pharmaceutics

MDPI

Article

QSAR and Chemical Read-Across Analysis of 370 Potential MGMT Inactivators to Identify the Structural Features Influencing Inactivation Potency

Guohui Sun [1,*], Peiying Bai [1], Tengjiao Fan [1,2], Lijiao Zhao [1], Rugang Zhong [1], R. Stanley McElhinney [3,†], T. Brian H. McMurry [3], Dorothy J. Donnelly [3], Joan E. McCormick [3], Jane Kelly [4] and Geoffrey P. Margison [4,5,*]

1 Beijing Key Laboratory of Environmental and Viral Oncology, Faculty of Environment and Life, Beijing University of Technology, Beijing 100124, China; baipy@emails.bjut.edu.cn (P.B.); fannie818@126.com (T.F.); zhaolijiao@bjut.edu.cn (L.Z.); lifesci@bjut.edu.cn (R.Z.)
2 Department of Medical Technology, Beijing Pharmaceutical University of Staff and Workers, Beijing 100079, China
3 Chemistry Department, Trinity College, D02 PN40 Dublin, Ireland; tmcmurry@tcd.ie (T.B.H.M.); dor.donnelly@gmail.com (D.J.D.)
4 Carcinogenesis Department, Paterson Institute for Cancer Research, Manchester M20 9BX, UK; janekelly99@hotmail.com
5 Epidemiology and Public Health Group, School of Health Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PG, UK
* Correspondence: sunguohui@bjut.edu.cn (G.S.); gmargison@manchester.ac.uk (G.P.M.); Tel.: +86-10-67391917 (G.S.); +44-1625-875367 (G.P.M.)
† This publication is dedicated to the memory of R. Stanley MacElhinney; R. Stanley McElhinney and Joan E. McCormick don't have email addresses.

Abstract: O^6-methylguanine-DNA methyltransferase (MGMT) constitutes an important cellular mechanism for repairing potentially cytotoxic DNA damage induced by guanine O^6-alkylating agents and can render cells highly resistant to certain cancer chemotherapeutic drugs. A wide variety of potential MGMT inactivators have been designed and synthesized for the purpose of overcoming MGMT-mediated tumor resistance. We determined the inactivation potency of these compounds against human recombinant MGMT using [^{3}H]-methylated-DNA-based MGMT inactivation assays and calculated the IC_{50} values. Using the results of 370 compounds, we performed quantitative structure–activity relationship (QSAR) modeling to identify the correlation between the chemical structure and MGMT-inactivating ability. Modeling was based on subdividing the sorted pIC_{50} values or on chemical structures or was random. A total of nine molecular descriptors were presented in the model equation, in which the mechanistic interpretation indicated that the status of nitrogen atoms, aliphatic primary amino groups, the presence of O-S at topological distance 3, the presence of Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X, the ionization potential and hydrogen bond donors are the main factors responsible for inactivation ability. The final model was of high internal robustness, goodness of fit and prediction ability (R^2_{pr} = 0.7474, Q^2_{Fn} = 0.7375–0.7437, CCC_{pr} = 0.8530). After the best splitting model was decided, we established the full model based on the entire set of compounds using the same descriptor combination. We also used a similarity-based read-across technique to further improve the external predictive ability of the model (R^2_{pr} = 0.7528, Q^2_{Fn} = 0.7387–0.7449, CCC_{pr} = 0.8560). The prediction quality of 66 true external compounds was checked using the "Prediction Reliability Indicator" tool. In summary, we defined key structural features associated with MGMT inactivation, thus allowing for the design of MGMT inactivators that might improve clinical outcomes in cancer treatment.

Keywords: MGMT; pseudosubstrates; inactivating agents; methyltransferase assay; MGMT activity determination; QSAR; read-across

check for updates

Citation: Sun, G.; Bai, P.; Fan, T.; Zhao, L.; Zhong, R.; McElhinney, R.S.; McMurry, T.B.H.; Donnelly, D.J.; McCormick, J.E.; Kelly, J.; et al. QSAR and Chemical Read-Across Analysis of 370 Potential MGMT Inactivators to Identify the Structural Features Influencing Inactivation Potency. *Pharmaceutics* **2023**, *15*, 2170. https://doi.org/10.3390/pharmaceutics15082170

Academic Editor: David Barlow

Received: 7 August 2023
Revised: 16 August 2023
Accepted: 19 August 2023
Published: 21 August 2023

1. Introduction

The DNA repair protein, O^6-methylguanine-DNA methyltransferase (MGMT; also known as O^6-alkylguanine-DNA alkyltransferase; AGT), can protect cells against the cytotoxic effects induced by DNA alkylating agents, such as the methylating antitumor drugs temozolomide (TMZ), procarbazine (PCB) and dacarbazine (DITC) and the chloroethylating antitumor drugs 1,3-bis(2-chloroethyl)-1-nitrosourea (BCNU), 1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea (CCNU) and 1-(4-amino-2-methyl-5-pyrimidinyl)methyl-3-(2-chloroethyl)-3-nitrosourea (ACNU) [1–4]. These alkylating agents predominantly exert their antitumor activity through the alkylation of DNA at the O^6-position of guanine, generating a highly cytotoxic lesion [1,3,5]. However, O^6-alkylguanine adducts can be repaired by MGMT that transfers an adducted alkyl group to its active center Cys145 residue in an irreversible and stoichiometric reaction [6]. Thus, MGMT is a "suicide enzyme" that acts only once, and further repair activity can be restored only by de novo protein synthesis [7].

The consensus repair mechanism by MGMT is shown in Figure 1. The formation of the *S*-alkyl adduct, at least in the case of methyl, causes a conformational change in MGMT, resulting in an increased recognition by ubiquitin ligase, targeting it for proteasome degradation [8–10].

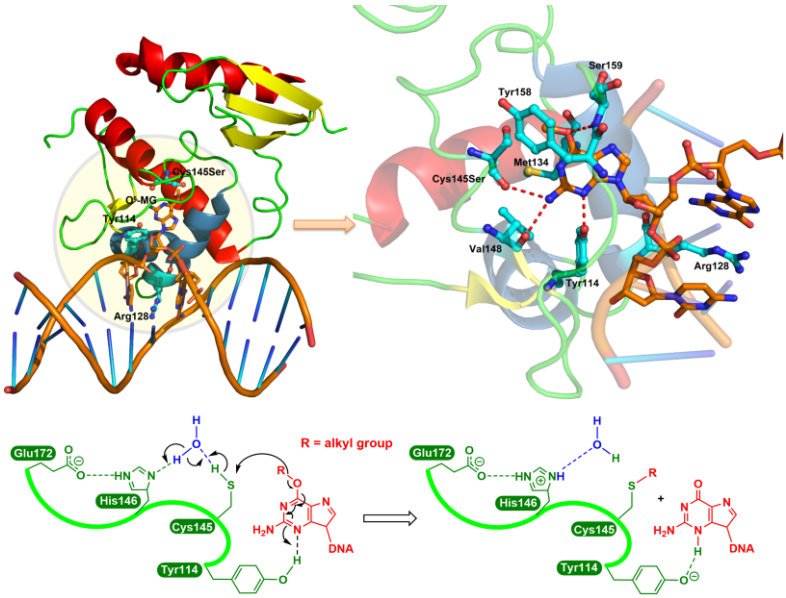

Figure 1. The consensus repair mechanism of O^6-alkylguanines by MGMT. In the crystal structure of the Cys145Ser MGMT mutant bound to O^6-MeG-containing DNA (upper left panel), the guanine moiety is flipped by Arg128 into the active site pocket and then forms hydrogen bonds (red dashed line; upper right panel) with Cys145Ser, Val148, Tyr114 and Ser159 residues. Lower panel: as a water-mediated general base, His146 deprotonates Cys145 (in the native protein), resulting in the transfer to the S atom of the O^6-alkyl carbon, while the N3 is protonated by the Tyr114 residue.

Given the role of MGMT in alkylating chemotherapeutic resistance and its ability to act on O^6-alkylguanines as free bases, various groups have synthesized a large number of such "pseudosubstrates" as potential inactivators of MGMT function [11–22]. Administering such agents prior to alkylating agents was proposed to ablate the protection provided by MGMT and hence increase the effectiveness of the chemotherapeutics [1,3,6].

Currently, only two MGMT inactivators, O^6-benzylguanine (O^6-BG) and O^6-(4-bromo thenyl)guanine (O^6-4-BTG; Lomeguatrib), have been used as potentiating agents in clinical trials [6,23–26]. Unfortunately, the combination greatly increased the systemic toxicity of the alkylating chemotherapeutic drugs, requiring a considerable reduction in their dose [1,6,23–25]. This dose reduction might explain, at least in part, why the MGMT inactivators did not improve the clinical outcome of chemotherapy. Other factors might

include: the rates of recovery of MGMT activity following depletion; tumor cell proliferation rates; the contribution of other protective mechanisms to cell survival; the relatively lower affinity of MGMT for free bases compared with O^6-alkylguanines in duplex DNA [9,27]; poor water solubility; low bioavailability, instability and/or catabolic processes and rapid plasma clearance [28].

Given that the systemic delivery of MGMT inactivators exacerbates collateral toxicities, the synthesis of tumor-targeting inactivating agents might be expected to circumvent this [1,3,21,29,30]. To build into an inactivating agent tumor-targeting moieties, or, indeed, any other structural moieties that may optimize in vivo effectiveness, it will be essential to know what structural features endow the greatest activity and which regions cannot be modified without a loss of function.

To achieve this, we performed quantitative structure–activity relationship (QSAR) modeling to establish the detailed relationship between the molecular structure and MGMT inactivation potency. Although we reported a QSAR model for MGMT inhibitors in a previous study, the focus was primarily on base analogs, and the dataset used was relatively smaller, consisting of 134 compounds [16]. In the current study, we used MGMT in vitro inactivation assay results for a total of 370 compounds, which provided IC_{50} values as the response endpoint, which included not only base analogs but also other types of molecules. Additionally, all the experimental values for the 370 compounds were directly determined in our laboratory, rather than relying on data from the literature. Furthermore, the chemical synthesis of the 370 compounds was conducted within our lab as well. Our model contributes to a definitive mechanistic interpretation but also provides a tool for predicting and rapidly designing new candidates for depleting MGMT activity, including the tumor-targeting MGMT inactivators.

2. Materials and Methods

2.1. Compound Design and Synthesis

The listed compounds were designed by R. Stanley McElhinney and T. Brian H. McMurry and synthesized by members of the Chemistry Department of Trinity College, Dublin. Typical examples of the methods for the synthesis and analysis of the compounds are presented elsewhere [18]. In addition to curiosity-driven compounds in pursuit of increasingly potent agents, others were designed with specific objectives in mind, among which there were: combination alkylators–inactivators; tumor receptor targeting agents or potential precursors and agents that would produce antimetabolites upon dealkylation (see Table S1).

2.2. MGMT Activity Assay

Compounds were assayed for their ability to inactivate human recombinant MGMT in vitro using [^{3}H]-methylated-DNA-based MGMT inactivation assays, as described in [18], at the Paterson Institute for Cancer Research, Manchester, U.K. The IC_{50} values were obtained for QSAR modeling. It should be noted that the in vitro assay does not differentiate between actual inactivation due to alkyl group transfer to Cys145 (see Figure 1) and competitive inhibition: alkyl group transfer to MGMT has been demonstrated for very few compounds, but we are not aware of any reports in which the mechanism of action has been proven to be competitive inhibition.

A flowchart of the methodology in the present study is shown in Figure 2.

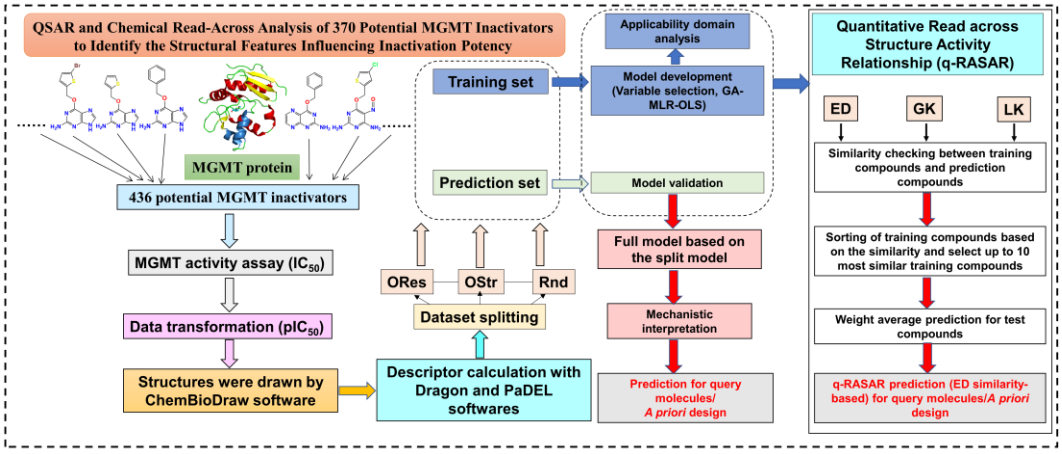

Figure 2. Workflow diagram for the methodology followed in the current study.

2.3. Dataset Preparation

The Organization for Economic Co-operation and Development (OECD) principle 1 states that a QSAR model should be associated with "a defined endpoint" [31–33], and in the present study, the IC_{50} value was used as the activity endpoint. All IC_{50} values (μM) were transformed into $-\log IC_{50}$ (pIC_{50}, mol/L), which is a common practice in QSAR modeling [32,34,35]; thus, the higher the pIC_{50} value, the more potent the MGMT inactivator.

2.4. Descriptor Calculation and Dataset Splitting

All the molecular structures were manually drawn using the ChemBioDraw Ultra 14.0 software (version 14.0, Cambridge soft, Cambridge, MA, USA) and geometrically optimized by energy minimization using its 3D module. After optimization, five quantum chemical descriptors including the dipole moment (μ), total energy (E), lowest unoccupied molecular orbital energy (E_{LUMO}), highest occupied molecular orbital energy (E_{HOMO}) and $E_{LUMO} - E_{HOMO}$ gap were calculated. Dragon software (version 7.0) [36] and PaDEL-Descriptor software (version 2.18) [37] were used to calculate the molecular descriptors. In order to avoid the occurrence of conformational complexity due to the inclusion of 3D descriptors, and for the ease of interpretability and reproductivity, only 2D descriptors with a definite physicochemical meaning were calculated. To remove redundant variables, we excluded the constant or near-constant descriptors (>80% compounds have the same value) and inter-correlated descriptors (>0.95) from the descriptor pool.

To avoid possible bias, the dataset was split into training sets and prediction sets in an approximately 3:1 ratio using QSARINS v2.2.4 software (Varese, Italy) [38,39], in which the training set was used to establish the model, while the prediction set was used for model validation. Three splitting techniques were used [38], and in each, the inactivators with the maximum and minimum response pIC_{50} values and principal component 1 (PC1) scores were always put into the training set to cover the range of the prediction set. Splitting was undertaken by the software and was based on (1) the sorted pIC_{50} values (ORes) or (2) the structure based on the PC1 score of descriptors (OStr) or (3) was random. For ORes splitting, compounds were sorted by their pIC_{50} values, and from the second molecules, every fourth compound was placed in the prediction set, and the remaining three of the four were put into the training set. For OStr splitting, compounds were sorted by their PC1 scores, and again, one of every four compounds was placed in the prediction set. The distribution of splitting (PC1 vs. PC2) was checked by the principal component analysis (PCA) using only descriptor variables (Figure S1).

The dataset splitting methods ensured that the selection procedure was unbiased. In order to develop a model with a wider applicability domain (AD), once the best variable combination was found by the splitting technique, the full model was obtained through

recalculation on the complete set (combining training and prediction sets), since all available experimental information was then considered [40,41].

2.5. Model Development and Validation

Variable selection from the large pool of descriptors is a very important step in the process of model development. Here, we used a Genetic Algorithm Variable Subset Selection (GA-VSS) tool of the QSARINS software [38] to conduct the variable selection. Initially, all the possible combinations of two descriptors were explored by all subset facilities to find the subset of descriptors encoding the response. Then, using the leave-one-out cross-validated correlation coefficient (Q^2_{LOO}) as a fitness function, GA-VSS was utilized to seek the new combinations with additional descriptors to yield the models. The generation per size, population size and mutation rate were given values of 2000, 200 and 20, respectively.

Depending on the empirical ratio [33,42], to reduce the possibility of chance correlation, the number of descriptors in the model should be less than one-fifth of the number of training compounds. QSAR models were established through Multiple Linear Regression (MLR) using the Ordinary Least Squares (OLS) approach implemented in the QSARINS software [38]. According to the OECD principle 2, a QSAR model should be associated with "an unambiguous algorithm" [42], which ensures the transparency of the model algorithm. It should be noted that the algorithmic information in commercial models is usually less publicly available.

Depending on the OECD principle 4, a QSAR model should be associated with "appropriate measures of goodness-of-fit, robustness and predictivity" [42]. The internal robustness and predictive ability of the model were assessed by the Q^2_{LOO}, Q^2_{LMO}, R^2 (including adjusted R^2_{adj}), root mean standard error ($RMSE_{tr}$) and mean absolute error (MAE_{tr}) [43,44]. In the leave more out (LMO) procedure, 30% of compounds were excluded from each calculation for 2000 iterations. A Y-randomization test (the dependent variable Y was randomly scrambled, while the independent variable matrix is unchanged) with 2000 iterations was also used for assessing the chance correlation between the model descriptors and response endpoint. In this test, the sequence of the response vector Y was randomly scrambled, while the descriptor variable X for each object was unchanged. In addition, we set the threshold of the QUIK (Q Under Influence of K) rule as 0.05 to exclude multi-co-linearity [42,45]. The external predictivity of the model was evaluated by the statistical parameters R^2_{pr}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, CCC_{pr}, $RMSE_{pr}$ and MAE_{pr} [44]. The detailed calculated formulae can be found elsewhere [44,46]: all the parameters are listed in Table S2 in the Supplementary Materials.

2.6. Best Model Selection by Multiple-Criteria Decision Making

On the basis of fitting and internal and external validation, the Multiple-Criteria Decision Making (MCDM) module implemented in QSARINS software [38] was utilized to rank the model performance as a score from 0 (the worst) to 1 (the best). The $MCDM_{fit}$ value was computed via the maximization of R^2, R^2_{adj} and CCC_{tr}, whereas the minimization of the $R^2 - R^2_{adj}$. $MCDM_{ext}$ value was computed via the maximization of Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC_{pr}. As a consequence, we selected the best QSAR model depending on both the $MCDM_{fit}$ and $MCDM_{ext}$ values. These models fulfill the OECD principles as well as various validation criteria [42]. It is accepted that the best model should be obtained with the lowest number of descriptors.

2.7. Applicability Domain (AD) Analysis

Depending on the OECD principle 3, a QSAR model should have "a defined domain of applicability" [42]. Only the compounds inside the AD of the model should provide reliable predictions. Here, we used both leverage and standardized residue approaches to define the AD [38,39]. Structural outliers were identified using the leverage approach. If a compound has a hat (h) value greater than the warning h^*, it will be identified as a

structural outlier. The warning h^* value was calculated by the formula of $3(p + 1)/n$, in which p is the number of variables in the model equation, and n is the number of training set compounds. If the standardized residual of a compound is more than three standard deviation units, it is identified as a response outlier.

We also prepared a true external set consisting of 66 compounds for checking the predictivity of the developed model. In order to visually show the prediction confidence for each molecule, an Insubria graph which plots the predicted values of the training/true external set against their hat values was generated [39]. The predictions for compounds with hat values greater than h^* should be considered to have low confidence.

2.8. Prediction Using a Similarity-Based Chemical Read-Across Technique

Read-Across (RA) is a completely similarity-based technique without the process of developing a statistical model, which is the most significant feature that is different from the classical QSAR methodology [47,48]. RA is widely used in qualitive predictions; however, the quantitative read-across technique was also reported in recent years. To further improve the external predictive ability, we used a novel approach called the quantitative read-across structure–activity relationship (q-RASAR) [49,50]. After completing the development of 2D-QSAR, the training set was divided into a subtraining set and subtest set, followed by the optimization of the hyperparameter using the Read-Across V4.1 tool (https://sites.google.com/jada vpuruniversity.in/dtc-lab-software/home) (accessed on 1 April 2023). The optimized hyperparameters were applied to the original dataset files as the input. In this study, the similarity determination between the training compounds and test compounds was determined based on the Euclidean distance, Laplacian kernel function and Gaussian kernel function. Then, we calculated the RASAR descriptors based on the selected descriptors in the 2D-QSAR model by the RASAR-Desc-Calc-v2.0 software (https://sites.google.com/jada vpuruniversity.in/dtc-lab-software/home) (accessed on 1 April 2023) using the optimized hyperparameters. The RASAR descriptors were combined with original 2D descriptors to develop the q-RASAR model using the same setting as 2D-QSAR. Finally, we obtained a q-RASAR model and q-RASAR-full model; the latter was also applied to the predictions of true external compounds.

3. Results and Discussion

3.1. MGMT Inactivation

The MGMT inactivation assay results, along with the compound name, number and structure, are listed in Table S1 in the Supplementary Materials.

3.2. Chemical Space Distribution

After processing the original data, we obtained 458 entries for MGMT inactivation (Table S1). For compounds also produced as salts, only the entry corresponding to the free compound was used. For hydrates, we deleted the water molecules in the descriptor calculations and model development.

The initial QSAR model development showed that 17 compounds were always response outliers that substantially influenced the linear fitting in different dataset splitting schemes. These compounds may be related to the activity cliffs [51,52]. It is suspected that the experimental IC_{50} values for these compounds may be erroneous, and they were therefore excluded. In addition, there were 49 compounds that were found to not inactivate MGMT at the highest concentration used in the assay and thus had no definitive IC_{50} values. Therefore, a total of 66 compounds were excluded from the training and prediction sets and selected as the true external set. Hence, in our modeling study (Table S1), the numbers for the training set, prediction set and true external set were 279, 91 and 66, respectively, in the best QSAR model (Tables S3 and S4).

Chemical space similarity is very important for evaluating the predictive performance of a model. Here, we used two commonly used physicochemical parameters: molecu-

lar weight (MW) and Ghose−Crippen $LogK_{ow}$ (ALogP), to explore the chemical space distribution [53–56] and plotted these as a scatter diagram (Figure 3).

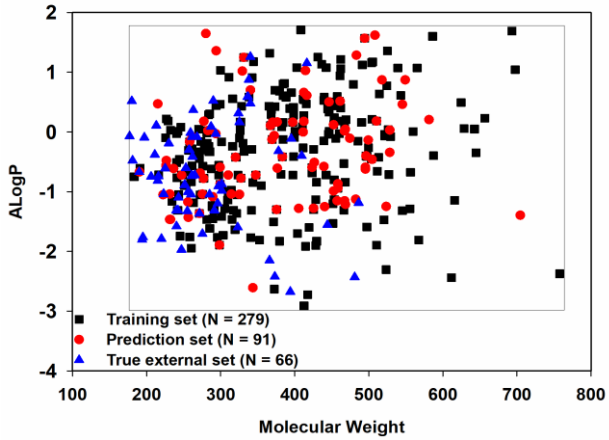

Figure 3. Chemical space distribution of the three datasets.

Given that the training set, prediction set and true external set, as expected, shared a similar chemical space, the models derived from the training set should have a broad applicability domain (AD) and thus a good degree of generalization.

3.3. QSAR Modeling of Potential MGMT Inactivators

3.3.1. Model Selection and Evaluation

According to the criteria recognized by Golbraikh and Tropsha [43], if a QSAR model meets the following thresholds for different statistical parameters: $Q^2_{LOO} > 0.5$, R^2 and $R^2_{pr} > 0.6$; $0.85 \leq k$ or $k' \leq 1.15$; $|R^2_0 - R'^2_0| < 0.3$, it should be considered an acceptable model. R^2_0 and R'^2_0 represent the correlation coefficients of regression of the predicted versus experimental values and experimental versus predicted values through the origin, respectively. K and k' represent the slopes of the corresponding regression lines for R^2_0 and R'^2_0, respectively.

Of the three splitting methods, that based on the ORes model had low values of Q^2_{LOO} and R^2 and did not meet the basic standard for an acceptable model [43]. This may be due to the model (Equation (1), Table 1) containing only three descriptors, and this cannot adequately simulate the biochemical response endpoint (pIC_{50}).

Although the OStr model (Equation (2), Table 1) is internally robust and stable (Q^2_{LOO} = 0.6496, R^2 = 0.6826), its external predictive performance is compromised ($R^2_{pr} < 0.6$) according to statistical criteria [43]. On the other hand, the OStr model included 13 molecular descriptors, which complicates the interpretation of the model.

It is remarkable that only the model derived from the Random splitting method (Rnd model or 2D-QSAR) fulfilled the Golbraikh and Tropsha criteria [43]. Furthermore, the 2D-QSAR model (Equation (3), Table 1) had the best predictivity for the prediction set (R^2_{pr} = 0.7474, Q^2_{Fn} = 0.7375~0.7437, CCC_{pr} = 0.8530), which met even the higher statistical standard proposed by Chirico and Gramatica [44], in which the thresholds of Q^2_{Fn}, R^2_{pr} and CCC_{pr} are 0.7, 0.7 and 0.85, respectively. Low values of Q^2_{Yscr} (−0.0424) and R^2_{Yscr} (0.0324) indicated that the model was not generated by chance correlation (Figure 4A).

Table 1. Statistical parameters for the internal and external validation of the developed QSAR models [+].

Division			**Fitting**	**Robustness**		**Chance Correlation**		**External Validation**					**Accuracy**			
Scheme	N_{tr}	N_{pr}	R^2	Q^2_{LOO}	Q^2_{LMO}	Q^2_{Yscr}	R^2_{Yscr}	R^2_{pr}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC_{pr}	$RMSE_{tr}$	$RMSE_{pr}$	MAE_{tr}	MAE_{pr}
ORes	278	92	0.5098	0.4968	0.4952	−0.0186	0.0108	0.5319	0.5271	0.5266	0.5658	0.6891	0.8669	0.8151	0.6666	0.6440
pIC_{50} = 3.9243 + 0.6729F09[O-S] + 0.121SaaN + 1.005MDEN-12 (k = 1.0, $\lvert R^2_0 - R'^2_0 \rvert$ = 0.4582) (1)																
OStr	278	92	0.6826	0.6496	0.6451	−0.0544	0.0432	0.5882	0.5721	0.5712	0.5814	0.7617	0.6917	0.7943	0.5514	0.5982
pIC_{50} = 7.1639 − 47.2151VE2sign_B(m) + 2.8662MATS6i − 1.5036GATS7p − 0.1826H-048 + 0.5367O-060 − 0.3698B08[N-O] + 0.3948F06[C-S] − 0.1856SsNH2 + 0.0537minHBint6 + 1.8451MDEN-12 + 0.1755MDEN-22 − 0.8906minaaCH (k = 1.0, $\lvert R^2_0 - R'^2_0 \rvert$ = 0.1428) (2)																
Random (2D-QSAR)	**279**	**91**	**0.6086**	**0.5743**	**0.5648**	**−0.0424**	**0.0324**	**0.7474**	**0.7377**	**0.7375**	**0.7437**	**0.8530**	**0.7682**	**0.6215**	**0.6114**	**0.5224**
pIC_{50} = 4.5562 + 2.5829MATS6i − 0.191nCp + 0.3196O-060 + 0.6746B03[O-S] − 0.2499SsNH2 + 2.4853maxHBd − 2.3712hmin + 1.1784MDEN-12 − 0.6509minaaCH (k = 1.0, $\lvert R^2_0 - R'^2_0 \rvert$ = 0.2436) **(3)**																
2D-QSAR-Full model	**370**	**—**	**0.6426**	**0.6202**	**0.6127**	**−0.0309**	**0.0248**	**—**	**—**	**—**	**—**	**—**	**0.7320**	**—**	**0.5855**	**—**
pIC_{50} = 4.7334 + 2.3826MATS6i − 0.2387nCp + 0.3401O-060 + 0.6301B03[O-S] − 0.248SsNH2 + 2.1364maxHBd − 2.3442hmin + 1.8332MDEN-12 − 0.6324minaaCH (4)																
q-RASAR	**279**	**91**	**0.6059**	**0.5957**	**0.5926**	**−0.0189**	**0.0103**	**0.7528**	**0.7389**	**0.7387**	**0.7449**	**0.8560**	**0.7708**	**0.6201**	**0.6144**	**0.4812**
pIC_{50} = −1.1683 + 0.9192RA function (ED) + 0.0718CATS2D_07_AL + 1.2875LLS_02 (5)																
q-RASAR-Full model	**370**		**0.6392**	**0.6322**	**0.6305**	**−0.0136**	**0.0083**	**—**	**—**	**—**	**—**	**—**	**0.7354**	**—**	**0.5799**	**—**
pIC_{50} = −1.1089 + 0.9149RA function (ED) + 0.0667CATS2D_07_AL + 1.3166LLS_02 (6)																

[+] All abbreviations are explained in the text. The bold typefaces indicate the best splitting model and the recalibrated full model using the same descriptors. The q-RASAR model was also established based on the best splitting model and recalibrated as the q-RASAR-Full model.

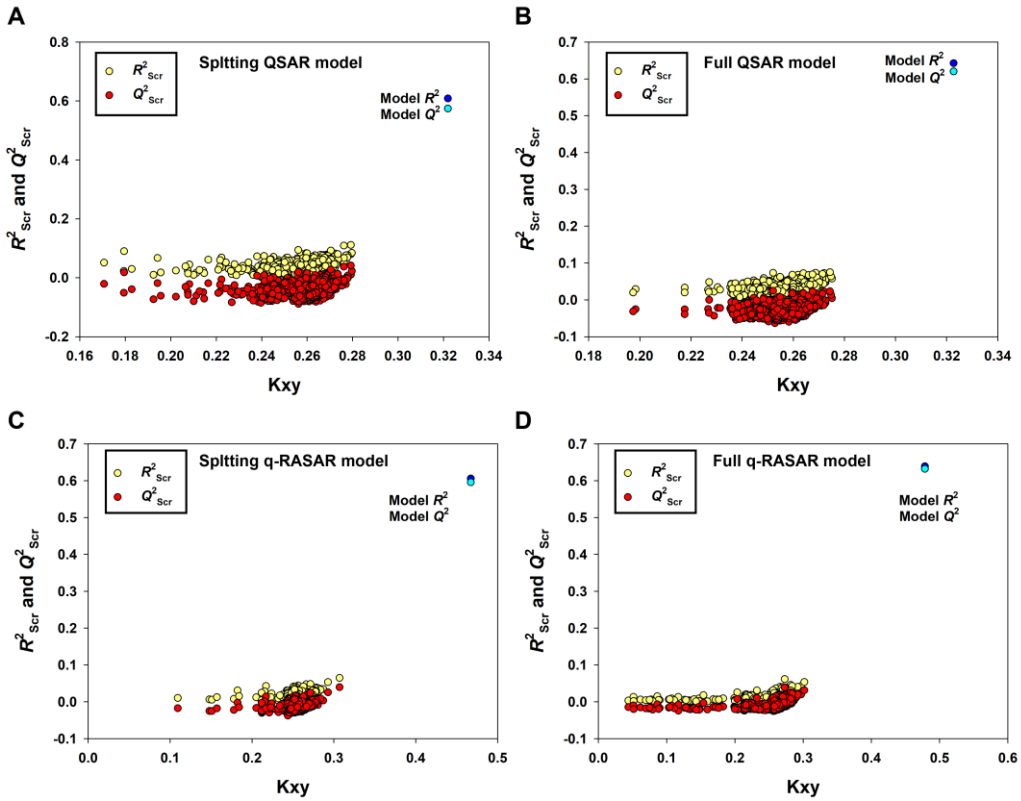

Figure 4. The results of Y-randomization for the Random splitting QSAR model (**A**), the Full QSAR model (**B**), the Random splitting q-RASAR model (**C**) and the Full q-RASAR model (**D**).

In fact, we also tried to use other dataset splitting methods such as Kennard–Stone (http://teqip.jdvu.ac.in/QSAR_Tools/) (accessed on 16 August 2023) and other modeling methods (PLS or stepwise MLR implemented in the Double Cross-Validation v2.0 Software Tool) [57] for model development. However, the quality of these models was not better than that of the Ran (2D-QSAR) model.

Figure 5 shows the graph of experimental versus predicted pIC_{50} values (Figure 5A) and the Williams Plot (Figure 5B) for the AD analysis of the 2D-QSAR model derived from the Random splitting method.

We found that the training and prediction set compounds were homogenously distributed around the trend line, indicating a good predictive ability for query molecules. Considering the AD of the 2D-QSAR model (Figure 5B), only four compounds in the prediction set and eleven compounds in the training set had hat values greater than the warning h^* value (0.108). These were identified as structural outliers, and they may thus be influential in the variable selection in the training set. However, we are not suggesting that these structural outliers cannot be predicted reliably. For example, compound **348** has the maximum hat value, but its predicted residual was very small (0.1601 log unit) (see the detailed data in Table S3). The four prediction set compounds (**299**, **301**, **302**, **336**) were also predicted accurately since their predicted residuals were also small (Table S3). Meanwhile, the predicted residuals of compounds **355**, **356** and **357** were relatively higher (~1 log unit) (Table S3). In contrast, only one compound, **257**, was identified as a response outlier because it had standardized residuals greater than 3.0 standard deviation units (Table S3).

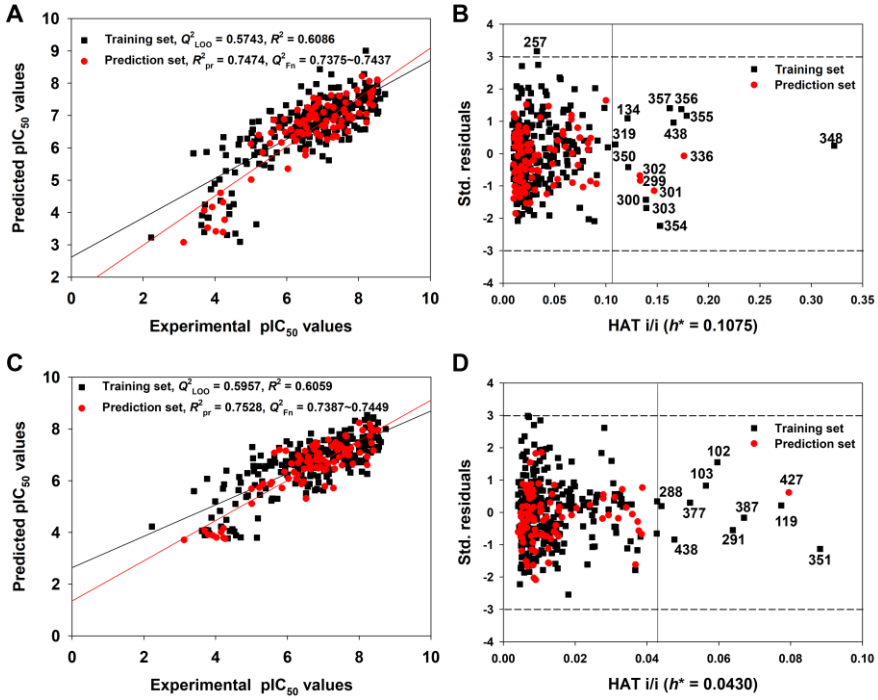

Figure 5. The graph of experimental versus predicted pIC_{50} values (**A**) and the Williams plot (**B**) for the 2D-QSAR model defined by Equation (3); the graph of experimental versus predicted pIC_{50} values (**C**) and the Williams plot (**D**) for the q-RASAR model defined by Equation (5) (Table 1).

Table 2 described the nine molecular descriptors selected by the GA-VSS that were present in the model equation along with their relative importance (Std. coefficient) and physicochemical definitions.

Table 2. Descriptors selected by GA-VSS with the standardized coefficient, range of values and physicochemical definitions.

Descriptors	Std. Coefficient	Range		Definition
	(Full Model)	Min	Max	
MATS6i	0.2152 (0.1946)	−0.212	0.371	Moran autocorrelation of lag 6 weighted by ionization potential (DRAGON)
nCp	−0.106 (−0.1331)	0	6	number of terminal primary C(sp3) (DRAGON)
O-060	0.1908 (0.2012)	0	4	Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X (Atom-centered fragments, Basic descriptors) (DRAGON)
B03[O-S]	0.2623 (0.2452)	0	1	Presence/absence of O-S at topological distance 3 (DRAGON)
SsNH2	−0.4636 (−0.4582)	0	11.662	Sum of atom-type E-State: $–NH_2$ (DRAGON)
maxHBd	0.2085 (0.185)	0	0.764	Maximum E-States for (strong) Hydrogen Bond donors (PaDEL)
hmin	−0.1766 (−0.1792)	−0.447	0.425	Minimum H E-State (PaDEL)
MDEN-12	0.6452 (0.6567)	0	2.515	Molecular distance edge between all primary and secondary nitrogens (PaDEL)
minaaCH	−0.1103 (−0.1063)	1.075	2.329	Minimum atom-type E-State: CH: (PaDEL)

3.3.2. Full Model

As described above, we have verified the external predictive ability of the 2D-QSAR model with the best combination of descriptor variables. Subsequently, based on the

same variables, the model (Equation (3), Table 1) was recalibrated using the entire set of compounds (N_{tr} = 370). The new model was called the 2D-QSAR-Full model (Equation (4), Table 1) and it considered all the available information in the training and test sets.

The endpoint, expressed as pIC_{50} ($-logIC_{50}$, mol/L), ranged from 2.22 to 8.74, spanning more than six log units and suggesting that the dataset is adequate for QSAR studies. As shown in Table 1, the 2D-QSAR-Full model showed satisfactory internal fitness (R^2 = 0.6426) and robustness (Q^2_{LOO} = 0.6202, Q^2_{LMO} = 0.6127). Again, the values of Q^2_{Yscr} (-0.0309) and R^2_{Yscr} (0.0248) were very low, indicating the absence of any chance correlation. The graph of experimental versus predicted pIC_{50} values (Figure 6A) and the Williams Plot (Figure 6B) are given below.

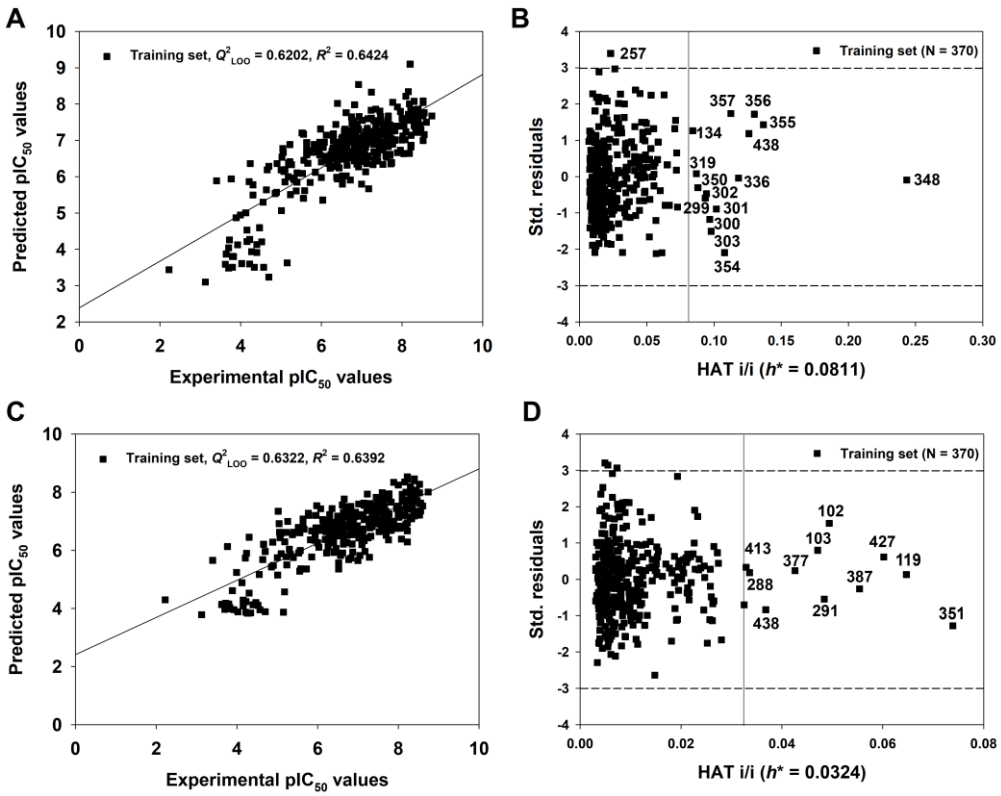

Figure 6. The graph of experimental versus predicted pIC_{50} values (**A**) and the Williams plot (**B**) for the 2D-QSAR-Full model defined by Equation (4) in Table 1; the graph of experimental versus predicted pIC_{50} values (**C**) and the Williams plot (**D**) for the q-RASAR-Full model defined by Equation (6) in Table 1.

Molecular descriptors were calculated from the 2D structural information using Dragon [36] and PaDEL [37] software. We emphasize that these descriptors capture the global properties of the molecular structure or encode for some specific groups or fragments, such as electronic accessibility (E-State), spatial autocorrelations (2D autocorrelation) or the presence or absence of a specific fragment. The scatter plot of each descriptor versus pIC_{50} was shown in Figure 7. The value of each descriptor was listed in Table S5 in the Supplementary Materials.

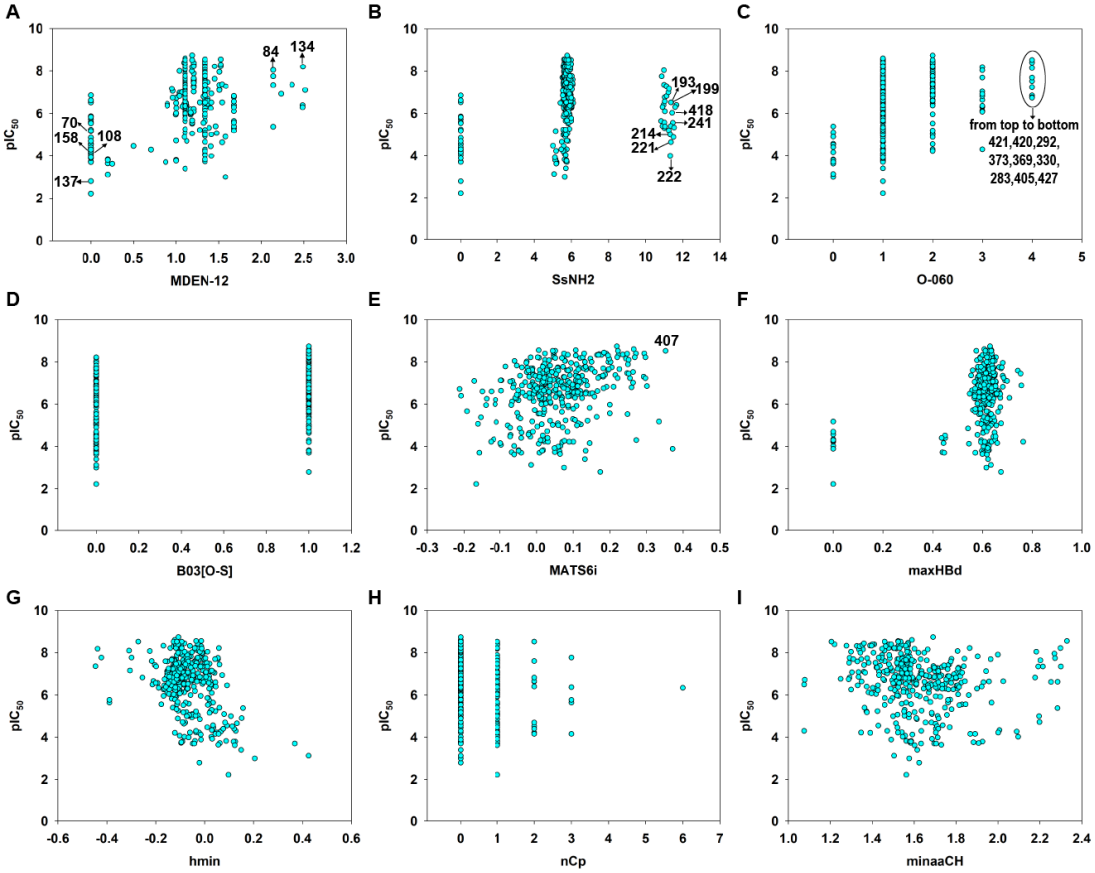

Figure 7. Variable scatter plot of each descriptor versus pIC_{50}. MDEN-12 (**A**), SsNH2 (**B**), O-040 (**C**), B03[O-S] (**D**), MATS6i (**E**), maxHBd (**F**), hmin (**G**), nCp (**H**) and minaaCH (**I**).

According to the model Equation (4) (Table 1) and the standardized coefficients of each variable (Table 2), the most important descriptors for MGMT inactivation were MDEN-12 (std. coefficient 0.6567) and SsNH2 (std. coefficient −0.4582). It should be noted that MDEN-12 was positively correlated (Figure 7A), while $SsNH_2$ was negatively correlated with the MGMT inactivation potency (Figure 7B). MDEN-12 is the molecular distance edge between all primary and secondary nitrogens [58]; for example, compounds **84** and **134** with high MDEN-12 values (2.144 and 2.490, respectively) were strong inactivators (pIC_{50} = 8.04 and 8.20, respectively). Figure S2 showed MDEN-12 descriptor values for the two benchmark inactivators Lomeguatrib and O^6-BG and the selected compounds. This descriptor also highlights the importance of the presence of $–NH_2$: compounds without 2′-NH_2 (such as **70**, **108**, **111** and **158**) were commonly less effective in MGMT inactivation. This is consistent with our previous study indicating that the 2′-NH_2 of guanine is essential for inactivation because it plays an important role in hydrogen bond formation with Cys145/Val148 residues of MGMT (see Figure 1) [3,8,59]. However, MDEN-12 as a single descriptor did not adequately model the MGMT inactivation potency in a general model; hence, the GA-VSS selected additional descriptors to obtain a model with higher predictivity. SsNH2 represents the sum of atom-type electrotopological states (E-State): $–NH_2$ [58], indicating that the aliphatic primary amino can compromise the MGMT inactivation potency to a certain extent, especially for the guanine derivatives **193**, **199**, **214**, **221**, **222**, **241** and **418,** which have an aliphatic amino in the N9 position. This was also supported by a previous study indicating that a large polar group at the N9 position of guanine was not well tolerated [14]. Figure S3 showed SsNH2 descriptor values for Lomeguatrib, O^6-BG and the selected compounds. Indeed, MDEN-12 and SsNH2 were two mutually balanced descriptors in the model, as indicated by the compounds **64**, **65** and

66, since they had low values (0) for the two opposite descriptors, but a moderate potency (pIC_{50} = 5.72, 6.52 and 4.55, respectively) (see Table S5).

O-060 (std. coefficient 0.2012) and B03[O-S] (std. coefficient 0.2452) are two descriptors that are related to the presence of a specific group or fragment [58], and on the basis of these coefficients, they were positive contributors to MGMT inactivation potency. O-060 belongs to the basic descriptors of atom-centered fragments, representing the presence of Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X fragments. In the entire dataset, this descriptor had discrete values of 0, 1, 2, 3 and 4, respectively. There were nine compounds (**283, 292, 330, 369, 373, 405, 420, 421** and **427**) that had the maximum O-060 value of four and a relatively high MGMT inactivation activity (pIC_{50} = 6.764–8.520) (Figure 7C). The values of the O-060 descriptor for Lomeguatrib, O^6-BG, and the selected compounds are shown in Figure S4. B03[O-S] indicates the presence or absence of O-S at topological distance 3, and it was clear that the thiophene group substituted on the guanine O^6 position in compounds like **112**, **401** and **402** (Tables S1 and S5) contributed substantially to their high inactivation potency (Figure 7D). Figure S5 shows B03[O-S] values for Lomeguatrib, O^6-BG and the selected compounds.

MATS6i (std. coefficient 0.1946) is the Moran autocorrelation of lag 6 weighted by the ionization potential [58]. It indicates the relative charge distribution of a molecule, i.e., the electron cloud, and thus may enhance charge or hydrogen bond interactions with the target. Because base analog-mediated MGMT inactivation is absolutely dependent on the ability to donate a carbocation to the active site of MGMT [1,3,6], MATS6i is generally a positive contributor to the response endpoint (Figure 7E). Figure S6 showed MATS6i values for Lomeguatrib, O^6-BG and the selected compound **407**, which had a high MATS6i value (0.352) and a high inactivation potency (pIC_{50} = 8.523) (Table S5).

The descriptor maxHBd (std. coefficient 0.1850) indicates the maximum E-States for (strong) hydrogen bond donors and clearly contributes to increasing the inactivation activity (Figure 7F). For example, compounds **411** and **90** were strong inactivators (pIC_{50} = 8.55 and 8.52, respectively) with high maxHBd values (0.629 and 0.637, respectively) (see Table S5). Figure S7 shows the values of the maxHBd descriptor for Lomeguatrib, O^6-BG and the selected compounds.

The last three descriptors were nCp (−0.1331), hmin (std. coefficient −0.1792) and minaaCH (−0.1063) (see Table 2), the latter two being E-state descriptors [58]. Individually, these three descriptors were less important in defining the model equation but supported the six main descriptors. The descriptor hmin indicates a minimum H E-State, which encodes for the minimum E-State of hydrogen atoms. The minaaCH descriptor represents the minimum atom-type E-State aromatic-CH-aromatic, in which atom-type E-state indices are computed by summing the E-state values of all atoms of the same atom type in a molecule [58]. These descriptors characterize the information related to the electronic accessibility of an atom and hence the probability of intermolecular interactions [60]. In our model, the two E-state descriptors were negatively correlated with the response endpoint, which was consistent with the aquatic toxicity models of pesticide and pharmaceuticals [41,61]. The nCp descriptor represents the number of terminal primary sp3 carbons [58], and this was also inversely related to the response according to its equation coefficient. The values of the three descriptors for Lomeguatrib, O^6-BG and the selected compounds are shown in Figure S8.

The developed model was derived from multivariable combinations based on a statistically driven procedure (i.e., GA-VSS-based selection). Thus, none of the descriptors can independently explain the distribution of the modeled endpoint, and only the combination of all selected descriptors can accurately model the response to be studied.

3.4. q-RASAR Analysis

After the development of the 2D-QSAR model, the same training and test set files were used as inputs for quantitative Read-Across predictions using three different similarity-based functions, namely, the Euclidean Distance, Gaussian Kernel function and Laplacean

Kernel function [47,48]. For the predictions of prediction set compounds, a default sigma value (σ) of 1 for the Gaussian kernel function and a default gamma value (γ) of 1 for the Laplacian kernel function were used, and the distance threshold value and similarity threshold value were set as 1 and 0, respectively. The number of the closest training compounds for activity prediction was six. It was found that the external validation parameters like Q^2_{F1} (0.7401), Q^2_{F2} (0.7399), $RMSE_{pr}$ (0.6187) and MAE_{pr} (0.4802) from quantitative Read-Across using the Euclidean Distance (see Table S6) were better than those of 2D-QSAR.

To establish QSAR-based Read-Across predictions, we performed the q-RASAR modeling [49,50]. The equation of the q-RASAR model (Equation (5)) is listed in Table 1. In Equation (5), the RA function (ED) variable was a Euclidean Distance-based Read-Across prediction function obtained from the original 2D descriptors. It can be accessed by the free online tool RASAR-Desc-Calc-v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) (accessed on 1 April 2023). The low difference between R^2 and Q^2_{LOO} indicated the robustness of the model and the higher values of R^2_{pr}, Q^2_{F1}, Q^2_{F2} and Q^2_{F3}, and the lower MAE_{pr} value suggested the good predictivity and transferability of the q-RASAR model. Due to the good internal robustness and external predictivity, we also constructed the q-RASAR-full model (Equation (6) in Table 1) using all the available information. Low values of Q^2_{Yscr} and R^2_{Yscr} (0.0324) indicated that the q-RASAR model and q-RASAR-Full model were not generated by chance correlation (Figure 4C,D).

The linear correlations for the q-RASAR and q-RASAR-Full models are shown in Figures 5C and 6C, respectively. Meanwhile, the AD analysis of the q-RASAR and q-RASAR-Full models is shown in Figures 5D and 6D. We found relatively fewer outliers in q-RASAR modeling compared to 2D-QSAR modeling.

The detailed information about the q-RASAR model is listed in Table S7 in the Supplementary Materials. The values of each variable in the q-RASAR-Full model are listed in Table S8 in the Supplementary Materials.

3.5. Application of the 2D-QSAR-Based Full Model and q-RASAR-Full Model

We constructed a true external set consisting of 66 unknown molecules. After calculating their descriptors, the 2D-QSAR-Full model was applied to predict their pIC_{50} values. As shown in Figure 8A, 12 of 66 true external compounds lay outside the model's AD, since their h values are greater than h^* (0.081), suggesting >80% prediction coverage. In particular, compound **36** has the highest h value (0.2664 >> h^*). If we defined the AD of the model using the PCA approach (Figure 8B), only one compound (again, compound **36**) in the true external set falls outside the AD, resulting in a more significant prediction coverage (98.5%).

Similarly, the q-RASAR-Full model was also applied to the true external compounds. As shown in Figure 8C, only one true external compound (**436**) lay outside the model's AD, suggesting >98% prediction coverage. In fact, compound **436** only has a slightly higher h value (0.0387) compared to the threshold value h^* (0.0324). Using the PCA approach (Figure 8D), only one compound (compound **308**) in the true external set falls outside the AD, also showing a considerable prediction coverage.

Subsequently, we also used the "Prediction Reliability Indicator" tool (http://dtclab.webs.com/software-tools) (accessed on 22 December 2022) [62] to check the prediction quality for each true external compound. Each compound is scored (composite score of 3, 2 or 1) based on the absolute prediction errors that correspond to "Good", "Moderate" or "Bad or Unreliable" prediction quality, respectively. As shown in Table S9, we found 56 "Good" compounds, 10 "Moderate" compounds and no "Bad or Unreliable" compounds derived from the 2D-QSAR-Full model. As for the q-RASAR-Full model, we found 61 "Good" compounds, 5 "Moderate" compounds and no "Bad or Unreliable" compounds (Table S10). The results suggest that our Full models, especially the latter, have a wide and reliable prediction scope and that they can be used to forecast the MGMT inactivation potency of untested compounds. A priori designed compounds would be identified by our validated

model and the most potent prioritized so that, time, money and resources would be saved. Of course, the multi-objective optimization modeling is also very important, especially when simultaneously considering the bioactivity, bioavailability and toxicity [63,64].

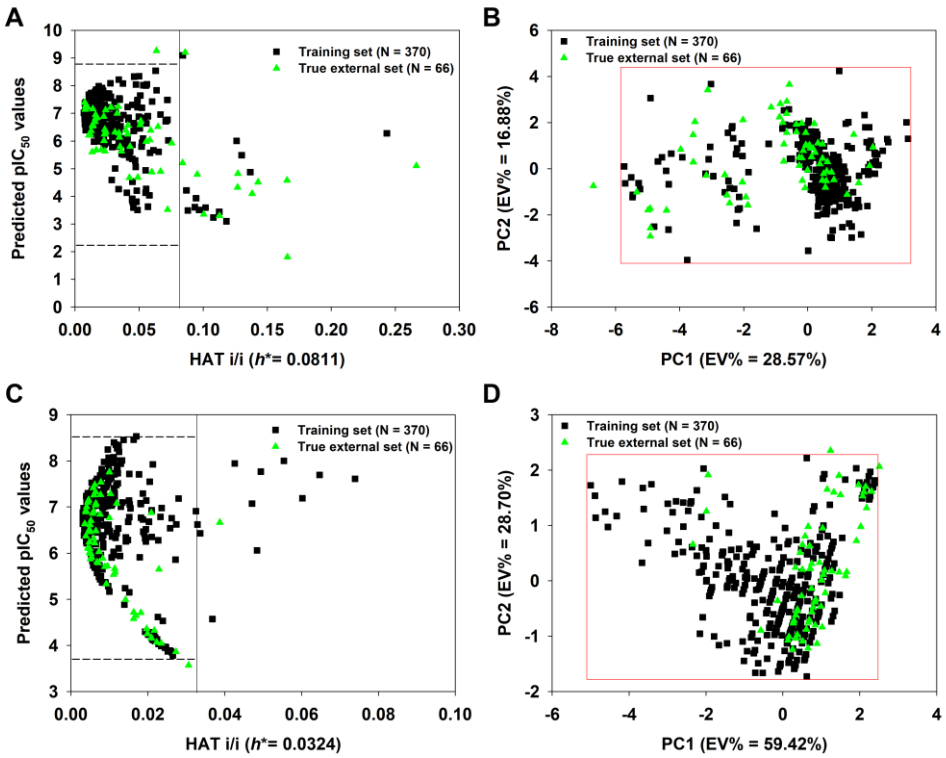

Figure 8. Insubria graph of the 2D-QSAR-Full model (**A**) and principal component analysis (PCA) plot based on the selected nine descriptors shown in Table 2 (**B**); Insubria graph of the q-RASAR-Full model (**C**) and PCA plot based on the three variables shown in Equation (6) (**D**). The four plots show the applicability domain (AD) when the two Full models are applied to true external compounds without experimental values.

4. Conclusions

In this study, using the experimental IC_{50} values for a total of 370 MGMT inactivators, we developed QSAR models using a GA-MLR method, and Dragon and PaDEL software were combined to calculate molecular descriptors for model establishment. Three splitting models were assessed for robustness, reliability, fitness and predictivity. After selecting the best splitting model, a 2D-QSAR-Full model was then recalibrated using all the available experimental information (370 compounds). The mechanistic interpretation indicated that the status of nitrogen atoms, aliphatic primary amino groups, the presence of O-S at topological distance 3, the presence of Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X, the ionization potential and hydrogen bond donors are the main factors controlling MGMT inactivation potency. Using the selected features in the 2D-QSAR and chemical Read-Across technique, we developed the q-RASAR model, which exhibited better external predictive ability. The AD analysis showed that the splitting 2D-QSAR model and 2D-QSAR-Full model had a significantly high coverage for the test set and true external set. In summary, the QSAR Full model developed in this study can be used for optimizing the design of novel MGMT inactivators. Thus, for novel untested compounds, we can predict their IC_{50} if they are located at the applicability domain, focus on compounds with a high inactivation potential and, hence, reduce unnecessary chemical synthesis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pharmaceutics15082170/s1. Figures S1–S8 and Tables S1–S10.

Author Contributions: Conceptualization, G.S. and G.P.M.; methodology, G.S. and P.B.; software, T.F. and L.Z.; validation, P.B.; formal analysis, G.S.; investigation, G.S., P.B., J.K., D.J.D., J.E.M., R.S.M. and T.B.H.M.; resources, R.Z.; data curation, G.S. and G.P.M.; writing—original draft preparation, G.S. and P.B.; writing—review and editing, G.P.M.; supervision, R.Z. and G.P.M.; project administration, G.S.; funding acquisition, G.S. and G.P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 82003599), The Project of Cultivation for Young Top-Notch Talents of Beijing Municipal Institutions (No. BPHR202203016), the Science and Technology General Project of Beijing Municipal Education Commission (No. KM202110005005), the Beijing Natural Science Foundation (No. 7222016) and the Beijing Science and Technology Plan (No. Z221100007122004), the International Research Cooperation Seed Fund of Beijing University of Technology (No. 2021B41), the Cancer Research Campaign, Cancer Research UK, the Christie Hospital Trust, Schering-Plough (now Merck & Co), the Irish Cancer Society, the Health Research Board (Ireland), the Leukemia Research Fund and the European Union Framework 6 program No. 037665 CHEMORES.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on request.

Acknowledgments: The authors thank P. Gramatica (University of Insubria, Varese, Italy) for authorizing the use of the QSARINS 2.2.4 software and K. Roy (Jadavpur University, Kolkata, India) for authorizing the use of the "Prediction Reliability Indicator" tools. T. Brian H. McMurry thanks Paul Murray, Sharon Bergin, Dawn Ronan, Una Hanafin, Christophe Carola and Celine Blais for the compound syntheses. This publication is dedicated to the memory of R. Stanley MacElhinney.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaina, B.; Margison, G.P.; Christmann, M. Targeting O^6-methylguanine-DNA methyltransferase with specific inhibitors as a strategy in cancer therapy. *Cell. Mol. Life Sci.* **2010**, *67*, 3663–3681. [CrossRef]
2. Nikolova, T.; Roos, W.P.; Kramer, O.H.; Strik, H.M.; Kaina, B. Chloroethylating nitrosoureas in cancer therapy: DNA damage, repair and cell death signaling. *Biochim. Biophys. Acta-Rev. Cancer* **2017**, *1868*, 29–39. [CrossRef]
3. Sun, G.H.; Zhao, L.J.; Zhong, R.G.; Peng, Y.Z. The specific role of O-6-methylguanine-DNA methyltransferase inhibitors in cancer chemotherapy. *Future Med. Chem.* **2018**, *10*, 1971–1996. [CrossRef]
4. Gnewuch, C.T.; Sosnovsky, G. A critical appraisal of the evolution of N-nitrosoureas as anticancer drugs. *Chem. Rev.* **1997**, *97*, 829–1013. [CrossRef]
5. Kaina, B.; Christmann, M. DNA repair in personalized brain cancer therapy with temozolomide and nitrosoureas. *DNA Repair* **2019**, *78*, 128–141. [CrossRef]
6. Pegg, A.E. Multifaceted roles of alkyltransferase and related proteins in DNA repair, DNA damage, resistance to chemotherapy, and research tools. *Chem. Res. Toxicol.* **2011**, *24*, 618–639. [CrossRef]
7. Lindahl, T.; Demple, B.; Robins, P. Suicide inactivation of the Escherichia-coli O^6-methylguanine-DNA methyltransferase. *EMBO J.* **1982**, *1*, 1359–1363. [CrossRef]
8. Daniels, D.S.; Woo, T.T.; Luu, K.X.; Noll, D.M.; Clarke, N.D.; Pegg, A.E.; Tainer, J.A. DNA binding and nucleotide flipping by the human DNA repair protein AGT. *Nat. Struct. Mol. Biol.* **2004**, *11*, 714–720. [CrossRef]
9. Daniels, D.S.; Mol, C.D.; Arvai, A.S.; Kanugula, S.; Pegg, A.E.; Tainer, J.A. Active and alkylated human AGT structures: A novel zinc site, inhibitor and extrahelical base binding. *EMBO J.* **2000**, *19*, 1719–1730. [CrossRef]
10. Xu-Welliver, M.; Pegg, A.E. Degradation of the alkylated form of the DNA repair protein, O^6-alkylguanine-DNA alkyltransferase. *Carcinogenesis* **2002**, *23*, 823–830. [CrossRef]
11. Dolan, M.E.; Morimoto, K.; Pegg, A.E. Reduction of O^6-alkylguanine-DNA alkyltransferase activity in HeLa-cells treated with O^6-alkylguanines. *Cancer Res.* **1985**, *45*, 6413–6417.
12. Dolan, M.E.; Moschel, R.C.; Pegg, A.E. Depletion of mammalian O^6-alkylguanine-DNA alkyltransferase activity by O^6-benzylguanine provides a means to evaluate the role of this protein in protection against carcinogenic and therapeutic alkylating-agents. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 5368–5372. [CrossRef]
13. Moschel, R.C.; McDougall, M.G.; Dolan, M.E.; Stine, L.; Pegg, A.E. Structural features of substituted purine derivatives compatible with depletion of human O^6-alkylguanine-DNA alkyltransferase. *J. Med. Chem.* **1992**, *35*, 4486–4491. [CrossRef]
14. Chae, M.Y.; McDougall, M.G.; Dolan, M.E.; Swenn, K.; Pegg, A.E.; Moschel, R.C. Substituted O^6-benzylguanine derivatives and their inactivation of human O^6-alkylguanine-DNA alkyltransferase. *J. Med. Chem.* **1994**, *37*, 342–347. [CrossRef]

15. Chae, M.Y.; Swenn, K.; Kanugula, S.; Dolan, M.E.; Pegg, A.E.; Moschel, R.C. 8-Substituted O^6-benzylguanine, substituted 6(4)-(benzyloxy)pyrimidine, and related derivatives as inactivators of human O^6-alkylguanine-DNA alkyltransferase. *J. Med. Chem.* **1995**, *38*, 359–365. [CrossRef]
16. Sun, G.H.; Fan, T.J.; Sun, X.D.; Hao, Y.X.; Cui, X.; Zhao, L.J.; Ren, T.; Zhou, Y.; Zhong, R.G.; Peng, Y.Z. In silico prediction of O^6-methylguanine-dna methyltransferase inhibitory potency of base analogs with QSAR and machine learning methods. *Molecules* **2018**, *23*, 2892. [CrossRef]
17. Dolan, M.E.; Roy, S.K.; Garbiras, B.J.; Helft, P.; Paras, P.; Chae, M.Y.; Moschel, R.C.; Pegg, A.E. O^6-alkylguanine-DNA alkyltransferase inactivation by ester prodrugs of O-6-benzylguanine derivatives and their rate of hydrolysis by cellular esterases. *Biochem. Pharmacol.* **1998**, *55*, 1701–1709. [CrossRef]
18. McElhinney, R.S.; Donnelly, D.J.; McCormick, J.E.; Kelly, J.; Watson, A.J.; Rafferty, J.A.; Elder, R.H.; Middleton, M.R.; Willington, M.A.; McMurry, T.B.H.; et al. Inactivation of O^6-alkylguanine-DNA alkyltransferase. 1. Novel O-6-(hetarylmethyl)guanines having basic rings in the side chain. *J. Med. Chem.* **1998**, *41*, 5265–5271. [CrossRef]
19. Terashima, I.; Kohda, K. Inhibition of human O^6-alkylguanine-DNA alkyltransferase and potentiation of the cytotoxicity of chloroethylnitrosourea by 4(6)-(benzyloxy)-2,6(4)-diamino-5-(nitro or nitroso)pyrimidine derivatives and analogues. *J. Med. Chem.* **1998**, *41*, 503–508. [CrossRef]
20. Griffin, R.J.; Arris, C.E.; Bleasdale, C.; Boyle, F.T.; Calvert, A.H.; Curtin, N.J.; Dalby, C.; Kanugula, S.; Lembicz, N.K.; Newell, D.R.; et al. Resistance-modifying agents. 8. Inhibition of O^6-alkylguanine-DNA alkyltransferase by O^6-alkenyl-, O^6-cycloalkenyl-, and O^6-(2-oxoalkyl)guanines and potentiation of temozolomide cytotoxicity in vitro by O-6-(1-cyclopentenylmethyl)guanine. *J. Med. Chem.* **2000**, *43*, 4071–4083. [CrossRef]
21. Reinhard, J.; Hull, W.E.; von der Lieth, C.W.; Eichhorn, U.; Kliem, H.C.; Kaina, B.; Wiessler, M. Monosaccharide-linked inhibitors of O-6-methylguanine-DNA methyltransferase (MGMT): Synthesis, molecular modeling, and structure-activity relationships. *J. Med. Chem.* **2001**, *44*, 4050–4061. [CrossRef] [PubMed]
22. Pauly, G.T.; Loktionova, N.A.; Fang, Q.M.; Vankayala, S.L.; Guida, W.C.; Pegg, A.E. Substitution of aminomethyl at the meta-position enhances the inactivation of O^6-alkylguanine-DNA alkyltransferase by O^6-benzylguanine. *J. Med. Chem.* **2008**, *51*, 7144–7153. [CrossRef] [PubMed]
23. Ranson, M.; Middleton, M.R.; Bridgewater, J.; Lee, S.M.; Dawson, M.; Jowle, D.; Halbert, G.; Waller, S.; McGrath, H.; Gumbrell, L.; et al. Lomeguatrib, a potent inhibitor of O-6-alkylguanine-DNA-alkyltransferase: Phase I safety, pharmacodynamic, and pharmacokinetic trial and evaluation in combination with temozolomide in patients with advanced solid tumors. *Clin. Cancer Res.* **2006**, *12*, 1577–1584. [CrossRef] [PubMed]
24. Warren, K.E.; Gururangan, S.; Geyer, J.R.; McLendon, R.E.; Poussaint, T.Y.; Wallace, D.; Balis, F.M.; Berg, S.L.; Packer, R.J.; Goldman, S.; et al. A phase II study of O^6-benzylguanine and temozolomide in pediatric patients with recurrent or progressive high-grade gliomas and brainstem gliomas: A Pediatric Brain Tumor Consortium study. *J. Neuro-Oncol.* **2012**, *106*, 643–649. [CrossRef]
25. Quinn, J.A.; Jiang, S.X.; Reardon, D.A.; Desjardins, A.; Vredenburgh, J.J.; Rich, J.N.; Gururangan, S.; Friedman, A.H.; Bigner, D.D.; Sampson, J.H.; et al. Phase II trial of temozolomide plus O^6-benzylguanine in adults with recurrent, temozolomide-resistant malignant glioma. *J. Clin. Oncol.* **2009**, *27*, 1262–1267. [CrossRef]
26. Watson, A.J.; Sabharwal, A.; Thorncroft, M.; McGown, G.; Kerr, R.; Bojanic, S.; Soonawalla, Z.; King, A.; Miller, A.; Waller, S.; et al. Tumor O(6)-methylguanine-DNA methyltransferase inactivation by oral lomeguatrib. *Clin. Cancer Res.* **2010**, *16*, 743–749. [CrossRef]
27. Pegg, A.E.; Kanugula, S.; Edara, S.; Pauly, G.T.; Moschel, R.C.; Goodtzova, K. Reaction of O^6-benzylguanine-resistant mutants of human O^6-alkylguanine-DNA alkyltransferase with O^6-benzylguanine in oligodeoxyribonucleotides. *J. Biol. Chem.* **1998**, *273*, 10863–10867. [CrossRef]
28. Dolan, M.E.; Chae, M.Y.; Pegg, A.E.; Mullen, J.H.; Friedman, H.S.; Moschel, R.C. Metabolism of O^6-benzylguanine, an inactivator of O^6-alkylguanine-dna alkyltransferase. *Cancer Res.* **1994**, *54*, 5123–5130.
29. Zhu, R.; Liu, M.C.; Luo, M.Z.; Penketh, P.G.; Baumann, R.P.; Shyam, K.; Sartorelli, A.C. 4-Nitrobenzyloxycarbonyl derivatives of O^6-benzylguanine as hypoxia-activated prodrug inhibitors of O^6-alkylguanine-DNA alkyltransferase (AGT), which produces resistance to agents targeting the O^6 position of DNA guanine. *J. Med. Chem.* **2011**, *54*, 7720–7728. [CrossRef]
30. Zhu, R.; Seow, H.A.; Baumann, R.P.; Ishiguro, K.; Penketh, P.G.; Shyam, K.; Sartorelli, A.C. Design of a hypoxia-activated prodrug inhibitor of O-6-alkylguanine-DNA alkyltransferase. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 6242–6247. [CrossRef]
31. Huang, T.; Sun, G.; Zhao, L.; Zhang, N.; Zhong, R.; Peng, Y. Quantitative structure-activity relationship (QSAR) studies on the toxic effects of nitroaromatic compounds (NACs): A Systematic Review. *Int. J. Mol. Sci.* **2021**, *22*, 8557. [CrossRef] [PubMed]
32. Khan, K.; Roy, K. Ecotoxicological risk assessment of organic compounds against various aquatic and terrestrial species: Application of interspecies i-QSTTR and species sensitivity distribution techniques. *Green Chem.* **2022**, *24*, 1458–1516. [CrossRef]
33. Gramatica, P. Principles of QSAR modeling: Comments and suggestions from personal experience. *Int. J. Quant. Struct. Prop. Relatsh.* **2020**, *5*, 61–97. [CrossRef]
34. Li, F.; Sun, G.; Fan, T.; Zhang, N.; Zhao, L.; Zhong, R.; Peng, Y. Ecotoxicological QSAR modelling of the acute toxicity of fused and non-fused polycyclic aromatic hydrocarbons (FNFPAHs) against two aquatic organisms: Consensus modelling and comparison with ECOSAR. *Aquat. Toxicol.* **2023**, *255*, 106393. [CrossRef] [PubMed]

35. Chen, S.; Sun, G.; Fan, T.; Li, F.; Xu, Y.; Zhang, N.; Zhao, L.; Zhong, R. Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): Assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods. *Sci. Total Environ.* **2023**, *876*, 162736. [CrossRef]
36. Kode Srl. Dragon Software for Molecular Descriptor Calculation V 7.0.6. Available online: https://chm.kode-solutions.net/ (accessed on 3 September 2017).
37. Yap, C.W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
38. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J. Comput. Chem.* **2013**, *34*, 2121–2132. [CrossRef]
39. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-Chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *J. Comput. Chem.* **2014**, *35*, 1036–1044. [CrossRef]
40. Gramatica, P.; Cassani, S.; Sangion, A. Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives' design. *Green Chem.* **2016**, *18*, 4393–4406. [CrossRef]
41. Sangion, A.; Gramatica, P. Hazard of pharmaceuticals for aquatic environment: Prioritization by structural approaches and prediction of ecotoxicity. *Environ. Int.* **2016**, *95*, 131–143. [CrossRef]
42. OECD (Organization for Economic Co-Operation and Development). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q) SAR] Models*; OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69; OECD: Paris, France, 2007. Available online: http://www.oecd.org/env/guidance-document-onthe-validation-of-quantitative-structure-activity-relationship-q-sar-models-9789264085442-en.htm (accessed on 12 March 2021).
43. Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graph.* **2002**, *20*, 269–276. [CrossRef] [PubMed]
44. Gramatica, P.; Sangion, A. A historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics and terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127–1131. [CrossRef] [PubMed]
45. Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: Theory development and its application in chemometrics. *Chemometr. Intell. Lab. Syst.* **1999**, *46*, 13–29. [CrossRef]
46. Roy, K.; Ambure, P.; Kar, S.; Ojha, P.K. Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? *J. Chemom.* **2018**, *32*, e2992. [CrossRef]
47. Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environ. Sci. Nano* **2022**, *9*, 189–203. [CrossRef]
48. Banerjee, A.; Chatterjee, M.; De, P.; Roy, K. Quantitative predictions from chemical read-across and their confidence measures. *Chemometr. Intell. Lab.* **2022**, *227*, 104613. [CrossRef]
49. Banerjee, A.; Roy, K. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. *Mol. Divers.* **2022**, *26*, 2847–2862. [CrossRef]
50. Banerjee, A.; De, P.; Kumar, V.; Kar, S.; Roy, K. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening endocrine disruptor chemicals using 2D-QSAR and chemical read-across. *Chemosphere* **2022**, *309*, 136579. [CrossRef]
51. Cruz-Monteagudo, M.; Medina-Franco, J.L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M.N.D.S.; Borges, F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **2014**, *19*, 1069–1080. [CrossRef]
52. Cruz-Monteagudo, M.; Medina-Franco, J.L.; Perera-Sardiña, Y.; Borges, F.; Tejera, E.; Paz-Y-Miño, C.; Pérez-Castillo, Y.; Sánchez-Rodríguez, A.; Contreras-Posada, Z.; Cordeiro, M.N. Probing the hypothesis of SAR continuity restoration by the removal of activity cliffs generators in QSAR. *Curr. Pharm. Des.* **2016**, *22*, 5043–5056. [CrossRef]
53. Sun, G.; Zhang, Y.; Pei, L.; Lou, Y.; Mu, Y.; Yun, J.; Li, F.; Wang, Y.; Hao, Z.; Xi, S.; et al. Chemometric QSAR modeling of acute oral toxicity of Polycyclic Aromatic Hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse. *Ecotoxicol. Environ. Saf.* **2021**, *222*, 112525. [CrossRef] [PubMed]
54. Li, F.F.; Fan, T.J.; Sun, G.H.; Zhao, L.J.; Zhong, R.G.; Peng, Y.Z. Systematic QSAR and iQCCR modelling of fused/non-fused aromatic hydrocarbons (FNFAHs) carcinogenicity to rodents: Reducing unnecessary chemical synthesis and animal testing. *Green Chem.* **2022**, *24*, 5304–5319. [CrossRef]
55. Hao, Y.X.; Sun, G.H.; Fan, T.J.; Tang, X.Y.; Zhang, J.; Liu, Y.D.; Zhang, N.; Zhao, L.J.; Zhong, R.G.; Peng, Y.Z. In vivo toxicity of nitroaromatic compounds to rats: QSTR modelling and interspecies toxicity relationship with mouse. *J. Hazard. Mater.* **2020**, *399*, 122981. [CrossRef] [PubMed]
56. Ghose, A.K.; Crippen, G.M. Atomic physicochemical parameters for 3-dimensional structure-directed quantitative structure-activity-relationships I. Partition-coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577. [CrossRef]
57. Roy, K.; Ambure, P. The "double cross-validation" software tool for MLR QSAR model development. *Chemometr. Intell. Lab. Syst.* **2016**, *159*, 108–126. [CrossRef]
58. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, Second, Revised and Enlarged Edition*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
59. Sun, G.H.; Fan, T.J.; Zhang, N.; Ren, T.; Zhao, L.J.; Zhong, R.G. Identification of the structural features of guanine derivatives as MGMT inhibitors using 3D-QSAR modeling combined with molecular docking. *Molecules* **2016**, *21*, 823. [CrossRef]
60. Hall, L.H.; Kier, L.B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045. [CrossRef]

61. Galimberti, F.; Moretto, A.; Papa, E. Application of chemometric methods and QSAR models to support pesticide risk assessment starting from ecotoxicological datasets. *Water Res.* **2020**, *174*, 115583. [CrossRef]
62. Roy, K.; Ambure, P.; Kar, S. How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* **2018**, *3*, 11392–11406. [CrossRef]
63. Sánchez-Rodríguez, A.; Pérez-Castillo, Y.; Schürer, S.C.; Nicolotti, O.; Mangiatordi, G.F.; Borges, F.; Cordeiro, M.N.D.S.; Tejera, E.; Medina-Franco, J.L.; Cruz-Monteagudo, M. From flamingo dance to (desirable) drug discovery: A nature-inspired approach. *Drug Discov. Today* **2017**, *22*, 1489–1502. [CrossRef]
64. Lambrinidis, G.; Tsantili-Kakoulidou, A. Challenges with multi-objective QSAR in drug discovery. *Expert. Opin. Drug Discov.* **2018**, *13*, 851–859. [CrossRef] [PubMed]