, x_2, x_3, x_4) were obtained using the Ordinary Least Squares (OLS) method (see for example [17]).

Table 2. Topological indices used in the present article (v_i– order of vertex *i,* (*i*, *j*) are pairs of connected by edges carbon atoms)

№	Topological Index	Definition
1	First order Randich index	$\chi^{(1)} = \sum_{(i,j)} 1/\sqrt{v_i v_j}$
2	First Zagreb index	$ZM_1 = \sum_i v_i^2$
3	Second Zagreb index	$ZM_2 = \sum_{(i,j)} v_i v_j$
4	Third Zagreb index (so called “forgotten index”)	$ZM_3 = \sum_i v_i^3$
5	First order informational content. n_k- number of vertices with definite *v*, $N = \sum_k n_k$	$IC_1 = -\sum_k \frac{n_k}{N} \log_2 \frac{n_k}{N}$
6	Informational index of distances in graph. r_k– number of routs with topological distances equal to *k*, $N_D = \sum_k r_k$	$InfD = -\sum_k \frac{r_k}{N_D} \log_2 \frac{r_k}{N_D}$

The residuals between given (experimental) values *Y* and those obtained by using regression equations (1-3) for training set (Y_i^{calc}) are calculated as follow:

$$e_i = Y_i - Y_i^{calc} \tag{4}$$

Also, the correspondence between the calculated and given values of the variable *Y* is usually described as a linear form which can be presented by two equivalent, but not identical, equations:

$$Y^{calc} = \beta_0 + \beta_1 Y \tag{5}$$

$$Y = \gamma_0 + \gamma_1 Y^{calc} \tag{6}$$

Of course, for the absolute (or “ideal”) correspondence between *Y* and Y^{calc} values, one can write ($\beta_0 = \gamma_0 = 0$, $\beta_1 = \gamma_1 = 1$):

$$Y = Y^{calc} \text{ and } Y^{calc} = Y \tag{7}$$

However, for the typical (realistic) situation of QSAR investigations, for the equation (5) one can write

$$Y^{calc} = (1 - R^2)\overline{Y} + R^2 Y \tag{8}$$

Where *R* is Pearson correlation coefficient, $\overline{Y} = \sum_i Y_i / n$ is mean value, and *n* is size of sample.

Further, according to known expression $\beta_1 \gamma_1 = R^2$ [18] for the eq. (6) we ***always*** have absolute correspondence in the sense of least square method ($\gamma_1 = 0$, $\gamma_1 = 1$).

$$Y = Y^{calc} \tag{9}$$

However, note that both equations (8) and (9) must be interpreted in the spirit of OLS and, of course, have the same coefficient of determination, R^2. A discussion of these issues can be found in [18, 19].

Hence, in the general case, the deviation of Y^{calc} from *Y* can be expressed in the following most formal way

$$\eta = F(Y, Y^{calc})\,. \tag{10}$$

Where the parameter η describes the quality of approximation for the selected regression model. Generally speaking, expression (10) implies the use of different metrics *F*.

The set of η – parameters calculated for the presented regression models based at (1-3) is presented in Table 3.

Table 3. Internal Validation parameters

№	**Parameter**	**Description**	**Best equation**
1	Root Mean Square Deviation, *RMSD*	$RMSD = \sqrt{\sum_i \frac{e_i^2}{n-p}}$	$RMSD \to 0$
2	Determination Coefficient, R^2. For the *LOO* procedure designated as Q^2	$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$	$R^2 \to 1$
3	Mean Absolute Error, *MAE*	$MAE = \frac{1}{n}\sum_i \lvert e_i \rvert$	$MAE \to 0$
4	Asymmetry of residuals, *Asymm*	$Asymm = \frac{1}{n}\sum_i e_i$	$Asymm \to 0$
5	Relative error of worst point, *WPt*	$WPt = \max_i \left\{ \lvert e_i / Y_i \rvert \right\}$	$WPt \to 0$
6	Kullback-Leibler divergence between *Y* and Y^{calc} distributions, $D_{KL}\left(Y \,\Vert\, Y^{calc}\right)$	$D_{KL}\left(Y \,\Vert\, Y^{calc}\right) = \sum_i \lvert \tilde{Y}_i \rvert \log_2 \frac{\lvert \tilde{Y}_i \rvert}{\lvert \tilde{Y}^{calc} \rvert}$ $\sum_i \lvert \tilde{Y}_i \rvert = 1$, $\sum_i \left\lvert \tilde{Y}_i^{calc} \right\rvert = 1$	$D_{KL}\left(Y \,\Vert\, Y^{calc}\right) \to 0$
7	Inhomogeneity of Residuals, *IhR*	$IhR = -\log_2 n - \frac{1}{n}\sum_i \log_2 \lvert e_i \rvert / \sum_k \lvert e_k \rvert$	$IhR \to 0$
8	Angle between ideal (7) and obtained lines (5, 8), $\Delta\varphi$	$\Delta\varphi = \arctan\left(\beta_1\right) - 0.7854$	$\Delta\varphi \to 0$

Here one can see several standard parameters. Among them the *RMSD* (p is number of regression coefficients) and determination coefficient – R^2. We have also included a parameter based at absolute values of error (*MAE)*. The advantages and disadvantages of parameters based at absolute values of error are discussed in details in [20-22]. Also we found a few simple parameters to be useful. *Asymm* is a measure of the under- or overestimation of the dependent variable *Y*. The parameters *WPt* is simple values that estimate the spread of the residual vector or can be treated as relative outlier of point.

We are using also parameters based at Kullback-Leibler informational theory [23-24]. The Kullback-Leibler divergence $D_{KL}\left(Y \,\Vert\, Z\right)$, also known as the relative entropy, in specific way describes how *Z* distribution differ from actual distribution *Y*. There are several interpretations of $D_{KL}\left(Y \,\Vert\, Z\right)$. One of them is designated as "informational lost" when *Z* used instead of *Y* (or *Z* approximates *Y*). The main properties of D_{KL} are: $D_{KL}\left(Y \,\Vert\, Y\right) = D_{KL}\left(Z \,\Vert\, Z\right) = 0$, and $D_{KL}\left(Y \,\Vert\, Z\right) \neq D_{KL}\left(Z \,\Vert\, Y\right)$. We use two indices based at Kullback-Leibler divergence. The first one is to describe divergence of *Y* from Y^{calc} $D_{KL}\left(Y \,\Vert\, Y^{calc}\right)$, and the second – to describe inhomogeneity of the residuals *IhR* (see Table 3). As an example of the use of the Kullback-Leibler informational theory in chemistry see [25].

And also we are calculating the angle ($\Delta\varphi$) between "ideal line" (7) and line that is result of the regression analysis (8). All the above mentioned parameters were calculated in two variants. The first one corresponds to the calculation of regression parameters for the full sample, and the second to the leave-one-out cross-validation procedure (*LOO*) [26-29].

Results of calculations and discussion

As a result of the OLS regression calculation for the additive scheme (Eq. 1, Table 1), the corresponding plot of the "theory-experiment" relationship is presented in Fig. 1. The red line is corresponding to "ideal" dependence (7). Here one can see, that dependence of Y^{calc} from experimental value $Y = Y^{experim}$, designated as green circle, too far from "ideal" dependence for additive scheme.

Among the results obtained with the simple one-parameteric equation (2), much better solutions can be found. Especially the equation with Randich index ($\chi^{(1)}$) demonstrated the best result.

$$BP(C^{\circ}) = -141.6 + 71.026\chi^{(1)}. \tag{11}$$

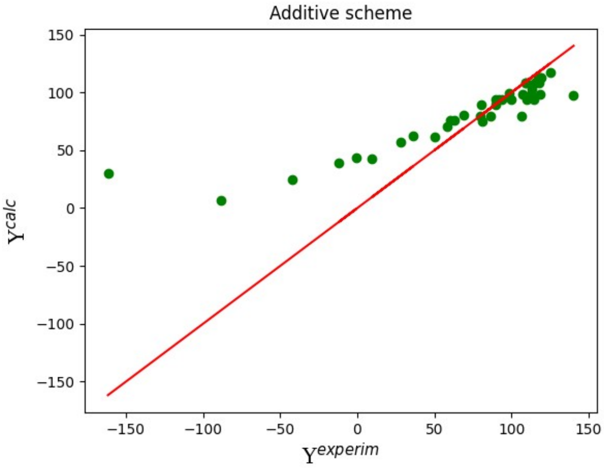
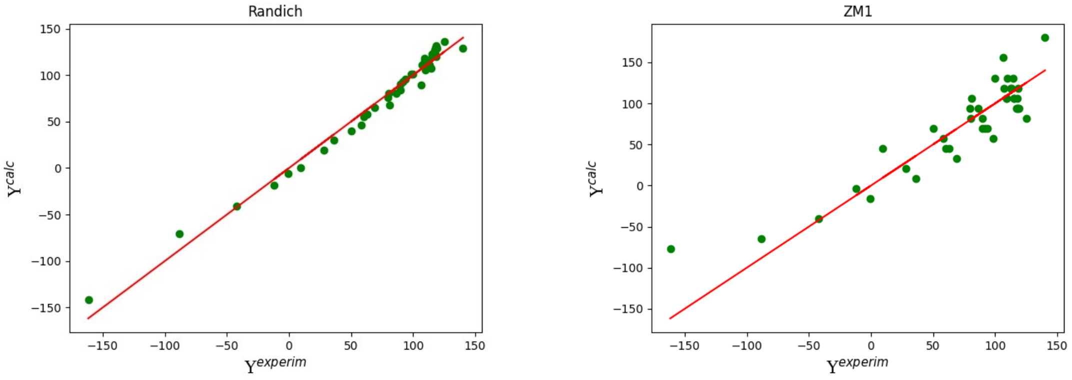


Figure 1. Dependence “theory-experiment” according to the additive scheme.

The internal validation parameters of the linear regressions obtained using the additive approach and those that follow eq. 2 are collected in Table 4. According to the obtained data, the equation based at $\chi^{(1)}$ is the best equation for all the parameters presented. Nevertheless, the choice of the following equations (in quality) depends on the chosen validation parameter. According to R^2 (and Q^2, MAE, $\Delta\varphi$) the next best equation is function from *ZM1*. The worst result was demonstrated by the *ZM3* index ($R^2 = 0.5265$ and a poor correspondence to the distribution of experimental data $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.5624$) with abnormal sensitivity to selected groups of molecules even compared to *InfD*. For the *InfD* the value $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.1445$ is significantly better than for the *ZM3* $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.3622$. It is also interesting that the average inhomogeneity of the residuals (*IhR*) is quite small for all indices (equations), even though there are large differences in *WPt* (relative outliers). However, *IhR* and *LOO IhR* for the additive scheme is noticeably greater than for all other regressions. Significant values of $\Delta\varphi$ and *LOO* $\Delta\varphi$ for additive approach are indicators of large difference between “ideal line” and actual.

For the two-parametric equations (3), which also include the number of carbon atoms in the molecule, the pictures are more optimistic. The quality of the equations in terms of validation parameters is much better (Fig. 3 and Table 5). Formally the best equation is:

$$BP(C^{\circ}) = -168.6 + 27.8Nc + 44.3InfD. \tag{12}$$



A

B

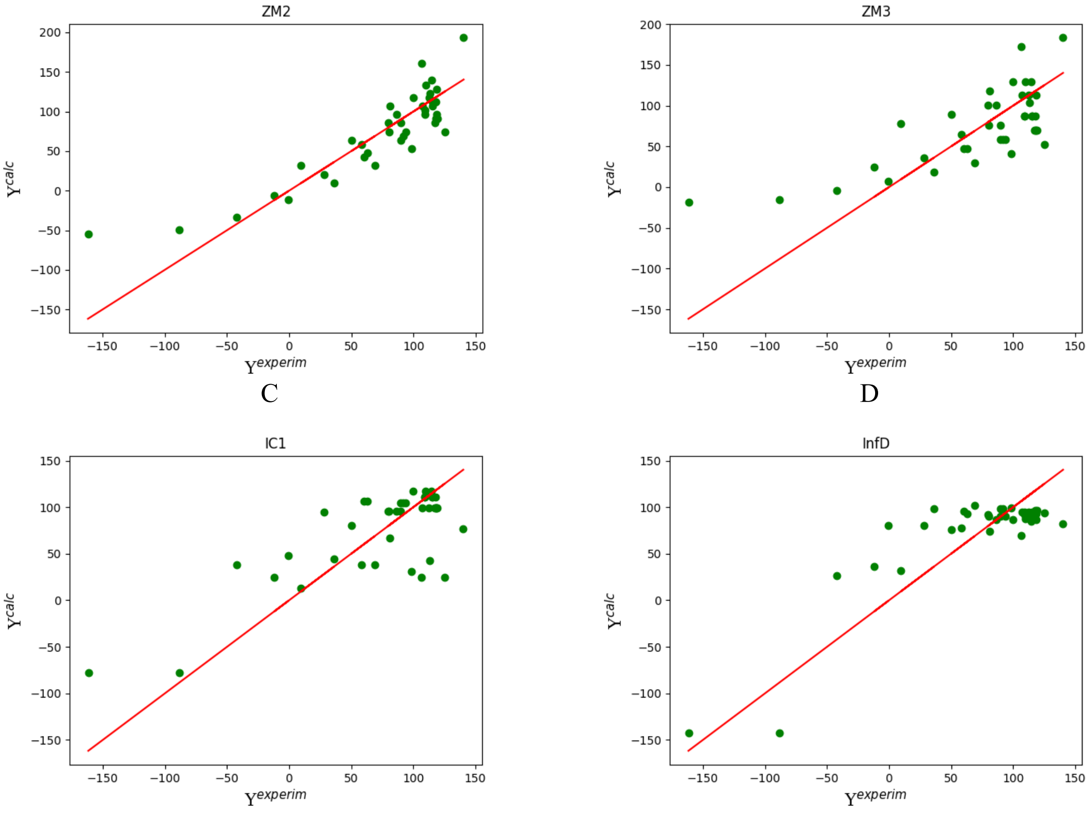


Figure 2. The QSAR estimations according to eq. 2 versus experimental data.

The equation is characterized by a large value of R^2, a smaller value of $D_{KL}\left(Y \,||\, Y^{calc}\right)$ and values of the other indices except inhomogeneity parameters *IhR* and *LOO IhR*. The *IhR* and *LOO IhR* values for *InfD* are only slightly worse than the corresponding values for other equations based on topological indices. The parameter *WPt*, characterizing the maximal relative outlier for (12), is much smaller than for the other equations. However, it can be noted visually that the best equation (12) is characterized by typical “steps” on the graph of the “calculation-experiment” dependence (Fig. 3, F). This indicates poor recognizability of some groups of molecules by regression model. This is in spite of high values of R^2, and low values of $D_{KL}\left(Y \,||\, Y^{calc}\right)$.

Table 4. The internal validation parameters for additive scheme (1) and regression equations (2).

	additive	$\chi^{(1)}$	ZM_1	ZM_2	ZM_3	IC_1	*InfD*
RMSD	41.2	8.7	26.2	29.7	43.1	40.0	33.0
LOO RMSD	61.1	9.8	29.2	33.1	47.0	43.5	36.7
R^2	0.5885	0.9808	0.8245	0.7740	0.5265	0.5914	0.7216
$Q^2 = LOO\ R^2$	0.0946	0.9754	0.7827	0.7201	0.4369	0.5169	0.6562
MAE	20.8	7.0	19.8	21.2	32.9	28.2	26.2
LOO MAE	26.1	7.6	21.3	22.9	35.1	30.1	28.4
Asymm	-0.93	10^{-14}	10^{-14}	10^{-13}	10^{-15}	10^{-14}	10^{-14}
LOO Asymm	-12.9	-0.03	-0.7	-0.9	-1.20	-0.7	0.6
WPt	80.2	9.3	27.9	20.1	14.5	88.7	147.4
LOO WPt	87.3	10.0	30.6	22.0	15.7	91.6	151.1
$D_{KL}(Y \,\|\, Y^{calc})$	0.5074	0.0454	0.1912	0.2539	0.5624	0.3622	0.1445
$LOO\ D_{KL}(Y \,\|\, Y^{pred})$	0.4042	0.0650	0.2445	0.3244	0.8227	0.4575	0.1314
IhR	1.17	0.48	0.64	0.79	0.55	0.70	0.55
LOO IhR	1.34	0.52	0.67	0.81	0.57	0.72	0.59
$\Delta\varphi$	-23.1	-0.6	-5.5	-7.3	-17.2	-14.4	-9.2
$LOO\ \Delta\varphi$	-36.2	-1.1	-6.8	-8.6	-19.3	-17.3	-8.7

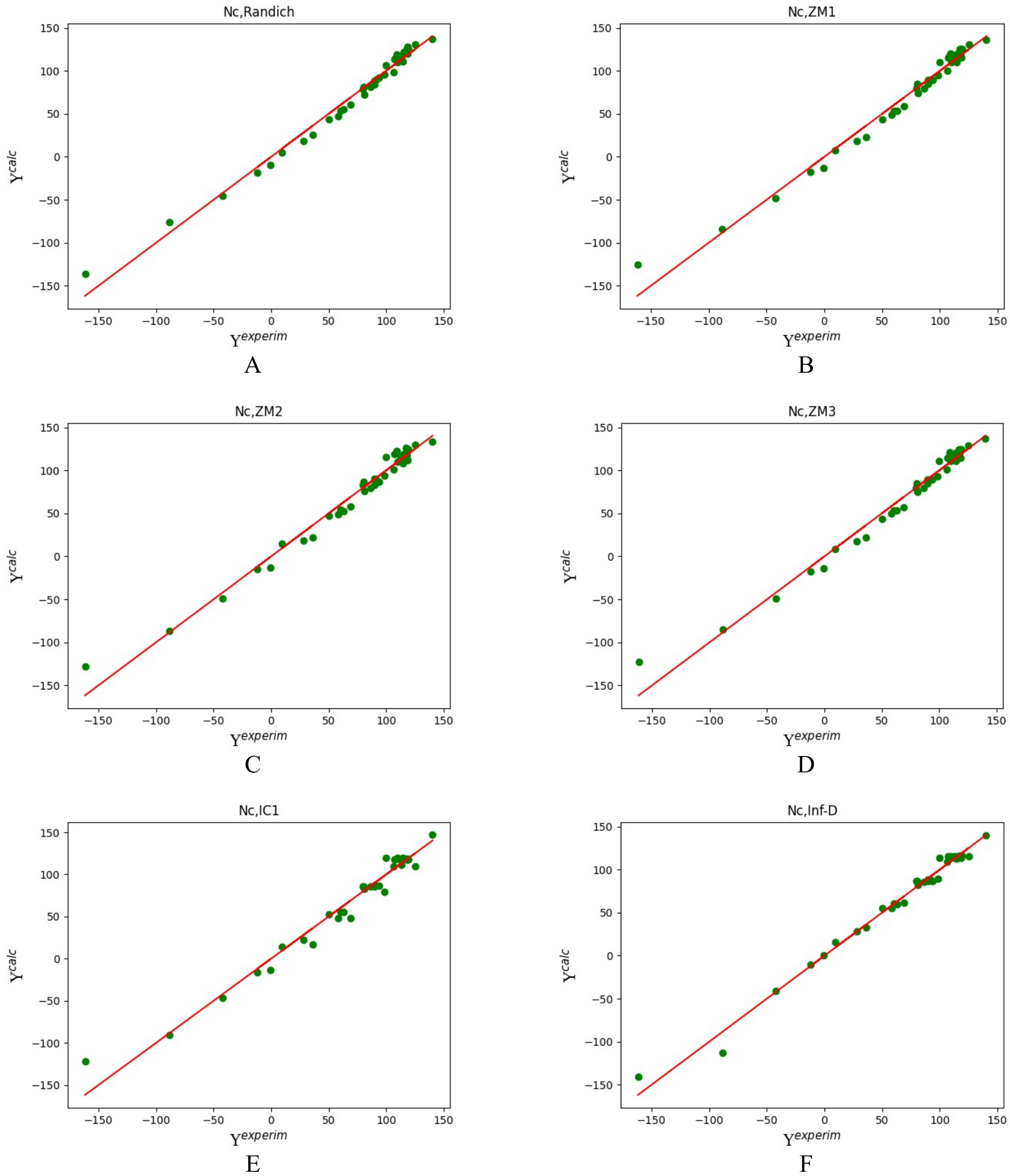


Figure 3. Data from QSAR regression (eq. 3) *vs* experimental data.

A similar by quality equation

$$BP(C^\circ) = -146.51 + 10.75Nc + 48.91\chi^{(1)}. \tag{13}$$

is visually free of this drawback (Fig. 3, A), but appears to have noticeable differences between estimated and theoretical distributions of the dependent variable ($D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.03$ and $LOO\ D_{KL}\left(Y \,||\, Y^{pred}\right) = 0.05$).

Table 5. Internal validation parameters for regression equations (3)

	$\chi^{(1)}$	ZM_1	ZM_2	ZM_3	IC_1	$InfD$
RMSD	7.9	9.2	9.4	9.6	11.0	7.2
LOO RMSD	10.1	11.4	11.6	11.8	13.3	11.0
R^2	0.9845	0.9791	0.9781	0.9772	0.9701	0.9871
$Q^2 = LOO\ R^2$	0.9747	0.9679	0.9668	0.9656	0.9554	0.9698
MAE	6.1	6.7	7.0	6.9	7.3	4.7
LOO MAE	7.0	7.5	7.8	7.7	8.3	5.8
Asymm	10^{-12}	10^{-12}	10^{-13}	10^{-13}	10^{-13}	10^{-13}
LOO Asymm	-0.39	-0.29	-0.28	-0.28	-0.25	0.011
WPt	17	22.2	23.5	24	24.3	2.4
LOO WPt	18.8	24.4	25.8	26.2	27.0	2.9
$D_{KL}(Y \parallel Y^{calc})$	0.0301	0.0299	0.0263	0.0310	0.0343	0.004
$LOO\ D_{KL}(Y \parallel Y^{pred})$	0.0531	0.0491	0.0461	0.0505	0.0555	0.004
IhR	0.52	0.51	0.59	0.46	0.83	0.98
LOO IhR	0.59	0.56	0.64	0.51	0.88	1.17
$\Delta\varphi$	-0.45	-0.61	-0.63	-0.66	-0.87	-0.37
$LOO\ \Delta\varphi$	-1.26	-1.29	-1.38	-1.33	-1.63	-0.47

Conclusion

To date, work on QSAR regression equation testing problems has shifted significantly toward external validation. However, the multitude of internal validation parameters is a useful tool for multilateral analysis of the resulting regression equations. In this paper, we have examined several internal validation parameters that are different in nature. It has been shown that such parameters can complement each other. In particular, parameters based on the Kullback-Leibler informational theory (indices $D_{KL}\left(Y \parallel Y^{calc}\right)$ and *IhR*) describe the correspondence of theoretical, based on the regression model, and experimental data, from a different perspective than the determination coefficient and other known parameters. Assessing the results of calculations, also it should be noted that the standard set of parameters is still not sufficient to identify the presence of "steps" in the "calculation-experiment" graph. Their influence on the quality of approximation still awaits quantitative assessment. At the same time, visual analysis of the "theory-experiment" graph remains important.

Note also that there is obviously no universal solution in choosing the best (accurate) regression model to describe the properties of the system in a situation where there is a large scatter in the initial data. In this case, it is difficult to avoid a wide range of parameters (standard deviation, coefficient of determination, etc.) when estimating multiple models.

Acknowledgment

This research was partly supported by the Ukrainian Minister of Education and Science. Grant № 0122U001485 "Design and optimization of functional nanodisperse systems: lyophilic aggregates, biocompatible ashes, hybrid materials, photoelectric converters".

References

1. Tropsha A., Gramatica P. and Gombar V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.* 2003, *22*, 69-77. https://doi.org/10.1002/qsar.200390007
2. Golbraikh A., Tropsha A. Beware of Q2! *Journal of Molecular Graphics and Modelling*. 2002, *20*, 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1
3. Alexander D. L. J., Tropsha A., and Winkler D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. of Chem. Inform. and Model.* 2015, *55*(7), 1316-1322. https://doi.org/10.1021/acs.jcim.5b00206
4. Joseph V. R. and Vakayil A. SPlit: An Optimal Method for Data Splitting. *Technometrics*. 2021, *64(2)*, 166-176. https://doi.org/10.1080/00401706.2021.1921037

5. Liu H., Cocea M. Semi-random partitioning of data into training and test sets in granular computing context. *Granul. Comput.* 2017, *2*, 357–386. https://doi.org/10.1007/s41066-017-0049-2
6. Joseph V. R. Optimal ratio for data splitting. *Stat. Anal. Data Min.: ASA Data Sci. J*. 2022, *15*, 531–538. https://doi.org/10.1002/sam.11583
7. Anscombe F. J. Graphs in Statistical Analysis. *Am. Stat.* 1973, *27*, 17-21. https://doi.org/10.2307/2682899
8. Besalu E., Julian-Ortiz J. V., Pogliani L. Trends and Plot Methods in MLR Studies. *J. Chem. Inf. Model.* 2007, *47,* 751-760. https://doi.org/10.1021/ci6004959
9. Mukwembi S., Nyabadza F. A new model for predicting boiling points of alkanes, *Scientific Reports*, 2021, *11*, 24261. https://doi.org/10.1038/s41598-021-03541-z
10. Mukwembi S, Nyabadza F. Predicting anti−cancer activity in flavonoids: a graph theoretic approach. *Scientific Reports*. 2023, *13*, 3309. https://doi.org/10.1038/s41598-023-30517-y
11. Zhen W., Khalid A., Ali P., Rehman H., Siddiqui M. K., Ullah H. Topological Study of Some Covid-19 Drugs by Using Temperature Indices. *Polycyclic Aromatic Compounds*. 2022. *43 (2)*, 1133-1144. 10.1080/10406638.2022.2025864
12. Zhang Y., Khalid A., Siddiqui M. K., Rehman H., Ishtiaq H., and Cancan M. On Analysis of Temperature Based Topological Indices of Some Covid-19 Drugs. *Polycyclic aromatic compounds*. 2023, *43*(4), 3810–3826. https://doi.org/10.1080/10406638.2022.2080238
13. https://www.rdkit.org/
14. https://www.chemeo.com/
15. Todeschini R., & Consonni V. (2000). *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH.
16. Devillers J., & Balaban A. T. (1999). *Topological Indices and Related Descriptors in QSAR and QSPR*. London: CRC Press
17. Roy K., Kar S., Das N.R. A Primer on QSAR/QSPR Modeling. Fundamental Concepts. Springer: 2015
18. M.R. Spiegel, John J. Schiller, R. A. Srinivasan Probability and Statistics. McGraw-Hill, New York, 2013, 424 p.
19. Besalú E., de Julián-Ortiz J. V., Iglesias M., Pogliani. L. An overlooked property of plot methods. *Journal of Mathematical Chemistry*. 2006, *39*, 475-484. https://doi.org/10.1007/s10910-005-9035-z
20. Hyndman R. J., Koehler A. B. Another look at measures of forecast accuracy, *International Journal of Forecasting*. 2006, *22*, 679– 688. https://doi.org/10.1016/j.ijforecast.2006.03.001
21. Hyndman R. J. Another Look at Forecast Accuracy Metrics for Intermittent Demand, *Foresight*. 2006, *4*, 43-46. https://robjhyndman.com/papers/foresight.pdf
22. Hewamalage H., Ackermann K., Bergmeir C. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Min Knowl Disc.* 2023, *37*, 788–832. https://doi.org/10.1007/s10618-022-00894-5
23. Kullback S. Information theory in statistics, Glouchester Mass, 1978, 399 p.
24. Kullback S. Leibler R.A. On information and sufficiency. *Ann. Math. Statist*. 1951, *22(1)*, 79 –86. https://doi.org/10.1214/aoms/1177729694
25. Hummer G., Garde S., Garcia A. E., Pohorille A., Prat L. R. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. USA.* 1996, *93*, 8951-8955. https://doi.org/10.1073/pnas.93.17.895
26. Arlot S., Celisse A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 2010, *4*, 40-79. https://doi.org/10.48550/arXiv.0907.4728
27. Quan N. T. The Prediction Sum of Squares as a General Measure for Regression Diagnostics. *J. Business & Economic Statistics*. 1988, *6(4)*, 501-504. https://doi.org/10.1080/07350015.1988.10509698
28. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*. 1974, *Ser. B, 36*, 111-133. http://www.jstor.org/stable/2984809
29. Cawley G. C., Talbot N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research.* 2010*, 11*, 2079-2107. http://jmlr.org/papers/v11/cawley10a.html

Received 19.05.2023 *Accepted 11.09.2023*

І. В. Христенко, В. В. Іванов. Параметри внутрішньої валідації рівнянь лінійної регресії в проблемі QSAR.

Харківський національний університет імені В. Н. Каразіна, майдан Свободи 4, Харків, 61022, Україна

У статті обговорюється набір внутрішніх параметрів валідації, які використовуються (або можуть бути використані) для опису якості регресійних моделей у задачах QSAR. Серед цих параметрів добре відомі коефіцієнт детермінації, залишкове середнє квадратичне відхилення, середня абсолютна похибка тощо. Також були досліджені індекси, засновані на дивергенції Кульбака-Лейблера як міри відстані між двома множинами. Всі параметри (індекси) були розраховані для декількох регресійних моделей, які описують температуру кипіння насичених вуглеводнів (алканів). Регресійні моделі включають чотирьохкомпонентну адитивну схему та рівняння, що описують температуру кипіння як функцію топологічних індексів. Два типи регресій на основі цих індексів - лінійні залежності тільки від одного топологічного індексу та лінійні залежності від кількості атомів вуглецю у вуглеводневій речовині та топологічного індексу.

Описано різні лінійні рівняння регресії з внутрішніми валідаційними параметрами, які оцінюють якість рівнянь з різних точок зору. Показано, що широкий набір тестових параметрів є не тільки додатковим, чи альтернативним описом регресійних моделей, а й забезпечує більш повніший опис прогностичних характеристик та якості отриманої регресійної моделі.

Ключові слова*: Кількісне співвідношення структура-властивість (QSAR), регресійні моделі, внутрішня валідація, топологічні дескриптори.*

Надіслано до редакції 19.05.2023 *Прийнято до друку 11.09.2023*

Kharkiv University Bulletin. Chemical Series. Issue 40 (63), 2023

Вісник Харківського національного університету, серія "Хімія", вип. 40 (63), 2023
https://doi.org/10.26565/2220-637X-2023-40-03 ISSN 2220-637X
УДК: 544.36:544.34

PROTON SOLVATION: COMPETITION BETWEEN ACETONE AND DIMETHYL SULFOXIDE

A. V. Lebed[a], N. O. Mchedlov-Petrossyan[b], D. Yu. Filatov[c], S. T. Goga[d]

V. N. Karazin Kharkiv National University, School of Chemistry, 4 Svobody sq., Kharkiv, 61022, Ukraine

a) alebed@karazin.ua https://orcid.org/0000-0001-5175-816X
b) mchedlov@karazin.ua https://orcid.org/0000-0001-6853-8411
c) d.filatov@prominvest.com.ua https://orcid.org/0000-0003-3758-4129
d) goga@karazin.ua https://orcid.org/0000-0002-2047-7119

This paper presents results of quantum-chemical study of proton exchange equilibrium between acetone (AC) and dimethyl sulfoxide (DMSO). Basing on the previous study of acid-base equilibria in AC with small additives of DMSO, the most probable composition of the lyonium ion is taken as $H(DMSO)_2^+$. The equilibrium geometries and energies of solvents molecules and all possible combinations of proton complexes with one or two solvent molecules were calculated in DFT level of theory using B3LYP functional and 6-31G++(d,p) basis set. The energies of solvated proton complexes were then calculated using PCM method. The energy change corresponding to the equilibrium $H(AC)_2^+ + 2\ DMSO \rightleftarrows H(DMSO)_2^+ + 2\ AC$ is –63.1 kJ mol^{-1} in the gas phase, which is in qualitative agreement with the standard Gibbs energy data, obtained experimentally in the gas phase by Kebarle's group,–77.6 kJ mol^{-1}, and with ΔG^{o}_{298} = –49.3 kJ mol^{-1} for AC solution with small additions of DMSO, determined by the Guss and Kolthoff method.

Keywords: *proton solvation; acetone; dimethyl sulfoxide; quantum-chemical study; energy of proton exchange.*

Introduction

This paper is devoted to the competition for proton between the molecules acetone and dimethyl sulfoxide (DMSO) using quantum-chemical calculations. The higher basicity of DMSO as compared with acetone is of common knowledge, see Table 1.

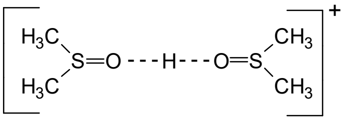
In addition to the numerical parameters gathered in Table 1, basing on experimental data, we decided to theoretically examine the proton exchange between these two solvents. Also, the same exchange equilibrium was estimated using our earlier published experimental results.

Table 1. Solvent descriptors characterizing the solvent basicity

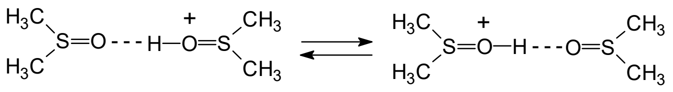
Parameter	Acetone	DMSO
DN^N	0.44	0.77
β	0.48	0.76
B	224	362
B_p	12.6	13.3
PA	831	897

Note. The data are taken from ref. [1]. DN^N is the normalized Gutmann's donor number; β is the solvatochromic Kamlet–Taft parameter; B is the Koppel–Palm parameter; B_p is the "polar basicity" according to Fawcett [2]; PA is the proton affinity, kJ mol^{-1}.

More than ten years ago, we studied the acid-base and related equilibria in acetone (AC) containing small additives of DMSO [3, 4]. It was demonstrated that the likely composition of the lyonium ion is $H(DMSO)_2^+$. In other words, the primary solvation shell can be represented in following manner:



or, more strictly:



(In this paper, we avoid discussing the likelihood of symmetric hydrogen bonds occurring.)

This was deduced from the dependence of the pK_a of picric acid on the DMSO content in the binary acetone – DMSO solvent and confirmed by the proximity of the limiting molar conductivities, Λ_0, at 25 °C of picric acid and tetraethylammonium picrate in acetone containing 5 mole % DMSO, a solvent with relative permittivity $\varepsilon_r = 22.19$. These values are 162.0±1.2 and 180.3±1.7 S mol^{-1} cm^2, respectively [3, 4].

At the same time, the linear size of the ion $H(DMSO)_2^+$, estimated as the distance between the most remote hydrogen atoms (Fig. 1a), is 900 pm. Crystallographic and various other experimental data available in the literature, as well as estimates using van der Waals radii of atoms, suggest that the diameter of the “pseudospherical” $N(C_2H_5)_4^+$ ion, which belongs to the “weakly solvated” category, is in the range from 674 to 800 pm [3, 4]. This amounts to (75–89) % of the linear size of the $H(DMSO)_2^+$ ion, which indicates the consistency of the proposed proton solvation model. In any case, the existence of the lyonium ion in the form of $H(DMSO)^+$ would be more difficult to reconcile with its lower mobility compared to the $N(C_2H_5)_4^+$ ion. As a result, we consider $H(DMSO)_2^+$ as the most probable composition of the lyonium ion.

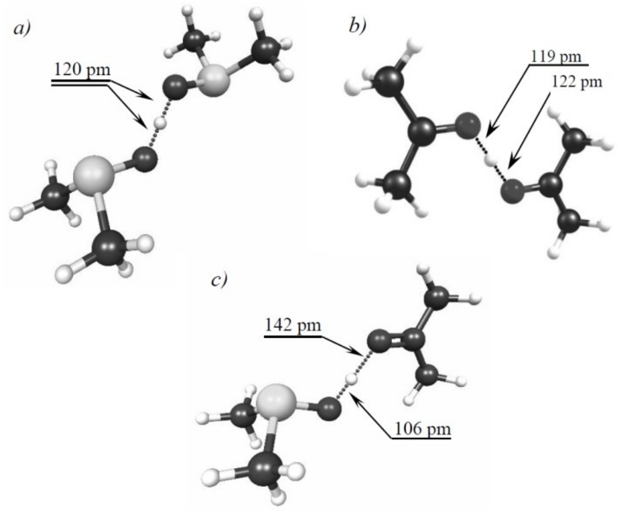


Fig. 1. Geometries of complexes a) $H(DMSO)_2^+$; b) $H(AC)_2^+$ and c) $H(AC,DMSO)^+$ from the quantum chemical calculations

In pure DMSO, this composition is preferred on the basis of data from IR spectroscopy [5, 6].

The geometric and energy parameters of proton complexes with acetone and DMSO molecules were also estimated using quantum chemical calculations. All calculations were provided using Gaussian 09 program set [7]. The geometries of individual solvent molecules, as well as complexes of a proton with one or two solvent molecules in vacuum (Fig. 1) were optimized using DFT (density functional theory) method with the B3LYP hybrid functional and 6-31G++(d,p) basis set. Additional keywords opt=tight and int=ultrafine were used to ensure accuracy of optimization procedure. The vibrational frequencies were also calculated and imaginary vibration frequencies were absent. The energies of these solvate complexes in acetone were then calculated using the polarized continuum model (PCM). Zero-point energy and thermal correction to 298 K were taken into account when calculating the energies.

According to our calculations, the O – O distance in the $H(DMSO)_2^+$ ion is 2.405 Å, which is in good agreement with the value of 2.403 Å calculated by Denisov *et al.* [8] at B3LYP/6-31G(d,p) level.

The transformation of structure 1a into structure 1b requires a cost of 20.9 kJ mol^{-1}. The greater basicity of the DMSO molecule is also clearly demonstrated by the bond lengths between the proton and oxygen atoms in two different structures. This result is very instructive, and a similar approach can be applied to other systems.

Energy changes corresponding to proton exchange between various proton solvates in acetone were also calculated:

$$H(AC)_2^+ + 2\,DMSO \rightleftarrows H(DMSO)_2^+ + 2\,AC;\ \Delta E_{2-2} = -63.1\ \text{kJ mol}^{-1}, \tag{1}$$

$$H(AC)^+ + DMSO \rightleftarrows H(DMSO)^+ + AC;\ \Delta E_{1-1} = -57.5\ \text{kJ mol}^{-1}. \tag{2}$$

The significantly higher basicity of DMSO compared to AC is also confirmed by the experimental data of the Kebarle group [9] on the values of the standard Gibbs energies of the indicated proton exchange reactions in the gas phase: ΔG°_{298} (1) = – 77.6 kJ mol^{-1} and ΔG°_{298} (2) = – 64.8 kJ mol^{-1}.

In addition, the proton exchange between acetone and DMSO in acetone solution can be examined using the approach developed by Guss and Kolthoff [10]. If only one solvate complex can be taken into account and ruling out the possibility of mixed proton solvates, such as H(AC, DMSO)$^+$, the following equilibria can be derived:

$$H(AC)_m^+ + n\text{DMSO} \rightleftarrows H(DMSO)_n^+ + m\text{AC};\quad K_r = \frac{a_{H(DMSO)_n^+}\, a^m_{AC}}{a_{H(AC)_m^+}\, a^n_{DMSO}} \tag{3}$$

In this case, according to Guss and Kolthoff, the following equation is valid:

$$K_a = K_a^{AC}\left(1 + K_r \frac{a^n_{DMSO}}{a^m_{AC}}\right) \tag{4}$$

Using the experimental pK_a values of picric acid in acetone – DMSO mixtures with molar fraction of DMSO from 0.005 to 0.06 [3, 4], taking pK_a^{AC} = 9.2 from the work by Foltin and Majer [11], and equating *n* and *m* to 2, we obtained a value $K_{r(2\text{-}2)} = (4.3\pm0.8)\times10^8$. Accordingly, $\Delta G^{\circ}_{298}(1) = -2.303\,RT\log K_{r(2\text{-}2)} = -49.3$ kJ mol^{-1}, which is in qualitative agreement with the results of quantum chemical calculations for the gas phase.

Similar calculations can be carried out for the case $n = 2$, $m = 1$ ($K_{r(1\text{-}2)}$), etc. However, as quantum chemical calculations show, the addition of a second DMSO molecule to H(DMSO)$^+$ and a second AC molecule to H(AC)$^+$ in acetone is accompanied by such significant energy gains (60.26 and 54.68 kJ·mol^{-1}, respectively), that it is quite justified to take into account only the solvates H(DMSO)$_2^+$ and H(AC)$_2^+$. A solvate of composition H(AC, DMSO)$^+$ is possible only in the case of DMSO deficiency, but under the conditions of our experiment, the proton concentration is 3–4 orders of magnitude lower than the DMSO concentration.

Thus, we have proven that in acetone with DMSO additives, the lyonium ion exists mainly in the form of H(DMSO)$_2^+$, and that the proton exchange constant between the two solvents is very high and amounts to over 10^8. Both theoretical and experimental data give new convincing support to the more strong basic properties of DMSO as compared with acetone.

Acknowledgements

This study was partially supported by the Ministry of Education and Science of Ukraine via grant number 0122U001485 and by the Simons Foundation (USA) via grant number 1030292.

References

1. Reichardt C., Welton T. Solvents and Solvent Effects in Organic Chemistry. Wiley-VCH; Weinheim, 2011.
2. Fawsett W.R. Acidity and basicity scales for polar solvents. *J. Phys. Chem.,* **1993**, 97 (37), 9540-9546. https://doi.org/10.1021/j100139a045
3. Lebed A.V., Mchedlov-Petrossyan N.O., Filatov D.Yu., Goga S.T. Acid-base equilibrium in a mixture of protophobic and protophilic polar non-hydrogen bond donor solvents: Stoichiometry of proton solvation by dimethyl sulfoxide in acetone. Proc. Nat. Acad. Sci. Ukraine. **2008**, no. 9, 121-125. http://dspace.nbuv.gov.ua/handle/123456789/5919
4. Mchedlov-Petrossyan N.O., Filatov D.Yu., Goga S.T., Lebed A.V. Ionic equilibrium in mixtures of protophobic and protophilic polar non-hydrogen bond donor solvents: Acids, salts, and indicators in acetone containing 5 mol % DMSO. *J. Phys. Org. Chem.,* **2010**, 23(5), 418-430. https://doi.org/10.1002/poc.1616

5. Williams J.M., Kreevoy M. Structure and infrared spectrum of the solvated proton in dimethyl sulfoxide. *J. Am. Chem. Soc.* **1967**, 89(21), 5499-5501. https://doi.org/10.1021/ja00997a062
6. Kampschulte-Scheuing I., Zundel G. Tunnel effect, infrared continuum, and solvate structure in aqueous and anhydrous acid solutions. *J. Phys. Chem.,* **1970**, 74(11), 2363-2368. https://doi.org/10.1021/j100705a021
7. Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2010.
8. Denisov G.S., Koll A., Lobadyuk V.I., Schreiber V.M., Shurukhina A.V., Spevak V.N. Hydrogen bonding in coordination compounds containing homoconjugated bis-dimethylsulfoxide cation. Ab initio and IR studies. *J. Mol. Struct.*, **2002**, 605(2-3), 221-226. https://doi.org/10.1016/S0022-2860(01)00783-9
9. Lau Y.K., Saluja P.P.S., Kebarle P. The proton in dimethyl sulfoxide and acetone. Results from -phase ion equilibria involving $(Me_2SO)_nH^+$ and $(Me_2CO)_nH^+$. *J. Amer. Chem. Soc.*, **1980**, 102(25), 7429-7433. https://doi.org/10.1021/ja00545a004
10. Guss L.S., Kolthoff I.M. The distribution of protons between water and other solvents. *J. Amer. Chem. Soc.,* **1940**, 62(6), 1494-1496. https://doi.org/10.1021/ja01863a047
11. Foltin M., Majer P. Determination of dissociation constant of picric acid and calibration of potentiometric cell in acetone. *Coll. Czech. Chem. Commun*., **1978**, 43(1), 95-102. https://doi.org/10.1135/cccc19780095

Received 25.05.2023 *Accepted 11.09.2023*

О.В. Лебідь, Н.О. Мчедлов-Петросян, Д.Ю. Філатов, С.Т. Гога. Сольватація протона: конкуренція між ацетоном та диметилсульфоксидом.

Харківський національний університет імені В.Н. Каразіна, хімічний факультет, майдан Свободи, 4, Харків, 61022, Україна

Ця стаття містить результати квантово-хімічного дослідження обміну протоном між ацетоном (AC) та диметилсульфоксидом (DMSO). Базуючись на попередніх дослідженнях кислотно-основних рівноваг в AC з малими домішками DMSO, як найімовірнішим складом іону ліонію прийнято $H(DMSO)_2^+$. Рівноважна геометрія та енергії молекул розчинника та усіх можливих комбінацій комплексів протону з однією або двома молекулами розчинників обчислені на рівні DFT з використанням функціоналу B3LYP ьа базису 6-31G++(d,p). Енергії сольватації комплексів протону обчислені за допомогою методу PCM. Зміна енергії, яка відповідає рівновазі $H(AC)_2^+ + 2\ DMSO \rightleftarrows H(DMSO)_2^+ + 2\ AC$, складає – 63.1 kJ mol^{-1} в газовій фазі. Значення цієї величини якісно узгоджуються зі значенням стандартної енергії Гіббса, – 77.6 kJ mol^{-1}, знайденої для газової фази експериментально групою Кебарле, та значенням ΔG^{o}_{298} = –49.3 kJ mol^{-1} для розчинів AC з малими домішками DMSO, знайденим методом Гасса і Кольтгофа.

Ключові слова: *сольватація протона; ацетон; диметилсульфоксид; квантово-хімічне дослідження; енергія обміну протоном.*

Надіслано до редакції 25.05.2023 *Прийнято до друку 11.09.2023*

ЕТИЧНІ НОРМИ ПУБЛІКАЦІЇ НАУКОВИХ РЕЗУЛЬТАТІВ ТА ЇХ ПОРУШЕННЯ. Редакційна колегія робить все можливе для дотримання етичних норм, прийнятих міжнародним науковим товариством, і для запобігання будь-яких порушень цих норм. Така політика є важливою умовою плідної участі журналу в розвитку цілісної системи знань в галузі хімії та суміжних галузях. Діяльність редакційної колегії значною мірою спирається на рекомендації Комітету з етики наукових публікацій (Committee of Publication Ethics), а також на цінний досвід міжнародних журналів та видавництв. Подання статті на розгляд означає, що вона містить отримані авторами нові нетривіальні наукові результати, які раніше не були опубліковані. Кожну статтю рецензують щонайменше два експерти, які мають усі можливості вільно висловити мотивовані критичні зауваження щодо рівня та ясності представлення матеріалу, його відповідності профілю журналу, новизни та достовірності результатів. Рекомендації рецензентів є основою для прийняття остаточного рішення щодо публікації статті. Якщо статтю прийнято, вона розміщується у відкритому доступі; авторські права зберігаються за авторами. За наявності будь-яких конфліктів інтересів (фінансових, академічних, персональних та інших), учасники процесу рецензування мають сповістити редакційну колегію про це. Всі питання, пов’язані з можливим плагіатом або фальсифікацією результатів ретельно обговорюються редакційною колегією, рівно як спори щодо авторства та доцільність роздроблення результатів на невеличкі статті. Доведені плагіат чи фальсифікація результатів є підставами для безумовного відхилення статті.

STATEMENT ON THE PUBLICATION ETHICS AND MALPRACTICE. The Editorial Board has been doing its best to keep the ethical standards adopted by the world scientific community and to prevent the publication malpractice of any kind. This policy is considered to be an imperative condition for the fruitful contribution of the journal in the development of the modern network of knowledge in chemistry and boundary fields. The activity of the Editorial Board in this respect is based, in particular, on the recommendations of the Committee of Publication Ethics and valuable practice of world-leading journals and publishers. The submission of a manuscript implies that it contains new significant scientific results obtained by authors that where never published before. Each paper is peer reviewed by at least two independent experts who are completely free to express their motivated critical comments on the level of the research, its novelty, reliability, readability and relevance to the journal scope. These comments are the background for the final decision about the paper. Once the manuscript is accepted, it becomes the open-access paper, and the copyright remains with authors. All participants of the review process are strongly asked to disclose conflicts of interest of any kind (financial, academic, personal, etc.). Any indication of plagiarism or fraudulent research receives extremely serious attention from the side of the Editorial Board, as well as authorship disputes and groundless subdivision of the results into several small papers. Confirmed plagiarism or fraudulent research entail the categorical rejection of the manuscript.

ІНФОРМАЦІЯ ДЛЯ АВТОРІВ. Журнал публікує статті російською, англійською та українською мовами. До публікації приймаються: огляди (за погодженням з редколегією); оригінальні статті, обсяг 6-10 журнальних сторінок; короткі повідомлення, обсяг до 3 журнальних сторінок. Крім звичайного списку літератури, в статті обов'язково повинен бути другий список, всі посилання якого дані латиницею. Правила підготовки цього списку наведені в розділі «Транслітерація» на сайті журналу. Обидва списки повинні бути повністю ідентичні. При рецензуванні статей один з критеріїв - наявність посилань на публікації останніх років. Стаття обов'язково повинна містити резюме російською, українською та англійською мовами. У всіх трьох необхідно вказати назву статті, прізвища авторів і ключові слова. Орієнтовний обсяг резюме - 1800 знаків (без урахування заголовку і ключових слів). Редакція приймає електронний (MS Word) і два роздрукованих (для харків'ян) тексту рукопису. Адреси вказані в розділі «Контакти» на сайті журналу. Супровідний лист до статті, виправленої відповідно до зауважень рецензента, повинен містити відповіді на всі зауваження. Подається електронний і один роздрукований (для харків'ян) варіант. Рукописи, які пройшли рецензування, прийняті до публікації і оформлені відповідно до правил для авторів, приймаються у форматі doc (не docx) електронною поштою (chembull@karazin.ua). Роздрукований варіант не потрібен. Докладніша інформація розміщена на сайті журналу http://chembull.univer.kharkov.ua.

INFORMATION FOR AUTHORS. Papers in Ukrainian, Russian and English are published. These may be invited papers; review papers (require preliminary agreement with Editors); regular papers; brief communications. In preparing the manuscript it is mandatory to keep the statement on the publication ethics and malpractice, which can be found on the web-site and in each issue. The article should contain summaries in English, Russian, and Ukrainian. In all three it is necessary to indicate the title of the article, the names of the authors and the keywords. The approximate volume of summary is 1800 characters (excluding the title and key words). The help in translation is provided by request for foreign authors.. Any style of references is acceptable, but all references within the paper must be given in the same style. In addition, the second, transliterated, list of references is required if at least one original reference is given in Cyrillic. See section "Transliteration" of the web-site for details. Please use papers of previous issues as samples when prepare the manuscript. The MS Word format is used. Standard fonts (Times New Roman, Arial, Symbol) are preferable. Figures and diagrams are required in vector formats. Figure captions are given separately. All figures, tables and equations are numbered. Please use MS Equation Editor or MathType to prepare mathematical equations and ISIS Draw to prepare chemical formulas and equations. The decimal point (not coma) is accepted in the journal. Please avoid any kind of formatting when prepare the manuscript. Manuscripts may be submitted to the Editor-in-Chief via e-mai chembull@karazin.ua. For more detailed information see the journal web-site http://chembull.univer.kharkov.ua.

Наукове видання

Вісник Харківського національного університету імені В.Н. Каразіна

Серія «Хімія»
Вип. 40 (63)
Збірник наукових праць
Українською та англійською мовами.

Технічний редактор:
Д.О. Анохін

Підписано до друку «05» жовтня 2023. Формат 60х84/8.
Ум.-друк. арк. 1,9 Обл.-вид. арк. 2,4.
Тираж 100 пр. Ціна договірна.

61022, Харків, майдан Свободи, 4
Харківський національний університет імені В.Н. Каразіна,
Видавництво Харківського національного університету імені В.Н. Каразіна

Надруковано: ХНУ імені В.Н. Каразіна
61022, м. Харків, майдан Свободи, 4.
Тел.: 705-24-32

Свідоцтво суб'єкта видавничої справи ДК № 3367 від 13.01.09