VILNIUS UNIVERSITY

Vytautas Raškevičius

COMPUTATIONAL STUDIES OF PROTEIN AND LIGAND INTERACTIONS

Summary of doctoral Dissertation
Technology Sciences, Chemical engineering (05T)

Vilnius, 2017

The dissertation work was carried out during 2012-2016 at the Vilnius University.

Scientific supervisor – dr. Visvaldas Kairys (Vilnius University, technological sciences, chemical engineering – 05T)

The dissertation defense is being held at the Council of Chemical Engineering science direction of Vilnius University:
Chairman – prof. habil. dr. Arūnas Ramanavičius (Vilnius University, physical sciences, chemistry – 03P)
Members:
dr. Saulius Gražulis (Vilnius University, technological sciences, chemical engineering – 05T);
dr. Vytautas Petrauskas (Vilnius University, technological sciences, chemical engineering – 05T);
prof. dr. Jolanta Sereikaitė (Vilnius Gediminas Technical University, technological sciences, chemical engineering – 05T);
dr. Artūras Žiemys (Houston Methodist Research Institute, USA, biomedical sciences, biophysics – 02 B);

The dissertation defense is being held at the Council of Chemical Engineering science direction on 2017-09-22 15:00 in Vilnius University Life Sciences Center Institute of Biotechnology R 401 auditorium.

Address: Saulėtekio al. 7, Vilnius, Lithuania.

The dissertation summary will be mailed on 2017-08-22 or earlier.

The dissertation is available at the Vilnius University library and VU website:
www.vu.lt/lt/naujienos/ivykiu-kalendorius

VILNIAUS UNIVERSITETAS

Vytautas Raškevičius

BALTYMŲ IR MAŽŲ MOLEKULIŲ SĄVEIKOS KOMPIUTERINIS MODELIAVIMAS

Daktaro disertacijos santrauka

Technologijos mokslai, chemijos inžinerija (05T)

Vilnius, 2017

Disertacija rengta 2012-2016 metais Vilniaus universitete.

Mokslinis vadovas – dr. Visvaldas Kairys (Vilniaus universitetas, technologijos mokslai, chemijos inžinerija – 05T).

Disertacija ginama viešame Gynimo tarybos posėdyje

Pirmininkas – prof. habil. dr. Arūnas Ramanavičius (Vilniaus universitetas, fiziniai mokslai, chemija – 03P).
Nariai:
dr. Saulius Gražulis (Vilniaus universitetas, technologijos mokslai, chemijos inžinerija – 05T);
dr. Vytautas Petrauskas (Vilniaus universitetas, technologijos mokslai, chemijos inžinerija – 05T);
prof. dr. Jolanta Sereikaitė (Vilniaus Gedimino technikos universitetas, technologijos mokslai, chemijos inžinerija – 05T);
dr. Artūras Žiemys (Hiustono metodistų tyrimų institutas, JAV, biomedicinos mokslai, biofizika – 02 B).

Disertacija bus ginama viešame Gynimo tarybos posėdyje 2017 m. rugsėjo mėn. 22 d. 15 val. Vilniaus universiteto Gyvybės mokslų centro Biotechnologijos instituto R 401 auditorijoje.

Adresas: Saulėtekio al. 7, Vilnius, Lietuva.

Disertacijos santrauka bus išsiuntinėta iki 2017 m. rugpjūčio mėn. 22 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu:
www.vu.lt/lt/naujienos/ivykiu-kalendorius

TABLE OF CONTENTS

Table of contents 5
Introduction 6
METHODS 10
RESULTS AND DISCUSSION 15
2.1. Study on antiproliferative activity of α-branched α,β-unsaturated ketones against human cancer cell lines by QSAR method 15
2.2. Computational analysis of carbonic anhydrase inhibition with benzensulfonamides using several computational methods 17
2.3. E-DRAGON descriptor-based QSAR 23
2.4. PHASE atom-based 3D-QSAR 27
2.5. Calculations of selectivity from separate affinity QSAR models compared to selectivity QSAR 32
2.6. Search for lead compounds similar to metabolites using the Tanimoto index 33
CONCLUSIONS 35
LIST OF PUBLICATIONS 36
CURRICULUM VITAE 37
REZIUMĖ 38
REFERENCES 39

INTRODUCTION

One of the most important practical applications and goals of biochemistry is the treatment of diseases. Rapid advances in science and technology disciplines such as functional genomics and proteomics greatly help in solving numerous challenges. Identification of proteins that could become potential drug targets is one of these challenges. Drugs acting on them would be able cure the disease or at least to alleviate condition. Another very important challenge is the discovery of drugs acting against those targets. The majority of drugs are small-molecular-mass compounds that activate or inhibit protein targets. It is useful to inhibit only proteins of disease-causing viruses and bacteria without altering the function of important human proteins. For example, anticancer drugs target proteins that are essential for cancer cell survival.

In modern drug discovery, one of the major tasks is identification of ligands (small-molecular-mass chemical compounds) that alter protein functions and improvement of their properties. There is a demand for flexible, fast, and efficient search strategies that can be used to find ligands acting on proteins. After the lead discovery, the compounds are further improved until they enter clinical trials and can be used as drugs in case the trials are successful. Better inhibitors can be discovered by simultaneous usage of several search methods [1]. Computational methods *in silico* (Latin for “in silicon”, which indicates that modeling was carried out using computers) calculate the interactions between small molecules and macromolecules, such as proteins, and evaluate their complementarity. *In silico* methods have widely been applied in discovery and improvement of ligands. The development of a number of currently used drugs such as HIV protease inhibitors has been largely based on the application of structural methods and computational search strategies [2]. However, in spite of the success stories, computational inhibitor search methods are still far from perfect, mostly due to difficulties in computing the correct ligand binding energies.

Making the inhibitors highly specific for just one enzyme isoform is one of the biggest challenges in the development of new drugs [3]. If selectivity is poor, the inhibitor will bind into multiple targets causing undesirable side effects. Many inhibitors, currently being used as drugs, are not highly selective for one enzyme isoform, but inhibit several isoforms [4]. Quantitative Structure-Activity Relationship (QSAR) concepts have long been used in the design of such inhibitors.

Carbonic anhydrases (CAs), also known as carbonate dehydratases (EC 4.2.1.1), belong to the family of metalloenzymes. CAs catalyze a reversible reaction between carbon dioxide and water [5]:

$$CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$$

CA II is known to be one of the most active catalysts (with k_{cat}(catalytic constant)/K_m (the Michaelis constant) $= 1.5 \times 10^8$ M^{-1} s^{-1}) [4]. CA is a target for approximately 30 drugs or compounds in clinical trials. Human CA inhibitors have already been used as diuretics and antiglaucoma drugs for some decades [3].

The computational study was carried out in three stages.

During the first stage of the work, QSAR modeling was carried out using a set of 62 ketones. Most of them possess antiproliferative activity against several cancer cell lines. The data about those newly synthesized compounds and their antiproliferative activity were obtained in collaboration with the scientists of the Faculty of Chemistry and the Institute of Biochemistry, Vilnius University. We developed QSAR models quantitatively describing the antiproliferative activities of the above compounds. These QSAR models can be useful for the development of new, not synthesized yet, chemical compounds exhibiting anticancer activity.

During the second stage of the work, CAs were selected for the comparison of several computational methods. The QSAR method was applied to suggest alterations in sulphonamide structures that could improve their selectivity. 2D-QSAR and 3D-QSAR were carried out for different CA isoforms with a data set of 40 inhibitors. CA XII selectivities (affinity ratios) were used as the target variables for QSAR modeling. A more traditional QSAR protocol in which the affinities for different CA isoforms were separately modeled was also used. These two QSAR protocols were compared for each CA isoform. Some of the QSAR models were improved using a newly created molecular descriptor. The created descriptor can be further modified to develop a large number of analogous descriptors.

A study comparing the calculated binding energies computed using several different computational methods (docking, linear interaction energy (LIE), metadynamics, and QSAR) was carried out using 40 CA inhibitors as a data set. The QSAR approach showed the best results among all the tested methods. The possible rationalization of the obtained results was presented. A new LIE-like equation modifying the LIE method was suggested, and this resulted in improved results compared to the original LIE method.

During the third stage of work we undertook a search for new cancer growth inhibitors. The theory was tested that it is possible to find new cancer growth inhibitors by searching for compounds that are similar to certain human metabolites. The Tanimoto chemical similarity index was used to search for drug candidates in large databases of purchasable compounds. A probability to find new inhibitors using this procedure was estimated by computing the Tanimoto similarity index between the known drug structures and human metabolites. Several compounds were suggested as cancer cell growth inhibitors, and two of them experimentally exhibited a previously unknown anticancer activity.

Aim of the study

The aim of this study was to search for new biologically active compounds using various theoretical approaches as well as to compare and possibly improve the computational methods used with a particular attention given to the QSAR method.

Specific objectives:

1. To develop QSAR models for a series of 62 α-branched α,β-unsaturated ketones with antiproliferative activity for three cancer cell lines (NB4, MCF–7, and A549).

2. To compare docking, linear interaction energy (LIE), metadynamics, and QSAR methods and to determine which of them is the most suitable for the discovery of CA II inhibitors.

3. To modify and improve the LIE approach so that it would become more applicable to the development of CA II inhibitors.

4. With the help of different approaches, to develop affinity and CA isoform selectivity QSAR models for a series of benzensulfonamide compounds with an attached pyrimidine ring.

Scientific novelty

Comparative and computational research on binding constants K_d and structures of 40 CA II inhibitors, i.e. benzensulfonamide derivatives with an attached pyrimidine ring, was carried out. Docking, LIE, metadynamics, and QSAR methods were used for the comparison. The computed K_d values were compared with the experimental data. This enabled the evaluation of advantages and disadvantages of the methods used when applied for the CA II receptor and its inhibitors. The best results were obtained with QSAR (coefficient R^2 between the experimental and predicted pK_d values varied from 0.83 to 0.89). Possible reasons for the observed results were explored. Moreover, a new improvement for the LIE method was suggested. With the proposed energy estimation equation, LIE results noticeably improved (R^2 between the experimental and predicted ΔG_{bind} values improved from 0.24 to 0.50).

Using the antiproliferative activity data of α-branched α,β-unsaturated ketones against NB4, MCF–7, and A549 cancer cell lines, several QSAR models using 5 descriptors were developed. They can be used to predict and explain the antiproliferative activities of compounds belonging to this class.

A probability to find novel inhibitors with a novel inhibitor search strategy by employing the Tanimoto similarity index between compound candidates and metabolites was estimated. This strategy led to a discovery of two compounds inhibiting cancer cell growth, using hitherto unknown modes of action.

METHODS

1.1. QSAR studies on antiproliferative activity of α-branched α,β-unsaturated ketones against human cancer cell lines

The QSAR method for a series of α-branched α,β-unsaturated ketones was computationally explored. A total data set of 62 molecules was divided into two sets of chemical compounds based on the presence of the terminal alkene group: Set 1 set without the terminal alkene group (35 molecular structures) and Set 2 with the terminal alkene group (27 molecular structures). The IC_{50} activities of compounds were converted to $pIC_{50} = -\log(IC_{50})$ for the sake of convenience. For computational purposes, compounds with IC_{50} of >100 μM (inactive) were assumed to have pIC_{50} equal to 4. Compound structures were sketched and afterward optimized in 3D using the AVOGADRO program (v. 1.1.1) [6]. In Set 1, the activity data of some compounds were determined for racemic mixtures, with unknown activities for individual enantiomers. For all these molecules, the same chirality was used to prevent introduction of noise. The generated molecular structure files were merged into a single sdf format file with the OPEN BABEL program (v.2.3.1) [7]. The combined sdf file was used as an input to the E-DRAGON program [8]. E-DRAGON calculates numerous descriptor families [9]. The resulting pool consisted of 1666 descriptors. Linear regression QSAR models were developed using the LEAPS package within R software environment. LEAPS warned about issues arising from the linear dependence of the descriptors, and such QSAR models were rejected. The final QSAR equations were built using 5 descriptors according to the generally accepted rule that at least 5 molecules should be present for each selected descriptor [10].

1.2. Benzensulfonamide data set

A total of 40 benzensulfonamides with experimental affinity data published previously [11] were used as the data set for the input. Of the 8 available PDB structures, 8 initial ligand structures were extracted in the crystallographic binding mode after carrying out protein structure alignment with PYMOL v.1.7.0.0. Hydrogen atoms were added with the AVOGADRO program (v.1.1.1) [6]. The structures of the remaining compounds were sketched and afterward optimized in 3D using AVOGADRO and maintaining the same alignment with the 3D structures of the initially prepared ligands. The separate molecular structure files were merged into a single sdf format file using OPEN BABEL (v.2.3.1) [7]. For further calculations, K_d values were converted into pK_d values (p$K_d = -\log(K_d)$).

Selectivity (p$K_{d,diff}$) toward CA XII was defined as the difference between the pK_d of CA XII and another CA. Combined selectivity for CA XII (∑CA XII) was defined simply as a sum of all selectivities for that compound. Even if a compound is highly selective for CA XII against just one other CA and not selective for CA XII when compared against the remaining CAs, it is still defined as selective for CA XII in general.

1.3. Docking

The same sdf file of chemical structures of CA II inhibitors was used as the input to carry out ligand-protein docking with ARGUSLAB [12]. The CA II structure (PDB ID: 3MYQ) prepared using the UCSF CHIMERA 1.8 "Dock Prep" procedure [13] was used as a docking target. A box on the docking target was set up to cover the extent of the known binding site. Docking precision was set to high precision, and the flexible ligand docking mode was employed. The same ligand set then was used as the input to carry out molecular docking with GLIDE [14] using the MAESTRO 2014.2 software package. The CA II receptor and all ligands were prepared using built-in preparation tools. Docking was performed using the standard default settings mode except donor-acceptor bonds between Zn^{2+} in CA II and the ligand. The donor-acceptor bond between the ligand and Zn^{2+} was not enforced by the program. Docking precision was set to extra precision.

1.4. Linear Interaction Energy

All molecular dynamics (MD) calculations for LIE were performed using GROMACS (v.4.6.2) [15]. The CA II structure (PDB ID: 3MYQ) was prepared using the UCSF CHIMERA 1.8 Dock Prep procedure [13]. The Amber ff99SB protein force field [16] and the general AMBER force field (GAFF) [17] were used for the protein and the ligand, respectively. To prevent the ligand from traveling away from the binding site during the MD simulation, a coordination bond was enforced between Zn and sulfonamide nitrogen. The non-bonded parameters R and ε for Zn were taken from ref. [18]. The formal charge on Zn was set to +1 in accordance with quantum chemical calculations [18]. The parameters for bonds, angles, and dihedrals for the zinc ion connected to the surrounding residues and the sulfonamide nitrogen atom were taken from references [19] and [20]. Na^+ or Cl^- counter ions were added to neutralize the system. The time step of the simulations was 2 fs. All simulations were carried out at 300 K. Prior to MD simulations, 800-step minimization and 80-ps equilibrations with the constrained protein heavy atoms were carried out. Two sets of 1-ns production runs were performed: one on a protein-ligand complex in water and the second on a separate ligand in a water box. During the production runs, pressure coupling was applied using the Parrinello-Rahman algorithm [21, 22], and temperature coupling was done with the V-rescale algorithm [23]. The particle-mesh Ewald [24] algorithm was used for long-range electrostatic interactions, with a cutoff of 9 Å. The same cutoff of 9 Å was used for Van der Waals interactions. The ΔG_{bind} was approximated using standard or modified LIE equations.

1.5. Metadynamics

The system was prepared for metadynamics calculations using the same procedure as for the LIE simulations. Metadynamics was performed using the PLUMED plugin (v.2.1.0) [25] compiled together with GROMACS (v.4.6.7). The donor-acceptor bond between the Zn ion and the ligand was not built, because the metadynamics procedure required a freely moving ligand being able to move away from the binding site. We assumed that the absence of the bond between the ligand and zinc should not pose a critical problem for the purposes of the ligand ranking. Metadynamics potential was set to act on only one distance collective variable (CV) between the Zn ion in CA II and the charged ligand nitrogen atom in the sulfonamide group. The rate of Gaussian deposition was set to one Gaussian every 500 steps, while the height and width of the Gaussian were set to 2 kJ/mol and 0.04 nm, correspondingly. The length of production runs was set to 4 ns in order to give enough time for the bias potential to fill the binding pocket. The PLUMED utility *sum_hills* was used to estimate the free energy as the function of metadynamics CVs directly from metadynamics bias potential.

1.6. Carbonic anhydrase II inhibitor QSAR

The same set of chemical structures of 40 CA II inhibitors as the sdf file was used as the input to compute molecular descriptors with E-DRAGON [8, 9]. The resulting pool consisted of 1666 structural descriptors. Appropriate descriptors were selected, and multiple linear regression QSAR models were developed using the LEAPS package within the R software environment. LEAPS warned about issues arising from the linear dependence of descriptors, and such QSAR models were rejected. The final QSAR equations were built using a data training set of 30 inhibitors, and the 10 remaining inhibitors were used as a test set for the model validation. Descriptors were included into the model according to the generally accepted rule that at least 5 chemical compounds with experimental data should be taken for each selected descriptor [10]. Mean absolute errors (MAE) were calculated according to the procedure described in [26]. The Applicability Domain is a QSAR model-relevant approach, which represents the chemical space from which a model is derived and where a prediction is considered to be reliable [27]. Here we used a simple standardization approach introduced by Roy et al. to determine the Applicability Domain of our 2D-QSAR models [28]. This approach is used to define outliers in the training set and the compounds residing outside the Applicability Domain in the test set of the built QSAR models.

1.7. PHASE atom-based 3D-QSAR

PHASE 3D-QSAR studies for CA affinities and CA XII selectivities were carried out using the PHASE v.4.4 module in the MAESTRO 10.3 molecular modeling package from Schrödinger, LLC [29, 30]. PHASE QSAR models may be atom-, field-, or property-based. The difference among them is that either all the atoms of the compounds are taken into molecular modeling or different pharmacophores, which confirm the hypothesis about what features determine biological activity. Structures of compounds were realigned using a common scaffold alignment in MAESTRO (Schrödinger, LLC). Atom-based 3D-QSAR models were built by correlating the actual and predicted pK_d or selectivities for a training set of 30 compounds using Partial Least Square (PLS) regression. The validation of the models was performed using the Leave

One Out (LOO) method, and by using a test set of 10 compounds. The test set for all calculations was kept the same in order to compare different QSAR protocols. PLS regression was carried out with a maximum number of *n/5* PLS factors (*n* = number of ligands in the training set) and a grid spacing of 1.0 Å. Validation of all models was performed by predicting pK_d or selectivity of the test set compounds. We used statistical metrics, and the X-ray crystal structures representing the active sites of CAs for the model validation. The Applicability Domain of PHASE models was not determined as there was no way to extract the molecular descriptors from the software.

RESULTS AND DISCUSSION

2.1. Study on antiproliferative activity of α-branched α,β-unsaturated ketones against human cancer cell lines by QSAR method

Using the activity data against the NB4, A549, and MCF-7 targets, several QSAR models were developed for compound Sets 1 and 2. The best models are presented below:

$$\mathrm{p}IC_{50}(NB4,Set1) = 5.111\,MATS1m - 1.83\,MATS5p - 2.47\,HOMA + 14.1\,E2u - 10.4\,E2a + 4.27 \qquad (1)$$

$R^2 = 0.86$, $R^2_{ADJ} = 0.83$, $Q^2 = 0.75$, $F(5,29) = 35.1$, $p < 10^{-5}$

$$\mathrm{p}IC_{50}(A549,Set1) = 0.968\,ATS6e + 1.55\,MATS8m + 0.370\,Mor05v - 0.484\,Mor11p + 0.765\,Mor20p + 1.59 \qquad (2)$$

$R^2 = 0.81$, $R^2_{ADJ} = 0.78$, $Q^2 = 0.75$, $F(5,29) = 25.3$, $p < 10^{-5}$

$$\mathrm{p}IC_{50}(MCF\text{-}7,Set1) = 0.0457\,RDF050u - 0.119\,RDF035p - 0.620\,Mor20m - 3.20\,H7u - 9.44\,HATS2u + 8.73 \qquad (3)$$

$R^2 = 0.92$, $R^2_{ADJ} = 0.90$, $Q^2 = 0.88$, $F(5,29) = 64.8$, $p < 10^{-5}$

$$\mathrm{p}IC_{50}(NB4,Set2) = -5.17\,MATS2m + 5.25\,MATS2e - 3.90\,E3u + 8.17\,G2s - 1.29\,R4v + 5.69 \qquad (4)$$

$R^2 = 0.86$, $R^2_{ADJ} = 0.82$, $Q^2 = 0.77$, $F(5,21) = 24.8$, $p < 10^{-5}$

$$\mathrm{p}IC_{50}(A549,Set2) = -25.0\,X3Av + 1.34\,Mor16u + 7.19\,G3u - 8.85\,G3p - 0.140\,Htv + 7.97 \qquad (5)$$

$R^2 = 0.80$, $R^2_{ADJ} = 0.75$, $Q^2 = 0.67$, $F(5,21) = 16.4$, $p < 10^{-5}$

$\mathrm{p}IC_{50}(MCF\text{-}7,Set2) = 0.0149\,CSI - 0.109\,UNIP - 0.0475\,MPC10 + 0.815\,X2 -$

$0.909\ GATS1v + 1.11$

$$(6)$$

$R^2 = 0.89$, $R^2_{ADJ} = 0.87$, $Q^2 = 0.83$, $F(5,21) = 34.7$, $p < 10^{-5}$

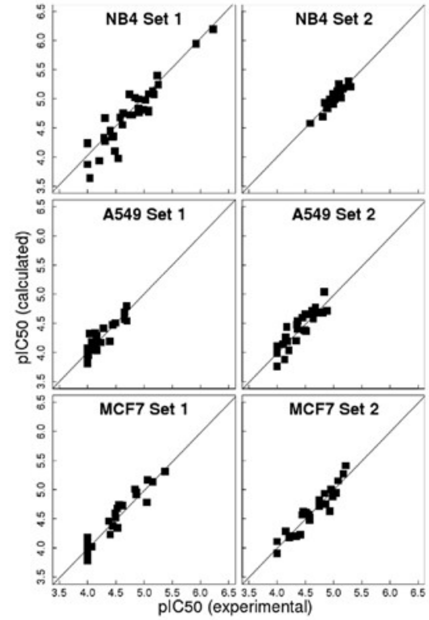

Fig. 1. Plots of the experimentally determined versus predicted pIC_{50} values. A diagonal straight line represents an ideal agreement between the experimental and calculated values. The scaling of x-and y-axes is the same for all subplots for a better comparison.

The models are described by several statistical parameters provided below each equation. The model was considered as acceptable based on correlation coefficients R^2, adjusted correlation coefficients R^2_{ADJ}, and F-test (F) values. The performance of the models was validated using the LOO method with variance Q^2. Plots of the experimentally determined versus predicted pIC_{50} values are presented in Fig. 1. A straight line represents an ideal agreement between the experimental and calculated values. It is worth noting that correlation coefficients R^2 were similar for the same cell line, regardless of the compound set, even though the QSAR equations were different. For the A549 cell line, the range of the measured pIC_{50} data for Sets 1 and 2 was only 0.7

and 0.9 log units, respectively, and this partly explains lower QSAR model quality for A549 compared to other cell lines. It was interesting to investigate how well the models were able to classify compounds as active and inactive. For the numerous points in Fig. 1 corresponding to inactive compounds with the experimental IC_{50} >100 μM (pIC_{50} = 4), QSAR models predicted pIC_{50} in the range from 3.76 to 4.25, i.e. the "true negatives" were reasonably well predicted by the QSAR equations. Only two compounds – 3ei and 3fn (Set 1, NB4 cell line) – were defined as "false negative" with the predicted pIC_{50} lower than 4.0, although the experimental pIC_{50} values were greater than 4.2, giving a relatively small error. Thus, we showed that QSAR models were able to distinguish between the active and inactive compounds, and they can be useful for designing molecules with improved antiproliferative properties.

2.2. Computational analysis of carbonic anhydrase inhibition with benzensulfonamides using several computational methods

2.2.1. Docking results

Linear regression between the experimentally determined pK_d values and the predicted energy values calculated using docking programs, namely ARGUSLAB and GLIDE, produced coefficients of determination (R^2) equal to 0.00 and 0.07, respectively. Root-mean-square deviation (RMSD) was calculated between the crystallographic binding modes and the conformations of the same ligands docked with GLIDE and ARGUSLAB. The binding modes of the 8 ligands with the known structures (1a, 1f, 1i, 1j, 2f, 2j, 4f, and 4g) were not reproduced with the GLIDE and ARGUSLAB docking programs (Fig. 2). Using GLIDE, ligand 1a was docked in a 180° reversed conformation (RMSD = 7.8 Å), and for other 6 ligands, the pyrimidine ring position was completely wrong. Only the docked conformation of ligand 2f was reasonable compared to the X-ray structure (RMSD = 2.2 Å). Similarly poor results were also obtained using ARGUSLAB. In conclusion, docking procedures with the GLIDE and ARGUSLAB poorly predicts the experimental binding modes for this system. Difficulties to correctly predict an experimental binding mode with the VDOCK program have been reported previously, and constrained docking has been suggested to solve this problem [31]. One suggestion to improve the docking procedure in cases like this, where the core structure is the same for all ligands and when at least one experimental pose is known, is as follows: to allow moving only the atoms that are different in the series and to fix the remaining (constant) part of the ligand structure in the crystallographic binding state. This could radically simplify the docking task. With the current versions of both ArgusLab and Glide docking software, it was not possible to try out this constrained docking procedure. Another issue is that half of those X-ray structures contain a dimethyl sulfoxide right inside the ligand binding pocket (PDB ID: 3SBH, 3SBI, 3S9T, 3SAX) and this may cause interference. Such artifact molecules are commonly deleted during the docking preparation procedure; thus, the docked ligand often takes a spot previously occupied by the artifact, conflicting with the experimental data.

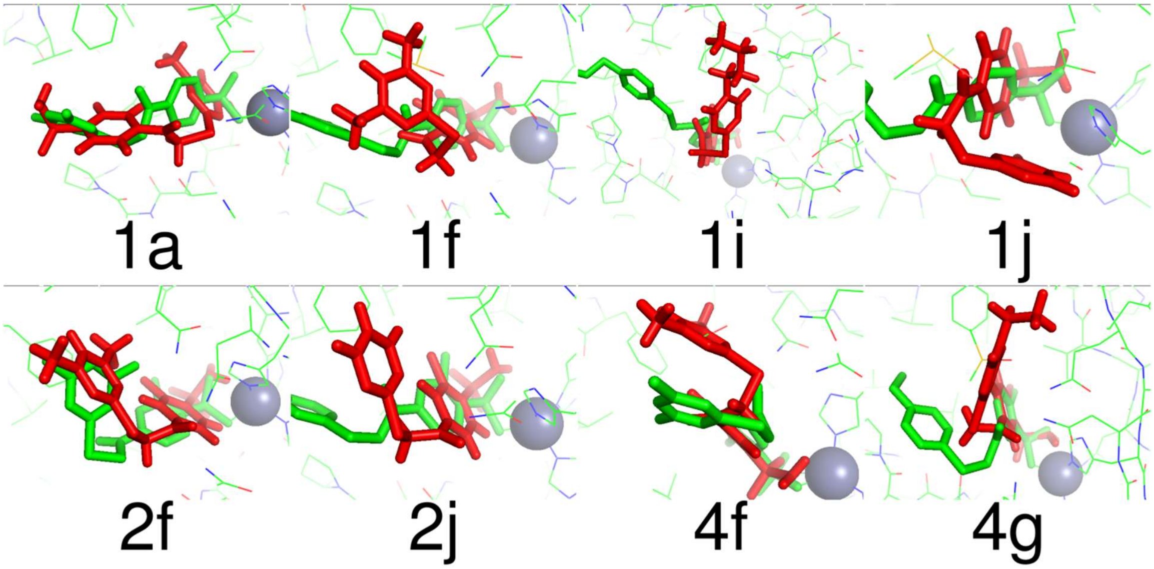

Fig. (2). Crystallographic (green) and docked (red) ligand binding modes. Zn ion is shown as a sphere.

2.2.2. LIE results

The MD simulation runs yielded data for 40 ligands. Because there was no correlation between the calculated LIE energies and the experimental affinities using the default fitting coefficients, new fitting coefficient values for the LIE method were derived as applied to CA II. The new fitting coefficients are shown in Table 1. The need to derive new fitting coefficients may arise because of the presence of the coordination bond between the ligand and the receptor in CA II. The free energy weight γ is an estimate of the effect of that bond.

Table 1. Derived fitting coefficients values for LIE.

Fitting Coefficient	Value
α	0.197
β	0.462
γ	10.4

Parameterization of the LIE equation with the new coefficients was not very successful either ($R^2 = 0.24$). While looking for an explanation for the poor performance of the LIE method, we observed that ligand intramolecular interactions play an important role in the investigated system and cannot be ignored.

Based on these observations, another approach inspired by LIE was developed to computationally approximate ΔG_{bind}. Two Coulomb and Lenard-Jones interactions were used in a similar fashion after observation that they separately show some low correlation with ΔG_{bind}:

$$\Delta G_{\text{bind}} = \alpha\langle V^{\text{el}}\rangle_{\text{lig-prot}} + \beta\langle V^{\text{vdw}}\rangle_{\text{lig-lig}} + \gamma, \qquad (7)$$

where $\langle V^{el}\rangle$ and $\langle V^{vdw}\rangle$ are the averages of electrostatic and van der Waals interactions, and the "lig-lig" and "lig-prot" subscripts refer to the ligand interacting with itself or with the protein. The parameterization of Eqn. 7 led to a better fit ($R^2 = 0.50$) compared with the typical LIE equation. The fitting coefficient values for LIE-like Eqn. 7 are shown in Table 2.

Table 2. Derived fitting coefficients values for the LIE-like method based on Eqn. 7.

Fitting Coefficient	Value
α	0.0884
β	0.689
γ	51.1

We also attempted to restrain a ligand position in order to maintain it the same as in the X-ray structure for a subset of 8 ligands with the known structures (PDB ID: 3S8X, 3S9T, 3SAP, 3SAX, 3SBH, 3SBI, 4KNI, and 4KNJ). The LIE results for the ligands with position restraints were not better compared with the results for unconstrained ligands. Interestingly, for this ligand subset, the LIE results without position restraints correlate with the experimental data ($R^2 = 0.54$). It shows that the correct initial ligand pose is beneficial for LIE calculations. It also shows that position restraints are not beneficial for LIE calculations.

2.2.3. Metadynamics results

The metadynamics simulation runs yielded data for a set of 40 ligands. The ligand leaving the protein binding pocket was observed during the production MD runs when simulation was biased with metadynamics potential. Estimates of free energy as the function of the distance between the Zn ion in CA II and sulfonamide nitrogen in the ligand were obtained for all ligands from 4-ns metadynamics simulations. An estimated difference in free energy between the bound and free ligand states did not correlate with the experimental data ($R^2 = 0.00$). The location of the energy minimum did not match the X-ray bound state, and for 17 most problematic ligands the energy minimum was in the bulk solvent (water). There could be a possibility that our initial assumption was wrong. Another important observation is that the ligand shifts away from its X-ray position in the binding pocket during the MD equilibration phase, even before the metadynamics production run, and this can cause problems both with the metadynamics and the LIE calculations. Therefore, the MD equilibration procedure or perhaps even the force field may need an improvement, because the ligand pose theoretically should remain stable during the short equilibration, and the ligand should leave the binding pocket only during metadynamics calculations.

2.2.4. QSAR results

Using the activity data against CA II, several QSAR models were developed. Three best models are presented below:

$$\mathrm{p}K_d(\mathrm{CA\ II}) = 2.05\ MATS7m + 1.02\ H5u + 4.49\ HATS8m + 4.86,$$

$$(8)$$

$R^2 = 0.89$, $R^2_{ADJ} = 0.87$, $Q^2 = 0.87$, $F(3,26) = 68$, $p < 10^{-9}$, $R^2_{TEST} = 0.57$

$$\mathrm{p}K_d(\mathrm{CA\ II}) = 0.0867\ G(N..S) + 1.23\ H5u + 68.7\ (R7p+) + 2.19, \tag{9}$$

$R^2 = 0.86$, $R^2_{ADJ} = 0.84$, $Q^2 = 0.83$, $F(3,26) = 53$, $p < 10^{-9}$, $R^2_{TEST} = 0.61$

$$\mathrm{p}K_d(\mathrm{CA\ II}) = 0.0106\ VAR + 2.26\ MATS7m - 2.65\ MATS8p + 4.30, \tag{10}$$

$R^2 = 0.83$, $R^2_{ADJ} = 0.81$, $Q^2 = 0.79$, $F(3,26) = 43$, $p < 10^{-9}$, $R^2_{TEST} = 0.63$

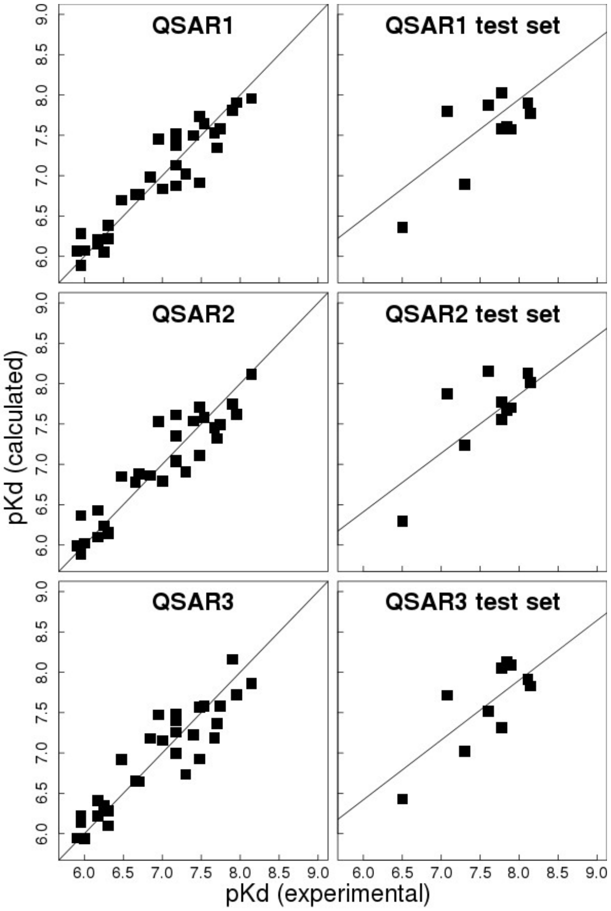

Fig. (3). QSAR plots of the experimentally determined versus predicted pK_d values.

The statistical parameters of the models are provided below each equation.

The models were considered acceptable based on high values of coefficients of determination R^2, adjusted coefficients of determination R^2_{ADJ}, and F-test (F). The performance of the models was validated using the LOO method with variance Q^2 and with a test set of 10 ligands. The plots of experimentally determined versus predicted pK_d values in all 3 QSAR models, using both training and test sets for all of them, are presented in Fig. (3).

2.2.5. Comparison of docking, LIE, metadynamics, and QSAR results

Five methods (ARGUSLAB docking, GLIDE docking, LIE, metadynamics, and QSAR) were applied to predict binding energy of the 40 ligand set to CA II with the experimentally determined pK_d values (Table 3). Of the 5 tested methods, only QSAR showed a sufficiently good correlation. One possible reason is that some ligands in the set have different tautomeric forms, and there are no available experimental data showing which form or forms are active in the binding site. This situation hinders all methods that strongly depend on ligand structure. We suggested an improvement for the LIE method in the form of modified Eqn. 7 that led to a significant improvement. We called this approach LIE-like. Another suggestion for docking, using a constrained scaffold, was not possible to test with the used docking software. The QSAR method most effectively helped quantify the subtle empirical relationship between the structure and the activity for the CA II target. The developed QSAR models have a potential to make predictions leading to the synthesis of novel ligands.

Table 3. Coefficients of determination R^2 between the computed binding affinities and the experimentally determined pK_d using various methods.

Method	R^2
LIE	0.24
New LIE-like	0.50
QSAR	0.83-0.89
ARGUSLAB docking	0.00

GLIDE docking	0.07
Metadynamics	0.00

2.3. E-DRAGON descriptor-based QSAR

Several QSAR models specifically targeting selectivities of the benzensulfonamides toward CA XII were developed. The best models for the compound selectivities for CA XII versus other isoforms are presented below:

$$\mathrm{p}K_{d,diff}(\mathrm{CA\ XII} - \mathrm{CA\ I}) = -0.566\ T(N..N) + 0.0733\ EPS0 + 21.0\ (R7m+) + 7.14, \tag{11}$$

$R^2 = 0.84$, $R^2_{ADJ} = 0.83$, $Q^2_{LOO} = 0.82$, $F(3,26) = 47$, $p < 10^{-9}$, $R^2_{TEST} = 0.79$, $Q^2_{TEST} = 0.67$, $MAE = 0.33$

$$\mathrm{p}K_{d,diff}(\mathrm{CA\ XII} - \mathrm{CA\ II}) = 8.76\ MATS1p - 1.41\ GATS8m - 1.03\ GATS3v + 4.57, \tag{12}$$

$R^2 = 0.81$, $R^2_{ADJ} = 0.79$, $Q^2_{LOO} = 0.77$, $F(3,26) = 38$, $p < 10^{-8}$, $R^2_{TEST} = 0.58$, $Q^2_{TEST} = 0.53$, $MAE = 0.28$

$$\mathrm{p}K_{d,diff}(\mathrm{CA\ XII} - \mathrm{CA\ VI}) = -12.1\ GNar + 5.01\ MATS7v - 2.74\ MATS7p + 23.5, \tag{13}$$

$R^2 = 0.82$, $R^2_{ADJ} = 0.80$, $Q^2_{LOO} = 0.78$, $F(3,26) = 39$, $p < 10^{-9}$, $R^2_{TEST} = 0.42$, $Q^2_{TEST} = 0.22$, $MAE = 0.19$

$$\mathrm{p}K_{d,diff}(\mathrm{CAXII} - \mathrm{CA\ VII}) = 2.04\ MATS8p + 53.3\ BELe5 - 0.214\ Mor04m - 107, \tag{14}$$

$R^2 = 0.83$, $R^2_{ADJ} = 0.81$, $Q^2_{LOO} = 0.78$, $F(3,26) = 42$, $p < 10^{-9}$, $R^2_{TEST} = 0.49$, $Q^2_{TEST} = 0.46$, $MAE = 0.45$

$$\mathrm{p}K_{d,diff}(\mathrm{CA\ XII} - \mathrm{CA\ XIII}) = -12.7\ MATS2v + 2.29\ MATS8p - 0.866\ H4p + 3.43, \tag{15}$$

$R^2 = 0.78$, $R^2_{ADJ} = 0.75$, $Q^2_{LOO} = 0.74$, $F(3,26) = 30$, $p < 10^{-7}$, $R^2_{TEST} = 0.71$, $Q^2_{TEST} =$

0.68, $MAE = 0.36$

$$\mathrm{p}K_{d,diff}(\Sigma\,\mathrm{CA\ XII}) = -5.40\ MATS8e + 0.308\ RDF055m - 41.9\ HATS8p - 0.0513, \tag{16}$$

$R^2 = 0.78$, $R^2_{ADJ} = 0.75$, $Q^2_{LOO} = 0.74$, $F(3{,}26) = 31$, $p < 10^{-7}$, $R^2_{TEST} = 0.58$, $Q^2_{TEST} = 0.56$, $MAE = 1.34$

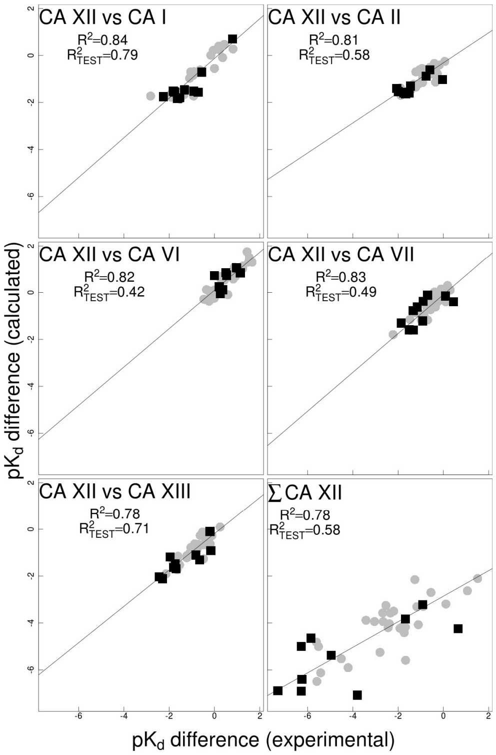

Fig. (4). Plots of selectivity-targeted QSAR using E-DRAGON descriptors. The x- and y-axes contain the experimental and calculated selectivities, respectively, for the training (gray circles) and the test set (black squares).

Statistical parameters for the selectivity-targeted QSAR models are provided below each equation. The models were considered acceptable based on high values of coefficients of determination R^2 and adjusted coefficients of determination R^2_{ADJ}. The predictive performance of the models was validated using the LOO method with variance Q^2 and by calculating R^2_{TEST}, Q^2_{TEST}, and MAE for a test set of 10 ligands. The plots of experimentally determined versus predicted pK_d differences in all computed QSAR models, using both training and test sets, are presented in Fig. (4).

Next, we also developed QSAR models using affinities for each of the individual isoforms. This resulted in the following QSAR models:

$$\mathrm{p}K_d(\mathrm{CA\ I}) = 1.14 \cdot MATS8e + 5.47 \cdot GATS5v - 2.98 \cdot GATS7p + 4.47, \qquad (17)$$

$R^2 = 0.94$, $R^2_{ADJ} = 0.93$, $Q^2_{LOO} = 0.93$, $F(3,26) = 138$, $p < 10^{-15}$, $R^2_{TEST} = 0.89$, $Q^2_{TEST} = 0.87$, $MAE = 0.21$

$$\mathrm{p}K_d(\mathrm{CA\ II}) = 2.05\ MATS7m + 1.02\ H5u + 4.49\ HATS8m + 4.86, \qquad (18)$$

$R^2 = 0.89$, $R^2_{ADJ} = 0.87$, $Q^2_{LOO} = 0.87$, $F(3,26) = 68$, $p < 10^{-11}$, $R^2_{TEST} = 0.57$, $Q^2_{TEST} = 0.48$, $MAE = 0.31$

$$\mathrm{p}K_d(\mathrm{CA\ VI}) = -0.883\ GATS5m + 0.264\ H2m + 13.1\ (R7m+) + 5.04, \qquad (19)$$

$R^2 = 0.79$, $R^2_{ADJ} = 0.76$, $Q^2_{LOO} = 0.72$, $F(3,26) = 32$, $p < 10^{-8}$, $R^2_{TEST} = 0.77$, $Q^2_{TEST} = 0.56$, $MAE = 0.20$

$$\mathrm{p}K_d(\mathrm{CA\ VII}) = 0.104\ T(S..Cl) - 6.54\ PCR + 1.59\ MATS7v + 15.2, \qquad (20)$$

$R^2 = 0.89$, $R^2_{ADJ} = 0.88$, $Q^2_{LOO} = 0.86$, $F(3,26) = 70$, $p < 10^{-11}$, $R^2_{TEST} = 0.87$, $Q^2_{TEST} = 0.74$, $MAE = 0.26$

$$\mathrm{p}K_d(\mathrm{CA\ XII}) = 1.47\ T(O..Cl) + 49.0\ MATS7m + 0.0473\ H6e + 4.24, \qquad (11a)$$

$R^2 = 0.82$, $R^2_{ADJ} = 0.80$, $Q^2_{LOO} = 0.78$, $F(3,26) = 40$, $p < 10^{-9}$, $R^2_{TEST} = 0.43$, $Q^2_{TEST} = 0.16$, $MAE = 0.44$

$$\mathrm{p}K_d(\mathrm{CA\ XII}) = 1.47\ H6e + 49.0\ (R7p+) + 0.0473\ T(OH..Cl) + 4.24, \qquad (21b)$$

$R^2 = 0.82$, $R^2_{ADJ} = 0.80$, $Q^2_{LOO} = 0.77$, $F(3,26) = 40$, $p < 10^{-9}$, $R^2_{TEST} = 0.60$, $Q^2_{TEST} = 0.51$, $MAE = 0.24$

$$pK_d(\text{CA XIII}) = 18.0\ EEig01r + 0.219\ G(N..S) + 0.0380\ Mor02u - 76.5, \qquad (22)$$

$R^2 = 0.88$, $R^2_{ADJ} = 0.86$, $Q^2_{LOO} = 0.85$, $F(3,26) = 63$, $p < 10^{-11}$, $R^2_{TEST} = 0.68$, $Q^2_{TEST} = 0.49$, $MAE = 0.32$

The statistical parameters of the QSAR models provided below each equation are the same as for Eqns. 11-16. The plots of experimentally determined versus predicted pK_d values in all individual affinity QSAR models, for both training and test sets, are presented in Fig. (5).

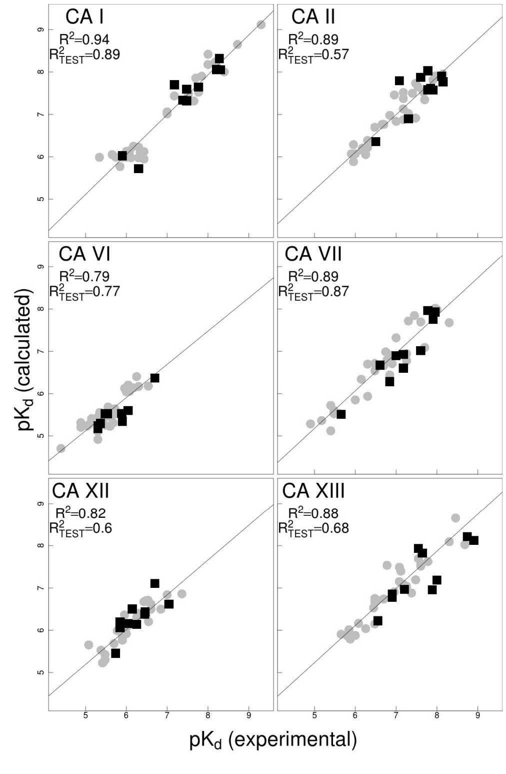

Fig. (5). Plots of affinity-targeted QSAR using E-DRAGON descriptors. The experimental and calculated affinities are plotted on the x- and y-axes, respectively, for the training (gray circles) and the test set (black squares). Six compounds (1d, 1g, 2g, 1h, 3h, and 1i) have the same K_d value for CA VI that equals to 5000 nM.

Affinity models of acceptable quality were developed using the original E-DRAGON descriptors for all CAs except CA XII (Eqn. 21a). Due to an unsatisfactory value of R^2_{TEST} in Eqn. 21a (0.36), we created a new ad hoc descriptor T(OH..Cl) designed to improve the quality of the CA XII QSAR affinity model (Eqn. 21b). It was defined as the sum of the topological distances between hydroxyl groups and chlorine atoms in the molecule. This descriptor is analogous to T(O..Cl) used in Eqn. 21a, except that in the latter the sum of topological distances between any oxygen atom and chlorine atom is used. T(OH..Cl) discerns better between hydroxyls and other oxygen atoms compared to T(O..Cl). This encouraging result may justify a more widespread use of similar, more refined "pharmacophore-like" descriptors for problems at hand. However, the new descriptor did not improve other QSAR models.

Interestingly, 6 compounds (1d, 1g, 2g, 1h, 3h, and 1i) have the same value of K_d (5000 nM) for CA VI. The QSAR approach found a descriptor R7m+, the value of which is approximately the same for these compounds (0.058-0.072) and varying more widely for the remaining compounds (0.045-0.120). All compounds in the training set were considered non-outliers and all compounds in the test set were shown to be inside the Applicability Domain of all QSAR models obtained, except compound 2j was determined to be an outlier in the CA II affinity model (Eqn. 18).

The MAE-based criterion for the developed 2D-QSAR models for the test set was found to be less than 0.15·(training set range) in all cases except ∑CA XII (Eqn. 16) and more than 0.25·(training set range)-3·σ in all cases except CA I (Eqn. 17). The ∑CA XII QSAR model (Eqn. 16) is totally unacceptable according to the MAE criterion.

2.4. PHASE atom-based 3D-QSAR

PHASE 3D-QSAR models were developed using either the selectivity data for CA XII vs. other isoforms (affinity ratios) or the affinity data. Differently from the E-DRAGON-based protocol, only two QSAR models of acceptable quality were obtained considering built-in statistical data: one for CA XII/CA I selectivity and one for CA I affinity.

The PHASE QSAR statistical parameters for the CA XII/CA I selectivity model were as follows: SD = 0.45, R^2 = 0.75, R^2 *CV* = 0.69, R^2 Scrambled = 0.27, Stability = 0.99, $F = 86$, $p < 10^{-9}$, RMSE = 0.52, $Q^2 = 0.61$, Pearson r = 0.82, and MAE = 0.39. The CA XII/CA I selectivity model was poor according to the MAE-based criterion that was calculated separately (not by PHASE), because compounds 1c and 4e in the test set had absolute prediction errors of 0.99 and 0.83 log units, respectively, in this model. The analogous statistical parameters for the CA I affinity model were as follows: SD = 0.40, $R^2 = 0.86$, R^2 CV = 0.82, R^2 Scrambled = 0.21, Stability = 0.99, $F = 168$, $p < 10^{-12}$, RMSE = 0.38, $Q^2 = 0.75$, Pearson r = 0.89, and MAE = 0.35.

In order to better understand PHASE results regarding CA I and CA XII selectivity, we explored interactions within the binding site of the two isoenzymes by employing manual docking using PyMOL 1.7.4.0. The isoenzymes CA I and CA XII were aligned, and the most CA I active compound 2d was manually placed in the active site based on the alignment with the available X-ray structures for the investigated series. The *tert*-butyl substituent on the pyrimidine ring of 2d is in contact with the hydrophobic Ala132, Ala135 and Leu131 side chains of CA I, and the hydroxyl substituent remains exposed to the water solvent (Fig. (6)). In contrast, CA XII has Ser130, Ser133, and Ala129 residues in the homologous positions and does not well accommodate the hydrophobic *tert*-butyl group of the ligand. There is a possibility that the pyrimidine ring could flip, and in that case, the hydroxyl group would make better contacts with these three residues in CA XII, but then the hydrophobic *tert*-butyl group would become unfavorably exposed to water.

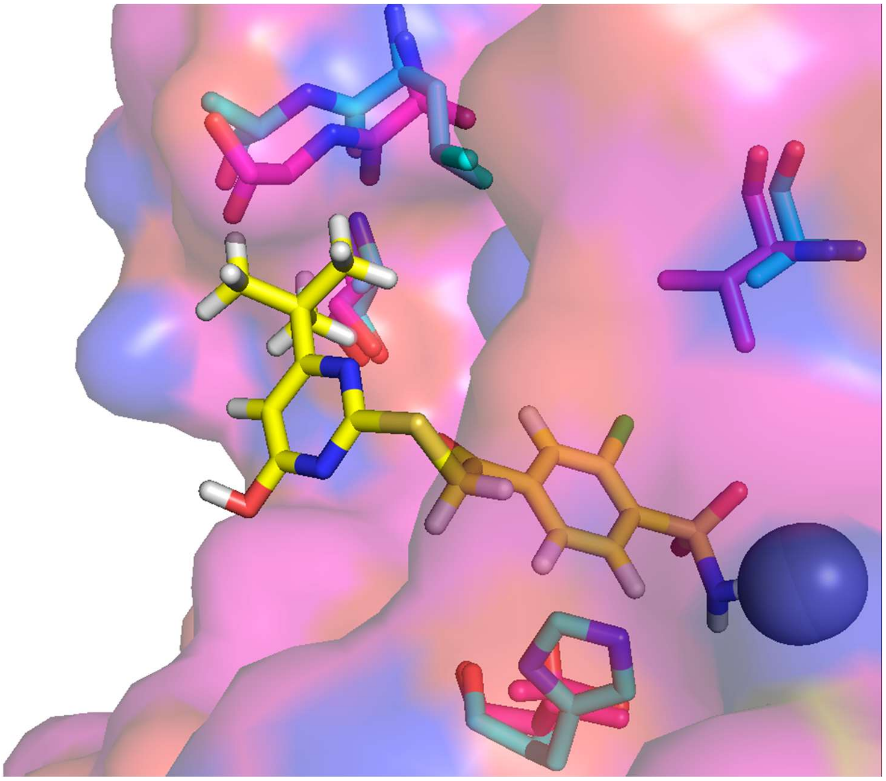

Fig. (6). The most CA I active compound 2d (yellow sticks) manually docked into the aligned CA I (azure, PDB ID: 4WR7) and CA XII (purple, PDB ID: 4KP8) receptors. Zinc ion is shown as a sphere.

PHASE in combination with MAESTRO software allows visualization of the impact of the various ligand groups on the target function (in this case, affinity or selectivity), based on the given QSAR model. The PHASE QSAR model for CA XII/ CA I selectivity is shown in Fig. (7), using the most selective compound 4c as an example. According to Fig. (7), the three main factors that affect the CA XII/CA I selectivity for 4c are as follows: the chlorine atom in the benzene ring at the *para* position with respect to the linker (blue zone in the lower part of Fig. (7c)), the neighboring sulfonamide group at the *meta* position (blue zone in the lower part of Fig. (7a) and (7b)), and to a lesser extent, the hydrophobic substituent in the pyrimidine ring (*meta* with respect to the linker) (top part of Fig. (7c)).

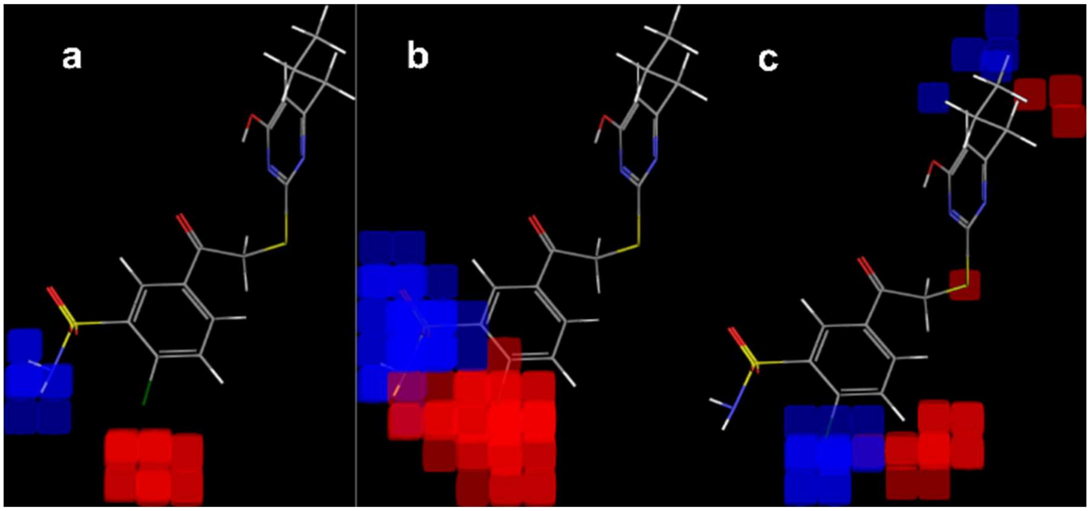

Fig. (7). PHASE atom-based 3D-QSAR for the CA XII/CA I selectivity model visualized in the context of the most selective compound 4c. Blue and red cubes depict favorable and unfavorable regions, respectively. (a) H-bond donors, (b) electron-withdrawing atoms (including H-bond acceptors), and (c) hydrophobic/non-polar groups. The cube coefficient visualization threshold was set to $\pm 2.5 \cdot 10^{-3}$.

The PHASE/Maestro visualization of the CA I affinity model using the most active compound 2d is shown in Fig. (8). Interestingly, Fig. (8) in several ways presents the "inverse" of Fig. 7: a blue zone in one visualization often corresponds to a red zone in the other, and vice versa. The blue zones near the benzene ring in Figs. 7(a,b) and 8(a,b) reflect sulfonamide positions in the best series 4 and 2 for the corresponding QSAR models. In agreement with the qualitative picture in Fig. (6), Fig. (8) shows the importance of the hydrophobic and hydrophilic substituents on the pyrimidine ring for the CA I affinity. One of the apparent drawbacks of ligand-only QSAR is that it does not take into account the variability of the actual binding modes. In this particular case, PHASE aligns benzene and pyrimidine rings as well as the linker. In reality, the most spatially constrained part of the ligand is the sulfonamide group that is attached to zinc, and the linker with the pyrimidine ring at the end may adopt several rather widely differing conformations [32].

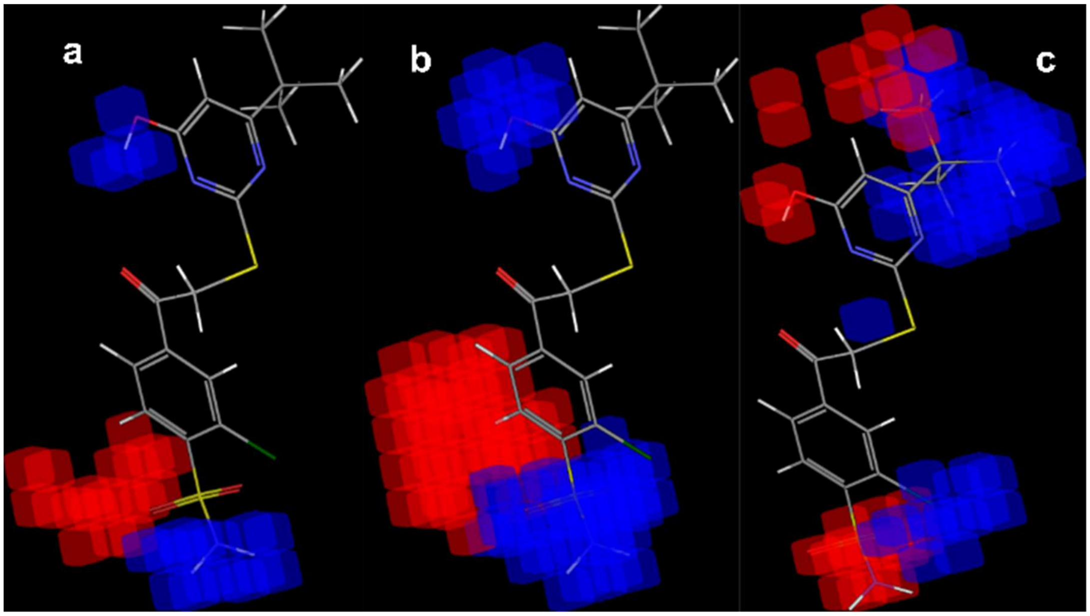

Fig. (8). The PHASE atom-based 3D-QSAR affinity model for CA I visualized in the context of the most CA I active compound 2d. Blue and red cubes depict favorable and unfavorable regions, correspondingly. (a) H-bond donors (b) electron-withdrawing atoms (including H-bond acceptors), (c) hydrophobic/non-polar groups. The cube coefficient visualization threshold was set to $\pm 2.5 \cdot 10^{-3}$.

For this reason, PHASE cannot always address the actual reasons determining the selectivity. For example, the constrained position of sulfonamide in series 4 causes the tail of the compound to move to a different zone of the binding pocket (not shown) compared to other series. In CA I, the linker part of the ligand interacts with the bulky His200 side chain (CA XII has Thr199 in the homologous position). This leads to an improved CA XII/CA I selectivity for series 4.

The lack of the significant QSAR contributions around the pyrimidine ring for the CA XII/CA I selectivity model (top part of Fig. (7), cf. Fig. (8)) also shows that the selectivity is mostly caused by the substituents of the benzene ring in agreement with the argument above. However, this also weakens the prospects of improving the substituents of the pyrimidine ring, targeting CA XII/CA I selectivity. If the binding modes between the ligands are very different, superposition of the ligand fragments loses part of its meaning: incorrectly superposed ligands may lead to a superposition of groups that could be in reality in different areas of space.

While PHASE generated a good quality QSAR model for CA XII/CA I, attempts to develop an acceptable QSAR model of CA XII inhibition by means of PHASE failed. Even though a good CA XII QSAR model was not developed, one can presumably use the CA XII/CA I selectivity model to help target CA XII affinity. The E-DRAGON descriptor-based QSAR results in much better models because E-DRAGON has many more criteria/descriptors used to build a model. Moreover, the variability of the binding modes within the series may have an effect on model quality. A great advantage of PHASE is that the influence of factors on affinities/selectivities can be easily visualized, while many E-DRAGON descriptors are somewhat obscure and not easy to understand.

2.5. Calculations of selectivity from separate affinity QSAR models compared to selectivity QSAR

In Sections 2.3 and 2.4, the selectivity was pre-calculated from the affinity data and then fed into the software as a target variable to compute the QSAR models. It is also possible to predict selectivity by calculating the predicted individual isoform affinities from the individual isoform affinity QSARs and afterward computing their ratio. Table 4 shows comparison of the R^2_{TEST} values for the selectivities computed using both methods for the E-DRAGON based descriptors.

Table 4. R^2_{TEST} values of the E-DRAGON descriptor-based QSAR test set for all CA XII binding selectivity models and for CA XII selectivity computed from separate affinity QSAR models. R^2_{TEST} values of > 0.40 are shown in bold.

	E-DRAGON descriptor-based QSAR R^2_{TEST} values: selectivity computed from separate affinity QSAR models/selectivity QSAR models
CA XII/CA I	0.85/0.79
CA XII/CA II	0.34/0.58

CA XII/CA VI	0.28/0.42
CA XII/CA VII	0.58/0.49
CA XII/CA XIII	0.65/0.71
∑CA XII	0.63/0.58

Using E-DRAGON descriptor-based selectivity QSAR, the acceptable R^2_{TEST} values were obtained in all cases. In two cases, namely CA XII/CA I and CA XII/CA XIII, the R^2_{TEST} value was greater than 0.70. When selectivities were calculated from the separate affinity models, acceptable R^2_{TEST} values were obtained only in four cases out of six. This approach failed to give useful results for CA XII/CA II and CA XII/CA VI selectivities (R^2_{TEST} was less than 0.40). The mean R^2_{TEST} value was also better for the selectivity-targeted QSAR compared to the selectivity computed from the predicted affinities (0.595 vs. 0.555), showing the advantage of the first approach.

2.6. Search for lead compounds similar to metabolites using the Tanimoto index

The Tanimoto index is a number that describes similarity between two series of binary digits (bits) in interval from 0 (no similarity at all) to 1 (high similarity/identity). Molecular structures can be converted into such series in one way or another, and then they are called molecular fingerprints.

The structures of the 1475 human metabolites were obtained from the KEGG database. For every metabolite structure, the Tanimoto index was calculated with Open Babel software using FP2 molecular fingerprints against every structure in the DrugBank database.

From the resulting data pool, 4231 pairs of human metabolite and DrugBank molecular structures with the Tanimoto index higher than 0.9 were extracted for further computational studies. For every selected DrugBank molecular structure, its target EC number was extracted from the DrugBank website. The same procedure was carried out for every selected human metabolite using the KEGG database. In 2817 pairs, the structures had defined targets in both cases; other pairs were rejected. Then the pairs in which the DrugBank molecule and the human metabolite obtained from KEGG were similar (Tanimoto index higher than 0.9) and least one of their targets matched were selected, resulting in 644 such cases.

These calculations showed that in cases where the structure of the drug is similar to the metabolite structure (Tanimoto index higher than 0.9), there is approximately a 23% (644/2817·100%) chance that both compounds will bind to the same receptor. This is just a rough estimate because many drugs and/or metabolites may have unknown targets, targets not listed in databases, targets without the EC number, and so on.

Moreover, there is a possibility that some targets might just match randomly. To estimate such chances, 4000 random pairs of human metabolite and DRUGBANK molecular structures were generated. For every randomly selected DRUGBANK molecular structure, its target EC number was extracted from the DRUGBANK website. Exactly the same procedure was carried out for every randomly selected human metabolite using the KEGG database. Calculation of the cases where the targets match for both structures revealed only approximately a 1% chance that when we randomly pick a pair of the human metabolite and DRUGBANK molecular structure, it will have exactly the same target.

This procedure was used to search for compounds that were similar to the metabolites, corresponding to the certain metabolic pathways, which were very active in cancer cells. A total of 14 candidate compounds were selected for further experimental testing, and two compounds were proved to be active against cancer cells.

CONCLUSIONS

1. Six QSAR models of α-branched α,β-unsaturated ketones developed for three cell lines (NB4, MCF–7, and A549) can differentiate between active and inactive compounds.

2. Of the 5 methods tested, the results of QSAR gave the best correlation with experimental data (R^2 increased from 0.83 to 0.89). The QSAR method can be used to predict K_d values describing the binding of benzensulfonamides to CA II.

3. Application of the new LIE-like method resulted in a considerably better correlation between experimental and predicted data as compared with the original LIE approach (R^2 increased from 0.24 to 0.50).

4. It was shown that when specialized QSAR models for inhibitor selectivity were developed higher statistical scores were obtained compared to selectivity predictions made from separate affinity QSAR models.

LIST OF PUBLICATIONS

1. Ieva Karpavičienė, Giedrė Valiulienė, Vytautas Raškevičius, Indrė Lebedytė, Algirdas Brukštus, Visvaldas Kairys, Rūta Navakauskienė, Inga Čikotienė. (2015) Synthesis and Antiproliferative Activity α–Branched α, β–Unsaturated Ketones in Human Hematological and Solid Cancer Cell Lines. *European Journal of Medicinal Chemistry*, **98**, 30-48.

2. Vytautas Raškevičius, Visvaldas Kairys. (2015) Comparison of performance of docking, LIE, metadynamics and QSAR in predicting binding affinity of benzenesulfonamides. *Current Computer-Aided Drug Design*, **11**, 237-244.

3. Vytautas Raškevičius, Visvaldas Kairys. (2017) Predicting isoform-specific binding selectivities of benzenesulfonamides using QSAR and 3D-QSAR. *Current Computer-Aided Drug Design*, **13**, 75-83.

CONFERENCE PRESENTATIONS

1. “Comparison performance docking, LIE, metadynamics and QSAR in predicting binding affinity benzensulfonamides." Coins Conference Natural and Life Sciences (Vilnius, Lithuania / March 3–7, **2015**).

2. “QSAR studies α–Branched α,β–Unsaturated Ketones Antiproliferative Activity in Human Hematological and Solid Cancer Cell Lines” 10, European Conference on Computational Chemistry – EuCO–CC 2015 (Fulda, Germany / August 31 – September 3, **2015**).

3. “Lead compound similar to metabolite discovery using Tanimoto score” Vita Scientia 2016 (Vilnius, Lithuania / January 4, **2016**).

4. “Benzensulfonamidų prisijungimo selektyvumo prie karboanhidrazės izoformų prognozavimas taikant QSAR ir 3D–QSAR metodus” Fizinių ir technologijos mokslų tarpdalykiniai tyrimai (Vilnius, Lithuania / February 10, **2016**).

CURRICULUM VITAE

Name	Vytautas Raškevičius
Date of birth	April 20, 1987
Phone	+37067520607
E-mail	vytautasrask@gmail.com
Education and professional background	
2006-2010	B.Sc. Biochemistry Vilnius University
2010-2012	M.Sc. Biochemistry Vilnius University
2012-2016	PhD student of chemical engineering at Institute of Biotechnology, Vilnius University
Since 2016	Junior Scientist at Laboratory of Cell Culture, Institute of Cardiology, Lithuanian University of Health Sciences

REZIUMĖ

Vienas iš svarbiausių biochemijos mokslo praktinių pritaikymų ir tikslų yra kova su ligomis. Spartus technologijų ir mokslinių metodų vystymasis padeda spręsti daugelį iššūkių. Vienas iš jų yra baltymų – vaistų taikinių, nustatymas, kuriuos paveikus, būtų išgydyta liga arba bent jau palengvinta jos eiga. Ne mažiau svarbus iššūkis yra vaistų prieš ligas paieška, kurių dauguma yra mažos molekulinės masės junginiai, kurie slopina arba aktyvuoja baltymus-taikinius. Naudinga slopinti tik ligas sukeliančių virusų ar bakterijų baltymus, neliečiant svarbių žmogaus organizmo baltymų. Kovojant su vėžiu, dažnai yra taikomasi į baltymus, kurie svarbūs vėžinių ląstelių išlikimui.

Darbe aprašomas tyrimas vyko trimis etapais.

Pirmoje darbo dalyje buvo atliekamas QSAR su naujai susintetintais ketonais, kurių dauguma yra antiproliferaciškai aktyvūs. Buvo sukurti QSAR modeliai, kiekybiškai aprašantys minėtų junginių antiproliferacinius aktyvumus. Generuotieji QSAR modeliai yra naudingi naujų, dar nesusintetintų, cheminių junginių, panašių į nagrinėtąją ketonų seriją, antiproliferaciniam aktyvumui prognozuoti, ir tai gali pasitarnauti, kuriant priešvėžinius vaistus.

Antroje darbo dalyje CA buvo pasirinkta kaip pagrindinis taikinys. Buvo pritaikytas QSAR metodas tam, kad būtų pasiūlyti pakeitimai sulfonamidų struktūrose, kurie pagerintų jų selektyvumą. Buvo atlikta QSAR analizė įvairioms CA izoformoms, naudojant slopiklių duomenų rinkinį. Du QSAR protokolai buvo palyginti, ir su naujai sukurtu QSAR deskriptoriumi buvo pagerinta CA XII QSAR modelių statistika.

Naudojant įvairius skaičiuojamuosius metodus, buvo atlikta palyginamoji cheminės struktūros ir biologinio aktyvumo studija, nagrinėjant CA slopiklių seriją. Darbui buvo pasirinkti dokinimo, LIE, metadinamikos ir QSAR metodai. Taip pat buvo pasiūlyta LIE metodo modifikacija, kuri galutiniams skaičiavimams naudoja alternatyvią „LIE-like“ lygtį. Dėl to pavyko gauti gerokai geresnius rezultatus nei naudojant originalų LIE metodą.

Trečioje darbo dalyje buvo ieškomi nauji vėžio augimo slopikliai. Buvo įvertinta tikimybė rasti slopiklius, naudojantis Tanimoto įverčiu, nagrinėjant žinomų

vaistų cheminės struktūros panašumus į žmogaus metabolitus. Galiausiai buvo pasiūlyti keli junginiai, galimai vėžio augimo slopiklių, poros iš kurių aktyvumas buvo vėliau patvirtintas eksperimentiniu keliu.

REFERENCES

[1] Bleicher, K.H.; Böhm, H.J.; Müller, K.; Alanine, A.I. Hit and lead generation: beyond high-throughput screening. Nat. Rev. Drug. Discov. 2003, 2, 369-378.

[2] Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug. Discov. 2004, 3, 935-949.

[3] Supuran, C.T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. Nat. Rev. Drug Discov. 2008, 7, 168-181.

[4] Supuran, C.T.; Scozzafava, A. Carbonic anhydrases as targets for medicinal chemistry. Bioorg. Med. Chem. 2007, 15, 4336.

[5] Badger, M.R.; Price, G.D. The role of carbonic anhydrase in photosynthesis. Annu. Rev. Biol. 1994, 45, 369-392.

[6] Hanwell, M.D.; Curtis, D.E.; Lonie, D.C.; Vandermeersch, T.; Zurek, E.; Hutchison, G.R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J. Cheminform. 2012, 4, 17.

[7] O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. J. Cheminform. 2011, 3, 33.

[8] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; Tanchuk, V.Y.; Prokopenko, V. V. J. Comput. Aided Mol. Des. 2005, 19, 453-463.[9] Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics (2 volumes), Wiley-VCH, Weinheim, 2009.

[10] Karelson, M.; Karelson, G.; Tamm, T.; Tulp, I.; Jänes, J.; Tämm, K.; Lomaka, A.;

Savchenko, D.; Dobchev, D. QSAR study of pharmacological permeabilities. ARKIVOC. 2009, 2, 218-238.

[11] Čapkauskaitė, E.; Zubrienė, A.; Smirnov, A.; Torresan, J.; Kišonaitė, M.; Kazokaitė, J.; Gylytė, J.; Michailovienė, V.; Jogaitė, V.; Manakova, E.; Gražulis, S.; Tumkevičius, S.; Matulis, D. Benzenesulfonamides with pyrimidine moiety as inhibitors of human carbonic anhydrases I, II, VI, VII, XII, and XIII. Bioorg. Med. Chem. 2013, 21, 6937-6947.

[12] ArgusLab, Mark A. Thompson, Planaria Software LLC, Seattle, http://www.ArgusLab.com.

[13] Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. 2004, 25, 1605-1612.

[14] Liu, J.; He, X.; Zhang, J. Z. Improving the scoring of protein–ligand binding affinity by including the effects of structural water and electronic polarization. J. Chem. Inf. Model. 2013, 53, 1306-1314.

[15] Berendsen, H.J.C.; van der Spoel, D.; van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. Comp. Phys. Comm. 1995, 91, 43-56.

[16] Salomon-Ferrer, R.; Case, D.A.; Walker, R.C. An overview of the Amber biomolecular simulation package. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2013, 3, 198-210.

[17] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. J. Comput. Chem. 2004, 25, 1157-1174.

[18] Hoops, S.C.; Anderson, K.W.; Merz Jr., K.M. Force field design for metalloproteins. J. Am. Chem. Soc. 1991, 113, 8262-8270.

[19] Peters, M.B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M.N.; Merz Jr., K.M. Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). J. Chem. Theory Comput. 2010, 6, 2935-2947.

[20] Lin, F.; Wang, R. Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. J. Chem. Theory Comput. 2010, 6, 1852-1870.

[21] Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. Mol. Phys. 1983, 50, 1055-1076.

[22] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. J. Appl. phys. 1981, 52, 7182-7190.

[23] Bussi, G.; Parrinello, M. Accurate sampling using Langevin dynamics. Phys. Rev. E Stat. Nonlin. Soft Matter. Phys. 2007, 75, 056707.

[24] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. J. Chem. phys. 1993, 98, 10089-10092.

[25] Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, R.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R.A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. Comput. Phys. Commun. 2009, 180, 1961-1972.

[26] Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr. Intell. Lab. 2016, 152, 18-33.

[27] Nicolotti, O.; Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A. Applicability domain for QSAR models: where theory meets reality. IJQSPR. 2016, 1, 45-63.

[28] Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. Chemometr. Intell. Lab. 2015, 145, 22-29.

[29] Dixon, S.L.; Smondyrev, A.M.; Knoll, E.H.; Rao, S.N.; Shaw, D.E.; Friesner, R.A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J. Comput. Aided Mol. Des. 2006, 20, 647-671.

[30] Dixon, S.L.; Smondyrev, A.M.; Rao S.N. PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem. Biol. Drug. Des. 2006, 67, 370-372.

[31] Čapkauskaitė, E.; Zubrienė, A.; Baranauskienė, L.; Tamulaitienė, G.; Manakova, E.; Kairys, V.; Gražulis, S.; Tumkevičius, S.; Matulis, D. Design of [(2-pyrimidinylthio) acetyl] benzenesulfonamides as inhibitors of human carbonic anhydrases. Eur. J. Med. Chem. 2012, 51, 259-270.

[32] Čapkauskaitė, E.; Zubrienė, A.; Smirnov, A.; Torresan, J.; Kišonaitė, M.; Kazokaitė, J.; Gylytė, J.; Michailovienė, V.; Jogaitė, V.; Manakova, E.; Gražulis, S.; Tumkevičius, S.; Matulis, D. Benzenesulfonamides with pyrimidine moiety as inhibitors of human carbonic anhydrases I, II, VI, VII, XII, and XIII. Bioorg. Med. Chem. 2013, 21, 6937-6947.