Mechanistic Analysis of Chemically Diverse Bromodomain-4 Inhibitors Using Balanced QSAR Analysis and Supported by X-ray Resolved Crystal Structures

Magdi E. A. Zaki 1,* , Sami A. Al-Hussain 1, Aamal A. Al-Mutairi 1, Vijay H. Masand 2,* , Abdul Samad 3 and Rahul D. Jawarkar 4

1 Department of Chemistry, Faculty of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia; sahussain@imamu.edu.sa (S.A.A.-H.); aamutairi@imamu.edu.sa (A.A.A.-M.)
2 Department of Chemistry, Vidyabharti Mahavidyalaya, Amravati 444602, India
3 Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Tishk International University, Erbil 44001, Iraq; abdul.samad@tiu.edu.iq
4 Department of Medicinal Chemistry, Dr. Rajendra Gode Institute of Pharmacy, University-Mardi Road, Amravati 444901, India; rahuljawarkar@gmail.com
* Correspondence: mezaki@imamu.edu.sa (M.E.A.Z.); vijaymasand@gmail.com (V.H.M.)

Abstract: Bromodomain-4 (BRD-4) is a key enzyme in post-translational modifications, transcriptional activation, and many other cellular processes. Its inhibitors find their therapeutic usage in cancer, acute heart failure, and inflammation to name a few. In the present study, a dataset of 980 molecules with a significant diversity of structural scaffolds and composition was selected to develop a balanced QSAR model possessing high predictive capability and mechanistic interpretation. The model was built as per the OECD (Organisation for Economic Co-operation and Development) guidelines and fulfills the endorsed threshold values for different validation parameters ($R^2_{tr} = 0.76$, $Q^2_{LMO} = 0.76$, and $R^2_{ex} = 0.76$). The present QSAR analysis identified that anti-BRD-4 activity is associated with structural characters such as the presence of saturated carbocyclic rings, the occurrence of carbon atoms near the center of mass of a molecule, and a specific combination of planar or aromatic nitrogen with ring carbon, donor, and acceptor atoms. The outcomes of the present analysis are also supported by X-ray-resolved crystal structures of compounds with BRD-4. Thus, the QSAR model effectively captured salient as well as unreported hidden pharmacophoric features. Therefore, the present study successfully identified valuable novel pharmacophoric features, which could be beneficial for the future optimization of lead/hit compounds for anti-BRD-4 activity.

Keywords: QSAR; BRD-4; pharmacophoric features; X-ray

1. Introduction

Cancer and heart failure are major causes of mortality [1], health complications, and social and economic problems for millions of people around the globe. Researchers have identified different chemotherapeutic methods to minimise heart failure as well as the onset, growth, and survival of cancer cells [1]. However, different serious health issues initiated or echoed by different anti-cancer and cardiac drugs are of great concerns. Therefore, the quest for a harmless and effective anti-cancer and cardiac drug is an important goal for the research and development laboratories of pharmaceutical companies and academic institutions. For this, researchers generally prefer to inhibit any irregularity occurring during a vital cellular process. A good number of recent studies have confirmed that reversible lysine acetylation (RAL) is a dynamic process responsible for protein post-translational modifications, transcriptional activation, and other cellular processes [2–11]. Therefore, any anomaly with RAL could lead to the initiation of malignancy or its survival [7,12]. RAL is regulated by three types of epigenetic regulatory proteins [12,13]: (1) histone acetyltransferases (HATs) acetyllysine, (2) Histone deacetylases (HDACs), and (3) bromodomain
(BRD) family of proteins. HATs are responsible for acetylation of lysine residues on histone tails and thereby behave as “writers”, whereas the reverse is true for HDACs and sirtuins, which work as “erasers” that are accountable for the elimination of the acetyl group from acetylated lysine (KAc) [14]. The Bromodomain and Extra-terminal (BET) family selectively recognises and links with acetylated lysine residues in histones H3 and H4 [15–17]; thus, they function as “readers”. The BET proteins, viz., BRD2, BRD3, and BRD-4, and bromodomain testis-specific protein (BRDT) are widely recognised as druggable target proteins for regulating cellular epigenetics [15]. Therefore, intruding interactions between BET proteins and acetylated lysine have attracted many researchers to develop better therapeutics for various human diseases including cancer, acute heart failure, and inflammation [2–11].

BRD-4, also called mitotic chromosomal-associated protein (MCAP), Fshr4, or Hunk1, is ubiquitously expressed and plays a crucial role in a number of DNA-centered processes [15]. It is generally localised in the nucleus and regulates transcription by RNA polymerase II through a positive transcriptional elongation factor complex [15]. Structurally, it comprises two highly conserved N-terminal bromodomains (BD1 and BD2), an ET domain, and a C-terminal domain (CTD) [7]. Furthermore, BRD-4 contains a set of four helices: αZ, αA, αB, and αC. αZ and αA helices are connected through the ZA loop, whereas the BC loop connects the αB and αC helices [11,18]. Together, the four helices and the two loops create an active acetyl-lysine binding pocket (see Figure 1) [11,18]. The active site also consists of a hydrophobic WPF shelf (Trp81, Pro82, and Phe83), ZA loop, Tyr97, Asn140, and Met149 [11,18]. The majority of BRD-4 inhibitors compete with histone H4 to imitate the interactions with Tyr97 and Asn140 [3]. The WPF shelf is believed to play an important role in deciding the selectivity for BET bromodomains [6].

Recent studies indicate that the inhibition of BRD-4 is a good strategy, and a good number of BRD-4 inhibitors are in clinical or pre-clinical trials (see Figure 2) [2,19,20]. However, the quest for a safer and effective BRD-4 inhibitor with an optimum ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profile with a retention of potency is still in progress. For this, it is essential to know the prominent and concealed pharmacophoric features associated with BRD-4 inhibitors. To achieve this goal, a good number of researchers have reported SAR (Structure Activity Relationships) and QSAR (Quantitative SAR) analyses of BRD-4 inhibitors. Tahir et al. [21] developed a CoMSIA (3D-QSAR) model with an $R^2_{tr}$ (coefficient of determination) = 0.982 and $R^2_{cv}$ (or $Q^2_{loo}$) (cross-validated coefficient of determination for leave-one-out) = 0.500 for a dataset of
60 quinolinone and quinazolinone derivatives as BRD-4 inhibitors. Tong et al. [22] reported four 3D-QSAR models possessing $R^2_{tr} = 0.912$ to 0.963 and $R^2_{cv} = 0.574$ to 0.759 for the BRD-4 inhibitory activity of 4,5-dihydro-[1,2,4]triazolo[4,3-f]pteridine derivatives. Obadawo and co-workers [23] performed QSAR modelling ($R^2_{tr} = 0.93$ and $R^2_{cv} = 0.70$) for 40 different substituted 4-Phenylisoquinolinones as potent BET bromodomain (BRD-4-BD1) inhibitors. Speck-Planche and Scotti [6] performed multi-target QSAR for bromodomain inhibitors using linear discriminant analysis and artificial neural networks. Their binary classification (active/inactive), which is based on a fragment-based topological approach, and analysis led to the identification of a good number of pharmacophoric features. However, the fragment-based topological approach involved the use of SMILES of molecules and thereby lacks the inclusion of 3D information. Thus, even though these studies are successful in identifying easily visible pharmacophoric features, they are based on small datasets with limited variations in structures, binary classification, lack thorough validation and general applicability, and provide partial mechanistic interpretations.

A literature survey reveals that BRD-4 inhibitors possess structural isomerism (positional, chain, etc.), variations in central scaffolds, and their chemical space is very broad [6,7]; therefore, many concealed or hidden correlations of pharmacophoric features cannot be identified by visual inspection [24]. In such a situation, there is a need to accomplish thorough QSAR analysis using a larger dataset of BRD-4 inhibitors. In the present work, we have performed QSAR analysis of 980 structurally diverse BRD-4 inhibitors. The developed QSAR model possesses a balance of excellent predictive ability with in-depth mechanistic interpretations, which are reinforced by reported X-ray-resolved structures of BRD-4 inhibitors with the target enzyme.

2. Results

The present QSAR analysis is based on a dataset covering a broad chemical space and data range owing to the inclusion of structurally diverse compounds with experimentally measured $IC_{50}$ in the range of 1 nM to 15 µM. Consequently, this helped us in developing an appropriately validated genetic algorithm multi-linear regression (GA-MLR) model for gathering or extending exhaustive information about the pharmacophoric traits that govern the desired bio-activity (Descriptive QSAR) and also possessing acceptable external predictive capability (Predictive QSAR) [25–27]. The seven variable-based GA-MLR QSAR model (see model-A), along with selected internal and external validation parameters (see Supplementary Material for additional parameters), is as follows.

**Model-A: $pIC_{50} = 4.27 \pm 0.156 + 0.093 (\pm 0.017) \ast fsp3CringC2B + 0.108 (\pm 0.014) \ast com_C_4A + 0.391 (\pm 0.066) \ast Saturated_Carbo_Rings + 0.428 (\pm 0.036) \ast fsulfonSaroC8B + 0.625 (\pm 0.059) \ast flipoacc3B + 0.921 (\pm 0.067) \ast fsp3OaroN6B—0.367 (\pm 0.036) \ast fsulfonSaroC8B$.**

![Chemical structures of selected BRD-4 inhibitors.](image-url)
Validation of Model-A: Method of splitting = Random, No. of descriptors = 7, Ntraining = 785, Ntest = 195, R²tr = 0.762, R²adj. = 0.760, RMSEtr = 0.389, MAEtr = 0.326, CCCtr = 0.865, s = 0.391, F = 355.446, R²cv (Q²loo) = 0.757, RMSEcv = 0.393, MAEcv = 0.329, CCCcv = 0.862, Q²LMO = 0.756, R²Yscr = 0.009, Q²-F2 = 0.760, Q²-F3 = 0.758, CCCex = 0.860. A multitude of statistical validation parameters and analysis of associated graphs has been recommended by different researchers to confirm the statistical robustness and external prediction ability of a QSAR model [28–39]. The same approach has been followed in the present work. A high value of R²tr, R²adj., R²cv (Q²loo), R²ex, Q²-Fn, CCCex, etc., and a small value of LOF (lack-of-fit), RMSEtr, MAEtr, R²Yscr (R² for Y-scrambling), etc., along with different graphs (Figure 3a–d) related to model-A support the external predictive ability, statistical robustness, and point outs the lack of chancy correlation in the present work. A high value of R²tr, R²adj., R²cv (Q²loo), R²ex, Q²-Fn, CCCex, etc., and a small value of LOF (lack-of-fit), RMSEtr, MAEtr, R²Yscr (R² for Y-scrambling), etc., along with different graphs (Figure 3a–d) related to model-A support the external predictive ability, statistical robustness, and point outs the lack of chancy correlation in the present work.

The descriptions of seven molecular descriptors constituting model-A have been tabulated in Table 1.
Table 1. Details of molecular descriptors present in model-A.

| Variable           | Description                                                                 | Software Used for Calculation |
|--------------------|-----------------------------------------------------------------------------|--------------------------------|
| fsp3CringC2B       | Frequency of occurrence of ring carbon atoms exactly at 2 bonds from sp<sup>3</sup>-hybridised carbon atoms | PyDescriptor [45]              |
| com_C_4A           | Total number of carbon atoms within 4 Å from centre of mass (com) of molecule | PyDescriptor                   |
| Saturated_Carbo_Rings | Total number of saturated rings containing carbon atoms only                | DataWarrior [46]               |
| fsulfonSaroC8B     | Frequency of occurrence of aromatic carbon atoms exactly at 8 bonds from sulphur atoms of Sulfone group | PyDescriptor                   |
| fsp3OaroN6B        | Frequency of occurrence of aromatic nitrogen atoms exactly at 6 bonds from sp<sup>3</sup>-hybridised oxygen atoms | PyDescriptor                   |
| flipoacc3B         | Frequency of occurrence of H-bond acceptor atoms exactly at 3 bonds from lipophilic atoms | PyDescriptor                   |
| fplaNN4B           | Frequency of occurrence of nitrogen atoms exactly at 4 bonds from planer nitrogen atoms | PyDescriptor                   |

Interestingly, five molecular descriptors, viz., com_C_4A, fsp3CringC2B, flipoacc3B, fsulfonSaroC8B, and Saturated_Carbo_Rings, comprise the presence of different types of carbon atoms, which indicates the importance of carbon atoms in deciding BRD-4 inhibitory activity. The same is true for nitrogen, which is a part of three molecular descriptors, viz., flipoacc3B, fplaNN4B, and fsp3OaroN6B. Since, in general, the presence of carbon increases lipophilicity whereas nitrogen is attributed to significantly influence the pharmacological and hydrophilic profile, therefore, a balance of an appropriate number of carbons for lipophilicity and nitrogen is necessary to obtain adequate BRD-4 inhibitory activity. Of the seven descriptors in model-A, six have positive coefficients and only one has a negative coefficient. The effects of descriptors and their role in deciding the BRD-4 inhibitory profile have been discussed in more detail with relevant examples in the Discussion section.

3. Discussion
Mechanistic Interpretation of QSAR Model

An appropriately validated relationship between prominent structural features or molecular descriptors of the molecules with the bioactivity enlarges knowledge about mechanism of action of molecules, reasons for their specificity, and pharmacophoric atoms/groups accountable for the desired bioactivity [20,26,39]. In the present analysis, although we have equated the IC<sub>50</sub> values of different molecules in a relationship with a specific molecular descriptor (or feature), a synergistic or reverse effect of other molecular descriptors or unknown factors having a superseding influence in deciding the overall IC<sub>50</sub> value of a molecule cannot be ignored. That is, a single molecular descriptor or feature neither decides nor completely explains the experimental IC<sub>50</sub> value for such a large and structurally diverse set of molecules. In other words, the effective use of a validated QSAR model depends on the synergetic consideration of constituent molecular descriptors. The newly developed QSAR model-A comprises seven descriptors.

The molecular descriptor fsp3CringC2B represents the frequency of occurrence of ring carbon atoms exactly at two bonds from sp<sup>3</sup>-hybridised carbon atoms. If the same ring carbon atom was also present at less than two bonds from any other sp<sup>3</sup>-hybridised carbon atoms, then it was excluded while calculating fsp3CringC2B. Its positive coefficient in model-A and also a correlation of 0.30 with pIC<sub>50</sub> indicate that increasing such a combination of ring and sp<sup>3</sup>-hybridised carbon atoms could lead to better inhibitory activities for BRD-4. For example, a comparison of molecule 736 with 737 indicates the significant influence of ring carbon atoms (shown using green dots in Figure 4a,b) at exactly two bonds from sp<sup>3</sup>-hybridised carbon atoms. This is further supported by their reported
X-ray resolved structures with BRD-4. Molecule 736 (pdb: 5z1s [47]) has an additional water-mediated interaction with receptors with a distance of 3.37 Å (see Figure 4c) due to the -OCH$_3$ group present in the benzoxazinone ring. The same -OCH$_3$ group is responsible for increasing the value of $fsp3C_{ringC2B}$ for 736, but it is absent in 737 (pdb: 5z1r [47]). The difference in $IC_{50}$ for the following pairs of molecules further support the influence of $fsp3C_{ringC2B}$ on the activity profile: 255 with 499, 725 with 716, 231 with 240, and 89 with 105, to mention a few.

From this discussion, it appears that ring carbon atoms alone are important. However, replacing $fsp3C_{ringC2B}$ by number of ring carbon atoms as a descriptor in model-A reduced the statistical performance from $R^2_{tr} = 0.76$ to 0.73. In addition, the number of ring carbon atoms has a correlation of 0.27 with $pIC_{50}$. Therefore, $fsp3C_{ringC2B}$ is a better descriptor than the number of ring carbon atoms.

![Figure 4. Comparison of BRD-4 inhibitory activity of 736 with 737 with respect to $fsp3C_{ringC2B}$: (a,b) 2D representation of 736 and 737; (c,d) X-ray-resolved structures of 737 (pdb: 5z1r) and 736 (pdb: 5z1s). Cyan coloured spheres represent water molecules, and dashed lines signify interactions and distances in Å.](#)

$com\_C\_4A$, which stands for total number of carbon atoms within 4 Å from centre of mass (com) of molecule, has a positive coefficient in model-A. Therefore, increasing the value of $com\_C\_4A$ leads to better inhibitory activities. This observation is supported
by the fact that it has a correlation of 0.404 with pIC_{50}, and molecules with IC_{50} < 10 nM (29 molecules) possess a high value of com_C_{4A}. In addition, a simple comparison of the following pairs of the molecules strengthens this observation: 620 with 621, 720 with 710, 724 with 717, 526 with 518, and 691 with 692, and 595 with 596. At first glance, it looks as if com_C_{4A} is pointing out the importance of the number of carbon atoms. However, nC (number of carbon atoms) has a correlation of 0.29 with pIC_{50} and substituting com_C_{4A} by nC led to a decrease in statistical performance of model-A from R_{tr}^2 = 0.76 to 0.69. Therefore, com_C_{4A} is a better choice as a variable for model-A.

As the presence of carbon is generally associated with the increased lipophilicity of a molecule, therefore, com_C_{4A} indicates that the lipophilic part must be concentrated near the com of the molecule for better activities. This in turn provides a crucial hint about the active site of BRD-4. It appears that a significant portion of the active site of BRD-4 is reasonably lipophilic in nature. This is supported by the fact that the active site of BRD-4 consists of a hydrophobic WPF shelf (see Figure 1) [11,18]. Thus, the findings of the present QSAR analysis are supported by the reported X-ray-resolved structure of BRD-4 enzyme. In addition, the pharmacophore model, depicted in Figure 5, generated using most active molecule 297, again points out the presence of a lipophilic region near the com of the molecule.

Another descriptor that also points out the importance of lipophilicity of a molecule is flipoacc3B, which represents the frequency of occurrence of H-bond acceptor atoms exactly at three bonds from lipophilic atoms. However, an acceptor atom was excluded while calculating flipoacc3B if it is also present within two or less bonds from same or any other lipophilic atom. Evidently, the lipophilic part of a molecule close to H-bond acceptor moiety (O or N atoms) plays a crucial role in deciding the inhibitory effect for BRD-4. This is once again visible in the pharmacophore model depicted in Figure 6. A simple comparison of the following pairs of molecules supports this observation: 812 with 823, 7 with 8, and 4 with 10.

An easily interpretable and influential molecular descriptor is Saturated_Carbo_Rings, which corresponds to total number of saturated carbocyclic rings. It has positive coefficient in model-A; therefore, increasing such rings is beneficial. A comparison of IC_{50} for 411 with 384 (see Figure 6), 137 with 127, 73 with 67, 131 with 124, 60 with 61, 570 with 573, and 572, 230 with 247 is in favour of this observation.

The importance of Saturated_Carbo_Rings in model-A indicates that the lipophilicity and flexibility of a molecule are the actual factors governing an activity profile. It is noteworthy that clogP, which represents molecular lipophilicity, has a correlation of 0.193 with pIC_{50}, whereas Saturated_Carbo_Rings has 0.240. Thus, Saturated_Carbo_Rings is a better choice, as it pinpoints the specific feature or part of the molecule (saturated carbocyclic rings), which is correlated with the activity due to its lipophilic nature, whereas clogP is a molecular property. A plausible reason could be the crucial role played by hydrophobic zones such as saturated carbocyclic rings at the periphery or outer part of...
molecule as a proxy for BRD4 selectivity through their interaction with the WPF shelf [6]. Therefore, saturated carbocyclic rings should be retained in future optimizations for better activity profiles. Thus, the present work is successful in identifying the significance of saturated carbocyclic rings as a novel unreported pharmacophore feature associated with BRD-4 inhibitory activity.

![Depiction of influence of Saturated_Carbo_Rings on IC50 using molecule 384 and 411 as representative examples](image)

The molecular descriptor \( \text{f}_{\text{sulfonSaroC8B}} \) (frequency of occurrence of aromatic carbon atoms exactly at eight bonds from sulphur atoms of Sulfone (-SO2-) group) has a positive coefficient in model-A. Consequently, increasing the value of \( \text{f}_{\text{sulfonSaroC8B}} \), favours binding with BRD-4. It is to be noted that if the same aromatic carbon atom is also present at \( \leq 7 \) bonds from a sulphur atom of same or different Sulfone group through any path, then it was excluded while calculating \( \text{f}_{\text{sulfonSaroC8B}} \). Obviously, this descriptor signifies the importance of the Sulfone group (a highly polar group) and its correlation with aromatic rings (a lipophilic moiety) in deciding the binding with BRD-4. This is clearly reflected in the difference in the activity of the following pairs of molecules: 715 with 718, 723 with 724, 714 with 715, 707 with 716, 936 with 941, and 942 with 943. Recent studies indicate that Sulfone moiety is present in a cleft near the ZA-loop and establishes an H-bond with the -CONH- (amide) of backbone [19].

\( \text{fsp3OaroN6B} \) stands for the frequency of occurrence of aromatic nitrogen atoms exactly at six bonds from \( \text{sp}^3 \)-hybridised oxygen atom. If the same aromatic nitrogen atom is also present at \( \leq 5 \) bonds from the same or any other \( \text{sp}^3 \)-hybridised oxygen atom through any path, then it was excluded while calculating \( \text{fsp3OaroN6B} \). For example, 79 with 609, 81 with 620, and 614 with 615, to mention a few. In our previous study [20], we identified a similar descriptor \( \text{notringO_acc_6B} \) (total number of all non-ring Oxygen atoms present within a distance of six bonds from H-bond acceptor atoms) as an important pharmacophoric feature that governs the binding affinity (K_i) of a molecule for BRD-4. Thus, a consensus between the previous and the present study indicates that a molecule must have an H-bond acceptor (preferably aromatic nitrogen) at a distance of six bonds from a \( \text{sp}^3 \)-hybridised oxygen atom (non-ring oxygen favoured). This observation is supported by the difference in the activity of molecule S1, S2, and S3 [48] (see Figure 7). To add further, the \( \text{sp}^3 \)-hybridised oxygen atom is present as a linker between two aromatic rings or as an -OR (alkoxy) group in a good number of molecules [20].
The only descriptor with a negative coefficient in model-A is fplaNN4B, which corresponds to frequency of occurrence of nitrogen atoms exactly at 4 bonds from planer nitrogen atoms. If the same nitrogen atom is also present at \( \leq 3 \) bonds from the same or any other planer nitrogen atom through any path, then it was excluded while calculating fplaNN4B. The following pairs of molecules have a significant difference in their activities, which could be attributed to the presence of fplaNN4B as a structural feature: 945 with 954, 936 with 944, 411 with 385, 170 with 171, 722 with 714, 563 with 577, 737 with 734, 239 with 240. Therefore, such a combination of nitrogen atoms should be avoided to have better inhibition of BRD-4.

4. Materials and Methods

The present work follows the standard procedure recommended by OECD and different researchers to perform QSAR analysis [49–51]. All the software were used with default settings to obtain a QSAR model possessing a balance of predictive ability and mechanistic interpretation; however, some settings were changed, which have been reported at appropriate places. The different steps are as follows.

Step-1: Collection of data and curation: The present work commenced with the collection of a large dataset of 2026 experimentally tested BRD-4 inhibitors from a free and publicly available database BindingDB (https://www.bindingdb.org/bind/index.jsp, accessed on 16 March 2022). A QSAR analysis is significantly influenced by the quality of data, its composition, and its appropriate curation before further processing [50,52–54]. Therefore, in the next step, data curation was performed [55], which involved the removal of duplicate entries, organometallic compounds, salts, molecules with ambiguous \( IC_{50} \) values, etc. This reduced the dataset to 980 molecules only. The reduced dataset still consists of molecules with experimental \( IC_{50} \) (nM) in the range 1 nM to 15 \( \mu \)M and the presence of diverse scaffolds such as heterocyclic rings, positional isomers, stereoisomers, etc., enhancing the chemical space and consequently widening the applicability of the newly developed model. The SMILES (Simplified Molecular Input Line Entry System) notations, including experimental \( IC_{50} \) and pIC\(_{50}\) (\( = -\log_{10}[IC_{50}] \)) of all the molecules used in the present work, are available in Supplementary Materials. For the sake of convenience, representative examples have been presented in Figure 8 to depict the structural diversity of the current dataset.

In Table 2, five most and least active molecules have been included as examples only along with their SMILES notation: \( IC_{50} \) (nM) and pIC\(_{50}\) (M).

Step-2: In the next step, SMILES notations were used to develop the optimised 3D structures (semi-empirical PM3 method) of the molecules, accomplished using OpenBabel 2.4 [56] and MOPAC 2012 (openmopac.net) using default settings.

Step-3: A QSAR model achieves a balance of mechanistic interpretation and predictive ability if a good number of diverse molecular descriptors are calculated, followed by adequate pruning to reduce the chances of overfitting from noisy redundant descriptors [57]. The next step involved the calculation of myriad number of 1D- to 3D-molecular descriptors including experimental IC\(_{50}\) and pIC\(_{50}\) (\( = -\log_{10}[IC_{50}] \)) of all the molecules used in the present work. The only descriptor with a negative coefficient in model-A is fplaNN4B, which corresponds to frequency of occurrence of nitrogen atoms exactly at 4 bonds from planer nitrogen atoms. If the same nitrogen atom is also present at \( \leq 3 \) bonds from the same or any other planer nitrogen atom through any path, then it was excluded while calculating fplaNN4B. The following pairs of molecules have a significant difference in their activities, which could be attributed to the presence of fplaNN4B as a structural feature: 945 with 954, 936 with 944, 411 with 385, 170 with 171, 722 with 714, 563 with 577, 737 with 734, 239 with 240. Therefore, such a combination of nitrogen atoms should be avoided to have better inhibition of BRD-4.

4. Materials and Methods

The present work follows the standard procedure recommended by OECD and different researchers to perform QSAR analysis [49–51]. All the software were used with default settings to obtain a QSAR model possessing a balance of predictive ability and mechanistic interpretation; however, some settings were changed, which have been reported at appropriate places. The different steps are as follows.

Step-1: Collection of data and curation: The present work commenced with the collection of a large dataset of 2026 experimentally tested BRD-4 inhibitors from a free and publicly available database BindingDB (https://www.bindingdb.org/bind/index.jsp, accessed on 16 March 2022). A QSAR analysis is significantly influenced by the quality of data, its composition, and its appropriate curation before further processing [50,52–54]. Therefore, in the next step, data curation was performed [55], which involved the removal of duplicate entries, organometallic compounds, salts, molecules with ambiguous \( IC_{50} \) values, etc. This reduced the dataset to 980 molecules only. The reduced dataset still consists of molecules with experimental \( IC_{50} \) (nM) in the range 1 nM to 15 \( \mu \)M and the presence of diverse scaffolds such as heterocyclic rings, positional isomers, stereoisomers, etc., enhancing the chemical space and consequently widening the applicability of the newly developed model. The SMILES (Simplified Molecular Input Line Entry System) notations, including experimental \( IC_{50} \) and pIC\(_{50}\) (\( = -\log_{10}[IC_{50}] \)) of all the molecules used in the present work, are available in Supplementary Materials. For the sake of convenience, representative examples have been presented in Figure 8 to depict the structural diversity of the current dataset.

In Table 2, five most and least active molecules have been included as examples only along with their SMILES notation: \( IC_{50} \) (nM) and pIC\(_{50}\) (M).

Step-2: In the next step, SMILES notations were used to develop the optimised 3D structures (semi-empirical PM3 method) of the molecules, accomplished using OpenBabel 2.4 [56] and MOPAC 2012 (openmopac.net) using default settings.

Step-3: A QSAR model achieves a balance of mechanistic interpretation and predictive ability if a good number of diverse molecular descriptors are calculated, followed by adequate pruning to reduce the chances of overfitting from noisy redundant descriptors [57]. The next step involved the calculation of myriad number of 1D- to 3D-molecular descriptors for all molecules. For this, PyDescriptor [45] and DataWarrior [46] were used, which
Figure 8. Representative examples to indicate the structural scaffold diversity in the present dataset.

Table 2. SMILES notation and IC$_{50}$ (nM) and pIC$_{50}$ (M) of five most and least active molecules of the selected data set.

| SN | Ligand SMILES                                                                 | IC$_{50}$ (nM) | pIC$_{50}$ (M) |
|----|-------------------------------------------------------------------------------|----------------|----------------|
| 207| Cn1cc2-c3cc(CS(C)(=O)=O)ccc3N(Cc3c[nH]c(=O)c1c23)c1ccccn1                    | 1              | 9              |
| 692| Cn1cc2-c3cc(CS(C)(=O)=O)ccc3N(Cc3c[nH]c(=O)c1c23)c1=O)c1ccc(F)cc1           | 1.5            | 8.824          |
| 158| CCS(=O)(=O)c1cc2Occc(F)ccccc(CCCCOc3cc(=O)n(C)c3-c2c1                      | 2              | 8.699          |
| 570| COc1cc2c(c1-c1c(C)nc1C)_[nH]c1nc(C)nc(Nc3cc(nn3C)C3CC3)c21                | 2              | 8.699          |
| 688| CC(=O)N1CCc2c(C1)c(nn2C1CCOCC1)-c1cccc2c(C)cncc12                          | 14,000         | 4.854          |
| 242| CC(=O)N1CCc2c(C1)c(N1ccccc(C)c1)nn2C1CC1                                  | 15,000         | 4.824          |
| 385| Cc1noc(C)c1-c1ccc(C)c1S(=O)(=O)N1CCNCC1                                    | 15,000         | 4.824          |
| 634| CCCc1nc2c(cccc12)-c1nn(C2CCCOCC2)2CCN(Cc12)C(C)=O                           | 15,000         | 4.824          |
| 721| CNC(=O)c1ccn(C)c1(O)c2cccce12                                              | 15,000         | 4.824          |

4.1. Splitting the Data Set into Training and External Sets and Subjective Feature Selection (SFS)

For developing a useful QSAR model and its appropriate validation, it is essential to divide the dataset into training and external (also called prediction or test set) sets [50,52-54,59]. Consequently, in the present work, the dataset was randomly divided into training (80% = 785 molecules) and external (20% = 195 molecules) sets to minimise any bias. The only purpose of the training set was to choose suitable number of variables.
was considered optimal for model building as there was no improvement in the statistical performance of model upon the inclusion of additional molecular descriptors. The analysis posed, i.e., external validation of the model (Predictive QSAR). Genetic Algorithm (GA) and validated QSAR model finds its usage for QSAR-based virtual screening, lead/hit optimization, decision making, etc. The following validation parameters and their recommended threshold values are usually used to assess a model: \( R^2_{tr} \) (coefficient of determination) \( \geq 0.6; Q^2_{loo} \) (cross-validated coefficient of determination for leave-one-out) \( \geq 0.5; Q^2_{LMO} \) (cross-validated coefficient of determination for leave-many-out) \( \geq 0.6, R^2 \geq Q^2; R^2_{ex} \) (external coefficient of determination) \( \geq 0.6; \) CCC (Concordance Correlation Coefficient) \( \geq 0.80; Q^2-F^1 \geq 0.60; \) high values of external validation parameters \( R^2_{ex}, Q^2_{F1}, Q^2_{F2}, \) and \( Q^2_{FI}; \) with low values of \( R^2_{Yscr} \) (coefficient of determination for Y-randomization); RMSE (Root mean square error); MAE (Mean absolute error); \( \text{RMSE}_{tr} < \text{RMSE}_{cv}. \) The formulae for calculating these statistical parameters are available in Supplementary Materials. In the present analysis, a Williams plot was used to assess the applicability domain of the newly developed QSAR model.

4.2. Building Regression Model and Its Validation

Appropriate validations involving cross/inter validation, external validation, Y-randomization analysis, and applicability domain (Williams plot) are necessary to estimate the reliability and general applicability of a QSAR model. A properly validated QSAR model finds its usage for QSAR-based virtual screening, lead/hit optimization, decision making, etc. The following validation parameters and their recommended threshold values are usually used to assess a model: \( R^2_{tr} \) (coefficient of determination) \( \geq 0.6; Q^2_{loo} \) (cross-validated coefficient of determination for leave-one-out) \( \geq 0.5; Q^2_{LMO} \) (cross-validated coefficient of determination for leave-many-out) \( \geq 0.6, R^2 \geq Q^2; R^2_{ex} \) (external coefficient of determination) \( \geq 0.6; \) CCC (Concordance Correlation Coefficient) \( \geq 0.80; Q^2-F^1 \geq 0.60; \) high values of external validation parameters \( R^2_{ex}, Q^2_{F1}, Q^2_{F2}, \) and \( Q^2_{FI}; \) with low values of \( R^2_{Yscr} \) (coefficient of determination for Y-randomization); RMSE (Root mean square error); MAE (Mean absolute error); \( \text{RMSE}_{tr} < \text{RMSE}_{cv}. \) The formulae for calculating these statistical parameters are available in Supplementary Materials. In the present analysis, a Williams plot was used to assess the applicability domain of the newly developed QSAR model.
4.3. Pharmacophore Model

For pharmacophore modelling, the 3D-optimised structure of the most active molecule, 207, was selected. The model was generated using LIQUID [62,63], a free and easy to use PyMOL plugin, using default settings, except that the contour region was set to 3 for H-bond donor/acceptor and hydrophobic regions.

4.4. Other Experimental Details

The reported X-ray resolved structures (pdb 5z1r and 5z1s) were downloaded from Protein Data Bank (www.rcsb.org accessed on 13 April 2022). PyMOL version 2.4 has been used for the depiction of molecular interactions between the compounds and the protein.

5. Conclusions

In the present study, a seven-descriptor-based and rigorously validated GA–MLR QSAR model with $R^2_{tr} = 0.79$, $Q^2_{LMO} = 0.79$, and $R^2_{ex} = 0.78$ was derived to identify the significant pharmacophoric features that influence BRD-4 inhibitory activity. As mentioned earlier, it is essential to perceive salient and visually unrecognizable pharmacophoric features linked with BRD-4 inhibitory activity for different chemical scaffolds. The analysis indicates that the presence of ring carbon and nitrogen atoms, occurrence of carbon atoms near the center of mass of a molecule, specific combination of planer nitrogen with ring carbon, donor and acceptor atoms, etc., are prominent features to be retained in future optimizations. On the other hand, a combination of nitrogen atoms with planer nitrogen atoms exactly at four bonds should be avoided for better BRD-4 inhibitory activity. The reported crystal structures of BRD-4 inhibitors strengthen these observations. The present study efficaciously captured and reported novel pharmacophoric features and has a good balance of predictive ability and mechanistic interpretations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ph15060745/s1.

Author Contributions: Conceptualization, V.H.M., M.E.A.Z. and A.S.; formal analysis and data curation: V.H.M., R.D.J. and S.A.A.-H.; writing, M.E.A.Z., V.H.M. and A.A.A.-M.; revisions, M.E.A.Z., V.H.M. and A.S.; editing and proofreading, V.H.M., M.E.A.Z., R.D.J. and A.A.A.-M.; funding and resources, S.A.A.-H. and M.E.A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia, for its support of this research through research group number RG-21-09-77.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article and Supplementary Material.

Acknowledgments: V. H. Masand is thankful to Paola Gramatica (Italy) and her team for providing the free copy of QSARINS 2.2.4.

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

Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| SMILES       | Simplified molecular-input line-entry system |
| GA           | Genetic algorithm |
| MLR          | Multiple linear regression |
| QSAR         | Quantitative structure–activity relationship |
| WHO          | World Health Organization |
| OLS          | Ordinary least square |
| QSARINS      | QSAR Insubria |
| OECD         | Organisation for Economic Co-operation and Development |
References

1. Boer, R.A.; Meijers, W.C.; Meer, P.; Veldhuisen, D.J. Cancer and heart disease: Associations and relations. Eur. J. Heart Fail. 2019, 21, 1515–1525. [CrossRef] [PubMed]

2. Fidanze, S.D.; Liu, D.; Mantie, R.A.; Hasvold, L.A.; Pratt, J.K.; Sheppard, G.S.; Wang, L.; Holms, J.H.; Dai, Y.; Aguirre, A.; et al. Discovery and optimization of novel constrained pyrrolopyridone BET family inhibitors. Bioorg. Med. Chem. Lett. 2018, 28, 1804–1810. [CrossRef] [PubMed]

3. Guest, E.E.; Pickett, S.D.; Hirst, J.D. Structural variation of protein–ligand complexes of the first bromodomain of BRD4. Org. Biomol. Chem. 2021, 19, 5632–5641. [CrossRef] [PubMed]

4. Liu, Z.; Chen, H.; Wang, P.; Li, Y.; Wold, E.A.; Leonard, P.G.; Joseph, S.; Brasier, A.R.; Tian, B.; Zhou, J. Discovery of Orally Bioavailable Chromone Derivatives as Potent and Selective BRD4 Inhibitors: Scaffold Hopping, Optimization, and Pharmacological Evaluation. J. Med. Chem. 2020, 63, 5242–5256. [CrossRef] [PubMed]

5. Alqahtani, A.; Choucair, K.; Ashraf, M.; Hammouda, D.M.; Allogghi, A.; Khan, T.; Senzer, N.; Nemunaitis, J. Bromodomain and extra-terminal motif inhibitors: A review of preclinical and clinical advances in cancer therapy. Future Sci. OA 2019, 5, FSO372. [CrossRef] [PubMed]

6. Speck-Planche, A.; Scotti, M.T. BET bromodomain inhibitors: Fragment-based in silico design using multi-target QSAR models. Mol. Divers 2018, 23, 555–572. [CrossRef] [PubMed]

7. Duan, Y.; Guan, Y.; Qin, W.; Zhai, X.; Yu, B.; Liu, H. Targeting Brd4 for cancer therapy: Inhibitors and degraders. MedChemComm 2018, 9, 1779–1802. [CrossRef]

8. Zhao, Y.; Bai, L.; Liu, L.; McEanchr, D.; Stuckey, J.A.; Meagher, J.L.; Yang, C.-Y.; Sun, X.; Zhou, B.; Hu, Y.; et al. Structure-Based Discovery of 4-(6-Methoxy-2-methyl-4-(quinolin-4-yl)-9H-pyrimido[4,5-b]indol-7-yl)-3,5-dimethylisoxazole (CD161) as a Potent and Orally Bioavailable BET Bromodomain Inhibitor. J. Med. Chem. 2017, 60, 3887–3901. [CrossRef]

9. Xing, J.; Lu, W.; Liu, R.; Wang, Y.; Xie, Y.; Zhang, H.; Shi, Z.; Jiang, H.; Liu, Y.-C.; Chen, K.; et al. Machine-Learning-Assisted Approach for Discovering Novel Inhibitors Targeting Bromodomain-Containing Protein 4. J. Chem. Inf. Model. 2017, 57, 1677–1690. [CrossRef]

10. Guo, J.; Tong, J.-B.; Luo, D.; Feng, Y.; Bian, S.; Zhang, X.; Wang, T.-H. Structural modification of 4, 5-dihydro-[1, 2, 4] triazolo [4, 3-f] pteridine derivatives as BRD4 inhibitors using 2D/3D-QSAR and molecular docking analysis. Mol. Divers. 2021, 25, 1855–1872. [CrossRef] [PubMed]
50. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* 2014, 57, 4977–5010. [CrossRef]

51. Huang, J.; Fan, X. Why QSAR fails: An empirical evaluation using conventional computational approach. *Mol. Pharm.* 2011, 8, 600–608. [CrossRef]

52. 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.; et al. QSAR without borders. *Chem. Soc. Rev.* 2020, 49, 3525–3564. [CrossRef]

53. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* 2014, 54, 1–4. [CrossRef] [PubMed]

54. Martin, T.M.; Harten, P.; Young, D.M.; Muratov, E.N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* 2012, 52, 2570–2578. [CrossRef] [PubMed]

55. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 2010, 50, 1189–1204. [CrossRef] [PubMed]

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

57. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* 2008, 48, 1733–1746. [CrossRef]

58. 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]

59. Tropsha, A.; Gramatica, P.; Gombar, V.K. The Importance of Being Earnest Validation is the Absolute Essential for Successful Application and Interpretation of QSIR Models. *QSAR Comb. Sci.* 2003, 22, 69–77. [CrossRef]

60. Zaki, M.E.A.; Al-Hussain, S.A.; Masand, V.H.; Sabnani, M.K.; Samad, A. Mechanistic and Predictive QSAR Analysis of Diverse Molecules to Capture Salient and Hidden Pharmacophores for Anti-Thrombotic Activity. *Int. J. Mol. Sci.* 2021, 22, 8352. [CrossRef]

61. Gramatica, P. Principles of QSAR models validation internal and external. *QSAR Comb. Sci.* 2007, 26, 694–701. [CrossRef]

62. Yuan, S.; Chan, H.C.S.; Hu, Z. Using PyMOL as a platform for computational drug design. *WIREs Comput. Mol. Sci.* 2017, 7, e1298. [CrossRef]

63. Tanrikulu, Y.; Nietert, M.; Scheffer, U.; Proschak, E.; Grabowski, K.; Schneider, P.; Weidlich, M.; Karas, M.; Gobel, M.; Schneider, G. Scaffold hopping by “fuzzy” pharmacophores and its application to RNA targets. *ChemBioChem* 2007, 8, 1932–1936. [CrossRef] [PubMed]