BEILSTEIN JOURNAL OF NANOTECHNOLOGY

AI-assisted models to predict chemotherapy drugs modified with C_{60} fullerene derivatives

Jonathan-Siu-Loong Robles-Hernández[1], Dora Iliana Medina[2], Katerin Aguirre-Hurtado[3], Marlene Bosquez[3], Roberto Salcedo*[3] and Alan Miralrio*[1]

Full Research Paper

Open Access

Address:
[1]Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey 64849, Mexico, [2]Tecnologico de Monterrey, Institute of Advanced Materials for Sustainable Manufacturing, Monterrey 64849, Mexico and [3]Instituto de Investigaciones en Materiales, Universidad Nacional Autónoma de México, Circuito exterior s/n, Ciudad Universitaria, Coyoacán, 04510, Ciudad de México, Mexico

Email:
Roberto Salcedo* - salcedo@unam.mx; Alan Miralrio* - miralrio@tec.mx

* Corresponding author

Keywords:
breast cancer; CXCR7; drug nanocarriers; QSAR

Beilstein J. Nanotechnol. **2024,** *15,* 1170–1188.
https://doi.org/10.3762/bjnano.15.95

Received: 28 May 2024
Accepted: 30 August 2024
Published: 19 September 2024

This article is part of the thematic issue "Symposium of Nanoscience and Nanomaterials 2024 (SNN 2024)".

Guest Editor: R. D. Cadena-Nava

Abstract

Employing quantitative structure–activity relationship (QSAR)/ quantitative structure–property relationship (QSPR) models, this study explores the application of fullerene derivatives as nanocarriers for breast cancer chemotherapy drugs. Isolated drugs and two drug–fullerene complexes (i.e., drug–pristine C_{60} fullerene and drug–carboxyfullerene C_{60}–COOH) were investigated with the protein CXCR7 as the molecular docking target. The research involved over 30 drugs and employed Pearson's hard–soft acid–base theory and common QSAR/QSPR descriptors to build predictive models for the docking scores. Energetic descriptors were computed using quantum chemistry at the density functional-based tight binding DFTB3 level. The results indicate that drug–fullerene complexes interact more with CXCR7 than isolated drugs. Specific binding sites were identified, with varying locations for each drug complex. Predictive models, developed using multiple linear regression and IBM Watson artificial intelligence (AI), achieved mean absolute percentage errors below 12%, driven by AI-identified key variables. The predictive models included mainly quantitative descriptors collected from datasets as well as computed ones. In addition, a water-soluble fullerene was used to compare results obtained by DFTB3 with a conventional density functional theory approach. These findings promise to enhance breast cancer chemotherapy by leveraging fullerene-based drug nanocarriers.

Introduction

Breast cancer is the most diagnosed cancer in women and the second leading cause of cancer-related mortality in women [1,2]. Heritage is the most critical risk factor, and 15 to 20% of breast cancer is familiar [3]. One of the characteristics of breast cancer is that it can be wholly cured given an early diagnosis [4]. The mortality rate from breast cancer has been reduced by 1.9% annually from 2002 to 2011 and 1.3% from 2011 to 2020 [5]. Diagnostics and treatments have continuously improved through the years. However, the situation is different in each country considering the costs and technological advances in each country. In the United States, 300,590 cases of breast cancer had been estimated for the year 2023, with a total of 43,700 deaths [6]. Latin America has over 210,000 new cases and around 60,000 deaths yearly [7]. For the year 2020, it was estimated that about 2.3 million breast cancer cases were diagnosed in women globally, and about 685,000 died from this disease [8]. A recurrent problem with standard treatments are the side effects. Regarding the use of chemotherapeutic drugs, such issues are nephrotoxicity of cisplatin, cardiotoxicity of doxorubicin, and pulmonary fibrosis from the use of bleomycin [9-11]. Besides, in the case of radiotherapy, fibrosis, atrophy, and neuronal damage caused by irradiation can occur [12,13]. Consequently, novel treatments try to reduce the secondary effects while retaining the benefits of standard approaches.

Chemotherapy is one of the most extensively applied treatments for breast cancer, with different drug targets depending on the type of cancer. Progesterone- or estrogen-receptor-positive tumors are related to cancers with low mortality [14]. Another common target in chemotherapy is human epidermal growth factor receptor 2 (HER2). Only 15 to 20% of all tumors are HER2-positive, overexpressing Erb-B2 receptor tyrosine kinase 2 (ERBB2) in the cell membrane. HER2 tumors are usually more aggressive than other ones, but the advantage is that their treatment is very effective [15]. Another chemotherapy target is the chemokine C-X-C motif receptor 7 (CXCR7) [16,17]. This G-protein is targeted because studies show a possible positive effect on inhibiting the metastasis of cervical cancer cells [18]. However, more clinical and preclinical studies on CXCR7 and its co-player CXCR4 are required since alterations have been detected in diseases such as cancer, central nervous system and cardiac disorders, and autoimmune diseases [16].

In recent years, nanomaterials have attracted the attention of different scientific communities by providing them with new solutions for drug delivery [19,20]. These nanotechnological applications have made it possible to obtain treatments that release substances at specific sites of interest, reducing the required drug amount and side effects. Nanostructures to form these drug delivery systems can be divided into organic and inorganic [19,20], with the latter one being the less extensively studied. One option currently considered in pharmacy and medicine is carbon-based nanomaterials because of their physicochemical, mechanical, electrical, thermal, and optical properties [19,20], as well as their capacity to modify existing drugs. Fullerene derivatives have been proposed recently, particularly those obtained from fullerene C_{60} [21]. The unmodified fullerene C_{60} is known as a “free radical sponge” because its double bonds tend to accept free radicals [22]. Because of its size, surface area, and capacity to extinguish or generate reactive oxygen species, C_{60} is very promising in medicine and clinical therapy [23,24]. It is also possible to modify pristine fullerenes by adding polar functional groups (e.g., –COOH, –OH, or $–NH_2$), to improve water solubility, antioxidant properties, and even biological activity [25]. For instance, polyhydroxy fullerenes (PHFs) exhibit properties suitable for biomedical applications, such as water solubility, biodegradability, biocompatibility, and hypoallergic response. It has been shown that PHFs can inhibit cancer tumor growth and positively regulate the immune system [26]. The same is valid for carboxylated fullerenes [27]; for instance, $C_{60}[C(COOH)_2]_3$ is well known for its high biological activity in plants [28] and within mitochondrial dynamics [29].

Since the evaluation of novel drugs is a task that requires significant human and material resources, innovative strategies have been formulated as alternatives. Quantitative structure–activity and quantitative structure–property relationships (QSAR/QSPR) are a paradigm that can be useful in choosing promising molecules, considering the information on inactive and active compounds, through in silico approaches. According to the QSAR/QSPR paradigm, a given activity/property, f, can be modeled using a set of quantitative descriptors, $x_1, x_2, x_3,..., x_n$, theoretically determined or measured by experiments [30]. A relationship $f(x_1, x_2, x_3,..., x_n)$ can be defined to predict the activity or property of molecules after the evaluation of their quantitative descriptors. However, the QSAR/QSPR paradigm does not explain how to select the descriptors or how to build the mathematical function. Consequently, the following paragraphs discuss basic concepts about selecting descriptors and regression techniques implemented in this manuscript.

Lipinski’s rule of five is a compendium of guidelines commonly used to determine if a molecule can be proposed as an orally delivered drug according to its physicochemical properties. According to this rule, a drug compound should have a molecular weight below 500 g/mol, a octanol–water partition coefficient (LogP) below 5, less than five hydrogen bond donor sites, and less than ten hydrogen bond acceptors sites. It is possible to add two other conditions, namely polar surface area (PSA) ≤ 140 Å^2 and less than ten rotatable bonds [31]. Taking advantage of the readiness of these quantities in public datasets, the current study proposes some of these quantities as potentially suitable descriptors for predictive models. Besides, Pearson's hard–soft acid–base (HSAB) theory suggests other descriptors to describe and predict the interactions between chemical species, such as those between a drug molecule as a ligand and a protein [32]. These quantitative values are based on the vertical ionization energy (I) and electron affinity (A). According to Koopmans' theorem, both can be approximated by $I = -E_{\text{HOMO}}$ and $A = -E_{\text{LUMO}}$, where E_{HOMO} is the energy of the highest occupied molecular orbital (HOMO), and E_{LUMO} is the energy of the lowest unoccupied molecular orbital (LUMO). It is advantageous to combine these properties to find out if an interaction between two species will occur and to obtain new quantitative relationships. Another helpful descriptor is the global electrophilicity, calculated as $\omega = \chi^2/2\eta$ [33]. Electrophilicity is related to the energetic stabilization that a species gains by obtaining an additional electron.

Methods

First, 42 drugs related to chemotherapy treatments for breast cancer were proposed. Although the most notable fullerene derivatives for biological applications are those with several hydrophilic groups, the carboxylic acid derivative C_{60}–COOH has been studied as well. Baglayan and coworkers carried out a conformation analysis within DFT to obtain the ground state structure for C_{60}–COOH [34]. In addition, they discussed its usage as a potential drug carrier for the antimetabolic and anticancer drug 5-fluoruracil [34]. Similarly, Parlak and Alver reported a theoretical study on the interactions and stability of paracetamol complexes with C_{60}–COOH [35]. Consequently, this work proposes the interaction of C_{60}–COOH fullerene with anticancer drugs. As a complement, a water-soluble fullerene predicted as stable at the normal human body temperature was proposed to study the interactions with doxorubicin and gemcitabine [36]. The water-soluble fullerene is introduced to avoid known mutagenic reactions related to breast cancer [36]. It was also studied as a potential carrier for bedaquiline, an agent against tuberculosis [37]. The current study only considered molecules and complexes formed with up to 100 atoms to be affordable with our computational resources.

A set of descriptors was chosen to build the dataset, including molecular weight and pK_a [38]. Also, LogP was included, as a descriptor associated with the concentration of a given substance in the aqueous phase of a two-phase octanol–water mixture [39]. Similarly, LogS, related to the water solubility of a substance, was considered. Besides, PSA, as molecular surface associated with charge accumulation due to heteroatoms and polar groups, as well as polarizability (α) associated with the tendency of a given molecule to acquire an electric dipole moment in the presence of an external electric field were taken into account. The mentioned QSAR/QSPR descriptors were obtained from the Drugbank dataset (https://go.drugbank.com). Initial drug structures and connectivity were also obtained from the simplified molecular input line entry specification (SMILES) retrieved from Drugbank.

Molecular mechanics and density functional-based tight binding (DFTB) with dispersion and solvation corrections were used to obtain the optimized structures of the molecules under study and to compute E_{HOMO}, E_{LUMO}, and ω as quantitative descriptors. As an alternative to the most robust but computationally more expensive density functional theory (DFT) method, DFTB was used. A reference electron density ρ_0 represents the sum of the neutral atomic densities [40]. Within the third-order approach DFTB3, the ground state density $\rho(r)$ is obtained as the reference density ρ_0 perturbed by density fluctuations $\delta\rho$, that is,

$$\begin{aligned} E_{\text{DFTB3}}\left[\rho_0+\delta\rho\right] &= E_0\left[\rho_0\right]+E_1\left[\rho_0+\delta\rho\right] \\ &\quad +E_2\left[\rho_0+\left(\delta\rho\right)^2\right]+E_3\left[\rho_0+\left(\delta\rho\right)^3\right]. \end{aligned} \quad (1)$$

For all calculations within DFTB3, the 3OB parameter set was used [41]. To carry out the global optimization procedure, Balloon 1.8.2 [42] and DFTB+ 17.1 [40] were used for the initial conformational study by genetic algorithms and final optimization at the DFTB3 level, respectively. London dispersion forces were considered in the DFTB3 and global optimization procedures by Lennard-Jones potentials, as implemented in UFF and MMFF94 force fields, respectively. The solvent effect was included by the Born solvation model within DFTB3. The study considered the chemotherapy drugs isolated and interacting with pristine C_{60} fullerene as well as its carboxylic acid derivative C_{60}–COOH. Eight initial drug–fullerene structures were proposed to obtain their global optimization by means of DFTB3. The drugs were initially set at 1.5 Å of minimal distance from the fullerene. Once the global optimization was done, the same steps as for the isolated drugs were carried out for the molecular docking. The datasets were modified to take into account the effect of the fullerenes. Also, the validation set was reduced because of the large size of the complexes.

The atypical chemokine receptor 3, also known as CXCR7 or G-protein-coupled receptor 159 (GPR159) [16,18,43], was selected as the target protein for molecular docking. The iterative assembly refinement server (I-Tasser) was used to produce an initial structure for the CXCR7 protein by the homology approach. The sequence was extracted from the UniProtKB/

Swiss-Prot dataset. From all homology structures produced by the I-Tasser server, the one with the highest confidence coefficients was selected to produce a reliable initial structure [44]. The lowest-energy structure, as in the study of Muthiah and coworkers [45], was validated using PROCHECK [46] to check the quality of the protein structure. The PDB produced with the previous step was subsequently optimized by an energy minimization through Amber force fields using the USCF Chimera 1.14 toolkit [47]. The secondary structural features were stabilized by TMpred [48] and HMMTOP [49] during energy minimization. Last, the protein was prepared by setting atomic charges and hydrogen atoms and merging the nonpolar groups. Once the structures were optimized, molecular docking was performed with the CXCR7 protein, using Autodock Vina 1.1, to obtain the docking score, established hydrogen bonds, and the binding site (pocket). The above was done for all drugs in the dataset and an external validation set.

IBM Watson AI was used to build the models and to predict the docking score through the Extra Trees regressor algorithm [50,51]. It was also used to obtain the most significant quantum descriptors used in each model. Extra Trees, an abbreviation of “extremely randomized trees”, is a mathematical method used to estimate a relationship between data and the covariates [52]. The Extra Trees algorithm creates many decision trees [52], but the sampling of each one is random. Thus, a dataset for each tree contains unique samples. The optimization of the hyperparameters associated with the decision trees obtained was performed by the derivative-free global search algorithm known as RBfOpt, which fits a radial basis function mode to accelerate the discovery of the hyperparameters [53]. All the above was used through the AutoAI tool within IBM Watson, an automatized routine to select the model with the best performance among those available in the platform. Since this method does not produce exportable mathematical models, another approach was used as detailed below [50].

Multiple linear regression (MLR) could be a tool to solve the problem in a complementary way to Extra Trees regression. MLR is a mathematical model that can be seen as an extension of linear regression. In terms of n input variables, $x_1, x_2, \ldots, x_n$, the outcome y can be expanded by the following linear expansion [54]:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \ldots + \beta_n \times x_n. \qquad (2)$$

In Equation 2, β_k are the partial regression coefficients, and β_0 is the value of y when all variables are set to zero.

To obtain the AI and MLR models, a fivefold approach was implemented, by using 80% of the data available to obtain the predictive model as training set and the remaining 20% as testing set. Supporting Information File 1 gives the results of the cross-validation for all the models reported in the current manuscript. Once the models were built, an additional external validation set was used to obtain evaluation metrics and to determine the most accurate models between methodologies. The metrics proposed to evaluate the performance of the predictive models were mean squared error (MSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean squared error (RMSE). These metrics were computed as follows:

$$\text{MAPE} = (1/n)\sum_{i=1}^{n}\left|(y_i - \hat{y}_i)/y_i\right| \times 100\%,$$

$$\text{MAE} = (1/n)\sum_{i=1}^{n}\left|(y_i - \hat{y}_i)\right|,$$

$$\text{MSE} = (1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

and

$$\text{RMSE} = [(1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]^{1/2}.$$

Here, y_i is the docking score for compound I, $\hat{y}_i$ is the estimated value of the docking score for compound I provided by the model. The workflow diagram in Figure 1 summarizes the procedure followed to obtain the models.

Results and Discussion

Table 1 presents the quantum descriptors proposed for the current study and the symbols used for them. The physical unit of each descriptor, as well as references to their usage in similar QSAR/QSPR models, were included as well.

Isolated drugs

A dataset containing all the descriptors of Table 1 for 33 drugs was created to obtain the predictive models. Also, another nine compounds were considered to build an external validation set, allowing for the comparison between methodologies (Supporting Information File 1, Table S1). In the case of the training set, the molecular weight was obtained with values between 130.08 and 915.4 g/mol. Water solubility values varied between 0.0004 mg/mL and 22.3 mg/mL. The LogP values varied between −2 and 6.54, whereas LogS ranged from −6 to −1.1. Besides, pK_a values varied between −8 and 14.55. The hydrogen acceptor count varied widely between 2 and 13, whereas the hydrogen donor count varied between 0 and 6. In addition, the polar surface area had variations between 12.47 and 221.29 Å^2. The cases of thiotepa and aldoxorubicin were not considered because they are part of the validation set. Rotatable bonds were obtained ranging from 0 to 15. The polarizability varied from 9.46 to 87.46 Å^3; everolimus was excluded as part of the

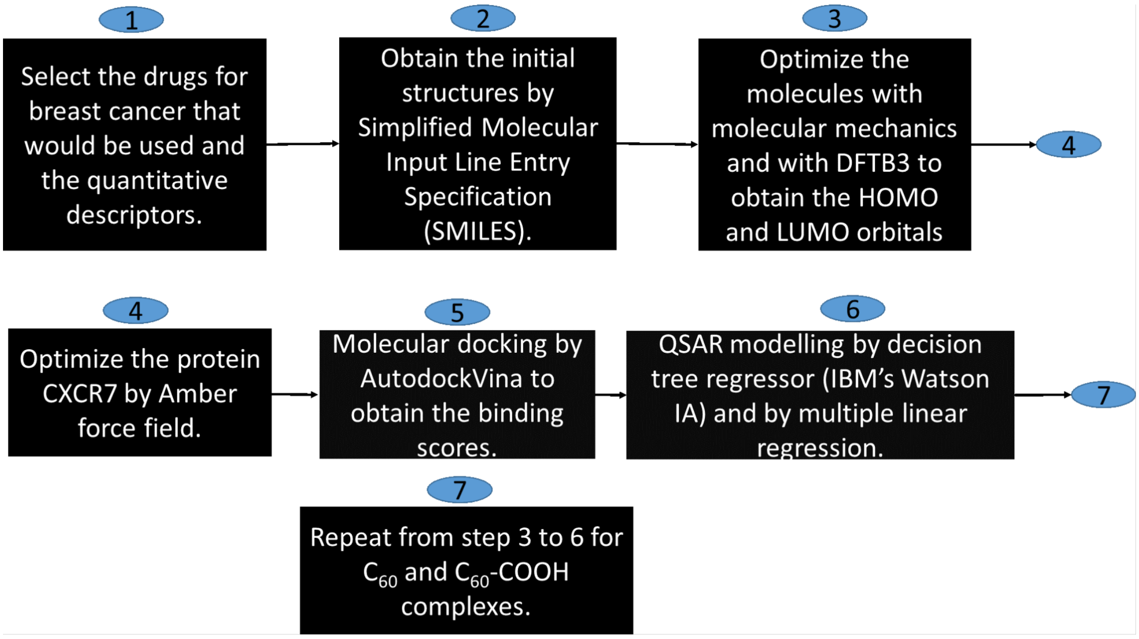
1
Select the drugs for breast cancer that would be used and the quantitative descriptors.
2
Obtain the initial structures by Simplified Molecular Input Line Entry Specification (SMILES).
3
Optimize the molecules with molecular mechanics and with DFTB3 to obtain the HOMO and LUMO orbitals
4
4
Optimize the protein CXCR7 by Amber force field.
5
Molecular docking by AutodockVina to obtain the binding scores.
6
QSAR modelling by decision tree regressor (IBM's Watson IA) and by multiple linear regression.
7
7
Repeat from step 3 to 6 for C60 and C60-COOH complexes.

Figure 1: Workflow diagram of the stages during modeling.

Table 1: Quantitative descriptors proposed to model the docking score of the isolated drugs, as well as of those modified with fullerenes C_{60} and C_{60}–COOH, interacting with the protein CXCR7.

Variable	Descriptor	Symbol	Unit	Reference
x_1	molecular weight	MW	g/mol	[55,56]
x_2	water solubility	WS	mg/mL	[57]
x_3	octanol–water partition coefficient	LogP	—	[58,59]
x_4	solubility coefficient	LogS	—	—
x_5	acid dissociation constant	pK_a	—	[60]
x_6	hydrogen acceptor count	Ac	—	[61]
x_7	hydrogen donor count	Dn	—	[61]
x_8	polar surface area	PSA	Å^2	[62]
x_9	rotatable bond count	RBC	—	[63]
x_{10}	polarizability	α	Å^3	[64]
x_{11}	number of rings	NOR	—	[65]
x_{12}	energy of HOMO	E_{HOMO}	eV	[66,67]
x_{13}	energy of LUMO	E_{LUMO}	eV	[66,67]
x_{14}	electrophilicity	ω	eV	[66,67]

external validation set. Also, values of number of rings were obtained from 0 to 9. The energy of the HOMO was computed ranging from −7.400 to −4.392 eV, and the LUMO energy from −5.341 to −0.889 eV. Finally, the electrophilicity varied from 2.12 to 180.39 eV. Figure S1 (Supporting Information File 1) shows the correlation matrix between the ten most relevant quantum descriptors used to obtain the mathematical models. There are significant correlations between the molecular weight and the polarizability of about 0.93 and between polarizability and the number of rings of about 0.88. Also, molecular weight and number of rings, as well as WS and LogS exhibited considerable correlations, with values of 0.87 and 0.73, respectively. However, all variables showed a correlation below 0.95. Once the drugs were optimized, blind molecular docking was performed with the CXCR7 protein to obtain the docking score, number of established hydrogen bonds, and the protein residues interacting with the ligands in a coordination sphere of 3 Å. The results obtained with Autodock Vina [47,68] for training–testing and validation sets are shown in Table 2.

The docking scores ranged from −10.1 to −4.6 kcal/mol for the training set. The molecule with the most significant bond strength, according to its docking score, was olaparib, whereas the one with the lowest bond strength was fluorouracil. The number of hydrogen bonds was computed ranging from 0 to 5. It is important to note that the number of hydrogen bonds is not directly related to the docking score since there are weak and strong hydrogen bonds. This assumption was proved by the analysis performed to obtain the predictive models, as discussed below.

According to the results of the interacting residues annotated in Table 2, two things can be highlighted. First, the analyzed isolated drugs bind inside the protein CXCR7 (Figure 2); second, the pocket is similar for several analyzed drugs. For example, the leucine residue Leu297 is shared by eleven drugs, indicating that their binding zone is close to each other. Comparing the interacting residues with those recently obtained by Muthiah et

Table 2: Docking score, number of established H-bonds, and protein–ligand interacting residue in three-letter symbol, up to 3 Å distance. Drugs marked with an asterisk were used as external validation set.

Ligand	Docking score (kcal/mol)	H-bonds	Interacting residues
doxorubicin	−8.4	5	Asp275, Phe294, Leu297, Tyr200, His298, Arg197, Ser198, Cys196, Asp179, Trp100
neratinib maleate	−8.5	2	Ile276, Leu297, Phe294, His298, Tyr 268, Gln30, Ser216, Ser198, Phe129, Asp179, Leu183, His121, Phe124
epirubicin	−8.1	2	Asp275, Leu297, Tyr200, Ser198, Arg197, Cys196, His121, Trp100, Leu104, His298, Phe294, Gln301,
lapatinib ditosylate	−8.9	1	Glu213, Ile276, Val272, Leu297, Tyr200, His298, Arg197, Ser198, Leu183, His121, Cys196,
fulvestrant	−7.9	0	Leu297, Ile276 Val272, Tyr268, Gln301, Trp100, Phe124, Cys196, Asp179, Ser198, Ser216, Arg197
dinaciclib	−8.4	1	Leu297, Gln301, Tyr268, Trp100, Cys196, Arg197, Phe124, Leu128, Ser198, Asp179, Phe129, Met212, Ser216
abemaciclib	−9.8	1	Asn108, Trp100, Trp110, Leu104, Arg197, Lys40, Asn36, Pro38, Leu43, His298
gemcitabine	−6.2	0	Asn36, Phe294, Met37, Pro38, Asn39, Leu43, Lys40, Leu104, Asn108
voruciclib	−8.8	0	Asn36, Pro38, Asn39, His298, Leu43, Lys40, Leu104, Asn108, Trp100, Ser103, Trp110, Arg197
fluorouracil	−4.6	1	Asn319, Asn321, Lys73, Ile83, His80
letrozole	−8.1	0	His298, Gln301, Arg197, Leu104, Trp100, Ser198, His121, Trp110, Leu183
olaparib	−10.1	2	Tyr268, Val272, Ile276, Glu213, Leu209, Ile205, Met212, Tyr200, Phe199, Ser198, Leu128, Phe124
paclitaxel	−9.6	0	His298, Phe294, Leu297, Tyr268, Gln301, Val272, Leu104, Trp100, Phe124, His121, Trp110, Arg197, Ser198, Leu209, Tyr200, Leu183
seliciclib	−7.6	0	Leu183, Asp179, Ser125, Phe124, Phe129, Leu128, Ser216, Tyr268, Leu297
ixabepilone	−8.2	0	Leu297, Val272, Tyr268, Trp265, Phe124, Arg197, His298, Leu128, Ser198, Leu209, Tyr200, Met212, Ser216
anastrozole	−7.7	0	Met212, Leu209, Ser216, Tyr200, Phe129, Asp179, Ser125, Ser198, Phe124, Leu183, His121
pentostatin	−6.3	3	Phe294, Lys40, Leu104, Trp100, Asn108, Trp110
alvocidib	−8.6	1	His298, Leu43, Lys40, Leu104, Trp100, Asn108, Ser103, Trp110, Arg197
methotrexate	−8.5	3	Leu104, Gln301, Tyr268, Val272, Trp100, Trp110, Cys196, Phe124, His121, Ser198, Leu183, Asp179, Phe129, Glu213, Ser216
milciclib	−9.2	0	Lys40, Phe294, Leu297, Leu104, Asn108, Tyr268, Arg197, Val272, His121
ribociclib	−9.5	1	Lys40, Leu104, Leu43, Tyr195, Cys196, Arg197, Phe124, His121, Ser198, Leu183, Asp179

Table 2: Docking score, number of established H-bonds, and protein–ligand interacting residue in three-letter symbol, up to 3 Å distance. Drugs marked with an asterisk were used as external validation set. (continued)

exemestane	−8.0	0	Phe294, Leu43, Lys40, Leu104, Trp110
tamoxifen	−7.9	1	Asn39, Pro38, Leu43, Lys40, Leu104, Trp110, Tyr195, Arg197
idarubicin	−8.8	1	Phe294, His298, Leu43, Leu104, Asn108, Ser103, Trp110, Cys196, His121
palbociclib	−8.3	1	Phe294, Asn36, Pro38, Leu43, Lys40, Leu104, Asn108, Arg197, Cys196
toremifene	−7.4	0	Leu43, Asn39, Leu104, Pro38, Lys40, Asn36, Phe294, Trp110, Tyr195, Arg197
vinblastine	−9.5	2	Leu43, Asn39, Leu104, His298, Phe294, Lys40, Asn108, Trp110, Arg197, Tyr195, Ile205, Leu209
pirarubicin	−9.4	1	Val272, Leu297, Tyr268, Gln301, Met212, Ile205, Leu209, Glu213, Arg197, Phe199, Trp110, Cys196, Phe199,
roniciclib	−8.4	1	Asn39, Met37, His298, Phe294, Pro38, Asn36, Trp100, Leu104, Lys40, Ser103, Asn108, Trp110,
capecitabine	−7.7	2	Gln301, Phe129, Phe124, Ser198, Arg197, His121, Cys196, Trp110, Leu183
sulfanilamide	−5.4	2	Tyr268, Val272, Glu213, Ser216, Leu128, Phe129, Met212, Ser125
zoledronic acid hydrate	−5.7	1	Tyr268, Glu213, Ser216, Phe129, Phe124, Ser125, Ser198, Met212, Asp179
pamidronic acid	−4.8	4	Val272, Tyr268, Trp265, Val217, Glu213, Ser216, Val215, Phe129, Met212, Val175, Phe124, Ser125, His121, Ser76, Tyr200, Asp179
docetaxel*	−9.5	2	Leu209, Glu213, Met212, Ser198, Ser216, Arg197, Leu183, Phe124, His121, Trp110, Ile276, Val272, Leu297, Tyr268, His298, Gln301, Leu104
topotecan*	−8.2	2	Tyr268, His298, Gln301, Trp100, Leu104, Trp110, Cys196, Phe124, Leu183, Ser198
tucatinib*	−10.1	1	Val272, Tyr268, Leu104, Arg197, Ser198, Trp110, His121, Phe124, Asp179, Leu183
aldoxorubicin*	−8.4	1	Ile205, Leu209, Tyr200, Ser198, Asp179, Arg197, Leu193, Asp179, His121, Phe124, Ser125, Trp110, Asn108, Leu104, His298, Gln301, Leu297, Val272, Tyr268
vinorelbine*	−9.1	0	Leu297, Tyr268, Phe294, His298, Gln301, Leu43, Leu47, Leu104, Arg197, Trp110, Ser198, Tyr200, Leu209
etoposide*	−9.3	2	Lys 40, Leu43, Asn108, Leu104, Trp100, Trp110, His121, Phe124, Leu183, Gln301
pemetrexed*	−9.1	2	Leu104, Ser103, Trp100, Trp110, Cys196, His121, Phe124, Leu183, Ser198, Tyr200, Val272, Glu213, Ser216
everolimus*	−9.2	3	Leu209, Met33, Asn36, Phe294, His298, Asn39, Leu43, Lys40, Leu104, Asn108, Tyr195, Arg197
thiotepa*	−3.9	1	Tyr268, Val272, Ser216, Phe124

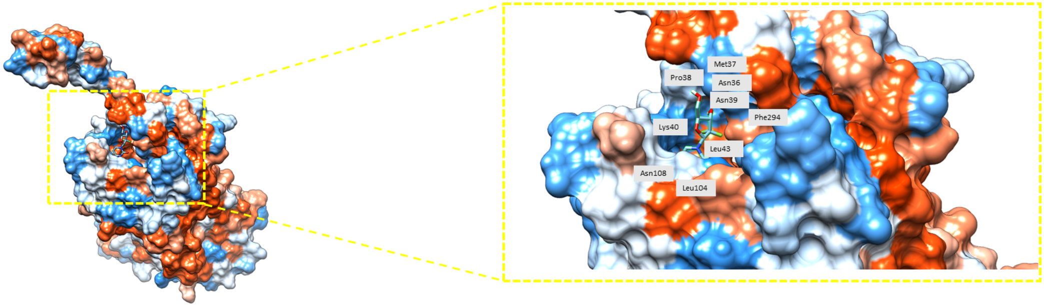

Figure 2: Gemcitabine after blind docking with the protein CXCR7. Blue color indicates the most hydrophilic sites and orange-red the most hydrophobic ones. Interacting residues are labeled as well.

al. [45], it is possible to conclude that the pockets are similar. For instance, doxorubicin was obtained in both cases with Asp179, Cys196, and Trp100 as interacting residues. Also, similar pockets could be obtained because the selected drugs are mostly designed to serve as chemotherapy agents for breast cancer. Thus, it is possible to assume that several drugs share a common mechanism of action and, subsequently, a common protein target, such as CXCR7. For instance, gemcitabine, shown in Figure 2, shares the pocket within CXCR7 with several drugs.

The quantitative descriptors included in the produced models are discussed next. Table 3 contains the used quantitative descriptors and the importance that each one has in the mathematical models. The docking score was the predicted variable in all cases. In addition, the correlation matrix (Supporting Information File 1, Figure S1) was used to build the models. Among the IBM Watson AI models, the first one was obtained using all the initially proposed quantitative descriptors (Table 1). In the second and fourth models, polarizability was not used because, as shown in the correlation matrix, it was found to be closely related to molecular weight. The third and fifth models did not consider molecular weight because of the same relationship with polarizability. Also, the descriptors pK_a, Ac, and PSA were not considered because their importance in the previous models was below 10% (Table 3). Although conceptually different, E_HOMO was discarded because it is related to E_LUMO and electrophilicity. The last model did not include α because of its relationship with MW. It is important to notice that the most important variables in the six models were NOR, polarizability, LogS, MW, and WS. The least important variables in the six models were E_LUMO, pK_a, PSA, E_HOMO, and Ac; they all had less than 10% importance in all models. Hence, the computed E_HOMO and E_LUMO values were not particularly useful in predicting the ligand–protein docking score and, subsequently, the docking score.

To compare the performance offered by the Extra Trees algorithm of Watson AI, a comparison with a family of MLR models was made. Supporting Information File 1 contains the cross-validation for all reported MLR models. Table S2 (Supporting Information File 1) shows the docking scores obtained for the validation set by using both methodologies. With these values, it is possible to appreciate the difference between the docking score obtained directly by molecular docking through Autodock Vina and the prediction of the mathematical models for the external validation set.

To clearly state the performance comparison between AI and MLR, Table 4 reports the values computed for the evaluation metrics proposed for each model. The MSE ranged from 0.30 to 1.73 kcal^2/mol^2, whereas the MAPE varied from 6.1 to 16.37%. Also, MAE values from 0.46 to 1.13 kcal/mol and RMSE values from 0.55 to 1.32 kcal/mol were obtained. The above shows that both AI and MLR approaches accurately model the protein–ligand docking score, yielding higher confidence in the case of Extra Tree regressor models. The best performance, according to the computed minimum errors, was obtained in the

Table 3: Input variables (IV) and output importance (OI) of six Extra Tree regressor models obtained from IBM Watson. Variables are annotated according to Table 1. The best model, according to the MAPE values, is highlighted in bold.

Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)
NOR	100	WS	100	**WS**	**100**	LogS	100	NOR	100	LogS	100
α	57	NOR	97	**LogS**	**99**	MW	53	α	74	NOR	15
LogS	43	LogP	94	**α**	**56**	NOR	26	LogS	49	E_LUMO	7
WS	38	LogS	87	**NOR**	**39**	WS	17	WS	31	MW	6
MW	33	MW	20	**ω**	**18**	ω	12	LogP	19	ω	5
LogP	32	Dn	14	**E_LUMO**	**6**	E_LUMO	5	Dn	13	RBC	0
Ac	8	RBC	12	**E_HOMO**	**5**	E_HOMO	3	E_LUMO	5	Dn	0
Dn	6	E_LUMO	7	**PSA**	**4**	PSA	3	RBC	0	Ws	0
E_HOMO	2	PSA	4	**pK_a**	**1**	pK_a	1	ω	0	LogP	0
RBC	1	E_HOMO	3	**LogP**	**1**	RBC	0	—	—	—	—
PSA	1	pK_a	2	**RBC**	**0**	Ac	0	—	—	—	—
ω	1	Ac	1	**Dn**	**0**	Dn	0	—	—	—	—
pK_a	0	Ω	0	**Ac**	**0**	LogP	0	—	—	—	—
E_LUMO	0	—	—	—	—	—	—	—	—	—	—

Table 4: Comparison metrics obtained by the use of AI and MLR in the case of isolated drugs. The best model, according to the MAPE values, is highlighted in bold. The values were computed relative to the validation set.

Error	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR
MSE (kcal^2/mol^2)	0.93	1.69	0.64	1.73	**0.30**	**1.02**	0.43	1.51	0.82	1.37	0.73	1.10
MAPE (%)	11.51	16.11	9.65	16.37	**6.17**	**11.98**	6.70	15.69	10.92	14.10	10.34	12.49
MAE (kcal/mol)	0.77	1.11	0.64	1.13	**0.46**	**0.82**	0.51	1.08	0.71	0.98	0.66	0.83
RMSE (kcal/mol)	0.97	1.30	0.80	1.32	**0.55**	**1.01**	0.66	1.23	0.91	1.17	0.85	1.05

case of Watson AI model 3. In contrast, the maximum error was obtained in MLR model 2. Thus, the variables denoted in this work can be used for other authors to propose novel chemotherapy drugs assuming CXCR7 as a target.

As mentioned above, the best model was AI model 3 with a MAPE of about 6.17%. Since the AI models are not exportable, our best model is represented by the following functional form: $\mathrm{AI3_{DRUG}} = f(\mathrm{WS}, \mathrm{LogS}, \alpha, \mathrm{NOR}, \omega, E_{\mathrm{LUMO}}, E_{\mathrm{HOMO}}, \mathrm{PSA}, \mathrm{p}K_a, \mathrm{LogP})$. The model with the lowest MAPE among those obtained by MLR, computed as 11.98%, is model 3, represented as follows:

$$\begin{aligned}\mathrm{MLR3_{DRUG}} = & -5.288 + 0.049\times\mathrm{WS} + 0.318\times\mathrm{LogS}\\ & +0.048\times\alpha - 0.637\times\mathrm{NOR} + 0.009\times\omega\\ & -0.054\times E_{\mathrm{HOMO}} + 0.113\times E_{\mathrm{LUMO}}\\ & -0.008\times\mathrm{PSA} - 0.234\times\mathrm{LogP} - 0.014\times\mathrm{p}K_a.\end{aligned} \quad (3)$$

Thus, the AI was useful in selecting the most relevant variables for the formulation of the linear, accurate, and exportable model. Thus, AI is a valuable guide to obtain mathematical models with other methodologies such as MLR. Another significant model is MLR model 6 with the second lowest MAPE of 12.49%. It is worth noting that the model is compact and includes the computed descriptors $\mathrm{E_{HOMO}}$ and ω, as well as experimentally determined ones.

$$\begin{aligned}\mathrm{MLR6_{DRUG}} = & -4.2179 + 0.4131\times\mathrm{LogS} - 0.5274\times\mathrm{NOR}\\ & +0.2869\times E_{\mathrm{HOMO}} + 0.0009\times\mathrm{MW}\\ & +0.0088\times\omega.\end{aligned} \quad (4)$$

Although the best AI model was model 3, model 4 is relevant as well. It has the following functional form: $\mathrm{AI4_{DRUG}} = f(\mathrm{LogS}, \mathrm{MW}, \mathrm{NOR}, \mathrm{WS}, \omega, E_{\mathrm{LUMO}}, E_{\mathrm{HOMO}}, \mathrm{PSA}, \mathrm{p}K_a)$. For this model, the MAPE is about 6.70%. The variables used can be mostly evaluated by computational methods, except for pK_a and LogS.

Drugs modified with C_{60}

Since this study aims to elucidate the potential use of AI suites, such as Watson, to predict the docking score of pristine and modified chemotherapy drugs, the following paragraphs detail the extension of our datasets and models to drugs modified with potential nanocarriers. First, a dataset with 28 drugs, extracted from public datasets or modified from the data annotated in the previous case, was built with the corresponding quantitative descriptors to study complexes of the drugs with fullerene C_{60} or a simple C_{60}–COOH derivative [29]. The resultant dataset is shown in Supporting Information File 1, Table S3. Complexes of the drugs with C_{60} or its derivative with more than one hundred atoms were excluded to save computational resources. Because of that, some maximums and minimums were modified in the dataset. Also, quantities such as the molecular weight or number of rings were shifted to the correspondent values for the drug–fullerene complexes since these modifications are only additive constants. Molecules interacting with C_{60} were studied in this subsection and those modified with the fullerene derivative are the described in the subsequent subsection.

In case of the complexes with C_{60}, the molecular weight is obtained with values above 907.1639 g/mol. The above imposes some restrictions on the usage of fullerene derivatives as drug nanocarriers, since it is accepted that common pharmacological agents applied in topical therapies are under 500 Da [69]. For simplicity, WS, LogP, LogS, pK_a, and α values were taken from isolated drugs. In the case of Dn, Ac, RBC, and PSA, values of the isolated drugs were also used. This is because fullerene is not expected to modify these descriptors since its chemical constitution lacks polar groups or donor/acceptor atoms. The number of rings was only modified with the addition of the fullerene rings from 32 to 39.

As stated in the Methods section, HSAB descriptors were computed for the drug–C_{60} complexes in their ground states while drugs with C_{60} were structurally optimized at the DFTB3 level. Since introducing a species known to act as an electron acceptor, such as fullerene C_{60}, could modify the electron structure of the modified species [70,71], the energies of the frontier orbitals were recomputed for the modified drugs in their ground states. The energy of the HOMO varied between −3.458 and −5.718 eV. The energy of LUMO ranged from −3.179 to −5.388 eV. The electrophilicity, computed through Koopman's theorem, varied from −134.88 to 280.8 eV. Once the drugs with fullerene C_{60} were globally optimized, molecular docking was conducted with the CXCR7 protein to acquire the docking score, number of established hydrogen bonds, and protein residues interacting with the complex at a distance of 3 Å. The results obtained with Autodock Vina for the 24 complexes and the validation set are presented in Table 5.

The protein–ligand docking score between drug–C_{60} and CXCR7 varied from −13.6 to −9.2 kcal/mol. These values are consistently higher than those computed for the isolated drugs.

Table 5: Docking score, number of established H-bonds after docking, and protein–ligand interacting residues up to 3 Å distance obtained for the drugs modified with C_{60} fullerene. Drugs with the asterisk were used as an external validation set.

Ligands	Score (kcal/mol)	H-bonds	Interacting residues
doxorubicin	–9.2	2	Arg323, Ser316, Val313, Leu326, Phe330, Leu340, Leu343, Val64, Ile57, Ile60
epirubicin	–9.2	0	Arg323, Ser316, Leu326, Tyr322, Leu340, Leu343, Lys342, Val64, Pro312, Ala61, Ile60
lapatinib ditosylate	–9.4	0	Arg323, Ser316, Phe330, Leu326, Tyr322, Leu340, Lys342, Leu343, Val64, Ala61, Ile60
fulvestrant	–11.7	0	Tyr322, Pro312, Leu326, Ala61, Val64, Ile60, Leu340, Leu343, Lys342
dinaciclib	–10.2	1	Val313, Ile60, Ala61, Val64, Leu326, Leu340, Thr340
abemaciclib	–10	0	Arg323, Leu326, Ser316, Tyr322, Leu340, Leu343, Lys342, Val64, Ala61, Ile60
gemcitabine	–10.9	1	Val313, Ile57, Ala61, Ile60, Val64, Leu326, Leu340, Leu343, Phe330
voruciclib	–10.2	0	Val64, Ala61, Ile57, Ile60, Val313, Leu326, Leu340, Thr341, Leu343
fluorouracil	–11.6	0	Pro312, Tyr322, Leu326, Leu340, Leu343, Lys342, Ala61, Val64, Ile60
olaparib	–12.4	2	Val313, Ala61, Ile60, Ile57, Arg323, Leu326, Leu340, Lys342
ixabepilone	–11.7	2	Arg323, Ser316, Val313, Ala61, Val64, Ile60, Leu326, Leu340, Lys342, Leu343
alvocidib	–12.1	0	Val313, Ser316, Tyr322, Leu326, Leu340, Leu343, Lys342, Val64, Ala61, Ile60, Ile57
methotrexate	–9.7	2	Ser316, Pro312, Tyr322, Leu326, Val65, Ala61, Val64, Ile60, Phe330, Leu340, Leu343, Thr341, Lys342
ribociclib	–12.4	0	Ser198, Arg197, Cys196, Tyr195, Trp110, Asn108, Leu104, Lys40, Ser90, Phe94, Asn36, Asp30, Val28
exemestane	–13.6	0	Val313, Pro312, Ser316, Tyr322, Val65, Ala61, Ile57, Val64, Ile60, Leu326, Leu340, Leu343, Lys34
tamoxifen	–9.8	0	Ile66, Leu84, Ile70, Ile88, Leu91, Trp92
idarubicin	–10.1	1	Ser316, Tyr322, Leu326, Leu340, Leu343, Lys342, Ile57, Ile60, Ala61, Val64
palbociclib	–11.2	0	Val313, Ala61, Ile57, Ile60, Val64, Leu343, Leu326
toremifene	–9.9	0	Val313, Ile57, Ile60, Val64, Leu326, Ser350, Leu340, Leu343
roniciclib	–11	2	Met33, Pro35, Asn36, Asp30, Val28, Asn39, Pro38, Ile27, Leu43, Lys40, His107
capecitabine	–9.5	0	Ser316, Tyr322, Leu326, Ala61, Val64, Ile60, Ile57, Leu340, Leu343, Lys342
sulfanilamide	–10.6	1	Cys165, Arg162, Ile166, Leu84, Ile70, Trp92
zoledronic acid hydrate	–9.8	2	Val313, Ile60, Ala61, Ile57, Leu326, Leu340, Lys342, Thr341
pamidronic acid	–9.9	1	Val313, Ser316, Tyr322, Pro312, Leu326, Val64, Ala61, Ile60, Leu340, Lys34, Leu343
topotecan*	–10.1	0	Lys342, Leu343, Leu340, Phe330, Leu326, Ser316, Val313, Ala61, Ile57, Val64, Ile60
tucatinib*	–11	0	Lys342, Leu343, Leu340, Tyr322, Leu326, Val64, Ala61, Ile60, Ile67
pemetrexed*	–11.5	0	Lys342, Thr341, Leu343, Leu340, Phe330, TTyr322, Val64, Ser316, Ala61, Ile60
thiotepa*	–9.9	0	Lys342, Leu343, Leu340, Tyr322, Leu326, Ser316, Val313, Val64, Ile60

In this case, examestane had the highest docking score, whereas epirubicine had the lowest docking score. The number of hydrogen bonds was between 0 and 2. Considering the results of the interacting residues, two things can be highlighted. First, drugs modified with fullerene C_{60} bind outside the protein; second, these binding sites are similar for the analyzed compounds (Figure 3). For example, the serine residue labeled as Ser316 is shared by 14 drugs, indicating that their binding zone is close to each other. Figure 3 shows the binding site between the protein and gemcitabine and the interacting residues.

The predictive QSAR/QSPR models were obtained as in the previous section. Table 6 shows the used quantum descriptors and the importance that each one has in the mathematical models for the drug–C_{60} complexes. The docking score was the predicted variable in all cases. The models for C_{60} were taken directly from the models of the isolated drugs. For example, model 1 for drug–C_{60} complexes considered variables with importance greater than zero from model 1 in the case of isolated drugs. The previous procedure was performed for all models. The most critical variables for the six models were LogP, pK_a, and PSA. The least important variables in the six models were WS and Ac because they have less than 10% importance in the models.

To give a quantitative reference about the performance of the predictive models, Supporting Information File 1, Table S4, shows the scores obtained for the validation set using both methods, the Extra Tree algorithm of IBM Watson and multiple linear regression. With these values, it is possible to appreciate the difference between the docking score obtained by molecular docking and the predictive models for the drug–C_{60} complexes. Table 7 shows the values obtained for the different evaluation metrics for each predictive model. The MSE ranges from 0.44 to 12.77 kcal^2/mol^2, and the MAPE varies from 4.97 to 31.5%. The MAE ranged from 0.5 to 3.26 kcal/mol, whereas the RMSE varied from 0.66 to 3.57 kcal/mol. The minimum

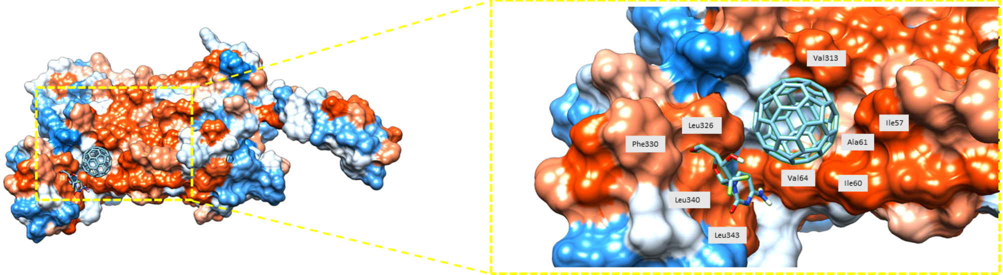

Figure 3: Binding site of gemcitabine–C_{60} and protein CXCR7. Blue color indicates the most hydrophilic sites and orange-red the most hydrophobic ones. Interacting residues are labeled as well.

Table 6: Input variables (IV) and output importance (OI) for six Extra Tree regressor models obtained from IBM Watson. Variables are annotated according to the notation introduced in Table 1. The best models, according to their MAPE values, are highlighted in bold.

Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)
LogP	100	pK_a	100	pK_a	100	**pK_a**	**100**	LogP	100	**MW**	**100**
PSA	97	LogP	55	PSA	69	**PSA**	**66**	Dn	58	**E_{LUMO}**	**96**
RBC	24	PSA	49	LogP	56	**LogS**	**32**	α	18	**ω**	**71**
E_{HOMO}	23	E_{LUMO}	29	E_{LUMO}	24	**E_{LUMO}**	**25**	E_{LUMO}	10	**NOR**	**34**
Dn	17	Dn	20	WS	8	**NOR**	**20**	NOR	6	**LogS**	**0**
ω	11	LogS	7	E_{HOMO}	5	**WS**	**8**	LogS	1	—	—
α	8	E_{HOMO}	1	ω	3	**E_{HOMO}**	**3**	WS	0	—	—
MW	4	Ac	0	α	3	**ω**	**1**	—	—	—	—
LogS	3	RBC	0	LogS	0	**MW**	**0**	—	—	—	—
NOR	1	MW	0	NOR	0	—	—	—	—	—	—
WS	1	NOR	0	—	—	—	—	—	—	—	—
Ac	0	WS	0	—	—	—	—	—	—	—	—

Table 7: Metrics obtained by using IA and MLR for drug–C_{60} complexes. The best models, according to their MAPE values, are highlighted in bold. The values were computed relative to the validation set.

Error	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR
MSE (kcal^2/mol^2)	1.62	12.77	3.9	0.86	3.61	3.75	2.38	**0.44**	1.96	2.22	**0.93**	1.77
MAPE (%)	10.99	31.5	18.62	7.58	17.3	17.36	12.96	**4.97**	12.6	12.8	**7.53**	11.94
MAE (kcal/mol)	1.17	3.26	1.97	0.77	1.8	1.8	1.38	**0.5**	1.36	1.34	**0.82**	1.24
RMSE (kcal/mol)	1.27	3.57	1.97	0.92	1.9	1.94	1.54	**0.66**	1.4	1.49	**0.96**	1.33

error for all four metrics was obtained in MLR model 4, while the maximum error was obtained in MLR model 1. Considering the MAPE, the best model, with a value of 4.97%, to predict the docking score is MLR model 4. The explicit form of this model is:

$$\begin{aligned}\mathrm{MLR4}_{\mathrm{DRUG+C60}} &= -10.253 + 0.179 \times \mathrm{p}K_\mathrm{a} - 0.003 \times \mathrm{PSA} \\ &+0.292 \times \mathrm{LogS} + 0.719 \times E_{\mathrm{LUMO}} \\ &-0.071 \times \mathrm{NOR} - 0.057 \times \mathrm{WS} \\ &-1.020 \times E_{\mathrm{HOMO}} + 0.006 \times \omega.\end{aligned} \tag{5}$$

This linear model exhibited a higher performance than the non-exportable approaches provided by the AI. However, one needs to remember that the variables included in model 4 were selected by the initial AI screening. Thus, the selection of variables using AI offers a significant improvement for modeling using other mathematical methods. In addition, all metrics obtained in the case of MLR model 4 are better than those calculated in the case of the best MLR model of the isolated drugs and are comparable to those of the best AI model. Another significant MLR model is the model 2 with a MAPE of 7.58% and the following linear function:

$$\begin{aligned}\mathrm{MLR2}_{\mathrm{DRUG+C60}} &= -12.884 + 0.141 \times \mathrm{p}K_\mathrm{a} + 0.139 \times \mathrm{LogP} \\ &-0.10 \times \mathrm{PSA} + 0.649 \times E_{\mathrm{LUMO}} \\ &-0.261 \times \mathrm{Dn} + 0.268 \times \mathrm{LogS} \\ &-0.828 \times E_{\mathrm{HOMO}}.\end{aligned} \tag{6}$$

The best AI model is model 6 with a MAPE of about 7.53% and the functional form $\mathrm{AI6}_{\mathrm{DRUG+C60}} = E(\mathrm{MW}, E_{\mathrm{LUMO}}, \omega, \mathrm{NOR})$; all variables can be evaluated by theoretical approaches without the necessity of experimental results. The other significant model from AI, model 1, yielded a higher MAPE value of about 10.99%. In this case, the functional form is $\mathrm{AI1}_{\mathrm{DRUG+C60}} = E(\mathrm{LogP}, \mathrm{PSA}, \mathrm{RBC}, E_{\mathrm{HOMO}}, \mathrm{Dn}, \omega, \alpha, \mathrm{MW}, \mathrm{LogS}, \mathrm{NOR}, \mathrm{WS})$. Despite the large number of variables, including theoretical and experimental ones, the error is larger than those of the previously discussed models.

Drugs modified with C_{60}–COOH

To elucidate the effect of a fullerene derivative, the carboxyfullerene C_{60}–COOH was chosen. A dataset with 19 drugs for the predictive model and four drugs as the validation set was built. The resultant dataset is shown in Supporting Information File 1, Table S5. As in the previous systems, the dataset was reduced to systems with less than 100 atoms. Because of this, ranges of the descriptors and their contributions to the predictive models were modified in the dataset. The molecular weights were increased to values ranging from 915.8022 to 1292.425 g/mol. As in the previous case, WS, LogS, pK_a, and LogP are the same as those obtained for the isolated drugs. The hydrogen acceptor count varied between 3 and 13, whereas the hydrogen donor count varied from 1 to 7. Since the carboxylic group is polar, polar surface area values, ranging from 49.77 to 243.22 Å^2, were modified. Also, after the introduction of the polar group, the RBC varied from 1 to 10.

The fullerene derivative C_{60}–COOH was expected to modify the electronic structure of the composed systems. In consequence, the energy of the HOMO of the complexes was recomputed for the globally optimized systems at the DBTB3 level with solvation effects; the results ranged from −3.504 to −5.164 eV. Similarly, the energy of the LUMO varied from −3.48 to −4.437 eV. The electrophilicity computed by Koopman's theorem had variations between 21.675 and 508.086 eV. Molecular docking was performed with the CXCR7 protein to obtain the docking score, number of established hydrogen bonds, and protein residues interacting with the complex at a distance of 3 Å. The results obtained with Autodock Vina for the 19 complexes and the validation set are shown in Table 8.

The resultant docking score between drug–C_{60}–COOH and CXCR7 ranged from −11.7 to −8.8 kcal/mol. As in the previous case, the water-soluble fullerene increased the docking score compared with the isolated drugs (Table 2). Modified ixabepilone had the highest docking score, and the capecitabine had the lowest. The numbers of hydrogen bonds were obtained in a narrow range from 0 to 3.

Table 8: Docking score, number of H-bonds established after docking, and interacting residues of CXCR7 with drug–fullerene C_{60}–COOH complexes at 3 Å distance.

Ligand	Score (kcal/mol)	H-bonds	Interacting residues
dinaciclib	−9.7	0	Lys342, Leu340, Phe330, Leu343, Leu326, Tyr322, Val64, Ile60, Ala61
gemcitabine	−10.6	0	Phe294, Asn36, Asp30, Leu104, Arg197. Trp100, Tyr195, Ser190, Asn108, Lys40, Leu43
voruciclib	−10.3	0	Lys342, Leu343, Leu340, Val64, Leu326, Ser316, Val313, Ala61, Ile60, Ile57, Val56
fluorouracil	−11.4	1	Lys342, Leu343, Leu340, Thr341, Phe330, Leu326, Tyr322, Ser316, Val64, Ile60
olaparib	−11.6	0	Lys342, Leu343, Leu340, Phe330, Val64, Leu326, Tyr322, Arg323, Ser316, Ile60, Ala61, Val313
ixabepilone	−11.7	0	Ser190, Glu193, His107, Lys40, Arg197, Ile205, Pro38, Val28, Ile27
alvocidib	−10.9	1	Lys342, Leu343, Leu340, Val64, Leu326, Tyr322, Ala61, Ile60, Ile57
methotrexate	−10.9	2	Asn108, Lys40, Tyr195, Ser190, Arg197, Ser198, Tyr200, Leu43, Gln301, Tyr268, His298, Phe294, Asn39, Pro38, Asn36, Met37, His291, Asp30, Val28
ribociclib	−11.2	0	Leu104, Asn36, Phe294, Met33, Val32, Arg197, Ser190, Trp110, Tyr195, Asp30, Ile205, Lys206
exemestane	−11.6	1	Lys342, Leu343, Leu340, Leu326, Phe330, Ala61, Ile60, Ile57, Val313
tamoxifen	−9.1	0	Leu340, Leu326, Phe330, Ser316, Ile60, Val313, Ala61, Ile60, Ile57, Val56
idarubicin	−10.6	0	Leu343, Leu340, Lys342, Val64, Tyr322, Pro312, Ser316, Val313, Phe330, Ile60, Ile57
palbociclib	−11.4	0	Ile166, Arg162, Cys165, Cys81, Asn85, Ile88, Trp92, Ile70, Val66
toremifene	−9	0	Leu343, Lys342, Ser350, Leu340, Val64, Leu326, Ala61, Ser316, Ile60, Ile57
roniciclib	−9.2	0	Lys342, Leu343, Leu340, Leu326, Arg323, Val64, Ala61, Ile60, Ile57
capecitabine	−8.8	0	Thr341, Leu343, Leu340, Val64, Phe330, Leu326, Ala61, Ile60, Ile57, Val313
sulfanilamide	−10.7	1	Lys342, Leu343, Thr341, Leu340, Leu326, Tyr322, Ser316, Pro312, Val64, Ala61, Ile60
zoledronic acid hydrate	−10	0	Lys342, Leu343, Val64, Val65, Leu326, Tyr322, Ser316, Val313, Pro312, Ala61, Ile57
pamidronic acid	−10.6	3	Leu122, Ile126, Ile123, Trp169, Trp92, Ile166, Asn85, Cys165
topotecan*	−12.6	2	Thr341, Phe330, Leu340, Leu343, Leu326, Val64, Tyr322, Ser316, Val65, Pro312, Val313, Ala61, Ile60, Ile57
tucatinib*	−10.9	1	Leu343, Lys342, Leu340, Thr341, Val64, Tyr322, Leu326, Ser316, Val313, Ile60, Ala61, Val56, Ile57
pemetrexed*	−9.9	1	Ile166, Arg162, Trp92, Ile88, Leu84, Asn85, Thr75, Cys81
thiotepa*	−9.4	0	Lys342, Leu343, Leu340, Thr341, Val64, Leu326, Phe330, Ile60

Considering the residues shown in Table 8 for the drug–fullerene complex with the protein, there are three possible binding sites. The first one is located inside the protein, near the pocket determined for isolated drugs. This binding site is near phenylalanine Phe294 and arginine Arg197 (Figure 4a) as prominent residues. The second possible binding site is outside the protein, near where drug–C_{60} binds; common residues are serine Ser316 and valine Val313 (Figure 4b). The third binding site is also at the outside of the protein and characterized by isoleucine Ile166 and arginine Arg162 (Figure 4c).

Table 9 shows the quantitative descriptors used and the importance that each one has in the predictive models for drug–C_{60}–COOH. The docking score was the predicted variable in all the cases. In addition, the variables used to obtain the models in the case of drug–C_{60} were initially considered. The variables to model the interaction with C_{60}–COOH were taken directly from those of drug–C_{60}. For example, AI model 1 with C_{60}–COOH used the variables that had an importance greater than zero from AI model 1 of drug–C_{60}. The most important variables in the six models were LogP, RBC, and Dn. The least important variable in the six models was WS because had less than 10% importance in all models.

Supporting Information File 1, Table S6 shows the docking scores obtained for the validation set using both methods, that

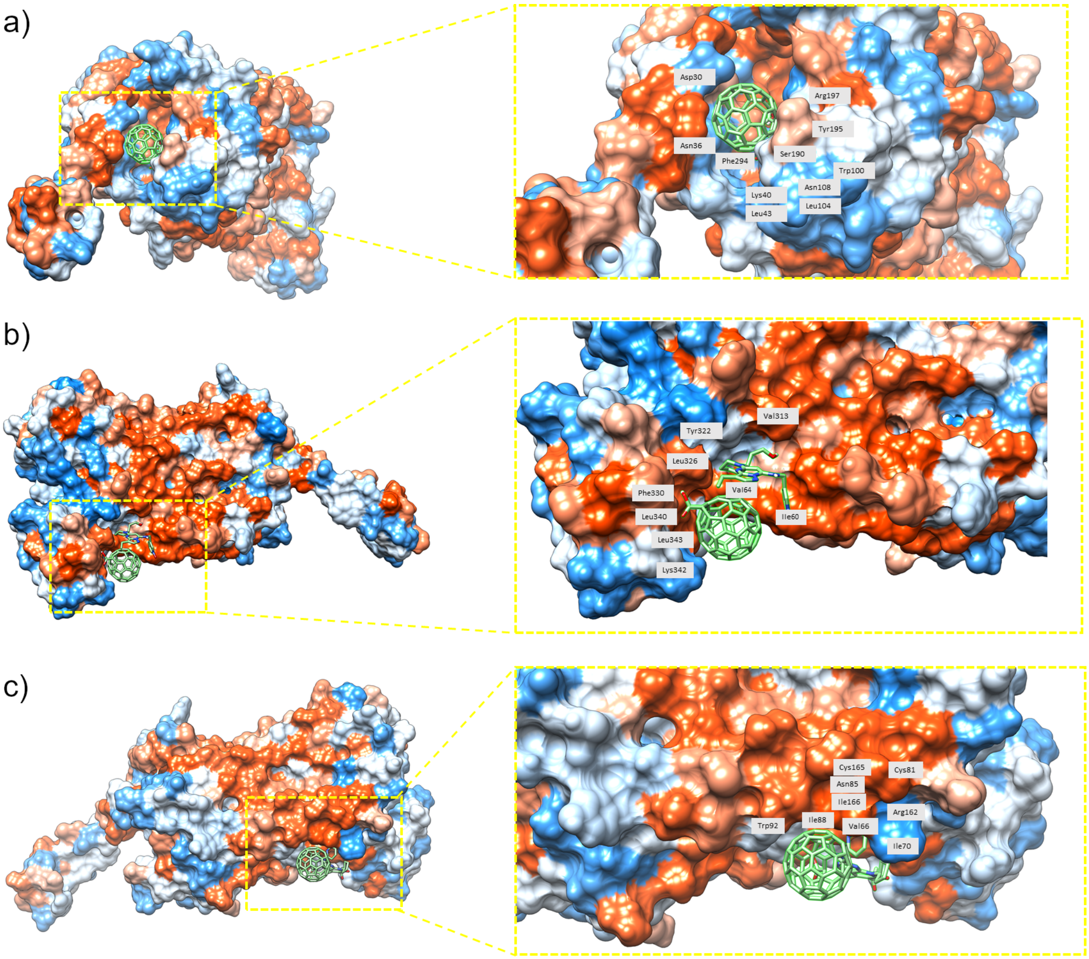

Figure 4: Binding sites for (a) gemcitabine, (b) dinaciclib, and (c) palbociclib with C_{60}–COOH and protein CXCR7. Blue color indicates the most hydrophilic sites and orange-red the most hydrophobic ones. Interacting residues are labeled as well.

Table 9: Input variables (IV) and output importance (OI) obtained for six Extra Tree regressor models obtained from IBM Watson. Variables are annotated according to the notation introduced in Table 1. The best models, according to their MAPE values, are highlighted in bold.

Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)	IV	OI (%)
RBC	**100**	pK_a	100	PSA	100	**Ω**	**100**	LogP	100	MW	100
LogP	**34**	LogP	96	LogP	77	**pK_a**	**93**	LogS	90	NOR	96
Dn	**20**	Dn	69	pK_a	72	**PSA**	**70**	α	83	E_{LUMO}	38
WS	**10**	LogS	64	A	42	**E_{LUMO}**	**31**	NOR	61	ω	0
LogS	**6**	E_{HOMO}	30	E_{HOMO}	32	**NOR**	**5**	Dn	52	—	—
NOR	**5**	PSA	12	E_{LUMO}	26	**E_{HOMO}**	**2**	E_{LUMO}	0	—	—
ω	**2**	E_{LUMO}	0	WS	3	**LogS**	**1**	—	—	—	—
MW	**1**	—	—	ω	0	**WS**	**0**	—	—	—	—
E_{HOMO}	**0**	—	—	—	—	—	—	—	—	—	—
PSA	**0**	—	—	—	—	—	—	—	—	—	—
α	**0**	—	—	—	—	—	—	—	—	—	—

is, the Extra Tree regressors implemented in IBM Watson and the MLR. With these values, it is possible to compare the difference in the docking score obtained by molecular docking and the prediction of the mathematical models for drug–C_{60}–COOH and the protein CXCR7. Table 10 shows the values obtained for the different types of errors in each of the models. The MSE ranged from 0.77 to 4.73 kcal^2/mol^2, whereas the MAPE varied from 6.70 to 16.22%. Also, the MAE was obtained ranging from 0.69 to 6.52 kcal/mol. Finally, the RMSE varied from 0.88 to 2.18 kcal/mol. Considering the MAPE, the best model, with a value of 6.7%, to predict the docking score is MLR model 1. Once again, the synergistic effect of using AI with a mathematical tool such as MLR is observed. The benefits are the predictive model's clarity, whereas AI was useful in determining the most important descriptors to be included in the QSAR/QSPR model. The explicit form of this model is:

$$\begin{aligned}\mathrm{MLR1}_{\mathrm{DRUG+C60\text{-}COOH}} &= -13.597 + 0.047 \times \mathrm{p}K_\mathrm{a} + 0.33 \times \mathrm{LogP} \\ &+ 0.18 \times \mathrm{Dn} + 0.284 \times \mathrm{LogS} \\ &- 0.696 \times E_{\mathrm{HOMO}} - 0.007 \times \mathrm{PSA}.\end{aligned} \tag{7}$$

The other significant model is MLR model 5 with a MAPE of 10.17%. In this model, experimental and theoretical descriptors were mixed. The following is the explicit form of MLR model 5:

$$\begin{aligned}\mathrm{MLR5}_{\mathrm{DRUG+C60\text{-}COOH}} &= 2.480 + 0.552 \times \mathrm{LogP} \\ &+ 0.21 \times \mathrm{LogS} + 0.002 \times \alpha \\ &- 0.402 \times \mathrm{NOR} + 0.186 \times \mathrm{Dn}.\end{aligned} \tag{8}$$

Considering the MAPE, the best AI model is model 4 with a value of 8.18% and the functional form $\mathrm{AI4}_{\mathrm{DRUG+C60\text{-}COOH}} = f$ (E_{HOMO}, ω, pK_a, PSA, E_{LUMO}, LogS, NOR). The other significant model obtained from AI is model 5, with a MAPE value of 8.69%; the functional form of this model is $\mathrm{AI5}_{\mathrm{DRUG+C60\text{-}COOH}} = f$ (LogP, α, LogS, NOR, Dn). Thus, although the Extra Trees algorithm was competitive in the case of drugs modified with a carboxyfullerene, this approach was surpassed by the MLR with the AI choosing the most important variables.

Doxorubicin and gemcitabine with a water-soluble fullerene

Finally, doxorubicin and gemcitabine were selected to compare the DFTB3 approach with the regular DFT method. In addition, their interactions with a water-soluble fullerene derivative were studied as well. Both anticancer agents are presented in Figure 5 interacting with a water-soluble fullerene [36,37]. Doxorubicin, an antibiotic that belongs to the family of tetracycline pharmaceutical agents, has gained popularity among chemotherapy agents and was recently modified with fullerene C_{60} [72-75]. Its anticarcinogenic activity comes from its ability to intercalate into DNA, inducing damage of the DNA strands and inhibiting its replication. Also, doxorubicin contributes to stopping the action of the enzyme topoisomerase II, leading to apoptosis of living tissues [71]; therefore, it is important to study its intrinsic chemical reactivity.

At the B3PW91/6-31G level of DFT, it is possible to appreciate that the periphery of the doxorubicin molecule is saturated by organic substituents. Its oxy, carbonyl, and carboxy terminal groups are active sites to interact with DNA or amino acids. The molecular orbital scheme of this molecule is shown in Supporting Information File 1, Figure S2, together with its molecular electrostatic potential (ESP). In the case of doxorubicin, from frontier molecular orbitals theory, HOMO and LUMO were found on the tetracycline moiety. However, the HOMO is confined to the quinoid ring, whereas the LUMO is completely delocalized (Supporting Information File 1, Figure S2). Oxygen atoms are the regions with the most negative ESP (red color in Supporting Information File 1, Figure S2). In contrast, a high electrostatic potential (blue color in Supporting Information File 1, Figure S2) is found on the lateral substituted cycle. Both central rings, the aromatic one and the quinoid one, are the main regions for the reactivity, including all the substituent oxygen atoms. The frontier molecular orbitals are similar, and it is ex-

Table 10: Metrics obtained by the use of AI and MLR in the case of drug–C_{60}–COOH. The best models, according to their MAPE values, are highlighted in bold. Values were computed relative to the validation set.

Error	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR	AI	MLR
MSE (kcal^2/mol^2)	0.97	**0.77**	1.67	1.77	1.80	4.73	**1.05**	1.53	1.41	2.08	1.46	1.80
MAPE (%)	8.75	**6.70**	10.51	12.03	10.71	16.22	**8.18**	10.26	8.69	10.17	10.86	11.39
MAE (kcal/mol)	0.93	**0.69**	1.13	1.29	1.17	6.52	**0.86**	1.10	0.93	0.99	1.16	1.18
RMSE (kcal/mol)	0.98	**0.88**	1.29	1.33	1.34	2.18	**1.02**	1.24	1.19	1.44	1.21	1.34

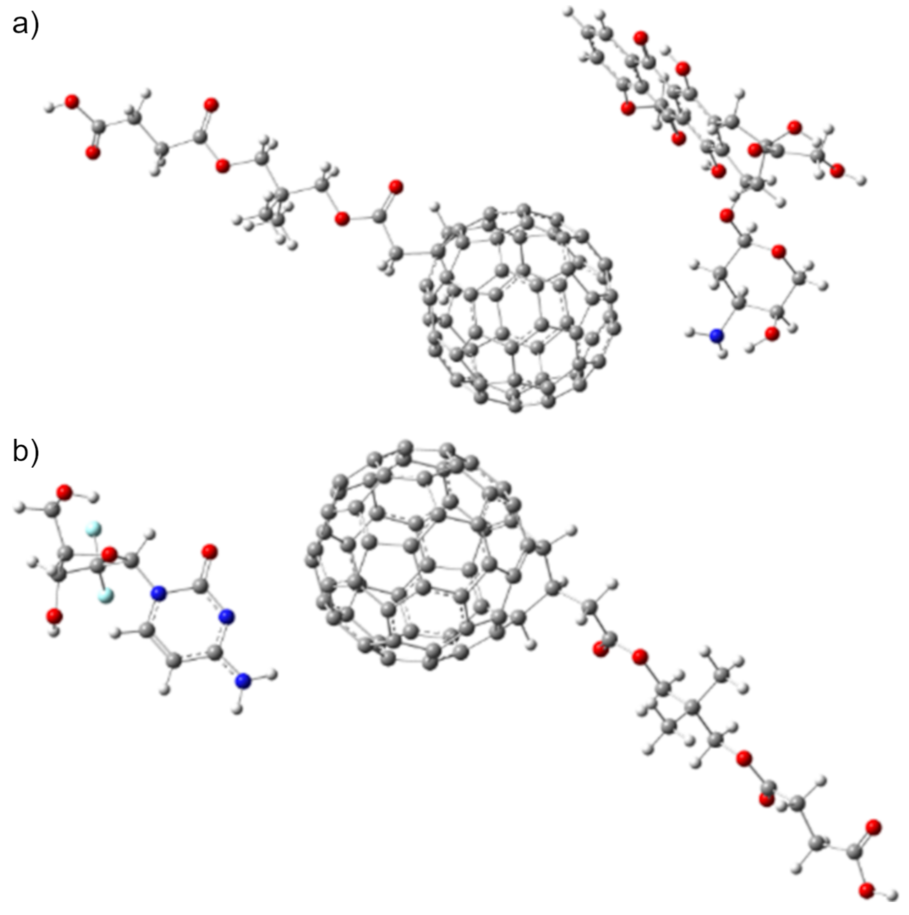

Figure 5: Ground state structures of (a) doxorubicin and (b) gemcitabine interacting with a water-soluble fullerene.

pected that electronic transit can occur in this region accepting and donating negative charges. The ESP map reinforces this suggestion showing negative density sites as well as a positive center, which can receive electrons. The energy of the HOMO was computed as −5.978 eV and that of the LUMO as −4.221 eV at the DFTB3 level. In comparison, the B3PW91 method yielded −6.116 and −3.242 eV, respectively. Thus, to consider the models obtained here, it is recommended to use DFTB3 to compute the electronic and energetic properties instead of DFT calculations.

Gemcitabine includes an active pyrimidinone fragment as a very reactive zone. The frontier molecular orbitals and the ESP map are shown Supporting Information File 1, Figure S2. Again, there is a strong polarization, which can induce a route for reaction. The ring nitrogen atom in alpha position concerning the carbonyl group is the more nucleophilic center, whereas there are two positive-density regions near the carbonyl group and in the C–C bond next to the amine-substituted carbon atom.

Both pharmaceutical agents are susceptible to interaction with fullerenes to form a force dispersion complex as it has been previously suggested. However, these complexes should be water-soluble to be delivered to their host. Considering all these factors, a water-soluble species was used to form such complexes, the structure of which [76] is shown in Figure 5. In both cases, a strong hydrogen bond is present; the distances are 1.97 Å for the doxorubicin complex and 2.25 Å for the gemcitabine complex. Furthermore, the energy of these interactions was calculated taking advantage of the Grimme module; they are 23.7 kcal/mol for doxorubicin and 18.9 kcal/mol for gemcitabine.

Conclusion

A QSAR/QSPR study of drugs commonly used for breast cancer chemotherapy modified with fullerene derivatives as drug nanocarriers was carried out. The CXCR7 protein was selected as a target for molecular docking calculations; the drugs were studied in the isolated form and modified with C_{60}

fullerene and with the water-soluble C_{60}–COOH fullerene derivative. An initial dataset was built by analyzing more than 30 drugs. The models to predict the docking score were obtained concerning Pearson's HSAB concept and common QSAR/QSPR descriptors. The energetic descriptors were computed quantum chemically by using density functional-based tight binding at the DFTB3 level. The highest docking score in the case of isolated drugs was −10.1 kcal/mol for olaparib. In contrast, in the case of the drugs modified with pristine C_{60} fullerene, it was −13.6 kcal/mol for exemestane. In the case of the drugs modified with the water-soluble fullerene derivative C_{60}–COOH, the maximum docking score was −11.7 kcal/mol for ixabepilone. Hence, the complexes are supposed to dock with stronger interactions with the CXCR7 protein than the isolated drugs. Also, characteristic binding sites were determined. The pocket of the isolated drugs was found within the protein, sharing residues including Trp100, Leu297, and Ser198. In the case of the drugs with fullerene C_{60}, the binding site was outside the protein with the complex pointing away from the pocket. The interacting residues included Arg323, Ser316, and Lys342. In the case of the drugs with C_{60}–COOH fullerene, there were three possible binding sites. The first two are the same as those in the previous cases The third binding site was found outside the protein, near the residues Ile166, Arg192, and Cys165. The docking score for the drug–fullerene complex is higher than that of the isolated drugs. QSAR/QSPR predictive models for the docking score were obtained from MLR and from IBM Watson artificial intelligence, yielding models with a MAPE of lower than 12% in all three cases. Although MLR exhibits the best evaluation metrics in the case of drug–C_{60} and drug–C_{60}–COOH complexes, an improvement is obtained based on the variables detected by the AI as the most important ones.

Supporting Information

Supporting Information File 1

Additional tables and figures.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-95-S1.pdf]

Acknowledgements

The authors thankfully acknowledge the computer resources, technical expertise, and support provided by Laboratorio de Supercómputo del Bajío, a member of the network of national laboratories. Also, the authors thank the Hasso Plattner Institute. Dedicated to the strongest and bravest woman ever, a survivor of breast cancer, my mom (A.M.).

Funding

A. Miralrio thanks the Challenge Based Research Funding program of Tecnológico de Monterrey.

Author Contributions

Jonathan-Siu-Loong Robles-Hernández: data curation; formal analysis; investigation; software; validation; visualization; writing – original draft. Dora Iliana Medina: formal analysis; funding acquisition; project administration; supervision; writing – review & editing. Katerin Aguirre-Hurtado: data curation; investigation; software; writing – original draft. Marlene Bosquez: data curation; formal analysis; investigation; software; writing – original draft. Roberto Salcedo: conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; supervision; writing – review & editing. Alan Miralrio: conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; writing – review & editing.

ORCID® iDs

Jonathan-Siu-Loong Robles-Hernández -
https://orcid.org/0009-0004-8982-0650
Dora Iliana Medina - https://orcid.org/0000-0001-5325-0079
Katerin Aguirre-Hurtado - https://orcid.org/0009-0000-5084-8008
Marlene Bosquez - https://orcid.org/0009-0008-1268-1153
Roberto Salcedo - https://orcid.org/0000-0003-1909-7757
Alan Miralrio - https://orcid.org/0000-0002-7992-3718

Data Availability Statement

The data that supports the findings of this study is available from the corresponding author upon reasonable request.

References

1. Giaquinto, A. N.; Miller, K. D.; Tossas, K. Y.; Winn, R. A.; Jemal, A.; Siegel, R. L. *Ca-Cancer J. Clin.* **2022,** *72,* 202–229. doi:10.3322/caac.21718
2. Miller, K. D.; Ortiz, A. P.; Pinheiro, P. S.; Bandi, P.; Minihan, A.; Fuchs, H. E.; Martinez Tyson, D.; Tortolero-Luna, G.; Fedewa, S. A.; Jemal, A. M.; Siegel, R. L. *Ca-Cancer J. Clin.* **2021,** *71,* 466–487. doi:10.3322/caac.21695
3. Wendt, C.; Margolin, S. *Acta Oncol.* **2019,** *58,* 135–146. doi:10.1080/0284186x.2018.1529428
4. Faruk, T.; Islam, M. K.; Arefin, S.; Haq, M. Z. *Clin. Breast Cancer* **2015,** *15,* 313–324. doi:10.1016/j.clbc.2015.01.002
5. Giaquinto, A. N.; Sung, H.; Miller, K. D.; Kramer, J. L.; Newman, L. A.; Minihan, A.; Jemal, A.; Siegel, R. L. *Ca-Cancer J. Clin.* **2022,** *72,* 524–541. doi:10.3322/caac.21754
6. Siegel, R. L.; Miller, K. D.; Wagle, N. S.; Jemal, A. *Ca-Cancer J. Clin.* **2023,** *73,* 17–48. doi:10.3322/caac.21763
7. Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. *Ca-Cancer J. Clin.* **2021,** *71,* 209–249. doi:10.3322/caac.21660

8. Arnold, M.; Morgan, E.; Rumgay, H.; Mafra, A.; Singh, D.; Laversanne, M.; Vignat, J.; Gralow, J. R.; Cardoso, F.; Siesling, S.; Soerjomataram, I. *Breast* **2022,** *66,* 15–23. doi:10.1016/j.breast.2022.08.010
9. Anderson, D. *Mutat. Res., Fundam. Mol. Mech. Mutagen.* **1995,** *329,* 37–47. doi:10.1016/0027-5107(95)00017-d
10. Hay, J.; Shahzeidi, S.; Laurent, G. *Arch. Toxicol.* **1991,** *65,* 81–94. doi:10.1007/bf02034932
11. Meyer, K. B.; Madias, N. E. *Miner Electrolyte Metab.* **1994,** *20,* 201–213.
12. Bentzen, S. M. *Nat. Rev. Cancer* **2006,** *6,* 702–713. doi:10.1038/nrc1950
13. Delanian, S.; Porcher, R.; Rudant, J.; Lefaix, J.-L. *J. Clin. Oncol.* **2005,** *23,* 8570–8579. doi:10.1200/jco.2005.02.4729
14. Dunnwald, L. K.; Rossing, M. A.; Li, C. I. *Breast Cancer Res.* **2007,** *9,* R6. doi:10.1186/bcr1639
15. Asif, H. M.; Sultana, S.; Ahmed, S.; Akhtar, N.; Tariq, M. *Asian Pac. J. Cancer Prev.* **2016,** *17,* 1609–1615. doi:10.7314/apjcp.2016.17.4.1609
16. Shi, Y.; Riese, D. J., II; Shen, J. *Front. Pharmacol.* **2020,** *11,* 574667. doi:10.3389/fphar.2020.574667
17. Sherif, M. F.; Ismail, I. M.; Ata, S. M. S. *Ann. Diagn. Pathol.* **2020,** *49,* 151621. doi:10.1016/j.anndiagpath.2020.151621
18. Xu, L.; Li, C.; Hua, F.; Liu, X. *Med. Oncol.* **2021,** *38,* 58. doi:10.1007/s12032-021-01481-2
19. Ding, C.; Li, Z. *Mater. Sci. Eng., C* **2017,** *76,* 1440–1453. doi:10.1016/j.msec.2017.03.130
20. Hossen, S.; Hossain, M. K.; Basher, M. K.; Mia, M. N. H.; Rahman, M. T.; Uddin, M. J. *J. Adv. Res.* **2019,** *15,* 1–18. doi:10.1016/j.jare.2018.06.005
21. Krätschmer, W.; Lamb, L. D.; Fostiropoulos, K.; Huffman, D. R. *Nature* **1990,** *347,* 354–358. doi:10.1038/347354a0
22. Krusic, P. J.; Wasserman, E.; Keizer, P. N.; Morton, J. R.; Preston, K. F. *Science* **1991,** *254,* 1183–1185. doi:10.1126/science.254.5035.1183
23. Markovic, Z.; Trajkovic, V. *Biomaterials* **2008,** *29,* 3561–3573. doi:10.1016/j.biomaterials.2008.05.005
24. Castro, E.; Garcia, A. H.; Zavala, G.; Echegoyen, L. *J. Mater. Chem. B* **2017,** *5,* 6523–6535. doi:10.1039/c7tb00855d
25. Lichota, A.; Krokosz, A. *Med. Pr.* **2016,** *67,* 817–831. doi:10.13075/mp.5893.00466
26. Krishna, V.; Singh, A.; Sharma, P.; Iwakuma, N.; Wang, Q.; Zhang, Q.; Knapik, J.; Jiang, H.; Grobmyer, S. R.; Koopman, B.; Moudgil, B. *Small* **2010,** *6,* 2236–2241. doi:10.1002/smll.201000847
27. Kumar, M.; Raza, K. *Pharm. Nanotechnol.* **2018,** 5. doi:10.2174/2211738505666170301142232
28. Panova, G. G.; Zhuravleva, A. S.; Khomyakov, Y. V.; Vertebnyi, V. E.; Ageev, S. V.; Petrov, A. V.; Podolsky, N. E.; Morozova, E. I.; Sharoyko, V. V.; Semenov, K. N. *J. Mol. Struct.* **2021,** *1235,* 130163. doi:10.1016/j.molstruc.2021.130163
29. Ye, S.; Zhou, T.; Cheng, K.; Chen, M.; Wang, Y.; Jiang, Y.; Yang, P. *Nanoscale Res. Lett.* **2015,** *10,* 246. doi:10.1186/s11671-015-0953-9
30. Danishuddin; Khan, A. U. *Drug Discovery Today* **2016,** *21,* 1291–1302. doi:10.1016/j.drudis.2016.06.013
31. Chen, X.; Li, H.; Tian, L.; Li, Q.; Luo, J.; Zhang, Y. *J. Comput. Biol.* **2020,** *27,* 1397–1406. doi:10.1089/cmb.2019.0323
32. Pearson, R. G. *J. Am. Chem. Soc.* **1963,** *85,* 3533–3539. doi:10.1021/ja00905a001
33. Chattaraj, P. K.; Giri, S. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2009,** *105,* 13. doi:10.1039/b802832j
34. Bağlayan, Ö.; Parlak, C.; Dikmen, G.; Alver, Ö. *Comput. Theor. Chem.* **2023,** *1221,* 114036. doi:10.1016/j.comptc.2023.114036
35. Parlak, C.; Alver, Ö. *J. Indian Chem. Soc.* **2022,** *99,* 100769. doi:10.1016/j.jics.2022.100769
36. Pérez-Manríquez, L.; Ramos, E.; Rangel, E.; Salcedo, R. *Mol. Simul.* **2013,** *39,* 612–620. doi:10.1080/08927022.2012.758845
37. Bosquez, M.; Fomina, L.; Salcedo, R. *Mol. Simul.* **2020,** *46,* 398–407. doi:10.1080/08927022.2020.1714042
38. Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. *Chem. Soc. Rev.* **2020,** *49,* 3525–3564. doi:10.1039/d0cs00098a
39. Selassie, C.; Verma, R. P. History of Quantitative Structure–Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery;* Abraham, D. J., Ed.; Wiley, 2010; pp 1–96. doi:10.1002/0471266949.bmc001.pub2
40. Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshaye, M. Y.; Dumitrică, T.; Dominguez, A.; Ehlert, S.; Elstner, M.; Van Der Heide, T.; Hermann, J.; Irle, S.; Kranz, J. J.; Köhler, C.; Kowalczyk, T.; Kubař, T.; Lee, I. S.; Lutsker, V.; Maurer, R. J.; Min, S. K.; Mitchell, I.; Negre, C.; Niehaus, T. A.; Niklasson, A. M. N.; Page, A. J.; Pecchia, A.; Penazzi, G.; Persson, M. P.; Řezáč, J.; Sánchez, C. G.; Sternberg, M.; Stöhr, M.; Stuckenberg, F.; Tkatchenko, A.; Yu, V. W.-z.; Frauenheim, T. *J. Chem. Phys.* **2020,** *152,* 124101. doi:10.1063/1.5143190
41. Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013,** *9,* 338–354. doi:10.1021/ct300849w
42. Vainio, M. J.; Johnson, M. S. *J. Chem. Inf. Model.* **2007,** *47,* 2462–2474. doi:10.1021/ci6005646
43. Wang, C.; Chen, W.; Shen, J. *Front. Pharmacol.* **2018,** *9,* 10.3389/fphar.2018.00641. doi:10.3389/fphar.2018.00641
44. Yang, J.; Zhang, Y. *Nucleic Acids Res.* **2015,** *43,* W174–W181. doi:10.1093/nar/gkv342
45. Muthiah, I.; Rajendran, K.; Dhanaraj, P.; Vallinayagam, S. *J. Biomol. Struct. Dyn.* **2021,** *39,* 4807–4815. doi:10.1080/07391102.2020.1783365
46. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993,** *26,* 283–291. doi:10.1107/s0021889892009944
47. Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. *J. Chem. Inf. Model.* **2021,** *61,* 3891–3898. doi:10.1021/acs.jcim.1c00203
48. Ganapathiraju, M.; Balakrishnan, N.; Reddy, R.; Klein-Seetharaman, J. *BMC Bioinf.* **2008,** *9,* S4. doi:10.1186/1471-2105-9-s1-s4
49. Tusnády, G. E.; Simon, I. *Bioinformatics* **2001,** *17,* 849–850. doi:10.1093/bioinformatics/17.9.849
50. Chen, Y.; Elenee Argentinis, J.; Weber, G. *Clin. Ther.* **2016,** *38,* 688–701. doi:10.1016/j.clinthera.2015.12.001
51. Geurts, P.; Ernst, D.; Wehenkel, L. *Mach. Learn.* **2006,** *63,* 3–42. doi:10.1007/s10994-006-6226-1
52. Suzuki, J. *Statistical Learning with Math and Python: 100 Exercises for Building Logic;* Springer Singapore: Singapore, 2021. doi:10.1007/978-981-15-7877-9
53. Costa, A.; Nannicini, G. *Math. Program. Comput.* **2018,** *10,* 597–629. doi:10.1007/s12532-018-0144-7

54. Methods in molecular biology. In *Topics in Biostatistics;* Ambrosius, W. T., Ed.; Humana Press: Totowa, N.J., 2007. doi:10.1007/978-1-59745-530-5
55. Yang, L.; Sang, C.; Wang, Y.; Liu, W.; Hao, W.; Chang, J.; Li, J. *Chemosphere* **2021,** *285,* 131456. doi:10.1016/j.chemosphere.2021.131456
56. Keshavarz, M. H.; Shirazi, Z.; Kiani Sheikhabadi, P. *Process Saf. Environ. Prot.* **2021,** *150,* 137–147. doi:10.1016/j.psep.2021.04.011
57. Lomba, L.; Ribate, M. P.; Zuriaga, E.; García, C. B.; Giner, B. *Ecotoxicol. Environ. Saf.* **2019,** *172,* 232–239. doi:10.1016/j.ecoenv.2019.01.081
58. Sadeghi, F.; Afkhami, A.; Madrakian, T.; Ghavami, R. *J. Iran. Chem. Soc.* **2021,** *18,* 2785–2800. doi:10.1007/s13738-021-02233-9
59. Zuriaga, E.; Giner, B.; Valero, M. S.; Gómez, M.; García, C. B.; Lomba, L. *Chemosphere* **2019,** *227,* 480–488. doi:10.1016/j.chemosphere.2019.04.054
60. Fadilah, F.; Arsianti, A.; Yanuar, A.; Andrajati, R.; Indah Paramita, R.; Hernawati Purwaningsih, E. *Orient. J. Chem.* **2018,** *34,* 2656–2660. doi:10.13005/ojc/340558
61. El Rhabori, S.; El Aissouq, A.; Chtita, S.; Khalil, F. *J. Indian Chem. Soc.* **2022,** *99,* 100675. doi:10.1016/j.jics.2022.100675
62. Er-Rajy, M.; ElFadili, M.; Mrabti, N. N.; Zarougui, S.; Elhallaoui, M. *Chin. J. Anal. Chem.* **2022,** *50,* 100163. doi:10.1016/j.cjac.2022.100163
63. Silva, A. M.; Martins-Gomes, C.; Silva, T. L.; Coutinho, T. E.; Souto, E. B.; Andreani, T. *Toxics* **2022,** *10,* 378. doi:10.3390/toxics10070378
64. Hansch, C.; Steinmetz, W. E.; Leo, A. J.; Mekapati, S. B.; Kurup, A.; Hoekman, D. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 120–125. doi:10.1021/ci020378b
65. Sudhakaran, S.; Calvin, J.; Amy, G. L. *Chemosphere* **2012,** *87,* 144–150. doi:10.1016/j.chemosphere.2011.12.006
66. Ogunyemi, B. T.; Latona, D. F.; Adejoro, I. A. *Sci. Afr.* **2020,** *8,* e00336. doi:10.1016/j.sciaf.2020.e00336
67. Bodun, D. S.; Omoboyowa, D. A.; Omotuyi, O. I.; Olugbogi, E. A.; Balogun, T. A.; Ezeh, C. J.; Omirin, E. S. *Comput. Biol. Chem.* **2023,** *104,* 107865. doi:10.1016/j.compbiolchem.2023.107865
68. Trott, O.; Olson, A. J. *J. Comput. Chem.* **2010,** *31,* 455–461. doi:10.1002/jcc.21334
69. Bos, J. D.; Meinardi, M. M. H. M. *Exp. Dermatol.* **2000,** *9,* 165–169. doi:10.1034/j.1600-0625.2000.009003165.x
70. Haddon, R. C. *Philos. Trans. R. Soc., A* **1993,** *343,* 53–62. doi:10.1098/rsta.1993.0040
71. Sergio, M.; Behzadi, H.; Otto, A.; van der Spoel, D. *Environ. Chem. Lett.* **2013,** *11,* 105–118. doi:10.1007/s10311-012-0387-x
72. Panchuk, R. R.; Prylutska, S. V.; Chumak, V. V.; Skorokhyd, N. R.; Lehka, L. V.; Evstigneev, M. P.; Prylutskyy, Y. I.; Berger, W.; Heffeter, P.; Scharff, P.; Ritter, U.; Stoika, R. S. *J. Biomed. Nanotechnol.* **2015,** *11,* 1139–1152. doi:10.1166/jbn.2015.2058
73. Butowska, K.; Kozak, W.; Zdrowowicz, M.; Makurat, S.; Rychłowski, M.; Hać, A.; Herman-Antosiewicz, A.; Piosik, J.; Rak, J. *Struct. Chem.* **2019,** *30,* 2327–2338. doi:10.1007/s11224-019-01428-4
74. Grebinyk, A.; Prylutska, S.; Grebinyk, S.; Prylutskyy, Y.; Ritter, U.; Matyshevska, O.; Dandekar, T.; Frohme, M. *Nanoscale Res. Lett.* **2019,** *14,* 61. doi:10.1186/s11671-019-2894-1
75. Liu, J.-H.; Cao, L.; Luo, P. G.; Yang, S.-T.; Lu, F.; Wang, H.; Meziani, M. J.; Haque, S. A.; Liu, Y.; Lacher, S.; Sun, Y.-P. *ACS Appl. Mater. Interfaces* **2010,** *2,* 1384–1389. doi:10.1021/am100037y
76. Martín, N.; Altable, M.; Filippone, S.; Martín-Domenech, A. *Synlett* **2007,** 3077–3095. doi:10.1055/s-2007-990939

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjnano.15.95