

GEOMETRY OPTIMIZATION IN WATER AND VACUUM: QPPR MODELING ON SELENAZOLE COMPOUNDS

Cozma C. Alina^{***}, Bolboacă D. Sorana^{**}, Jäntschi Lorentz^{****}

* University of Oradea, Department of Chemistry, Universității Street no. 1, 410087 Oradea, Romania, e-mail: acozma@uoradea.ro

** Iuliu Hațieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, Louis Pasteur Street no. 6, 400349 Cluj-Napoca, Romania, e-mail: sbolboaca@umfcluj.ro

*** Technical University of Cluj-Napoca, Department of Physics and Chemistry, Muncii Bvd. 103-105, 400641 Cluj-Napoca, Romania, e-mail: lorentz.jantschi@gmail.com

Abstract

A sample of fifteen selenazole related compounds were structural modeled in order to find quantitative property-property relationships, with main goal as assessment of the effect of the same properties calculated on water and vacuum environments. Three properties derived from chromatographic measurements were subject to analysis: retention factor, specific surface area of the solvent and chromatographic hydrophobicity index. The structural models of the compounds was optimized using Hartree-Fock (restricted, RHF) method with 6-31G as basis set, in both vacuum and water (SM8 model) environments. Furthermore, for each compound and geometry (vacuum and water environments) energy calculations were involved, and from several properties have been also derived, by keeping untouched the level of theory. Full search approach was used to find simple and multiple linear regression models. No reliable model in regards of internal validity was obtained for specific surface area of the solvent. The models with higher performances in estimation and prediction expressed as determination coefficients for retention factor and chromatographic hydrophobicity index proved to catch the additive effect of structure based properties calculated in both water and vacuum at the Hartree-Fock level of theory.*

Key words: geometry optimization, organic compounds with selenium, quantitative property-property relationships (QPPR)

INTRODUCTION

The properties of material are determined by their electronic structure that is closely related to the field of quantum mechanics. Correlations are generated through software provided by various computational approximations (Le and Winkler, 2015). These approximations appear in computational chemistry that become an important field of researcher due to realistic simulations of complex materials (Li and Hall, 2015; Suvitha et al., 2015).

Geometry optimization of chemical compounds is an important step in preparation of molecule for modeling whenever structural features of the molecules are translated in descriptors (Rinnan et al., 2010). The optimum

geometry is the geometry that minimizes the strain on a given system is known as an optimum geometry and any perturbation from optimum will induce the system to change so as to reduce this perturbation unless prevented by external forces. Optimization methods could be classified as methods based on quantum mechanics and methods based on classical mechanics. Optimization of geometry could be done manually or automatically to assure fewer errors (Jäntschi, 2011; Pencheva et al., 2008). Hartree-Fock (HF) models, method limited to 30 atoms, are used to calculate energies and wavefunctions state geometries and vibrational frequencies with STO-3G (Ernest and David, 1986), 3-21G, 6-31G* (Ditchfield et al., 1971) and 6-311+G** (McLean and Chandler, 1980) basis sets (Hehre, 2003; Hehre et al., 1986). The HF/6-31G* models are appropriate for equilibrium and transition-state structure determinations of organic and main-group inorganic molecules (except molecules with transition metals) as well as for calculation of reaction energies (except reactions involving bond making or breaking).

The aim of our research was to assess how optimization on water and vacuum using the same approach and basis set is reflected on the association between the structure and the parameters derived from chromatographic measurements on the sample of hidrazinoselenazole compounds.

MATERIAL AND METHOD

Selenazole related compounds dataset

The structures of fifteen selenazole related compounds, of which structures are depicted in Fig.1 was previously identified and analyzed using RP-HPLC (Reverse Phase - High-Performance Liquid Chromatography), see (Zaharia et al., 2013; Grozav et al., 2013) and was included in this analysis. For these compounds, three properties derived from chromatographic measurements (Cozma, 2011; 2012) were investigated, being presented in Table 1.

Table 1

Selenazole compounds and their properties from chromatographic measurements															
Compound	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15
log(Rf)	5.19	4.2	4	4.75	4.71	4.95	3.51	4.22	2.83	3.82	2.31	2.18	1.67	3.08	4.15
S·100	-5.52	-4.79	-4.67	-5.6	-6.13	-5.63	-4.6	-4.71	-4.39	-5.15	-4.43	-3.6	-3.22	-4.62	-5.54
$\varphi_0/10$	9.41	8.75	8.56	8.48	7.68	8.8	7.63	8.96	6.44	7.41	5.21	6.07	5.2	6.66	7.49

Rf = retention factor; S = specific surface area of the solvent; φ_0 = chromatographic hydrophobicity index

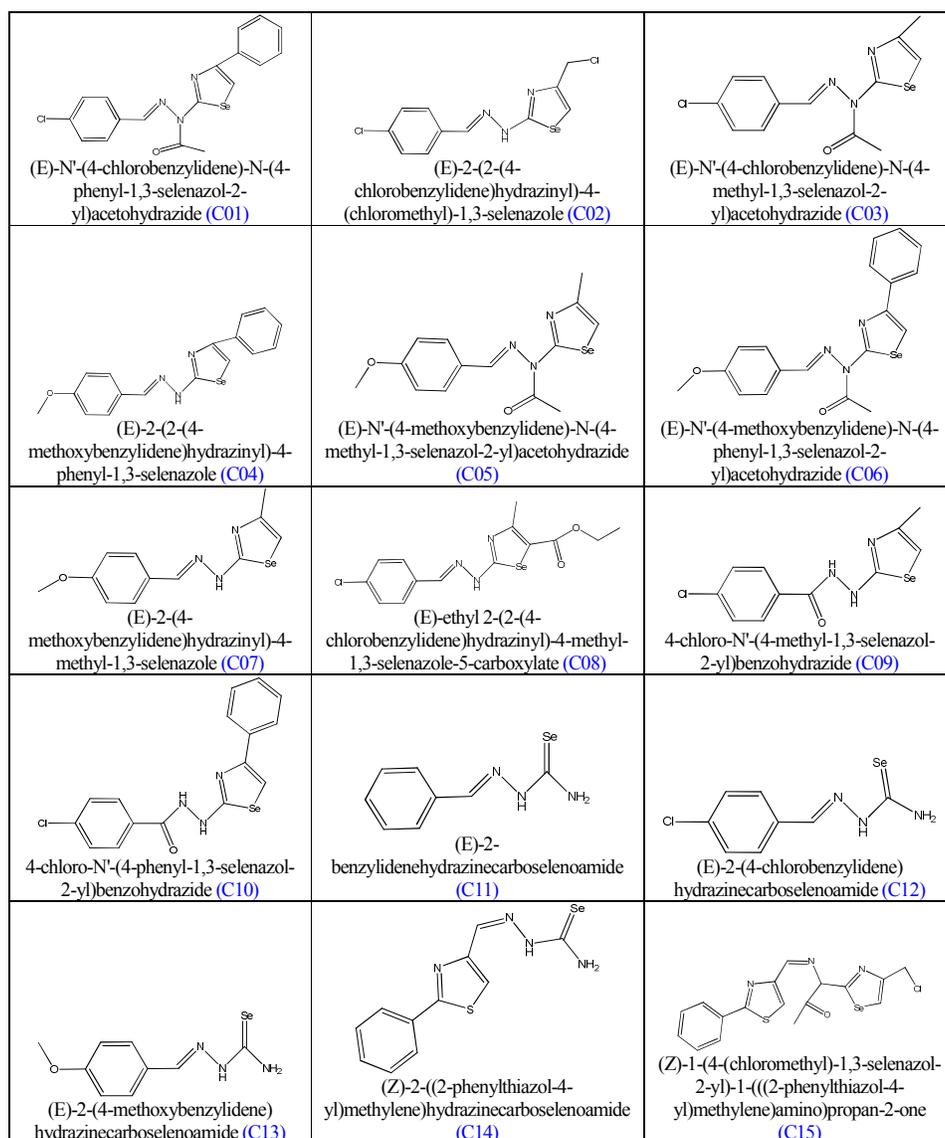


Fig. 1. Investigated class of compounds (C01-C15)

Geometry optimization and calculation of properties

The structures of the investigated chemicals were built with ChemDraw software (v. 6.0) and subject to optimization of geometry with Spartan software (v. 8) using equilibrium geometry at ground state with Hartree-Fock RHF method with 6-31G* basis set, both in vacuum and water (SM8 model).

The calculated properties were from energy calculations at same

theory level as follows: Hartree-Fock energy (HFE), temperature correction (cHv), entropy correction (cSv), free energy of solvation - water SM8 (FESM8), and standard thermodynamic quantities at 298.15 K and 1.00 atm: zero-point energy (ZPE, kJ/mol), enthalpy (H, au = atomic units), entropy (S, J/mol), heat capacity at constant volume (Cv, J/mol) for total vibrations (Htv, Stv, Cvv), ideal gas (Hig), translation (Htr, Str, Cvt), rotation (Hr, Sr, Cvr), and their totals (Ht, St, Cvt). For the properties calculated for the dataset optimized in water the labels have additionally '_w' in the name of the property, '_v' for the ones optimized in vacuum. Free energy of solvation - water SM8 (FESM8) was calculated for molecules optimized in water.

Modeling approach

The three properties derived from previously conducted chromatographic measurements from Table 1 were subject of modeling using the pool of structural based derived properties, named descriptors in the next. In the pool were 17 descriptors derived from vacuum models and 18 from water models (additionally is Free energy of solvation, FESM8 in the next).

The pool of descriptors was filtered before modeling and all with determination coefficient (r^2) smaller than 0.001 (with measured property subject to analysis) or higher than 0.999 in between (any other one) were deleted from the pool of the input data in the first stage of analysis.

In the second stage of analysis, a systematic search for all possible associations of descriptors able to explain the measured properties were conducted, when simple linear regressions (with one dependent variable) and multiple linear regressions (with two and three dependent variables) were listed as possible (candidate) relationships able to explain the association.

A home-made program was used to systematically search for these models and their results are reproducible when any other well known software for statistical analysis is used.

The models with the highest goodness-of-fit were assessed in terms of estimation (determination coefficient, standard error of estimate, F and t-statistics and associated values) and prediction (leave-one-out analysis and percentage prediction error) abilities (Bolboacă and Jäntschi, 2008; 2013).

RESULTS AND DISSCUSIONS

From the thirty-five descriptors coming from the calculations in water (17) and vacuum (18) environments, after applying the first stage of filtering, it remained each time (three times, for each chromatographic parameter) 14 regressors (40%).

Following calculated properties were deleted due to their redundant character:

- ÷ Ht_v having with ZPE_v $R^2 > 0.9997$
- ÷ Cvt_v having with Cvv_v $R^2 > 0.9999$
- ÷ cHv_v having with ZPE_v $R^2 > 0.9997$
- ÷ cSv_v having with ZPE_v $R^2 > 0.9995$
- ÷ ZPE_w having with ZPE_v $R^2 > 0.9999$
- ÷ Str_w having with Str_v $R^2 > 0.9999$
- ÷ Ht_w having with ZPE_v $R^2 > 0.9998$
- ÷ Cvt_w having with Cvv_w $R^2 > 0.9999$
- ÷ cHv_w having with ZPE_v $R^2 > 0.9998$
- ÷ cSv_w having with ZPE_v $R^2 > 0.9993$
- ÷ HFE_w having with HFE_v $R^2 > 0.9999$

The pool of regressors after these deletes comprised the following calculated properties: ZPE_v, Htv_v, Stv_v, Cvv_v, Str_v, Sr_v, St_v, HFE_v, Htv_w, Stv_w, Cvv_w, Sr_w, St_w, and FESM8.

The models with one, two, and three descriptors are properly to be designed by taking into account the number of compounds - see for details (Hawkins, 2004).

Significant models with one, two and respectively three regressors identified are presented in Table 2 and show the additive effect between the calculated properties for molecules optimized in both water and vacuum.

The models with highest determination coefficient among those with one independent variable are obtained in two out of three cases with properties calculated in vacuum (Eq(4) and Eq(7)). Models with highest goodness-of-fit for all measured properties proved to combine properties calculated on both water and vacuum (Eq(3), Eq(6), and Eq(9)).

Table 2

QPPR models with highest goodness-of-fit according to number of regressors

Prop.	Eq	Model	R ² (%)
log(Rf)	1	0+Cvv_w*1.61e-2(t= 27.13)	75.43
	2	-1.46e+1(t= 3.77)+FESM8* 2.29e-2(t= 3.90)+Sr_w* 1.34e-1(t= 5.24)	86.15
	3	0+FESM8* 2.65e-2(t= 4.64)+Stv_w* -2.29e-1(t= 3.65)+Htv_v* 1.48e+0(t= 4.04)	90.71
S	4	-1.94e-2(t= 3.15)+ZPE_v* -4.37e-5(t= 4.81)	64.02
	5	-1.72e-2(t= 3.15)+Stv_v* 1.72e-3(t= 2.55)+Htv_v* -1.09e-2(t= 2.79)	74.12
	6	-1.48e-2(t= 3.07)+Cvv_w* 1.44e-3(t= 2.27)+Stv_v* 4.47e-3(t= 3.33)+Htv_v* -3.82e-2(t= 3.06)	82.39
φ ₀	7	-7.82e+2(t= 4.74)+Str_v* 4.74(t= 5.20)	67.51
	8	-5.35e+2(t= 3.78)+FESM8* 2.82e-1(t= 3.43)+Str_v* 3.46e+0(t= 4.48)	83.58
	9	-6.18e+2(t= 4.37)+FESM8* 5.33e-1(t= 6.34)+Stv_w* -2.51e+0(t= 4.47)+St_v* 2.30e+0(t= 4.80)	91.96

Log(Rf) = logarithm of retention factor; S = specific surface area of the solvent; φ₀ = chromatographic hydrophobicity index; R² = determination coefficient; t = t-statistic associated to intercept or slope; _v = calculations in vacuum; _w = calculations in water;

Main characteristics of models with highest determination coefficient are presented in Table 2. Graphical representation of models detailed in Table 3 are presented in Fig. 2.

Table 3

Characteristics of models with highest goodness-of-fit

Eq	R ² _{adj} (%)	se	F(p-value)	Q ² (%)	se _{loo}	F _{loo} (p _{loo})
3	80.82	0.3550	39 (3.71e-6)	86.81	0.4436	24 (4.03e-5)
6	77.59	0.0037	17 (1.85e-4)	58.69	0.0059	5 (2.40e-2)
9	89.77	4.3225	42 (2.59e-6)	85.73	5.8206	21 (6.62e-5)

R²_{adj} = adjusted determination coefficient; se = standard error of estimate; loo = leave-one-out analysis; F(p-value) = F-statistic and associated p-value; Q² = determination coefficient in leave-one-out analysis

The analysis of the models with highest goodness-of-fit showed that two calculated properties able to explain retention factors – log(Rf) (free energy of solvation – water SM8 and entropy for total vibrations) are also able to explain chromatographic hydrophobicity index (φ_0) which sustain the consistency of the models since $\varphi_0 = \log(Rf)/S$. The retention factors are explained by free energy of solvation (SM8 model), and entropy (water) and enthalpy (vacuum) for total vibrations. Specific surface area of solvent is explained by heat capacity at constant volume (water) and entropy (vacuum) and enthalpy (vacuum) for total vibrations.

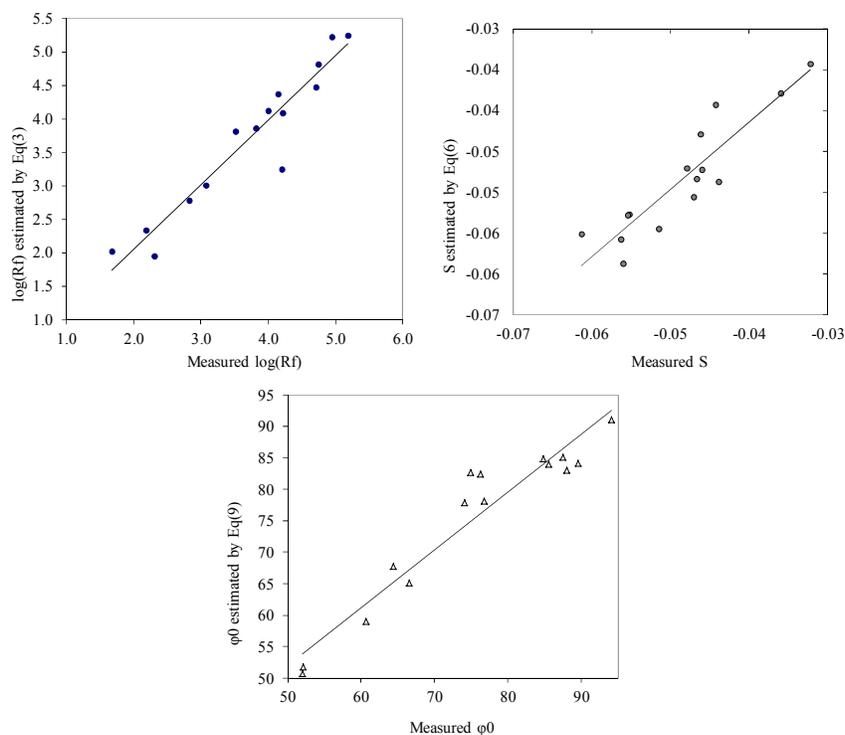


Fig. 2. Measured properties estimated by models with three regressors

The internal validation analysis showed that the best performing model defined in terms of smallest difference between determination

coefficient of the model and determination coefficient in leave-one-out analysis is given by Eq(3) followed by Eq(9). The model of the surface area of solvent, which contain two properties calculated in vacuum and one calculated in water, proved smallest internal validity, with a difference in determination of 23.7%. In this regards, the model for surface area of the solvent could not be considered valid; this property seems not to be linearly related with calculated properties.

Relations between surface area of the solvent are reported in the scientific literature in relation with different properties/activities (such as dimer stability on polycyclic aromatic hydrocarbons (Lowe et al., 2015), or oxidation of organic micropollutants (Sudhakaran and Amy, 2013)). Cozma et al. (2012) showed on the same class of compounds that basic feature can be extracted from RP-HPLC (reversed phase-high performance liquid chromatography).

CONCLUSIONS

Two out of three measured properties of investigated compounds proved linearly related with calculated properties after optimization of molecules in water and vacuum.

No valid model was obtained for the specific surface area of the solvent while models with performance in terms of goodness-of-fit were obtained for the logarithm of retention factor and chromatographic hydrophobicity index.

Our results showed additive effects between calculated properties on water and vacuum with an important contribution of free energy of solvation (SM8 model).

Acknowledgments

This paper was published under the frame of European Social Found, Human Resources Development Operational Programme 2007-2013, project no. POSDRU/159/1.5/138776.

The home-made programs were developed and used in the FreePascal environment, a free, easy to use and fast developing environment, here being acknowledged this.

REFERENCES

1. Bolboacă, S.D., Jäntschi, L., 2008, Modelling the property of compounds from structure: statistical methods for models validation, *Environ. Chem. Lett.*, 6, 175-181.
2. Bolboacă, S.D., Jäntschi, L., 2013, Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example, *International Journal on Mathematical Methods and Models in Biosciences (BIOMATH)*, 2(1), 130908.
3. Cozma, A., 2011, The study of lipophilicity for new thiazole and selenazol

- derivatives with biological potential by reversed – phase liquid chromatography, PhD Thesis, Babeş-Bolyai University Cluj-Napoca, Romania.
4. Cozma, A., Vlase, L., Ignat, A., Zaharia, V., Goca, S., Măruțoiu, C., Fodor, A., 2012, Lipophilicity Study of New Selenazole Derivatives by RP-HPLC, *Rev. Chim. (Bucharest)*, 63(7), 651-655.
 5. Ditchfield, R., Hehre, W.J., Pople, J.A., 1971, Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules, *J. Chem. Phys.*, 54(2), 724-728.
 6. Ernest, D., David, F., 1986, Basis set selection for molecular calculations, *Chem. Rev.*, 86(4), 681-696.
 7. Grozav, A., Gaina, L., Kuete, V., Silaghi-Dumitrescu, L., Efferth, T., Zaharia, V., 2013, Microwave-Assisted Synthesis of New Selenazole Derivatives with Antiproliferative Activity, *Molecules*, 18(4), 4679-4688.
 8. Hawkins, D.M., 2004, The problem of overfitting, *J. Chem. Inf. Comput. Sci.*, 44(1), 1-12.
 9. Hehre W.J., 2003, A Guide to Molecular Mechanics and Quantum Chemical Calculations, Wavefunction, Inc., Irvine, CA.
 10. Hehre, W.J., Radom, L., Schleyer, P.V.R., Pople, J.A., 1986, *Ab Initio Molecular Orbital Theory*, Wiley, New York.
 11. Jäntschi, L., 2011, Computer Assisted Geometry Optimization for in silico Modeling, *Appl. Med. Inform.*, 29(3), 11-18.
 12. Le, T.C., Winkler, D.A., 2015, A Bright Future for Evolutionary Methods in Drug Design, *ChemMedChem*, 10(8), 1296-1300.
 13. Li, H., Hall, M.B., 2015 Computational mechanistic studies on reactions of transition metal complexes with noninnocent pincer ligands: Aromatization-dearomatization or not, *ACS Catalysis*, 5(3), 1895-1913.
 14. Lowe, J.S., Lai, J.Y.W., Elvati, P., Violi, A., 2015, Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity, *Proceedings of the Combustion Institute*, 35(2), 1827-1832.
 15. McLean, A.D., Chandler, G.S., 1980, Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18, *J. Chem. Phys.*, 72, 5639-5648.
 16. Pencheva, T., Lagorce, D., Pajeva, I., Villoutreix, B.O., Miteva, M.A., 2008, AMMOS: Automated Molecular Mechanics Optimization tool for in silico Screening, *BMC Bioinformatics*, 9, 438.
 17. Rinnan, A., Christensen, N.J., Engelsen, S.B., 2010, How the energy evaluation method used in the geometry optimization step affect the quality of the subsequent QSAR/QSPR models. *J. Comput. Aided Mol Des.*, 24(1), 17-22.
 18. Sudhakaran, S., Amy, G.L., 2013, QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification, *Water Res.*, 47(3), pp. 1111-1122.
 19. Suvitha, A., Periandy, S., Gayathri, P., 2015, NBO, HOMO-LUMO, UV, NLO, NMR and vibrational analysis of veratrole using FT-IR, FT-Raman, FT-NMR spectra and HF-DFT computational methods, *Spectrochim. Acta A*, 138, 357-369.
 20. Zaharia, V., Ignat, A., Ngameni, B., Kuete, V., Mounang, M.L., Fokunang, C.N., Vasilescu, M., Palibroda, N., Cristea, C., Silaghi-Dumitrescu, L., Ngadjui, B.T., 2013, Heterocycles 23: Synthesis, characterization and anticancer activity of new hydrazinoselenazole derivatives, *Med. Chem. Res.*, 22(12), 5670-5679.