



Virtual screening of PEBP1 inhibitors by combining 2D/3D-QSAR analysis, hologram QSAR, homology modeling, molecular docking analysis, and molecular dynamic simulations

Mourad Stitou¹ · Hamid Toufik¹ · Taoufik Akabli¹ · Fatima Lamchouri¹

Received: 20 October 2021 / Accepted: 19 April 2022 / Published online: 12 May 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Human phosphatidylethanolamine binding protein 1 (hPEBP1) is a novel target affecting many cellular signaling pathways involved in the formation of metastases. It can be used in the treatment of many cases of cancer. For these reasons, pharmaceutical companies use computational approaches, including multi-QSAR (2D, 3D, and hologram QSAR) analysis, homology modeling, molecular docking analysis, and molecular dynamic simulations, to speed up the drug discovery process. In this paper, QSAR modeling was conducted using two quantum chemistry optimization methods (AM1 and DFT levels). As per PLS results, we found that the DFT/B3LYP method presents high predictability according to 2D-QSAR, CoMFA, CoMSIA, and hologram QSAR studies, with Q^2 of 0.81, 0.67, 0.79, and 0.67, and external power with R^2_{pred} of 0.78, 0.58, 0.66, and 0.56, respectively. This result has been validated by CoMFA/CoMSIA graphics, which suggests that electrostatic fields combined with hydrogen bond donor/acceptor fields are beneficial to the antiproliferative activity. While the hologram QSAR models show the contributions of each fragment in improving the activity. The results from QSAR analyses revealed that ursolic acids with heterocyclic rings could improve the activities. Ramachandran plot validated the modeled PEBP1 protein. Molecular docking and MD simulations revealed that the hydrophobic and hydrogen bond interactions are dominant in the PEBP1's pocket. These results were used to predict *in silico* structures of three new compounds with potential anticancer activity. Similar molecular docking stability studies and molecular dynamics simulations were conducted.

Keywords Ursolic acid heterocyclic · Multi-QSAR · DFT/B3LYP levels · PEBP1 target · Molecular dynamic simulations (MDs)

Abbreviations

hPEBP1	Human phosphatidylethanolamine binding protein 1
QSARs	Quantitative structure–activity relationships
2D	Two dimensional
3D	Three dimensional
DFT	Density functional theory
AM1	Austin model 1
MDs	Molecular dynamics
CoMFA	Comparative molecular field analysis

CoMSIA	Comparative molecular similarity indices analysis
HQSAR	Hologram quantitative structure–activity relationships
GMQE score	Global Model Quality Estimation
OECD	Organization for Economic Co-operation and Development

Introduction

According to GLOBOCAN 2018 (Global Cancer Statistics), cancer continues to spread at an alarming rate around the world, with 18.1 million new cases and 9.6 million deaths in 2018 [1]. Human phosphatidylethanolamine-binding protein 1 (hPEBP1), a novel member of the PEBP family of newly genes recently updated in the Universal Protein Resource UniProt (<https://www.uniprot.org/>), has been implicated in a variety of cellular processes, including “signal transduction,

✉ Hamid Toufik
hamid.toufik@usmba.ac.ma

¹ Laboratory of Natural Substances, Pharmacology, Environment, Modeling, Health & Quality of Life (SNAMOPEQ), Polydisciplinary Faculty of Taza, Sidi Mohamed Ben Abdellah University of Fez, Taza Gare, B.P 1223 Taza, Morocco

cell cycle, inflammation, adhesion, proliferation, differentiation, apoptosis, autophagy, circadian rhythm, and mitotic spindle checkpoint” [2]. Furthermore, this gene encodes a member of the PEBP1 family of proteins and regulates multiple signaling pathways, namely the MAPK, NF-kappa B [3], and glycogen synthase kinase-3 (GSK-3 signaling pathways) [4]. In addition, it can also inhibit the RAF1 kinase activity through inhibiting its activation and by dissociating the RAF1/MEK complex and acting as a competitive inhibitor of MEK phosphorylation [5]. Also, it is associated with many human cancers by acting as a metastasis suppressor gene. Likewise, with multiple query sequences in the genome, it is necessary for the commitment of scientists and organizations to explore novel approaches for discovering the specific role of PEBP1 protein, which would give us new ways in cancer prevention and therapy.

Quantitative structure–activity relationship (QSAR) is a mathematical model describing the relationship between biological activities and the descriptors. The descriptors include physicochemical properties, various stereoelectronic characteristics, topological indices, and presence/absence of functional groups, and quantum chemistry (QC) descriptors [6]. The selection of accurate quantum chemical methods is of central importance in QSAR studies [7]. Semiempirical methods (like AM1) use empirical or experimental parameters to deal with the Schrödinger equation and omit some molecular integral calculations, so they are much faster than the DFT method, which is why it is utilized more widely. But the accuracy of their results is inferior to the DFT method. The QC calculation methods are rapidly improving as computer technology and quantum chemistry advance. Higher level methods such as DFT are being used more and more widely now. AM1 does not include the effects of instant electronic correlation, while the DFT-B3LYP method takes them into account [8]. Therefore, the choice of the quantum chemical method may produce errors in the calculation of the molecular descriptor and, ultimately, a lack of adjustment in QSAR model development. Whereas in the CoMFA approach, we calculate the steric and electrostatic potentials around the molecules and we relate the variations of the field calculated to the biological activity [9]. While the CoMSIA strategy is introduced as an extension of the CoMFA analysis [10], with the five similarity indices, namely hydrophobic effects and hydrogen-bond donor/acceptor potentials, however, as a result of the CoMFA and CoMSIA analyses, contour areas are undertaken to locate the favored or unfavored regions surrounding molecules that would be needed to increase the biological activity. On the other hand, CoMFA and CoMSIA require the best alignment of molecules to get better results, which poses a major problem in many cases of 3D-QSAR studies. To overcome this problem, the newly developed hologram QSAR (HQSAR) approach [11], which does not take into account molecular alignment in the QSAR model development, is used solve this problem.

Homology modeling is a commonly known as one of the structure-based methods that can predict with high accuracy the 3D structure of a protein from its amino acid sequence. Therefore, computational structural determination methods are needed to bridge this growing gap between the number of available sequences and the 3D structures of experimentally resolved proteins [12]. Furthermore, to explore the binding affinity of the comparative protein model, molecular docking simulation was used to select the best ligand–protein interactions [13], while molecular dynamics simulation was investigated their stability into binding pocket [14].

Ursolic acid (3 β -hydroxy-urs-12-en-28-oic acid) is a pentacyclic terpenoid, usually present in the stem bark, leaves, or fruit peel [15]. Recently, it has been reported that ursolic acid is one of the naturally abundant pentacyclic triterpene acids and has vast pharmacological activities including antidiabetic, antioxidative [16], antiHIV, antihepatodamage, antimalarial, antimicrobial, cardiovascular, immunomodulatory effects, anti-inflammatory, and antitumor activities [17, 18]. Ursolic acid has the following characteristics: low toxicity, liver protection, and potential anti-cancer metastasis [19, 20]. In addition, recent progress in the clinical trial proved that ursolic acid exhibits many anticancer effects with acceptable toxicity [21–23]. However, it is imperative to continue the research into the mechanism of action and signaling pathway studies of ursolic acid as a new anticancer drug.

The aim of this study was to select a good quantum chemistry (QC) method to obtain the most relevant descriptors and establish the best 2D-QSAR model. Meanwhile, the 3D-QSAR analysis has been undertaken to select the structural requirements needed to enhance the antiproliferative activities of ursolic acids through CoMFA, CoMSIA, and HQSAR analyses based on the studied quantum chemistry methods. Whereas, the homology modeling was carried out on the PEBP1 query sequence as a new target protein and its binding pocket was investigated by molecular docking simulation to explore the potential affinity of ursolic acid heterocyclic derivatives against the modeled PEBP1 protein.

At the end of this study, we predicted the design of three new heterocyclic ursolic acid derivatives with higher in silico activities than the most active compound (M30) of the series studied, and then evaluated their binding affinity and stability, respectively through molecular docking and molecular dynamics simulations.

Materials and methods

2D-QSAR modeling

Dataset treatment

A series of new synthesized ursolic acid (UA) derivatives (32 compounds of which only 30 have determined antiproliferative activities) has been selected from the literature

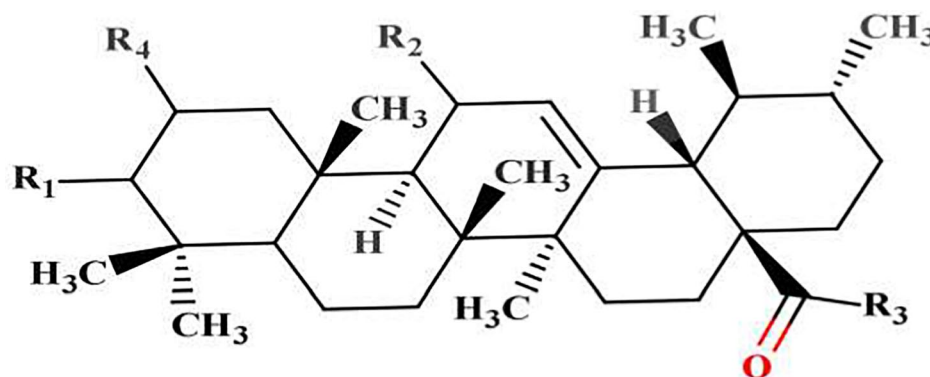
[24] in order to develop, among others, 2D/3D and hologram-QSAR models. In this study, the antiproliferative activity IC_{50} values (in μM) were converted to $-\text{Log } IC_{50}$, as reported in Table 1. Then, all compounds were sketched and initially optimized by the MM2 force field in ChemBioOffice software (15.1.0, PerkinElmer, Waltham, MA, USA). After that, we randomly selected 25 compounds as training set for QSAR development and 5 compounds as the test set to check the predictability of the retained QSAR models.

Computational details

Ursolic acids geometry optimization was performed with Gaussian 09 software [25] using two methods. The first optimisation was carried out by the semi-empirical method (AM1), followed by calculating excitation energies through the semi-empirical ZINDO method for structures optimized

by the AM1 method. The second method adopted is empiric with the optimization that was determined by the DFT (B3LYP) method [26] with the 6-31G basis set [27, 28], and the time-dependent density functional theory (TD-DFT) method [29, 30], which was used to calculate the absorption energy. Based on the calculation methods used, the descriptors determined are the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the energy of the highest occupied molecular orbital (E_{HOMO}), the frontier orbital energy (E_{gap}), the dipole moment (DM), the energies $E_{\text{(ZINDO)}}$ and $E_{\text{(TD-DFT)}}$, the oscillator strength (f), the excitation energy (ΔE), and the maximum wavelength (λ_{max}) (Supplementary Data: Table S1). Supplementary topological descriptors were determined by using the ChemBioOffice 15.1 program. Consequently, two separate matrices were obtained, each comprising 38 descriptors, and then statistical tools were employed to select the most relevant among them. Those

Table 1 Molecular structure of ursolic acid derivatives with their antiproliferative activities (pIC_{50})



ID	R1	R2	R3	R4	pIC_{50}	ID	R1	R2	R3	R4	pIC_{50}
1	OH	H	OH	H	4.89	16	=O	H	Imidazole	H	4.81
2	OAc	H	OH	H	4.14	17	=O	H	1-Methyl-Imidazole	H	4.70
3	OAc	=O	OH	H	4.47	18	=O	H	Triazole	H	4.71
4	=O	H	OH	H	4.61	19	=O	=O	Imidazole	H	4.93
5*	=O	=O	OH	H	4.26	20	=O	=O	1-Methyl-Imidazole	H	5.29
6	=O	H	OH	OH	4.87	21	=O	=O	Triazole	H	5.00
7*	OH	H	OMe	H	4.75	22	=O	H	Imidazole	Imidazole	5.29
8	=O	=O	OMe	OH	5.04	23*	=O	H	1-Methyl-Imidazole	1-Methyl-Imidazole	4.94
9	=O	H	OMe	OH	5.14	24	=O	H	Triazole	Imidazole	5.40
10	OAc	H	Imidazole	H	4.81	25	=O	=O	OMe	Imidazole	5.63
11	OAc	H	1-Methyl-Imidazole	H	4.69	26	=O	=O	OMe	1-Methyl-Imidazole	5.67
12	OAc	H	Triazole	H	4.63	27*	=O	=O	OMe	Triazole	5.09
13	OAc	=O	Imidazole	H	4.85	28	=O	H	OMe	Imidazole	5.23
14	OAc	=O	1-Methyl-Imidazole	H	4.81	29*	=O	H	OMe	1-Methyl-Imidazole	5.67
15	OAc	=O	Imidazole	H	4.99	30	=O	H	OMe	Triazole	5.72

*Test set

data sets were applied to determine the influence of quantum chemical methods on the QSAR models' performance.

Statistical analysis

Partial least squares (PLS) method Statistical analyses were conducted through QSARINS [31] and XLSTAT software. Then 2D-QSAR studies were investigated using PLS analysis at the 99% confidence level. The model's quality was assessed with R-squared (R^2 coefficient), R^2 adj (R^2 adjustment), the RMSE (root mean square error), the Student-test and its probability (t -test, p -value), and the Fisher-test (F value, p -value) [32, 33]. Furthermore, the collinearity between descriptors in the model was fixed by the correlation matrix and the variance inflation factor (VIF) [32, 34].

2D-QSAR model assessment Before any employment of the 2D-QSAR study in the prediction of new compounds, the relevant statistical metrics must check the predictive power of the retained 2D-QSAR model [35, 36]. The criteria commonly employed to check the internal reliability of the QSAR model are the leave-one-out (LOO) strategy [37]. The latter allows the Q^2 coefficient to be determined as shown in Eq. 1. For an acceptable 2D-QSAR model, the Q^2 must be greater than 0.5 [37]. Furthermore, the Y -randomization test [38] was examined to avoid chance correlation between the retained descriptors and antiproliferative activity in the selected QSAR model. However, the external predictive capability was tested by the R^2_{pred} metric (Eq. 2).

$$Q^2 = 1 - \left[\frac{\sum_i (y_{i\text{train}} - y_{i\text{pred(LOO)}})^2}{\sum_i (y_{i\text{train}} - \bar{y}_{(\text{LOO})})^2} \right] \quad (1)$$

$$R^2_{\text{Pred}} = 1 - \left[\frac{\sum_i (y_{i\text{test}} - y_{i\text{pred(test)}})^2}{\sum_i (y_{i\text{test}} - y_{i\text{train}})^2} \right] \quad (2)$$

Here, y_i and $y_{i\text{pred}}$ are respectively the experimental and predicted pIC_{50} values of each molecule (training set/test set), \bar{y}_{Train} represents the average pIC_{50} values of the molecules in the training set.

Applicability domain (AD) Before any in silico prediction of the activities of new molecules, we should define the AD as reported by the Organization for Economic Cooperation and Development (OECD) QSAR validation. The AD of the developed QSAR model is defined as the chemical space in which the biological activity of new molecules can be predicted with confidence [39]. The prediction is considered reliable only if the predicted

compound is situated inside the domain of applicability. The data set was examined for the possible existence of outliers through William plot analysis. Herein, the plot shows the standardized residuals in relation to leverage. In this case, if the high lever (h) is greater than the warning lever (h^*), the compounds have a significant influence on the predicted model [40, 41].

3D-QSAR methods

Molecular alignment

Since the quality of a 3D-QSAR model is based on the choice of the correct molecular alignment [42], for the comparison and the development of the optimal 3D-QSAR models, we have investigated alignment by two different optimization methods to reach the best quantum level that will allow having the best alignment. Alignment I, as depicted in Fig. 1a, was studied by the semi-empirical (AM1) method, and the DFT/B3LYP optimization method was investigated to get alignment II (Fig. 1b). From both alignments, the distill-rigid strategy as introduced in SYBYL-X 2.0 software [43] was carried out to select the common scaffold. For alignment, the highly active compound 30 ($pIC_{50} = 5.72$) was used as a molecule model, and the remaining 29 molecules were superimposed on it.

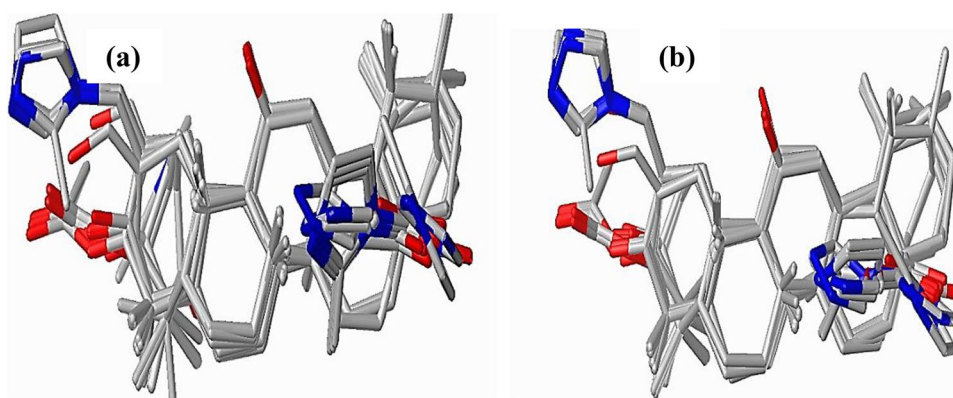
CoMFA strategy

The CoMFA approach was used to develop the reliable 3D-QSAR model based on the aligned compounds. The steric (S) field was calculated by Lennard–Jones potentials and Coulombic (electrostatic) potentials were calculated in a 3D grid regularly spaced (2.0 Å between two consecutive points) [40, 44]. A carbon-probe sp^3 hybridized with +1 charge was fixed to calculate the steric (S)/electrostatic (E) potential effects [40]. The column filtering was set at 1.0 kcal/mol to reduce noise and improve the results of the constructed 3D-QSAR model. The cut-off value for steric and electrostatic fields was set to 30 kcal/mol.

CoMSIA strategy

The CoMSIA strategy was done with the standard grid box as employed in CoMFA studies. Besides the (S) and (E) potentials, the CoMSIA hydrophobic (H) and hydrogen-bond donor (HBD) and hydrogen-bond acceptor (HBA) descriptors were calculated to assess the similarity between compounds. The CoMSIA descriptors were calculated using a carbon-probe [40] with a +1 charge for HBD and HBA at each lattice point. We have set

Fig. 1 Alignment of compounds used in 3D-QSAR analysis. **a** AM1-based method, **b** DFT-based method



column filtering at 1.0 kcal/mol to improve the performance of the CoMSIA model.

Partial least squares (PLS) analysis and 3D-QSAR models assessment

The PLS method was used to correlate the CoMFA and CoMSIA descriptors to biological activities to develop reliable 3D-QSAR models [40]. The Q^2 metric (Eq. 1) has been employed to set the optimum number of components (NOC) to achieve the high predictive power of the established 3D-QSAR models. As a result, the Q^2 metric ($Q^2 > 0.5$) allowed us to check the internal reliability of 3D-QSAR models [45, 46]. Furthermore, the external assessment of the developed 3D-QSAR models was assessed by the external set (test set: 5 molecules). Therefore, the predictive R-squared coefficient (R^2_{pred}) was calculated to assess the external reliability of the retained 3D-QSAR model.

Hologram QSAR (HQSAR)

3D-QSAR methods have been demonstrated to be reliable tools in drug design. However, the complexity of the choice of optimization methods (stable geometries) and the structural alignment method causes serious problems in the development of the 3D-QSAR model [11]. Furthermore, hologram QSAR (HQSAR) is a new approach developed to solve the problems facing 3D QSAR analysis. HQSAR uses molecular holograms (fragment fingerprints) as descriptors to determine which fragments contribute to the activity [11]. In this study, HQSAR analysis was carried out by using the following parameters: hologram length (HL) values (53 to 401) and the fragment distinction; atoms (A), bonds (B), connections (C), hydrogen atoms (H), chirality (Ch), and donor/acceptor (D/A). The setting atom count in fragments was set at between 4 and 7 atoms in each fragment. The PLS

technique was used to develop the HQSAR model and then validated by Q^2 , R^2 , and $RMSE$ values.

Homology modeling

Templates searching and modeled PEBP1 protein quality

The amino acid sequence of PEBP1 homo sapiens (ID: P30086) has been uploaded from further homology modeling. The BLASTp program was enrolled to search the best template protein with a high identity with the query sequence. After an alignment search, we have chosen the crystalized structure (PDB ID: 6ens) as the template protein. This template was uploaded from PDB the database (<https://www.rcsb.org/>). The homology model of PEBP1 was built with the SWISS-MODEL web server [47], and the accuracy of the developed homology model was checked by Ramachandran plot [48], the GMQE score (Global Model Quality Estimation), and Qmean score [49].

PEBP1 model — ursolic acid docking

Once the homology model validation is completed, we have used the CavityPlus web server to identify the binding sites necessary to get the best binding affinity. The cavity pocket is constructed with the following amino acid residues (Fig. 2): Asp-70, Asp-72, Ala-73, Pro-74, Ser-75, Lys-80, Tyr-81, Arg-82, Trp-84, His-85, His-86, Val-107, Pro-111, Thr-115, Gly-108, Gly-110, Ser-109, Pro-112, His-118, Tyr-120, Ser-142, Gly-143, Asp-144, Arg-146, Gly-147, Lys-148, Leu-180, Tyr-181, Leu-184, and Ser-185. Subsequently, the highly active compounds 30 ($pIC_{50} = 5.72$), 29 ($pIC_{50} = 5.67$), and 26 ($pIC_{50} = 5.67$) and less active compound 1 (ursolic acid without heterocyclic ring) were taken as template ligands for docking studies. Based on the obtained results from 2D, 3D, and hologram QSAR models, we selected optimized structures by DFT/B3LYP method to elucidate the best poses of the studied ligands through the binding pocket. While, Autodock Tools was

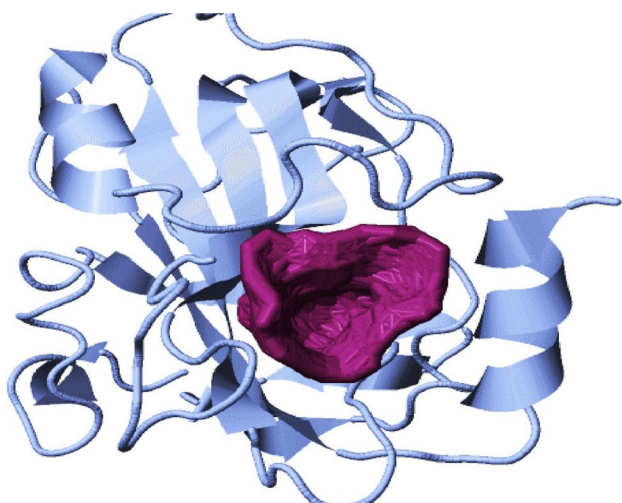


Fig. 2 Binding pocket of the modeled PEBP1 protein

used to prepare the entry PDBQT files for the Autodock vina docking program. Grid box was set to cover the maximum part of the predicted binding pocket of the homologous PEBP1 protein and to cover the studied triterpenoid heterocyclic compounds. The binding location was defined by a grid box including the whole amino acids included in the predicted binding pocket, with dimensions XYZ of $60 \text{ \AA} \times 60 \text{ \AA} \times 60 \text{ \AA}$, with a grid spacing of 0.375 \AA and grid center was designated at dimensions (Center $x = 17.0$, Center $y = -6.928$, Center $z = 20.733$). AutoDock/Vina was employed for docking using protein and ligand informations along with grid box properties in the configuration file. The best poses were carried out based on energy score (E in Kcal/mol). The generated conformations of PEBP1 protein and the studied compounds were analyzed using Discovery Studio molecular visualization software. The PEBP1-ursolic acid interactions were visualized using the Discovery Studio software.

Molecular dynamic simulations (MDs) and system building

For better understanding of the stability of the PEBP1-ligand complex, molecular dynamic simulations (MDs) were carried out using the conformations obtained by molecular docking studies. The molecular dynamics (MD) simulations were carried out for the most active compound, M30 ($pIC_{50} = 5.72$) and the novel predicted ligands (X1, X2, and X3) in the binding pocket of the modeled PEBP1 receptor. Molecular dynamic (MD) simulation studies were carried out using the software package GROMACS [50] with the CHARMM36 force field [51, 52]. The topology for the ligands was created by the CHARMM General Force Field (CGenFF) server [51]. The protein topology files were

developed using the `pdb2gmx` module. The protein ligand complex system was built inside a dodecahedron box and was solvated with three-point water model (TIP3 water) [53]. Na^+ ions were added as counter-ions to neutralize the charge of the system. Steepest descent minimization was performed to remove bad van der Waals contacts. To remove any steric clashes or unusual geometry, which would artificially raise the energy of the system, we must relax the structure by running an energy minimization (EM) algorithm. After that, the system was relaxed through energy minimization process by using steepest descent until reaching a tolerance of 1000 kJ/mol . Restrained constant number of particles, volume, and temperature (NVT) ensemble equilibration was performed for 100 picoseconds (ps) at 300 K with Berendsen thermostat temperature coupling [54]. A constant number of particles, pressure, and temperature (NPT) ensemble equilibration of 100 picoseconds (ps) followed this. Parrinello-Rahman barostat was used for pressure coupling at 1 bar [55]. Particle Mesh Ewald (PME) was applied for calculations of the long-range electrostatics [56]. The SHAKE method [57], with a tolerance of 10^{-5} \AA , was applied to constrain all covalent bonds involving hydrogen atoms. Linear Constraint Solver (LINCS) algorithm was applied to constraint the covalent bonds, including heavy atom-H bonds during the molecular dynamics (MD) simulations. Finally, the systems were simulated in water under the biological conditions, namely 300 K, approximately 1000 kg/m^3 water density and average pressure of 1 bar. The run time for each mode of complex was 10 ns. The `gmx rms`, `rmsf`, and `gyrate` were used for the calculation of RMSD, RMSF, and Rg.

Results and discussion

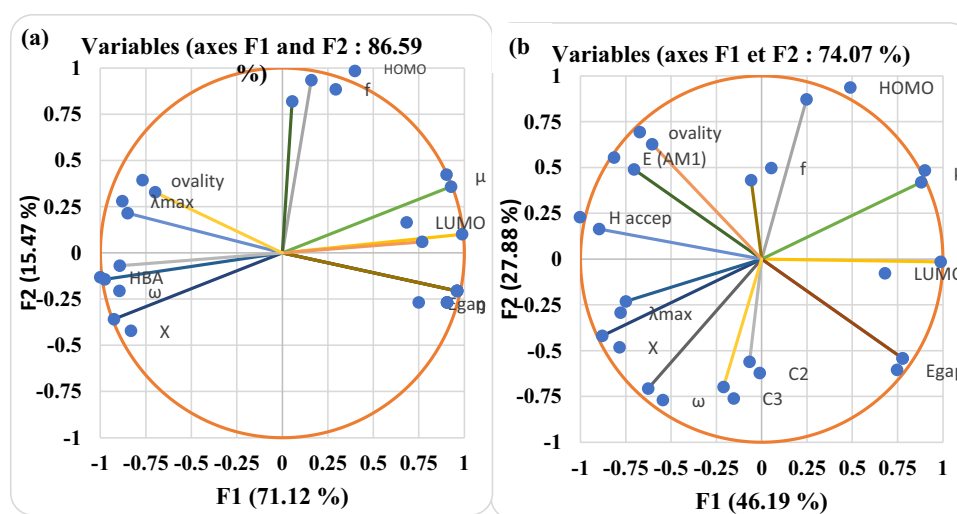
2D-QSAR analysis

We calculated a data set of 38 descriptors based on quantum mechanics methods (AM1 and DFT) and molecular mechanic methods (MM2) (for details, see Table A1, ESI[†]). The partial least squares (PLS) method allowed us to select a set of 14 descriptors from the 38 initial descriptors.

Selection of descriptors

Due to the multicollinearity problem inherent in the development of 2D-QSAR model, correlation analysis is an essential tool for refining the most correlated descriptors with the antiproliferative activity. From the analysis of the results in Fig. 3a, we can deduce that the two principal axes, F1 and F2 obtained by the DFT-based method represent 71.12% and 15.47%, respectively, of the total variance, while those obtained by the AM1-based method represent only 46.19 and 27.88, respectively (Fig. 3b). The correlation matrices

Fig. 3 Correlation circles of the studied methods: **a** DFT-based method, **b** AM1-based method



for both (AM1, DFT) methods are shown in Electronic Supplementary Material (ESI†: Tables A2 and A3).

Based on those results, we have selected four descriptors determined by the AM1-based method and the topological index: Egap, C3, HBA, and TV connectivity, whereas, four descriptors were selected by using the DFT and topological index methods: Egap, E_{TD-DFT} , f , and HBA. The matrix dataset for the retained descriptors in both models is tabulated in Supplementary Material (Table A4).

PLS regression

Using the antiproliferative activity values as the dependent variable and the quantum mechanics descriptors as independent variables, the established 2D-QSARs models through the PLS regression method are shown below:

AM1-based method:

$$pIC_{50} = 7.98 - 0.41 * Egap + 1.108 * C3 + 0.155 * HBA + 1.13 * TVConn(\times 10^6) \quad (3)$$

$$N = 25 \quad R^2 = 0.77 \quad R^2 \text{ adjusted} = 0.73 \quad RMSE = 0.19 \quad F = 17.86 \quad \alpha = 1\%$$

DFT-based method:

$$pIC_{50} = 8.50 - 0.418 * Egap + 0.00005 * E_{TD-DFT} + 26.25 * f + 0.19 * HBA \quad (4)$$

$$N = 25 \quad R^2 = 0.87 \quad R^2 \text{ adjusted} = 0.85 \quad RMSE = 0.14 \quad F = 34.41 \quad \alpha = 1\%$$

With N representing the number of molecules (in the training set), R^2 is the coefficient of determination, R^2_{adj} represents the coefficient of determination adjusted, F is the global significance of the retained 2D-QSAR model, and $RMSE$ its root mean square error. The significance level α was set at 1%.

From the above results, we can observe that the statistical quality of the model in Eq. (4) is more significant than that obtained for the model in Eq. (3). The coefficients R^2 and R^2_{adj} in the DFT-based model are remarkably superior to those obtained in the AM1-based model with a lower error RMSE too. However, the F value of each model indicates that the selected model in Eq. (4) is more significant than the observed model in Eq. (3). The values of the Student test and its resulting probability (p -value) for each descriptor included in the retained 2D-QSAR models are listed in Table 2. We can note that the p values of the t -test are all less than 0.01. Therefore, the coefficients of both models are acceptable. Moreover, the VIF values in Eqs. (3) and (4) indicate that there is no collinearity between descriptors in both 2D-QSAR models (Table 2). The observed versus predicted activities from both retained 2D-QSAR models are shown in Fig. 4. The predicted antiproliferative activities of the compounds with their residues from Eqs. (3) and (4) are given in Supplementary Material (Table A5).

Predictive powers of retained 2D-QSAR models

The internal predictive capacity of the two selected 2D-QSAR models was tested by a cross validation (LOO) procedure. Q^2 value (0.55) of Eq. (3) is relatively lower than that observed in Eq. (4) (0.81). These results show that the AM1-based model has lower predictive power than the DFT-based model. The correlation between observed and predicted activity values through internal and external validations are illustrated in Fig. 4.

The randomization test was used to avoid the chance correlation between activities and descriptors [58]. As we can see in Table 3, the initial models show high stability compared to randomized models, which allows us to conclude that there is a low probability of finding a chance correlation

Table 2 Significance of all descriptors by using the Student test and the VIF values

Descriptors	Coeff	Standard error	<i>t</i> -test	<i>p</i> -value	VIF
AM1-based model					
Egap	-0.41	0.12	-3.22	0.004	2.81
HBA	0.15	0.04	3.07	0.006	1.91
TVCConn	1,135,377	396,908	2.86	0.01	2.13
C3	1.11	0.39	2.78	12	1.21
DFT-based model					
Egap	-0.41	0.07	-5.89	<0.001	2.27
E _{TD-DFT}	0.000051	0.000015	3.43	0.001	2.45
f	26.25	7.07	3.71	0.001	1.13
HBA	0.19	0.05	3.70	0.003	3.88

in the retained 2D-QSAR models. However, the selected test set was used in this study to prove the external predictive power of both selected 2D-QSAR models. The model obtained by DFT/B3LYP level has better external predictive power ($R^2_{\text{pred}}=0.78$) than that obtained by the AM1 level ($R^2_{\text{pred}}=0.68$).

The theory level of descriptor calculation is a critical step in 2D-QSAR model development. Thus, from the above presented results, we can see the effect of the chosen level of theory on the performance of 2D-QSAR models. The DFT-based method (Eq. 4) produced the best statistical quality with an R-squared (R^2) value of 0.87 and the best predictive internal and external power with $Q^2=0.81$ and $R^2_{\text{pred}}=0.78$, respectively. The second 2D-QSAR model is based on the AM1-based method, with the significant coefficients $R^2=0.77$, $Q^2=0.55$, and $R^2_{\text{pred}}=0.68$. The significant difference in the results obtained for both developed 2D-QSAR models could probably be explained by the fact that the semi-empirical AM1 method does not include the effects of instantaneous electronic correlation

Table 3 Values of Q^2 and R^2 metrics after multiple Y-randomization test

Iteration	Randomization test for both obtained 2D-QSAR models			
	DFT-based model		AM1-based model	
	R^2	Q^2	R^2	Q^2
Initial model	0.87	0.81	0.77	0.57
Random 1	0.29	-0.07	0.08	-0.32
Random 2	0.29	-0.09	0.14	-0.59
Random 3	0.19	-0.16	0.16	-0.20
Random 4	0.06	-0.14	0.08	-1.02
Random 5	0.09	-0.29	0.19	-1.89
Random 6	0.06	-0.29	0.04	-1.33
Random 7	0.02	-0.17	0.16	-2.54
Random 8	0.14	-0.69	0.14	-0.49
Random 9	0.34	-0.01	0.12	-0.24
Random 10	0.17	-0.78	0.06	-0.46

[59], compared to DFT/B3LYP method. Our results are in agreement with those of the other comparative studies [60, 61], which have proved that the DFT/B3LYP method is the most appropriate method to determine the various properties of molecules.

From both 2D-QSAR models, the energy Egap is negatively correlated with antiproliferative activity, indicating that a decrease in Egap of ursolic acid derivatives is favorable to increase the antiproliferative activity. To ensure the significance of the energy $E_{\text{gap}}=E_{\text{LUMO}}-E_{\text{HOMO}}$ in the DFT-based model. As shown in Fig. 5, the E_{HOMO} (AM1) energies are weakly correlated with E_{HOMO} (DFT) energies ($R^2=0.40$), while the relation between E_{LUMO} (AM1) and E_{LUMO} (DFT) energies is highly correlated ($R^2=0.95$).

The main objective of this analysis is to prove that the E_{LUMO} energies calculated by the two methods (AM1

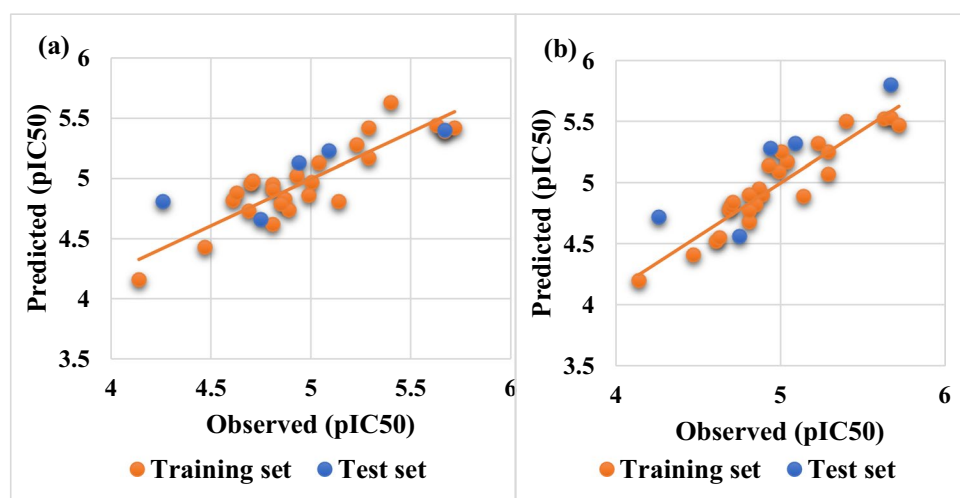
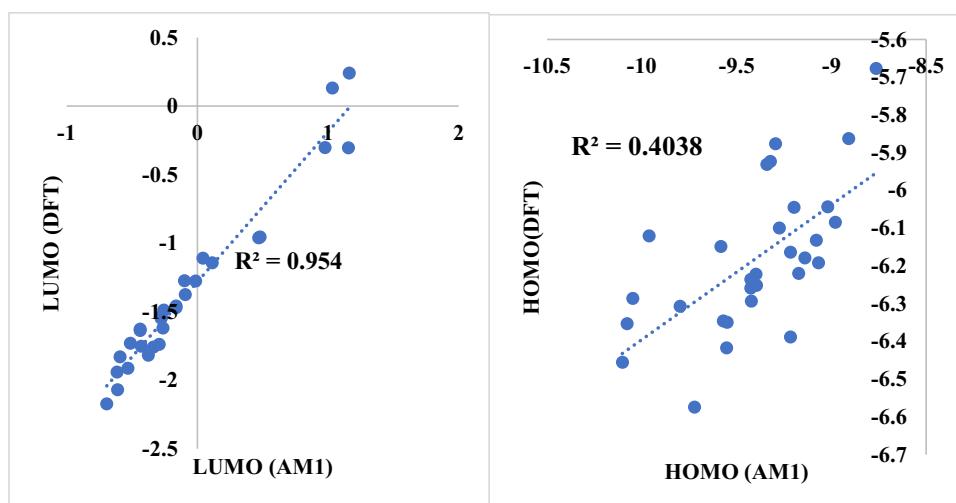
Fig. 4 Plots of observed vs. predicted activities for both 2D-QSAR models by using internal and external validation: **a** AM1-based method, **b** DFT-based method

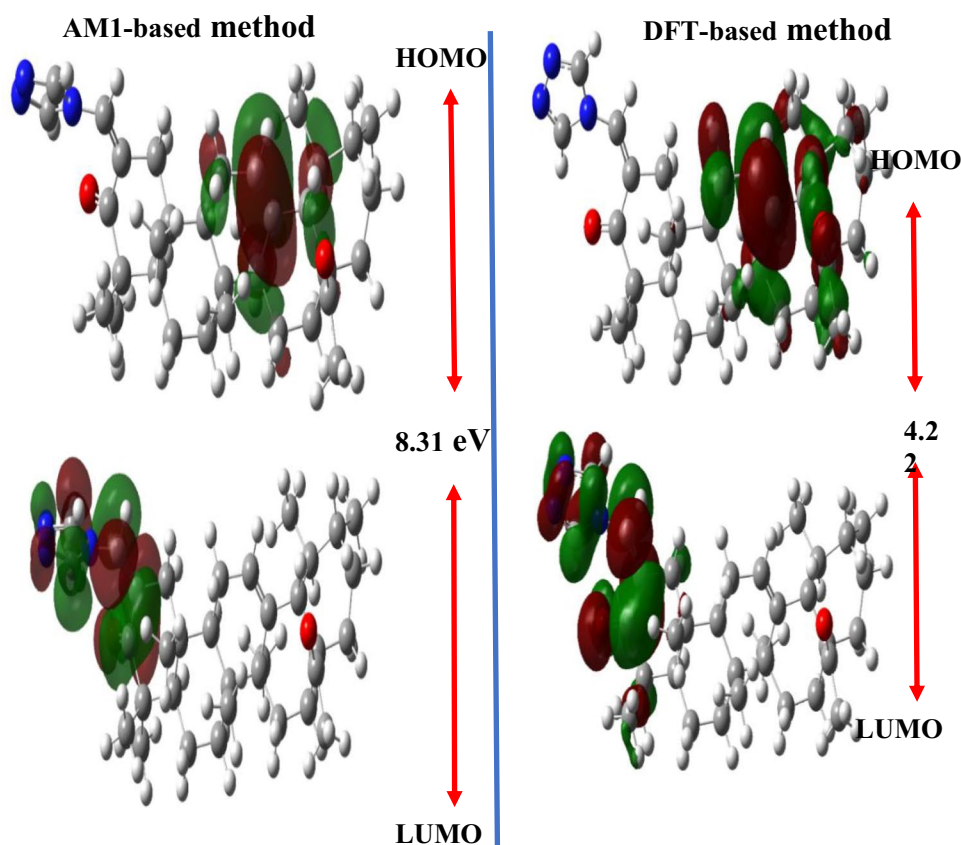
Fig. 5 Correlation between $E_{\text{LUMO}}(\text{AM1})/E_{\text{LUMO}}(\text{DFT})$ and $E_{\text{HOMO}}(\text{AM1})/E_{\text{HOMO}}(\text{DFT})$



and DFT) do not affect on the variation of E_{gap} energies. Consequently, the observed variations are due to the variation of E_{HOMO} energies. As we see in Fig. 6, for the most active compound 30 ($pIC_{50} = 5.72$) as the template molecule to elucidate the molecular orbitals (LUMO and HOMO), the orbital LUMO at both quantum levels is located around the heterocyclic group (acceptor) at R4 position, while the orbital HOMO (donor) is shown around the ursolic acid scaffold near to R3

position. From this analysis, we can deduce that the significance of E_{gap} in the DFT-based method is more accurate than that calculated in the AM1-based method. To assess the importance of the heterocyclic group in the calculation of the E_{gap} energy, another comparative study based on the DFT method was carried out with or without the heterocyclic group. Ursolic acid (compound 1) without heterocyclic groups has a high E_{gap} value of 6.28 eV compared to compound 30 ($E_{\text{gap}} = 4.22$ eV),

Fig. 6 Frontier molecular orbitals of the active compound 30 based on AM1 and DFT methods



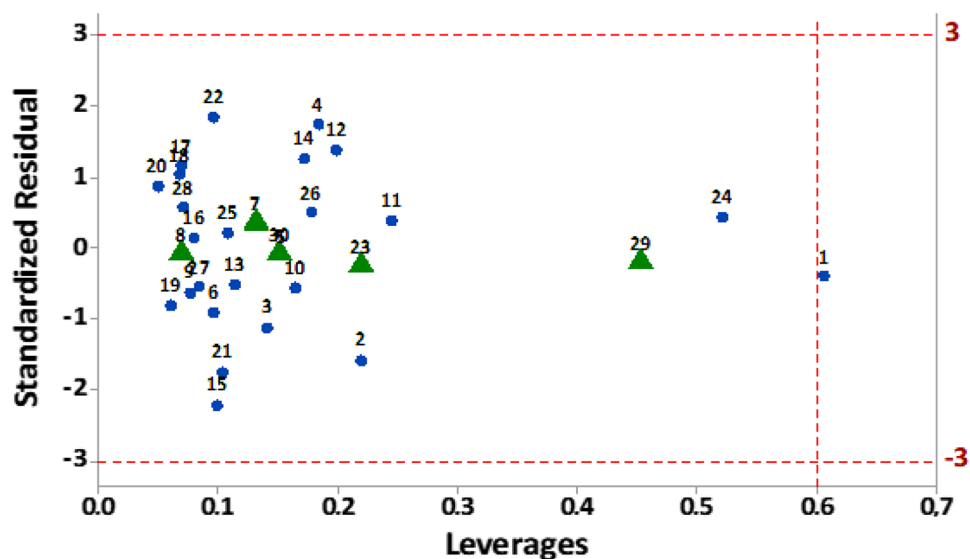
which indicates that the presence of heterocyclic groups like imidazole is favorable for reducing E_{gap} energy, thus leading to an increase in the antiproliferative activity. In addition, the $E_{\text{TD-DFT}}$ descriptor presents a positive contribution to the antiproliferative activity, which means that the lower the energy of the molecule, the more the molecule is stable. The active compound 30 could validate this analysis, which presents the lowest $E_{\text{TD-DFT}} = -44,620$ eV compared to the less active compound M1 ($E_{\text{TD-DFT}} = -38,035$ eV). The observed $E_{\text{TD-DFT}}$ energies of the two compounds could be due to heterocyclic rings grafted to ursolic acid (compound 1), which contribute favorably to improving its antiproliferative activities.

Other topological descriptor retained in the model (Eq. (3)) is the Mulliken charge (C3), calculated by AM1 method, negatively correlated to pIC_{50} , indicating that the presence of the N-alkyltriazole in R4 conjugated to an unsaturated ketone in R1 affords a positive charge of the compound 30, which is favorable to improving the antiproliferative activity. Whereas TVconn and oscillator strength (f) correlate positively to pIC_{50} . Another descriptor involved in the AM1-based model is the hydrogen bond acceptor (HBA), which is a topological descriptor derived from the acceptor of nitrogen and oxygen atoms (in this study, the heterocyclic group).

Applicability domain (AD)

The plot of the applicability domain of the developed 2D-QSAR model-based DFT method is depicted in Fig. 7. The analysis of this figure shows that compound 1 has leverage above the threshold h^* ($h^* = 0.6$) and therefore, these compounds are considered outside of the applicability domain.

Fig. 7 Applicability domain of 2D-QSAR model-based DFT method: training set (blue), test set (green)



3D-QSAR analysis

CoMFA results

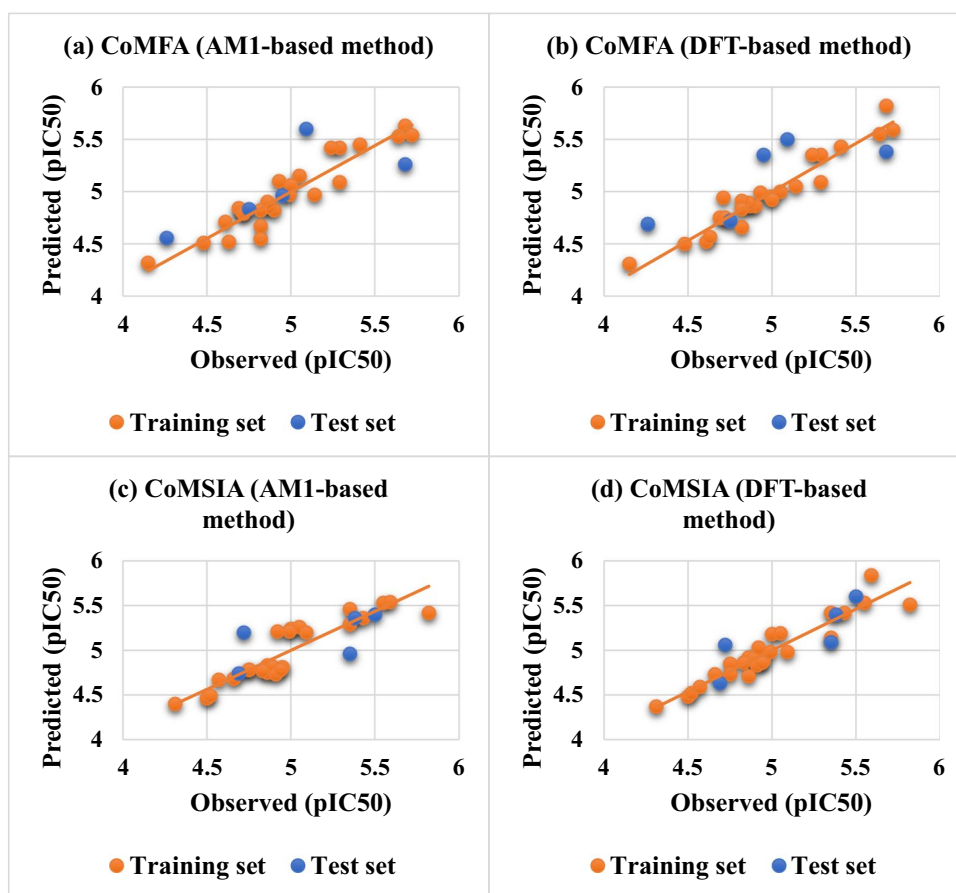
The statistical metrics of the obtained CoMFA models for both the studied alignment methods are illustrated in Table 4. The CoMFA model generated by align II (DFT-based optimization method) shows better performance than that obtained by align I (AM1-based optimization method), with values of Q^2 , R^2 , and R^2_{pred} , respectively, equal to 0.67, 0.92, and 0.58, and a low value of the $RMSE_{\text{train}}$ of 0.1. In both alignment methods, we find a slight difference between the steric and electrostatic fields' contributions. The observed and predicted activities with their residual values are presented in Supplementary Material (Table A6) and the relationship between the observed versus predicted pIC_{50} values of the training/test set is shown in Fig. 8.

CoMSIA results

In the case of the CoMSIA study, the statistical results are presented in Table 4. From the five similarity indices of CoMSIA, we have identified that E, HBD, and HBA descriptors are favorable for predicting the antiproliferative activity of ursolic acid derivatives. As we can see in Table 4, the align II showed better results than align I with Q^2 , NOC , R^2 , F , $RMSE_{\text{train}}$, and R^2_{pred} of 0.79, 2, 0.88, 86.46, 0.12, and 0.66, respectively, and a slight difference observed between E, HBD, and HBA field contributions. The observed and predicted activities with their residual values are presented in the supplementary material (Table A6) and the relationship between the observed versus predicted pIC_{50} values of training and test sets is shown in Fig. 8.

Table 4 Results of CoMFA, CoMSIA, and HQSAR studies

Statistical parameters	Align I (AM1-based)			Align II (DFT-based)		
	CoMFA	CoMSIA	HQSAR	CoMFA	CoMSIA	HQSAR
Q^2	0.62	0.70	0.67	0.67	0.79	0.67
R^2	0.88	0.84	0.86	0.92	0.88	0.87
NOC	5	2	5	3	2	5
F	53.84	62.7	83 (HL)	58.7	86.49	83 (HL)
$RMSE_{train}$	0.12	0.15	0.15	0.1	0.12	0.15
R^2_{pred}	0.54	0.62	0.56	0.58	0.66	0.56
Probability of R^2_{cv}	0	0		0	0	-
Fields contributions						
Steric (S)	0.48	-	-	0.53		-
Electrostatic (E)	0.51	0.4	-	0.46	0.45	-
H-bond donor (HBD)	-	0.32	-	-	0.3	-
H-bond acceptor (HBA)	-	0.27	-	-	0.23	-

Fig. 8 Relationships between the observed and predicted pIC_{50} values: **a** CoMFA AM1-based, **b** CoMFA DFT-based, **c** CoMSIA AM1-based, **d** CoMSIA DFT-based

HQSAR results

According to HQSAR analysis (Table 4), the identification of the fragment that best predicts the antiproliferative activity indicates that the model using the best fragment distinction (atoms (A), bonds (B), connections (C), and donor/

acceptor (DA)) gives significant results with a hologram length equal to 83. The selected hologram shows high-validated coefficients with values of Q^2 , NOC , R^2 , $RMSE$, and R^2_{pred} equal to 0.67, 4, 0.87, 0.15, and 0.56 respectively. As per Table 4, we can see that there is no significant change in the observed results of the two alignment methods. From

this analysis, we can conclude that the HQSAR model is independent of molecular alignment, which is in accordance with the literature [11, 62]. The plot of the observed versus predicted activities of training/test sets is shown in Fig. 9.

Contour maps of CoMFA, CoMSIA, and HQSAR

The CoMFA and CoMSIA contour plots were shaped to show the contributions of each selected field in enhancing the antiproliferative activity of ursolic acid derivatives. The maps obtained represent regions with standard coefficients of over 80% (favorable) or under 20% (unfavorable). The active compound 30 ($pIC_{50}=5.72$) is taken as a template molecule to elucidate the contour maps.

CoMFA contour plots As per the CoMFA contour areas, the green colors show the regions near to the R4 position, indicating that this region must be occupied by bulky steric groups (Fig. 10a), which would enhance the antiproliferative activity. Another green region is observed around the R3 position. These results could be validated, on the one hand, by the active conformation 29 ($pIC_{50}=5.67$) that contains the 1-methylimidazole group at R4 and the methyl group at R3, which explains its potential activity. On the other hand, the lack of steric groups at the R3 and R4 positions considerably reduces the antiproliferative activity. Furthermore, the least active compounds (1, 2, 3, 4, and 5) contain no steric groups at these positions, reflecting the weaker antiproliferative activities. As a result, the yellow region at R1 indicates an area where an increase in steric volume would reduce antiproliferative activity. In the case of the CoMFA electrostatic contour maps, the blue-colored areas represent the favorable regions and the red-colored areas represent the unfavorable regions for antiproliferative activity. As you see in Fig. 10b, the big contour of blue is around the triterpenoid scaffold, which indicates that hydrogen bonds would be the

reason for the high electropositive interactions around the triterpenoid. The red region near to R3 and R1 positions in the active compound 30 denotes that the electronegative group (ketone group) has likely contributed to the enhanced activity.

CoMSIA contour maps In Fig. 11a, we observe that the CoMSIA/E contour map is located at the R3 position, which indicates that electropositive groups like methyl is favorable (blue color) to the increase of the activity. Similarly, the electronegative (red color) region observed near R4 position provides that the presence of nitrogen atoms (N atom) provided by the heterocyclic rings (imidazole, triazole, 1-methyl imidazole) is favorable to enhance the conducive activities enhancement. Moreover, the favored regions of CoMSIA/HBD are displayed in cyan, while the unfavorable regions are shown in purple. Figure 11b shows a big cyan region around the R4 position suggesting that heterocyclic rings like the imidazole group may contribute favorably to the HBD interactions, consequently improving the antiproliferative activity. Furthermore, the CoMSIA/HBA presented in Fig. 11c indicates that the favored areas are visualized in magenta and the red color represents the unfavorable regions for the activity. From Fig. 11c, we can deduce that the ketone group at R1 and R3 positions are favorable to enhancing the activity, while the nitrogen atom acceptor is detrimental for the activity.

HQSAR contribution maps In addition to the CoMFA and CoMSIA contour maps, HQSAR provides complete contribution maps showing where molecule fragments could contribute to the antiproliferative activity of ursolic acid derivatives. Figure 12 shows the colored contributions for the highly active compounds 30 ($pIC_{50}=5.72$) and 26 ($pIC_{50}=5.67$). The contribution map is categorized by different colors. The extreme green color represents positive

Fig. 9 The relationships between the observed and predicted pIC_{50} values by hologram QSAR (HQSAR)

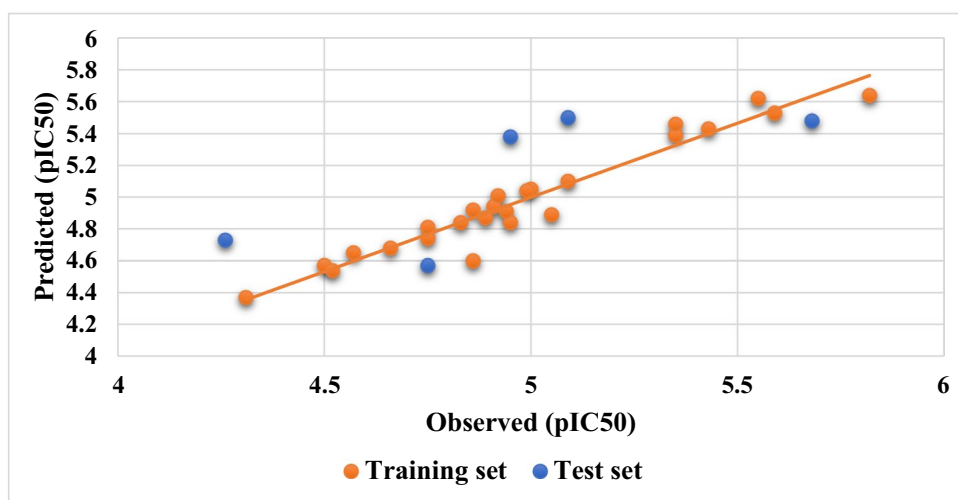
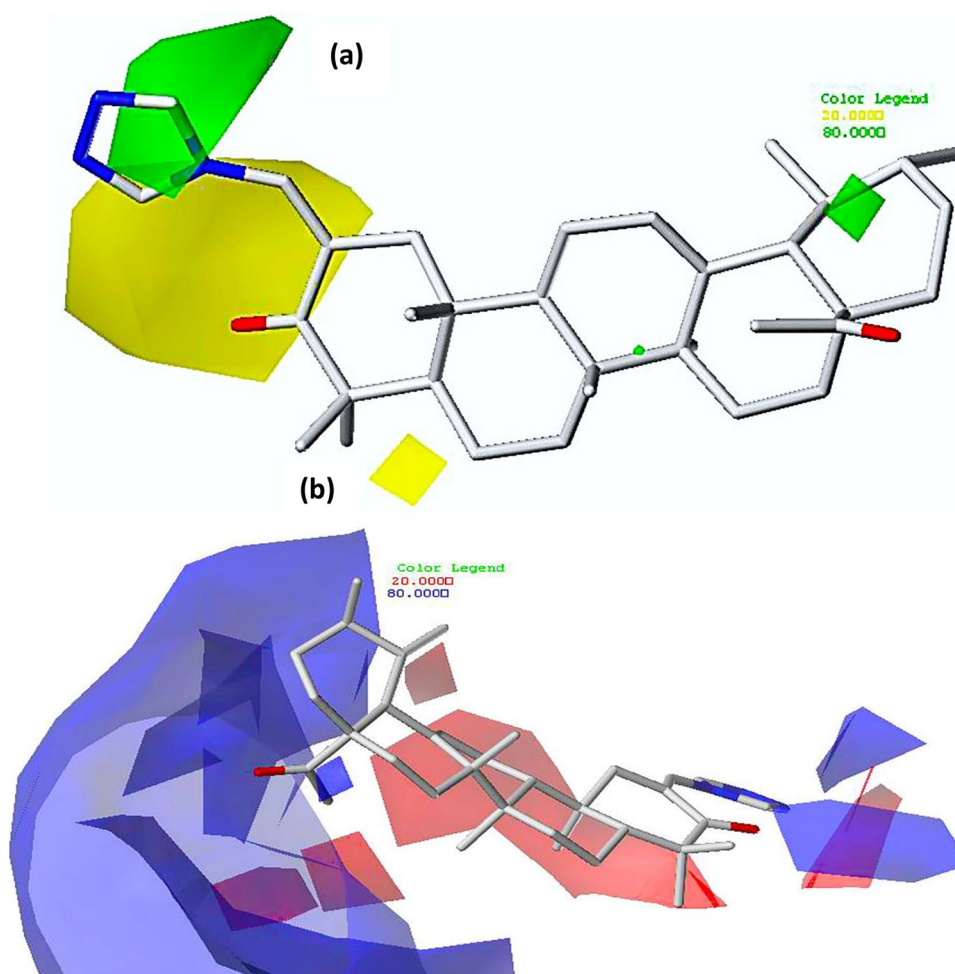


Fig. 10 Contour plots obtained by CoMFA analysis: **a** CoMFA steric, **b** CoMFA electrostatic



contributions to the antiproliferative activity, the extreme red color represents negative contributions, and the observed white color represents neutral contributions to the activity.

Homology-based modeling

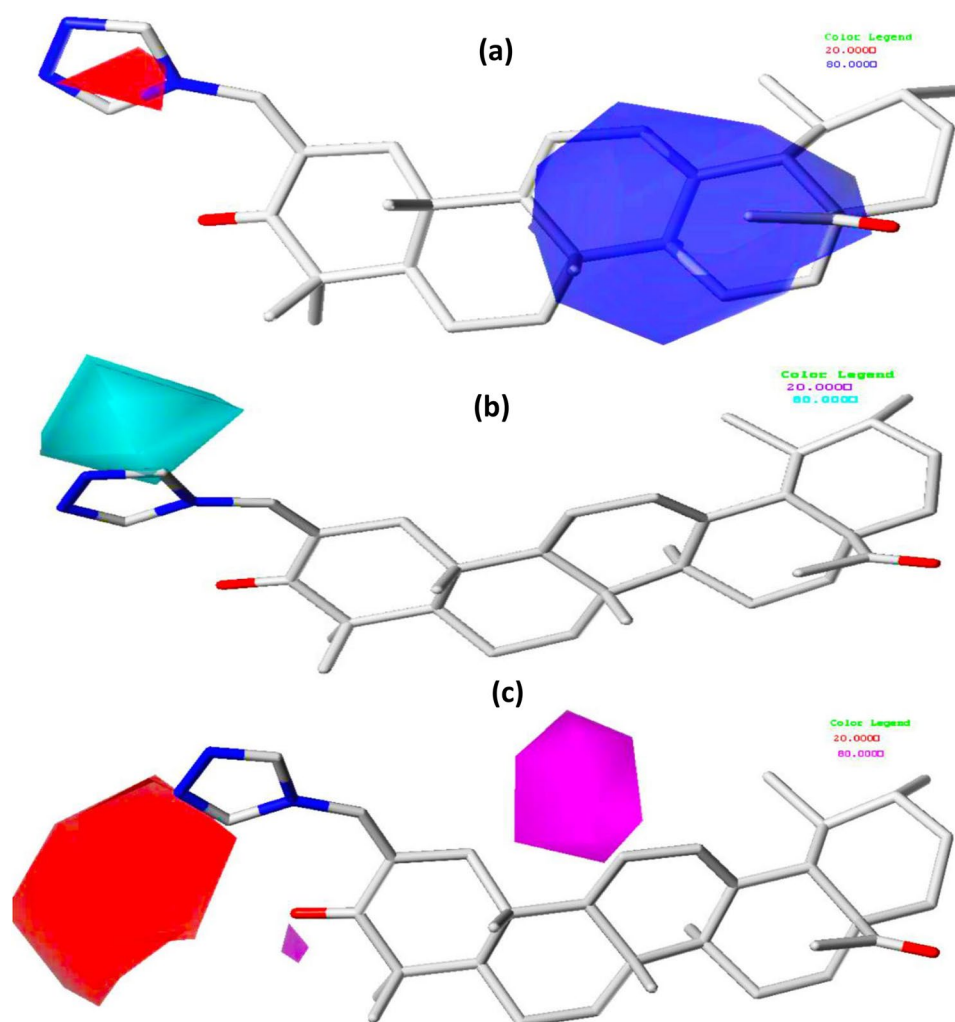
The sequence identity between the PEBP1 target and the protein (template PDB ID: 6ENS) has a value of 86.49%, which fulfills the condition proposed by Rost.B et al. (identity > 25%) [63]. Therefore, the crystal structure of the template is more suitable for homology modeling of the PEBP1 sequence. The initial alignment of the PEBP1 sequence with the template protein was examined using T-Coffee (<http://tcoffee.org.cat/>) (Fig. 13a). The quality of the PEBP1 model was examined by the Ramachandran plot (Fig. 13b) and Table A7 (Electronic Supplementary Material, ESI[†]), indicating that 100% of residues are positioned in the favored regions. In addition, the GMQE score of 0.97, which is close to 1, indicates the accuracy of the alignment to build the PEBP1 model. In addition, the QMEAN value of -0.10 indicates the best accuracy of the obtained PEBP1 model with the experimental data. Those results indicate that the PEBP1

model is stable and reliable to explain the further binding modes of ursolic acids using molecular docking.

Molecular docking studies

Docking simulation is the main approach used to screen the binding modes of the retained homology PEBP1 model (Fig. 13c). Hence, it was employed in this study to elucidate the predicted binding pocket using a set of compounds of ursolic acid derivatives. To investigate the binding affinity of the ursolic acid heterocyclic derivatives, we selected the active compounds 30, 29, and 26 (with heterocyclic ring) and the less active compounds 1, (without heterocyclic ring). As shown in Table 5, the active compounds 30 ($pIC_{50} = 5.72$), 29 ($pIC_{50} = 5.67$), 26 ($pIC_{50} = 5.67$), and the less active compound 1 ($pIC_{50} = 4.89$) showed higher binding energies, which are equal to -10 , -9.1 , -9.2 , and -8.4 kcal/mol, respectively. Molecule 30 has the best binding energy (-10 kcal/mol) with the highest antiproliferative activity compared to other compounds. To better understand the interactions in the selected liaison pocket, we have

Fig. 11 Contour plots retained by CoMSIA analysis: **a** CoMSIA electrostatic, **b** CoMSIA hydrogen-bond donor, **c** CoMSIA hydrogen-bond acceptor



represented graphically the most active compound 30 to elucidate the different interactions relevant to molecular docking.

Figure 14a shows the complex PEBP1-M30 with the dominance of hydrophobic and hydrogen bond interactions. As we see, the hydrophobic interactions have been revealed with the key residues Pro112, Tyr181, Tyr81, Pro74, Leu184, Trp84, and Ala73. Meanwhile, the residues Trp173, Lys113, Trp84, Gly141, Gly147, and Trp84 contribute to the hydrogen bond interactions. Overall, residues Pro112, Tyr181, Tyr81, Pro74, and Trp84 were conserved throughout docking simulations. The binding mode of the PEBP1-M30 is very strong and the bond distances are Trp173 (2.76 Å), Lys113 (2.06 Å), Trp84 (1.90 Å), Pro112 (4.51 Å, 4.18 Å) and 4.03 Å, Tyr181 (5.10 Å and 5.19 Å), Tyr81 (4.45 Å and 4.93 Å), and Pro74 (4.51 Å, 5.42 Å and 5.37 Å) (see Table 5).

Structurally, residue Trp84 interacts with the carbonyl group at the R3 position of compound M30, while residues

Trp173 and Lys113 interact with the imidazole group at the R3 position via hydrogen bond interactions. A strong hydrophobic network is observed around the triterpenoid scaffold by Pi-Pi T-shaped, alkyl, and Pi-alkyl interactions. On the other hand, the less active compound (1) shows Electronic Supplementary Material (Fig. A6) no hydrogen bond interactions, which could explain its low antiproliferative activity compared to the active compound 30. These results are in accordance with the 2D-QSAR and 3D-QSAR analyses, which suggest that the heterocycle groups like imidazole contribute favorably to improving the antiproliferative activity. The 2D and 3D representations of the docked molecules M26 and M29 are displayed in the ESI[†] (Figs. A1 and A2).

Molecular electrostatic potential (MEP) analysis

The importance of MEP lies in its ability to predict the binding affinity of new candidate drugs [64]. This approach indicates the intermolecular specificity of

Table 5 Molecular docking results of the homology PEBP1 model

Cpds ID	Energies score (Kcal/mol)	Antiproliferative activities (pIC_{50})	Amino acid residues H-bond interactions	Amino acid residues hydrophobic interactions
30	-10	5.72	Trp173 (2.76 Å), Lys113 (2.06 Å), Trp84 (1.90 Å)	Pro112 (4.51 Å, 4.18 Å) and 4.03 Å, Tyr181 (5.10 Å and 5.19 Å), Tyr81 (4.45 Å and 4.93 Å), Pro74 (4.51 Å, 5.42 Å and 5.37 Å)
29	-9.2	5.67	Gly141 (5.44 Å)	Tyr181 (5.13 Å), Leu184 (5.83 Å), Trp84 (4.95 Å, 4.13 Å and 5.81 Å), Leu184 (5.83 Å), Ala 73 (5.06 Å and 5.40 Å), Tyr81 (5.10 Å)
26	-9.1	5.67	Gly147 (3.65 Å), Trp84 (5.22 Å)	Tyr181 (4.81 Å, 4.28 Å and 5.88 Å), Pro112 (4.07 Å and 5.35 Å), Pro74 (5.28 Å), Tyr81 (4.76 Å and 6.69 Å), Trp84 (6.38 Å)
UA (1)	-8.4	4.89	-	Tyr81 (4.09 Å, 5.87 Å), Trp84 (5.43 Å, 5.23 Å), Pro (4.08 Å, 4.48 Å)

molecules, necessary to predict their chemical reactivity and also some types of ligand–protein interactions [65]. Electrostatic molecular potential maps are formed by moving a charged probe across the surface of the molecule. At each point, the potential electrostatic energy of the probe is computed, and the surface of the molecule is coded accordingly by the specific colors. The blue color indicates the extreme electropositive region susceptible to reacting with a nucleophilic group, and the red color represents the extreme electronegative region likely to react with an electrophile group. The observed medium color regions represent the middle of the charge between both extremes' cases. In this study, the molecular potential electrostatic maps (MEP) were generated for M1: ursolic acid, less compound (pIC_{50} = 4.89) and for the active compound in the series (30: pIC_{50} = 5.72) using DFT (B3LYP/6-31G(d,p)) level. Analysis of the results shows, on the one hand, that the ESP surfaces of compound 30 show that extremely negative regions exist around N38, N39, O29, and O26 atoms, while the extremely positive surfaces are near H85, H83, and H42 atoms (see Fig. 15a). Moreover, the observed electronegative maps on atoms N38, N39, and O26 are connected with residues Trp173, Lys113, and Trp84, respectively, of the homology model PEBP1, and the electropositive surfaces observed on atoms H85, H83, and H42 are grafted with the amino acids of Pro112 and Tyr181, which is in agreement with the molecular docking study. On the other hand, the MEP maps of compound 1 show that extreme electropositive surfaces are around atoms H62 and H69 (see electronic supplementary material, Fig. A7), while electronegative surfaces are observed around atoms O22 and O27. From this analysis, we can see that heterocyclic ursolic acids (highest active compound 30) keep better stability in the active pocket compared to the less active compound 1 (ursolic acid without

heterocycle group), these results are in accordance with CoMFA and CoMSIA analysis.

Newly predicted compound with potent anticancer activity

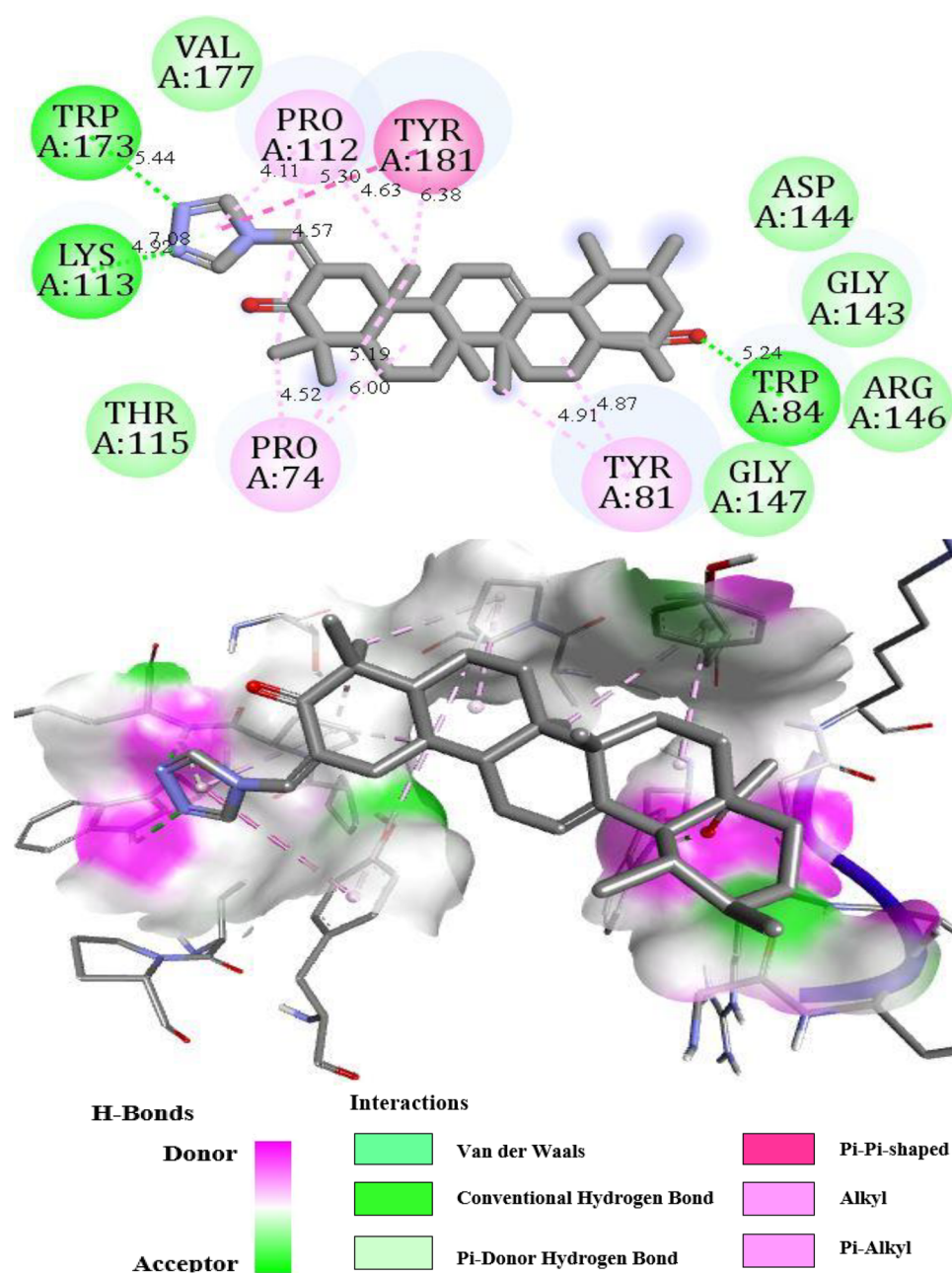
Based on the results obtained by 2D-QSAR, 3D-QSAR, hologram QSAR (HQSAR), and molecular docking simulations, a set of new molecules were predicted based on DFT/B3LYP optimization method as depicted in Table 6. Also, the predicted compounds (X1, X2, X3) were subjected to molecular docking simulations to test their stability against the modeled PEBP1 protein. The results of these simulations are presented in Table 6. From the analysis of this table, we can notice that all predicted compounds (X1, X2, X3) present a higher binding affinity than that observed for the compound with the highest activity in the series of molecules studied (M30). While the stability of those predicted compounds was also performed by ulterior molecular dynamics simulations. The 3D and 2D interactions of the predicted compounds with PEBP1 protein were shown in Figs. A3, A4, and A5 (Supplementary materials).

Molecular dynamic simulations (MDs)

Structural deviation analysis

To evaluate the stability of the studied complexes and characterize the molecular interactions between ligands and PEBP1, we calculated the root-mean-square deviation (RMSD) through 10,000 ps of simulations. The RMSD calculation was based upon the complex backbone atoms over a 10,000 ps of molecular dynamic simulations. In our analysis, we have seen that all the complexes reach the equilibration state and produce a stable trajectory through 10,000 ps of MD simulations

Fig. 14 2D-3D representations of molecular docking results of the active compound 30



(Fig. 16a). The average RMSD value for PEBP1-M30, PEBP1-X1, PEBP1-X2, and PEBP1-X3 was 0.12, 0.09, 0.11, and 0.10 nm, respectively. In the comparison of all predicted hits, the PEBP1-X1 showed a lower RMSD value. It means that this is a more stable complex than other hits. Compared to the active compound (M30), the RMSD graphs reveal that complexes with predicted ligands (X1, X2, X3) tend to have fewer deviations than complexes with the active compound (M30). In addition, the results of the present study do not show any larger conformational changes in the structure of the

PEBP1 protein when subjected to molecular dynamic simulations.

Compactness analysis

Radius of gyration (R_g) was determined to understand the level of compactness in the structure of PEBP1 in the presence of three hits (X1, X2, X3) and the most active compound in the series (M30). The R_g is defined as the mass-weighted root mean square distance of a collection of atoms from their common center of mass [66]. The R_g values for each protein–ligand complex have been calculated as shown in Fig. 16b. The average R_g values for

Fig. 15 Electrostatic potential maps (ESP) of the active compound M30

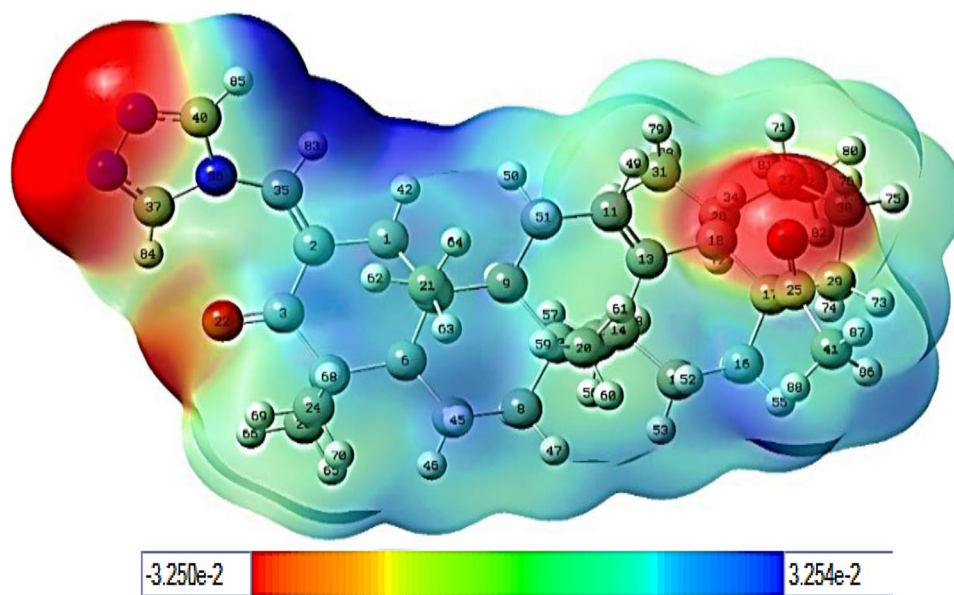
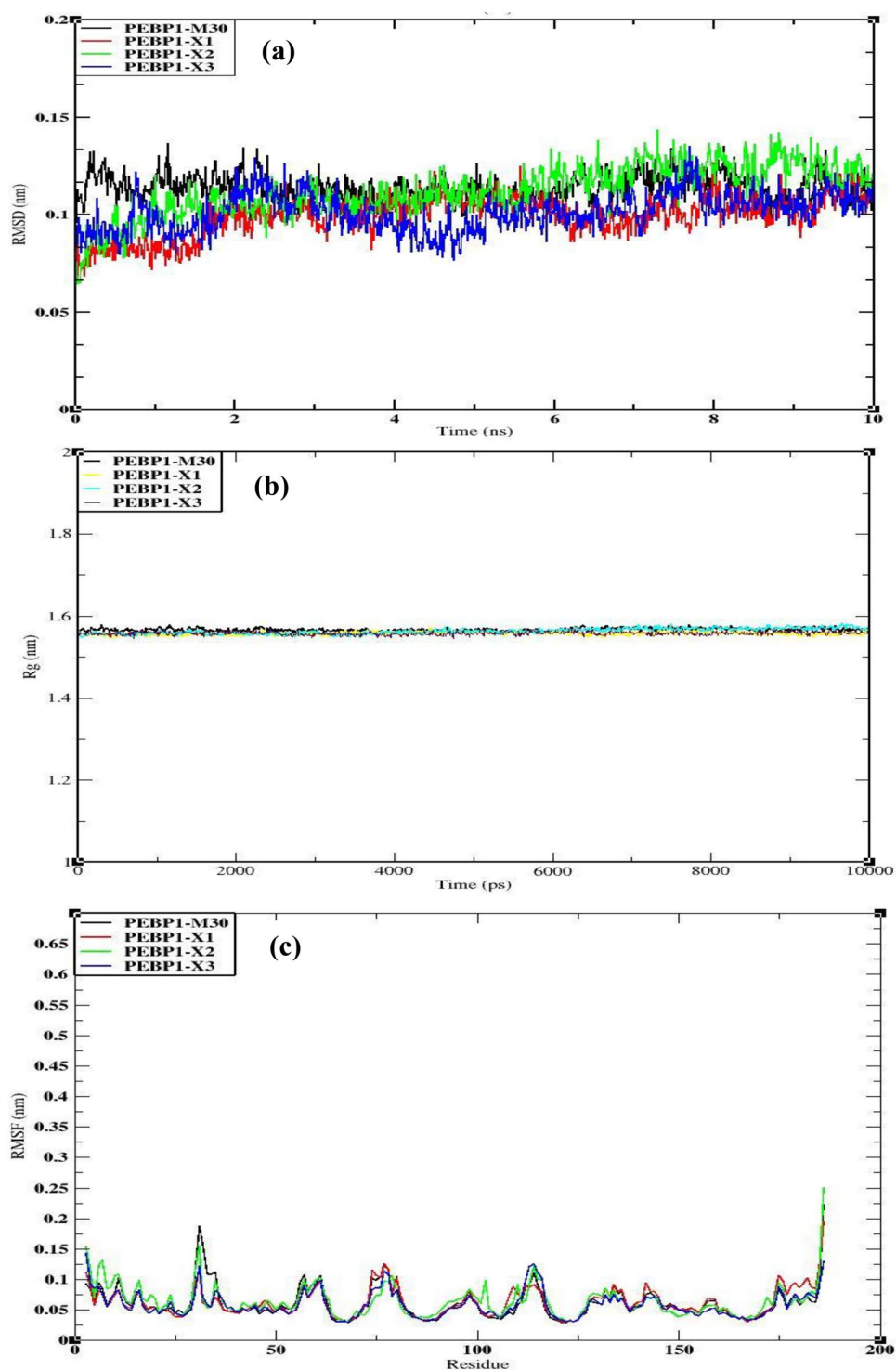


Table 6 New predicted compounds with their anticancer activities

Id	Structure	HQSAR pIC ₅₀	2D-QSAR pIC ₅₀	Binding energy (Kcal/mol)	Residues interactions
X 1		6.12	7.11	-10.30	Trp84, Leu184, Asp70, Ala73, Tyr181
X 2		6.43	6.56	-11.9	Trp84, Ala73, Tyr120, Asp70, Gly110, Tyr81
X 3		5.82	6.48	-11.1	Trp84, Tyr81, Tyr181, Ala73, Pro74, Pro112, Pro178

Fig. 16 Molecular dynamics simulations through the 10 ns MDs at 300 K of each complex: **a** RMSD vs time of C α backbone atoms for 10 ns; **b** radius of gyration values for 10 ns of simulations; **c** RMSF vs time of Ca atoms for 10 ns. The color code for each system can be seen in the legend box

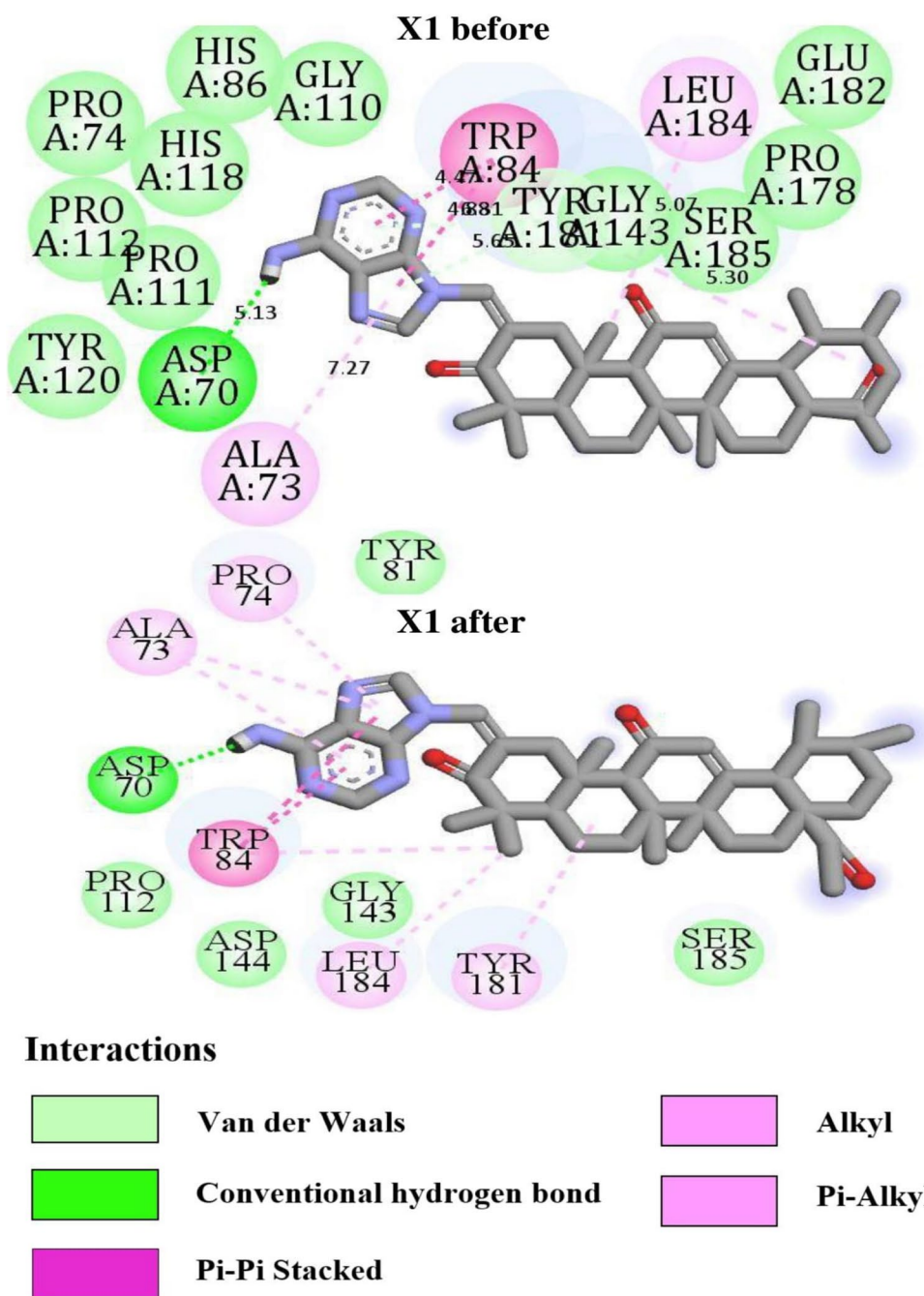


PEBP1-M30, PEBP1-X1, PEBP1-X2, and PEBP1-X3 were 1.99, 1.56, 1.54, 1.53, and 1.54 nm, respectively. From the figure and average Rg value, we have clearly seen that the difference between Rg values cannot be considered significant, which indicates the stable complexes during the MD simulations.

Flexibility analysis

The root-mean-square fluctuations (RMSF) of the C α atoms of all residues of protein were calculated by taking an average of all the conformations obtained during the entire simulations to

Fig. 17 Comparison between the conformation of the predicted ligand (X1) before and after the molecular dynamics simulations



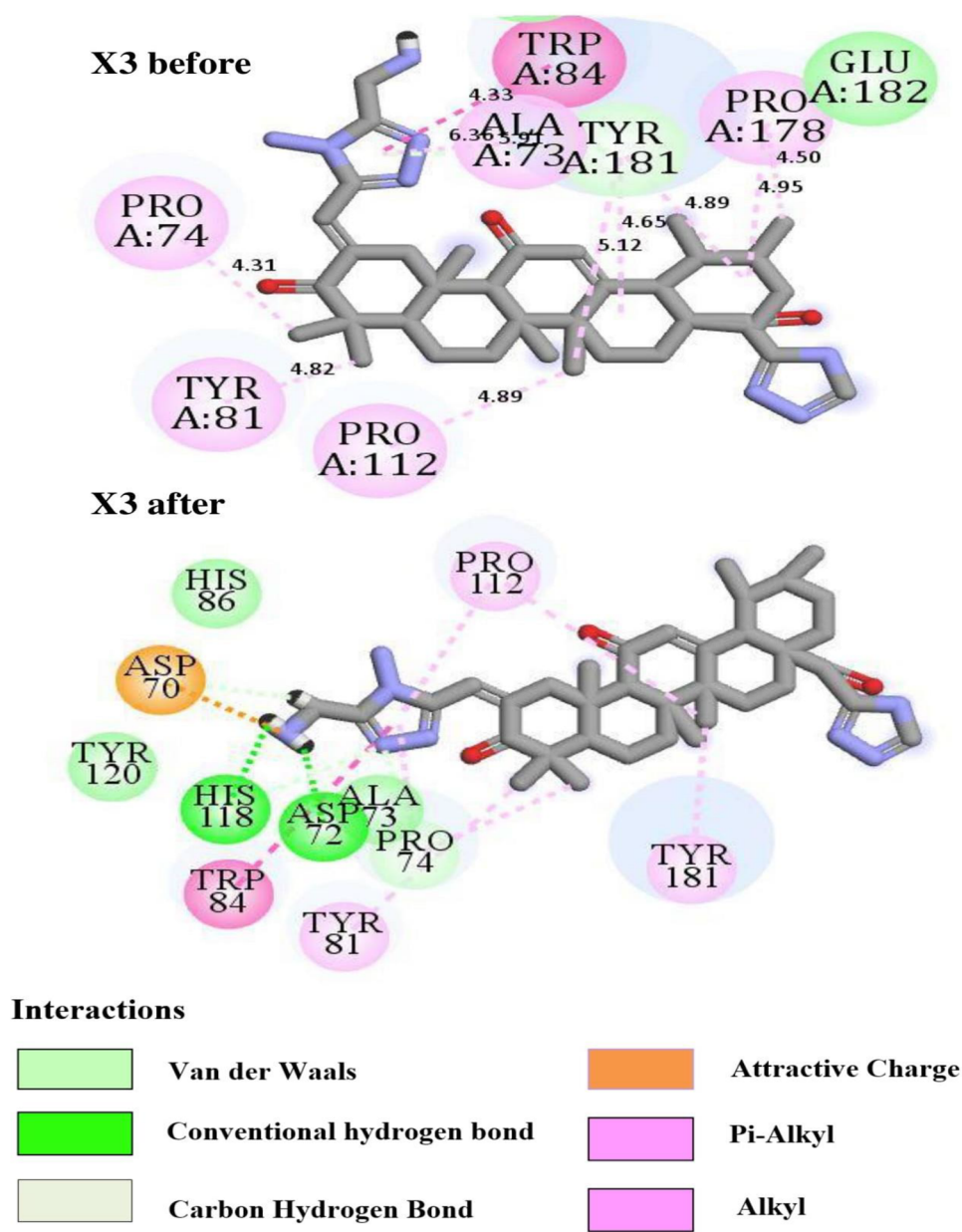
determine the dynamic behavior of residues. RMSF provides useful insights regarding the structural flexibility and the fluctuations in different regions of the protein. We have calculated the root mean square fluctuations (RMSF) as shown in Fig. 16c. The average RMSF values for PEBP1-M30, PEBP1-X1, PEBP1-X2, and PEBP1-X3 were 0.064, 0.058, 0.06, and 0.06 nm, respectively. The average value showed that the PEBP1-X1 showed a very low RMSF value compared to other complexes. From this analysis, we have seen that all the protein–ligand complexes are good in terms of fluctuation analysis. However, N-terminal and C-terminal residues showed large conformational changes. This

is due to the hanging position of the terminal residues that tend to fluctuate. The nominal variations in RMSF may be due to the dynamic nature of ligands inside the binding pocket.

Protein–ligands interactions analysis

To investigate the stability of the conformation obtained during molecular docking analysis, we proceeded to compare the binding interactions before and after MD simulations. This analysis is employed to explore more details about interactions between

Fig. 18 Comparison between the conformation of the predicted ligands (X3) before and after the molecular dynamics simulations



the modeled PEBP1 protein and lead compounds (X1, X3). Figures 17 and 18 show that all of the studied ligands are bound to the model PEBP1 protein with the same amino acids obtained from molecular docking analysis, indicating that this pocket is the binding site. The predicted compound (X1), with a binding energy of -10.3 kcal/mol, has shown the same interactions with the amino acid residues as Trp84, Leu184, Asp70, Ala73, and Tyr181 during the MD simulations. Meanwhile, the predicted compound (X3) with a binding energy of -11.1 kcal/mol showed better stability with the amino acid residues as Trp84, Tyr81, Tyr181, Ala73, Pro74, Pro112, and Pro178. As you can see in Fig. 18, the PEBP1-X3 complex obtained after molecular dynamics simulations shows additional hydrogen bond interactions with His118, and Asp72 residues and charge-charge

interaction with Asp70 residue, which indicates more stability in the binding site of the studied PEBP1 protein. This analysis indicates that the conformations obtained by molecular docking analysis are kept constant and conserving their stability before and after molecular dynamic simulations.

Conclusion

In this work, we studied the anticancer activity of a series of 30 molecules of ursolic acid derivatives using the semi-empirical method AM1 and the empirical method DFT. These methods have allowed us to build predictive 2D, 3D,

and hologram QSAR models that link biological activities to molecular descriptors, using statistical analysis. The retained 2D-QSAR results proved that the DFT (B3LYP(d, p)) method is more accurate than the AM1 method, with a high Q^2 value of 0.81 and a maximum external validity R^2_{pred} value of 0.78. These results suggest that the DFT (B3LYP(d, p)) method is the most appropriate approach to predict the antiproliferative activities of ursolic acid derivatives. Thus, the same quantum level was used to select the appropriate molecular alignment, giving the best 3D and hologram QSAR models. The results found that the DFT/B3LYP-based method is more appropriate than AM1-based method to build the 3D-QSAR model. However, the contour maps of the CoMFA and CoMSIA analysis proved that steric, electrostatic, and hydrogen bond donor/acceptor fields are beneficial to the antiproliferative activity. Moreover, we found no significant change in the performance of the HQSAR model for the two theoretical levels. Homology modeling was used to build the 3D structure of target protein PEBP1 and then validate it with Ramachandran plots and Qmean scores for further molecular docking studies. As a result, the best binding affinities between the active ursolic acid derivatives and the modeled PEBP1 protein were obtained and validated by the molecular potential maps to select each atomic contribution responsible for the observed binding affinity. Consequently, the above studies allowed us to predict three new triterpenoid derivatives with potent anticancer activity. In addition, the molecular docking and molecular dynamics simulations suggested the high conformational stability of the predicted compounds against the modeled PEBP1 protein.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00894-022-05143-6>.

Author contribution All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

Funding This work was supported by the National Center for Scientific and Technical Research (CNRST — Morocco) as part of the Research Excellence Awards Program (no. 34USMBA2017).

Data availability The data is integrated into the manuscript.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no competing interests.

References

- Bray F, Ferlay J, Soerjomataram I et al (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clinicians* 68:394–424. <https://doi.org/10.3322/caac.21492>
- Schoentgen F, Jonic S (2018) PEBP1/RKIP: from multiple functions to a common role in cellular processes. *arXiv preprint arXiv:1802.02378*
- Beshir AB, Ren G, Magpusao AN et al (2010) Raf kinase inhibitor protein suppresses nuclear factor- κ B-dependent cancer cell invasion through negative regulation of matrix metalloproteinase expression. *Cancer Lett* 299:137–149. <https://doi.org/10.1016/j.canlet.2010.08.012>
- Al-Mulla F, Bitar MS, Al-Maghrebi M et al (2011) Raf kinase inhibitor protein RKIP enhances signaling by glycogen synthase kinase-3 β . *Cancer Res* 71:1334–1343. <https://doi.org/10.1158/0008-5472.CAN-10-3102>
- Rath O, Park S, Tang H et al (2008) The RKIP (Raf-1 kinase inhibitor protein) conserved pocket binds to the phosphorylated N-region of Raf-1 and inhibits the Raf-1-mediated activated phosphorylation of MEK. *Cell Signal* 20:935–941. <https://doi.org/10.1016/j.cellsig.2008.01.012>
- Yan X-F, Xiao H-M, Gong X-D, Ju X-H (2006) A comparison of semiempirical and first principle methods for establishing toxicological QSARs of nitroaromatics. *J Mol Struct (Theochem)* 764:141–148. <https://doi.org/10.1016/j.theochem.2006.02.018>
- Puzyn T, Leszczynski J, Cronin MTD (2010) Recent advances in QSAR studies: methods and applications. Challenges and advances in computational chemistry and physics. Springer, Dordrecht, New York
- Cronin MTD, Livingstone D (2004) Predicting chemical toxicity and fate. CRC Press
- Kubinyi H (1998) Comparative Molecular Field Analysis (CoMFA). *Encyclopedia Comput Chem* 1:448–460
- Ghemtio L, Zhang Y, Xhaard H (2012) CoMFA/CoMSIA and pharmacophore modeling as a powerful tools for efficient virtual screening: application to anti-leishmanial betulin derivatives. In: *Virtual Screening*, 55–82
- Lowis DR (1997) HQSAR: a new, highly predictive QSAR technique. *Tripos Tech Notes* 1:17
- Muhammed MT, Aki-Yalcin E (2019) Homology modeling in drug discovery: overview, current applications, and future perspectives. *Chem Biol Drug Des* 93:12–20. <https://doi.org/10.1111/cbdd.13388>
- Safarizadeh H, Garkani-Nejad Z (2019) Molecular docking, molecular dynamics simulations and QSAR studies on some of 2-arylethylquinoline derivatives for inhibition of Alzheimer's amyloid-beta aggregation: insight into mechanism of interactions and parameters for design of new inhibitors. *J Mol Graph Model* 87:129–143. <https://doi.org/10.1016/j.jmgn.2018.11.019>
- Muneeswaran G, Pandiaraj M, Kartheeswaran S et al (2018) Molecular dynamics simulation approach to explore atomistic molecular mechanism of peroxidase activity of apoptotic cytochrome c mutants. *Informatics Med Unlocked* 11:51–60. <https://doi.org/10.1016/j.imu.2018.04.003>
- Woźniak Ł, Skapska S, Marszałek K (2015) Ursolic acid—a pentacyclic triterpenoid with a wide spectrum of pharmacological activities. *Molecules* 20:20614–20641. <https://doi.org/10.3390/molecules201119721>
- Mourya A, Akhtar A, Ahuja S et al (2018) Synergistic action of ursolic acid and metformin in experimental model of insulin resistance and related behavioral alterations. *Eur J Pharmacol* 835:31–40. <https://doi.org/10.1016/j.ejphar.2018.07.056>

17. López-Hortas L, Pérez-Larrán P, González-Muñoz MJ et al (2018) Recent developments on the extraction and application of ursolic acid. A review. *Food Res Int* 103:130–149. <https://doi.org/10.1016/j.foodres.2017.10.028>
18. Hua S-X, Huang R-Z, Ye M-Y et al (2015) Design, synthesis and in vitro evaluation of novel ursolic acid derivatives as potential anticancer agents. *Eur J Med Chem* 95:435–452. <https://doi.org/10.1016/j.ejmech.2015.03.051>
19. Tang Q, Liu Y, Li T et al (2016) A novel co-drug of aspirin and ursolic acid interrupts adhesion, invasion and migration of cancer cells to vascular endothelium via regulating EMT and EGFR-mediated signaling pathways: multiple targets for cancer metastasis prevention and treatment. *Oncotarget* 7. <https://doi.org/10.18632/oncotarget.12232>
20. Mlala S, Oyedeji AO, Gondwe M, Oyedeji OO (2019) Ursolic acid and its derivatives as bioactive agents. *Molecules* 24:2751. <https://doi.org/10.3390/molecules24152751>
21. Qian Z, Wang X, Song Z et al (2015) A phase I trial to evaluate the multiple-dose safety and antitumor activity of ursolic acid liposomes in subjects with advanced solid tumors. *Biomed Res Int* 2015:1–7. <https://doi.org/10.1155/2015/809714>
22. Sultana N (2011) Clinically useful anticancer, antitumor, and anti-wrinkle agent, ursolic acid and related derivatives as medicinally important natural product. *J Enzyme Inhib Med Chem* 26:616–642. <https://doi.org/10.3109/14756366.2010.546793>
23. Iqbal J, Abbasi BA, Ahmad R et al (2018) Ursolic acid a promising candidate in the therapeutics of breast cancer: current status and future implications. *Biomed Pharmacother* 108:752–756. <https://doi.org/10.1016/j.biopha.2018.09.096>
24. Leal AS, Wang R, Salvador JAR, Jing Y (2012) Synthesis of novel ursolic acid heterocyclic derivatives with improved abilities of antiproliferation and induction of p53, p21waf1 and NOXA in pancreatic cancer cells. *Bioorg Med Chem* 20:5774–5786. <https://doi.org/10.1016/j.bmc.2012.08.010>
25. Gaussian 09 (2009) R.A.: 1, mj frisch, gw trucks, hb schlegel, ge scuseria, ma robb, jr cheeseman, g. Scalmani, v. Barone, b. Mennucci, ga petersson et al., gaussian. Inc Wallingford CT. 121, 150
26. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140:A1133–A1138. <https://doi.org/10.1103/PhysRev.140.A1133>
27. Lazrak M, Toufik H, Bouzzine SM et al (2018) The computational study of bridge effect in D- π -A photosensitive dyes, based on triphenylamine. *IOP Conf Ser Earth Environ Sci* 161:012021. <https://doi.org/10.1088/1755-1315/161/1/012021>
28. Lazrak M, Toufik H, Bouzzine SM, Lamchouri F (2020) Bridge effect on the charge transfer and optoelectronic properties of triphenylamine-based organic dye sensitized solar cells: theoretical approach. *Res Chem Intermed* 46:3961–3978. <https://doi.org/10.1007/s11164-020-04184-x>
29. Casida ME (1995) Time-dependent density functional response theory for molecules. In: *Recent Advances in Computational Chemistry*. WORLD SCIENTIFIC 155–192. https://doi.org/10.1142/9789812830586_0005
30. Ennehary S, Toufik H, Bouzzine SM, Lamchouri F (2020) Effect of the alkyl chain length on the optoelectronic properties of organic dyes: theoretical approach. *J Comput Electron* 19:840–848. <https://doi.org/10.1007/s10825-020-01486-6>
31. Gramatica P, Chirico N, Papa E et al (2013) QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34:2121–2132. <https://doi.org/10.1002/jcc.23361>
32. Stitou M, Toufik H, Akabli T et al (2018) 2D-QSAR method of lupane-type saponins the treatment of cancer cell line. *RHAZES: Green Appl Chem* 2:33–45
33. Akabli T, Toufik H, Yasri A et al (2018) Combining ligand-based and structure-based drug design approaches to study the structure-activity relationships of a β -carboline derivative series. *Struct Chem* 29:1637–1645. <https://doi.org/10.1007/s11224-018-1141-1>
34. O'Brien RM, (2007) A caution regarding rules of thumb for variance inflation factors. *Qual Quant* 41:673–690. <https://doi.org/10.1007/s11135-006-9018-6>
35. Golbraikh A, Tropsha A (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers*. 231-243. <https://doi.org/10.1023/A:1021372108686>
36. ReenuVikas, (2016) Evaluating the role of electron-correlation in the external prediction of the toxicity of nitrobenzenes towards *Tetrahymena pyriformis*. *New J Chem* 40:2343–2353. <https://doi.org/10.1039/C5NJ02552D>
37. Stitou M, Toufik H, Bouachrine M et al (2019) Machine learning algorithms used in quantitative structure-activity relationships studies as new approaches in drug discovery. In: *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, IEEE, Taza, Morocco, 1–8. <https://doi.org/10.1109/ISACS48493.2019.9068917>
38. Rücker C, Rücker G, Meringer M (2007) γ -randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357. <https://doi.org/10.1021/ci700157b>
39. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29. <https://doi.org/10.1016/j.chemolab.2015.04.013>
40. Pourbasheer E, Aalizadeh R, Shokouhi Tabar S et al (2014) 2D and 3D quantitative structure-activity relationship study of hepatitis C virus NS5B polymerase inhibitors by comparative molecular field analysis and comparative molecular similarity indices analysis methods. *J Chem Inf Model* 54:2902–2914. <https://doi.org/10.1021/ci500216c>
41. Hadni H, Elhallaoui M (2020) 2D and 3D-QSAR, molecular docking and ADMET properties *in silico* studies of azaaurones as anti-malarial agents. *New J Chem* 44:6553–6565. <https://doi.org/10.1039/C9NJ05767F>
42. Halim SA, Zaheer-ul-Haq, (2015) Structure based 3D-QSAR studies of interleukin-2 inhibitors: comparing the quality and predictivity of 3D-QSAR models obtained from different alignment methods and charge calculations. *Chem Biol Interact* 238:9–24. <https://doi.org/10.1016/j.cbi.2015.05.018>
43. SYBYL/QSAR. Molecular Modelling Software, Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63944, USA
44. Stitou M, Toufik H, Bouachrine M, Lamchouri F (2020) Quantitative structure-activity relationships analysis, homology modeling, docking and molecular dynamics studies of triterpenoid saponins as Kirsten rat sarcoma inhibitors. *J Biomol Struct Dyn* 39:152–170. <https://doi.org/10.1080/07391102.2019.1707122>
45. Matysiak J, Niewiadomy A (2017) QSAR models of antiproliferative activity of imidazo[2,1-b][1,3,4]thiadiazoles in various cancer cell lines. *Mol Divers* 21:211–218. <https://doi.org/10.1007/s11030-016-9705-8>
46. Wang X, Yan J, Wang M et al (2018) Synthesis and three-dimensional quantitative structure-activity relationship study of quinoxaline derivatives containing a 1,3,4-oxadiazole moiety as efficient inhibitors against *Xanthomonas axonopodis* pv. *citri*. *Mol Divers* 22:791–802. <https://doi.org/10.1007/s11030-018-9837-0>
47. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201. <https://doi.org/10.1093/bioinformatics/bti770>
48. Ye W-L, Zhang L-X, Guan Y-D et al (2019) Virtual screening and experimental validation of eEF2K inhibitors by combining homology modeling, QSAR and molecular docking from FDA

- approved drugs. *New J Chem* 43:19097–19106. <https://doi.org/10.1039/C9NJ02600B>
49. Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: A comprehensive scoring function for model quality assessment. *Proteins Struct Funct Bioinformatics* 71:261–277. <https://doi.org/10.1002/prot.21715>
50. Van Der Spoel D, Lindahl E, Hess B et al (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26:1701–1718. <https://doi.org/10.1002/jcc.20291>
51. Vanommeslaeghe K, Hatcher E, Acharya C et al (2009) CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem NA-NA*. <https://doi.org/10.1002/jcc.21367>
52. Huang J, MacKerell AD (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* 34:2135–2145. <https://doi.org/10.1002/jcc.23354>
53. Jorgensen WL, Tirado-Rives J The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 110:1657–1666. <https://doi.org/10.1021/ja00214a001>
54. Berendsen HJC, Postma JPM, van Gunsteren WF et al (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690. <https://doi.org/10.1063/1.448118>
55. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys* 52:7182–7190. <https://doi.org/10.1063/1.328693>
56. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an $N\log(N)$ method for Ewald sums in large systems. *J Chem Phys* 98:10089–10092. <https://doi.org/10.1063/1.464397>
57. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 23:327–341. [https://doi.org/10.1016/0021-9991\(77\)90098-5](https://doi.org/10.1016/0021-9991(77)90098-5)
58. Walton IM, Cox JM, Benson CA et al (2016) The role of atropisomers on the photo-reactivity and fatigue of diarylethene-based metal–organic frameworks. *New J Chem* 40:101–106. <https://doi.org/10.1039/C5NJ01718A>
59. Frimand K, Jalkanen KJ (2002) SCC-TB, DFT/B3LYP, MP2, AM1, PM3 and RHF study of ethylene oxide and propylene oxide structures, VA and VCD spectra. *Chem Phys* 279:161–178. [https://doi.org/10.1016/S0301-0104\(02\)00457-3](https://doi.org/10.1016/S0301-0104(02)00457-3)
60. Pasha FA, Srivastava HK, Singh PP (2005) Comparative QSAR study of phenol derivatives with the help of density functional theory. *Bioorg Med Chem* 13:6823–6829. <https://doi.org/10.1016/j.bmc.2005.07.064>
61. Deeb O, Clare BW (2008) Comparison of AM1 and B3LYP-DFT for inhibition of MAO-A by phenylisopropylamines: a QSAR study. *Chem Biol Drug Design* 71:352–362. <https://doi.org/10.1111/j.1747-0285.2008.00643.x>
62. Myint KZ, Xie X-Q (2010) Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *IJMS* 11:3846–3866. <https://doi.org/10.3390/ijms11103846>
63. Rost B, Sander C (1996) Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 25:113–136
64. Weiner PK, Langridge R, Blaney JM et al (1982) Electrostatic potential molecular surfaces. *Proc Natl Acad Sci* 79:3754–3758. <https://doi.org/10.1073/pnas.79.12.3754>
65. Rathi PC, Ludlow RF, Verdonk ML (2019) Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network. *J Med Chem* 63:8778–8790. <https://doi.org/10.1021/acs.jmedchem.9b01129>
66. Monajjemil M, Oliaey AR (2009) Gyration radius and energy study at different temperatures for acetylcholine receptor protein in gas phase by Monte Carlo, molecular and Langevin dynamics simulations. *J Phys Theor Chem Islamic Azad Univ Iran* 5:195–201

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.